Framework

AWS Framework Well-Architected

AWS Framework Well-Architected: Framework

Table of Contents

Framework AWS Well-Architected

Date de publication : 6 novembre 2024 (Révisions du document)

Le cadre Framework AWS Well-Architected vous aide à comprendre les avantages et les inconvénients des décisions que vous prenez lors du développement de systèmes sur AWS. En utilisant ce cadre, vous apprendrez les bonnes pratiques architecturales pour concevoir et exploiter des systèmes fiables, sécurisés, efficaces, économiques et durables dans le cloud.

Introduction

AWS Well-Architected Framework vous aide à comprendre les avantages et les inconvénients des décisions que vous prenez lors du développement de systèmes sur AWS. En utilisant ce framework, vous apprenez les bonnes pratiques architecturales en matière de conception et d'exploitation de charges de travail fiables, sécurisées, efficaces, économiques et durables dans le AWS Cloud. Il vous permet d'évaluer vos architectures par rapport aux bonnes pratiques et d'identifier les points à améliorer. Le processus de vérification d'une architecture est une discussion constructive autour de décisions architecturales. Ce n'est pas un mécanisme d'audit. Nous pensons que l'adoption de systèmes Well-Architected augmente considérablement les chances de réussite d'une entreprise.

Les architectes de solutions AWS ont des années d'expérience en architecture de produits sur une très grande variété de segments verticaux commerciaux et de cas d'utilisation. Nous avons contribué à la conception et à la vérification de milliers d'architectures client sur AWS. Sur la base de cette expérience, nous avons identifié les bonnes pratiques et les principales stratégies d'architecture de systèmes dans le Cloud.

Well-Architected Framework AWS documente un ensemble de questions de base afin de vous aider à évaluer si une architecture spécifique respecte bien les bonnes pratiques du cloud. Le cadre offre une approche cohérente pour évaluer les systèmes par rapport aux qualités attendues des systèmes modernes basés sur le Cloud, ainsi que les corrections requises pour atteindre ces qualités. Comme AWS évolue en permanence et que nous ne cessons d'apprendre en collaborant avec nos clients, nous continuerons à affiner la définition de « well-architected ».

Le présent outil est conçu pour ceux qui occupent des postes technologiques, comme les directeurs de la technologie, les architectes, les développeurs et les membres de l'équipe d'exploitation. Il décrit les stratégies et les bonnes pratiques AWS à utiliser lors de la conception et de l'exécution d'une charge de travail dans le cloud, et fournit des liens vers d'autres détails de mise en œuvre et modèles d'architecture. Pour plus d'informations, consultez la page d'accueil AWS Well-Architected.

AWS fournit également un service gratuit pour vérifier vos charges de travail. L'outil AWS Well-Architected Tool (AWS WA Tool) est un service dans le cloud qui fournit un processus uniforme pour qui vous aide à vérifier et mesurer votre architecture à l'aide du Cadre AWS Well-Architected. L'AWS WA Tool fournit des recommandations pour améliorer la fiabilité, la sécurité, l'efficacité et la rentabilité de vos charges de travail.

Pour vous aider à appliquer les bonnes pratiques, nous avons créé AWS Well-Architected Labs, qui vous fournit un référentiel de code et de documentation et vous offre une expérience concrète de la mise en œuvre des bonnes pratiques. Nous avons également fait équipe avec des partenaires du réseau de partenaires AWS sélectionnés, eux-mêmes membres du programme de partenariat AWS Well-Architected. Ces partenaires AWS disposent de connaissances approfondies sur AWS et peuvent vous aider à vérifier et améliorer vos charges de travail.

Définitions

Chaque jour, les experts AWS aident les clients à concevoir des systèmes afin de tirer parti des bonnes pratiques dans le cloud. Nous collaborons avec vous pour parvenir à des compromis architecturaux au fur et à mesure que vos conceptions évoluent. Lorsque vous déployez ces systèmes dans des environnements réels, nous découvrons leurs performances réelles ainsi que les conséquences de ces compromis.

Grâce aux enseignements acquis, nous avons créé AWS Well-Architected Framework, qui fournit un ensemble cohérent de bonnes pratiques pour les clients et les partenaires, afin d'évaluer les architectures. Il inclut également un ensemble de questions dont vous pouvez vous inspirer pour évaluer le degré de conformité d'une architecture aux bonnes pratiques AWS.

Le cadre AWS Well-Architected Framework repose sur six piliers, à avoir l'Excellence opérationnelle, la Sécurité, la Fiabilité, l'Efficacité des performances, l'Optimisation des coûts et la Durabilité.

Tableau 1 Les piliers d'AWS Well-Architected Framework

Name (Nom)	Description
Excellence opérationnelle	Capacité de soutenir le développement et de gérer efficacement les charges de travail, de recueillir des informations sur leurs opération s et d'améliorer continuellement les processus

Name (Nom)	Description
	et procédures de soutien afin de fournir de la valeur ajoutée.
Sécurité	Le pilier de sécurité décrit comment tirer parti des technologies cloud pour protéger les données, les systèmes et les actifs de manière à améliorer votre posture de sécurité.
Fiabilité	Le pilier Fiabilité englobe la capacité d'une charge de travail à exécuter sa fonction de manière correcte et cohérente et ce, en temps utile. Cela inclut la possibilité d'exploiter et de tester la charge de travail tout au long de son cycle de vie. Ce livre blanc fournit des bonnes pratiques détaillées pour la mise en œuvre de charges de travail fiables sur AWS.
Efficacité des performances	Capacité à utiliser efficacement les ressource s informatiques pour satisfaire aux exigences système et à maintenir cette efficacité au fur et à mesure que la demande change et que les technologies évoluent.
Optimisation des coûts	Capacité à exécuter des systèmes de façon à offrir une valeur business au niveau de prix le plus bas.
Durabilité	Capacité d'améliorer continuellement les impacts sur la durabilité via la réduction de la consommation d'énergie et l'amélioration de l'efficacité de tous les composants d'une charge de travail en maximisant les avantages des ressources allouées et en minimisant les ressources totales requises.

Dans le cadre AWS Well-Architected Framework, nous utilisons les termes suivants :

- Composant : code, configuration et ressources AWS qui, ensemble, répondent à une exigence. Un composant est souvent une unité de propriété technique. Il est découplé des autres composants.
- Le terme charge de travail est utilisé pour désigner un ensemble de composants qui collaborent pour apporter une valeur métier. La charge de travail représente généralement le niveau de détails dont discutent les responsables métier et techniques.
- Nous considérons l'architecture comme la façon dont les composants fonctionnent ensemble dans une charge de travail. Les schémas d'architecture se concentrent souvent sur la manière dont les composants communiquent et interagissent entre eux.
- Les étapes signalent les modifications importantes de votre architecture à mesure de son évolution tout au long du cycle de vie du produit (conception, mise en place, tests, préproduction et production).
- Au sein d'une organisation, le portefeuille technologique est l'ensemble des charges de travail qui sont nécessaires pour l'exécution des activités.
- Le niveau d'effort consiste à catégoriser le temps, les efforts et la complexité qu'une tâche nécessite pour sa mise en œuvre. Chaque organisation doit tenir compte de la taille et de l'expertise de l'équipe et de la complexité de la charge de travail comme contexte supplémentaire afin de catégoriser correctement son niveau d'effort.
  - Élevé : le projet peut prendre plusieurs semaines, voire plusieurs mois. Il pourrait être divisé en plusieurs scénarios, versions et tâches.
  - Moyen : le projet peut prendre plusieurs jours, voire plusieurs semaines. Il pourrait être divisé en plusieurs versions et tâches.
  - Faible : le projet peut prendre plusieurs heures, voire plusieurs jours. Il pourrait être divisé en plusieurs tâches.

Lorsque vous concevez des charges de travail, vous faites des compromis entre des piliers en fonction de votre contexte commercial. Ces décisions professionnelles peuvent orienter vos priorités en matière d'ingénierie. Vous pouvez opter pour l'optimisation afin d'améliorer l'impact sur la durabilité et de réduire les coûts au détriment de la fiabilité dans les environnements de développement, ou, pour les solutions stratégiques, vous pouvez optimiser la fiabilité avec des coûts plus élevés et un impact plus important sur la durabilité. Dans les solutions d'e-commerce, les performances peuvent affecter les revenus et la propension des clients à acheter les produits. La sécurité et l'excellence opérationnelle ne donnent généralement pas lieu à des compromis avec les autres piliers.

Sur l’architecture

Dans les environnements sur site, les clients possèdent souvent une équipe centrale pour l’architecture technologique, qui supervise les autres équipes dédiées aux produits ou aux fonctionnalités afin de vérifier qu’elles ont adopté les bonnes pratiques. Les équipes d’architecture technologique comptent généralement un ensemble de rôles, notamment : architecte technique (infrastructure), architecte de solutions (logiciel), architecte de données, architecte de mise en réseau et architecte de sécurité. Ces équipes utilisent souvent TOGAF ou le Zachman Framework dans le cadre d’une capacité d’architecture d’entreprise.

Chez AWS, nous préférons répartir les fonctionnalités entre plusieurs équipes dédiées plutôt que de ne recourir qu’à une seule équipe centralisée pour toutes les fonctionnalités. Il existe des risques lorsque vous choisissez de répartir les décisions. Par exemple, il faut vérifier que les équipes respectent les normes internes. Nous réduisons ces risques de deux manières. Tout d’abord, nous avons des pratiques (façons de procéder, processus, règles et autres normes acceptées) qui visent à permettre à chaque équipe de se charger de cette fonctionnalité comme il se doit, et nous avons recours à des experts afin de vérifier que les équipes dépassent les exigences. Ensuite, nous mettons en œuvre des mécanismes qui effectuent des vérifications automatisées pour vérifier que les normes sont respectées.

> « Les bonnes intentions ne suffisent pas, il faut de bons mécanismes pour tout rendre possible », Jeff Bezos.

Cela implique d’avoir recours à des mécanismes (souvent automatisés) qui vérifient la conformité aux règles ou aux processus plutôt que de faire appel à la bonne volonté des employés. Cette approche distribuée est soutenue par les principes de leadership d’Amazon et établit une culture à travers tous les rôles qui partent du client. Le travail à rebours est un élément fondamental de notre processus d’innovation. Nous commençons par les clients et ce dont ils ont besoin, afin de définir et de guider nos efforts. Les équipes obsédées par le client fabriquent des produits en s’appuyant sur les besoins des clients.

Pour l’architecture, cela signifie que nous prévoyons que chaque équipe soit capable de créer des architectures et de suivre les bonnes pratiques. Pour aider les nouvelles équipes à s’approprier ces fonctionnalités, ou les équipes existantes à mettre la barre plus haut, nous activons l’accès à une communauté virtuelle d’ingénieurs en chef qui peuvent vérifier leurs conceptions et les aider à comprendre ce que sont les bonnes pratiques AWS. La communauté des ingénieurs principaux

travaille pour rendre les bonnes pratiques visibles et accessibles. Une solution pourrait être, par exemple, d'organiser des discussions durant le déjeuner qui se concentrent sur l'application de bonnes pratiques à de véritables exemples. Ces discussions sont enregistrées et peuvent être utilisées dans le cadre de documents d'accueil pour les nouveaux membres de l'équipe.

Les bonnes pratiques AWS naissent de notre expérience dans l'exécution de milliers de systèmes à l'échelle d'Internet. Nous préférons utiliser des données pour définir les bonnes pratiques, mais utilisons également des experts fonctionnels en tant qu'ingénieurs principaux pour les définir. Lorsque les ingénieurs en chef voient l'émergence de nouvelles bonnes pratiques, ils collaborent afin de vérifier que les équipes les suivent. Avec le temps, ces bonnes pratiques sont officialisées dans nos processus d'évaluation internes, ainsi que dans les mécanismes qui renforcent la conformité. Le cadre Well-Architected est l'implémentation destinée aux clients de notre processus d'évaluation interne, où nous avons codifié notre réflexion sur l'ingénierie principale à travers les rôles tels que l'architecture de solutions et les équipes d'ingénieurs internes. Le cadre Well-Architected est un mécanisme évolutif qui vous permet de tirer parti de ces connaissances.

En suivant l'approche d'une communauté d'ingénierie principale avec une propriété d'architecture distribuée, nous pensons qu'une architecture d'entreprise Well-Architected reposant sur les besoins du client peut émerger. Nos leaders en matière de technologie (tels que les directeurs techniques ou les directeurs de développement), qui effectuent des évaluations Well-Architected sur toutes vos charges de travail, vous permettront de mieux comprendre les risques au sein de votre portefeuille de technologies. Cette approche vous permet d'identifier les thèmes, pour différentes équipes, que votre organisation pourrait aborder via les mécanismes, les formations, ou les discussions de midi où vos ingénieurs principaux peuvent partager leurs idées sur les domaines spécifiques avec plusieurs équipes.

Principes généraux de conception

Le cadre Well-Architected identifie un ensemble de principes généraux de conception destinés à faciliter la bonne conception dans le Cloud :

- Cessez de deviner vos besoins en capacité : si vous prenez une mauvaise décision en matière de capacité lors du déploiement d'une charge de travail, il se peut que vous vous retrouviez face à des ressources inutilisées onéreuses, ou que vous deviez traiter les implications relatives aux performances d'une capacité limitée. Grâce au cloud computing, vous n'avez plus de soucis à vous faire. Vous pouvez utiliser autant ou aussi peu de capacité que vous le souhaitez, et la mise à l'échelle se fait automatiquement.

- Tester les systèmes à l'échelle de la production : dans le cloud, vous pouvez créer un environnement de tests à l'échelle de la production et à la demande, exécuter les tests, puis mettre les ressources hors service. Puisque vous ne payez l'environnement de test que lorsqu'il s'exécute, vous pouvez simuler votre environnement réel pour une fraction du coût d'un test sur site.
- Automatiser en gardant à l'esprit l'expérimentation architecturale : l'automatisation vous permet de créer et de répliquer vos charges de travail à un coût peu élevé et d'éviter les frais de main-d'œuvre. Vous pouvez suivre les modifications apportées à l'automatisation, auditer l'impact et rétablir les paramètres antérieurs si nécessaire.
- Prenons l'exemple des architectures évolutives : dans un environnement traditionnel, les décisions d'architecture sont souvent mises en place comme des événements statiques et fixes, avec quelques versions majeures d'un système pendant sa durée de vie. Tandis que l'entreprise et son contexte continuent à évoluer, ces décisions initiales peuvent entraver la capacité du système à satisfaire des exigences métier variables. Dans le cloud, la capacité d'automatiser et de tester les éléments à la demande réduit le risque d'impact des modifications de conception. Les systèmes peuvent ainsi évoluer au fil du temps, de telle sorte que les entreprises peuvent tirer profit des innovations dans le cadre d'une pratique standard.
- Piloter les architectures à l'aide de données : dans le cloud, vous pouvez collecter des données sur la façon dont vos choix architecturaux affectent le comportement de votre charge de travail. Cela vous permet de prendre des décisions basées sur les faits sur la façon d'améliorer votre charge de travail. Votre infrastructure cloud est codée. Vous pouvez donc utiliser ces données pour alimenter vos choix architecturaux et des améliorations au fil du temps.
- S'améliorer au fil des jours de match : testez les performances de votre architecture et de vos processus en programmant régulièrement des tests de simulation de pannes, pour simuler des événements durant la production. Cela vous aidera à comprendre où apporter des améliorations et à développer une expérience de gestion des événements au sein de votre organisation.

Les cinq piliers du framework

La création d'un logiciel est similaire à celle d'un bâtiment. Si la fondation n'est pas solide, des problèmes structurels peuvent saper l'intégrité et la fonction du bâtiment. Lorsque vous concevez des solutions technologiques, si vous négligez les six piliers de l'excellence opérationnelle, à savoir la sécurité, la fiabilité, l'efficacité des performances, l'optimisation des coûts et la durabilité, il peut s'avérer difficile de créer un système qui répond à vos attentes et à vos exigences. Le fait d'intégrer ces domaines à votre architecture vous aidera à produire des systèmes stables et efficaces. Vous pouvez ainsi vous concentrer sur d'autres aspects de la conception, tels que les exigences fonctionnelles.

Piliers

- Excellence opérationnelle
- Sécurité
- Fiabilité
- Efficacité des performances
- Optimisation des coûts
- Durabilité

Excellence opérationnelle

L'excellence opérationnelle (OE) est un engagement à concevoir correctement un logiciel tout en offrant constamment une expérience client de qualité. Le pilier Excellence opérationnelle inclut les bonnes pratiques pour organiser votre équipe, concevoir votre charge de travail, la faire fonctionner à grande échelle et la faire évoluer au fil du temps.

Le pilier Excellence opérationnelle fournit une vue d'ensemble des principes de conception, des bonnes pratiques et des questions. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Excellence opérationnelle.

Rubriques

- Principes de conception
- Définition
- Bonnes pratiques

- Ressources

Principes de conception

Voici les principes de conception pour l'excellence opérationnelle dans le cloud :

- Organisation des équipes en fonction des résultats commerciaux : la capacité d'une équipe à atteindre des résultats commerciaux repose sur une vision du leadership, des opérations efficaces et un modèle opérationnel aligné sur l'entreprise. Les dirigeants doivent être pleinement investis et engagés dans la transformation des opérations dans le cloud (CloudOps) avec un modèle d'exploitation cloud adapté qui encourage les équipes à travailler de la manière la plus efficace possible et à atteindre les résultats commerciaux. Le bon modèle d'exploitation utilise les ressources humaines, les processus et les capacités technologiques pour évoluer, optimiser la productivité et se différencier grâce à l'agilité, à la réactivité et à l'adaptation. La vision à long terme de l'organisation se traduit par des objectifs qui sont communiqués au sein de l'entreprise aux parties prenantes et aux consommateurs de vos services cloud. Les objectifs et les indicateurs de performance clés (KPI) opérationnels sont harmonisés à tous les niveaux. Cette pratique maintient la valeur à long terme dérivée de la mise en œuvre des principes de conception suivants.
- Mise en œuvre de l'observabilité des informations exploitables : faites-vous une idée précise du comportement, des performances, de la fiabilité, des coûts et de l'état de la charge de travail. Établissez des indicateurs de performance clés (KPI) et tirez parti de la télémétrie de l'observabilité pour prendre des décisions éclairées et agir rapidement lorsque les résultats de l'entreprise sont menacés. Améliorez de manière proactive les performances, la fiabilité et les coûts sur la base de données d'observabilité exploitables.
- Automatisation sécurisée si possible : dans le cloud, vous pouvez appliquer la même discipline d'ingénierie que celle que vous utilisez pour le code d'application dans l'ensemble de l'environnement. Vous pouvez définir l'ensemble de votre charge de travail et de ses opérations (applications, infrastructure, configuration et procédures) sous forme de code et les mettre à jour. Vous pouvez ensuite automatiser les opérations de votre charge de travail en les lançant en réponse à des événements. Dans le cloud, vous pouvez utiliser la sécurité de l'automatisation en configurant des barrières de protection, notamment le contrôle du débit, les seuils d'erreur et les approbations. Grâce à une automatisation efficace, vous pouvez obtenir des réponses cohérentes aux événements, limiter les erreurs humaines et réduire la charge de travail des opérateurs.
- Réalisation de modifications fréquentes, légères et réversibles : concevez des charges de travail évolutives et à couplage faible pour permettre la mise à jour régulière des composants. Les techniques de déploiement automatisé associées à des modifications mineures et incrémentielles

réduisent le rayon d'impact et permettent de faire marche arrière plus rapidement en cas de problème. Cela renforce la confiance dans la possibilité d'apporter des modifications positives à votre charge de travail tout en maintenant la qualité et en s'adaptant rapidement à l'évolution des conditions du marché.

- Affinez fréquemment les procédures opérationnelles : au fur et à mesure que vous mettez à l'échelle vos charges de travail, faites évoluer vos opérations de manière appropriée. Tout en utilisant des procédures opérationnelles, cherchez le moyen de les améliorer. Passez régulièrement en revue les procédures et assurez-vous qu'elles sont efficaces et maîtrisées par les équipes. Lorsque des lacunes sont identifiées, actualisez les procédures en conséquence. Communiquez les mises à jour des procédures à toutes les parties prenantes et équipes. Transformez vos opérations en jeu pour partager les bonnes pratiques et former les équipes.
- Anticipation des défaillances : optimisez le succès opérationnel en élaborant des scénarios de défaillance afin de comprendre le profil de risque de la charge de travail et son impact sur les résultats de votre entreprise. Testez l'efficacité de vos procédures et la réponse de votre équipe face à ces défaillances simulées. Prenez des décisions éclairées pour gérer les risques ouverts identifiés lors de vos tests.
- Enseignements à tirer de l'ensemble des événements et métriques opérationnels : favorisez l'amélioration grâce aux leçons tirées de tous les événements et pannes liés aux opérations. Communiquez ce qui a été appris aux équipes et à l'ensemble de l'entreprise. Les enseignements tirés devraient mettre en lumière des données et des anecdotes sur la façon dont les opérations contribuent aux résultats commerciaux.
- Utilisation de services gérés : réduisez la charge opérationnelle en utilisant des services gérés AWS dans la mesure du possible. Élaborez des procédures opérationnelles autour des interactions avec ces services.

Définition

Il existe quatre domaines de bonnes pratiques pour l'excellence opérationnelle dans le cloud :

- Organisation
- Préparation
- Exploitation
- Évolution

La direction de votre organisation définit les objectifs opérationnels. Votre organisation doit comprendre les besoins et les priorités et les utiliser pour organiser et mener des travaux visant à soutenir l'obtention des résultats opérationnels. Votre charge de travail doit émettre les informations nécessaires pour la prendre en charge. La mise en œuvre de services permettant l'intégration, le déploiement et la distribution de votre charge de travail générera un flux accru de changements bénéfiques dans la production en automatisant les processus répétitifs.

Il peut exister des risques inhérents à l'exploitation de votre charge de travail. Vous devez comprendre ces risques et prendre une décision avisée lors de la mise en production. Vos équipes doivent pouvoir prendre en charge votre charge de travail. Les métriques économiques et opérationnelles dérivées des résultats commerciaux souhaités vous permettront de comprendre l'état de votre charge de travail et de vos activités opérationnelles, et de réagir aux incidents. Vos priorités évolueront en fonction des besoins de votre entreprise et des changements dans l'environnement de votre entreprise. Utilisez-les comme une boucle de rétroaction afin d'améliorer continuellement votre organisation et le fonctionnement de votre charge de travail.

Bonnes pratiques

> **Note**
>
> Toutes les questions relatives à l'excellence opérationnelle ont le préfixe OPS comme raccourci du pilier.

Rubriques

- Organisation
- Préparation
- Exploitation
- Évolution

Organisation

Vos équipes doivent avoir une compréhension commune de l'ensemble de votre charge de travail, de leur rôle dans celle-ci et de leurs objectifs économiques communs afin de fixer les priorités qui permettent la réussite de l'entreprise. Des priorités bien définies maximiseront les bénéfices tirés de vos efforts. Évaluez les besoins des clients internes et externes en impliquant les principales parties prenantes, notamment les équipes commerciales, de développement et d'exploitation, afin

de déterminer où il est nécessaire de concentrer les efforts. L'évaluation des besoins des clients vous permet de vous assurer que vous avez une compréhension approfondie du soutien nécessaire pour atteindre les résultats économiques. Assurez-vous de connaître les lignes directrices ou les obligations définies par la gouvernance de votre entreprise, ainsi que les facteurs externes, tels que les exigences de conformité réglementaire et les normes sectorielles, qui peuvent imposer un objectif spécifique ou mettre l'accent sur ce dernier. Vérifiez que vous disposez de mécanismes permettant d'identifier les changements apportés à la gouvernance interne et aux exigences de conformité externe. Si aucune exigence n'est identifiée, assurez-vous d'avoir effectué les vérifications préalables dans cette détermination. Revoyez régulièrement vos priorités afin qu'elles puissent être mises à jour en fonction de l'évolution des besoins.

Évaluez les menaces pesant sur l'entreprise (par exemple, les risques et les responsabilités de l'entreprise, et les menaces sur la sécurité des informations) et conservez ces informations dans un registre des risques. Évaluez l'impact des risques et les compromis entre des intérêts concurrents ou des approches alternatives. Par exemple, l'accélération de la mise sur le marché de nouvelles fonctionnalités peut être privilégiée par rapport à l'optimisation des coûts, ou vous pouvez choisir une base de données relationnelle pour les données non relationnelles afin de simplifier la migration d'un système sans restructuration. Gérez les avantages et les risques afin de prendre des décisions éclairées lorsqu'il s'agit de déterminer où il est nécessaire de concentrer les efforts. Certains risques ou choix peuvent être acceptables pendant un certain temps, il peut être possible d'atténuer les risques associés, ou il peut devenir inacceptable de laisser un risque subsister, auquel cas vous prendrez des mesures pour y remédier.

Vos équipes doivent comprendre leur rôle dans l'obtention des résultats de l'entreprise. Les équipes doivent comprendre leur rôle dans la réussite des autres équipes, le rôle des autres équipes dans leur réussite, et avoir des objectifs communs. Comprendre la responsabilité, la manière dont les décisions sont prises et qui a le pouvoir de prendre des décisions vous aide à concentrer les efforts et à maximiser les avantages de vos équipes. Les besoins d'une équipe seront déterminés par le client qu'elle soutient, son organisation, la composition de l'équipe et les caractéristiques de sa charge de travail. Il n'est pas raisonnable de s'attendre à ce qu'un modèle d'exploitation unique puisse soutenir toutes les équipes et leurs charges de travail dans votre entreprise.

Assurez-vous qu'il existe des propriétaires identifiés pour chaque application, charge de travail, plateforme et composant d'infrastructure, et que chaque processus et procédure a un propriétaire identifié responsable de sa définition, et des propriétaires responsables de leur performance.

La compréhension de la valeur ajoutée de chaque composant, processus et procédure, de la raison pour laquelle ces ressources sont en place ou ces activités exécutées, et de la raison pour laquelle

cette propriété existe, éclaire les actions des membres de votre équipe. Définissez clairement les responsabilités des membres de l'équipe afin qu'ils puissent agir de manière appropriée et disposer de mécanismes permettant d'identifier la responsabilité et la propriété. Mettez en œuvre des mécanismes permettant de demander des ajouts, des modifications et des exceptions afin de ne pas entraver l'innovation. Définissez des accords entre les équipes décrivant la manière dont elles travaillent ensemble pour se soutenir mutuellement et soutenir les résultats de votre entreprise.

Fournissez un soutien aux membres de votre équipe afin qu'ils puissent être plus efficaces dans leur action et soutenir les résultats de votre entreprise. Les dirigeants engagés doivent fixer des attentes et mesurer le succès. Les principaux dirigeants devraient être le parrain, l'avocat et le moteur de l'adoption des bonnes pratiques et de l'évolution de l'organisation. Permettez aux membres de l'équipe d'agir lorsque les résultats sont menacés afin de minimiser l'impact et de les encourager à remonter jusqu'aux décideurs et aux parties prenantes lorsqu'ils estiment qu'il existe un risque afin de pouvoir le traiter et éviter les incidents. Fournissez en temps utile des communications claires et exploitables sur les risques connus et les événements prévus afin que les membres de l'équipe puissent prendre des mesures appropriées en temps opportun.

Encouragez l'expérimentation pour accélérer la formation et maintenir l'intérêt et l'engagement des membres de l'équipe. Les équipes doivent développer leurs compétences pour adopter les nouvelles technologies, et pour soutenir l'évolution des besoins et des responsabilités. Soutenez et encouragez cette démarche en accordant du temps structurel à la formation. Assurez-vous que les membres de votre équipe disposent des ressources, à la fois des outils et des membres de l'équipe, nécessaires à la réussite et à la mise à l'échelle pour soutenir les résultats de l'entreprise. Exploitez la diversité inter-organisationnelle pour rechercher des perspectives multiples et uniques. Utilisez cette perspective pour accroître l'innovation, remettre en question vos hypothèses et réduire le risque de biais de confirmation. Développez l'inclusion, la diversité et l'accessibilité au sein de vos équipes afin d'obtenir des perspectives bénéfiques.

Si des exigences réglementaires ou de conformité externes s'appliquent à votre organisation, vous devez utiliser les ressources fournies par AWS Cloud Compliance pour former vos équipes afin qu'elles puissent déterminer l'impact sur vos priorités. Le cadre Well-Architected met l'accent sur la formation, la mesure et l'amélioration. Il offre une approche cohérente pour évaluer les architectures et mettre en œuvre des conceptions qui seront mises à l'échelle dans le temps. AWS fournit l'outil AWS Well-Architected Tool pour vous aider à vérifier votre approche avant le développement et l'état de vos charges de travail avant et pendant la production. Vous pouvez comparer les charges de travail aux bonnes pratiques architecturales AWS les plus récentes, surveiller leur état global et obtenir des informations sur les risques. AWS Trusted Advisor est un outil qui permet d'accéder à un ensemble de vérifications de base qui recommandent des optimisations pouvant vous aider

à définir vos priorités. Les clients du Business and Enterprise Support ont accès à des contrôles supplémentaires axés sur la sécurité, la fiabilité, les performances, l'optimisation des coûts et la durabilité qui peuvent les aider à définir leurs priorités.

AWS peut vous aider à former vos équipes à AWS et à ses services afin qu'elles comprennent mieux comment leurs choix peuvent avoir un impact sur votre charge de travail. Utilisez les ressources fournies par AWS Support (Centre de connaissances AWS, forums de discussion AWS et AWS Support Center) et la documentation AWS pour former vos équipes. Contactez AWS Support via AWS Support Center pour obtenir des réponses à vos questions AWS. AWS partage également les bonnes pratiques et les modèles que nous avons appris grâce à l'exploitation d'AWS dans Amazon Builders' Library. Un grand nombre d'autres informations utiles sont disponibles sur le blog AWS et sur le podcast AWS officiel. AWS Training and Certification offre une formation par le biais de cours en ligne d'autoformation sur les principes fondamentaux d'AWS. Vous pouvez également vous inscrire à une formation dirigée par un formateur afin de soutenir le développement des compétences AWS de vos équipes.

Utilisez des outils ou des services qui permettent de gérer de manière centralisée vos environnements dans plusieurs comptes, comme AWS Organizations, pour gérer vos modèles d'exploitation. Des services tels que AWS Control Tower élargissent cette fonctionnalité de gestion en vous permettant de définir des plans (soutenant vos modèles d'exploitation) pour la configuration des comptes, d'appliquer une gouvernance continue en utilisant AWS Organizations et d'automatiser l'allocation de nouveaux comptes. Les fournisseurs de services gérés tels que AWS Managed Services, les partenaires AWS Managed Services ou les fournisseurs de services gérés du réseau de partenaires AWS offrent une expertise dans la mise en œuvre des environnements cloud et soutiennent vos exigences de sécurité et de conformité, ainsi que vos objectifs métier. L'ajout de services gérés à votre modèle d'exploitation peut vous faire gagner du temps et économiser des ressources, et vous permet de maintenir vos équipes internes réduites et concentrées sur les résultats stratégiques qui différencieront votre entreprise, plutôt que de développer de nouvelles compétences et capacités.

Les questions suivantes sont axées sur ces quelques considérations relatives à l'excellence opérationnelle. (Pour obtenir la liste des questions et des bonnes pratiques d'excellence opérationnelle, consultez l'annexe.)

OPS 1 : comment déterminer vos priorités ?

Chacun doit comprendre le rôle qu'il joue dans la réussite de l'entreprise. Établissez des objectifs partagés afin de définir des priorités pour les ressources. Cela permet de maximiser le fruit de vos efforts.

OPS 2 : comment structurer votre entreprise pour soutenir les résultats métier ?

Vos équipes doivent comprendre leur rôle dans l'obtention des résultats de l'entreprise. Les équipes doivent comprendre leur rôle dans la réussite des autres équipes, le rôle des autres équipes dans leur réussite, et avoir des objectifs communs. La compréhension de la responsab ilité, de la propriété de la manière dont les décisions sont prises et qui a le pouvoir de prendre des décisions vous aide à concentrer les efforts et à maximiser les avantages de vos équipes.

OPS 3 : comment votre culture organisationnelle soutient-elle vos résultats commerciaux ?

Offrez du soutien aux membres de votre équipe afin qu'ils puissent agir plus efficacement et soutenir les résultats commerciaux.

Vous pouvez décider à un moment donné de mettre l'accent sur un petit sous-ensemble de vos priorités. Utilisez une approche équilibrée sur le long terme pour garantir le développement des capacités nécessaires et de la gestion des risques. Vérifiez régulièrement les priorités opérationnelles et mettez-les à jour en fonction de l'évolution de vos besoins. Lorsque la responsabilité et la propriété sont indéfinies ou inconnues, vous risquez à la fois de ne pas effectuer les actions nécessaires en temps utile et de déployer des efforts redondants et potentiellement conflictuels pour répondre à ces besoins. La culture organisationnelle a un impact direct sur la satisfaction professionnelle et la fidélisation des membres de l'équipe. Stimulez l'engagement et l'exploitation des capacités des membres de votre équipe pour assurer la réussite de votre entreprise. L'expérimentation est nécessaire pour que l'innovation se produise et transforme les idées en résultats. Admettez qu'un résultat non désiré est une expérience positive qui a identifié un chemin qui ne mène pas au succès.

Préparation

Pour vous préparer à l'excellence opérationnelle, il est nécessaire de comprendre vos charges de travail et les comportements attendus. Vous pourrez ensuite les concevoir pour fournir des informations sur leur statut et créer les procédures nécessaires pour les prendre en charge.

Concevez votre charge de travail de manière à ce qu'elle vous fournisse les informations nécessaires pour comprendre son état interne (par exemple, les mesures, les journaux, les événements et les traces) dans tous ses composants à des fins d'observabilité et de résolution des problèmes. L'observabilité va au-delà de la simple surveillance. Elle fournit une compréhension complète du fonctionnement interne d'un système sur la base de ses résultats externes. Enracinée dans les métriques, les journaux et les données de suivi, l'observabilité propose des informations approfondies sur le comportement et la dynamique du système. Grâce à une observabilité efficace, les équipes peuvent identifier les modèles, les anomalies et les tendances, ce qui leur permet de résoudre les problèmes potentiels de manière proactive et de maintenir un état optimal du système. L'identification des indicateurs clés de performance (KPI) est essentielle pour garantir l'alignement entre les activités de surveillance et les objectifs commerciaux. Cet alignement garantit que les équipes prennent des décisions basées sur les données en utilisant des indicateurs réellement importants, optimisant à la fois les performances du système et les résultats commerciaux. En outre, l'observabilité permet aux entreprises d'être proactives plutôt que réactives. Les équipes peuvent comprendre les relations de cause à effet au sein de leurs systèmes, prévoir et prévenir les problèmes au lieu de simplement y réagir. À mesure que les charges de travail évoluent, il est essentiel de revoir et d'affiner la stratégie d'observabilité, afin de s'assurer qu'elle reste pertinente et efficace.

Adoptez des approches qui améliorent le flux des changements en production et qui permettent la restructuration, un retour d'information rapide sur la qualité et la correction des bugs. Ces approches accélèrent l'entrée des modifications bénéfiques dans l'environnement de production, limitent les problèmes déployés et permettent d'identifier et de corriger rapidement les problèmes introduits par les activités de déploiement ou découverts dans vos environnements.

Adoptez des approches qui fournissent un retour d'information rapide sur la qualité et permettent une reprise rapide à la suite de changements qui n'offrent pas les résultats escomptés. L'utilisation de ces pratiques diminue l'impact des problèmes découlant du déploiement des modifications. Prévoyez les modifications qui échouent afin de pouvoir réagir plus rapidement si nécessaire, et testez et validez les changements que vous apportez. Tenez compte des activités planifiées dans vos environnements afin de pouvoir gérer le risque des modifications affectant les activités planifiées. Mettez l'accent sur les modifications fréquentes, minimes et réversibles pour limiter leur portée. Ainsi, vous facilitez

la résolution des problèmes et les corrections avec la possibilité d'annuler une modification. Cela signifie également que vous pouvez tirer profit plus souvent de modifications importantes.

Évaluez l'état de préparation opérationnelle de votre charge de travail, de vos processus, de vos procédures et de votre personnel afin de comprendre les risques opérationnels liés à votre charge de travail. Utilisez un processus cohérent (y compris des listes de contrôle manuelles ou automatisées) pour déterminer quand vous êtes prêt à mettre en service votre charge de travail ou un changement. Cela vous permet également d'identifier tous les domaines d'amélioration nécessaire. Dotez-vous de dossiers d'exploitation qui documentent vos activités de routine, et de playbooks qui guident vos processus pour la résolution des problèmes. Déterminez les avantages et les risques afin de prendre des décisions éclairées pour autoriser les changements dans l'environnement de production.

AWS vous permet de visualiser l'ensemble de votre charge de travail (applications, infrastructure, politique, gouvernance et opérations) en tant que code. Cela signifie que vous pouvez appliquer la même discipline d'ingénierie que celle que vous utilisez pour le code d'application à chaque élément de votre pile et partager ces éléments entre les équipes ou les organisations afin d'amplifier les avantages des efforts de développement. Utilisez les opérations en tant que code dans le cloud et testez-les en toute sécurité pour développer votre charge de travail, vos procédures opérationnelles et la pratique de l'échec. L'utilisation de AWS CloudFormation vous permet de disposer d'environnements de test (sandbox), de développement, de test autres que sandbox et de production cohérents et modélisés, avec des niveaux de contrôle des opérations toujours plus élevés.

Les questions suivantes sont axées sur ces quelques considérations relatives à l'excellence opérationnelle.

OPS 4 : comment mettre en œuvre l'observabilité dans votre charge de travail ?

Intégrez l'observabilité à votre charge de travail afin de comprendre son état et de prendre des décisions basées sur les données en fonction des exigences de l'entreprise.

OPS 5 : Comment réduire les défauts, faciliter les corrections et améliorer le flux dans la productio n ?

Adoptez des approches qui améliorent l'entrée des modifications en production et qui permettent la refactorisation, un retour rapide sur la qualité et la correction de bogues. Cela permet d'accélér er l'entrée des modifications bénéfiques en production, de limiter le déploiement de problèmes et d'identifier et de corriger rapidement les problèmes introduits par les activités de déploiement.

OPS 6 : comment réduire les risques liés au déploiement ?

Adoptez des approches qui fournissent un retour d'information rapide sur la qualité et permettent une reprise rapide à la suite de changements qui n'offrent pas les résultats escomptés. L'utilisation de ces pratiques diminue l'impact des problèmes découlant du déploiement des modifications.

OPS 7 : comment savoir si vous êtes prêt à gérer une charge de travail ?

Évaluez la disponibilité opérationnelle de votre charge de travail, des processus et des procédure s, ainsi que le personnel pour comprendre les risques opérationnels liés à votre charge de travail.

Investissez dans la mise en œuvre des activités opérationnelles en tant que code pour maximiser la productivité du personnel opérationnel, minimiser les taux d'erreur et automatiser les réponses. Adoptez des « pre-mortems » pour anticiper les défaillances, et créez des procédures si nécessaire. Appliquez des métadonnées à l'aide des balises de ressource et de AWS Resource Groups en suivant une stratégie de balisage cohérente pour permettre l'identification de vos ressources. Balisez vos ressources pour l'organisation, la comptabilité analytique, les contrôles d'accès et le ciblage de l'exécution des activités d'opérations automatisées. Adoptez des pratiques de déploiement qui tirent parti de l'élasticité du cloud pour faciliter les activités de développement, et le prédéploiement des systèmes pour accélérer les mises en œuvre. Lorsque vous apportez des modifications aux listes de contrôle que vous utilisez pour évaluer votre charge de travail, planifiez les opérations que vous allez exécuter pour les systèmes en service qui ne sont plus conformes.

Exploitation

L'observabilité vous permet de vous concentrer sur les données pertinentes et de comprendre les interactions et les résultats de votre charge de travail. En vous concentrant sur les informations essentielles et en éliminant les données inutiles, vous maintenez une approche simple pour comprendre les performances des charges de travail. Il est essentiel non seulement de collecter des données, mais également de les interpréter correctement. Définissez des bases de référence claires, spécifiez des seuils d'alerte appropriés et surveillez activement tout écart. Un changement au niveau d'une métrique clé, en particulier lorsqu'elle est corrélée à d'autres données, contribue à identifier des problèmes spécifiques. Grâce à l'observabilité, vous êtes mieux équipé pour prévoir et relever les défis potentiels, veillant ainsi à ce que votre charge de travail fonctionne sans heurts et réponde aux besoins de l'entreprise.

Le bon fonctionnement d'une charge de travail se mesure à l'aune des résultats obtenus par les entreprises et les clients. Définissez les résultats attendus, déterminez comment le succès sera mesuré et identifiez les paramètres qui seront utilisés dans ces calculs pour déterminer le succès de votre charge de travail et des opérations. L'état opérationnel comprend à la fois l'état de la charge de travail et l'état et le succès des activités opérationnelles menées pour soutenir la charge de travail (par exemple, déploiement et réponse aux incidents). Établissez des métriques de référence pour l'amélioration, l'investigation et l'intervention, collectez et analysez vos métriques, puis validez votre compréhension du succès des opérations et de leur évolution dans le temps. Utilisez les métriques collectées pour déterminer si vous satisfaites vos clients et vos besoins commerciaux, et pour identifier les points à améliorer.

Une efficacité opérationnelle et une gestion efficace des événements sont requises pour atteindre une excellence opérationnelle. Cela s'applique à la fois aux événements opérationnels planifiés et imprévus. Utilisez les dossiers d'exploitation établis pour les événements bien compris, et utilisez les playbooks pour faciliter l'investigation et la résolution des problèmes. Priorisez les réponses aux événements en fonction de leur impact sur l'entreprise et les clients. Assurez-vous que, si une alerte est générée en réponse à un événement, il existe un processus associé à exécuter, avec un propriétaire spécifiquement identifié. Définissez à l'avance le personnel requis pour résoudre un événement et inclure des processus de remontée pour engager du personnel supplémentaire, si nécessaire, en fonction de l'urgence et de l'impact. Identifiez et engagez des personnes habilitées à prendre une décision sur les mesures à prendre lorsqu'une réponse à un événement non traité auparavant a un impact opérationnel.

Communiquez l'état opérationnel des charges de travail au moyen de tableaux de bord et de notifications adaptés au public cible (par exemple, clients, entreprises, développeurs, opérations) afin qu'il puisse prendre les mesures appropriées, que leurs attentes soient gérées et qu'il soit informé lorsque les opérations normales reprennent.

Dans AWS, vous pouvez générer des vues de tableau de bord de vos métriques collectées à partir des charges de travail et nativement depuis AWS. Vous pouvez tirer profit de CloudWatch ou d'applications tierces pour regrouper et présenter des vues d'activités au niveau de l'entreprise, de la charge de travail ou des opérations. AWS fournit des informations sur les charges de travail par le biais de fonctionnalités de journalisation, notamment AWS X-Ray, CloudWatch, CloudTrail et les journaux de flux VPC pour identifier les problèmes de charges de travail en soutien à l'analyse des causes premières et à la résolution.

Les questions suivantes sont axées sur ces quelques considérations relatives à l'excellence opérationnelle.

OPS 8 : comment exploiter l'observabilité de la charge de travail dans votre organisation ?

Garantissez un état optimal de la charge de travail en tirant parti de l'observabilité. Utilisez des métriques, des journaux et des données de suivi pertinents pour obtenir une vue complète des performances de votre charge de travail et résoudre les problèmes de manière efficace.

OPS 9 : comment comprendre l'état de vos opérations ?

Définissez, capturez et analysez les métriques des opérations pour obtenir une visibilité sur les événements opérationnels afin de pouvoir prendre des mesures appropriées.

OPS 10 : comment gérer les événements relatifs aux charges de travail et aux opérations ?

Préparez et validez des procédures de réponse aux événements afin de réduire leur effet disruptif sur votre charge de travail.

Toutes les métriques que vous recueillez doivent être alignées sur un besoin métier et les résultats qu'elles prennent en charge. Développez des réponses scriptées aux événements bien compris et automatisez leur exécution en réponse à la reconnaissance de l'événement.

Évolution

Apprenez, partagez et progressez continuellement pour maintenir l'excellence opérationnelle. Consacrez des cycles de travail à la réalisation quasi continuelle d'améliorations supplémentaires. Effectuez une analyse post-incident de tous les événements ayant un impact sur le client. Identifiez les facteurs contributifs et les mesures préventives pour limiter ou empêcher la récurrence. Communiquez les facteurs contributifs aux communautés concernées, le cas échéant. Évaluez régulièrement et priorisez les possibilités d'amélioration (par exemple, les demandes de fonctionnalités, la correction des problèmes et les exigences de conformité), y compris la charge de travail et les procédures opérationnelles.

Introduisez des boucles de rétroaction au sein de vos procédures pour identifier rapidement les domaines d'amélioration et tirer des enseignements de l'exécution d'opérations.

Partagez les leçons retenues et leurs avantages entre les équipes. Analysez les tendances dans les leçons apprises et effectuez une analyse rétrospective entre les équipes des métriques des opérations pour identifier les opportunités et les méthodes d'amélioration. Mettez en œuvre les changements destinés à apporter des améliorations et évaluez les résultats pour déterminer le succès.

Sur AWS, vous pouvez exporter vos données de journal vers Amazon S3 ou envoyer les journaux directement vers Amazon S3 pour un stockage longue durée. Avec AWS Glue, vous pouvez découvrir et préparer vos données de journaux dans Amazon S3 pour l'analytique et stocker les métadonnées associées dans AWS Glue Data Catalog. Grâce à son intégration native à AWS Glue, vous pouvez ensuite utiliser Amazon Athena pour analyser vos données de journaux, en les interrogeant à l'aide de SQL standard. En utilisant un outil de veille économique comme Amazon QuickSight, vous pouvez visualiser, explorer et analyser vos données. Découvrez les tendances et les événements d'intérêt qui peuvent entraîner une amélioration.

La question suivante est axée sur ces considérations relatives à l'excellence opérationnelle.

OPS 11 : comment faire évoluer les opérations ?

Consacrez du temps et des ressources à l'amélioration incrémentielle presque continue pour contribuer à l'évolution de l'efficacité et de l'efficience de vos opérations.

L'évolution réussie des opérations repose sur de fréquentes améliorations minimes, la fourniture d'environnements sûrs et le temps pour expérimenter, développer, tester les améliorations et les environnements dans lesquels on encourage à tirer les leçons des échecs. La prise en charge des opérations pour les environnements de test (sandbox), de développement, de test autres que sandbox et de production, avec un niveau croissant de contrôles opérationnels, facilite le développement et augmente la prévisibilité des résultats positifs des changements déployés en production.

Ressources

Veuillez vous référer aux ressources suivantes pour en savoir plus sur les bonnes pratiques en matière d'excellence opérationnelle.

Documentation

- DevOps et AWS

Livre blanc

- Pilier Excellence opérationnelle

Vidéo

- DevOps at Amazon

Sécurité

Le pilier Sécurité présente la capacité de protéger les données ainsi que les systèmes et les ressources pour tirer parti des technologies du cloud et améliorer votre sécurité.

Il fournit une vue d'ensemble des principes de conception, des bonnes pratiques et des questions. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Sécurité.

Rubriques

- Principes de conception
- Définition
- Bonnes pratiques
- Ressources

Principes de conception

Dans le cloud, il existe un certain nombre de principes qui peuvent vous aider à renforcer la sécurité de vos charges de travail :

- Mettez en place une base d'identité solide : mettez en œuvre le principe du moindre privilège et appliquez la séparation des tâches avec les autorisations appropriées pour chaque interaction avec vos AWS ressources. Centralisez la gestion des identités et visez l'élimination de la dépendance aux informations d'identification statiques de longue durée.
- Maintien de la traçabilité : supervisez, alertez et auditez les actions et les modifications apportées à votre environnement en temps réel. Intégrez la collecte des journaux et des métriques aux systèmes pour effectuer des analyses et prendre des mesures automatiquement.
- Appliquer la sécurité à toutes les couches : appliquez une approche défensive en profondeur avec plusieurs contrôles de sécurité. Appliquer à toutes les couches (par exemple, la périphérie du

réseauVPC, l'équilibrage de charge, chaque instance et service de calcul, le système d'exploitation, l'application et le code).

- Automatiser les bonnes pratiques de sécurité : les mécanismes de sécurité automatisés basés sur le logiciel améliorent votre capacité à évoluer plus rapidement et de manière plus économique en toute sécurité. Créez des architectures sécurisées, y compris avec mise en œuvre des contrôles définis et gérés en tant que code dans les modèles de contrôle de versions.
- Protéger les données en transit et au repos : classez vos données selon différents niveaux de sensibilité et utilisez des mécanismes, tels que le chiffrement, la création de jetons et le contrôle d’accès, le cas échéant.
- Protéger l’accès aux données : utilisez des mécanismes et des outils pour réduire ou éliminer le besoin d’accès direct ou le traitement manuel des données. Cette approche permet de réduire les risques de mauvaise manipulation ou de modification ainsi que les erreurs humaines lors d’interventions sur des données sensibles.
- Se préparer aux événements de sécurité : préparez-vous à un incident en mettant en place une politique et des processus de gestion et d’investigation en matière d’incidents qui correspondent aux exigences de votre organisation. Exécutez des simulations de réponse aux incidents et utilisez des outils d’automatisation pour améliorer votre vitesse de détection, d’investigation et de récupération.

Définition

Il existe sept domaines de bonnes pratiques en matière de sécurité dans le cloud :

- Bases de la sécurité
- Gestion des identités et des accès
- Détection
- Protection de l’infrastructure
- Protection des données
- Intervention en cas d’incidents
- Sécurité des applications

Vous devez mettre en place des pratiques qui influent sur la sécurité avant de concevoir l’architecture d’une charge de travail. Vous voudrez contrôler qui peut faire quoi. De plus, vous voulez être en mesure d’identifier les incidents de sécurité, de protéger vos systèmes et services, et de maintenir

la confidentialité et l'intégrité des données via la protection de données. Vous devez disposer d'un processus bien défini et utilisé pour répondre aux incidents de sécurité. Ces outils et techniques sont importants, car ils soutiennent des objectifs tels que la prévention des pertes financières ou le respect des obligations réglementaires.

Le modèle de responsabilité AWS partagée aide les entreprises qui adoptent le cloud à atteindre leurs objectifs de sécurité et de conformité. Parce que l'infrastructure qui prend en charge nos services cloud est sécurisée AWS physiquement, en tant que AWS client, vous pouvez vous concentrer sur l'utilisation des services pour atteindre vos objectifs. Le AWS cloud fournit également un meilleur accès aux données de sécurité et une approche automatisée pour répondre aux événements de sécurité.

Bonnes pratiques

Rubriques

- Sécurité
- Gestion des identités et des accès
- Détection
- Protection de l'infrastructure
- Protection des données
- Intervention en cas d'incidents
- Sécurité des applications

Sécurité

La question suivante est axée sur ces considérations relatives à la sécurité. (Pour obtenir la liste des questions et des bonnes pratiques liées à la sécurité, consultez l'annexe.)

SEC1 : Comment gérez-vous votre charge de travail en toute sécurité ?

Pour gérer votre charge de travail en toute sécurité, vous devez appliquer les bonnes pratiques générales à tous les domaines de sécurité. Prenez les exigences et les processus que vous avez définis dans le cadre de l'excellence opérationnelle au niveau de l'organisation et de la charge de travail, et appliquez-les à tous les domaines.

SEC1 : Comment gérez-vous votre charge de travail en toute sécurité ?

En vous tenant au courant des recommandations provenant de sources AWS sectorielles et des informations sur les menaces, vous pouvez faire évoluer votre modèle de menace et vos objectifs de contrôle. L'automatisation des processus de sécurité, des tests et de la validation vous permet de mettre à l'échelle vos opérations de sécurité.

En effet AWS, il est recommandé de séparer les différentes charges de travail par compte, en fonction de leur fonction et des exigences de conformité ou de sensibilité des données.

Gestion des identités et des accès

La gestion des identités et des accès est un élément essentiel d'un programme de protection des informations. En effet, elle garantit que seuls les utilisateurs et les composants autorisés et authentifiés sont en mesure d'accéder à vos ressources, et uniquement comme vous le décidez. Par exemple, vous devez définir des principaux (c'est-à-dire les comptes, les utilisateurs, les rôles et les services qui peuvent effectuer des actions dans votre compte), élaborer des stratégies conformes à ces principaux et mettre en œuvre une gestion solide des informations d'identification. Ces éléments de gestion des privilèges constituent la base de l'authentification et de l'autorisation.

Dans AWS, la gestion des privilèges est principalement prise en charge par le service AWS Identity and Access Management (IAM), qui vous permet de contrôler l'accès des utilisateurs et des programmatiques aux AWS services et aux ressources. Vous devez appliquer des stratégies précises qui attribuent des autorisations à un utilisateur, un groupe, un rôle ou une ressource. Vous pouvez également exiger des pratiques strictes en matière de mots de passe, telles que le niveau de complexité, la prévention de la réutilisation et l'application de l'authentification multifactorielle (). MFA Vous pouvez utiliser la fédération avec votre service d'annuaire existant. Pour les charges de travail auxquelles les systèmes doivent avoir accès AWS, IAM permet un accès sécurisé via des rôles, des profils d'instance, une fédération d'identité et des informations d'identification temporaires.

Les questions suivantes sont axées sur ces quelques considérations liées à la sécurité.

SEC2 : Comment gérez-vous les identités des personnes et des machines ?

Il existe deux types d'identités que vous devez gérer lorsque vous abordez des AWS charges de travail sécurisées. La compréhension du type d'identité que vous devez gérer et pour lequel vous devez autoriser l'accès vous permet de vérifier que les identités appropriées ont accès aux ressources adéquates, dans les bonnes conditions.

SEC2 : Comment gérez-vous les identités des personnes et des machines ?

Identités humaines : vos administrateurs, développeurs, opérateurs et utilisateurs finaux ont besoin d'une identité pour accéder à vos AWS environnements et à vos applications. Il s'agit de membres de votre organisation ou d'utilisateurs externes avec lesquels vous collaborez et qui interagissent avec vos AWS ressources via un navigateur Web, une application cliente ou des outils de ligne de commande interactifs.

Identités des machines : vos applications de service, vos outils opérationnels et vos charges de travail ont besoin d'une identité pour envoyer des demandes aux AWS services, par exemple pour lire des données. Ces identités incluent les machines exécutées dans votre AWS environnement, telles que les EC2 instances ou les AWS Lambda fonctions Amazon. Vous pouvez également gérer les Identités machine pour les parties externes qui ont besoin d’un accès. En outre, il se peut AWS que des machines extérieures aient besoin d'accéder à votre AWS environnement.

SEC3 : Comment gérez-vous les autorisations pour les personnes et les machines ?

Gérez les autorisations pour contrôler l'accès aux personnes et aux identités des machines qui nécessitent un accès à votre charge de travail AWS et à celle-ci. Les autorisations régissent les ressources accessibles et les conditions d’accès.

Les informations d’identification ne doivent être partagées entre aucun utilisateur ou système. L'accès des utilisateurs doit être accordé selon une approche basée sur le moindre privilège, avec les meilleures pratiques, y compris les exigences relatives aux mots de passe, et MFA leur application doit être appliquée. L'accès programmatique, y compris API les appels aux AWS services, doit être effectué à l'aide d'informations d'identification temporaires et à privilèges limités, telles que celles émises par le. AWS Security Token Service

Les utilisateurs ont besoin d'un accès programmatique s'ils souhaitent interagir avec AWS l'extérieur du AWS Management Console. La manière d'accorder un accès programmatique dépend du type d'utilisateur qui y accède AWS.

Pour accorder aux utilisateurs un accès programmatique, choisissez l’une des options suivantes.

Quel utilisateur a besoin d'un accès programmatique ?	Pour	Par
Identité de la main-d'œuvre (Utilisateurs gérés dans IAM Identity Center)	Utilisez des informations d'identification temporaires pour signer les demandes programmatiques adressées au AWS CLI AWS SDKs, ou AWS APIs.	Suivez les instructions de l'interface que vous souhaitez utiliser. • Pour le AWS CLI, voir Configuration du AWS CLI à utiliser AWS IAM Identity Center dans le guide de AWS Command Line Interface l'utilisateur. • Pour AWS SDKs, outils, et AWS APIs, voir Authentification IAM Identity Center dans le guide de référence AWS SDKs et Tools.
IAM	Utilisez des informations d'identification temporaires pour signer les demandes programmatiques adressées au AWS CLI AWS SDKs, ou AWS APIs.	Suivez les instructions de la section Utilisation d'informations d'identification temporaires avec les AWS ressources du Guide de IAM l'utilisateur.
IAM	(Non recommandé) Utilisez des informations d'identification à long terme pour signer des demandes programmatiques adressées au AWS CLI AWS SDKs, ou AWS APIs.	Suivez les instructions de l'interface que vous souhaitez utiliser. • Pour le AWS CLI, voir Authentification à l'aide des informations IAM d'identification utilisateur dans le Guide de AWS Command Line Interface l'utilisateur.

Quel utilisateur a besoin d'un accès programmatique ?	Pour	Par
		• Pour les outils AWS SDKs et, voir Authentifier à l'aide d'informations d'identification à long terme dans le guide de référence des outils AWS SDKs et. • Pour AWS APIs, voir Gestion des clés d'accès pour IAM les utilisateurs dans le Guide de IAM l'utilisateur.

AWS fournit des ressources qui peuvent vous aider à gérer les identités et les accès. Pour vous aider à découvrir les bonnes pratiques, explorez nos ateliers pratiques sur la gestion des informations d'identification et de l'authentification, le contrôle de l'accès humain et le contrôle de l'accès par programmation.

Détection

Vous pouvez utiliser les contrôles de détection pour identifier une menace ou un incident de sécurité potentiel. Ils constituent un élément essentiel des cadres de gouvernance et peuvent être utilisés pour soutenir un processus de qualité, une obligation légale ou de conformité et pour identifier les menaces et renforcer les moyens d'intervention. Il existe différents types de contrôles de détection. Par exemple, la réalisation d'un inventaire des ressources et de leurs attributs détaillés favorise une prise de décision plus efficace (et des contrôles du cycle de vie) pour contribuer à établir des bases de référence opérationnelles. Vous pouvez également utiliser un audit interne, un examen des contrôles liés aux systèmes d'informations, pour vérifier que les pratiques répondent aux stratégies et aux exigences, et que vous avez défini les notifications d'alerte automatique correctes en fonction des conditions définies. Ces contrôles sont des facteurs réactifs importants qui peuvent aider votre organisation à identifier et à comprendre la portée des activités anormales.

Dans AWS, vous pouvez mettre en œuvre des contrôles de détection en traitant les journaux, les événements et en surveillant, ce qui permet d'effectuer des audits, des analyses automatisées et des alarmes. CloudTrail enregistre, AWS API appelle, surveille les CloudWatch métriques avec

des alarmes et AWS Config fournit un historique de configuration. Amazon GuardDuty est un service géré de détection des menaces qui surveille en permanence les comportements malveillants ou non autorisés afin de vous aider à protéger vos AWS comptes et vos charges de travail. Des journaux de niveau de service sont également disponibles. Vous pouvez par exemple utiliser Amazon Simple Storage Service (Amazon S3) pour consigner les demandes d'accès.

La question suivante est axée sur ces considérations relatives à la sécurité.

SEC4 : Comment détectez-vous et étudiez-vous les événements de sécurité ?

Capturez et analysez les événements à partir des journaux et des métriques pour gagner en visibilité. Prenez des mesures en cas d'événements de sécurité et de menaces potentielles afin de sécuriser votre charge de travail.

La gestion des journaux est essentielle dans le cadre d'une charge de travail Well-Architected, pour des raisons allant de la sécurité ou de l'analyse aux exigences réglementaires ou légales. Il est essentiel que vous analysiez les fichiers journaux et que vous y répondiez pour pouvoir identifier les incidents de sécurité éventuels. AWS fournit des fonctionnalités qui simplifient l'implémentation de la gestion des journaux en vous offrant la possibilité de définir un cycle de vie de conservation des données ou de définir à quel emplacement les données seront conservées, archivées ou éventuellement supprimées. La gestion des données fiables et prévisibles en devient plus simple et plus économique.

Protection de l'infrastructure

La protection de l'infrastructure comprend des méthodologies de contrôle, telles que la défense en profondeur, nécessaires au respect des bonnes pratiques et des obligations organisationnelles ou réglementaires. L'utilisation de ces méthodologies est essentielle au succès des opérations en cours, que ce soit dans le cloud ou sur site.

Dans AWS, vous pouvez implémenter l'inspection dynamique et apatride des paquets, soit en utilisant les technologies AWS natives, soit en utilisant les produits et services partenaires disponibles via le. AWS Marketplace Vous devez utiliser Amazon Virtual Private Cloud (AmazonVPC) pour créer un environnement privé, sécurisé et évolutif dans lequel vous pouvez définir votre topologie, notamment les passerelles, les tables de routage et les sous-réseaux publics et privés.

Les questions suivantes sont axées sur ces quelques considérations liées à la sécurité.

SEC5 : Comment protégez-vous les ressources de votre réseau ?

Pour toute charge de travail ayant une forme quelconque de connectivité réseau, qu'il s'agisse d'Internet ou d'un réseau privé, plusieurs couches de défense sont nécessaires pour vous protéger contre les menaces externes et internes basées sur le réseau.

SEC6 : Comment protégez-vous vos ressources informatiques ?

Les ressources informatiques de votre charge de travail nécessitent plusieurs couches de protection contre les menaces externes et internes. Les ressources informatiques incluent les EC2 instances, les conteneurs, AWS Lambda les fonctions, les services de base de données, les appareils IoT, etc.

Plusieurs couches de défense sont conseillées dans tout type d'environnement. Dans le cas de la protection de l'infrastructure, la plupart des concepts et méthodes sont valides pour les modèles cloud et sur site. L'application d'une protection de la périphérie, la surveillance des points d'entrée et de sortie, la journalisation complète, la supervision et les alertes, sont toutes essentielles à un plan de sécurité de l'information efficace.

AWS les clients peuvent personnaliser ou renforcer la configuration d'un conteneur ou d'une AWS Elastic Beanstalk instance Amazon Elastic Compute Cloud (AmazonEC2), Amazon Elastic Container Service (AmazonECS), et conserver cette configuration sur une Amazon Machine Image (AMI) immuable. Ensuite, qu'ils soient lancés par Auto Scaling ou manuellement, tous les nouveaux serveurs virtuels (instances) lancés avec cette AMI solution reçoivent la configuration renforcée.

Protection des données

Avant de concevoir l'architecture d'un système, les pratiques de base qui influent sur la sécurité doivent être en place. Par exemple, la classification des données permet de classer les données organisationnelles en fonction des niveaux de sensibilité, et le chiffrement protège les données en les rendant incompréhensibles en cas d'accès non autorisé. Ces outils et techniques sont importants, car ils soutiennent des objectifs tels que la prévention des pertes financières ou le respect des obligations réglementaires.

Dans AWS, les pratiques suivantes facilitent la protection des données :

- En tant que AWS client, vous gardez le contrôle total de vos données.

- AWS vous permet de chiffrer plus facilement vos données et de gérer les clés, y compris la rotation régulière des clés, que vous pouvez facilement automatiser AWS ou gérer.
- La journalisation détaillée qui contient des informations importantes, telles que l'accès aux fichiers et les modifications apportées, est disponible.
- AWS a conçu des systèmes de stockage offrant une résilience exceptionnelle. Par exemple, Amazon S3 Standard, S3 Standard-IA, S3 One Zone-IA et Amazon Glacier sont tous conçus pour fournir une durabilité des objets de 99,999999999 % sur une période d'un an. Ce niveau de durabilité correspond à une perte moyenne annuelle de 0,000000001 % des objets.
- La gestion des versions, qui peut faire partie d'un processus de gestion du cycle de vie des données plus étendu, assure une protection contre les remplacements ou suppressions accidentels et les dommages similaires.
- AWS n'initie jamais le mouvement de données entre les régions. Le contenu affecté à une région restera dans celle-ci, à moins que vous n'utilisiez explicitement une fonctionnalité ou que vous n'exploitiez un service qui fournit cette fonctionnalité.

Les questions suivantes sont axées sur ces quelques considérations liées à la sécurité.

SEC7 : Comment classez-vous vos données ?

La classification des données fournit un moyen de classer les données en fonction de leur importance et de leur sensibilité afin de vous aider à déterminer les contrôles de protection et de conservation appropriés.

SEC8 : Comment protégez-vous vos données au repos ?

Protégez vos données au repos en mettant en place plusieurs contrôles, afin de réduire le risque d'accès non autorisé ou de mauvaise gestion.

SEC9 : Comment protégez-vous vos données en transit ?

Protégez vos données en transit en mettant en place plusieurs contrôles, afin de réduire le risque d'accès non autorisé ou de perte.

AWS fournit plusieurs moyens de chiffrer les données au repos et en transit. Nous intégrons à nos services des fonctionnalités qui facilitent le chiffrement de vos données. Par exemple, nous avons implémenté le chiffrement côté serveur (SSE) pour Amazon S3 afin de vous permettre de stocker plus facilement vos données sous forme cryptée. Vous pouvez également faire en sorte que l'ensemble du processus de HTTPS chiffrement et de déchiffrement (généralement appelé SSL terminaison) soit géré par Elastic Load Balancing (ELB).

Intervention en cas d’incidents

Même avec des contrôles de détection et de prévention extrêmement matures, votre organisation doit toujours mettre en place des processus pour répondre et atténuer l’impact potentiel d’incidents de sécurité. L’architecture de votre charge de travail affecte fortement la capacité de vos équipes à fonctionner efficacement lors d’un incident, à isoler ou à contenir des systèmes et à restaurer les opérations dans un état correct connu. La mise en place des outils et de l’accès avant un incident de sécurité, puis la mise en pratique régulière de la réponse aux incidents pendant les journées de jeu, vous aideront à vérifier que votre architecture peut prendre en charge des enquêtes et des restaurations rapides.

Dans AWS, les pratiques suivantes facilitent une réponse efficace aux incidents :

- La journalisation détaillée qui contient des informations importantes, telles que l’accès aux fichiers et les modifications, est disponible.
- Les événements peuvent être traités automatiquement et lancer des outils qui automatisent les réponses grâce à l'utilisation de AWS APIs.
- Vous pouvez préprovisionner des outils et une « salle blanche » à l’aide de AWS CloudFormation. Cela vous permet d’effectuer des analyses dans un environnement sécurisé et isolé.

La question suivante est axée sur ces considérations relatives à la sécurité.

SEC10 : Comment anticipez-vous les incidents, comment y répondez-vous et comment vous en remettez-vous ?

La préparation est essentielle pour une enquête rapide et efficace, une réponse et une reprise en cas d’incidents de sécurité, afin de minimiser les perturbations pour votre organisation.

Vérifiez que vous disposez d’un moyen permettant d’accorder rapidement l’accès à votre équipe de sécurité. Automatisez également l’isolation des instances ainsi que la saisie de données et l’état des analyses.

Sécurité des applications

La sécurité des applications (AppSec) décrit le processus global de conception, de création et de test des propriétés de sécurité des charges de travail que vous développez. Vous devez disposer de personnes correctement formées dans votre organisation, comprendre les propriétés de sécurité de votre infrastructure de création et de diffusion, et utiliser l’automatisation pour identifier les problèmes de sécurité.

L'adoption de tests de sécurité des applications dans le cadre du cycle de vie de développement de vos logiciels (SDLC) et des processus post-publication permet de valider que vous disposez d'un mécanisme structuré pour identifier, corriger et empêcher les problèmes de sécurité des applications de pénétrer dans votre environnement de production.

Votre méthodologie de développement d’applications doit inclure des contrôles de sécurité lors de la conception, de l’élaboration, du déploiement et de l’exploitation de vos charges de travail. Ce faisant, alignez le processus afin de limiter les défauts en continu et de minimiser la dette technique. Par exemple, l’utilisation de la modélisation des menaces au cours de la phase de conception permet de découvrir rapidement les défauts de conception, ce qui les rend plus faciles et moins coûteux à corriger que d’attendre et de les atténuer plus tard.

Le coût et la complexité de la résolution des défauts diminuent généralement à mesure que vous entrez tôt dans leSDLC. Le moyen le plus simple de résoudre les problèmes est de ne pas en avoir. C’est pourquoi le fait de commencer par élaborer un modèle de menace permet de se concentrer sur les bons résultats dès la phase de conception. Au fur et à mesure que votre AppSec programme évolue, vous pouvez augmenter le nombre de tests effectués grâce à l'automatisation, améliorer la fidélité des commentaires transmis aux concepteurs et réduire le temps nécessaire aux examens de sécurité. Toutes ces actions améliorent la qualité du logiciel que vous créez et accélèrent la mise en production des fonctionnalités.

Ces directives de mise en œuvre se concentrent sur quatre domaines : organisation et culture, sécurité du pipeline, sécurité dans le pipeline et gestion des dépendances. Chaque domaine fournit un ensemble de principes que vous pouvez mettre en œuvre et fournit un end-to-end aperçu de la façon dont vous concevez, développez, développez, déployez et gérez les charges de travail.

Dans AWS, il existe un certain nombre d'approches que vous pouvez utiliser pour aborder votre programme de sécurité des applications. Certaines de ces approches reposent sur la technologie,

tandis que d'autres se concentrent sur les aspects humains et organisationnels de votre programme de sécurité des applications.

La question suivante est axée sur ces considérations relatives à la sécurité des applications.

SEC11 : Comment intégrer et valider les propriétés de sécurité des applications tout au long du cycle de vie de conception, de développement et de déploiement ?

La formation du personnel, le test à l'aide de l'automatisation, la compréhension des dépendanc es et la validation des propriétés de sécurité des outils et des applications contribuent à réduire la probabilité de problèmes de sécurité dans les charges de travail de production.

Ressources

Consultez les ressources suivantes pour en savoir plus sur nos bonnes pratiques relatives à la sécurité.

Documentation

- AWS Sécurité dans le cloud
- ConformitéAWS
- AWS Blog sur la sécurité
- Modèle de maturité de sécuritéAWS

Livre blanc

- Pilier Sécurité
- AWS Aperçu de la sécurité
- AWS Risque et conformité

Vidéo

- AWS État de sécurité de l'Union
- Présentation de la responsabilité partagée

Fiabilité

Le pilier Fiabilité englobe la capacité d'une charge de travail à exécuter sa fonction de manière correcte et cohérente et ce, en temps utile. Cela inclut la possibilité d'exploiter et de tester la charge de travail tout au long de son cycle de vie. Ce livre blanc fournit des bonnes pratiques détaillées pour la mise en œuvre de charges de travail fiables sur AWS.

Le pilier fiabilité fournit une vue d'ensemble des principes de conception, des bonnes pratiques et des questions. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Fiabilité.

Rubriques

Principes de conception

Il existe cinq principes de conception pour la fiabilité dans le cloud :

- Récupération automatique après une panne : en contrôlant les indicateurs de rendement clés d'une charge de travail, vous pouvez déclencher l'automatisation en cas de transgression d'un seuil. Ces KPI doivent couvrir la valeur commerciale, et non des aspects techniques du fonctionnement du service. Cela permet la notification et le suivi automatiques des pannes, et l'exécution de processus de récupération automatique qui contournent ou corrigent la panne. Une automatisation plus sophistiquée rend possible l'anticipation et la correction des pannes avant qu'elles ne se produisent.
- Test des procédures de récupération : dans un environnement sur site, des tests sont souvent conduits pour prouver que la charge de travail fonctionne dans un scénario particulier. Ces tests ne sont généralement pas utilisés pour valider les stratégies de récupération. Dans le cloud, vous pouvez tester de quelle façon votre charge de travail cesse de fonctionner et valider vos procédures de récupération. Vous pouvez utiliser l'automatisation pour simuler différentes pannes ou recréer les scénarios qui ont déjà conduit à des pannes. Cette approche expose les chemins de défaillance que vous pouvez tester et corriger avant qu'un scénario de défaillance réelle ne se produise et réduire ainsi les risques.

- Mise à l'échelle horizontale pour augmenter la disponibilité de la charge de travail : remplacez une ressource volumineuse par plusieurs petites ressources pour réduire l'impact d'une défaillance unique sur la charge de travail globale. Répartissez les demandes entre plusieurs ressources plus petites pour vérifier qu'elles ne partagent pas un point de panne commun.
- Une capacité réellement adaptée à vos besoins : une cause courante de défaillance des charges de travail sur site est la saturation des ressources, lorsque les demandes ciblant une charge de travail en dépassent la capacité (c'est souvent l'objectif des attaques par déni de service). Dans le cloud, vous pouvez contrôler la demande et l'utilisation de la charge de travail. Vous pouvez aussi automatiser l'ajout ou la suppression de ressources afin de maintenir le niveau plus efficace de satisfaction de la demande sans surallocation ou sous-allocation. Des limites demeurent, mais certains quotas peuvent être contrôlés et d'autres gérés (consultez Gestion des Service Quotas et contraintes de service).
- Gestion des changements avec l'automatisation : les modifications apportées à l'infrastructure doivent être appliquées via l'automatisation. Les modifications à gérer incluent celles apportées à l'automatisation et qui peuvent ensuite être suivies et vérifiées.

Définition

Il existe quatre domaines de bonnes pratiques en matière de fiabilité dans le cloud :

- Fondations
- Architecture de charge de travail
- Gestion des modifications
- Gestion des défaillances

Pour la fiabilité, vous devez commencer par les bases, c'est-à-dire un environnement où les Service Quotas et la topologie réseau s'adaptent à la charge de travail. L'architecture de la charge de travail du système distribué doit être conçue pour prévenir et atténuer les défaillances. La charge de travail doit gérer les modifications au niveau de la demande ou des exigences. Elle doit être conçue pour détecter les défaillances et se réparer automatiquement.

Bonnes pratiques

Rubriques

- Fondations

- Architecture de charge de travail
- Gestion des modifications
- Gestion des défaillances

Fondations

Les exigences de base sont celles dont le champ d'application s'étend au-delà d'une seule charge de travail ou d'un seul projet. Avant de concevoir l'architecture d'un système, les exigences de base qui influent sur la fiabilité doivent être mises en place. Par exemple, vous devez avoir une bande passante du réseau suffisante pour votre centre de données.

Avec AWS, la plupart de ces exigences élémentaires sont déjà intégrées ou sont gérées au cas par cas. Le cloud est conçu pour être presque illimité. Il est donc de la responsabilité d'AWS de satisfaire l'exigence d'une capacité suffisante de mise en réseau et de calcul, ce qui vous permet de modifier la taille des ressources et les allocations à la demande.

Les questions suivantes sont axées sur ces quelques considérations relatives à la fiabilité. (Pour obtenir la liste des questions et bonnes pratiques en matière de fiabilité, consultez l'annexe.)

FIA 1 : comment gérer les Service Quotas et les contraintes de service ?

Pour les architectures de charge de travail basées sur le Cloud, il existe des Service Quotas (également appelés limites de service). Ces quotas permettent d'éviter de fournir accidentellement plus de ressources que nécessaire et de limiter les taux de demande des opérations d'API afin de protéger les services de tout abus. Il existe également des contraintes de ressources, par exemple la vitesse à laquelle les bits peuvent être transmis par un câble à fibre optique ou la quantité de données stockées sur un disque physique.

FIA 2 : comment planifier la topologie de votre réseau ?

Les charges de travail existent souvent dans plusieurs environnements. Il s'agit notamment de plusieurs environnements Cloud (accessibles au public et privés) et éventuellement de votre infrastructure de centre de données existante. Les plans doivent inclure des considérations relatives au réseau, telles que la connectivité intra- et inter-systèmes, la gestion des adresses IP publiques, la gestion des adresses IP privées et la résolution des noms de domaine.

Architecture de charge de travail

Pour garantir la fiabilité d'une charge de travail, il faut commencer par choisir le bon logiciel et la bonne infrastructure. Vos choix d'architecture ont un impact sur le comportement des charges de travail sur les différents piliers Well-Architected. Pour des raisons de fiabilité, vous devez suivre des modèles spécifiques.

Avec AWS, les développeurs de charges de travail peuvent choisir les langages et les technologies à utiliser. Les kits AWS SDK éliminent la complexité du codage en fournissant des API propres au langage pour les services AWS. Ces kits SDK, ainsi que le choix des langages, permettent aux développeurs de mettre en œuvre les bonnes pratiques de fiabilité répertoriées ici. Les développeurs peuvent également découvrir comment Amazon conçoit et exploite des logiciels dans Amazon Builders' Library.

Les questions suivantes sont axées sur ces quelques considérations relatives à la fiabilité.

FIA 3 : comment concevoir l'architecture de service de votre charge de travail ?

Créez des charges de travail hautement évolutives et fiables à l'aide d'une architecture orientée services (SOA) ou d'une architecture de microservices. L'architecture orientée services (SOA) consiste à rendre les composants logiciels réutilisables via les interfaces de service. L'archite cture des microservices va plus loin, en particulier en rendant les composants plus petits et plus simples.

FIA 4 : comment concevoir des interactions dans un système distribué pour éviter les défaillan ces ?

Les systèmes distribués s'appuient sur des réseaux de communication pour interconnecter les composants, comme les serveurs ou les services. Votre charge de travail doit fonctionner de manière fiable malgré la perte de données ou la latence sur ces réseaux. Les composants du système distribué doivent fonctionner de manière à ne pas avoir d'impact négatif sur les autres composants ou sur la charge de travail. Ces bonnes pratiques permettent d'éviter les défaillances et d'améliorer le temps moyen entre défaillances (MTBF).

FIA 5 : comment concevoir des interactions dans un système distribué pour atténuer les défaillances ou y résister ?

Les systèmes distribués s'appuient sur des réseaux de communication pour interconnecter des composants (tels que des serveurs ou des services). Votre charge de travail doit fonctionner de manière fiable malgré la perte de données ou la latence sur ces réseaux. Les composants du système distribué doivent fonctionner de manière à ne pas avoir d'impact négatif sur les autres composants ou sur la charge de travail. Ces bonnes pratiques permettent aux charges de travail de résister aux contraintes ou aux défaillances, de s'en remettre plus rapidement et d'atténuer l'impact de ces altérations. Il en résulte une amélioration du temps moyen de récupération (MTTR).

Gestion des modifications

Les modifications apportées à votre charge de travail ou à son environnement doivent être anticipées et prises en compte pour assurer un fonctionnement fiable de la charge de travail. Les modifications incluent celles imposées à votre charge de travail telles que les pics de demande, ainsi que celles venant de l'intérieur comme les déploiements de fonctionnalités et les correctifs de sécurité.

Avec AWS, vous pouvez surveiller le comportement d'une charge de travail et automatiser la réponse aux KPI. Par exemple, votre charge de travail peut ajouter des serveurs supplémentaires à mesure que des utilisateurs supplémentaires s'y ajoutent. Vous pouvez contrôler les personnes qui ont l'autorisation d'apporter des modifications à la charge de travail et auditer l'historique de ces modifications.

Les questions suivantes sont axées sur ces quelques considérations relatives à la fiabilité.

FIA 6 : comment surveiller les ressources de charges de travail ?

Les journaux et les métriques sont des outils puissants qui permettent de mieux comprendre l'état de santé de votre charge de travail. Vous pouvez configurer votre charge de travail pour qu'elle surveille les journaux et les métriques et envoie des notifications lorsque des seuils sont franchis ou que des événements importants se produisent. La surveillance permet à votre charge de travail de reconnaître quand des seuils de faibles performances sont franchis ou quand des défaillances se produisent, afin d'y répondre par une récupération automatique.

FIA 7 : comment concevoir votre charge de travail pour qu'elle s'adapte aux changements de demande ?

Une charge de travail évolutive permet d'ajouter ou de supprimer automatiquement des ressource s de manière à ce qu'elles correspondent à la demande actuelle à un moment donné.

FIA 8 : comment implémenter les modifications ?

Des modifications contrôlées sont nécessaires pour déployer de nouvelles fonctionnalités et vérifier que les charges de travail et l'environnement d'exploitation fonctionnent avec des logiciels connus et peuvent être corrigés ou remplacés de manière prévisible. Si ces modifications ne sont pas maîtrisées, il devient difficile d'en prévoir les effets ou de résoudre les problèmes qui en découlent.

Lorsque vous concevez l'architecture d'une charge de travail de manière à ajouter ou supprimer automatiquement des ressources en réponse aux évolutions de la demande, cela accroît la fiabilité et garantit également que la réussite commerciale ne devient pas un poids. Une fois la surveillance en place, votre équipe est automatiquement avertie lorsque les KPI cessent de correspondre aux valeurs attendues. La journalisation automatique des modifications apportées à votre environnement vous permet d'auditer et d'identifier rapidement les actions susceptibles d'avoir un impact sur la fiabilité. Les contrôles de la gestion des modifications certifient que vous appliquez les règles offrant la fiabilité dont vous avez besoin.

Gestion des défaillances

Des pannes peuvent survenir dans tous les systèmes présentant un niveau de complexité raisonnable. Pour que votre charge de travail soit fiable, vous devez avoir connaissance des défaillances au moment où elles se produisent et prendre des mesures pour éviter qu'elles aient un impact sur la disponibilité. Les charges de travail doivent être en mesure de résister aux défaillances et de résoudre automatiquement les problèmes.

Avec AWS, vous pouvez tirer profit de l'automatisation pour réagir aux données de surveillance. Par exemple, lorsqu'une métrique particulière franchit un seuil, vous pouvez lancer une action automatique pour corriger le problème. De même, plutôt que de tenter de diagnostiquer et de corriger une ressource défaillante qui fait partie de votre environnement de production, vous pouvez la remplacer par une nouvelle ressource et exécuter l'analyse de cette ressource hors production.

Comme le cloud vous permet de maintenir les versions temporaires d'un système complet à bas coût, vous pouvez utiliser les tests automatiques pour vérifier les processus complets de récupération.

Les questions suivantes sont axées sur ces quelques considérations relatives à la fiabilité.

FIA 9 : Comment sauvegarder des données ?

Sauvegardez les données, les applications et la configuration pour répondre à vos exigences en matière d'objectifs de délai de reprise (RTO) et de points de reprise (RPO).

FIA 10 : Comment utiliser l'isolation des pannes pour protéger votre charge de travail ?

L'isolation des défaillances limite l'impact de la défaillance d'un composant ou d'un système à une limite définie. Si l'isolation est correcte, les composants situés en dehors de cette limite ne sont pas affectés par la défaillance. L'exécution de votre charge de travail au-delà de plusieurs limites d'isolation des défaillances peut la rendre plus résistante aux défaillances.

FIA 11 : comment concevoir votre charge de travail pour la rendre résistante aux défaillances de composants ?

Les charges de travail exigeant une haute disponibilité et un faible temps moyen de récupération (MTTR) doivent être conçues pour être résilientes.

FIA 12 : comment tester la fiabilité ?

Une fois que vous avez conçu votre charge de travail pour qu'elle soit résiliente aux sollicitations de la production, les tests sont le seul moyen de s'assurer qu'elle fonctionne comme prévu et d'obtenir la résilience voulue.

FIA 13 : comment planifier la reprise après sinistre (DR) ?

La mise en place de sauvegardes et de composants de charge de travail redondants constitue le début de votre stratégie de DR. L'objectif de délai de reprise (RTO) et l'objectif de point de reprise

FIA 13 : comment planifier la reprise après sinistre (DR) ?

(RPO) sont vos objectifs pour la restauration de votre charge de travail. Définissez-les en fonction des besoins de l'entreprise. Mettez en œuvre une stratégie pour atteindre ces objectifs, en particuli er en tenant compte de l'emplacement et de la fonction des données et des ressources de charge de travail. La probabilité d'une perturbation et le coût de la reprise sont également des facteurs clés qui permettent de déterminer la valeur opérationnelle de la reprise après sinistre d'une charge de travail.

Sauvegardez régulièrement vos données et testez vos fichiers de sauvegarde pour vérifier que vous pouvez récupérer après des erreurs logiques ou physiques. La clé de la gestion des pannes réside dans des tests réguliers et automatiques des charges de travail afin de créer des pannes, et dans l'observation de la façon dont ces charges reprennent. Effectuez ces opérations régulièrement et vérifiez que de tels tests sont également lancés après des modifications significatives de la charge de travail. Suivez activement les KPI, ainsi que l'objectif de délai de reprise (RTO) et l'objectif de point de reprise (RPO) pour évaluer la résilience d'une charge de travail (notamment au cours de scénarios de test de panne). Le suivi des KPI vous aidera à identifier et à atténuer les points de défaillance uniques. L'objectif est de tester intégralement vos processus de reprise de charge de travail de telle sorte que vous soyez assuré de récupérer l'ensemble de vos données et de continuer à servir vos clients, même en présence de problèmes persistants. Vos processus de reprise doivent être aussi bien maîtrisés que vos processus de production habituels.

Ressources

Consultez les ressources suivantes pour en savoir plus sur nos bonnes pratiques relatives à la fiabilité.

Documentation

- Documentation AWS
- Infrastructure mondiale AWS
- AWS Auto Scaling : Fonctionnement des plans de dimensionnement
- Présentation de AWS Backup

Livre blanc

- Pilier Fiabilité : AWS Well-Architected

- Implémentation des microservices sur AWS

Efficacité des performances

Le pilier Efficacité des performances englobe la capacité à utiliser efficacement les ressources du cloud pour satisfaire aux exigences de performances et à maintenir cette efficacité au fur et à mesure que la demande change et que les technologies évoluent.

Le pilier Efficacité des performances fournit une vue d'ensemble des principes de conception, des bonnes pratiques et des questions. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Efficacité en matière de performance.

Rubriques

- Principes de conception
- Définition
- Bonnes pratiques
- Ressources

Principes de conception

Il existe cinq principes de conception pour l'efficacité des performances dans le cloud :

- Démocratisation des technologies avancées : simplifiez la mise en œuvre de technologies avancées pour votre équipe en déléguant des tâches complexes à votre fournisseur de cloud. Plutôt que de demander à votre équipe informatique de s'informer sur l'hébergement et l'exploitation d'une nouvelle technologie, envisagez de consommer la technologie en tant que service. Par exemple, l'absence SQL de bases de données, le transcodage multimédia et l'apprentissage automatique sont autant de technologies qui nécessitent une expertise spécialisée. Dans le cloud, ces technologies deviennent des services que votre équipe peut consommer, ce qui lui permet de se dédier au développement de produits plutôt qu'à l'allocation et à la gestion des ressources.
- Passez à l'international en quelques minutes : le déploiement de votre charge de travail dans plusieurs AWS régions du monde vous permet de réduire le temps de latence et d'offrir une meilleure expérience à vos clients à moindre coût.
- Utilisation d'architectures sans serveur : les architectures sans serveur vous évitent d'exécuter et de gérer des serveurs physiques pour les activités traditionnelles de calcul. Par exemple, les

services de stockage sans serveur peuvent agir comme des sites Web statiques (éliminant le besoin de serveurs Web), et les services d'événements peuvent héberger du code. Ainsi, vous supprimez la charge opérationnelle de gestion des serveurs physiques et réduisez les coûts des transactions, car les services gérés fonctionnent à l'échelle du cloud.

- Expérimentation plus fréquente : avec des ressources virtuelles et automatisables, vous pouvez rapidement exécuter des tests comparatifs à l'aide de différents types d'instances, de stockage ou de configurations.
- Prévision de la compréhension technique : comprenez comment les services cloud sont consommés et utilisez toujours l'approche technologique qui correspond le mieux à vos objectifs de charges de travail. Par exemple, tenez compte des modèles d'accès aux données lorsque vous sélectionnez les approches de stockage ou de base de données.

Définition

Les bonnes pratiques en matière d'efficacité des performances dans le cloud sont au nombre de cinq :

- Choix d'architecture
- Informatique et matériel
- Gestion des données
- Réseau et diffusion de contenu
- Processus et culture

Optez pour une approche orientée données lors de la création d'une architecture à hautes performances. Collectez des données sur tous les aspects de l'architecture, depuis la conception générale jusqu'à la sélection et la configuration des types de ressources.

L'examen régulier de vos choix confirme que vous tirez parti de l'évolution constante du AWS Cloud. La surveillance vous offre la garantie d'être informé de tout écart par rapport aux performances attendues. Effectuez des compromis dans votre architecture pour améliorer les performances, comme l'utilisation de la compression, la mise en cache ou l'abaissement des exigences de cohérence.

Bonnes pratiques

Rubriques

- Choix d’architecture
- Informatique et matériel
- Gestion des données
- Réseau et diffusion de contenu
- Processus et culture

Choix d’architecture

La solution optimale pour une charge de travail peut varier, et les solutions combinent souvent plusieurs approches. Les charges de travail Well-Architected utilisent plusieurs solutions et permettent d’exploiter différentes fonctionnalités pour améliorer les performances.

AWS les ressources sont disponibles dans de nombreux types et configurations, ce qui permet de trouver plus facilement une approche qui correspond le mieux à vos besoins. Vous pouvez également rechercher des options qui ne sont pas facilement accessibles avec une infrastructure sur site. Par exemple, un service géré tel qu'Amazon DynamoDB fournit une base de données SQL sans base de données entièrement gérée avec une latence d'un chiffre en millisecondes, quelle que soit l'échelle.

La question suivante est axée sur ces quelques considérations relatives à l’efficacité des performances. (Pour obtenir la liste des questions et bonnes pratiques en matière d’efficacité des performances, consultez l’annexe.)

PERF1 : Comment sélectionnez-vous les ressources cloud et les modèles d'architecture adaptés à votre charge de travail ?

Plusieurs approches sont souvent nécessaires pour obtenir de meilleures performances sur une charge de travail. Les systèmes Well-Architected utilisent plusieurs solutions et fonctions pour améliorer les performances.

Informatique et matériel

Le choix d’une solution de calcul optimale pour une charge de travail particulière peut varier selon la conception de l’application, les modèles d’utilisation et les paramètres de configuration. Les architectures peuvent utiliser différentes solutions de calcul pour divers composants et permettent

différentes fonctionnalités pour améliorer les performances. Le choix d'une solution de calcul inadaptée à une architecture peut nuire à ses performances.

Dans AWS, le calcul est disponible sous trois formes : instances, conteneurs et fonctions :

- Les instances sont des serveurs virtualisés qui vous permettent de modifier leurs fonctionnalités à l'aide d'un bouton ou d'un API appel. Comme les décisions relatives aux ressources dans le cloud ne sont pas figées, vous pouvez expérimenter avec différents types de serveurs. Chez AWS, ces instances de serveur virtuel se déclinent en différentes familles et tailles, et elles offrent une grande variété de fonctionnalités, notamment des disques SSD (SSDs) et des unités de traitement graphique (GPUs).
- Les conteneurs sont une méthode de virtualisation du système d'exploitation qui vous permet d'exécuter une application et ses dépendances dans le cadre de processus isolés en termes de ressources. AWS Fargate est le calcul sans serveur pour les conteneurs ou Amazon EC2 peut être utilisé si vous avez besoin de contrôler l'installation, la configuration et la gestion de votre environnement informatique. Vous pouvez également choisir parmi plusieurs plateformes d'orchestration de conteneurs : Amazon Elastic Container Service (ECS) ou Amazon Elastic Kubernetes Service (). EKS
- Les fonctions extraient l'environnement d'exécution depuis le code à appliquer. Par exemple, vous AWS Lambda permet d'exécuter du code sans exécuter d'instance.

La question suivante est axée sur ces quelques considérations relatives à l'efficacité des performances.

PERF2 : Comment sélectionnez-vous et utilisez-vous les ressources informatiques dans votre charge de travail ?

La solution de calcul la plus efficace pour une charge de travail varie en fonction de la conceptio n de l'application, des modèles d'utilisation et des paramètres de configuration. Les architect ures peuvent utiliser différentes solutions de calcul pour divers composants et activer différentes fonctions pour améliorer les performances. La sélection d'une solution de calcul inadaptée à une architecture peut nuire à ses performances.

Gestion des données

La solution de gestion des données optimale pour un système donné varie en fonction du type de données (bloc, fichier ou objet), des modèles d'accès (aléatoires ou séquentiels), du débit requis, de la fréquence d'accès (en ligne, hors ligne, archivage), de la fréquence de mise à jour (WORMdynamique) et des contraintes de disponibilité et de durabilité. Les charges de travail Well-Architected utilisent des magasins de données sur mesure qui intègrent différentes fonctionnalités pour améliorer les performances.

Dans AWS, le stockage est disponible sous trois formes : objet, bloc et fichier :

- Le stockage d’objets fournit une plateforme avec capacité de mise à l’échelle et durable pour rendre les données accessibles depuis n’importe quel emplacement Internet pour le contenu généré par l’utilisateur, l’archivage actif, le calcul sans serveur, le stockage de données big data ou la sauvegarde et la restauration. Amazon Simple Storage Service (Amazon S3) est un service de stockage d’objets qui offre une capacité de mise à l’échelle, une disponibilité des données, une sécurité et des performances de pointe. Amazon S3 offre une durabilité de 99,999999999 % (11 « 9 ») et stocke des données pour des millions d’applications pour des entreprises du monde entier.
- Le stockage par blocs fournit un stockage par blocs à haute disponibilité, cohérent et à faible latence pour chaque hôte virtuel. Il est analogue au stockage en attachement direct (DAS) ou à un réseau de stockage (). SAN Amazon Elastic Block Store (AmazonEBS) est conçu pour les charges de travail qui nécessitent un stockage persistant accessible par des EC2 instances, ce qui vous permet d'ajuster les applications avec la capacité de stockage, les performances et les coûts appropriés.
- Le stockage de fichiers permet d’accéder à un système de fichiers partagés sur plusieurs systèmes. Les solutions de stockage de fichiers telles qu'Amazon Elastic File System (AmazonEFS) sont idéales pour les cas d'utilisation tels que les référentiels de contenu volumineux, les environnements de développement, les magasins de médias ou les répertoires personnels des utilisateurs. Amazon FSx permet de lancer et d'exécuter des systèmes de fichiers courants de manière efficace et rentable, afin que vous puissiez tirer parti des riches fonctionnalités et des performances rapides des systèmes de fichiers open source et sous licence commerciale largement utilisés.

La question suivante est axée sur ces quelques considérations relatives à l’efficacité des performances.

PERF3 : Comment stockez-vous, gérez-vous et accédez-vous aux données de votre charge de travail ?

La solution de stockage la plus efficace pour un système varie en fonction du type d'opération d'accès (bloc, fichier ou objet), des modèles d'accès (aléatoire ou séquentiel), du débit requis, de la fréquence d'accès (en ligne, hors ligne, archivage), de la fréquence de mise à jour (WORMdyna mique) et des contraintes de disponibilité et de durabilité. Les systèmes Well-Architected utilisent plusieurs solutions de stockage et activent différentes fonctionnalités pour améliorer les performan ces et utiliser efficacement les ressources.

Réseau et diffusion de contenu

La solution de mise en réseau optimale pour une charge de travail varie en fonction de la latence, des exigences de débit, de l’instabilité et de la bande passante. Le choix des options d’emplacement est tributaire des contraintes physiques telles que les ressources pour utilisateur ou sur site. Ces contraintes peuvent être compensées avec les emplacements périphériques ou le placement des ressources.

Activé AWS, le réseau est virtualisé et est disponible dans un certain nombre de types et de configurations différents. Il est ainsi plus facile de répondre à vos besoins en matière de mise en réseau. AWS propose des fonctionnalités de produit (par exemple, mise en réseau améliorée, instances optimisées pour le EC2 réseau Amazon, accélération des transferts Amazon S3 et Amazon dynamique CloudFront) pour optimiser le trafic réseau. AWS propose également des fonctionnalités réseau (par exemple, le routage de latence Amazon Route 53, les VPC points de terminaison Amazon AWS Direct Connect, etc. AWS Global Accelerator) pour réduire la distance ou l'instabilité du réseau.

La question suivante est axée sur ces quelques considérations relatives à l’efficacité des performances.

PERF4 : Comment sélectionnez-vous et configurez-vous les ressources réseau de votre charge de travail ?

Cette question inclut des conseils et des bonnes pratiques pour concevoir, configurer et exploiter des solutions de mise en réseau et de diffusion de contenu efficaces dans le cloud.

Processus et culture

Lors de la création de l’architecture des charges de travail, vous pouvez adopter certains principes et certaines pratiques pour optimiser l’exécution de charges de travail cloud efficaces et performantes. Pour adopter une culture qui favorise l’efficacité des performances des charges de travail dans le cloud, tenez compte des principes et pratiques clés suivants.

Tenez compte de ces principes clés pour développer cette culture :

- Infrastructure en tant que code : définissez votre infrastructure en tant que code à l'aide d'approches telles que AWS CloudFormation des modèles. L’utilisation de modèles vous permet de placer votre infrastructure en mode de contrôle de code source parallèlement au code et aux configurations de votre application. Vous pouvez ainsi appliquer les pratiques utilisées pour développer des logiciels à votre infrastructure et itérer rapidement.
- Pipeline de déploiement : utilisez un pipeline de déploiement d’intégration continue (CI) et de livraison continue (CD) (par exemple, référentiel de code source, systèmes de génération, déploiement et automatisation des tests) pour déployer votre infrastructure. Vous pouvez ainsi déployer de manière reproductible et cohérente, le tout à un faible coût, à mesure que vous itérez.
- Mesures bien définies : configurez et surveillez les mesures pour capturer les indicateurs de performance clés (KPIs). Nous vous recommandons d’utiliser des métriques techniques, mais aussi des métriques commerciales. Pour les sites Web ou les applications mobiles, les indicateurs clés sont la capture time-to-first-byte ou le rendu. D’autres mesures généralement applicables comprennent le nombre de threads, le taux de récupérateur de mémoire et les états d’attente. Les métriques commerciales, telles que les coûts cumulés agrégés par demande, peuvent vous permettre d’identifier des solutions pour réduire vos coûts. Réfléchissez bien à la façon dont vous prévoyez d’interpréter les métriques. Par exemple, vous pouvez choisir le maximum ou le 99e centile plutôt que la moyenne.
- Tests de performance automatiques : dans le cadre de votre processus de déploiement, des tests de performance peuvent se déclencher automatiquement une fois les tests en cours d’exécution bien effectués. L’automatisation doit créer un environnement, configurer des conditions initiales (comme des données de test), puis exécuter une série d’analyses comparatives et de tests de charge. Les résultats de ces tests doivent être rattachés à la version de génération afin que vous puissiez suivre l’évolution des performances dans le temps. Pour les tests de longue durée, vous pouvez rendre cette partie du pipeline asynchrone par rapport au reste de la compilation. Vous pouvez également exécuter des tests de performance pendant la nuit à l'aide d'Amazon EC2 Spot Instances.

- Génération de charge : vous devez créer une série de scripts qui reproduisent des parcours utilisateur synthétiques ou préenregistrés. Ces scripts doivent être idempotents et non couplés. Il se peut que vous deviez aussi inclure à cette série des scripts de préparation pour obtenir des résultats valides. Dans la mesure du possible, vos scripts de test doivent pouvoir répliquer le comportement d'utilisation en production. Vous pouvez utiliser des logiciels ou des solutions software-as-a-service (SaaS) pour générer la charge. Envisagez d'utiliser les solutions AWS Marketplace et les instances Spot : elles peuvent être des moyens économiques de générer la charge.
- Visibilité des performances : les métriques clés doivent être visibles pour votre équipe, en particulier pour chaque version. Vous pouvez ainsi identifier les tendances positives ou négatives significatives au fil du temps. Vous devez également afficher les métriques sur le nombre d'erreurs ou d'exceptions pour vous assurer que vous testez un système fonctionnel.
- Visualisation : utilisez des techniques de visualisation qui permettent d'identifier clairement l'origine des problèmes de performances, les points chauds, les états d'attente ou les taux d'utilisation faibles. Superposez les métriques de performance sur les schémas d'architecture, des graphiques ou codes d'appel qui peuvent vous aider à identifier rapidement les problèmes.
- Processus d'examen régulier : les architectures qui présentent des performances médiocres sont généralement le résultat d'un processus d'évaluation des performances inexistant ou interrompu. Si votre architecture est peu performante, la mise en œuvre d'un processus d'évaluation des performances vous permet de procéder à des améliorations itératives.
- Optimisation continuelle : adoptez une culture permettant d'optimiser en permanence l'efficacité des performances de votre charge de travail dans le cloud.

La question suivante est axée sur ces quelques considérations relatives à l'efficacité des performances.

PERF5 : Quel processus utilisez-vous pour améliorer l'efficacité des performances de votre charge de travail ?

Lors de la création de l'architecture des charges de travail, vous pouvez adopter certains principes et certaines pratiques pour optimiser l'exécution de charges de travail cloud efficaces et performan tes. Pour adopter une culture qui favorise l'efficacité des performances des charges de travail dans le cloud, tenez compte des principes et pratiques clés suivants.

Ressources

Consultez les ressources suivantes pour en savoir plus sur nos bonnes pratiques relatives à l'efficacité des performances.

Documentation

- Optimisation des performances Amazon S3
- Amazon EBS Volume Performance

Livre blanc

- Pilier Efficacité des performances

Vidéo

- AWS re:Invent 2019 : les EC2 fondations d'Amazon (-R2) CMP211
- AWS RE:Invent 2019 : Séance de direction : État du stockage dans l'union (01-L) STG2
- AWS re:Invent 2019 : Séance sur le leadership : bases de données AWS spécialement conçues (09-L) DAT2
- AWS re:Invent 2019 : Connectivité AWS et architectures AWS réseau hybrides (NET317-R1)
- AWS re:Invent 2019 : Au service d'EC2Amazon de nouvelle génération : plongée en profondeur dans le système Nitro (03-R2) CMP3
- AWS re:Invent 2019 : passer à vos 10 premiers millions d'utilisateurs (ARC211-R)

Optimisation des coûts

Le pilier Optimisation des coûts comprend la possibilité d'exécuter des systèmes pour offrir une valeur métier au prix le plus bas.

Le pilier que représente l'optimisation des coûts fournit une vue d'ensemble des principes de conception, des bonnes pratiques et des interrogations. Vous trouverez des conseils prescriptifs sur la mise en œuvre dans le livre blanc Pilier Optimisation des coûts.

Rubriques

Principes de conception

Il existe cinq principes de conception pour l'optimisation des coûts dans le cloud :

- Mettre en œuvre la gestion financière en cloud : mettre en œuvre la gestion financière du cloud : pour atteindre le succès financier et accélérer la réalisation de la valeur ajoutée dans le cloud, vous devez investir dans la gestion financière du cloud/l'optimisation des coûts. Votre organisation doit consacrer du temps et des ressources au renforcement des capacités dans ce nouveau domaine de la gestion des technologies et des usages. Comme pour les piliers Sécurité ou Excellence opérationnelle, vous devez renforcer vos capacités par l'acquisition de connaissances, de programmes, de ressources et de processus pour devenir une entreprise rentable.
- Adopter un modèle basé sur votre consommation : payez uniquement pour les ressources de calcul dont vous avez besoin et augmentez ou diminuez l'utilisation en fonction des exigences opérationnelles, sans recourir à des prévisions élaborées. Par exemple, les environnements de développement et de test ne sont généralement utilisés que huit heures par jour pendant la semaine de travail. Vous pouvez désactiver ces ressources lorsqu'elles ne sont pas utilisées et réduire les coûts de 75 % (40 heures au lieu de 168 heures).
- Mesurer l'efficacité globale : mesurez le rendement opérationnel de la charge de travail et les coûts associés à sa distribution. Utilisez ces informations pour déterminer les avantages que vous pouvez tirer de l'augmentation du rendement et de la réduction des coûts.
- Arrêtez de dépenser de l'argent pour des tâches lourdes indifférenciées : AWS elle se charge des opérations les plus lourdes des centres de données, telles que le montage en rack, l'empilage et l'alimentation des serveurs. Cela supprime également la charge opérationnelle liée à la gestion des systèmes d'exploitation et des applications avec des services gérés. Ainsi, vous pouvez vous concentrer sur vos clients et vos projets professionnels plutôt que sur l'infrastructure informatique.
- Analyser et attribuer les dépenses : le cloud facilite l'identification précise de l'utilisation et du coût des systèmes, ce qui permet ensuite d'attribuer de manière transparente les coûts informatiques aux différents propriétaires des charges de travail. Cela permet de mesurer le retour sur investissement (ROI) et donne aux responsables de la charge de travail la possibilité d'optimiser leurs ressources et de réduire les coûts.

Définition

Il existe cinq domaines de bonnes pratiques pour l’optimisation des coûts dans le cloud :

- Pratiques en matière de gestion financière du cloud
- Sensibilisation aux dépenses et à l’utilisation
- Ressources rentables
- Gérer la demande et les sources d’approvisionnement
- Optimiser dans le temps

Comme pour les autres piliers du Well-Architected Framework, il y a des compromis à prendre en compte, par exemple, pour savoir s'il faut optimiser en fonction des coûts ou en fonction des coûts. speed-to-market Dans certains cas, il est plus efficace d’optimiser la vitesse, de commercialiser rapidement, de livrer de nouvelles fonctionnalités ou de respecter un délai, plutôt que d’investir dans l’optimisation des coûts initiaux. Les décisions de conception sont parfois prises avec vitesse et non selon les données, et il est tentant de surcompenser « au cas où », plutôt que de consacrer du temps à des essais comparatifs pour un déploiement le plus optimal en matière de coût. Cela peut entraîner des déploiements sur-approvisionnés et sous-optimisés. Cependant, il s’agit d’un choix raisonnable lorsque vous avez besoin de « lift-and-shift » depuis votre environnement sur site vers le cloud, puis d’optimiser par la suite. En investissant dès le départ les efforts nécessaires dans une stratégie d’optimisation des coûts, vous pourrez profiter plus facilement des avantages économiques du cloud en respectant les bonnes pratiques et en évitant un surdimensionnement inutile. Les sections suivantes présentent les techniques et les bonnes pratiques pour la mise en œuvre initiale et continue de la gestion financière dans le cloud et l’optimisation des coûts de vos charges de travail.

Bonnes pratiques

Rubriques

Pratiques en matière de gestion financière du cloud

Avec l’adoption du cloud, les équipes technologiques innovent plus rapidement grâce à la réduction des cycles d’approbation, d’achat et de déploiement des infrastructures. Une nouvelle approche de la gestion financière dans le cloud est nécessaire pour générer de la valeur ajoutée et connaître le succès financier. Cette approche, appelée « gestion financière dans le cloud », permet de renforcer les capacités de votre organisation en mettant en œuvre des programmes, des ressources et des processus de renforcement des connaissances à l’échelle de l’organisation.

De nombreuses organisations sont composées de nombreuses unités différentes avec des priorités différentes. La capacité d’aligner votre organisation sur un ensemble d’objectifs financiers convenus et de lui fournir les mécanismes nécessaires pour les atteindre crée une organisation plus efficace. Une organisation sera capable d’innover et de créer plus rapidement, d’être plus agile et de s’adapter à tous les facteurs internes ou externes.

AWS Vous pouvez utiliser Cost Explorer, et éventuellement Amazon Athena et Amazon QuickSight with the Cost and Usage Report (CUR), pour informer l'ensemble de votre organisation sur les coûts et l'utilisation. AWS Budgets fournit des notifications proactives concernant les coûts et l’utilisation. Les AWS blogs fournissent des informations sur les nouveaux services et fonctionnalités afin de vérifier que vous êtes au courant des nouvelles versions de services.

La question suivante est axée sur ces quelques considérations relatives à l’optimisation des coûts. (Pour obtenir la liste des questions et bonnes pratiques en matière d’optimisation des coûts, consultez l’Annexe.)

COST1 : Comment mettez-vous en œuvre la gestion financière dans le cloud ?

La mise en œuvre de la gestion financière dans le cloud aide les entreprises à tirer parti de la valeur commerciale et à réussir sur le plan financier en optimisant leurs coûts et leur utilisation et en les adaptant AWS.

Lorsque vous créez une fonction d'optimisation des coûts, faites appel à des membres et complétez l'équipe avec des experts en CFM optimisation des coûts. Les membres actuels de l’équipe comprendront comment l’entreprise fonctionne actuellement et détermineront la manière de mettre en œuvre rapidement des améliorations. Pensez également à inclure des personnes disposant de compétences supplémentaires ou spécialisées, telles que dans les domaines de l’analytique et de la gestion de projet.

Lorsque vous mettez en œuvre la sensibilisation aux coûts dans votre entreprise, améliorez ou appuyez-vous sur les programmes et processus existants. Il est beaucoup plus rapide d'ajouter des intégrations aux processus et programmes existants, que d'en créer de nouveaux. Ainsi, les résultats sont beaucoup plus rapides.

Sensibilisation aux dépenses et à l'utilisation

La flexibilité et la souplesse accrues que permet le cloud favorisent l'innovation, ainsi que le développement et le déploiement à un rythme soutenu. Il réduit les processus manuels et les délais associés au provisionnement d'une infrastructure sur site, y compris l'identification des spécifications matérielles, la négociation des devis, la gestion des bons de commande, la planification des livraisons et le déploiement des ressources. Cependant, la facilité d'utilisation et la capacité illimitée et à la demande nécessitent une nouvelle façon d'envisager les dépenses.

De nombreuses entreprises sont composées de plusieurs systèmes, dirigés par diverses équipes. La possibilité de répartir les coûts des ressources entre les différentes organisations ou les différents responsables de produits permet un comportement d'utilisation efficace et contribue à réduire le gaspillage. La répartition précise des coûts permet d'identifier les produits réellement rentables, et de prendre des décisions en connaissance de cause quant à la répartition du budget.

Dans AWS, vous créez une structure de compte avec AWS Organizations ou AWS Control Tower, qui assure la séparation et vous aide à répartir vos coûts et votre utilisation. Vous pouvez également utiliser le balisage des ressources pour appliquer les informations de l'entreprise à votre utilisation et à vos coûts. Utilisez-le AWS Cost Explorer pour avoir une visibilité sur vos coûts et votre utilisation, ou créez des tableaux de bord et des analyses personnalisés avec Amazon Athena et Amazon. QuickSight Le contrôle de vos coûts et de votre utilisation se fait par le biais de notifications via AWS les budgets, et de contrôles à l'aide de AWS Identity and Access Management (IAM) et de Quotas de Service.

Les questions suivantes sont axées sur ces quelques considérations relatives à l'optimisation des coûts.

COST2 : Comment régissez-vous l'utilisation ?

Définissez des stratégies et des mécanismes pour vous assurer que les coûts appropriés sont facturés tout en atteignant les objectifs. En utilisant une checks-and-balances approche, vous pouvez innover sans trop dépenser.

COST3 : Comment surveillez-vous l'utilisation et les coûts ?

Définissez des stratégies et des procédures pour surveiller et allouer vos coûts de manière appropriée. Cela vous permet d'évaluer et d'améliorer la rentabilité de cette charge de travail.

COST4 : Comment mettez-vous hors service les ressources ?

Mettez en œuvre le contrôle du changement et la gestion des ressources depuis le début du projet jusqu'à end-of-life. Cela facilite l'arrêt des ressources inutilisées afin de réduire le gaspillage.

Vous pouvez utiliser des balises de répartition des coûts pour catégoriser et suivre votre utilisation et vos coûts AWS . Lorsque vous appliquez des balises à vos AWS ressources (telles que des EC2 instances ou des compartiments S3), vous AWS générez un rapport sur les coûts et l'utilisation avec votre utilisation et vos balises. Vous pouvez appliquer des balises qui représentent des catégories de l'organisation (telles que les centres de coûts, les noms des charges de travail ou les propriétaires) pour organiser vos coûts dans plusieurs services.

Veillez à utiliser le niveau de détail et la granularité appropriés dans les rapports et la surveillance des coûts et de l'utilisation. Pour obtenir des informations de haut niveau et des tendances générales, utilisez la granularité quotidienne avec AWS Cost Explorer. Pour une analyse et une inspection plus approfondies, utilisez la granularité horaire dans AWS Cost Explorer, ou Amazon Athena et QuickSight Amazon avec le rapport sur les coûts et l'utilisation CUR () selon une granularité horaire.

La combinaison de ressources balisées et d'une fonction de suivi du cycle de vie des entités (employés, projets) permet d'identifier les ressources orphelines ou les projets qui ne génèrent plus de valeur pour l'organisation et qui doivent être mis hors service. Vous pouvez configurer des alertes de facturation pour être averti des dépassements de dépenses prévisibles.

Ressources rentables

L'utilisation d'instances et de ressources adaptées à votre charge de travail est l'élément essentiel à la réalisation d'économies. Par exemple, un processus de reporting peut prendre jusqu'à cinq heures pour s'exécuter sur un petit serveur, mais seulement une heure sur un serveur plus grand et deux fois plus cher. Vous obtiendrez les mêmes résultats avec les deux serveurs, mais le plus petit implique un coût plus élevé au fil du temps.

Une charge de travail Well-Architected utilise les ressources les plus rentables, ce qui peut avoir un impact économique positif et significatif. Vous pouvez également utiliser des services gérés pour réduire les coûts. Par exemple, plutôt que d'entretenir des serveurs pour envoyer des e-mails, vous pouvez utiliser un service effectuant une facturation au message.

AWS propose une variété d'options tarifaires flexibles et économiques pour acquérir des instances auprès d'Amazon EC2 et d'autres services de manière à mieux répondre à vos besoins. Les instances à la demande vous permettent de payer la capacité de calcul à l'heure, sans aucun engagement minimum. Avec les Savings Plans et les instances réservées, vous bénéficiez d'économies allant jusqu'à 75 % par rapport à la tarification à la demande. Avec les instances Spot, vous pouvez tirer parti de la EC2 capacité Amazon inutilisée et faire des économies allant jusqu'à 90 % par rapport à la tarification à la demande. Les instances Spot sont appropriées lorsque le système peut tolérer l'utilisation d'un parc de serveurs où les serveurs individuels peuvent entrer et sortir de manière dynamique, comme les serveurs Web apatrides, le traitement par lots ou lors de l'utilisation HPC de mégadonnées.

Une sélection de services appropriée peut également réduire l'utilisation et les coûts, par exemple CloudFront pour minimiser le transfert de données, ou pour réduire les coûts, par exemple en utilisant Amazon Aurora sur Amazon RDS pour supprimer les coûts élevés de licence de base de données.

Les questions suivantes sont axées sur ces quelques considérations relatives à l'optimisation des coûts.

COST5 : Comment évaluez-vous le coût lorsque vous sélectionnez des services ?

Amazon EC2EBS, Amazon et Amazon S3 sont des services de base. AWS Les services gérés, tels qu'Amazon RDS et Amazon DynamoDB, sont des services de niveau supérieur, ou de niveau application. AWS En sélectionnant les services fondamentaux et les services gérés appropriés, vous pouvez optimiser cette charge de travail en matière de coûts. Par exemple, en utilisant des services gérés, vous pouvez réduire ou supprimer une grande partie de votre traitement administr atif et opérationnel, et vous dégagez ainsi du temps pour travailler sur les applications et les activités liées aux activités.

COST6 : Comment atteignez-vous les objectifs de coûts lorsque vous sélectionnez le type, la taille et le nombre de ressources ?

Veillez à choisir la taille et le nombre de ressources qui conviennent pour la tâche à accomplir. En choisissant le type, la taille et le nombre les plus rentables, vous réduisez le gaspillage.

COST7 : Comment utilisez-vous les modèles de tarification pour réduire les coûts ?

Utilisez le modèle de tarification qui convient le mieux à vos ressources pour réduire les dépenses.

COST8 : Comment prévoyez-vous les frais de transfert de données ?

Veillez à planifier et à surveiller les frais de transfert de données afin de pouvoir prendre des décisions architecturales pour minimiser les coûts. Une modification architecturale minime, mais efficace, peut réduire de façon spectaculaire vos coûts d'exploitation.

En tenant compte des coûts lors de la sélection des services, en utilisant des outils tels que Cost Explorer et en AWS Trusted Advisor révisant régulièrement votre AWS utilisation, vous pouvez surveiller activement votre utilisation et ajuster vos déploiements en conséquence.

Gérer la demande et les sources d'approvisionnement

Lorsque vous migrez vers le cloud, vous ne payez que ce dont vous avez besoin. Vous pouvez fournir des ressources qui correspondent à la demande de la charge de travail au moment où elles sont nécessaires, ce qui réduit la nécessité d'un sur-approvisionnement coûteux et gaspilleur. Vous pouvez également modifier la demande à l'aide d'une limitation, d'une mémoire tampon ou d'une file d'attente pour la lisser et la gérer avec moins de ressources, ce qui réduit les coûts, ou la traiter ultérieurement avec un service de traitement par lots.

Dans AWS, vous pouvez automatiquement allouer des ressources pour répondre à la demande de charge de travail. Auto Scaling utilisant des approches basées sur la demande ou sur le temps vous permet d'ajouter et de supprimer des ressources en fonction des besoins. Si vous pouvez anticiper l'évolution de la demande, vous pouvez économiser plus et faire en sorte que vos ressources répondent aux besoins de la charge de travail. Vous pouvez utiliser Amazon API Gateway pour

implémenter la régulation, ou Amazon SQS pour implémenter une file d'attente dans votre charge de travail. Ces deux éléments vous permettent de modifier la demande sur les composants de votre charge de travail.

La question suivante est axée sur ces quelques considérations relatives à l’optimisation des coûts.

COST9 : Comment gérez-vous la demande et les ressources d'approvisionnement ?

Pour une charge de travail dont les dépenses et les performances sont équilibrées, assurez-vous que tout ce que vous payez est utilisé et évitez une sous-utilisation importante des instances. Un indicateur d'utilisation biaisé dans les deux sens a un impact négatif sur votre organisation, que ce soit en termes de coûts opérationnels (dégradation des performances due à une surutilisation) ou de AWS dépenses inutiles (en raison d'un provisionnement excessif).

Lorsque vous concevez dans le but de modifier la demande et l’offre de ressources, pensez activement aux modèles d’utilisation, au temps nécessaire pour allouer de nouvelles ressources et à la prévisibilité du modèle de la demande. Lors de la gestion de la demande, veillez à disposer d’une file d’attente ou d’une mémoire tampon correctement dimensionnée et à répondre à la demande de la charge de travail dans le délai requis.

Optimiser dans le temps

À AWS mesure que de nouveaux services et fonctionnalités sont lancés, il est recommandé de revoir vos décisions architecturales existantes afin de vérifier qu'elles restent les plus rentables. Lorsque vos besoins changent, n’hésitez pas à mettre hors service les ressources, les services entiers et les systèmes devenus inutiles.

La mise en œuvre de nouvelles fonctionnalités ou de nouveaux types de ressources peut optimiser votre charge de travail de façon progressive, tout en minimisant les efforts requis pour mettre en œuvre la modification. Ainsi, vous améliorez continuellement votre efficacité au fil du temps et vous restez au fait des technologies les plus récentes afin de réduire vos coûts d’exploitation. Vous pouvez également remplacer ou ajouter de nouveaux composants à la charge de travail avec de nouveaux services. Cela peut accroître considérablement l’efficacité. Il est donc essentiel de vérifier régulièrement votre charge de travail et de mettre en œuvre de nouveaux services et de nouvelles fonctionnalités.

Les questions suivantes sont axées sur ces quelques considérations relatives à l’optimisation des coûts.

COST10 : Comment évaluez-vous les nouveaux services ?

À AWS mesure que de nouveaux services et fonctionnalités sont lancés, il est recommandé de revoir vos décisions architecturales existantes afin de vérifier qu'elles restent les plus rentables.

En réexaminant régulièrement vos déploiements, évaluez dans quelle mesure des services plus récents peuvent vous permettre de réaliser des économies. Par exemple, Amazon Aurora sur Amazon RDS peut réduire les coûts des bases de données relationnelles. L’utilisation des technologies sans serveur, telle que Lambda, peut éviter d’exploiter et de gérer des instances pour exécuter du code.

COST11 : Comment évaluez-vous le coût de l'effort ?

Évaluez le coût des opérations dans le cloud, passez en revue vos opérations cloud fastidieuses et automatisez-les afin de réduire les efforts humains et les coûts en adoptant des AWS services connexes, des produits tiers ou des outils personnalisés.

Ressources

Consultez les ressources suivantes pour en savoir plus sur nos bonnes pratiques relatives à l’optimisation des coûts.

Documentation

- Documentation AWS

Livre blanc

- Cost Optimization Pillar

Durabilité

Le pilier Durabilité se concentre sur les impacts environnementaux, notamment la consommation et l’efficacité énergétiques, qui sont des leviers importants permettant aux architectes de recueillir

des informations sur les actions directes afin d'utiliser moins de ressources. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Durabilité.

Rubriques

Principes de conception

Il existe six principes de conception pour la durabilité dans le cloud :

- Comprendre votre impact : mesurez l'impact de votre charge de travail sur le cloud et modélisez l'impact futur de votre charge de travail. Incluez toutes les sources d'impact, y compris les impacts résultant de l'utilisation de vos produits par les clients, et ceux découlant de leur éventuelle mise hors service. Comparez le rendement productif à l'impact total de vos charges de travail sur le cloud en évaluant les ressources et les émissions nécessaires par unité de travail. Utilisez ces données pour établir des indicateurs de performance clés (KPIs), évaluer les moyens d'améliorer la productivité tout en réduisant l'impact, et estimer l'impact des modifications proposées au fil du temps.
- Établir des objectifs de durabilité : pour chaque charge de travail dans le cloud, établissez des objectifs de durabilité à long terme, tels que la réduction des ressources de calcul et de stockage nécessaires par transaction. Modélisez le retour sur investissement des améliorations durables pour les charges de travail existantes et donnez aux propriétaires les ressources dont ils ont besoin pour investir dans leurs objectifs de durabilité. Planifiez en vue d'une croissance et concevez l'architecture de vos charges de travail afin que la croissance entraîne une intensité de l'impact moindre mesurée par rapport à une unité appropriée, par utilisateur ou par transaction par exemple. Les objectifs vous aident à soutenir les cibles de durabilité plus larges de votre entreprise ou organisation, identifier les régressions et privilégier les zones pouvant être améliorées.
- Optimiser l'utilisation : dimensionnez correctement les charges de travail et intégrez une conception efficace pour assurer une forte utilisation et optimiser l'efficacité énergétique du matériel sous-jacent. Deux hôtes s'exécutant à 30 % de leur utilisation sont moins efficaces qu'un seul hôte s'exécutant à 60 % du fait de la consommation énergétique de base par hôte. Éliminez ou minimisez également les ressources, le traitement et le stockage inactifs afin de réduire l'énergie totale nécessaire pour alimenter votre charge de travail.

- Anticiper et adopter de nouvelles offres matérielles et logicielles plus efficaces : soutenez les améliorations en amont de vos partenaires et fournisseurs pour permettre de réduire l'impact de vos charges de travail sur le cloud. Contrôlez et évaluez de façon continue des offres matérielles et logicielles neuves et plus efficaces. Concevez de manière flexible afin de permettre l'adoption rapide de nouvelles technologies efficaces.
- Utiliser des services gérés : le partage des services auprès d'une clientèle importante permet de maximiser l'utilisation des ressources, ce qui réduit la quantité d'infrastructure nécessaire pour soutenir les charges de travail dans le cloud. Par exemple, les clients peuvent partager l'impact des composants courants des centres de données tels que l'alimentation et le réseau en migrant les charges de travail vers les services gérés AWS Cloud et en adoptant des services gérés, tels que AWS Fargate pour les conteneurs sans serveur, qui AWS opère à grande échelle et est responsable de leur fonctionnement efficace. Utilisez des services gérés qui peuvent vous aider à minimiser votre impact, tels que le transfert automatique des données rarement consultées vers un stockage à froid avec les configurations Amazon S3 Lifecycle ou Amazon EC2 Auto Scaling pour ajuster la capacité en fonction de la demande.
- Réduisez l'impact en aval de vos charges de travail dans le cloud : réduisez la quantité d'énergie ou de ressources nécessaires pour utiliser vos services. Réduisez ou supprimez le besoin pour les clients de mettre à niveau leurs appareils afin d'utiliser vos services. Réalisez des tests à l'aide de Device Farms pour comprendre l'impact attendu et auprès de clients pour comprendre l'impact réel que représente l'utilisation de vos services.

Définition

Il existe six domaines de bonnes pratiques en matière de durabilité dans le cloud :

- Sélection d'une région
- Alignement à la demande
- Logiciels et architecture
- Données
- Matériel et services
- Processus et culture

La durabilité dans le cloud est un effort continu axé principalement sur la réduction d'énergie et l'efficacité de tous les composants d'une charge de travail en tirant le meilleur parti possible des ressources allouées et en minimisant les ressources totales requises. Cet effort peut inclure la

sélection initiale d'un langage de programmation efficace, l'adoption d'algorithmes modernes, l'utilisation de techniques de stockage de données performantes, le déploiement sur une infrastructure de calcul correctement dimensionnée et efficace, et la réduction des besoins en matériel de grande puissance pour les utilisateurs finaux.

Bonnes pratiques

Rubriques

Sélection d'une région

Le choix de la région pour votre charge de travail influe de manière significative sur celle-ciKPIs, notamment en termes de performances, de coûts et d'empreinte carbone. Pour les améliorerKPIs, vous devez choisir des régions pour vos charges de travail en fonction à la fois des exigences commerciales et des objectifs de durabilité.

La question suivante est axée sur les considérations relatives à la durabilité. (Pour obtenir la liste des questions et bonnes pratiques liées à la durabilité, consultez l'Annexe.)

SUS1 : Comment sélectionnez-vous les régions pour votre charge de travail ?

Le choix de la région pour votre charge de travail influe de manière significative sur celle-ciKPIs, notamment en termes de performances, de coûts et d'empreinte carbone. Pour les améliorerKPIs, vous devez choisir des régions pour vos charges de travail en fonction à la fois des exigences commerciales et des objectifs de durabilité.

Alignement à la demande

La façon dont les utilisateurs et les applications consomment vos charges de travail et d'autres ressources peut vous aider à identifier les améliorations nécessaires pour atteindre vos objectifs de

durabilité. Mettez à l'échelle l'infrastructure pour répondre en permanence à la demande et vérifiez que vous n'utilisez que les ressources minimales requises pour prendre en charge vos utilisateurs. Alignez les niveaux de service sur les besoins des clients. Positionnez des ressources afin de limiter le réseau nécessaire aux utilisateurs et aux applications pour les consommer. Supprimez les ressources inutilisées. Fournissez aux membres de votre équipe des appareils qui répondent à leurs besoins et minimisent leur impact en matière de durabilité.

La question suivante est axée sur les considérations relatives à la durabilité :

SUS2 : Comment adaptez-vous les ressources du cloud à votre demande ?

La façon dont les utilisateurs et les applications consomment vos charges de travail et d'autres ressources peut vous aider à identifier les améliorations nécessaires pour atteindre vos objectifs de durabilité. Mettez à l'échelle l'infrastructure pour répondre en permanence à la demande et vérifiez que vous n'utilisez que les ressources minimales requises pour prendre en charge vos utilisateurs. Alignez les niveaux de service sur les besoins des clients. Positionnez des ressource s afin de limiter le réseau nécessaire aux utilisateurs et aux applications pour les consommer. Supprimez les ressources inutilisées. Fournissez aux membres de votre équipe des appareils qui répondent à leurs besoins et minimisent leur impact en matière de durabilité.

Mettre à l'échelle l'infrastructure avec la charge de l'utilisateur : identifiez les périodes d'utilisation faible ou nulle, et mettez vos ressources à l'échelle afin de réduire toute capacité excédentaire et de gagner en efficacité.

S'SLAsaligner sur les objectifs de durabilité : définissez et mettez à jour les accords de niveau de service (SLAs) tels que la disponibilité ou les périodes de conservation des données afin de minimiser le nombre de ressources nécessaires pour soutenir votre charge de travail tout en continuant à répondre aux exigences commerciales.

Réduire la création et la maintenance des actifs inutilisés : analyser les actifs des applications (tels que les rapports précompilés, les jeux de données et les images statiques) et les schémas d'accès aux actifs pour identifier la redondance, la sous-utilisation et les cibles potentielles de déclassement. Consolidez les ressources générées avec le contenu redondant (par exemple, des rapports mensuels avec des jeux de données et des résultats se chevauchant ou courants) pour réduire les ressources consommées lors de la duplication des résultats. Mettez hors service les ressources inutilisées (par exemple, des images de produits qui ne sont plus vendus) afin de libérer des ressources consommées et réduire le nombre de ressources utilisées afin de soutenir la charge de travail.

Optimiser l’emplacement géographique des charges de travail en fonction de la localisation des utilisateurs : analysez les modèles d’accès au réseau pour identifier les lieux de connexion de vos clients. Choisissez des régions et des services qui réduisent la distance que le trafic du réseau doit parcourir afin de diminuer le nombre total de ressources réseau nécessaires pour assurer votre charge de travail.

Optimiser les ressources des membres de l’équipe pour les activités réalisées : optimisez les ressources fournies aux membres de l’équipe pour réduire l’impact sur la durabilité tout en répondant à leurs besoins. Par exemple, effectuer des opérations complexes, telles que le rendu et la compilation, sur des bureaux partagés en nuage très utilisés plutôt que sur des systèmes à utilisateur unique sous-utilisés et très puissants.

Logiciels et architecture

Mettez en œuvre des modèles permettant de lisser les charges et de conserver une haute utilisation constante des ressources déployées afin de réduire les ressources consommées. Les composants peuvent devenir inactifs s’ils ne sont pas utilisés à la suite de changements de comportement des utilisateurs dans le temps. Révisez les modèles et l’architecture afin de consolider les composants sous-utilisés et d’augmenter l’utilisation globale. Mettez hors service les composants qui ne sont plus nécessaires. Comprenez les performances des composants de vos charges de travail et optimisez les composants qui consomment le plus de ressources. Soyez conscient des appareils que vos clients utilisent pour accéder à vos services et mettez en œuvre des modèles qui réduisent le besoin de mettre à niveau ces appareils.

La question suivante est axée sur les considérations relatives à la durabilité :

SUS3 : Comment tirer parti des modèles de logiciels et d'architecture pour atteindre vos objectifs de durabilité ?

Mettez en œuvre des modèles permettant de lisser les charges et de conserver une haute utilisation constante des ressources déployées afin de réduire les ressources consommées. Les composants peuvent devenir inactifs s’ils ne sont pas utilisés à la suite de changements de comportement des utilisateurs dans le temps. Révisez les modèles et l’architecture afin de consolider les composants sous-utilisés et d’augmenter l’utilisation globale. Mettez hors service les composants qui ne sont plus nécessaires. Comprenez les performances des composants de vos charges de travail et optimisez les composants qui consomment le plus de ressources. Soyez conscient des appareils que vos clients utilisent pour accéder à vos services et mettez en œuvre des modèles qui réduisent le besoin de mettre à niveau ces appareils.

Optimiser les logiciels et l'architecture pour les tâches asynchrones et prévues : utilisez des conceptions et des architectures logicielles efficaces pour réduire les ressources moyennes nécessaires par unité de travail. Mettez en œuvre des mécanismes qui entraînent une utilisation uniforme des composants pour réduire les ressources inactives entre deux tâches et réduire l'impact des pics de charge.

Supprimer ou refactoriser les composants de charges de travail faiblement utilisés ou inutilisés : surveillez l'activité de la charge de travail pour identifier des changements dans l'utilisation des composants individuels dans le temps. Supprimez les composants utilisés et qui ne sont plus nécessaires, et refactorisez les composants peu utilisés afin de limiter le gaspillage des ressources.

Optimiser les sections de votre code les plus longues ou qui consomment le plus de ressources : contrôlez l'activité de la charge de travail pour identifier les composants de l'application qui consomment le plus de ressources. Optimisez le code exécuté dans ces composants pour réduire l'utilisation des ressources tout en optimisant la performance.

Optimiser l'impact sur les appareils et les équipements des clients : ayez une compréhension des appareils et du matériel utilisés par vos clients pour consommer vos services, leur cycle de vie prévu et l'impact financier et durable que représente le remplacement de ces composants. Mettez en œuvre des modèles et des architectures logiciels pour réduire le besoin pour les clients de remplacer les appareils et de mettre à niveau leur matériel. Par exemple, mettez en œuvre de nouvelles fonctions en utilisant du code compatible avec du matériel et des versions de systèmes d'exploitation plus récents, ou gérez la taille des charges utiles afin qu'elles n'excèdent pas la capacité de stockage de l'appareil cible.

Utiliser des modèles et des architectures logicielles qui prennent en charge le plus efficacement possible les modèles d'accès aux données et de stockage : comprenez comment les données sont utilisées au sein de votre charge de travail, comment elles sont consommées par vos utilisateurs, transférées et stockées. Sélectionnez des technologies afin de réduire le traitement des données et les exigences de stockage.

Gestion des données

La question suivante est axée sur les considérations relatives à la durabilité :

SUS4 : Comment tirer parti des politiques et modèles de gestion des données pour atteindre vos objectifs de durabilité ?

Mettez en œuvre des pratiques de gestion des données afin de réduire le stockage alloué nécessaire pour assurer votre charge de travail et les ressources nécessaires à son utilisati on. Comprenez vos données et utilisez les technologies de stockage et les configurations qui soutiennent le plus efficacement la valeur commerciale des données et la façon dont elles sont utilisées. Adoptez un cycle de vie des données offrant un stockage plus efficace et moins performant quand les exigences baissent et supprimez les données qui ne sont plus nécessaires.

Mettre en œuvre une politique de classification des données : classez les données afin de déterminer leur importance pour les résultats commerciaux. Utilisez ces informations afin de déterminer quand déplacer vos données vers un stockage plus économe en énergie ou les supprimer en toute sécurité.

Utiliser des technologies qui prennent en charge les modèles d’accès aux données et de stockage : utilisez le stockage qui prend en charge le plus efficacement la manière dont vos données sont accédées et stockées afin de minimiser les ressources provisionnées tout en prenant en charge votre charge de travail. Par exemple, les périphériques SSD (SSDs) consomment plus d'énergie que les lecteurs magnétiques et ne doivent être utilisés que pour les cas d'utilisation de données actives. Utilisez un stockage de classe d’archivage économe en énergie pour les données rarement consultées.

Utiliser des politiques de cycle de vie pour supprimer les données inutiles : gérez le cycle de vie de toutes vos données et appliquez automatiquement des délais de suppression pour réduire l’ensemble des besoins de stockage de votre charge de travail.

Réduire le sur-approvisionnement dans le stockage par bloc : pour réduire au minimum le stockage alloué total, créez un stockage par bloc avec des allocations de taille adaptées à la charge de travail. Utilisez des volumes Elastic pour agrandir le stockage au fur et à mesure que les données augmentent sans avoir à redimensionner le stockage attaché aux ressources de calcul. Examinez régulièrement les volumes Elastic et réduisez les volumes sur-alloués pour qu’ils correspondent à la taille actuelle des données.

Supprimer les données inutiles ou redondantes : dupliquez les données uniquement lorsque cela s’avère nécessaire pour réduire le stockage total consommé. Utilisez des technologies de sauvegarde qui dédupliquent les données au niveau du fichier et du bloc. Limitez l'utilisation de

configurations de réseaux redondants de disques indépendants (RAID), sauf lorsque cela est nécessaire pour les respecterSLAs.

Utiliser des systèmes de fichiers partagés ou le stockage d'objets pour accéder aux données courantes : adoptez le stockage partagé et des sources uniques de confiance pour éviter la duplication des données et réduire l'ensemble des besoins en stockage pour votre charge de travail. Récupérez les données à partir du stockage partagé uniquement en fonction des besoins. Détachez les volumes inutilisés afin de libérer des ressources. Réduisez au minimum les déplacements des données entre les réseaux : utilisez le stockage partagé et accédez aux données des magasins de données régionaux pour réduire les ressources de réseaux totales nécessaires à la prise en charge des mouvements des données pour votre charge de travail.

Sauvegarder des données uniquement lorsqu'elles sont difficiles à recréer : afin de réduire la consommation de stockage, sauvegardez uniquement les données ayant une valeur opérationnelle ou nécessaires pour répondre aux exigences en matière de conformité. Examinez les politiques de sauvegarde et excluez tout magasin éphémère n'apportant aucune valeur dans un scénario de récupération.

Matériel et services

Recherchez des possibilités de réduire les impacts en matière de durabilité de la charge de travail en modifiant vos pratiques de gestion du matériel. Réduisez la quantité de matériel nécessaire à allouer et à déployer, et sélectionnez le matériel et les services les plus efficaces pour votre charge de travail individuelle.

La question suivante est axée sur les considérations relatives à la durabilité :

SUS5 : Comment sélectionnez-vous et utilisez-vous le matériel et les services cloud dans votre architecture pour atteindre vos objectifs de durabilité ?

Recherchez des possibilités de réduire les impacts en matière de durabilité de la charge de travail en modifiant vos pratiques de gestion du matériel. Réduisez la quantité de matériel nécessaire à allouer et à déployer, et sélectionnez le matériel et les services les plus efficaces pour votre charge de travail individuelle.

Utiliser la quantité minimale de matériel pour répondre à vos besoins : en utilisant les fonctionnalités du cloud, vous pouvez apporter régulièrement des modifications à vos mises en œuvre de charges de travail. Mettez à jour les composants déployés à mesure que vos besoins évoluent.

Utiliser les types d’instance ayant le moins d’impact : contrôlez de façon continue le lancement de nouveaux types d’instances et profitez d’améliorations de l’efficacité énergétique, y compris les types d’instances conçus pour soutenir des charges de travail spécifiques comme l’entraînement et l’inférence du machine learning et le transcodage vidéo.

Utiliser des services gérés : les services gérés transfèrent la responsabilité du maintien d'un taux d'utilisation moyen élevé et de l'optimisation de la durabilité du matériel déployé à AWS. Utilisez des services gérés pour distribuer l’impact de la durabilité du service sur tous les locataires du service, ce qui réduit votre contribution individuelle.

Optimisez votre utilisation de GPUs : les unités de traitement graphique (GPUs) peuvent être une source de forte consommation d'énergie, et de nombreuses GPU charges de travail sont très variables, telles que le rendu, le transcodage, la formation et la modélisation par machine learning. Exécutez les GPUs instances uniquement pendant le temps nécessaire et mettez-les hors service grâce à l'automatisation lorsque cela n'est pas nécessaire afin de minimiser la consommation de ressources.

Processus et culture

Recherchez des opportunités de réduire votre impact en matière de durabilité modifiant vos pratiques de développement, de test et de déploiement.

La question suivante est axée sur les considérations relatives à la durabilité :

SUS6 : Comment vos processus organisationnels soutiennent-ils vos objectifs de développement durable ?

Recherchez des opportunités de réduire votre impact en matière de durabilité modifiant vos pratiques de développement, de test et de déploiement.

Adopter des opérations qui peuvent rapidement présenter des améliorations en matière de durabilité : testez et validez les améliorations potentielles avant de les déployer en production. Tenez compte du coût des tests lors du calcul des avantages futurs potentiels d’une amélioration. Développer des opérations d’essai à faible coût pour favoriser la mise en œuvre de petites améliorations.

Maintenez votre charge de travail à jour : les systèmes Up-to-date d'exploitation, les bibliothèques et les applications peuvent améliorer l'efficacité de la charge de travail et favoriser l'adoption de technologies plus efficaces. Up-to-dateles logiciels peuvent également inclure des fonctionnalités

permettant de mesurer plus précisément l'impact de votre charge de travail sur le développement durable, car les fournisseurs proposent des fonctionnalités répondant à leurs propres objectifs de durabilité.

Augmenter l'utilisation de vos environnements de création : utilisez l'automatisation et l'infrastructure en tant que code pour mettre en place des environnements de préproduction lorsque cela est nécessaire et les arrêter lorsqu'ils ne sont pas utilisés. Un modèle courant consiste à planifier des périodes de disponibilité qui coïncident avec les heures de travail des membres de votre équipe de développement. La mise en veille prolongée est un outil pratique pour préserver l'état et mettre rapidement des instances en ligne uniquement lorsque cela est nécessaire. Utilisez des types d'instance avec une capacité de débordement, des instances Spot, des services de base de données Elastic, des conteneurs et d'autres technologies pour harmoniser la capacité de développement et de test avec l'utilisation.

Utiliser des tests Device Farms gérés pour effectuer les tests : les tests Device Farms gérés répartissent l'impact en matière de durabilité de la fabrication de matériel et de l'utilisation des ressources sur plusieurs locataires. Les parcs d'appareils gérés offrent divers types d'appareils, ce qui vous permet de prendre en charge du matériel plus ancien et moins populaire, et d'éviter l'impact sur la durabilité des clients des mises à niveau inutiles des appareils.

Ressources

Consultez les ressources suivantes pour en savoir plus sur nos bonnes pratiques relatives à la durabilité.

Livre blanc

- Pilier de durabilité

Vidéo

- The Climate Pledge

Processus de vérification

L’évaluation des architectures doit être effectuée de manière cohérente, avec une approche sans faute qui invite à étudier la situation en profondeur. Il doit s’agir d’un processus léger (qui dure des heures et non des jours), car il s’agit d’une conversation et non d’un audit. L’objectif de l’évaluation d’une architecture est d’identifier les problèmes critiques qui doivent être gérés, ou les domaines qui peuvent être améliorés. Le résultat de la révision est un ensemble d’actions ayant pour objectif d’améliorer l’expérience d’un client à l’aide de la charge de travail.

Comme indiqué dans la section « Sur l’architecture », il faut que chaque membre de l’équipe assume la responsabilité de la qualité de son architecture. Nous recommandons que les membres de l’équipe qui construisent une architecture utilisent le cadre Well-Architected pour revoir continuellement leur architecture, au lieu d’organiser une réunion formelle de révision. Une approche presque continue permet aux membres de votre équipe de mettre à jour des réponses au fur et à mesure que l’architecture évolue, et d’améliorer l’architecture lorsque vous fournissez des fonctions.

Le AWS Well-Architected Framework est aligné sur la manière dont les systèmes et services sont AWS examinés en interne. Il repose sur un ensemble de principes de conception qui influencent l'approche architecturale et sur des questions visant à vérifier que les gens ne négligent pas les domaines souvent mentionnés dans Root Cause Analysis ()RCA. Chaque fois qu'un problème important survient avec un système interne, un AWS service ou un client, nous examinons le problème RCA pour voir si nous pouvons améliorer les processus d'évaluation que nous utilisons.

Les évaluations doivent être effectuées à des étapes clés du cycle de vie du produit, dès le début de la phase de conception afin d’éviter les portes à sens unique difficiles à changer, puis avant la date de mise en service. (De nombreuses décisions concernent les portes bidirectionnelles réversibles. Ces décisions peuvent être prises à l’aide d’un processus léger. Les portes unidirectionnelles sont difficiles, voire impossibles, à inverser et nécessitent une inspection plus poussée avant de les fabriquer.) Après le lancement de la production, votre charge de travail continuera à évoluer à mesure que de nouvelles fonctions seront ajoutées et que les implémentations technologiques seront modifiées. L’architecture d’une charge de travail change au fil du temps. Vous devez suivre les bonnes pratiques d’hygiène pour empêcher ses caractéristiques architecturales de se dégrader au fur et à mesure de son évolution. Lorsque vous apportez des modifications d’architecture significatives, vous devez suivre un ensemble de processus d’hygiène et procéder à une évaluation Well-Architected.

Si vous souhaitez utiliser l’évaluation en tant qu’instantané unique ou mesure indépendante, vous devez vous assurer que vous avez inclus toutes les bonnes personnes dans la conversation.

Souvent, nous constatons que c'est lors des évaluations qu'une équipe comprend vraiment, pour la première fois, ce qu'elle a mis en œuvre. Une approche qui fonctionne bien lors de l'évaluation d'une autre charge de travail d'équipe est d'avoir une série de conversations informelles sur leur architecture, durant laquelle vous pouvez recueillir les réponses à la plupart des questions. Vous pouvez ensuite effectuer un suivi avec une ou deux réunions où vous pouvez gagner en clarté, ou des informations complètes sur les domaines d'ambiguïté ou les risques perçus.

Voici quelques suggestions pour faciliter vos réunions :

- Une salle de réunion avec des tableaux blancs
- Des impressions de tous les schémas ou notes de conception
- Liste d'actions contenant des questions auxquelles des out-of-band recherches sont nécessaires pour répondre (par exemple, « Avons-nous activé le chiffrement ou non ? »)

Une fois que vous avez effectué une évaluation, vous devez avoir une liste de questions que vous pouvez hiérarchiser en fonction de votre environnement métier. Vous devez également tenir compte de l'impact de ces problèmes sur le day-to-day travail de votre équipe. Si vous traitez ces problèmes tôt, vous pourrez libérer du temps pour travailler sur la création d'une valeur métier au lieu de résoudre des problèmes récurrents. Au fur et à mesure que vous traitez les problèmes, vous pouvez mettre à jour votre vérification pour voir comment l'architecture s'améliore.

Bien que la valeur ajoutée d'une évaluation d'architecture soit claire une fois l'exercice terminé, il est possible que vous rencontriez de la résistance de la part d'une nouvelle équipe au début. Voici quelques objections qui peuvent être traitées grâce à la sensibilisation de l'équipe sur les avantages d'une révision :

- « Nous sommes trop occupés ! » (Souvent déclaré lorsque l'équipe se prépare pour un lancement conséquent.)
  - Si vous préparez un grand lancement, vous voudrez qu'il se passe bien. La révision vous permettra de comprendre les problèmes que vous pourriez avoir manqués.
  - Nous vous recommandons de réaliser des révisions tôt dans le cycle de vie du produit pour découvrir les risques et développer un plan d'atténuation aligné avec la fonctionnalité de route de livraison.
- « Nous n'avons pas le temps de faire quoi que ce soit avec les résultats ! » (Souvent déclaré lorsqu'ils ciblent un événement fixe, comme le Super Bowl.)
  - Ces événements ne peuvent pas être déplacés. Voulez-vous vraiment vous aventurer dans cette entreprise sans connaître les risques liés à votre architecture ? Même si vous ne traitez pas

tous ces problèmes, vous pouvez toujours avoir des stratégies en place pour les gérer s'ils se concrétisent.

- « Nous ne voulons pas que les autres connaissent les secrets de l'implémentation de notre solution ! »
  - Si vous orientez l'équipe vers les questions qui figurent dans l'annexe du livre blanc sur Well-Architected Framework, elle verra qu'aucune des questions ne révèle d'informations propriétaires de nature technique ou commerciale.

Au fur et à mesure que vous effectuez des vérifications avec les équipes au sein de votre entreprise, vous pourrez identifier les problèmes récurrents. Par exemple, vous pourrez voir qu'un groupe d'équipes a rencontré divers problèmes liés à un pilier ou sujet particulier. Il est conseillé d'examiner toutes vos révisions d'une manière globale, et d'identifier tous les mécanismes, formations, ou les discussions d'ingénierie principale qui pourraient aider à traiter ces questions thématiques.

Conclusion

Le AWS Well-Architected Framework fournit les meilleures pratiques architecturales à travers les six piliers pour concevoir et exploiter des systèmes fiables, sécurisés, efficaces, rentables et durables dans le cloud. Ce cadre fournit un ensemble de questions qui vous permettent d'évaluer une architecture existante ou proposée. Il fournit également un ensemble de AWS bonnes pratiques pour chaque pilier. L'utilisation du cadre dans votre architecture vous aidera à produire des systèmes stables et efficaces, qui vous permettent de vous concentrer sur vos exigences fonctionnelles.

Collaborateurs

Les personnes et organisations suivantes ont contribué à l'élaboration du présent document :

- Brian Carlson, responsable des opérations de l'équipe Well-Architected, Amazon Web Services
- Ben Potter, responsable sécurité de l'équipe Well-Architected, Amazon Web Services
- Seth Eliot, responsable de la fiabilité de l'équipe Well-Architected, Amazon Web Services
- Eric Pullen, architecte de solutions senior, Amazon Web Services
- Rodney Lester, architecte principal des solutions, Amazon Web Services
- Jon Steel, responsable senior des comptes techniques, Amazon Web Services
- Max Ramsay, architecte principal de solutions de sécurité, Amazon Web Services
- Callum Hughes, architecte de solutions, Amazon Web Services
- Ben Mergen, architecte senior de solutions en charge des coûts, Amazon Web Services
- Chris Kozlowski, responsable senior des comptes techniques, support aux entreprises, Amazon Web Services
- Alex Livingstone, architecte principal de solutions spécialisé, opérations cloud, Amazon Web Services
- Paul Moran, technologue principal, support aux entreprises, Amazon Web Services
- Peter Mullen, consultant, services professionnels, Amazon Web Services
- Chris Pates, responsable senior des comptes techniques, support aux entreprises, Amazon Web Services
- Arvind Raghunathan, responsable principal des comptes techniques, support aux entreprises, Amazon Web Services
- Sam Mokhtari, architecte senior de solutions en matière d'efficacité, Amazon Web Services

Suggestions de lecture

Centre d’architecture AWS

Conformité dans le cloud AWS

AWS Programme de partenariat Well-Architected

AWS Well-Architected Tool

AWS Page d'accueil de Well-Architected

Livre blanc du pilier Excellence opérationnelle

Livre blanc du pilier Sécurité

Livre blanc du pilier Fiabilité

Livre blanc du pilier Efficacité des performances

Livre blanc du pilier Optimisation des coûts

Livre blanc du pilier Durabilité

Bibliothèque Amazon Builders’ Library

Révisions du document

Pour être informé des mises à jour de ce livre blanc, abonnez-vous au flux RSS.

Modification	Description	Date
Mise à jour majeure	Les bonnes pratiques ont été mises à jour avec de nouveaux conseils en matière de fiabilité, de sécurité, d'excellence opérationnelle, de durabilité et d'efficacité des performances. Le pilier Fiabilité a fait l'objet d'une actualisation et d'une mise à jour à grande échelle de nombreuses bonnes pratiques . Les conseils en matière de sécurité et d'excellence opérationnelle ont été mis à jour et affinés avec de nouveaux services et des suggestions d'IA générative. La durabilité a fait l'objet de plusieurs mises à jour basées sur les services AWS et une nouvelle bonne pratique.	6 novembre 2024
Mise à jour majeure	Des mises à jour à grande échelle des bonnes pratiques ont été effectuées dans l'ensemble des piliers. La sécurité et les coûts ont tous deux fait l'objet de nouvelles bonnes pratiques.	27 juin 2024

Mise à jour majeure	Mises à jour majeures des piliers.	3 octobre 2023
Mise à jour majeure	Les bonnes pratiques ont été mises à jour avec des recommandations et de nouvelles bonnes pratiques. De nouvelles questions ont été ajoutées aux piliers Sécurité et Optimisation des coûts.	10 avril 2023
Mise à jour mineure	Ajout d'une définition du niveau d'effort et mise à jour des bonnes pratiques dans l'annexe.	20 octobre 2022
Mise à jour majeure	Ajout du pilier Durabilité et mise à jour des liens.	2 décembre 2021
Mise à jour majeure	Ajout du pilier Durabilité dans le framework.	20 novembre 2021
Mise à jour mineure	Suppression du langage non inclusif.	22 avril 2021
Mise à jour mineure	Correction de plusieurs liens.	10 mars 2021
Mise à jour mineure	Modifications rédactionnelles mineures du document.	15 juillet 2020
Mise à jour majeure	Vérification et réécriture de la plupart des questions et réponses.	8 juillet 2020

Livre blanc mis à jour	Ajout du système AWS Well-Architected Tool et de liens vers les ateliers AWS Well-Architected et les partenair es AWS Well-Architected, correctifs mineurs pour prendre en charge la version multilingue du framework.	1 juillet 2019
Livre blanc mis à jour	Révision et reformulation de la plupart des questions et réponses, afin que les questions soient axées sur un thème à la fois. Certaines questions précédentes ont été divisées en plusieurs questions. Ajout de termes courants aux définitions (charge de travail, composant , etc.). Modification de la présentation des questions dans le corps principal pour inclure un texte descriptif.	1er novembre 2018
Livre blanc mis à jour	Mises à jour pour simplifier le texte des questions, normalise r les réponses et améliorer la lisibilité.	1er juin 2018
Livre blanc mis à jour	L’excellence opérationnelle est déplacée devant les piliers et récrite afin d’encadrer les autres domaines. Actualisation des autres piliers pour refléter l’évolution d’AWS.	1er novembre 2017

Livre blanc mis à jour	Mise à jour du cadre pour inclure un domaine d'excellence opérationnelle, révision et mise à jour des autres piliers pour réduire le dédoublement et intégrer les découvertes obtenues à partir des révisions avec des milliers de clients.	1er novembre 2016
Mises à jour mineures	Mise à jour de l'annexe avec les informations récentes d'Amazon CloudWatch Logs.	1er novembre 2015
Publication initiale	Publication d'AWS Well-Architected Framework.	1er octobre 2015

> **Note**
>
> Pour vous abonner aux mises à jour RSS, un plug-in RSS doit être activé pour le navigateur que vous utilisez.

Versions du cadre :

- 27-06-2024
- 03-10-2023
- 10-04-2023
- 31-03-2022

Annexe : questions et bonnes pratiques

Cette annexe résume toutes les questions et les bonnes pratiques du cadre AWS Well-Architected.

Piliers

- Excellence opérationnelle
- Sécurité
- Fiabilité
- Efficacité des performances
- Optimisation des coûts
- Durabilité

Excellence opérationnelle

L'excellence opérationnelle (OE) est un engagement à concevoir correctement un logiciel tout en offrant constamment une expérience client de qualité. Le pilier Excellence opérationnelle inclut les bonnes pratiques pour organiser votre équipe, concevoir votre charge de travail, la faire fonctionner à grande échelle et la faire évoluer au fil du temps. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Excellence opérationnelle.

Domaines de bonnes pratiques

- Organisation
- Préparation
- Exploitation
- Évolution

Organisation

Questions

- OPS 1. Comment déterminer vos priorités ?
- OPS 2. Comment structurer l'organisation pour soutenir les résultats de l'entreprise ?
- OPS 3. Comment votre culture organisationnelle soutient-elle vos résultats opérationnels ?

OPS 1. Comment déterminer vos priorités ?

Chacun doit comprendre le rôle qu'il a à jouer dans la réussite de l'entreprise. Établissez des objectifs partagés afin de définir des priorités pour les ressources. Cela permet de maximiser le fruit de vos efforts.

Bonnes pratiques

- OPS01-BP01 Évaluer les besoins des clients externes
- OPS01-BP02 Évaluer les besoins des clients internes
- OPS01-BP03 Évaluer les exigences de gouvernance
- OPS01-BP04 Évaluer les exigences de conformité
- OPS01-BP05 Évaluer le paysage des menaces
- OPS01-BP06 Évaluer les compromis tout en gérant les avantages et les risques

OPS01-BP01 Évaluer les besoins des clients externes

Impliquez les principales parties prenantes, notamment les équipes commerciales, de développement et d'exploitation, pour déterminer où il est nécessaire de concentrer les efforts sur les besoins des clients externes. Cela vous donnera une compréhension approfondie du soutien opérationnel nécessaire pour atteindre les résultats opérationnels souhaités.

Résultat escompté :

- Vous travaillez à rebours à partir des résultats des clients.
- Vous comprenez comment vos pratiques opérationnelles soutiennent les résultats et les objectifs de l'entreprise.
- Vous impliquez toutes les parties concernées.
- Vous disposez de mécanismes pour capturer les besoins des clients externes.

Anti-modèles courants :

- Vous avez décidé de ne pas bénéficier du service client en dehors des heures de bureau, mais vous n'avez pas examiné les données historiques des demandes d'assistance. Vous ne savez pas si cela aura un impact sur vos clients.

- Vous développez une nouvelle fonctionnalité, mais n'avez pas contacté vos clients pour déterminer si elle est souhaitée, sous quelle forme, et sans expérimentation pour valider le besoin et la méthode de distribution.

Avantages liés au respect de cette bonne pratique : les clients dont les besoins sont satisfaits sont beaucoup plus susceptibles de rester fidèles. L'évaluation et la compréhension des besoins des clients externes vous permettent d'établir des priorités dans vos efforts pour apporter de la valeur ajoutée à votre entreprise.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Identification des besoins de l'entreprise : le succès de l'entreprise repose sur des objectifs communs et une compréhension partagée entre les parties prenantes, y compris les équipes commerciales, de développement et d'exploitation.

Révision des objectifs de l'entreprise, des besoins et des priorités des clients externes : impliquez les acteurs clés, notamment, les équipes commerciales, de développement et d'exploitation, pour discuter des objectifs, besoins et priorités des clients externes. Cela permet de vérifier que vous comprenez bien le soutien opérationnel requis pour atteindre les résultats de l'entreprise et des clients.

Établissement d'une compréhension commune : établissez une compréhension commune des fonctions opérationnelles de la charge de travail, des rôles de chacune des équipes dans l'exploitation de la charge de travail, et de la manière dont ces facteurs soutiennent les objectifs opérationnels partagés chez les clients internes et externes.

Ressources

Bonnes pratiques associées :

- OPS11-BP03 Mettre en œuvre des boucles de rétroaction

OPS01-BP02 Évaluer les besoins des clients internes

Impliquez les principales parties prenantes, notamment les équipes commerciales, de développement et d'exploitation, lorsqu'il s'agit de déterminer où il est nécessaire de concentrer les efforts sur les besoins des clients internes. Ainsi, vous aurez une connaissance approfondie du soutien opérationnel requis pour atteindre les résultats opérationnels.

Résultat escompté :

- Tenez compte des priorités que vous avez établies pour concentrer vos efforts d'amélioration là où ils auront le plus d'impact (par exemple, le développement des compétences de l'équipe, l'amélioration des performances des charges de travail, la réduction des coûts, l'automatisation des runbooks ou encore l'amélioration de la surveillance).
- Mettez à jour vos priorités en fonction des besoins.

Anti-modèles courants :

- Vous avez décidé de modifier l'attribution des adresses IP de vos équipes de produits sans les consulter, afin de faciliter la gestion de votre réseau. Vous ne connaissez pas l'impact que cela aura sur vos équipes de produits.
- Vous mettez en place un nouvel outil de développement, mais vous n'avez pas demandé à vos clients internes s'ils en ont besoin ou s'il est compatible avec leurs pratiques existantes.
- Vous mettez en place un nouveau système de surveillance, mais vous demandez à vos clients internes s'ils ont des besoins en matière de surveillance ou de rapports à prendre en compte.

Avantages liés au respect de cette bonne pratique : l'évaluation et la compréhension des besoins des clients internes vous permettent d'établir des priorités dans vos efforts pour apporter de la valeur ajoutée à votre entreprise.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

- Identifiez les besoins de l'entreprise : la réussite repose sur des objectifs et une compréhension partagés entre les différents acteurs, y compris les équipes commerciales, de développement et d'opérations.
- Analysez les objectifs, les besoins et les priorités des clients internes : impliquez les acteurs clés, notamment, les équipes commerciales, du développement et des opérations, pour discuter des objectifs, besoins et priorités des clients internes. Cela permet de vérifier que vous comprenez bien le soutien opérationnel requis pour atteindre les résultats de l'entreprise et des clients.
- Établir une compréhension commune : établissez une compréhension commune des fonctions opérationnelles de la charge de travail, des rôles de chacune des équipes dans l'exploitation de la charge de travail, et de la manière dont ces facteurs soutiennent les objectifs opérationnels partagés chez les clients internes et externes.

Ressources

Bonnes pratiques associées :

- OPS11-BP03 Implémenter des boucles de rétroaction

OPS01-BP03 Évaluer les exigences de gouvernance

La gouvernance désigne l'ensemble des politiques, règles ou cadres qu'une entreprise utilise pour atteindre ses objectifs commerciaux. Les exigences en matière de gouvernance sont générées au sein de votre organisation. Elles peuvent affecter les types de technologies que vous choisissez ou influencer la façon dont vous gérez votre charge de travail. Incorporez les exigences de gouvernance organisationnelle dans votre charge de travail. La conformité désigne la capacité à prouver que vous avez mis en œuvre les exigences de gouvernance.

Résultat escompté :

- Les exigences de gouvernance sont intégrées à la conception architecturale et au fonctionnement de votre charge de travail.
- Vous pouvez fournir la preuve que vous avez suivi les exigences de gouvernance.
- Les exigences en matière de gouvernance sont régulièrement revues et mises à jour.

Anti-modèles courants :

- Votre organisation exige que le compte racine dispose d'une authentification multi-facteur. Vous n'avez pas mis en œuvre cette exigence et le compte racine est compromis.
- Lors de la conception de votre charge de travail, vous choisissez un type d'instance qui n'est pas approuvé par le service informatique. Vous ne parvenez pas à lancer votre charge de travail et devez procéder à une refonte.
- Vous êtes tenu de préparer un plan de reprise après sinistre. Vous n'en avez pas créé et votre charge de travail subit une interruption prolongée.
- Votre équipe souhaite utiliser de nouvelles instances, mais vos exigences de gouvernance n'ont pas été mises à jour pour les autoriser.

Avantages liés au respect de cette bonne pratique :

- Le respect des exigences de gouvernance permet d'aligner votre charge de travail sur les politiques de l'organisation dans son ensemble.
- Les exigences en matière de gouvernance reflètent les normes industrielles et les bonnes pratiques de votre organisation.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Identifiez les besoins en matière de gouvernance en travaillant avec les parties prenantes et les organisations de gouvernance. Incorporez les exigences de gouvernance à votre charge de travail. Soyez en mesure de prouver que vous avez respecté les exigences de gouvernance.

Exemple client

Chez AnyCompany Retail, l'équipe des opérations cloud travaille avec les parties prenantes de l'organisation pour développer les exigences de gouvernance. Par exemple, ils interdisent SSH l'accès aux EC2 instances Amazon. Si les équipes doivent accéder au système, elles doivent utiliser AWS Systems Manager Session Manager. L'équipe chargée des opérations dans le cloud met régulièrement à jour les exigences de gouvernance à mesure que de nouveaux services sont disponibles.

Étapes d'implémentation

1. Identifiez les parties prenantes de votre charge de travail, y compris toute équipe centralisée.
2. Travaillez avec les parties prenantes pour identifier les exigences de gouvernance.
3. Une fois que vous avez dressé une liste, classez les points à améliorer par ordre de priorité et commencez à les mettre en œuvre dans votre charge de travail.
   a. Utilisez des services tels que AWS Configpour créer governance-as-code et valider le respect des exigences de gouvernance.
   b. Si vous utilisez AWS Organizations, vous pouvez tirer parti des stratégies de contrôle des services pour mettre en œuvre les exigences de gouvernance.
4. Fournissez la documentation qui valide la mise en œuvre.

Niveau d'effort du plan d'implémentation : moyen. La mise en œuvre des exigences de gouvernance manquantes peut entraîner une refonte de votre charge de travail.

Ressources

Bonnes pratiques associées :

- OPS01-BP04 Évaluer les exigences de conformité – La conformité est similaire à la gouvernance, mais elle émane de l'extérieur de l'organisation.

Documents connexes :

- AWS Guide de gestion et de gouvernance de l'environnement cloud
- Meilleures pratiques en matière AWS Organizations de politiques de contrôle des services dans un environnement multi-comptes
- La gouvernance dans le AWS Cloud : le juste équilibre entre agilité et sécurité
- Qu'est-ce que la gouvernance, le risque et la conformité (GRC) ?

Vidéos connexes :

- AWS Gestion et gouvernance : configuration, conformité et audit - Discussions techniques AWS en ligne
- AWS RE:inForce 2019 : La gouvernance à l'ère du cloud (-R1) DEM12
- AWS re:Invent 2020 : Garantir la conformité sous forme de code en utilisant AWS Config
- AWS re:Invent 2020 : la gouvernance agile sur AWS GovCloud (US)

Exemples connexes :

- AWS Config Exemples de packs de conformité

Services connexes :

- AWS Config
- AWS Organizations - Politiques de contrôle des services

OPS01-BP04 Évaluer les exigences de conformité

Les exigences en matière de conformité réglementaire, sectorielle et interne constituent un facteur important pour définir les priorités de votre organisation. Votre cadre de conformité peut vous

empêcher d’utiliser des technologies ou des emplacements géographiques spécifiques. Appliquez les principes de diligence raisonnable si aucun cadre de conformité externe n’est identifié. Générez des audits ou des rapports qui valident la conformité.

Si vous mettez en avant le fait que votre produit respecte des normes de conformité spécifiques, vous devez mettre en place un processus interne pour assurer une conformité constante. Des exemples de normes de conformité incluent PCIDSS, FedRAMP, etHIPAA. Les normes de conformité applicables sont déterminées par divers facteurs, tels que les types des données stockées ou transmises par la solution et les régions géographiques prises en charge par la solution.

Résultat escompté :

- Les exigences en matière de conformité réglementaire, industrielle et interne sont intégrées dans le choix de l’architecture.
- Vous pouvez valider la conformité et générer des rapports d’audit.

Anti-modèles courants :

- Une partie de votre charge de travail est régie par le cadre de la norme de sécurité des données de l'industrie des cartes de paiement (PCI-DSS), mais votre charge de travail stocke les données des cartes de crédit de manière non cryptée.
- Vos développeurs et architectes de logiciels ne connaissent pas le cadre de conformité auquel votre organisation doit se conformer.
- L'audit annuel du contrôle des systèmes et des organisations (SOC2) de type II aura lieu prochainement et vous n'êtes pas en mesure de vérifier que les contrôles sont en place.

Avantages liés au respect de cette bonne pratique :

- L’évaluation et la compréhension des exigences de conformité qui s’appliquent à votre charge de travail détermineront la façon dont vous priorisez vos efforts pour produire de la valeur ajoutée.
- Vous choisissez les bons sites et les bonnes technologies, en accord avec votre cadre de conformité.
- La conception de votre charge de travail en vue de son auditabilité vous aide à prouver que vous adhérez à votre cadre de conformité.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

La mise en œuvre de cette bonne pratique signifie que vous intégrez les exigences de conformité dans votre processus de conception de l’architecture. Les membres de votre équipe connaissent le cadre de conformité requis. Vous validez la conformité conformément au cadre.

Exemple client

AnyCompany Informations sur les cartes de crédit des magasins de détail pour les clients. Les développeurs de l'équipe de stockage par carte savent qu'ils doivent se conformer au PCI DSS framework. Ils ont pris des mesures pour vérifier que les informations relatives aux cartes de crédit sont stockées et accessibles en toute sécurité conformément au PCI DSS cadre. Chaque année, ils travaillent avec leur équipe de sécurité pour valider la conformité.

Étapes d’implémentation

1. Travaillez avec vos équipes de sécurité et de gouvernance pour déterminer les cadres de conformité sectoriels, réglementaires ou internes auxquels votre charge de travail doit se conformer. Incorporez les cadres de conformité à votre charge de travail.
   a. Validez la conformité continue des AWS ressources avec des services tels que AWS Compute Optimizeret AWS Security Hub.
2. Informez les membres de votre équipe sur les exigences de conformité afin qu’ils puissent travailler et faire évoluer la charge de travail en fonction de celles-ci. Les exigences de conformité doivent être incorporées aux choix architecturaux et technologiques.
3. En fonction du cadre de conformité, vous pouvez être amené à générer un audit ou un rapport de conformité. Travaillez avec votre organisation pour automatiser ce processus autant que possible.
   a. Utilisez des services comme AWS Audit Manager pour valider des rapports de conformité et générer des rapports d’audit.
   b. Vous pouvez télécharger les documents AWS de sécurité et de conformité avec AWS Artifact.

Niveau d’effort du plan d’implémentation : moyen. La mise en œuvre de cadres de conformité peut s’avérer difficile. La génération de rapports d’audit ou de documents de conformité ajoute un niveau de complexité supplémentaire.

Ressources

Bonnes pratiques associées :

- SEC01-BP03 Identifier et valider les objectifs de contrôle - Les objectifs de contrôle de sécurité jouent un rôle important dans la conformité globale.
- SEC01-BP06 Automatisez les tests et la validation des contrôles de sécurité dans les pipelines - Dans le cadre de vos pipelines, validez les contrôles de sécurité. Vous pouvez également générer des documents de conformité pour les nouvelles modifications.
- SEC07-BP02 Définir les contrôles de protection des données - De nombreux cadres de conformité sont basés sur des politiques de gestion et de stockage des données.
- SEC10-BP03 Préparer les capacités de criminalistique - Les capacités de criminalistique peuvent parfois être utilisées pour auditer la conformité.

Documents connexes :

- AWS Centre de conformité
- AWS Ressources relatives à la conformité
- AWS Livre blanc sur les risques et la conformité
- AWS Modèle de responsabilité partagée
- AWS services visés par les programmes de conformité

Vidéos connexes :

- AWS re:Invent 2020 : Garantir la conformité sous forme de code en utilisant AWS Compute Optimizer
- AWS re:Invent 2021 - Conformité, assurance et audit du cloud
- AWS Sommet ATL 2022 - Mise en œuvre de la conformité, de l'assurance et de l'audit le AWS (COP202)

Exemples connexes :

- PCIDSSet les meilleures pratiques de sécurité AWS fondamentales sur AWS

Services connexes :

- AWS Artifact
- AWS Audit Manager

- AWS Compute Optimizer
- AWS Security Hub

OPS01-BP05 Évaluer le paysage des menaces

Évaluez les menaces pesant sur l'entreprise (par exemple, la concurrence, les risques commerciaux et les responsabilités, les risques opérationnels et les menaces sur la sécurité des informations) et tenez à jour les informations dans un registre des risques. Incluez l'impact des risques pour déterminer où concentrer les efforts.

Le cadre Well-Architected met l'accent sur la formation, la mesure et l'amélioration. Il fournit une approche cohérente qui vous permet d'évaluer les architectures et de mettre en œuvre des conceptions qui évolueront au fil du temps. AWS fournit les informations AWS Well-Architected Toolnécessaires pour vous aider à revoir votre approche avant le développement, l'état de vos charges de travail avant la production et l'état de vos charges de travail en production. Vous pouvez les comparer aux meilleures pratiques AWS architecturales les plus récentes, surveiller l'état général de vos charges de travail et avoir un aperçu des risques potentiels.

AWS les clients peuvent bénéficier d'un examen guidé par Well-Architected de leurs charges de travail critiques afin de mesurer leurs architectures par rapport aux meilleures pratiques. AWS Les clients Enterprise Support sont éligibles à une vérification des opérations conçue pour les aider à identifier les failles de leur approche d'exécution dans le cloud.

L'implication des équipes dans ces vérifications contribue à établir une compréhension partagée de vos charges de travail et de la façon dont les rôles de chacun contribuent à la réussite de l'équipe. Les besoins identifiés par la vérification peuvent vous aider à définir vos priorités.

AWS Trusted Advisor est un outil qui donne accès à un ensemble de base de vérifications qui recommandent des optimisations pouvant vous aider à définir vos priorités. Les clients du Business and Enterprise Support ont accès à des contrôles supplémentaires axés sur la sécurité, la fiabilité, les performances et l'optimisation des coûts qui peuvent les aider à définir leurs priorités.

Résultat escompté :

- Vous révisez et agissez régulièrement sur Well-Architected Trusted Advisor et ses résultats
- Vous êtes au courant de l'état des derniers correctifs de vos services.
- Vous comprenez le risque et l'impact des menaces connues et vous agissez en conséquence.
- Vous mettez en œuvre des mesures d'atténuation si nécessaire.

- Vous communiquez les actions et le contexte.

Anti-modèles courants :

- Vous utilisez une ancienne version d'une bibliothèque de logiciels dans votre produit. Vous n'êtes pas au courant des mises à jour de sécurité de la bibliothèque pour les questions qui peuvent avoir un impact involontaire sur votre charge de travail.
- Votre concurrent vient de lancer une version de son produit qui répond aux nombreuses plaintes de vos clients concernant votre produit. Vous n'avez pas priorisé la résolution de ces problèmes connus.
- Les régulateurs ont poursuivi des entreprises comme la vôtre qui ne respectaient pas les exigences légales de conformité réglementaire. Vous n'avez pas priorisé la résolution des vos exigences de conformité en suspens.

Avantages liés au respect de cette bonne pratique : l'identification et la compréhension des menaces qui pèsent sur votre organisation et votre charge de travail vous permettent de déterminer les menaces à traiter, leur priorité et les ressources nécessaires pour y parvenir.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

- Évaluation des menaces existantes : évaluez les menaces qui pèsent sur l'entreprise (par exemple, la concurrence, les risques commerciaux et les responsabilités, les risques opérationnels et les menaces sur la sécurité des données) afin de pouvoir tenir compte de leur impact lorsque vous déterminez où concentrer vos efforts.
  - Derniers bulletins de sécuritéAWS
  - AWS Trusted Advisor
- Gestion d'un modèle de menace : établissez et gérez un modèle de menace identifiant les menaces potentielles, les mesures d'atténuation prévues et en place, et leur priorité. Examinez la probabilité que les menaces se manifestent par des incidents, le coût de la récupération après ces incidents, le préjudice attendu et le coût de la prévention de ces incidents. Modifiez les priorités au fur et à mesure que le contenu du modèle de menace change.

Ressources

Bonne pratique associée :

- SEC01-BP07 Identifier les menaces et prioriser les mesures d'atténuation à l'aide d'un modèle de menace

Documents connexes :

- ConformitéAWS Cloud
- Derniers bulletins de sécuritéAWS
- AWS Trusted Advisor

Vidéos connexes :

- AWS re:Inforce 2023 - A tool to help improve your threat modeling

OPS01-BP06 Évaluer les compromis tout en gérant les avantages et les risques

Les intérêts divergents de plusieurs parties peuvent compliquer la hiérarchisation des efforts, la création de solutions et l'obtention de résultats conformes aux stratégies commerciales. Par exemple, il peut vous être demandé d'accélérer la mise en place speed-to-market de nouvelles fonctionnalités plutôt que d'optimiser les coûts de l'infrastructure informatique. Cela peut mettre deux parties intéressées en conflit l'une avec l'autre. Dans ces situations, les décisions doivent être portées devant une autorité supérieure pour résoudre le conflit. Des données sont nécessaires pour écarter l'attachement émotionnel du processus de prise de décision.

Le même défi peut se présenter au niveau tactique. Par exemple, le choix entre l'utilisation de technologies de base de données relationnelle ou non relationnelle peut avoir un impact significatif sur le fonctionnement d'une application. Il est essentiel de comprendre les résultats prévisibles des différents choix.

AWS peut vous aider à sensibiliser vos équipes à ses services afin qu'elles comprennent mieux comment leurs choix peuvent avoir un impact sur votre charge de travail. AWS Utilisez les ressources fournies par Support (Centre de connaissances AWS, forums de discussion AWS et Support Center) et la documentation AWS pour former vos équipes. Pour toute autre question, contactez Support.

AWS partage également les meilleures pratiques et modèles opérationnels dans The Amazon Builders' Library. Une grande variété d'autres informations utiles sont disponibles sur le AWS blog et le AWS podcast officiel.

Résultat escompté : vous disposez d'un cadre clairement défini de gouvernance de prise de décision pour faciliter les décisions importantes à tous les niveaux au sein de votre organisation de fourniture de cloud. Ce cadre comprend des fonctionnalités telles qu'un registre des risques, des rôles définis autorisés à prendre des décisions et un modèle défini pour chaque niveau de décision pouvant être prise. Ce cadre définit à l'avance comment les conflits sont résolus, quelles données doivent être présentées et comment les options sont hiérarchisées, de sorte qu'une fois les décisions prises, vous puissiez vous engager sans délai. Le cadre décisionnel comprend une approche normalisée pour examiner et évaluer les avantages et les risques de chaque décision afin de comprendre les compromis. Cela peut inclure des facteurs externes, tels que le respect des exigences de conformité réglementaires.

Anti-modèles courants :

- Vos investisseurs vous demandent de démontrer votre conformité aux normes de sécurité des données du secteur des cartes de paiement (PCIDSS). Vous n'envisagez pas les compromis entre la satisfaction de leur demande et la poursuite de vos efforts de développement actuels. Au lieu de cela, vous poursuivez vos efforts de développement sans en démontrer la conformité. Vos investisseurs cessent de soutenir votre entreprise en raison de préoccupations concernant la sécurité de votre plate-forme et de leurs investissements.
- Vous avez décidé d'inclure une bibliothèque que l'un de vos développeurs a trouvée sur Internet. Vous n'avez pas évalué les risques d'adoption de cette bibliothèque d'une source inconnue et ne savez pas si elle contient des vulnérabilités ou du code malveillant.
- La justification commerciale initiale de votre migration reposait sur la modernisation de 60 % des charges de travail de vos applications. Cependant, en raison de difficultés techniques, il a été décidé de ne moderniser que 20 %. Cela a entraîné une réduction des avantages prévus à long terme, une augmentation de la charge de travail des opérateurs pour la prise en charge manuelle des systèmes hérités par les équipes d'infrastructure et une plus grande dépendance au développement de nouvelles compétences au sein de vos équipes d'infrastructure qui ne prévoyaient pas ce changement.

Avantages de l'établissement de cette bonne pratique : harmonisation et prise en charge intégrales des priorités commerciales du conseil d'administration, compréhension des risques liés au succès, prise de décisions éclairées et action appropriée lorsque les risques entravent les chances de réussite. La compréhension des implications et des conséquences de vos décisions vous aide à hiérarchiser vos options et à amener les dirigeants à se mettre d'accord plus rapidement, ce qui se traduit par de meilleurs résultats commerciaux. En identifiant les avantages de vos choix et en étant

conscient des risques auxquels votre organisation est exposée, vous pouvez prendre des décisions fondées sur des données, plutôt que de vous fier à des anecdotes.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

La gestion des avantages et des risques doit être définie par un organe directeur qui gère les exigences relatives à la prise de décisions clés. Les décisions doivent être prises et classées par ordre de priorité en fonction de leurs avantages pour l’organisation, en comprenant les risques encourus. Des informations précises sont essentielles à la prise de décisions organisationnelles. Cet élément doit être basé sur des mesures solides et défini par les pratiques courantes du secteur en matière d’analyse des coûts par rapport aux avantages. Pour prendre ce type de décisions, il faut trouver un équilibre entre l’autorité centralisée et l’autorité décentralisée. Les compromis sont nécessaires, et il est important de comprendre l’impact de chaque choix sur les stratégies définies et les résultats commerciaux souhaités.

Étapes d’implémentation

1. Formalisez les pratiques de mesure des avantages dans un cadre de gouvernance du cloud holistique.
   a. Trouvez un juste milieu entre le contrôle décisionnel central et l’autorité décentralisée pour certaines décisions.
   b. Comprenez que les processus décisionnels fastidieux imposés à chaque décision peuvent vous ralentir.
   c. Intégrez des facteurs externes à votre processus de prise de décision (comme les exigences de conformité).
2. Établissez un cadre décisionnel convenu pour les différents niveaux de décision. Il doit notamment préciser qui est tenu de débloquer les décisions qui sont sujettes à des conflits d’intérêts.
   a. Centralisez les décisions à sens unique qui peuvent être irréversibles.
   b. Permettez aux responsables organisationnels de niveau inférieur de prendre des décisions bidirectionnelles.
3. Comprenez et gérez les avantages et les risques. Équilibrez les avantages des décisions par rapport aux risques impliqués.
   a. Identification des avantages : identifiez les avantages en fonction des objectifs, des besoins et des priorités de l’entreprise. Les exemples incluent l'impact sur l'analyse de time-to-market rentabilisation, la sécurité, la fiabilité, les performances et les coûts.

   b. Identification des risques : identifiez les risques en fonction des objectifs, des besoins et des priorités de l'entreprise. Les exemples incluent la sécurité time-to-market, la fiabilité, les performances et les coûts.
   c. Évaluation des avantages par rapport aux risques et prise de décisions avisées : déterminez l'impact des avantages et des risques en fonction des objectifs, des besoins et des priorités de vos parties prenantes clés, notamment les équipes commerciales, le développement et les opérations. Évaluez la valeur ajoutée de l'avantage par rapport à la probabilité de réalisation du risque et au coût de son impact. Par exemple, mettre l'accent speed-to-market sur la fiabilité peut apporter un avantage concurrentiel. Toutefois, cela peut entraîner une réduction du temps de fonctionnement en cas de problèmes de fiabilité.
4. Appliquez de manière programmatique les décisions clés qui automatisent le respect des exigences de conformité.
5. Tirez parti des cadres et capacités industriels connus, tels que l'analyse de la chaîne de valeur etLEAN, pour établir une base de référence des performances actuelles, des indicateurs commerciaux et définir des itérations des progrès réalisés en vue d'améliorer ces indicateurs.

Niveau d'effort du plan d'implémentation : moyen-élevé

Ressources

Bonnes pratiques associées :

- OPS01-BP05 Évaluer le paysage des menaces

Documents connexes :

- Éléments de la culture d'Amazon dès le premier jour | Prise de décisions rapides et éclairées
- Gouvernance du cloud
- Management and Governance Cloud Environment
- Gouvernance dans le cloud et à l'ère du numérique : première et deuxième parties

Vidéos connexes :

- Podcast | Jeff Bezos | On how to make decisions

Exemples connexes :

- Prendre des décisions éclairées en utilisant les données (The DevOps Sagas)
- Utiliser la cartographie de la chaîne de valeur du développement pour identifier les contraintes pesant sur les DevOps résultats

OPS 2. Comment structurer l'organisation pour soutenir les résultats de l'entreprise ?

Vos équipes doivent comprendre leur rôle dans l'obtention des résultats de l'entreprise. Les équipes doivent comprendre leur rôle dans la réussite des autres équipes, le rôle des autres équipes dans leur réussite, et avoir des objectifs communs. La compréhension de la responsabilité, de la manière dont les décisions sont prises et qui a le pouvoir de prendre des décisions vous aide à concentrer les efforts et à maximiser les avantages de vos équipes.

Bonnes pratiques

- OPS02-BP01 Les ressources ont identifié les propriétaires
- OPS02-BP02 Les processus et procédures ont des propriétaires identifiés
- OPS02-BP03 Les activités opérationnelles ont identifié les propriétaires responsables de leur performance
- OPS02-BP04 Des mécanismes existent pour gérer les responsabilités et la propriété
- OPS02-BP05 Des mécanismes existent pour demander des ajouts, des modifications et des exceptions
- OPS02-BP06 Les responsabilités entre les équipes sont prédéfinies ou négociées

OPS02-BP01 Les ressources ont identifié les propriétaires

Les ressources de votre charge de travail doivent disposer de propriétaires identifiés pour le contrôle des modifications, le dépannage et d'autres fonctions. Des propriétaires sont désignés pour les charges de travail, les comptes, l'infrastructure, les plateformes et les applications. La propriété est enregistrée à l'aide d'outils tels qu'un registre central ou des métadonnées attachées aux ressources. La valeur commerciale des composants informe les processus et les procédures qui leur sont appliqués.

Résultat escompté :

- Les ressources disposent de propriétaires identifiés à l'aide de métadonnées ou d'un registre central.
- Les membres de l'équipe peuvent identifier le propriétaire des ressources.

- Les comptes disposent d’un propriétaire unique dans la mesure du possible.

Anti-modèles courants :

- Les contacts alternatifs pour vous ne Comptes AWS sont pas renseignés.
- Les ressources manquent de balises permettant d’identifier les équipes qui les possèdent.
- Vous avez une ITSM file d'attente sans mappage d'e-mails.
- Deux équipes se partagent la propriété d’un élément d’infrastructure critique.

Avantages liés au respect de cette bonne pratique :

- Le contrôle des modifications pour les ressources est simple et la propriété est attribuée.
- Vous pouvez impliquer les bons propriétaires lors du dépannage des problèmes.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Définissez ce que signifie la propriété pour les cas d’utilisation des ressources dans votre environnement. La propriété peut signifier qui supervise les modifications apportées à la ressource, qui prend en charge la ressource pendant le dépannage, ou qui est financièrement responsable. Précisez et enregistrez les propriétaires des ressources, y compris, le nom, les coordonnées, l’organisation et l’équipe.

Exemple client

AnyCompany Le commerce de détail définit la propriété comme l'équipe ou l'individu responsable des changements et du soutien aux ressources. Ils tirent parti AWS Organizations de leur Comptes AWS. Les autres contacts de comptes sont configurés via des boîtes de réception de groupe. Chaque ITSM file d'attente est mappée à un alias d'e-mail. Les balises identifient les propriétaires AWS des ressources. Pour les autres plateformes et infrastructures, ces personnes disposent d’une page wiki qui identifie les propriétaires et les informations de contact.

Étapes d’implémentation

1. Commencez par définir la propriété dans votre organisation. La propriété peut impliquer qui est responsable du risque pour la ressource, qui est responsable des modifications apportées à la

ressource, ou qui prend en charge la ressource lors du dépannage. La propriété peut également impliquer la propriété financière ou administrative de la ressource.

2. Utilisez AWS Organizations pour gérer les comptes. Vous pouvez gérer les autres contacts de vos comptes de manière centralisée.
   a. Grâce aux adresses e-mail et aux numéros de téléphone appartenant à l'entreprise, vous pourrez y accéder même si les personnes qui les consultent ne font plus partie de votre entreprise. Par exemple, créez des listes de distribution d'e-mails distinctes pour la facturation, les opérations et la sécurité, et configurez-les en tant que contacts Facturation, Sécurité et Opérations dans chaque Compte AWS actif. Plusieurs personnes recevront des AWS notifications et pourront y répondre, même si une personne est en vacances, change de rôle ou quitte l'entreprise.
   b. Si un compte n'est pas géré par AWS Organizations, d'autres contacts de compte aident AWS à contacter le personnel approprié si nécessaire. Configurez les autres contacts du compte pour qu'ils pointent vers un groupe plutôt que vers un individu.
3. Utilisez des balises pour identifier les propriétaires des AWS ressources. Vous pouvez indiquer les deux propriétaires et leurs coordonnées dans des balises distinctes.
   a. Vous pouvez utiliser des règles AWS Config pour garantir que les ressources possèdent les balises de propriété requises.
   b. Pour obtenir des conseils détaillés sur la façon d'élaborer une stratégie de balisage pour votre organisation, consultez le livre blanc sur les bonnes pratiques en matière de balisage AWS.
4. Utilisez Amazon Q Business, un assistant conversationnel qui utilise l'IA générative pour améliorer la productivité du personnel, répondre aux questions et effectuer des tâches en fonction des informations contenues dans les systèmes de votre entreprise.
   a. Connectez Amazon Q Business à la source de données de votre entreprise. Amazon Q Business propose des connecteurs prédéfinis vers plus de 40 sources de données prises en charge, notamment Amazon Simple Storage Service (Amazon S3), SharePoint Microsoft, Salesforce et Atlassian Confluence. Pour plus d'informations, consultez la section Connecteurs Amazon Q Business.
5. Pour les autres ressources, plateformes et infrastructures, créez une documentation qui identifie la propriété. Tous les membres de l'équipe doivent y avoir accès.

Niveau d'effort du plan d'implémentation : faible Utilisez les informations de contact et les tags du compte pour attribuer la propriété des AWS ressources. Pour les autres ressources, vous pouvez

utiliser quelque chose d'aussi simple qu'un tableau dans un wiki pour enregistrer les informations de propriété et de contact, ou utiliser un ITSM outil pour cartographier les propriétaires.

Ressources

Bonnes pratiques associées :

- OPS02-BP02 Les processus et les procédures ont identifié les propriétaires
- OPS02-BP04 Des mécanismes existent pour gérer les responsabilités et la propriété

Documents connexes :

- Gestion des comptes AWS : mise à jour des informations de contact
- AWS Organizations - Mise à jour des contacts alternatifs au sein de votre organisation
- Livre blanc des Bonnes pratiques de balisage AWS
- Créez des applications d'IA génératives d'entreprise privées et sécurisées avec Amazon Q Business and AWS IAM Identity Center
- Amazon Q Business, désormais disponible pour le grand public, contribue à améliorer la productivité du personnel grâce à l'IA générative
- AWS Cloud Blog sur les opérations et les migrations - Mise en œuvre de contrôles de balisage automatisés et centralisés avec AWS Config et AWS Organizations
- AWS Blog de sécurité - Étendez vos hooks de pré-validation avec AWS CloudFormation Guard
- AWS DevOps Blog - Intégration AWS CloudFormation Guard dans les pipelines CI/CD

Ateliers connexes :

- Atelier AWS : étiquetage

Exemples connexes :

- AWS Config Rules - Amazon EC2 avec les balises obligatoires et les valeurs valides

Services connexes :

- AWS Config Rules - étiquettes obligatoires
- AWS Organizations

OPS02-BP02 Les processus et procédures ont des propriétaires identifiés

Déterminez qui est propriétaire de la définition des différents processus et procédures individuels, pourquoi ces processus et procédures sont utilisés et pourquoi cette propriété existe. La compréhension des raisons pour lesquelles des processus et des procédures spécifiques sont utilisés permet d'identifier les possibilités d'amélioration.

Résultat escompté : votre organisation dispose d'un ensemble défini et géré de processus et de procédures pour les tâches opérationnelles. Le processus et les procédures sont stockés dans un emplacement central et mis à la disposition des membres de votre équipe. Le processus et les procédures sont fréquemment mis à jour, par un propriétaire clairement désigné. Dans la mesure du possible, les scripts, les modèles et les documents d'automatisation sont implémentés sous forme de code.

Anti-modèles courants :

- Les processus ne sont pas documentés. Des scripts fragmentés peuvent exister sur les postes de travail des opérateurs isolés.
- La connaissance de l'utilisation des scripts est détenue par quelques personnes ou de manière informelle en tant que connaissance d'équipe.
- Un ancien processus doit être actualisé, mais la propriété de l'actualisation est incertaine et l'auteur d'origine ne fait plus partie de l'organisation.
- Les processus et les scripts ne sont pas détectables, ils ne sont donc pas facilement disponibles en cas de besoin (par exemple, pour répondre à un incident).

Avantages liés au respect de cette bonne pratique :

- Les processus et les procédures dynamisent vos efforts pour gérer vos charges de travail.
- Les nouveaux membres de l'équipe deviennent efficaces plus rapidement.
- Réduction du temps nécessaire pour atténuer les incidents.
- Différents membres de l'équipe (et différentes équipes) peuvent utiliser les mêmes processus et procédures de manière cohérente.
- Les équipes peuvent mettre à l'échelle leurs processus à l'aide de processus reproductibles.
- Les processus et procédures normalisés contribuent à atténuer l'impact du transfert des responsabilités liées à la charge de travail entre les équipes.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

- Les processus et procédures ont un propriétaire identifié qui est responsable de leur définition.
  - Identifiez les activités des opérations réalisées à l’aide de vos charges de travail. Documentez ces activités dans un emplacement détectable.
  - Identifiez de façon unique l’individu ou l’équipe responsable de la spécification d’une activité. Il incombe à l’individu ou à l’équipe de vérifier qu’elle peut être exécutée avec succès par un membre de l’équipe disposant des autorisations, des accès et des outils appropriés. En cas de problème lié à l’exécution de l’activité, les membres de l’équipe chargés de cette tâche sont tenus de fournir les commentaires détaillés nécessaires à son amélioration.
  - Capturez la propriété des métadonnées de l’artefact d’activité par le biais de services tels qu’AWS Systems Manager, via des documents, et AWS Lambda. Capturez la propriété des ressources à l’aide de balises ou de groupes de ressources, en spécifiant les informations de propriété et de contact. Utilisez AWS Organizations pour créer des stratégies de balisage et capturer les informations de propriété et de contact.
- Au fil du temps, ces procédures doivent évoluer pour être exécutables sous forme de code, ce qui réduit la nécessité d’une intervention humaine.
  - Réfléchissez par exemple aux fonctions AWS Lambda, aux modèles CloudFormation ou aux documents d’automatisation AWS Systems Manager.
  - Effectuez le contrôle des versions dans les référentiels appropriés.
  - Incluez un balisage approprié des ressources afin que les propriétaires et la documentation puissent être facilement identifiés.

Exemple client

AnyCompany Retail définit la propriété comme l’équipe ou la personne qui possède les processus d’une application ou de groupes d’applications (qui partagent des pratiques et des technologies architecturales communes). Dans un premier temps, le processus et les procédures sont documentés sous forme de guides détaillés dans le système de gestion de documents, détectables à l’aide de balises sur le Compte AWS qui héberge l’application et sur des groupes spécifiques de ressources du compte. Ces personnes utilisent AWS Organizations pour gérer leurs Comptes AWS. Au fil du temps, ces processus sont convertis en code et les ressources sont définies à l’aide de l’infrastructure sous forme de code (par exemple, les modèles CloudFormation ou AWS Cloud Development Kit (AWS CDK)). Les processus opérationnels deviennent des documents d’automatisation dans AWS Systems

Manager ou des fonctions AWS Lambda, que vous pouvez lancer en tant que tâches planifiées, en réponse à des événements tels que des alarmes AWS CloudWatch ou des événements AWS EventBridge, ou que vous pouvez démarrer par des demandes au sein d'une plateforme de gestion des services informatiques (ITSM). Tous les processus comportent des balises pour identifier leur propriété. La documentation relative à l'automatisation et au processus est conservée dans les pages wiki générées par le référentiel de code pour le processus.

Étapes d'implémentation

1. Documentez les processus et procédures existants.
    a. Révisez-les et veillez à leur actualisation.
    b. Identifiez un propriétaire pour chaque processus ou procédure.
    c. Placez-les sous le contrôle des versions.
    d. Dans la mesure du possible, partagez les processus et les procédures entre les charges de travail et les environnements qui ont des conceptions architecturales en commun.
2. Mettez en place des mécanismes de commentaires et d'amélioration.
    a. Définissez des politiques relatives à la fréquence à laquelle les processus doivent être révisés.
    b. Définissez les processus pour les réviseurs et les approbateurs.
    c. Consignez les problèmes ou établissez des files d'attente de tickets afin que les commentaires puissent être transmis et faire l'objet d'un suivi.
    d. Dans la mesure du possible, les processus et procédures doivent faire l'objet d'une approbation préalable et d'une classification des risques par un comité d'approbation des modifications (CAB).
3. Vérifiez que les processus et les procédures sont accessibles et détectables par ceux qui ont besoin de les exécuter.
    a. Utilisez des balises pour indiquer où le processus et les procédures sont accessibles pour la charge de travail.
    b. Utilisez des messages d'erreur et d'événements significatifs afin d'indiquer les processus ou procédures appropriés pour résoudre un problème.
    c. Utilisez les wikis et la gestion des documents, et veillez à ce que les processus et les procédures puissent être consultés par l'ensemble de l'organisation.
4. Utilisez Amazon Q Business, un assistant conversationnel qui utilise l'IA générative pour améliorer la productivité du personnel, répondre aux questions et effectuer des tâches en fonction des informations contenues dans les systèmes de votre entreprise.

   a. Connectez Amazon Q Business à la source de données de votre entreprise. Amazon Q Business propose des connecteurs prédéfinis vers plus de 40 sources de données prises en charge, notamment Amazon S3, Microsoft SharePoint, Salesforce et Atlassian Confluence. Pour plus d'informations, consultez Connecteurs Amazon Q.

5. Automatisez le cas échéant.
   a. Les automatisations doivent être développées lorsque les services et les technologies fournissent une API.
   b. Formez de manière adéquate aux processus. Développez les témoignages d'utilisateurs et les exigences pour automatiser ces processus.
   c. Mesurez l'utilisation réussie de vos processus et procédures, en tenant compte des problèmes permettant une amélioration itérative.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS02-BP01 Les ressources ont des propriétaires identifiés
- OPS02-BP04 Des mécanismes sont en place pour gérer les responsabilités et qui est responsable de quoi
- OPS11-BP04 Gestion des connaissances

Documents connexes :

- Livre blanc AWS : présentation du DevOps sur AWS
- Livre blanc AWS : bonnes pratiques en matière de balisage des ressources AWS
- Livre blanc AWS : organisation de votre environnement AWS à l'aide de comptes multiples
- Blog sur les migrations et opérations cloud AWS Cloud : utilisation d'Amazon Q Business pour rationaliser vos opérations
- Blog sur les opérations et les migrations AWS Cloud : Build a Cloud Automation Practice for Operational Excellence: Best Practices from AWS Managed Services
- Blog sur les opérations et les migrations AWS Cloud : mise en œuvre de contrôles de balisage automatisés et centralisés avec AWS Config et AWS Organizations

- Blog de sécurité AWS : extension de vos hooks de pré-validation avec AWS CloudFormation Guard
- Blog DevOps AWS : intégration de AWS CloudFormation Guard dans des pipelines CI/CD

Ateliers connexes :

- Atelier sur l'excellence opérationnelle Well-Architected AWS
- Atelier AWS : étiquetage

Vidéos connexes :

- Comment automatiser des opérations informatiques sur AWS
- AWS re:Invent 2020 - Automate anything with AWS Systems Manager
- AWS re:Inforce 2022 - Automating patch management and compliance using AWS (NIS306)
- Supports You - Diving Deep into AWS Systems Manager

Services connexes :

- AWS Systems Manager : automatisation
- Connecteur AWS Service Management

OPS02-BP03 Les activités opérationnelles ont identifié les propriétaires responsables de leur performance

Déterminez qui est chargé d'exécuter des activités spécifiques sur des charges de travail définies et pourquoi cette responsabilité existe. La détermination de la personne responsable de l'exécution des activités indique qui va mener l'activité, valider le résultat et fournir des commentaires au propriétaire de l'activité.

Résultat escompté :

Votre organisation définit clairement les responsabilités relatives à l'exécution d'activités spécifiques sur des charges de travail définies et répond aux événements générés par la charge de travail. L'organisation documente la propriété des processus et de leur exécution et rend ces informations détectables. Vous passez en revue et mettez à jour les responsabilités lorsque des changements organisationnels se produisent, et les équipes suivent et mesurent les performances des activités

d'identification des défauts et des inefficacités. Vous mettez en œuvre des mécanismes de rétroaction pour suivre les défauts et les améliorations et soutenir l'amélioration itérative.

Anti-modèles courants :

- Vous ne documentez pas les responsabilités.
- Il existe des scripts fragmentés sur les postes de travail d'opérateurs isolés. Seules quelques personnes savent comment les utiliser ou les qualifier de manière informelle de connaissances d'équipe.
- Un ancien processus doit être mis à jour, mais personne ne sait qui en a la responsabilité, et l'auteur d'origine ne fait plus partie de l'organisation.
- Les processus et les scripts ne sont pas détectables, ils ne sont donc pas facilement disponibles en cas de besoin (par exemple, pour répondre à un incident).

Avantages liés au respect de cette bonne pratique :

- Vous savez qui est responsable de l'exécution d'une activité, qui avertir lorsqu'une action est nécessaire et qui exécute l'action, qui valide le résultat et qui fournit des commentaires au responsable de l'activité.
- Les processus et les procédures dynamisent vos efforts pour gérer vos charges de travail.
- Les nouveaux membres de l'équipe deviennent efficaces plus rapidement.
- Vous réduisez le temps nécessaire pour atténuer les incidents.
- Les différentes équipes utilisent les mêmes processus et procédures pour effectuer les tâches de manière cohérente.
- Les équipes peuvent mettre à l'échelle leurs processus à l'aide de processus reproductibles.
- Les processus et procédures normalisés contribuent à atténuer l'impact du transfert des responsabilités liées à la charge de travail entre les équipes.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Pour commencer à définir les responsabilités, commencez par la documentation existante, comme les matrices de responsabilité, les processus et les procédures, les rôles et les responsabilités, ainsi que les outils et l'automatisation. Passez en revue et animez des discussions sur les responsabilités

relatives aux processus documentés. Passez en revue les responsabilités avec les équipes pour identifier les incohérences entre les responsabilités et les processus des documents. Discutez des services proposés avec les clients internes de cette équipe afin d'identifier les écarts entre les équipes en matière d'attentes.

Analysez et corrigez les écarts. Identifiez les opportunités d'amélioration et recherchez les activités gourmandes en ressources et fréquemment demandées, qui sont généralement de bonnes candidates à l'amélioration. Explorez les bonnes pratiques, les modèles et les conseils prescriptifs pour simplifier et standardiser les améliorations. Enregistrez les opportunités d'amélioration et suivez les améliorations jusqu'à leur achèvement.

Au fil du temps, ces procédures doivent évoluer pour être exécutées sous forme de code, ce qui réduit la nécessité d'une intervention humaine. Par exemple, les procédures peuvent être initiées sous forme de AWS Lambda fonctions, AWS CloudFormation de modèles ou de documents AWS Systems Manager d'automatisation. Vérifiez que ces procédures sont contrôlées par version dans les référentiels appropriés et incluez un balisage des ressources adéquat afin que les équipes puissent identifier facilement les personnes responsables et la documentation. Documentez la responsabilité de l'exécution des activités, puis surveillez les automatisations pour garantir un démarrage et un fonctionnement réussis, ainsi que la performance des résultats souhaités.

Exemple client

AnyCompany Le commerce de détail définit la propriété comme l'équipe ou l'individu responsable des processus d'une application ou de groupes d'applications partageant des pratiques et des technologies architecturales communes. Dans un premier temps, l'entreprise documente les processus et les procédures sous forme de step-by-step guides dans le système de gestion des documents. Ils permettent de découvrir les procédures à l'aide de balises situées sur le Compte AWS serveur qui héberge l'application et sur des groupes de ressources spécifiques au sein du compte, AWS Organizations afin de les Comptes AWS gérer. Au fil du temps, AnyCompany Retail convertit ces processus en code et définit les ressources en utilisant l'infrastructure sous forme de code (via des services tels que CloudFormation des AWS Cloud Development Kit (AWS CDK) modèles). Les processus opérationnels deviennent des documents d'automatisation dans AWS Systems Manager ou dans AWS Lambda les fonctions, qui peuvent être lancés sous forme de tâches planifiées en réponse à des événements tels que des CloudWatch alarmes Amazon ou des EventBridge événements Amazon ou par des demandes au sein d'une plateforme de gestion des services informatiques (ITSM). Tous les processus ont des balises pour identifier qui en est le responsable. Les équipes gèrent la documentation relative à l'automatisation et au processus dans les pages wiki générées par le référentiel de code pour ce processus.

Étapes d’implémentation

1. Documentez les processus et procédures existants.
   a. Vérifiez et vérifiez qu'ils le sont up-to-date.
   b. Vérifiez que chaque processus ou procédure est associé à un responsable.
   c. Placez les procédures sous contrôle des versions.
   d. Dans la mesure du possible, partagez les processus et les procédures entre les charges de travail et les environnements qui ont des conceptions architecturales en commun.
2. Mettez en place des mécanismes de commentaires et d’amélioration.
   a. Définissez des politiques relatives à la fréquence à laquelle les processus doivent être révisés.
   b. Définissez les processus pour les réviseurs et les approbateurs.
   c. Mettez en œuvre une file d’attente de problèmes ou de tickets pour fournir et suivre les commentaires.
   d. Dans la mesure du possible, fournissez une approbation préalable et une classification des risques pour les processus et les procédures par un comité d'approbation des modifications (CAB).
3. Rendez les processus et les procédures accessibles et détectables par les utilisateurs qui ont besoin de les exécuter.
   a. Utilisez des balises pour indiquer où le processus et les procédures sont accessibles pour la charge de travail.
   b. Utilisez des messages d’erreur et d’événements significatifs afin d’indiquer les processus ou procédures appropriés pour résoudre le problème.
   c. Utilisez les wikis ou la gestion de documents pour rendre les processus et les procédures consultables de manière cohérente dans l’ensemble de l’organisation.
4. Recourez à l’automatisation lorsque cela est approprié.
   a. Là où les services et les technologies fournissent et développent des automatisations. API
   b. Vérifiez que les processus sont bien compris et développez les témoignages d’utilisateurs et les exigences pour automatiser ces processus.
   c. Mesurez l’utilisation réussie des processus et des procédures, avec un suivi des problèmes pour favoriser une amélioration itérative.

Niveau d’effort du plan d’implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS02-BP01 Les ressources ont identifié les propriétaires
- OPS02-BP02 Les processus et les procédures ont identifié les propriétaires
- OPS02-BP04 Des mécanismes existent pour gérer les responsabilités et la propriété
- OPS02-BP05 Des mécanismes existent pour identifier les responsabilités et les droits de propriété
- OPS11-BP04 Effectuer la gestion des connaissances

Documents connexes :

- AWS Livre blanc | Présentation de DevOps AWS
- AWS Livre blanc | Bonnes pratiques en matière de balisage des ressources AWS
- AWS Livre blanc | Organisation de votre AWS environnement à l'aide de plusieurs comptes
- AWS Cloud Blog sur les opérations et les migrations | Créez une pratique d'automatisation du cloud pour l'excellence opérationnelle : les meilleures pratiques de AWS Managed Services
- Atelier AWS : étiquetage
- AWS Service Management Connector

Vidéos connexes :

- AWS Knowledge Center Live | Ressources de balisage AWS
- AWS re:Invent 2020 | Automatisez tout avec Systems Manager AWS
- AWS Re:inForce 2022 | Automatisation de la gestion des correctifs et de la conformité à l'aide de (06) AWS NIS3
- Support C'est vous | En savoir plus sur AWS Systems Manager

Exemples connexes :

- AWS Atelier sur l’excellence opérationnelle Well-Architected

OPS02-BP04 Des mécanismes existent pour gérer les responsabilités et la propriété

L'identification des responsabilités de votre rôle et de la manière dont vous contribuez aux résultats de l'entreprise permet de définir les priorités de vos tâches et de comprendre pourquoi votre rôle est important. Cette approche permet aux membres de l'équipe d'identifier les besoins et d'y répondre de manière appropriée. Lorsque les membres de l'équipe connaissent leur rôle, ils savent qui est propriétaire, ils identifient les opportunités d'amélioration et ils comprennent comment influencer ou apporter les changements appropriés.

Il arrive qu'une responsabilité ne soit pas clairement attribuée à une personne en particulier. Dans ce cas, concevez un mécanisme permettant de combler cette lacune. Créez un chemin hiérarchique bien défini qui renvoie vers une personne habilitée à attribuer la responsabilité à un rôle spécifique ou à prévoir le nécessaire pour répondre à ce besoin.

Résultat escompté : les équipes de votre organisation ont des responsabilités clairement définies qui incluent la manière dont elles sont liées aux ressources, aux actions à effectuer, aux processus et aux procédures. Ces responsabilités correspondent aux responsabilités et aux objectifs de l'équipe, ainsi qu'à celles des autres équipes. Vous documentez les chemins hiérarchiques de manière cohérente et transparente, et vous intégrez ces décisions dans des artefacts de documentation, tels que des matrices de responsabilité, des définitions d'équipes ou des pages wiki.

Anti-modèles courants :

- Les responsabilités de l'équipe sont ambiguës ou mal définies.
- L'équipe n'attribue pas les responsabilités à des rôles spécifiques.
- L'équipe n'aligne pas ses buts et ses objectifs sur ses responsabilités, ce qui rend difficile la mesure du succès.
- Les responsabilités des membres de l'équipe ne correspondent pas à celles de l'équipe et de l'organisation dans son ensemble.
- Votre équipe ne conserve pas ses responsabilités up-to-date, ce qui les rend incompatibles avec les tâches effectuées par l'équipe.
- Les chemins hiérarchiques permettant de déterminer les responsabilités ne sont pas définis ou ne sont pas clairs.
- Les chemins hiérarchiques n'ont pas de responsable de thread unique pour garantir une réponse rapide.
- Les rôles, les responsabilités et les chemins hiérarchiques ne sont pas détectables, et ils ne sont donc pas facilement disponibles en cas de besoin (par exemple, en réponse à un incident).

Avantages liés au respect de cette bonne pratique :

- Lorsque vous savez qui est responsable ou propriétaire, vous pouvez contacter l'équipe ou le membre de l'équipe concerné pour faire une demande ou transférer une tâche.
- Pour réduire le risque d'inaction et de besoins non satisfaits, vous avez identifié une personne habilitée à attribuer la responsabilité ou la propriété.
- Lorsque vous définissez clairement l'étendue d'une responsabilité, les membres de votre équipe gagnent en autonomie et en propriété.
- Vos responsabilités éclairent les décisions que vous prenez, les actions que vous effectuez et vos activités de transfert à leurs véritables propriétaires.
- Il est facile d'identifier des responsabilités abandonnées, car vous comprenez clairement ce qui ne relève pas de la responsabilité de votre équipe, ce qui vous permet de demander des éclaircissements.
- Les équipes évitent la confusion et les tensions, et elles gèrent leurs charges de travail et leurs ressources de manière plus adéquate.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Identifiez les rôles et responsabilités des membres de l'équipe et assurez-vous qu'ils comprennent les attentes de leur rôle. Rendez ces informations accessibles afin que les membres de votre organisation sachent qui contacter, que ce soit une équipe ou une personne, pour des besoins spécifiques. À mesure que les organisations cherchent à tirer parti des opportunités de migration et de modernisation AWS, les rôles et les responsabilités peuvent également changer. Tenez vos équipes et leurs membres conscients de leurs responsabilités et formez-les de manière appropriée pour qu'ils s'acquittent de leurs tâches pendant ce changement.

Déterminez le rôle ou l'équipe qui doit recevoir les remontées hiérarchiques afin d'identifier les responsabilités et la propriété. Cette équipe peut dialoguer avec différentes parties prenantes pour prendre une décision. Cependant, elle doit être responsable de la gestion du processus de prise de décision.

Fournissez des mécanismes accessibles aux membres de votre organisation pour découvrir et identifier la propriété et la responsabilité. Ces mécanismes leur indiquent à qui s'adresser pour des besoins spécifiques.

Exemple client

AnyCompany Le commerce de détail a récemment effectué une migration des charges de travail d'un environnement sur site vers sa zone d'atterrissage en utilisant une AWS approche « lift and shift ». Cette société a effectué un examen des opérations afin de réfléchir à la manière d'accomplir les tâches opérationnelles courantes, et a vérifié que sa matrice de responsabilité existante reflétait les opérations dans le nouvel environnement. Lorsqu'ils ont migré de l'infrastructure sur site vers l'infrastructure AWS, ils ont réduit les responsabilités des équipes chargées de l'infrastructure en ce qui concerne le matériel et l'infrastructure physique. Cette décision a également révélé de nouvelles opportunités de faire évoluer le modèle opérationnel pour ses charges de travail.

Tout en identifiant, en abordant et en documentant la majorité des responsabilités, elle a également défini des chemins hiérarchiques pour toutes les responsabilités qui n'ont pas été respectées ou qui pourraient changer à mesure que les pratiques opérationnelles évoluent. Pour explorer de nouvelles opportunités de standardisation et d'amélioration de l'efficacité de vos charges de travail, donnez accès à des outils opérationnels tels que AWS Systems Manager et à des outils de sécurité tels qu' AWS Security Hub Amazon. GuardDuty AnyCompanyLe commerce de détail prépare un examen des responsabilités et de la stratégie en fonction des améliorations qu'il souhaite apporter en premier lieu. Au fur et à mesure que l'entreprise adopte de nouvelles méthodes de travail et de nouveaux modèles technologiques, elle met à jour sa matrice de responsabilité en conséquence.

Étapes d'implémentation

1. Commencez par la documentation existante. Certains documents sources classiques peuvent inclure les éléments suivants :
   a. Responsabilité ou matrices responsables, responsables, consultées et informées (RACI)
   b. Définitions des équipes ou pages wiki
   c. Définitions et offres de services
   d. Descriptions de rôle ou de poste
2. Passez en revue les responsabilités documentées et organisez des discussions à ce sujet :
   a. Passez en revue les responsabilités avec les équipes pour identifier les incohérences entre les responsabilités documentées et les responsabilités que l'équipe assume habituellement.
   b. Discutez des services potentiels proposés par les clients internes afin d'identifier les écarts d'attentes entre les équipes.
3. Analysez et corrigez les écarts.
4. Identifiez les opportunités d'amélioration.
   a. Identifiez les demandes fréquentes gourmandes en ressources, qui sont généralement de bonnes candidates à l'amélioration.

   b. Recherchez les bonnes pratiques, les modèles et les conseils prescriptifs, et simplifiez et standardisez les améliorations grâce à ces conseils.
   c. Enregistrez les opportunités d'amélioration et suivez-les jusqu'à leur réalisation.
5. Si aucune équipe n'est encore chargée de la gestion et du suivi de l'attribution des responsabilités, identifiez un membre de l'équipe qui assumera cette responsabilité.
6. Définissez un processus permettant aux équipes de demander des éclaircissements sur les responsabilités.
   a. Passez en revue le processus et vérifiez qu'il est clair et simple à utiliser.
   b. Assurez-vous que quelqu'un contrôle les remontées hiérarchiques et en assure le suivi jusqu'à leur conclusion.
   c. Établissez des métriques opérationnelles pour mesurer l'efficacité.
   d. Créez un mécanisme de rétroaction pour vérifier que les équipes peuvent mettre en avant les opportunités d'amélioration.
   e. Mettez en place un mécanisme de vérification périodique.
7. Stockez les documents à un endroit détectable et accessible.
   a. Les wikis ou les portails de documentation sont des choix courants.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS01-BP06 Évaluer les compromis
- OPS03-BP02 Les membres de l'équipe sont habilités à agir lorsque les résultats sont menacés
- OPS03-BP03 L'escalade est encouragée
- OPS03-BP07 Ressources appropriées pour les équipes
- OPS09-BP01 Mesurer les objectifs opérationnels et avec des métriques KPIs
- OPS09-BP03 Examiner les indicateurs des opérations et prioriser les améliorations
- OPS11-BP01 Disposer d'un processus d'amélioration continue

Documents connexes :

- AWS Livre blanc - Présentation d'on DevOps AWS

- AWS Livre blanc - Cadre d' AWS Cloud adoption : perspective opérationnelle
- Excellence opérationnelle du cadre AWS Well-Architected : topologies du modèle d’exploitation au niveau de la charge de travail
- Conseils prescriptifs AWS : création de votre modèle d’exploitation cloud
- AWS Conseils prescriptifs - Création d'une RASCI matrice RACI OR pour un modèle d'exploitation cloud
- AWS Cloud Blog sur les opérations et les migrations - Créer de la valeur commerciale grâce aux équipes de la plateforme cloud
- AWS Cloud Blog sur les opérations et les migrations - Pourquoi un modèle d'exploitation dans le cloud ?
- AWS DevOps Blog - Comment les entreprises se modernisent pour les opérations cloud

Vidéos connexes :

- AWS Summit Online - Cloud Operating Models for Accelerated Transformation
- AWS re:Invent 2023 - Future-proofing cloud security: A new operating model

OPS02-BP05 Des mécanismes existent pour demander des ajouts, des modifications et des exceptions

Vous pouvez adresser des demandes aux propriétaires des processus, des procédures et des ressources. Les demandes comprennent les ajouts, les modifications et les exceptions. Ces demandes sont soumises à un processus de gestion des modifications. Prenez des décisions avisées pour approuver les demandes lorsque celles-ci sont viables et appropriées après une évaluation des avantages et des risques.

Résultat escompté :

- Vous pouvez faire des demandes de modification des processus, des procédures et des ressources en fonction de la propriété attribuée.
- Les modifications sont réalisées de manière délibérée, en pesant les avantages et les risques.

Anti-modèles courants :

- Vous devez mettre à jour la façon dont vous déployez votre application, mais il n’existe aucun moyen de demander à l’équipe chargée des opérations de modifier le processus de déploiement.

- Le plan de reprise après sinistre doit être mis à jour, mais il n'y a aucun propriétaire désigné à qui demander des modifications.

Avantages liés au respect de cette bonne pratique :

- Les processus, les procédures et les ressources peuvent évoluer au fur et à mesure que les exigences évoluent.
- Les propriétaires peuvent décider en connaissance de cause du moment où il convient d'apporter des modifications.
- Les modifications sont réalisées de manière délibérée.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Pour mettre en œuvre cette bonne pratique, vous devez être en mesure de demander des modifications des processus, des procédures et des ressources. Le processus de gestion des modifications peut être léger. Documentez le processus de gestion des modifications.

Exemple client

AnyCompany Le commerce de détail utilise une matrice d'attribution des responsabilités (RACI) pour identifier à qui appartient les modifications apportées aux processus, aux procédures et aux ressources. La société dispose d'un processus de gestion des modifications documenté, léger et facile à suivre. À l'aide de la RACI matrice et du processus, n'importe qui peut soumettre des demandes de modification.

Étapes d'implémentation

1. Identifiez les processus, les procédures et les ressources pour votre charge de travail et les responsables de chacun d'entre eux. Documentez-les dans votre système de gestion des connaissances.
   a. Si vous n'avez pas implémentés OPS02-BP01 Les ressources ont identifié les propriétaires, OPS02-BP02 Les processus et procédures ont des propriétaires identifiés ou OPS02-BP03 Les activités opérationnelles ont identifié les propriétaires responsables de leur performance, commencez par là.

2. Travaillez avec les parties prenantes de votre organisation pour élaborer un processus de gestion des modifications. Le processus doit couvrir les ajouts, les modifications et les exceptions pour les ressources, les processus et les procédures.
   a. Vous pouvez utiliser AWS Systems Manager Change Manager comme plateforme de gestion des modifications pour les ressources de charge de travail.
3. Documentez le processus de gestion des modifications dans votre système de gestion des connaissances.

Niveau d'effort du plan d'implémentation : moyen. L'élaboration d'un processus de gestion des modifications nécessite un alignement avec les multiples parties prenantes de votre organisation.

Ressources

Bonnes pratiques associées :

- OPS02-BP01 Les ressources ont identifié les propriétaires : les ressources ont besoin de propriétaires identifiés avant la mise en place d'un processus de gestion du changement.
- OPS02-BP02 Les processus et procédures ont des propriétaires identifiés : les processus ont besoin de propriétaires identifiés avant la mise en place d'un processus de gestion du changement.
- OPS02-BP03 Les activités opérationnelles ont identifié les propriétaires responsables de leur performance : les activités opérationnelles ont besoin de propriétaires identifiés avant la mise en place d'un processus de gestion du changement.

Documents connexes :

- AWS Conseils prescriptifs - Manuel de base pour les AWS grandes migrations : création de matrices RACI
- Livre blanc sur la gestion des modifications dans le cloud

Services connexes :

- AWS Systems Manager Gestionnaire du changement

OPS02-BP06 Les responsabilités entre les équipes sont prédéfinies ou négociées

Utilisez des accords définis ou négociés entre les équipes, accords qui décrivent la manière dont elles travaillent ensemble et se soutiennent mutuellement (par exemple, les temps de réponse, les

objectifs de niveau de service ou les contrats de niveau de service). Les canaux de communication entre équipes sont documentés. La compréhension de l'impact du travail des équipes sur les résultats opérationnels et les résultats des autres équipes et organisations indique la priorité de leurs tâches et les aide à répondre de manière appropriée.

Lorsque la responsabilité et la propriété ne sont pas définies ou sont inconnues, vous risquez de ne pas traiter les activités nécessaires en temps opportun et de déployer des efforts redondants et potentiellement contradictoires pour répondre à ces besoins.

Résultat escompté :

- Des accords de travail ou de soutien entre équipes sont convenus et documentés.
- Les équipes qui se soutiennent ou travaillent les unes avec les autres ont défini des canaux de communication et des attentes en matière de réponse.

Anti-modèles courants :

- Un problème survient en production et deux équipes distinctes commencent à le résoudre indépendamment l'une de l'autre. Leurs efforts cloisonnés prolongent la panne.
- L'équipe chargée des opérations a besoin de l'aide de l'équipe de développement mais aucun délai de réponse n'a été convenu. La demande est bloquée dans le backlog.

Avantages liés au respect de cette bonne pratique :

- Les équipes savent comment interagir et se soutenir mutuellement.
- Les attentes en matière de réactivité sont connues.
- Les canaux de communication sont clairement définis.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

La mise en œuvre de cette bonne pratique signifie qu'il n'y a aucune ambiguïté sur la façon dont les équipes travaillent les unes avec les autres. Les accords formels codifient la manière dont les équipes travaillent ensemble ou se soutiennent mutuellement. Les canaux de communication entre équipes sont documentés.

Exemple client

AnyCompany SREL'équipe de Retail a conclu un accord de niveau de service avec son équipe de développement. Chaque fois que l'équipe de développement émet une demande dans son système de tickets, elle peut s'attendre à recevoir une réponse dans les quinze minutes. En cas de panne du site, l'SREéquipe prend la tête de l'enquête avec le soutien de l'équipe de développement.

Étapes d'implémentation

1. En collaboration avec les parties prenantes de votre organisation, élaborez des accords entre les équipes sur la base de processus et de procédures.
   a. Si un processus ou une procédure est partagé entre deux équipes, élaborez un runbook sur la manière dont les équipes travailleront ensemble.
   b. S'il existe des dépendances entre les équipes, acceptez une réponse SLA aux demandes.
2. Documentez les responsabilités dans votre système de gestion des connaissances.

Niveau d'effort du plan d'implémentation : moyen. Si rien n'est convenu entre les équipes, il peut être difficile de parvenir à un accord avec les parties prenantes de votre organisation.

Ressources

Bonnes pratiques associées :

- OPS02-BP02 Les processus et procédures ont des propriétaires identifiés : la propriété du processus doit être identifiée avant la conclusion d'accords entre les équipes.
- OPS02-BP03 Les activités opérationnelles ont identifié les propriétaires responsables de leur performance : la propriété des activités opérationnelles doit être identifiée avant la conclusion d'accords entre les équipes.

Documents connexes :

- AWS Executive Insights - Favoriser l'innovation avec l'équipe de Two-Pizza
- Présentation de On DevOps AWS - Two-Two-Pizza Teams

OPS 3. Comment votre culture organisationnelle soutient-elle vos résultats opérationnels ?

Offrez du soutien aux membres de votre équipe afin qu'ils puissent agir plus efficacement et soutenir les résultats commerciaux.

Bonnes pratiques

- OPS03-BP01 Fournir un parrainage exécutif
- OPS03-BP02 Les membres de l'équipe sont habilités à agir lorsque les résultats sont menacés
- OPS03-BP03 L'escalade est encouragée
- OPS03-BP04 Les communications sont rapides, claires et exploitables
- OPS03-BP05 L'expérimentation est encouragée
- OPS03-BP06 Les membres de l'équipe sont encouragés à maintenir et à développer leurs compétences
- OPS03-BP07 Ressources appropriées pour les équipes

OPS03-BP01 Fournir un parrainage exécutif

À l’échelon le plus élevé de l’entreprise, la haute direction agit en tant que parrain exécutif pour définir clairement les attentes et l’orientation des résultats de l’organisation, y compris en évaluant son succès. Ce parrain préconise et favorise l’adoption des bonnes pratiques et l’évolution de l’organisation.

Résultat escompté : les organisations qui s’efforcent d’adopter, de transformer et d’optimiser leurs opérations cloud établissent des lignes de direction et de responsabilité claires pour obtenir les résultats souhaités. L’organisation comprend chaque capacité requise pour atteindre un nouveau résultat et attribue la propriété du développement aux équipes fonctionnelles. Le leadership définit activement cette orientation, attribue la responsabilité, définit le travail. Les membres de l’organisation peuvent ainsi se mobiliser, se sentir inspirés et travailler activement à la réalisation des objectifs souhaités.

Anti-modèles courants :

- Les responsables de charges de travail doivent migrer les charges de travail vers AWS sans parrain ni plan précis pour les opérations cloud. Par conséquent, les équipes ne collaborent pas consciemment pour améliorer et faire mûrir leurs capacités opérationnelles. L’absence de normes en matière de bonnes pratiques opérationnelles submerge les équipes (telles que la quantité de travail des opérateurs, les astreintes et la dette technique), ce qui limite l’innovation.
- Un nouvel objectif à l’échelle de l’organisation a été fixé pour adopter une technologie émergente sans assurer le parrainage de la direction ni fournir aucune stratégie. Les équipes interprètent les objectifs différemment, ce qui ne permet pas de savoir où concentrer les efforts, pourquoi ils

sont importants et comment mesurer l'impact. Par conséquent, l'organisation perd son élan dans l'adoption de la technologie.

Avantages de l'établissement de cette bonne pratique : lorsque le parrainage exécutif communique et partage clairement la vision, l'orientation et les objectifs, les membres de l'équipe savent ce que l'on attend d'eux. Les individus et les équipes commencent à concentrer intensément leurs efforts dans la même direction pour atteindre les objectifs définis lorsque les dirigeants sont activement engagés. Par conséquent, l'organisation maximise sa capacité à réussir. Lorsque vous évaluez le succès, vous pouvez mieux identifier les obstacles au succès afin de pouvoir les surmonter grâce à l'intervention du sponsor exécutif.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

- À chaque étape de la transition vers le cloud (migration, adoption ou optimisation), la réussite passe par une implication active au plus haut niveau de la direction, avec un parrain exécutif désigné. Le parrain exécutif aligne l'état d'esprit, les compétences et les méthodes de travail de l'équipe sur la stratégie définie.
  - Explication de la raison : apportez de la clarté et expliquez le raisonnement qui sous-tend la vision et la stratégie.
  - Définition des attentes : spécifiez et publiez des objectifs pour vos organisations, y compris la façon dont le progrès et la réussite sont évalués.
  - Suivi de la réalisation des objectifs : mesurez régulièrement la réalisation progressive des objectifs (non seulement l'achèvement des tâches). Partagez les résultats afin que les actions appropriées puissent être effectuées si les résultats sont menacés.
  - Fourniture des ressources nécessaires pour atteindre vos objectifs : réunissez les personnes et les équipes pour qu'elles collaborent et élaborent les bonnes solutions qui produisent les résultats définis. Cette approche permet de réduire ou d'éliminer les frictions organisationnelles.
  - Défense de vos équipes : restez impliqué avec vos équipes afin de comprendre comment elles évoluent et de savoir s'il existe des facteurs externes qui les affectent. Identifiez les obstacles qui entravent la progression de vos équipes. Agissez au nom de vos équipes pour surmonter les obstacles et éliminer les charges inutiles. Lorsque vos équipes sont affectées par des facteurs externes, réévaluez les objectifs et ajustez les cibles le cas échéant.

- Être un moteur de l’adoption des bonnes pratiques : acceptez les bonnes pratiques qui apportent des avantages quantifiables et montrez de la reconnaissance pour les créateurs et les adoptants. Encouragez une adoption plus large pour amplifier les avantages obtenus.
- Encouragement de l’évolution de vos équipes : créez une culture d’amélioration continue et apprenez de manière proactive des progrès réalisés et des échecs. Encouragez la croissance et le développement personnels et organisationnels. Utilisez les données et des anecdotes pour faire évoluer la vision et la stratégie.

Exemple client

AnyCompany Le commerce de détail est en train de transformer son activité en réinventant rapidement l'expérience client, en améliorant la productivité et en accélérant la croissance grâce à l'IA générative.

Étapes d’implémentation

1. Établissez un leadership unique et désignez un sponsor exécutif principal pour diriger et piloter la transformation.
2. Définissez les résultats commerciaux clairs de votre transformation et attribuez les responsabilités aux parties prenantes. Donnez au sponsor exécutif principal le pouvoir nécessaire pour diriger et prendre des décisions critiques.
3. Vérifiez que votre stratégie de transformation est très claire et qu’elle est largement communiquée par le sponsor exécutif à tous les niveaux de l’organisation.
   a. Établissez des objectifs commerciaux clairement définis pour les initiatives informatiques et cloud.
   b. Documentez les métriques commerciales clés pour favoriser la transformation de l’informatique et du cloud.
   c. Communiquez la vision de manière cohérente à toutes les équipes et à toutes les personnes responsables de divers aspects de la stratégie.
4. Élaborez des matrices de planification de la communication qui spécifient le message à transmettre à des dirigeants, des responsables et des contributeurs individuels spécifiques. Spécifiez la personne ou l’équipe qui devra transmettre ce message.
   a. Exécutez les plans de communication de manière cohérente et fiable.
   b. Définissez et gérez les attentes en organisant régulièrement des événements en personne.

   c. Acceptez les retours sur l'efficacité des communications, ajustez les communications et planifiez en conséquence.
   d. Planifiez des événements de communication pour comprendre de manière proactive les défis rencontrés par les équipes et établissez une boucle de rétroaction cohérente qui permettra de corriger le cap si nécessaire.
5. Lancez activement chaque initiative du point de vue de la direction afin de vérifier que toutes les équipes concernées comprennent les résultats qu'elles sont tenues d'atteindre.
6. À chaque réunion sur l'état d'avancement, les sponsors exécutifs doivent rechercher les obstacles, examiner les métriques établies, les anecdotes ou le retour des équipes, et mesurer les progrès réalisés par rapport aux objectifs.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS03-BP04 Les communications sont rapides, claires et exploitables
- OP11-BP01 Disposer d'un processus d'amélioration continue
- OPS11-BP07 Réaliser des examens des métriques opérationnelles

Documents connexes :

- Untangling Your Organisational Hairball: Highly Aligned
- The Living Transformation: Pragmatically approaching changes
- Becoming a Future-Ready Enterprise
- 7 pièges à éviter lors de la construction d'un CCOE
- Navigating the Cloud: Key Performance Indicators for Success

Vidéos connexes :

- AWS re:INVENT 2023 : Guide des leaders sur l'IA générative : utiliser l'histoire pour façonner le futur (04) SEG2

Exemples connexes :

- Prosci : rôle et importance du parrain principal

OPS03-BP02 Les membres de l'équipe sont habilités à agir lorsque les résultats sont menacés

Un comportement culturel axé sur la responsabilisation inculqué par la direction donne à chaque employé le sentiment d'être habilité à agir au nom de l'ensemble de l'entreprise au-delà de son mandat et de ses responsabilités définis. Les employés peuvent agir pour identifier les risques de manière proactive à mesure qu'ils apparaissent et prendre les mesures appropriées. Une telle culture permet aux employés de prendre des décisions importantes en ayant connaissance de la situation.

Par exemple, Amazon utilise les principes de leadership comme directives pour inciter les employés à adopter le comportement souhaité afin d'avancer dans les situations, de résoudre les problèmes, de gérer les conflits et de prendre des mesures.

Résultat escompté : le leadership a influencé une nouvelle culture qui permet aux individus et aux équipes de prendre des décisions critiques, même aux niveaux inférieurs de l'organisation (les décisions à long terme étant définies par des autorisations vérifiables et des mécanismes de sécurité). L'échec n'est pas découragé et les équipes apprennent de manière itérative à améliorer leurs prises de décisions et leurs réactions afin de pouvoir faire face à des situations similaires à l'avenir. Si les actions d'une personne entraînent une amélioration susceptible pouvant bénéficier à d'autres équipes, les leçons tirées de ces actions sont partagées avec ces équipes. La direction mesure les améliorations opérationnelles et incite l'individu et l'organisation à adopter de tels modèles.

Anti-modèles courants :

- Il n'existe pas de directives ni de mécanismes clairs au sein d'une organisation, indiquant la marche à suivre lorsqu'un risque est identifié. Par exemple, lorsqu'un employé remarque une attaque de phishing, il n'en informe pas l'équipe de sécurité, ce qui entraîne une propagation de l'attaque dans une grande partie de l'organisation. Cela entraîne une violation de données.
- Vos clients se plaignent de l'indisponibilité du service, qui est principalement due à l'échec des déploiements. Votre SRE équipe est responsable de l'outil de déploiement, et une annulation automatique des déploiements figure dans sa feuille de route à long terme. Lors du récent déploiement d'une application, l'un des ingénieurs a conçu une solution permettant de restaurer automatiquement la version précédente de son application. Bien que leur solution puisse devenir un modèle pour les SRE équipes, d'autres équipes ne l'adoptent pas, car il n'existe aucun processus permettant de suivre ces améliorations. L'organisation continue de faire face à des

échecs de déploiements qui ont un impact sur les clients et suscitent encore davantage de sentiments négatifs.

- Afin de rester en conformité, votre équipe infosec supervise un processus établi de longue date qui consiste à alterner régulièrement SSH les clés partagées pour le compte des opérateurs qui se connectent à leurs instances Amazon EC2 Linux. Les équipes de sécurité de l'information mettent plusieurs jours à effectuer la rotation des clés, ce qui vous empêche de vous connecter à ces instances. Personne, à l'intérieur ou à l'extérieur d'Infosec, ne suggère d'utiliser d'autres options AWS pour obtenir le même résultat.

Avantages de l'établissement de cette bonne pratique : en décentralisant l'autorité chargée de prendre des décisions et en habilitant vos équipes à prendre les décisions clés, vous êtes en mesure de résoudre les problèmes plus rapidement avec des taux de réussite accrus. De plus, les équipes commencent à ressentir un sentiment d'appartenance et à réaliser que les échecs sont acceptables. L'expérimentation devient un pilier de la culture. Les responsables et les directeurs n'ont pas l'impression d'être microgérés dans tous les aspects de leur travail.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

1. Développez une culture où l'on s'attend à ce que des échecs se produisent.
2. Définissez clairement la responsabilité et la propriété pour les différents domaines fonctionnels de l'organisation.
3. Communiquez la propriété et la responsabilité à tous afin que chaque personne sache qui peut l'aider à prendre des décisions décentralisées.
4. Définissez vos décisions à sens unique et bidirectionnelles pour aider les individus à déterminer quand ils doivent faire remonter un problème à des niveaux hiérarchiques supérieurs.
5. Sensibilisez l'organisation au fait que tous les employés sont habilités à agir à différents niveaux lorsque les résultats sont menacés. Fournissez aux membres de votre équipe de la documentation sur la gouvernance, les niveaux d'autorisation, les outils et les opportunités de mettre en pratique les compétences nécessaires pour réagir efficacement.
6. Donnez aux membres de votre équipe l'occasion de mettre en pratique les compétences nécessaires pour répondre à diverses décisions. Une fois les niveaux de décision définis, effectuez des journées de simulation pour vérifier que tous les contributeurs comprennent le processus et peuvent le démontrer.

   a. Fournissez d’autres environnements sûrs où les processus et les procédures peuvent être testés et auxquels les parties prenantes peuvent être formées en toute sécurité.
   b. Soulignez et faites prendre conscience que les membres de l’équipe ont le pouvoir de prendre des mesures lorsque le résultat présente un niveau de risque prédéfini.
   c. Définissez le pouvoir des membres de l’équipe de prendre des mesures en leur attribuant des autorisations et un accès aux charges de travail et aux composants qu’ils prennent en charge.
7. Permettez aux équipes de partager les leçons tirées (réussites et échecs opérationnels).
8. Donnez aux équipes les moyens de remettre en question le statu quo et fournissez des mécanismes permettant de suivre et de mesurer les améliorations, ainsi que leur impact sur l’organisation.

Niveau d’effort du plan d’implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS01-BP06 Évaluer les compromis tout en gérant les avantages et les risques
- OPS02-BP05 Des mécanismes existent pour identifier les responsabilités et les propriétés

Documents connexes :

- Article de blog AWS | The agile enterprise
- Article de blog AWS | Measuring success : A paradox and a plan
- Article de blog AWS | Letting go : Enabling autonomy in teams
- Centralize or Decentralize?

Vidéos connexes :

- re:Invent 2023 | Comment ne pas saboter votre transformation (01) SEG2
- re:Invent 2021 - Amazon Builders’ Library: Operational Excellence at Amazon
- Centralization vs. Decentralization

Exemples connexes :

- Utilisation d’enregistrements de décisions architecturales pour rationaliser la prise de décisions techniques dans le cadre d’un projet de développement logiciel

OPS03-BP03 L'escalade est encouragée

Les membres de l’équipe sont encouragés par la direction à faire part des problèmes et des préoccupations aux décideurs et aux parties prenantes de niveau supérieur s’ils estiment que les résultats souhaités sont menacés et que les normes attendues ne sont pas respectées. Il s’agit d’une caractéristique de la culture de l’entreprise, qui est encouragée à tous les niveaux. Les remontées doivent être effectuées tôt et souvent afin que les risques puissent être identifiés et les incidents évités La direction ne réprimande pas les personnes qui font remonter un problème.

Résultat escompté : les membres de l’organisation sont à l’aise pour porter les problèmes à leur niveau de direction immédiat ou supérieur. La direction a délibérément et consciemment fixé des attentes pour que ses équipes se sentent en sécurité lorsqu’il s’agit de faire remonter un problème. Un mécanisme est en place pour faire remonter les problèmes à chaque niveau de l’organisation. Lorsque les employés font remonter un problème à leur responsable, ils décident conjointement du niveau d’impact et de la nécessité ou non de faire remonter ce problème à un niveau supérieur. Pour lancer une remontée hiérarchique, les employés doivent inclure une recommandation de plan de travail visant à résoudre le problème. Si le supérieur direct ne prend pas de mesures en temps opportun, les employés sont encouragés à faire remonter les problèmes au niveau hiérarchique le plus élevé au sein de la direction s’ils sont convaincus que les risques pour l’organisation justifient une telle démarche.

Anti-modèles courants :

- Les dirigeants ne posent pas suffisamment de questions approfondies lors de la réunion sur l’état d’avancement de votre programme de transformation du cloud pour identifier les problèmes et les blocages. Seules les bonnes nouvelles sont présentées dans l’état d’avancement. Elle CIO a clairement indiqué qu'elle n'aimait entendre que les bonnes nouvelles, car tout défi soulevé laisse CEO penser que le programme est un échec.
- Vous êtes ingénieur des opérations cloud et vous remarquez que le nouveau système de gestion des connaissances n’est pas largement adopté par les équipes d’application. L’entreprise a investi un an et plusieurs millions de dollars pour mettre en œuvre ce nouveau système de gestion des connaissances, mais les utilisateurs continuent de créer leurs runbooks localement et de les partager sur un partage cloud organisationnel. Cette approche rend difficile la recherche de connaissances pertinentes pour les charges de travail prises en charge. Vous essayez d’attirer l’attention de la direction sur ce point, car une utilisation cohérente de ce système contribuerait à

améliorer l’efficacité opérationnelle. Lorsque vous expliquez la situation à la directrice qui gère la mise en œuvre du système de gestion des connaissances, elle vous réprimande, car cela remet en question l’investissement.

- L'équipe infosec chargée du renforcement des ressources informatiques a décidé de mettre en place un processus qui nécessite d'effectuer les analyses nécessaires pour garantir que les EC2 instances sont entièrement sécurisées avant que l'équipe de calcul ne libère la ressource pour utilisation. Cela a créé un délai d'une semaine supplémentaire pour le déploiement des ressources, ce qui les interromptSLA. L’équipe informatique craint de faire remonter ce problème au vice-président via le cloud, car cela donnerait une mauvaise image du vice-président de la sécurité de l’information.

Avantages liés au respect de cette bonne pratique :

Les problèmes complexes ou critiques sont résolus avant d’avoir un impact sur l’entreprise. Les pertes de temps sont réduites. Les risques sont minimisés. Les équipes deviennent plus proactives et plus axées sur les résultats lorsqu’elles résolvent des problèmes.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

La volonté et la capacité de faire remonter librement un problème à tous les niveaux de l’entreprise constituent un fondement organisationnel et culturel qui doit être développé consciemment en mettant l’accent sur la formation, la communication avec la direction, la définition des attentes et le déploiement de mécanismes dans l’ensemble de l’organisation à tous les niveaux.

Étapes d’implémentation

1. Définissez les stratégies, les normes et les attentes de votre organisation.
   a. Assurez l’adoption et la compréhension à grande échelle des stratégies, des attentes et des normes.
2. Encouragez les employés, formez-les et donnez-leur les moyens de faire remonter les problèmes rapidement et fréquemment lorsque les normes ne sont pas respectées.
3. Reconnaissez sur le plan organisationnel que la bonne pratique consiste à faire remonter les informations rapidement et fréquemment. Acceptez le fait que les remontées puissent être sans fondement et qu’il est préférable d’avoir la possibilité d’éviter un incident plutôt que de devoir y faire face parce que vous n’avez pas fait remonter l’information.

   a. Construisez un mécanisme d’escalade (comme un système à cordes Andon).
   b. Disposez de procédures documentées définissant quand et comment la remontée doit avoir lieu.
   c. Définissez la série de personnes ayant un pouvoir croissant pour prendre ou approuver des mesures, ainsi que les coordonnées de chaque partie prenante.
4. Toute remontée hiérarchique doit rester ouverte jusqu’à ce que le membre de l’équipe soit convaincu que le risque a été atténué grâce aux mesures prises par la direction.
   a. Les remontées hiérarchiques doivent inclure les détails suivants :
      i. Description de la situation et de la nature du risque
      ii. Sévérité de la situation
      iii. Qui est concerné (ou quoi)
      iv. Ampleur de l’impact
      v. Urgence en cas d’impact direct
      vi. Solutions et plans d’atténuation suggérés
   b. Protégez les employés qui font remonter les problèmes. Définissez une stratégie qui protège les membres de l’équipe contre les représailles s’ils font remonter un problème auprès d’un décideur ou d’une partie prenante non réceptifs. Mettez en place des mécanismes permettant d’identifier si cela se produit et de répondre de manière appropriée.
5. Encouragez une culture basée sur des boucles de rétroaction pour l’amélioration continue dans tout ce que l’organisation produit. Les boucles de rétroaction s’apparentent à des remontées mineures adressées aux personnes responsables. Elles identifient les opportunités d’amélioration, même lorsque la remontée n’est pas nécessaire. Les cultures d’amélioration continue obligent tout le monde à être plus proactif.
6. La direction doit périodiquement rappeler l’importance des stratégies, des normes et des mécanismes, ainsi que le souhait de faire remonter les problèmes de manière ouverte et d’encourager des boucles de rétroaction continues, sans représailles.

Niveau d’effort du plan d’implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS02-BP05 Des mécanismes existent pour demander des ajouts, des modifications et des exceptions

Documents connexes :

- How do you foster a culture of continuous improvement and learning from Andon and escalation systems?
- The Andon Cord (IT Revolution)
- AWS DevOps Conseils | Établissez des voies d'escalade claires et encouragez les désaccords constructifs

Vidéos connexes :

- Jeff Bezos on how to make decisions (& increase velocity)
- Toyota Product System: Stopping Production, a Button, and an Andon Electric Board
- Andon Cord dans le secteur manufacturier LEAN

Exemples connexes :

- Utilisation des plans d’escalade dans Incident Manager

OPS03-BP04 Les communications sont rapides, claires et exploitables

La direction est responsable de la création de communications solides et efficaces, en particulier lorsque l’organisation adopte de nouvelles stratégies, technologies ou méthodes de travail. Les dirigeants doivent fixer des attentes pour que l’ensemble du personnel travaille à la réalisation des objectifs de l’entreprise. Concevez des mécanismes de communication qui sensibilisent à long terme les équipes responsables de l’exécution des plans financés et parrainés par la direction. Tirez parti de la diversité interorganisationnelle et écoutez attentivement les divers points de vue uniques. Utilisez cette perspective pour accroître l’innovation, remettre en question vos hypothèses et réduire le risque de biais de confirmation. Favorisez l’inclusion, la diversité et l’accessibilité au sein de vos équipes afin d’obtenir des perspectives bénéfiques.

Résultat escompté : votre organisation conçoit des stratégies de communication pour faire face à l’impact du changement sur l’organisation. Les équipes restent informées et motivées pour continuer à travailler les unes avec les autres plutôt que les unes contre les autres. Chaque personne comprend à quel point son rôle est important pour atteindre les objectifs fixés. Le courrier électronique n’est qu’un mécanisme passif de communication et est utilisé en conséquence. Les responsables passent du temps avec leurs collaborateurs individuels pour les aider à comprendre leurs responsabilités, les tâches à accomplir et la manière dont leur travail contribue à la mission

globale. Lorsque cela est nécessaire, les dirigeants mobilisent les intéressés directement dans des lieux plus restreints pour transmettre certains messages et vérifier qu'ils sont transmis efficacement. Grâce à de bonnes stratégies de communication, l'organisation obtient des résultats égaux ou supérieurs aux attentes de la direction. La direction encourage et sollicite la diversité des opinions au sein des équipes et entre elles.

Anti-modèles courants :

- Votre organisation dispose d'un plan quinquennal pour migrer toutes les charges de travail vers AWS. L'analyse de rentabilisation du cloud inclut la modernisation de 25 % de toutes les charges de travail afin de tirer parti de la technologie sans serveur. Il CIO communique cette stratégie à ses subordonnés directs et attend de chaque leader qu'il diffuse cette présentation aux responsables, aux directeurs et aux contributeurs individuels sans aucune communication en personne. Il prend CIO du recul et s'attend à ce que son organisation mette en œuvre la nouvelle stratégie.
- La direction ne fournit ni n'utilise de mécanisme de commentaires, et l'écart entre les attentes se creuse, ce qui entraîne le blocage des projets.
- Il vous est demandé d'apporter une modification à vos groupes de sécurité, mais aucune information ne vous est donnée quant à la modification à apporter, à l'impact qu'elle pourrait avoir sur l'ensemble des charges de travail et à quel moment elle devrait avoir lieu. Le responsable transmet un e-mail du vice-président de InfoSec et ajoute le message « Make this happen ».
- Des modifications ont été apportées à votre stratégie de migration afin de faire passer le nombre de modernisations prévues de 25 % à 10 %. Cela a des répercussions en aval sur l'organisation des opérations. Il n'a pas été informé de ce changement stratégique et ne dispose donc pas de suffisamment de main-d'œuvre qualifiée pour gérer la migration en lift-and-shift d'un plus grand nombre de charges de travail vers AWS.

Avantages liés au respect de cette bonne pratique :

- Votre organisation est bien informée sur les stratégies nouvelles ou modifiées, et elle agit en conséquence avec une forte motivation pour s'aider mutuellement à atteindre les objectifs généraux et les métriques fixés par la direction.
- Des mécanismes existent et sont utilisés pour informer en temps opportun les membres de l'équipe des risques connus et des événements planifiés.
- Les nouvelles méthodes de travail (y compris les changements apportés aux parties prenantes ou à l'organisation, aux processus ou à la technologie), ainsi que les compétences requises,

sont adoptées plus efficacement par l'organisation, et votre organisation bénéficie d'avantages commerciaux plus rapidement.

- Les membres de l'équipe disposent du contexte nécessaire pour les communications reçues et peuvent être plus efficaces dans leur travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Pour mettre en œuvre cette bonne pratique, vous devez travailler avec les parties prenantes de votre organisation pour convenir de normes de communication. Diffusez ces normes dans votre organisation. Pour toute transition informatique importante, une équipe de planification bien établie est mieux à même de gérer l'impact du changement sur ses collaborateurs qu'une organisation qui ignore cette pratique. Les organisations de grande envergure peuvent avoir plus de difficulté à gérer le changement, car il est essentiel d'obtenir l'adhésion de tous les contributeurs individuels par rapport à une nouvelle stratégie. En l'absence d'une équipe de planification de la transition, la direction assume 100 % de la responsabilité liée à une communication efficace. Lorsque vous mettez en place une équipe de planification de la transition, demandez aux membres de l'équipe de travailler avec tous les dirigeants de l'organisation afin de définir et de gérer des communications efficaces à tous les niveaux.

Exemple client

AnyCompany Retail s'est inscrit au Support aux AWS entreprises et dépend d'autres fournisseurs tiers pour ses opérations dans le cloud. L'entreprise utilise le chat et le chatops comme principal moyen de communication pour ses activités opérationnelles. Les alertes et autres informations alimentent des canaux spécifiques. Lorsque quelqu'un doit agir, il indique clairement le résultat souhaité et, dans de nombreux cas, il reçoit un runbook ou un playbook à utiliser. Il planifie les modifications majeures des systèmes de production à l'aide d'un calendrier des modifications.

Étapes d'implémentation

1. Mettez en place une équipe centrale au sein de l'organisation chargée d'élaborer et de lancer des plans de communication pour les changements qui se produisent à plusieurs niveaux de l'organisation.
2. Nommez des responsables uniques pour assurer la supervision. Donnez aux équipes individuelles la capacité d'innover de manière indépendante et trouvez un juste milieu avec l'utilisation de mécanismes cohérents, afin d'obtenir le bon niveau d'inspection et de vision directionnelle.

3. Travaillez avec les parties prenantes de votre organisation pour convenir de normes, de pratiques et de plans de communication.
4. Vérifiez que l'équipe centrale des communications collabore avec la direction de l'organisation et du programme pour rédiger des messages destinés au personnel concerné au nom des dirigeants.
5. Créez des mécanismes de communication stratégiques pour gérer le changement par le biais d'annonces, de calendriers partagés, de réunions informelles, de rencontres en personne ou de one-on-one méthodes afin que les membres de l'équipe aient des attentes appropriées quant aux mesures à prendre.
6. Fournissez le contexte, les détails et l'heure nécessaires (si possible) pour déterminer si une action est nécessaire. Lorsqu'une action est nécessaire, précisez l'action requise et son impact.
7. Mettez en œuvre des outils qui facilitent les communications tactiques, tels que le chat interne, le courrier électronique et la gestion des connaissances.
8. Mettez en œuvre des mécanismes pour mesurer et vérifier que toutes les communications aboutissent aux résultats souhaités.
9. Mettez en place une boucle de commentaires qui mesure l'efficacité de toutes les communications, en particulier lorsque les communications sont liées à la résistance aux changements dans l'ensemble de l'organisation.
10. Pour tous Comptes AWS, établissez des contacts alternatifs pour la facturation, la sécurité et les opérations. Idéalement, chaque contact devrait correspondre à une liste de distribution par e-mail et non à un contact individuel spécifique.
11. Établissez un plan de communication d'escalade et d'escalade inverse pour interagir avec vos équipes internes et externes, y compris le AWS support et les autres fournisseurs tiers.
12. Initiez et mettez en œuvre des stratégies de communication de manière cohérente tout au long de la durée de vie de chaque programme de transformation.
13. Définissez la priorité des actions qui sont reproductibles dans la mesure du possible pour procéder à une automatisation à grande échelle en toute sécurité.
14. Lorsque des communications sont requises dans les scénarios où les actions sont automatisées, leur objectif doit être d'informer les équipes à des fins d'audit ou dans le cadre du processus de gestion du changement.
15. Analysez les communications de vos systèmes d'alerte pour détecter les faux positifs ou les alertes créées en permanence. Supprimez ou modifiez ces alertes afin qu'elles se déclenchent lorsqu'une intervention humaine est requise. Si une alerte est déclenchée, fournissez un runbook ou un playbook.

a. Vous pouvez utiliser AWS Systems Manager Documents pour créer des playbooks et des runbooks pour les alertes.

16. Des mécanismes sont en place pour notifier les risques ou les événements prévus d'une manière claire et exploitable, avec un préavis suffisant pour permettre des réponses appropriées. Utilisez des listes d'e-mails ou des canaux de conversation instantanée pour envoyer des notifications avant les événements prévus.

    a. AWS Chatbot peut être utilisé pour envoyer des alertes et répondre à des événements au sein de la plateforme de messagerie de votre organisation.

17. Fournissez une source d'informations accessible où les événements planifiés peuvent être découverts. Envoyez des notifications d'événements planifiés à partir du même système.

    a. AWS Le calendrier des modifications de Systems Manager peut être utilisé pour créer des fenêtres de modification lorsque des modifications peuvent survenir. Cela permet aux membres de l'équipe de savoir quand ils peuvent apporter des modifications en toute sécurité.

18. Surveillez les notifications de vulnérabilités et les informations sur les correctifs pour comprendre les failles dangereuses et les risques potentiels associés aux éléments de votre charge de travail. Envoyez une notification aux membres de l'équipe afin qu'ils puissent agir.

    a. Vous pouvez vous abonner aux bulletins AWS de sécurité pour recevoir des notifications de vulnérabilités sur AWS.

19. Recherche d'opinions et de perspectives variées : encouragez les contributions de chacun. Offrez des opportunités de communication aux groupes sous-représentés. Effectuez une rotation des rôles et des responsabilités lors des réunions.

    a. Élargissement des rôles et des responsabilités : offrez aux membres de l'équipe la possibilité d'assumer des rôles qu'ils n'auraient pas autrement. Ils pourront ainsi acquérir de l'expérience et façonner leur perspective grâce à leur rôle et à leurs interactions avec de nouveaux membres de l'équipe avec lesquels ils n'auraient peut-être pas eu d'interaction autrement. Ils pourront également apporter leur expérience et leur perspective au nouveau rôle et aux nouveaux membres de l'équipe avec lesquels ils interagissent. À mesure que les perspectives s'élargissent, identifiez les opportunités commerciales émergentes ou les nouvelles opportunités d'amélioration. Demandez aux membres d'une équipe d'effectuer des tâches communes que d'autres exécutent habituellement afin de comprendre les exigences et l'impact de leur exécution.

    b. Offrir un environnement sûr et accueillant : établissez une politique et des contrôles qui protègent la sécurité mentale et physique des membres de l'équipe au sein de votre organisation. Les membres de l'équipe doivent être en mesure d'interagir sans craindre de

représailles. Lorsque les membres de l'équipe se sentent en sécurité et sont les bienvenus, ils sont plus susceptibles d'être impliqués et productifs. Plus votre organisation est diversifiée, mieux vous pouvez comprendre les personnes que vous soutenez, y compris vos clients. Lorsque les membres de votre équipe sont à l'aise, se sentent libres de parler et sont sûrs d'être entendus, ils sont plus susceptibles de partager des informations précieuses (par exemple, les possibilités de marketing, les besoins d'accessibilité, les segments de marché délaissés et les risques non reconnus dans votre environnement).

c. Encourager les membres de l'équipe à participer pleinement : fournissez les ressources nécessaires pour que vos employés puissent participer pleinement à toutes les activités liées à leur travail. Les membres de l'équipe confrontés à des défis quotidiens développent des compétences pour les surmonter. Ces compétences développées de manière unique peuvent apporter des avantages considérables à votre organisation. Accompagnez les membres de l'équipe avec les ajustements nécessaires pour accroître les avantages que vous pouvez tirer de leurs contributions.

Ressources

Bonnes pratiques associées :

- OPS03-BP01 Fournir un parrainage exécutif
- OPS07-BP03 Utiliser des runbooks pour exécuter des procédures
- OPS07-BP04 Utiliser des playbooks pour étudier les problèmes

Documents connexes :

- Article de blog AWS | La responsabilisation et l'autonomisation sont essentielles pour des organisations agiles performantes
- Executive Insights AWS | Apprenez à développer l'innovation, et non la complexité | Single-Thread Leaders
- Bulletins de sécuritéAWS
- Ouvert CVE
- Support Application dans Slack pour gérer les demandes de support
- Gérez AWS les ressources de vos chaînes Slack avec AWS Chatbot

Exemples connexes :

- Ateliers Well-Architected : inventaire et gestion des correctifs (niveau 100)

Services connexes :

- AWS Chatbot
- AWS Calendrier des modifications de Systems Manager
- AWS Documents relatifs aux Systems Manager

OPS03-BP05 L'expérimentation est encouragée

L’expérimentation est un catalyseur qui permet de transformer de nouvelles idées en produits et en fonctionnalités. Elle accélère la formation et permet aux membres de l’équipe de s’intéresser et d’être engagés. Les membres de l’équipe sont encouragés à expérimenter souvent pour stimuler l’innovation. Même lorsqu’un résultat indésirable se produit, il est bon de savoir ce qu’il ne faut pas faire. Les membres de l’équipe ne sont pas sanctionnés pour les expérimentations réussies produisant des résultats indésirables.

Résultat escompté :

- Votre organisation encourage l’expérimentation pour favoriser l’innovation.
- Les expériences sont utilisées comme une occasion d’apprendre.

Anti-modèles courants :

- Vous souhaitez effectuer un test A/B mais il n’existe aucun mécanisme pour réaliser l’expérience. Vous déployez une modification de l’interface utilisateur sans pouvoir la tester. Il en résulte une expérience négative pour le client.
- Votre entreprise ne dispose que d’un environnement d’étape et de production. Il n’existe pas d’environnement de test (sandbox) pour expérimenter de nouvelles fonctionnalités ou de nouveaux produits. Vous devez donc expérimenter dans l’environnement de production.

Avantages liés au respect de cette bonne pratique :

- L’expérimentation est le moteur de l’innovation.
- Vous pouvez réagir plus rapidement aux commentaires des utilisateurs grâce à l’expérimentation.
- Votre organisation développe une culture de l’apprentissage.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Les expériences doivent être menées en toute sécurité. Exploitez plusieurs environnements pour expérimenter sans mettre en péril les ressources de production. Utilisez les tests A/B et les indicateurs de fonctionnalités pour tester les expériences. Donnez aux membres de l’équipe la possibilité de mener des expériences dans un environnement de test (sandbox).

Exemple client

AnyCompany Le commerce de détail encourage l'expérimentation. Les membres de l’équipe peuvent utiliser 20 % de leur semaine de travail pour expérimenter ou apprendre de nouvelles technologies. Ils disposent d’un environnement de test (sandbox) où ils peuvent innover. Les tests A/B sont utilisés pour les nouvelles fonctionnalités afin de les valider en fonction des commentaires réels des utilisateurs.

Étapes d’implémentation

1. Travaillez avec les dirigeants de votre organisation pour soutenir l’expérimentation. Les membres de l’équipe doivent être encouragés à réaliser des expériences en toute sécurité.
2. Offrez aux membres de votre équipe un environnement où ils peuvent expérimenter en toute sécurité. Ils doivent avoir accès à un environnement similaire à celui de la production.
   a. Vous pouvez utiliser un environnement distinct Compte AWS pour créer un environnement de bac à sable à des fins d'expérimentation. AWS Control Towerpeut être utilisé pour approvisionner ces comptes.
3. Utilisez des indicateurs de fonctionnalités et des tests A/B pour expérimenter en toute sécurité et recueillir les commentaires des utilisateurs.
   a. AWS AppConfig Feature Flags permet de créer des drapeaux de fonctionnalités.
   b. Amazon CloudWatch Evidently peut être utilisé pour exécuter des tests A/B sur un déploiement limité.
   c. Vous pouvez utiliser des versions AWS Lambda pour déployer une nouvelle version d’une fonction à des fins de test bêta.

Niveau d’effort du plan d’implémentation : élevé La fourniture aux membres de l’équipe d’un environnement dans lequel expérimenter et d’un moyen sûr de mener des expériences peut nécessiter un investissement important. Il se peut également que vous deviez modifier le code de l’application pour utiliser des indicateurs de fonctionnalités ou prendre en charge les tests A/B.

Ressources

Bonnes pratiques associées :

- OPS11-BP02 Réaliser une analyse post-incident : les leçons tirées des incidents constituent un moteur important de l'innovation, tout comme de l'expérimentation.
- OPS11-BP03 Implémenter des boucles de rétroaction : les boucles de commentaires jouent un rôle important dans l'expérimentation.

Documents connexes :

- An Inside Look at the Amazon Culture: Experimentation, Failure, and Customer Obsession
- Bonnes pratiques pour créer et gérer des comptes sandbox dans AWS
- Créez une culture d'expérimentation rendue possible par le cloud
- Permettre l'expérimentation et l'innovation dans le cloud chez SulAm Érica Seguros
- Expérimentez plus, échouez moins
- Organisation de votre AWS environnement à l'aide de plusieurs comptes - Sandbox OU
- Utilisation des indicateurs AWS AppConfig de fonctionnalité

Vidéos connexes :

- AWS Sur Air ft. Amazon CloudWatch Evidently | Événements AWS
- AWS En direct, San Francisco Summit 2022 ft. AWS AppConfig Intégration de Feature Flags à Jira
- AWS re:Invent 2022 - Un déploiement n'est pas une version : contrôlez vos lancements à l'aide d'indicateurs de fonctionnalité (05-R) BOA3
- Créez un avec par programmation Compte AWSAWS Control Tower
- Configurez un AWS environnement multi-comptes qui utilise les meilleures pratiques pour AWS Organizations

Exemples connexes :

- AWS Bac à sable innovant
- End-to-endPersonnalisation 101 pour le commerce électronique

Services connexes :

- Amazon, CloudWatch évidemment
- AWS AppConfig
- AWS Control Tower

OPS03-BP06 Les membres de l'équipe sont encouragés à maintenir et à développer leurs compétences

Les équipes doivent accroître leurs compétences pour adopter les nouvelles technologies, et pour faire face à l'évolution de la demande et des responsabilités afin de supporter votre charge de travail. Le développement des compétences dans les nouvelles technologies est souvent une source de satisfaction pour les membres de l'équipe et favorise l'innovation. Aidez les membres de votre équipe à obtenir et à conserver des certifications sectorielles qui valident et reconnaissent leurs compétences croissantes. Mettez en place la formation croisée pour promouvoir le transfert de connaissances et réduire le risque d'impact significatif lorsque vous perdez des membres d'équipe qualifiés et expérimentés ayant un savoir institutionnel. Mettez en place des créneaux dédiés à la formation.

AWS fournit des ressources, notamment le centre de ressources AWS Getting Started, des AWS blogs, des conférences techniques AWS en ligne, AWS des événements et des webinaires, ainsi que les AWS Well-Architected Labs, qui fournissent des conseils, des exemples et des procédures détaillées pour former vos équipes.

Des ressources telles que Support, (AWS Re:Post, Support Center) et la documentation AWS permettent d'éliminer les obstacles techniques et d'améliorer les opérations. Contactez le Support Centre pour Support obtenir de l'aide pour répondre à vos questions.

AWS partage également les meilleures pratiques et les modèles que nous avons appris grâce au fonctionnement de AWSThe Amazon Builders' Library et d'une grande variété d'autres supports pédagogiques utiles via le AWS blog et le podcast officiel. AWS

AWS Training et la certification inclut une formation gratuite par le biais de cours numériques adaptés à votre rythme, ainsi que des plans d'apprentissage par rôle ou par domaine. Vous pouvez également vous inscrire à une formation dispensée par un instructeur afin de renforcer le développement des compétences de vos équipes AWS .

Résultat escompté : votre organisation évalue constamment les lacunes en matière de compétences et les comble grâce à un budget et à des investissements structurés. Les équipes encouragent et incitent leurs membres grâce à des activités de renforcement des compétences, telles que l'obtention de certifications de premier plan dans le secteur. Les équipes tirent parti de programmes de partage

de connaissances dédiés lunch-and-learns, tels que des journées d'immersion, des hackathons et des journées de jeu. Votre organisation conserve ses systèmes de connaissances pertinents pour la formation polyvalente up-to-date des membres de l'équipe, y compris les formations d'intégration des nouveaux employés.

Anti-modèles courants :

- En l’absence d’un programme de formation et d’un budget structurés, les équipes sont confrontées à l’incertitude lorsqu’elles tentent de suivre le rythme de l’évolution technologique, ce qui entraîne une augmentation de l’attrition.
- Dans le cadre de la migration vers le cloud AWS, votre entreprise présente des lacunes en matière de compétences et des différences dans la maîtrise du cloud au sein des équipes. Sans effort de renforcement des compétences, les équipes se retrouvent surchargées par la gestion inefficace et traditionnelle de l’environnement cloud, ce qui accroît la quantité de travail accrue pour les opérateurs. Cette surcharge de travail accroît le mécontentement des employés.

Avantages de la mise en place de cette bonne pratique : lorsque votre organisation investit consciemment dans l’amélioration des compétences de ses équipes, elle contribue également à accélérer et à mettre à l’échelle l’adoption et l’optimisation du cloud. Les programmes de formation ciblés stimulent l’innovation et renforcent la capacité opérationnelle des équipes à se préparer à gérer les événements. Les équipes investissent consciemment dans la mise en œuvre et l’évolution des bonnes pratiques. Le moral de l’équipe est au beau fixe et ses membres apprécient leur contribution à l’entreprise.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Pour adopter de nouvelles technologies, stimuler l’innovation et suivre l’évolution de la demande et des responsabilités afin de pouvoir gérer efficacement vos charges de travail, investissez continuellement dans le développement professionnel de vos équipes.

Étapes d’implémentation

1. Utilisation de programmes structurés de promotion du cloud :AWS Skills Guild propose des formations consultatives pour renforcer la confiance en matière de compétences cloud et renforcer la culture de l’apprentissage continu.
2. Fourniture de ressources de formation : fournissez un temps structuré dédié, l’accès à des supports de formation, des ressources d’atelier et la possibilité de se joindre à des conférences et

à des organisations professionnelles qui offrent des possibilités de formation auprès de formateurs et de pairs. Donnez aux membres de votre équipe junior l'accès à des membres seniors de l'équipe en tant que mentors, ou permettez aux membres juniors de suivre le travail de leurs seniors et de découvrir leurs méthodes et leurs compétences. Encouragez l'apprentissage du contenu qui n'est pas directement lié au travail afin d'avoir une perspective plus large.

3. Encouragement de l'utilisation de ressources techniques spécialisées : tirez parti de ressources comme AWS Re:post pour accéder à des connaissances sélectionnées et à une communauté dynamique.
4. Créez et gérez un référentiel de up-to-date connaissances : utilisez des plateformes de partage de connaissances telles que les wikis et les runbooks. Créez votre propre source de connaissances d'experts réutilisable avec AWS Re:Post Private pour rationaliser la collaboration, améliorer la productivité et accélérer l'intégration des employés.
5. Formation des équipes et engagement entre équipes : planifiez les besoins de formation continue des membres de votre équipe. Offrez aux membres de l'équipe la possibilité de rejoindre d'autres équipes (temporairement ou définitivement) pour partager les compétences et les bonnes pratiques au profit de l'ensemble de votre organisation.
6. Soutien à l'obtention et au maintien des certifications industrielles : soutenez les membres de votre équipe dans l'acquisition et le maintien de certifications industrielles qui valident ce qu'ils ont appris et reconnaissent leurs réalisations.

Niveau d'effort du plan d'implémentation : élevé

Ressources

Bonnes pratiques associées :

- OPS03-BP01 Fournir un parrainage exécutif
- OPS11-BP04 Effectuer la gestion des connaissances

Documents connexes :

- AWS Livre blanc | Cadre d'adoption du cloud : le point de vue des personnes
- Investissement dans l'apprentissage continu pour développer le futur de votre organisation
- AWS Skills Guild
- AWS Training et certification

- Support
- AWS Re : Publier
- AWS Mise en route avec le Centre de ressources
- Blogs AWS
- ConformitéAWS Cloud
- Documentation AWS
- Le AWS podcast officiel.
- AWS Online Tech Talks
- AWS Événements et webinaires
- Ateliers AWS Well-Architected
- Bibliothèque Amazon Builders' Library

Vidéos connexes :

- AWS re:INVENT 2023 | Reskilling at the speed of cloud: Turning employees into entrepreneurs
- WS re:Invent 2023 | Building a culture of curiosity through gamification

OPS03-BP07 Ressources appropriées pour les équipes

Allouez le nombre approprié de membres d'équipe compétents et fournissez les outils et les ressources nécessaires pour répondre à vos besoins en matière de charge de travail. La surcharge des membres de l'équipe accroît le risque d'erreur humaine. Les investissements dans des outils et des ressources, tels que l'automatisation, peuvent mettre à l'échelle l'efficacité de votre équipe et l'aider à gérer efficacement un plus grand nombre de charges de travail sans avoir besoin de capacité supplémentaire.

Résultat escompté :

- Vous avez doté votre équipe du personnel approprié pour acquérir les compétences nécessaires pour gérer les charges de travail conformément AWS à votre plan de migration. Au fur et à mesure que votre équipe s'est développée au cours de votre projet de migration, elle a acquis des compétences dans les AWS technologies de base que l'entreprise prévoit d'utiliser lors de la migration ou de la modernisation de ses applications.
- Vous avez soigneusement aligné votre plan de dotation en personnel afin d'utiliser efficacement les ressources en tirant parti de l'automatisation et du flux de travail. Une équipe plus petite

peut désormais gérer une plus grande partie de l’infrastructure pour le compte des équipes de développement d’applications.

- Compte tenu de l’évolution des priorités opérationnelles, toutes les contraintes en matière de ressources humaines sont identifiées de manière proactive afin de protéger le succès des initiatives commerciales.
- Les métriques opérationnelles qui font état du labeur opérationnel (comme la fatigue liée au travail d’astreinte ou les appels excessifs) sont passés en revue pour vérifier que le personnel n’est pas dépassé.

Anti-modèles courants :

- Votre personnel n'a pas développé ses AWS compétences alors que vous approchez de votre plan pluriannuel de migration vers le cloud, ce qui risque de supporter les charges de travail et de réduire le moral des employés.
- L’ensemble de votre organisation informatique est en train de passer à des méthodes de travail agiles. L’entreprise donne la priorité au portefeuille de produits et définit des métriques pour les fonctionnalités qui doivent être développées en premier. Votre processus agile n’oblige pas les équipes à attribuer des points d’histoire à leurs plans de travail. Par conséquent, il est impossible de connaître le niveau de capacité requis pour le prochain volume de travail ou de déterminer si vous possédez les compétences appropriées pour le travail à accomplir.
- Vous demandez à un AWS partenaire de migrer vos charges de travail et vous n'avez pas de plan de transition de support pour vos équipes une fois que le partenaire a terminé le projet de migration. Vos équipes ont du mal à gérer les charges de travail de manière efficiente et efficace.

Avantages de la mise en place de cette bonne pratique : vous disposez au sein de votre organisation de membres d’équipe possédant les compétences nécessaires pour gérer les charges de travail. L’allocation des ressources peut s’adapter à l’évolution des priorités sans affecter les performances. Les équipes sont donc capables de gérer efficacement les charges de travail tout en maximisant le temps nécessaire pour se concentrer sur l’innovation pour les clients, ce qui augmente la satisfaction des employés.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

La planification des ressources pour votre migration vers le cloud doit être effectuée à un niveau organisationnel qui correspond à votre plan de migration, ainsi qu’au modèle opérationnel souhaité

mis en œuvre pour prendre en charge votre nouvel environnement cloud. Cela devrait inclure la compréhension des technologies cloud déployées pour les équipes commerciales et de développement d'applications. La direction de l'infrastructure et des opérations doit planifier l'analyse des lacunes en matière de compétences, la formation et la définition des rôles des ingénieurs qui dirigent l'adoption du cloud.

Étapes d'implémentation

1. Définissez les critères de réussite de l'équipe à l'aide de mesures opérationnelles pertinentes telles que la productivité du personnel (par exemple, le coût de prise en charge d'une charge de travail ou les heures passées par l'opérateur lors d'incidents).
2. Définissez des mécanismes de planification et d'inspection de la capacité en matière de ressources pour confirmer que la quantité appropriée de capacités qualifiées est disponible en cas de besoin et qu'elle pourra être ajustée au fil du temps.
3. Créez des mécanismes (par exemple, l'envoi d'une enquête mensuelle aux équipes) pour comprendre les défis liés au travail qui ont un impact sur les équipes (comme l'augmentation des responsabilités, les changements technologiques, la perte de personnel ou l'augmentation du nombre de clients pris en charge).
4. Utilisez ces mécanismes pour interagir avec les équipes et identifier les tendances susceptibles de contribuer aux problèmes de productivité des employés. Lorsque vos équipes sont affectées par des facteurs externes, réévaluez les objectifs et ajustez les cibles le cas échéant. Identifiez les obstacles qui entravent la progression de votre équipe.
5. Vérifiez régulièrement si les ressources allouées restent suffisantes ou si des ressources supplémentaires sont nécessaires, et apportez les ajustements appropriés pour soutenir les équipes.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS03-BP06 Les membres de l'équipe sont encouragés à maintenir et à développer leurs compétences
- OPS09-BP03 Examiner les indicateurs des opérations et prioriser les améliorations
- OPS10-BP01 Utiliser un processus pour la gestion des événements, des incidents et des problèmes

- OPS10-BP07 Automatiser les réponses aux événements

Documents connexes :

- AWS Cloud Cadre d'adoption : point de vue des personnes
- Becoming a Future-Ready Enterprise
- Prioritize your Employees' Skills to Drive Business Growth
- High performing organization - the Amazon Two-Pizza team
- How Cloud-Mature Enterprises Succeed

Préparation

Questions

- OPS 4. Comment mettre en œuvre l'observabilité dans votre charge de travail ?
- OPS 5. Comment réduire les défauts, faciliter les corrections et améliorer le flux dans la production ?
- OPS 6. Comment réduire les risques liés au déploiement ?
- OPS 7. Comment savoir si vous êtes prêt à assurer une charge de travail ?

OPS 4. Comment mettre en œuvre l'observabilité dans votre charge de travail ?

Intégrez l'observabilité à votre charge de travail afin de comprendre son état et de prendre des décisions basées sur les données en fonction des exigences de l'entreprise.

Bonnes pratiques

- OPS04-BP01 Identifier les indicateurs de performance clés
- OPS04-BP02 Implémenter la télémétrie des applications
- OPS04-BP03 Implémenter la télémétrie de l'expérience utilisateur
- OPS04-BP04 Implémenter la télémétrie des dépendances
- OPS04-BP05 Mettre en œuvre le traçage distribué

OPS04-BP01 Identifier les indicateurs de performance clés

La mise en œuvre de l’observabilité dans votre charge de travail commence par la compréhension de son état et par la prise de décisions basées sur les données en fonction des exigences de l’entreprise. L'un des moyens les plus efficaces de garantir l'alignement entre les activités de surveillance et les objectifs commerciaux consiste à définir et à suivre des indicateurs de performance clés (KPIs).

Résultat escompté : pratiques d’observabilité efficaces qui sont étroitement alignées sur les objectifs commerciaux, garantissant que les efforts de surveillance sont toujours au service de résultats commerciaux tangibles.

Anti-modèles courants :

- Non défini KPIs : le fait de travailler sans clarté KPIs peut entraîner une surveillance trop importante ou insuffisante, ce qui peut entraîner l'absence de signaux vitaux.
- Statique KPIs : ne pas revoir ou affiner au KPIs fur et à mesure de l'évolution de la charge de travail ou des objectifs commerciaux.
- Désalignement : se concentrer sur des métriques techniques qui ne sont pas directement corrélées aux résultats commerciaux ou qui sont plus difficiles à corréler aux problèmes réels.

Avantages liés au respect de cette bonne pratique :

- Facilité d'identification des problèmes : les entreprises identifient KPIs souvent les problèmes plus clairement que les indicateurs techniques. Une baisse d'activité KPI permet d'identifier un problème plus efficacement que de passer au crible de nombreux indicateurs techniques.
- Cohérence des activités : garantit que les activités de surveillance soutiennent directement les objectifs commerciaux.
- Efficacité : la priorité est donnée à la surveillance des ressources et l’attention est concentrée sur les métriques déterminantes.
- Proactivité : identifiez et traitez les problèmes avant qu’ils n’aient des implications commerciales plus larges.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Pour définir efficacement la charge de travail KPIs :

1. Commencement par les résultats commerciaux : avant de vous plonger dans les métriques, déterminez les résultats commerciaux souhaités. S'agit-il d'une augmentation des ventes, d'un engagement plus élevé des utilisateurs ou d'une réduction des temps de réponse ?
2. Corrélation des métriques techniques avec les objectifs commerciaux : les métriques techniques n'ont pas toutes un impact direct sur les résultats commerciaux. Identifiez ceux qui le font, mais il est souvent plus simple d'identifier un problème dans le cadre d'une entrepriseKPI.
3. Utilisez Amazon CloudWatch : Employ CloudWatch pour définir et surveiller les indicateurs qui représentent votreKPIs.
4. Révision et mise à jour régulières KPIs : au fur et à mesure de l'évolution de votre charge de travail et de votre activité, restez KPIs pertinent.
5. Impliquer les parties prenantes : Impliquer les équipes techniques et commerciales dans la définition et la révisionKPIs.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- the section called "OPS04-BP02 Implémenter la télémétrie des applications"
- the section called "OPS04-BP03 Implémenter la télémétrie de l'expérience utilisateur"
- the section called "OPS04-BP04 Implémenter la télémétrie des dépendances"
- the section called "OPS04-BP05 Mettre en œuvre le traçage distribué"

Documents connexes :

- AWS Meilleures pratiques en matière d'observabilité
- CloudWatch Guide de l'utilisateur
- AWS Cours de renforcement des compétences en observabilité

Vidéos connexes :

- Developing an observability strategy

Exemples connexes :

- Un atelier sur l’observabilité

OPS04-BP02 Implémenter la télémétrie des applications

La télémétrie de l’application est la pierre angulaire de l’observabilité de votre charge de travail. Il est essentiel de diffuser des données télémétriques fournissant des informations exploitables sur l’état de votre application et sur son taux de réussite par rapport aux résultats techniques et commerciaux. Qu'il s'agisse de résoudre des problèmes, de mesurer l'impact d'une nouvelle fonctionnalité ou de garantir l'alignement sur les indicateurs de performance clés de l'entreprise (KPIs), la télémétrie des applications vous permet de créer, d'exploiter et de faire évoluer votre charge de travail.

Les métriques, les journaux et les données de suivi constituent les trois principaux piliers de l’observabilité. Ils servent d’outils de diagnostic qui décrivent l’état de votre application. Au fil du temps, ils contribuent à créer des points de référence et à identifier les anomalies. Cependant, pour garantir l'alignement entre les activités de surveillance et les objectifs commerciaux, il est essentiel de les définir et de les surveillerKPIs. KPIsLes entreprises facilitent souvent l'identification des problèmes par rapport aux seuls indicateurs techniques.

D'autres types de télémétrie, tels que la surveillance des utilisateurs réels (RUM) et les transactions synthétiques, complètent ces sources de données principales. RUMfournit des informations sur les interactions des utilisateurs en temps réel, tandis que les transactions synthétiques simulent les comportements potentiels des utilisateurs, aidant ainsi à détecter les goulets d'étranglement avant que les utilisateurs réels ne les rencontrent.

Résultat escompté : obtenez des informations exploitables sur les performances de votre charge de travail. Ces informations vous permettront de prendre des décisions proactives concernant l’optimisation des performances, d’accroître la stabilité de la charge de travail, de rationaliser les processus CI/CD et d’utiliser efficacement les ressources.

Anti-modèles courants :

- Observabilité incomplète : le fait de négliger d’intégrer l’observabilité à chaque niveau de la charge de travail entraîne des angles morts susceptibles de masquer des informations essentielles sur les performances et le comportement du système.
- Vue fragmentée des données : lorsque les données sont dispersées entre plusieurs outils et systèmes, il devient difficile de conserver une vision globale de l’état et des performances de la charge de travail.

- Problèmes signalés par les utilisateurs : cela indique que la détection proactive des problèmes par le biais de la télémétrie et de la KPI surveillance des activités fait défaut.

Avantages liés au respect de cette bonne pratique :

- Prise de décision éclairée : grâce aux informations issues de la télémétrie et des activités commercialesKPIs, vous pouvez prendre des décisions basées sur les données.
- Efficacité opérationnelle améliorée : l'utilisation des ressources axée sur les données est source de rentabilité.
- Stabilité accrue de la charge de travail : détection et résolution plus rapides des problèmes, ce qui améliore la disponibilité.
- Processus CI/CD rationalisés : les informations issues des données de télémétrie facilitent l'affinement des processus et la livraison fiable du code.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Pour implémenter la télémétrie des applications pour votre charge de travail, utilisez AWS des services tels qu'Amazon CloudWatch et. AWS X-Ray Amazon CloudWatch fournit une suite complète d'outils de surveillance, qui vous permet d'observer vos ressources et vos applications dans AWS et sur site. Il collecte, suit et analyse les métriques, consolide et surveille les données des journaux, et répond à l'évolution de vos ressources, vous permettant ainsi de mieux comprendre le fonctionnement de votre charge de travail. En tandem, vous AWS X-Ray permet de suivre, d'analyser et de déboguer vos applications, ce qui vous permet de mieux comprendre le comportement de votre charge de travail. Des fonctionnalités telles que les cartes de service, les distributions de latence et les chronologies de suivi AWS X-Ray fournissent des informations sur les performances de votre charge de travail et les obstacles qui l'affectent.

Étapes d'implémentation

1. Identification des données à collecter : déterminez les métriques, les journaux et les données de suivi essentiels qui fourniraient des informations substantielles sur l'état, les performances et le comportement de votre charge de travail.
2. Déployez l'CloudWatchagent : l' CloudWatch agent joue un rôle essentiel dans l'obtention de métriques et de journaux du système et des applications à partir de votre charge de travail et de

son infrastructure sous-jacente. L' CloudWatch agent peut également être utilisé pour collecter OpenTelemetry ou radiographier des traces et les envoyer à X-Ray.

3. Mettez en œuvre la détection des anomalies pour les journaux et les métriques : utilisez la détection CloudWatch des anomalies des journaux et la détection CloudWatch des anomalies des métriques pour identifier automatiquement les activités inhabituelles dans les opérations de votre application. Ces outils utilisent des algorithmes de machine learning pour détecter les anomalies et émettre des alertes en cas d'anomalie, ce qui améliore vos capacités de surveillance et accélère le temps de réponse en cas de perturbations ou de menaces de sécurité potentielles. Configurez ces fonctionnalités pour gérer de manière proactive l'intégrité et la sécurité des applications.
4. Sécurisez les données de journal sensibles : utilisez la protection des données Amazon CloudWatch Logs pour masquer les informations sensibles contenues dans vos journaux. Cette fonctionnalité permet de préserver la confidentialité et la conformité grâce à la détection automatique et au masquage des données sensibles avant leur accès. Mettez en œuvre le masquage des données pour gérer et protéger en toute sécurité les informations sensibles telles que les informations personnelles identifiables (PII).
5. Définissez et surveillez les activités KPIs : établissez des indicateurs personnalisés qui correspondent aux résultats de votre entreprise.
6. Instrumentez votre application avec AWS X-Ray : Outre le déploiement de l' CloudWatchagent, il est essentiel d'instrumenter votre application pour émettre des données de trace. Ce processus peut fournir des informations supplémentaires sur le comportement et les performances de votre charge de travail.
7. Standardisation de la collecte de données dans l'ensemble de votre application : standardisez les pratiques de collecte de données dans l'ensemble de votre application. L'uniformité facilite la corrélation et l'analyse des données, fournissant ainsi une vue complète du comportement de votre application.
8. Mettez en œuvre l'observabilité entre comptes : améliorez l'efficacité de la surveillance sur plusieurs comptes grâce à l'observabilité entre comptes Comptes AWS Amazon CloudWatch . Grâce à cette fonctionnalité, vous pouvez consolider les métriques, les journaux et les alarmes de différents comptes en une seule vue, ce qui simplifie la gestion et améliore les temps de réponse aux problèmes identifiés dans l' AWS environnement de votre entreprise.
9. Analysez les données et agissez en conséquence : une fois que la collecte et la normalisation des données sont en place, utilisez Amazon CloudWatch pour l'analyse des métriques et des journaux, ainsi que AWS X-Raypour l'analyse des traces. Une telle analyse peut fournir des informations cruciales sur l'état, les performances et le comportement de votre charge de travail, orientant ainsi votre processus décisionnel.

Niveau d’effort du plan d’implémentation : élevé

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Définir la charge de travail KPIs
- OPS04-BP03 Implémenter la télémétrie de l'activité des utilisateurs
- OPS04-BP04 Implémenter la télémétrie des dépendances
- OPS04-BP05 Mettre en œuvre la traçabilité des transactions

Documents connexes :

- Bonnes pratiques AWS en matière d’observabilité
- Guide de l’utilisateur CloudWatch
- AWS X-Ray Manuel du développeur
- Instrumentation des systèmes distribués au profit de la visibilité opérationnelle
- Cours de renforcement des compétences en observabilitéAWS
- Nouveautés d'Amazon CloudWatch
- Quoi de neuf avec AWS X-Ray

Vidéos connexes :

- AWS re:Invent 2022 - Meilleures pratiques en matière d'observabilité sur Amazon
- AWS re:Invent 2022 - Élaboration d'une stratégie d'observabilité

Exemples associés :

- Un atelier sur l’observabilité
- AWS Bibliothèque de solutions : surveillance des applications avec Amazon CloudWatch

OPS04-BP03 Implémenter la télémétrie de l'expérience utilisateur

Il est essentiel d’obtenir des informations approfondies sur les expériences des clients et leurs interactions avec votre application. La surveillance des utilisateurs réels (RUM) et les transactions synthétiques constituent de puissants outils à cette fin. RUMfournit des données sur les interactions

réelles des utilisateurs, offrant une perspective non filtrée de la satisfaction des utilisateurs, tandis que les transactions synthétiques simulent les interactions des utilisateurs, aidant à détecter les problèmes potentiels avant même qu'ils n'affectent les utilisateurs réels.

Résultat escompté :une vision globale de l’expérience client, une détection proactive des problèmes et une optimisation des interactions avec les utilisateurs pour proposer des expériences numériques fluides.

Anti-modèles courants :

- Applications sans véritable surveillance des utilisateurs (RUM) :
  - Détection différée des problèmes : sans celaRUM, vous ne vous rendrez peut-être pas compte de l'existence de problèmes ou de problèmes de performances tant que les utilisateurs ne se seront pas plaints. Cette approche réactive peut entraîner l’insatisfaction des clients.
  - Manque d'informations sur l'expérience utilisateur : si vous RUM ne l'utilisez pas, vous perdez des données cruciales qui montrent comment les utilisateurs réels interagissent avec votre application, ce qui limite votre capacité à optimiser l'expérience utilisateur.
- Applications sans transactions synthétiques :
  - Cas marginaux manqués : les transactions synthétiques vous aident à tester des chemins et des fonctions qui ne sont pas toujours fréquemment utilisés par les utilisateurs ordinaires, mais qui sont essentiels à certaines fonctions commerciales. Sans ces transactions synthétiques, ces chemins pourraient mal fonctionner et passer inaperçus.
  - Recherche de problèmes lorsque l’application n’est pas utilisée : des tests synthétiques réguliers permettent de simuler les situations où les utilisateurs réels n’interagissent pas activement avec votre application, garantissant ainsi le bon fonctionnement du système.

Avantages liés au respect de cette bonne pratique :

- Détection proactive des problèmes : identifiez et résolvez les problèmes potentiels avant qu’ils n’affectent les utilisateurs réels.
- Expérience utilisateur optimisée : le feedback continu RUM permet d'affiner et d'améliorer l'expérience utilisateur globale.
- Informations sur les performances de l’appareil et du navigateur : comprenez le fonctionnement de votre application sur différents appareils et navigateurs, afin de l’affiner davantage.
- Flux de travail validés : des transactions synthétiques régulières garantissent que les fonctionnalités de base et les chemins critiques restent opérationnels et efficaces.

- Performances améliorées des applications : exploitez les informations recueillies à partir de données sur les utilisateurs réels pour améliorer la réactivité et la fiabilité des applications.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Pour exploiter RUM et synthétiser les transactions à des fins de télémétrie de l'activité des utilisateurs, AWS propose des services tels qu'Amazon et CloudWatch RUMAmazon CloudWatch Synthetics. Les métriques, les journaux et les données de suivi, associés aux données d’activité des utilisateurs, fournissent une vue complète de l’état de fonctionnement de l’application et de l’expérience utilisateur.

Étapes d’implémentation

1. Déployez Amazon CloudWatch RUM : intégrez votre application CloudWatch RUM pour collecter, analyser et présenter des données utilisateur réelles.
   a. Utilisez la CloudWatch RUM JavaScript bibliothèque pour l'intégrer RUM à votre application.
   b. Configurez des tableaux de bord pour visualiser et surveiller les données sur les utilisateurs réels.
2. Configurer CloudWatch Synthetics : créez des canaris, ou des routines scriptées, qui simulent les interactions des utilisateurs avec votre application.
   a. Définissez les flux de travail et les chemins d’application critiques.
   b. Concevez des canaris à l'aide CloudWatch de scripts Synthetics pour simuler les interactions des utilisateurs sur ces trajectoires.
   c. Planifiez et surveillez les scripts canary pour qu’ils fonctionnent à des intervalles spécifiés, afin de garantir des contrôles de performance cohérents.
3. Analysez les données et agissez en fonction de celles-ci : utilisez les données issues RUM des transactions synthétiques pour obtenir des informations et prendre des mesures correctives lorsque des anomalies sont détectées. Utilisez des CloudWatch tableaux de bord et des alarmes pour rester informé.

Niveau d’effort du plan d’implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identifier les indicateurs de performance clés
- OPS04-BP02 Implémenter la télémétrie des applications
- OPS04-BP04 Implémenter la télémétrie des dépendances
- OPS04-BP05 Mettre en œuvre le traçage distribué

Documents connexes :

- CloudWatch RUMGuide Amazon
- Guide Amazon CloudWatch Synthetics

Vidéos connexes :

- Optimisez les applications grâce aux informations sur les utilisateurs finaux avec Amazon CloudWatch RUM
- AWS sur Air ft. Surveillance des utilisateurs réels pour Amazon CloudWatch

Exemples connexes :

- Un atelier sur l’observabilité
- Référentiel Git pour Amazon CloudWatch RUM Web Client
- Utilisation d'Amazon CloudWatch Synthetics pour mesurer le temps de chargement des pages

OPS04-BP04 Implémenter la télémétrie des dépendances

La télémétrie des dépendances est essentielle pour surveiller l’état et les performances des services et composants externes sur lesquels repose votre charge de travail. Il fournit des informations précieuses sur l'accessibilité, les délais d'attente et d'autres événements critiques liés aux dépendances telles que les bases de données ou DNS les tiers. APIs Lorsque vous instrumentez votre application de sorte à émettre des métriques, des journaux et des données de suivi concernant ces dépendances, vous identifiez plus facilement les goulets d’étranglement potentiels, les problèmes de performances ou les défaillances susceptibles d’avoir un impact sur votre charge de travail.

Résultat escompté : assurez-vous que les dépendances sur lesquelles repose votre charge de travail fonctionnent comme prévu, ce qui vous permet de résoudre les problèmes de manière proactive et de garantir des performances de charge de travail optimales.

Anti-modèles courants :

- Omission des dépendances externes : se concentrer uniquement sur les métriques internes des applications tout en négligeant les métriques liées aux dépendances externes.
- Absence de surveillance proactive : attendre l'apparition de problèmes au lieu de surveiller en permanence l'état et les performances des dépendances.
- Surveillance cloisonnée : utiliser des outils de surveillance divers et variés qui peuvent donner lieu à des visions fragmentées et incohérentes de l'état des dépendances.

Avantages liés au respect de cette bonne pratique :

- Fiabilité améliorée de la charge de travail : en garantissant que les dépendances externes sont constamment disponibles et fonctionnent de manière optimale.
- Détection et résolution plus rapides des problèmes : en identifiant et en résolvant de manière proactive les problèmes liés aux dépendances avant qu'ils n'affectent la charge de travail.
- Vue globale : grâce à une visibilité complète des composants internes et externes qui influencent l'état de la charge de travail.
- Meilleure capacité de mise à l'échelle de la charge de travail : grâce à une meilleure compréhension des limites de la capacité de mise à l'échelle et des caractéristiques de performance des dépendances externes.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Mettez en œuvre la télémétrie des dépendances en commençant par identifier les services, l'infrastructure et les processus sur lesquels repose votre charge de travail. Quantifiez ce à quoi les conditions favorables ressemblent lorsque ces dépendances fonctionnent comme prévu, puis déterminez les données nécessaires pour les mesurer. Ces informations vous permettront de créer des tableaux de bord et des alertes qui fourniront à vos équipes opérationnelles des informations sur l'état de ces dépendances. Utilisez AWS des outils pour découvrir et quantifier les impacts lorsque les dépendances ne peuvent pas répondre aux besoins. Revoyez continuellement votre stratégie en tenant compte de l'évolution des priorités, des objectifs et des connaissances acquises.

Étapes d'implémentation

Pour implémenter efficacement la télémétrie des dépendances :

1. Identification des dépendances externes : collaborez avec les parties prenantes pour identifier les dépendances externes sur lesquelles repose votre charge de travail. Les dépendances externes peuvent englober des services tels que des bases de données externes, des services tiersAPIs, des routes de connectivité réseau vers d'autres environnements et des DNS services. La première étape à suivre pour assurer l'efficacité de la télémétrie des dépendances consiste à comprendre parfaitement ce que sont ces dépendances.
2. Élaboration d'une stratégie de suivi : une fois que vous avez une idée précise de vos dépendances externes, élaborez une stratégie de surveillance qui leur est adaptée. Cela implique de comprendre le caractère critique de chaque dépendance, son comportement attendu et tous les accords ou cibles de niveau de service associés (SLAou). SLTs Configurez des alertes proactives pour vous informer des changements d'état ou des écarts de performance.
3. Utilisation de la surveillance du réseau : utilisez Internet Monitor et Network Monitor, qui fournissent des informations complètes sur l'état mondial de l'Internet et du réseau. Ces outils vous aident à comprendre les pannes, les interruptions ou les dégradations de performances qui affectent vos dépendances externes et à y répondre.
4. Restez informé avec AWS Health Dashboard: il fournit des alertes et des conseils de résolution en cas AWS d'événements susceptibles d'avoir un impact sur vos services.
    a. Surveillez AWS Health les événements EventBridge selon les règles d'Amazon ou intégrez-les par programmation AWS Health API pour automatiser les actions lorsque vous recevez des AWS Health événements. Il peut s'agir d'actions générales, telles que l'envoi de tous les messages relatifs aux événements du cycle de vie planifiés vers une interface de discussion, ou d'actions spécifiques, telles que le lancement d'un flux de travail dans un outil de gestion des services informatiques.
    b. Si vous en utilisez AWS Organizations, regroupez AWS Health les événements entre les comptes.
5. Instrumentez votre application avec AWS X-Ray: AWS X-Ray fournit des informations sur les performances des applications et de leurs dépendances sous-jacentes. En suivant les requêtes du début à la fin, vous pouvez identifier les goulets d'étranglement ou les défaillances des services ou composants externes sur lesquels repose votre application.
6. Utilisez Amazon DevOps Guru : ce service basé sur l'apprentissage automatique identifie les problèmes opérationnels, prédit le moment où des problèmes critiques peuvent survenir et recommande des mesures spécifiques à prendre. Il s'agit d'un outil inestimable qui permet de mieux comprendre les dépendances et de déterminer qu'elles ne sont pas à l'origine de problèmes opérationnels.

7. Surveillance régulière : surveillez en permanence les métriques et les journaux liés aux dépendances externes. Configurez des alertes en cas de comportement inattendu ou de dégradation des performances.
8. Validation après les modifications : chaque fois qu'une dépendance externe est mise à jour ou modifiée, validez ses performances et vérifiez qu'elle correspond aux exigences de votre application.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Définir la charge de travail KPIs
- OPS04-BP02 Implémenter la télémétrie des applications
- OPS04-BP03 Implémenter la télémétrie de l'activité des utilisateurs
- OPS04-BP05 Mettre en œuvre la traçabilité des transactions
- OP08-BP04 Création d'alertes exploitables

Documents connexes :

- Guide de AWS Health Dashboard l'utilisateur Amazon Personal
- Guide de l'utilisateur d'AWS Internet Monitor
- AWS X-Ray Manuel du développeur
- AWS DevOpsGuide de l'utilisateur Guru

Vidéos connexes :

- Visibility into how internet issues impact app performance
- Présentation d'Amazon DevOps Guru
- Gérez les événements liés au cycle de vie des ressources à grande échelle avec AWS Health

Exemples connexes :

- Obtenir des informations opérationnelles AIOps grâce à Amazon DevOps Guru

- AWS Health Conscient
- Utilisation du filtrage basé sur des balises pour gérer la AWS Health surveillance et les alertes à grande échelle

OPS04-BP05 Mettre en œuvre le traçage distribué

Le suivi distribué permet de surveiller et de visualiser les requêtes lorsqu'elles traversent les différents composants d'un système distribué. En capturant les données de suivi provenant de plusieurs sources et en les analysant dans une vue unifiée, les équipes peuvent mieux comprendre le flux des requêtes, les endroits où les goulots d'étranglement ont lieu et les domaines dans lesquels les efforts d'optimisation doivent se concentrer.

Résultat escompté : bénéficiez d'une vue globale des requêtes circulant dans votre système distribué, ce qui permet un débogage précis, des performances optimisées et une meilleure expérience utilisateur.

Anti-modèles courants :

- Instrumentation incohérente : les services d'un système distribué ne sont pas tous instrumentés pour le suivi.
- Ignorer la latence : se concentrer uniquement sur les erreurs et ne pas tenir compte de la latence ou de la dégradation progressive des performances.

Avantages liés au respect de cette bonne pratique :

- Vue d'ensemble complète du système : visualisation du parcours complet des requêtes, de l'entrée à la sortie.
- Débogage amélioré : identification rapide des défaillances ou des problèmes de performance.
- Expérience utilisateur améliorée : surveillance et optimisation basées sur des données sur les utilisateurs réels, afin de garantir que le système répond aux exigences du monde réel.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Commencez par identifier tous les éléments de votre charge de travail qui nécessitent de l'instrumentation. Une fois que tous les composants sont pris en compte, utilisez des outils tels

que AWS X-Ray et OpenTelemetry pour collecter des données de trace à des fins d'analyse avec des outils tels que X-Ray et Amazon CloudWatch ServiceLens Map. Participez à des évaluations régulières avec les développeurs et complétez ces discussions avec des outils tels qu'Amazon DevOps Guru, X-Ray Analytics et X-Ray Insights pour vous aider à découvrir des résultats plus approfondis. Définissez des alertes à partir des données de suivi pour envoyer une notification lorsque les résultats, tels que décrits dans le plan de surveillance de la charge de travail, sont menacés.

Étapes d’implémentation

Pour mettre en œuvre efficacement le suivi distribué :

1. Adoption de AWS X-Ray : intégrez X-Ray à votre application pour mieux comprendre son comportement, interpréter ses performances et identifier les goulots d’étranglement. Utilisez X-Ray Insights pour l’analyse automatique des données de suivi.
2. Instrumentez vos services : vérifiez que chaque service, qu'il s'agisse d'une AWS Lambdafonction ou d'une EC2instance, envoie des données de suivi. Plus vous instrumentez de services, plus la end-to-end vue est claire.
3. Intégrez la surveillance des utilisateurs CloudWatch réels et la surveillance synthétique : intégrez la surveillance des utilisateurs réels (RUM) et la surveillance synthétique avec X-Ray. Cela permet de capturer des expériences utilisateur réelles et de simuler les interactions des utilisateurs afin d’identifier les problèmes potentiels.
4. Utiliser l'CloudWatch agent : l'agent peut envoyer des traces à partir de X-Ray ou OpenTelemetry pour améliorer la profondeur des informations obtenues.
5. Utilisez Amazon DevOps Guru : DevOps Guru utilise les données de X-Ray CloudWatch, AWS Config, et AWS CloudTrail pour fournir des recommandations exploitables.
6. Analyse des traces : passez régulièrement en revue les données de suivi pour identifier les tendances, les anomalies ou les goulots d’étranglement susceptibles d’avoir un impact sur les performances de votre application.
7. Configurez des alertes : configurez les alarmes en fonction CloudWatchde modèles inhabituels ou de latences prolongées, ce qui permet de résoudre les problèmes de manière proactive.
8. Amélioration continue : revoyez votre stratégie de suivi au fur et à mesure que des services sont ajoutés ou modifiés afin de capturer tous les points de données pertinents.

Niveau d’effort du plan d’implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identifier les indicateurs de performance clés
- OPS04-BP02 Implémenter la télémétrie des applications
- OPS04-BP03 Implémenter la télémétrie de l'expérience utilisateur
- OPS04-BP04 Implémenter la télémétrie des dépendances

Documents connexes :

- AWS X-Ray Guide du développeur
- Guide de CloudWatch l'utilisateur d'Amazon Agent
- Guide de l'utilisateur Amazon DevOps Guru

Vidéos connexes :

- Utilisez AWS X-Ray Insights
- AWS sur Air ft. Observabilité : Amazon CloudWatch et AWS X-Ray

Exemples connexes :

- Instrumentation de votre application pour AWS X-Ray

OPS 5. Comment réduire les défauts, faciliter les corrections et améliorer le flux dans la production ?

Adoptez des approches qui améliorent l’entrée des modifications en production et qui permettent une refactorisation, des retours rapides sur la qualité et la correction de bogues. Cela permet d’accélérer l’entrée des modifications bénéfiques en production, de limiter le déploiement de problèmes et d’identifier et de corriger rapidement les problèmes introduits par les activités de déploiement.

Bonnes pratiques

- OPS05-BP01 Utilisation du contrôle de version
- OPS05-BP02 Tester et valider les modifications
- OPS05-BP03 Utiliser des systèmes de gestion de la configuration

- OPS05-BP04 Utiliser des systèmes de gestion du développement et du déploiement
- OPS05-BP05 Effectuer la gestion des correctifs
- OPS05-BP06 Partager les normes de conception
- OPS05-BP07 Mise en œuvre de pratiques visant à améliorer la qualité du code
- OPS05-BP08 Utilisation de plusieurs environnements
- OPS05-BP09 Procéder à des modifications fréquentes, mineures et réversibles
- OPS05- BP1 0 Automatisez entièrement l'intégration et le déploiement

OPS05-BP01 Utilisation du contrôle de version

Utilisez le contrôle de version pour activer le suivi des modifications et des versions.

De nombreux services AWS offrent des fonctionnalités de contrôle de version. Utilisez un système de révision ou de contrôle des sources comme Git pour gérer le code et d'autres artefacts, tels que les modèles AWS CloudFormation de contrôle de versions de votre infrastructure.

Résultat escompté : vos équipes collaborent sur le code. Une fois fusionné, le code est cohérent et aucune modification n'est perdue. Les erreurs sont facilement corrigées grâce à une gestion des versions appropriée.

Anti-modèles courants :

- Vous avez développé et stocké le code sur votre poste de travail. Un problème de stockage s'est produit sur le poste de travail et vous avez perdu le code.
- Après avoir remplacé le code existant par vos modifications, vous redémarrez votre application et elle n'est plus utilisable. Vous ne pouvez pas annuler la modification.
- Vous disposez d'un verrou d'écriture sur un fichier de rapport que quelqu'un d'autre doit modifier. Il vous contacte pour vous demander d'arrêter de travailler dessus afin qu'il puisse effectuer ses tâches.
- Votre équipe de recherche a travaillé sur une analyse détaillée qui façonnera vos futurs travaux. Quelqu'un a accidentellement enregistré sa liste d'achats sur le rapport final. Vous ne pouvez pas annuler la modification et vous devrez recréer le rapport.

Avantages liés au respect de cette bonne pratique : en utilisant les fonctionnalités de contrôle de version, vous pouvez revenir facilement aux bons états connus et aux versions précédentes, et limiter le risque de perte de ressources.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Maintenez les ressources dans des référentiels avec contrôle de version. Cela permet le suivi des modifications, le déploiement de nouvelles versions, la détection des modifications apportées aux versions existantes, et le retour à des versions antérieures (par exemple, la restauration à un état correct connu en cas de défaillance). Intégrez les fonctionnalités de contrôle de version de vos systèmes de gestion de la configuration dans vos procédures.

Ressources

Bonnes pratiques associées :

- OPS05-BP04 Utiliser des systèmes de gestion du développement et du déploiement

Vidéos connexes :

- AWS re:Invent 2023 - How Lockheed Martin builds software faster, powered by DevSecOps
- AWS re:Invent 2023 - How GitHub operationalizes AI for team collaboration and productivity

OPS05-BP02 Tester et valider les modifications

Chaque changement déployé doit être testé pour éviter des erreurs de production. Cette bonne pratique est axée sur les tests des changements du contrôle des versions à la création d’artefacts. En plus des changements du code de l’application, les tests doivent inclure l’infrastructure, la configuration, les contrôles de sécurité et les procédures opérationnelles. Les tests prennent de nombreuses formes, des tests unitaires à l'analyse des composants logiciels (SCA). Pousser les tests encore plus loin dans le processus d’intégration et de livraison de logiciels entraîne une plus grande certitude de la qualité des artefacts.

Votre organisation doit développer des normes de test pour tous les artefacts logiciels. Les tests automatisés réduisent la quantité de travail et évitent les erreurs de test manuels. Des tests manuels peuvent être nécessaires dans certains cas. Les développeurs doivent avoir accès aux résultats des tests automatisés pour créer des boucles de rétroaction qui améliorent la qualité du logiciel.

Résultat escompté : les changements apportés au logiciel sont testés avant d’être livrés. Les développeurs ont accès aux résultats des tests et aux validations. Votre organisation a une norme de test qui s’applique à tous les changements apportés au logiciel.

Anti-modèles courants :

- Vous déployez un nouveau changement apporté au logiciel sans aucun test. Il ne s'exécute pas en production, ce qui entraîne une panne.
- Les nouveaux groupes de sécurité sont déployés AWS CloudFormation sans être testés dans un environnement de pré-production. Les groupes de sécurité empêchent les clients d'atteindre votre application.
- Une méthode est modifiée mais il n'existe aucun test d'unité. Le logiciel échoue quand il est déployé en production.

Avantages liés au respect de cette bonne pratique : le taux d'échec des changements dans les déploiements de logiciels est réduit. La qualité du logiciel s'améliore. Les développeurs ont une meilleure connaissance de la viabilité de leur code. Des politiques de sécurité peuvent être déployées en toute confiance pour soutenir la conformité de l'organisation. Les changements apportés à l'infrastructure, tels que les mises à jour de la politique de mise à l'échelle automatique, sont testés à l'avance pour répondre aux besoins du trafic.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Des tests sont réalisés sur tous les changements, du code de l'application à l'infrastructure, dans le cadre de votre pratique d'intégration continue. Les résultats des tests sont publiés afin que les développeurs disposent d'une rétroaction rapide. Votre organisation a une norme de test que tous les changements doivent respecter.

Utilisez la puissance de l'IA générative avec Amazon Q Developer pour améliorer la productivité des développeurs et la qualité du code. Amazon Q Developer comprend la génération de suggestions de code (basées sur de grands modèles de langage), la production de tests unitaires (y compris les conditions limites) et l'amélioration de la sécurité du code par la détection et la correction des vulnérabilités de sécurité.

Exemple client

Dans le cadre de son pipeline d'intégration continue, AnyCompany Retail effectue plusieurs types de tests sur tous les artefacts logiciels. L'entreprise pratique le développement axé sur les tests afin que tous les logiciels bénéficient de tests d'unités. Une fois l'artefact créé, ils exécutent end-to-end des tests. Une fois cette première série de tests terminée, elle exécute une analyse de la sécurité des applications statiques qui cherchent des vulnérabilités connues. Les développeurs reçoivent des

messages indiquant que chaque palier de test est validé. Une fois tous les tests terminés, l'artefact logiciel est stocké dans un référentiel d'artefacts.

Étapes d'implémentation

1. Collaborez avec les parties prenantes dans votre organisation pour développer une norme de test pour les artefacts logiciels. Quels tests standards tous les artefacts doivent-ils valider ? Des exigences en matière de conformité ou de réglementation doivent-elles être incluses dans la couverture des tests ? Faut-il réaliser des tests de qualité du code ? Qui doit être informé de la fin des tests ?
    1. L'architecture de référence du pipeline de déploiement AWS contient une liste officielle des types de tests qui peuvent être réalisés sur des artefacts logiciels dans le cadre d'un pipeline d'intégration.
2. Instrumentalisez votre application avec les tests nécessaires en fonction de la norme de test de votre logiciel. Chaque ensemble de tests doit être réalisé en moins de dix minutes. Les tests doivent être exécutés dans le cadre d'un pipeline d'intégration.
    a. Utilisez Amazon Q Developer, un outil d'IA générative qui peut vous aider à créer des cas de tests unitaires (y compris des conditions limites), à générer des fonctions à l'aide de code et de commentaires, et à implémenter des algorithmes connus.
    b. Utilisez Amazon CodeGuru Reviewer pour tester le code de votre application afin de détecter d'éventuels défauts.
    c. AWS CodeBuild vous permet de réaliser des tests sur les artefacts logiciels.
    d. AWS CodePipeline peut orchestrer vos tests logiciels dans un pipeline.

Ressources

Bonnes pratiques associées :

- OPS05-BP01 Utiliser le contrôle de version
- OPS05-BP06 Partager les normes de conception
- OPS05-BP07 Mettre en œuvre des pratiques pour améliorer la qualité du code
- OPS05- BP1 0 Automatisez entièrement l'intégration et le déploiement

Documents connexes :

- Adopter une approche de développement piloté par les tests

- Accélération du cycle de développement de vos logiciels avec Amazon Q
- Amazon Q Developer, désormais disponible pour le grand public, inclut des aperçus de nouvelles fonctionnalités destinées à réinventer l'expérience des développeurs
- L'aide-mémoire ultime pour utiliser Amazon Q Developer dans votre IDE
- Shift-Left Workload, tirant parti de l'IA pour la création de tests
- Centre de développement Amazon Q
- 10 façons de créer des applications plus rapidement avec Amazon CodeWhisperer
- Au-delà de la couverture du code avec Amazon CodeWhisperer
- Bonnes pratiques pour une ingénierie rapide avec Amazon CodeWhisperer
- Pipeline AWS CloudFormation de tests automatisés avec TaskCat et CodePipeline
- Création d'un pipeline end-to-end AWS DevSecOps CI/CD avec des outils et des SCA logiciels SAST open source DAST
- Démarrer avec les applications de test sans serveur
- Le pipeline d'intégration et de livraison continues comme pierre angulaire de la cohérence du code
- Livre blanc Mise en pratique de l'intégration continue/livraison continue sur le livre blanc AWS

Vidéos connexes :

- Implémenter un agent de développement API avec Amazon Q pour le développement de logiciels
- Installation, configuration et utilisation d'Amazon Q Developer avec JetBrains IDEs (mode d'emploi)
- Maîtriser l'art d'Amazon CodeWhisperer - playlist YouTube
- AWS re:Invent 2020 : Infrastructure testable : tests d'intégration sur AWS
- AWS Sommet ANZ 2021 - Mettre en place une stratégie axée sur les tests grâce à un développement piloté par CDK les tests
- Tester votre infrastructure sous forme de code avec AWS CDK

Ressources connexes

- Création d'applications à l'aide de l'IA générative avec Amazon CodeWhisperer
- CodeWhisperer Atelier Amazon
- Architecture de référence du pipeline de déploiement AWS – Application
- AWS Pipeline Kubernetes DevSecOps

- Atelier Politique en tant que code : développement axé sur les tests
- Exécutez des tests unitaires pour une application Node.js à GitHub l'aide de AWS CodeBuild
- Utilisation de Serverspec pour le développement axé sur les tests du code d'infrastructure

Services connexes :

- Amazon Q Developer
- CodeGuru Réviseur Amazon
- AWS CodeBuild
- AWS CodePipeline

OPS05-BP03 Utiliser des systèmes de gestion de la configuration

Utilisez des systèmes de gestion de la configuration pour effectuer et suivre les modifications de la configuration. Ces systèmes réduisent les erreurs causées par les processus manuels et diminuent le niveau d'effort nécessaire au déploiement des modifications.

La gestion de la configuration statique définit des valeurs lors de l'initialisation d'une ressource. Elles doivent rester cohérentes tout au long de la durée de vie de cette ressource. La gestion dynamique de la configuration définit des valeurs à l'initialisation qui peuvent ou sont censées changer pendant la durée de vie d'une ressource. Par exemple, vous pouvez définir une fonctionnalité pour activer les fonctionnalités de votre code par le biais d'une modification de configuration, ou modifier le niveau de détail du journal lors d'un incident.

Les configurations doivent être déployées dans un état connu et cohérent. Vous devez utiliser l'inspection automatisée pour surveiller en permanence les configurations des ressources dans les environnements et les régions. Ces contrôles doivent être définis sous forme de code et de gestion automatisés afin de garantir que les règles sont appliquées de manière cohérente dans tous les environnements. Les modifications apportées aux configurations doivent être mises à jour par le biais de procédures de contrôle des modifications convenues et appliquées de manière cohérente, dans le respect du contrôle des versions. La configuration des applications doit être gérée indépendamment du code de l'application et de l'infrastructure. Cela permet un déploiement cohérent dans plusieurs environnements. Les modifications de configuration n'entraînent pas la reconstruction ou le redéploiement de l'application.

Résultat escompté : vous effectuez la configuration, la validation et le déploiement dans le cadre de votre pipeline d'intégration et de livraison continues (CI/CD). Vous assurez la surveillance pour

vérifier que les configurations sont correctes. Cela permet de minimiser l'impact sur les utilisateurs finaux et les clients.

Anti-modèles courants :

- Vous mettez manuellement à jour la configuration des serveurs Web de votre flotte, et un certain nombre de serveurs ne répondent plus en raison d'erreurs de mise à jour.
- Vous mettez à jour manuellement votre flotte de serveurs d'applications pendant plusieurs heures. L'incohérence de la configuration pendant la modification entraîne des comportements inattendus.
- Quelqu'un a mis à jour vos groupes de sécurité et vos serveurs Web ne sont plus accessibles. Sans savoir ce qui a changé, vous passez beaucoup de temps à enquêter sur la question, ce qui prolonge votre temps de reprise.
- Vous mettez en production une configuration de préproduction via le pipeline CI/CD sans validation. Vous exposez les utilisateurs et les clients à des données et à des services incorrects.

Avantages liés au respect de cette bonne pratique : l'adoption de systèmes de gestion de la configuration réduit le niveau d'effort nécessaire pour effectuer et suivre les changements, ainsi que la fréquence des erreurs causées par les procédures manuelles. Les systèmes de gestion de la configuration fournissent des garanties en matière de gouvernance, de conformité et d'exigences réglementaires.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les systèmes de gestion de la configuration sont utilisés pour suivre et mettre en œuvre les modifications apportées aux configurations des applications et de l'environnement. Ils sont également utilisés pour réduire les erreurs causées par les processus manuels, pour rendre les modifications de configuration reproductibles et vérifiables, et pour réduire le niveau d'effort.

Sur AWS, vous pouvez utiliser AWS Config pour surveiller continuellement vos configurations de ressources AWS à travers les comptes et les régions. Il vous permet de suivre leur historique de configuration, de comprendre comment une modification de la configuration affecterait d'autres ressources et de les auditer par rapport aux configurations attendues ou souhaitées avec AWS Config Rules et AWS Config Conformance Packs.

Pour les configurations dynamiques de vos applications exécutées sur des instances Amazon EC2, AWS Lambda, des conteneurs , des applications mobiles ou des appareils IoT, vous pouvez

les utiliser AWS AppConfig pour les configurer, les valider, les déployer et les surveiller dans vos environnements.

Étapes d'implémentation

1. Identifiez les responsables de la configuration.
   a. Informez les responsables de la configuration de tout besoin en matière de conformité, de gouvernance ou de réglementation.
2. Identifiez les éléments de configuration et les livrables.
   a. Les éléments de configuration sont toutes les configurations d'application et d'environnement concernées par un déploiement au sein de votre pipeline CI/CD.
   b. Les livrables incluent les critères de réussite, la validation et ce qui doit être surveillé.
3. Sélectionnez les outils de gestion de la configuration en fonction des besoins de votre entreprise et de votre pipeline de livraison.
4. Envisagez des déploiements pondérés tels que les déploiements canary pour les modifications de configuration importantes, afin de minimiser l'impact des configurations incorrectes.
5. Intégrez la gestion de votre configuration dans votre pipeline CI/CD.
6. Validez toutes les modifications apportées.

Ressources

Bonnes pratiques associées :

- OPS06-BP01 Planifier les modifications infructueuses
- OPS06-BP02 Déploiements de tests
- OPS06-BP03 Adoption de stratégies de déploiement sûres
- OPS06-BP04 Automatiser les tests et les annulations

Documents connexes :

- AWS Control Tower
- Accélérateur de zone de destination AWS
- AWS Config
- Présentation de AWS Config

- AWS AppConfig
- Présentation de AWS CloudFormation
- Outils pour développeurs AWS
- AWS CodeBuild
- AWS CodePipeline
- AWS CodeDeploy

Vidéos connexes :

- AWS re:Invent 2022 - Proactive governance and compliance for AWS workloads
- AWS re:Invent 2020: Achieve compliance as code using AWS Config
- Gérer et déployer des configurations d'applications avec AWS AppConfig

OPS05-BP04 Utiliser des systèmes de gestion du développement et du déploiement

Utilisez des systèmes de gestion du développement et du déploiement. Ces systèmes réduisent les erreurs causées par les processus manuels et diminuent le niveau d'effort nécessaire au déploiement des modifications.

Dans AWS, vous pouvez créer des pipelines d'intégration continue/de déploiement continu (CI/CD) à l'aide de services tels que les outils pour développeurs AWS (par exemple, AWS CodeBuild, AWS CodePipeline et AWS CodeDeploy).

Résultat escompté : vos systèmes de gestion du développement et du déploiement prennent en charge le système d'intégration et de livraison continues (CI/CD) de votre entreprise, qui fournit des fonctionnalités permettant d'automatiser des déploiements sécurisés avec les configurations appropriées.

Anti-modèles courants :

- Après avoir compilé votre code sur votre système de développement, vous copiez l'exécutable sur vos systèmes de production et il ne démarre pas. Les fichiers journaux locaux indiquent qu'il n'a pas fonctionné en raison de dépendances manquantes.
- Vous créez avec succès votre application avec de nouvelles fonctionnalités dans votre environnement de développement et soumettez le code à l'assurance qualité (QA). L'assurance qualité échoue, car il manque des ressources statiques.

- Vendredi, après de nombreux efforts, vous avez réussi à créer manuellement votre application dans votre environnement de développement, y compris vos nouvelles fonctionnalités codées. Lundi, vous ne pouvez pas répéter les étapes qui vous ont permis de créer votre application avec succès.
- Vous effectuez les tests que vous avez créés pour votre nouvelle version. Ensuite, vous passez la semaine suivante à configurer un environnement de test et à exécuter tous les tests d'intégration existants, suivis des tests de performances. Le nouveau code a un impact inacceptable sur les performances et doit être redéveloppé, puis retesté.

Avantages liés au respect de cette bonne pratique : en fournissant des mécanismes pour gérer les activités de construction et de déploiement, vous réduisez le niveau d'effort nécessaire pour effectuer des tâches répétitives, vous libérez les membres de votre équipe pour qu'ils puissent se concentrer sur leurs tâches créatives de grande valeur et vous limitez l'introduction d'erreurs provenant des procédures manuelles.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les systèmes de gestion du développement et du déploiement sont utilisés pour suivre et mettre en œuvre les modifications, réduire les erreurs causées par les processus manuels et limiter le niveau d'effort requis pour des déploiements sûrs. Automatisez entièrement le pipeline d'intégration et de déploiement à partir du code d'enregistrement et par le biais du développement, des tests, du déploiement et de la validation. Cela permet de réduire les délais, de diminuer les coûts, d'augmenter la fréquence des modifications, de limiter le niveau d'effort et d'accroître la collaboration.

Étapes d’implémentation

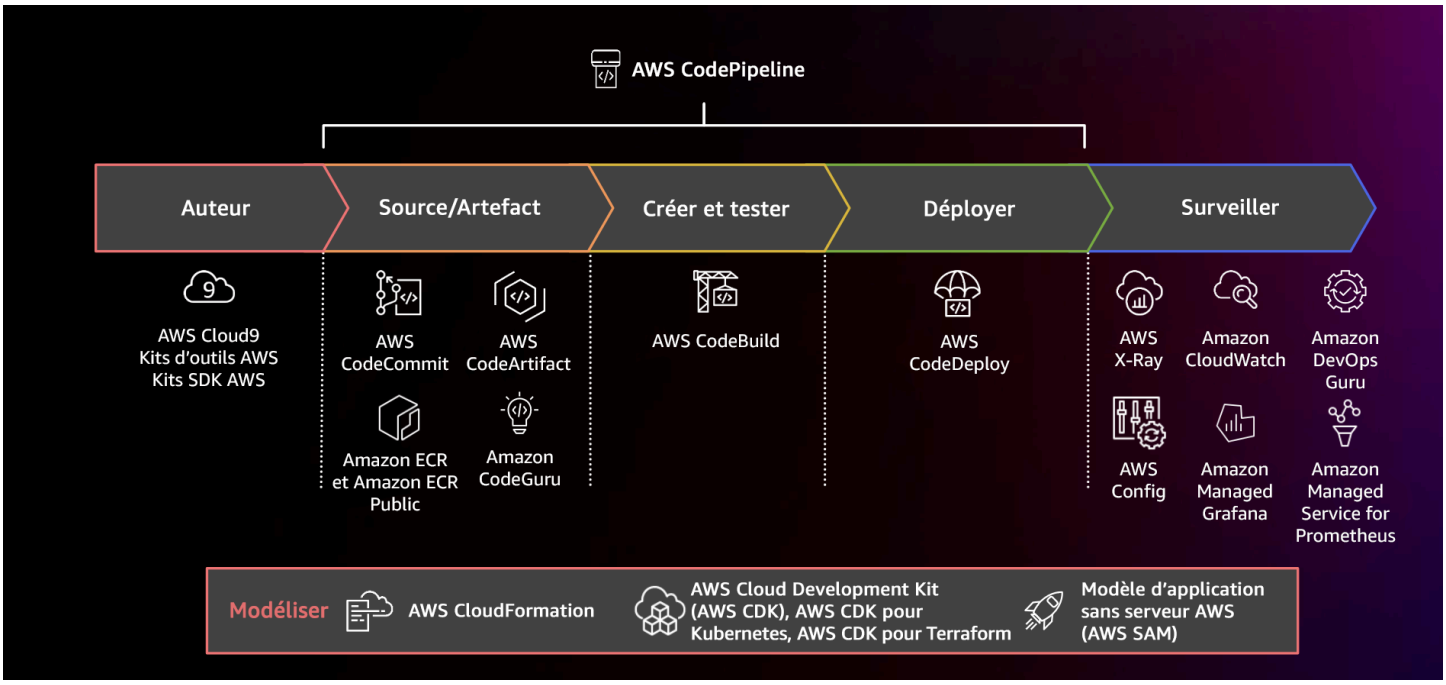


Schéma illustrant un pipeline CI/CD utilisant AWS CodePipeline et des services connexes

1. Utilisez un système de contrôle de version pour stocker et gérer les ressources (tels que des documents, du code source et des fichiers binaires).
2. Utilisez CodeBuild pour compiler votre code source, exécute des tests unitaires et produit des artefacts prêts à être déployés.
3. Utilisez CodeDeploy comme un service de déploiement qui automatise les déploiements d’applications vers des instances Amazon EC2, des instances sur site, des fonctions AWS Lambda sans serveur ou Amazon ECS.
4. Surveillez vos déploiements.

Ressources

Bonnes pratiques associées :

- OPS06-BP04 Automatiser les tests et les annulations

Documents connexes :

- Outils pour développeurs AWS

- Présentation de AWS CodeBuild
- AWS CodeBuild
- Présentation de AWS CodeDeploy

Vidéos connexes :

- AWS re:Invent 2022 - AWS Well-Architected best practices for DevOps on AWS

OPS05-BP05 Effectuer la gestion des correctifs

Procédez à la gestion des correctifs afin de profiter des fonctionnalités, de résoudre les problèmes et de rester conforme à la gouvernance. Automatisez la gestion des correctifs pour réduire les erreurs causées par les processus manuels, pour permettre la mise à l’échelle et pour réduire le niveau d’efforts nécessaire aux correctifs.

La gestion des correctifs et des vulnérabilités fait partie de vos activités de gestion des bénéfices et des risques. Il est préférable d’avoir des infrastructures immuables et de déployer des charges de travail dans des états de bon fonctionnement connus et vérifiés. Lorsque cela n’est pas viable, l’application de correctifs est la seule solution.

Amazon EC2 Image Builder fournit des pipelines pour mettre à jour les images des machines. Dans le cadre de la gestion des correctifs, envisagez qu'Amazon Machine Images (AMIs) utilise un pipeline d'AMIimages ou des images de conteneur avec un pipeline d'images Docker, tout en AWS Lambda fournissant des modèles pour des environnements d'exécution personnalisés et des bibliothèques supplémentaires pour supprimer les vulnérabilités.

Vous devez gérer les mises à jour des images Amazon Machine Images pour Linux ou Windows Server à l'aide d'Amazon EC2 Image Builder. Vous pouvez utiliser Amazon Elastic Container Registry (AmazonECR) avec votre pipeline existant pour gérer les ECS images Amazon et gérer les EKS images Amazon. Lambda comprend les fonctionnalités de gestion des versions.

L’application de correctifs ne doit pas être effectuée sur les systèmes de production sans avoir effectué un test préalable dans un environnement sûr. Les correctifs ne doivent être appliqués que s’ils favorisent la réalisation d’un résultat opérationnel ou métier. Activé AWS, vous pouvez utiliser AWS Systems Manager Patch Manager pour automatiser le processus d'application des correctifs aux systèmes gérés et planifier l'activité à l'aide de Systems Manager Maintenance Windows.

Résultat escompté : vos images AMI et celles du conteneur sont corrigées up-to-date et prêtes à être lancées. Vous pouvez suivre l’état de toutes les images déployées et déterminer la conformité des

correctifs. Vous êtes en mesure de rendre compte de l'état actuel et de disposer d'un processus pour répondre à vos besoins en matière de conformité.

Anti-modèles courants :

- On vous demande d'appliquer tous les nouveaux correctifs de sécurité dans un délai de deux heures, ce qui entraîne de multiples pannes dues à l'incompatibilité de l'application avec les correctifs.
- Une bibliothèque non corrigée entraîne des conséquences imprévues, car des parties inconnues y utilisent des failles pour accéder à votre charge de travail.
- Vous corrigez automatiquement les environnements de développement sans en informer les développeurs. Vous recevez plusieurs réclamations des développeurs indiquant que leur environnement ne fonctionne plus correctement.
- Vous n'avez pas appliqué de correctif au off-the-shelf logiciel commercial sur une instance persistante. Lorsque vous rencontrez un problème avec le logiciel et que vous contactez le fournisseur, celui-ci vous informe que la version n'est pas prise en charge et que vous devez effectuer appliquer un correctif à un niveau spécifique pour recevoir de l'aide.
- Un correctif récemment publié pour le logiciel de chiffrement que vous avez utilisé présente des améliorations significatives de performances. Votre système non corrigé présente des problèmes de performances qui persistent suite à l'absence de correctifs.
- Vous êtes averti d'une vulnérabilité de type « jour zéro » nécessitant une correction d'urgence et vous devez corriger manuellement tous vos environnements.

Avantages liés au respect de cette bonne pratique : en établissant un processus de gestion des correctifs, y compris vos critères d'application des correctifs et la méthodologie de distribution dans vos environnements, vous pouvez mettre à l'échelle les niveaux de correctifs et créer des rapports sur ces niveaux. Cela fournit des garanties concernant les correctifs de sécurité et assure une visibilité claire sur l'état des correctifs connus en cours de mise en place. Cela encourage aussi l'adoption des fonctions et fonctionnalités désirées, l'élimination rapide des problèmes et le respect durable de la gouvernance. Mettez en œuvre des systèmes de gestion des correctifs et d'automatisation pour réduire le niveau d'effort nécessaire au déploiement des correctifs et limiter les erreurs causées par les processus manuels.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d’implémentation

Appliquez des correctifs aux systèmes pour corriger les problèmes, obtenir des fonctionnalités souhaitées et rester conforme à la politique de gouvernance et aux exigences d’assistance du fournisseur. Dans les systèmes immuables, déployez avec l’ensemble de correctifs approprié pour obtenir le résultat souhaité. Automatisez le mécanisme de gestion des correctifs afin de réduire le temps écoulé avant l’application des correctifs, d’éviter les erreurs causées par les processus manuels et de limiter le niveau d’efforts nécessaire pour appliquer les correctifs.

Étapes d’implémentation

Pour Amazon EC2 Image Builder :

1. À l'aide EC2 d'Amazon Image Builder, spécifiez les détails du pipeline :
    a. Créez un pipeline d’images et nommez-le.
    b. Définissez le calendrier et le fuseau horaire du pipeline.
    c. Configurez toutes les dépendances.
2. Choisissez une recette :
    a. Sélectionnez une recette existante ou créez-en une.
    b. Sélectionnez le type d’image.
    c. Donnez un nom et une version à votre recette.
    d. Sélectionnez votre image de base.
    e. Ajoutez des composants de build et incluez-les dans le registre cible.
3. Facultatif : définissez la configuration de votre infrastructure.
4. Facultatif : définissez les paramètres de configuration.
5. Révisez les paramètres.
6. Gérez régulièrement l’hygiène des recettes.

Pour le gestionnaire de correctifs de Systems Manager :

1. Créez un référentiel de correctifs.
2. Sélectionnez une méthode d'opérations d'application de correctifs.
3. Activez le reporting et l’analyse de conformité.

Ressources

Bonnes pratiques associées :

- OPS06-BP04 Automatiser les tests et les annulations

Documents connexes :

- Qu'est-ce qu'Amazon EC2 Image Builder
- Créez un pipeline d'images à l'aide d'Amazon EC2 Image Builder
- Création d’un pipeline d’images de conteneurs
- AWS Systems Manager Patch Manager
- Utilisation du gestionnaire de correctifs
- Utilisation des rapports de conformité des correctifs
- AWS Outils pour développeurs

Vidéos connexes :

- CI/CD pour applications sans serveur sur AWS
- Design with Ops in Mind

Exemples connexes :

- Ateliers Well-Architected : inventaire et gestion des correctifs
- AWS Systems Manager Tutoriels de Patch Manager

OPS05-BP06 Partager les normes de conception

Partagez les bonnes pratiques entre les équipes pour sensibiliser et maximiser les bénéfices des efforts de développement. Documentez-les et mettez-les à jour au fur et à mesure de l’évolution de votre architecture. Si votre organisation applique des normes partagées, il est essentiel de prévoir des mécanismes permettant de demander des ajouts, des modifications et des exceptions aux normes. Sans cette possibilité, les normes deviennent une contrainte à l’innovation.

Résultat escompté : les normes de conception sont partagées par toutes les équipes de vos organisations. Ils sont documentés et conservés au up-to-date fur et à mesure de l'évolution des meilleures pratiques.

Anti-modèles courants :

- Deux équipes de développement ont chacune créé un service d'authentification des utilisateurs. Vos utilisateurs doivent conserver un ensemble distinct d'informations d'identification pour chaque partie du système à laquelle ils veulent accéder.
- Chaque équipe gère sa propre infrastructure. Une nouvelle exigence de conformité impose une modification de votre infrastructure et chaque équipe la met en œuvre de manière différente.

Avantages liés au respect de cette bonne pratique : l'utilisation de normes communes favorise l'adoption de bonnes pratiques et maximise les avantages des efforts de développement. La documentation et la mise à jour des normes de conception permettent à votre organisation de up-to-date respecter les meilleures pratiques et les exigences de sécurité et de conformité.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Partagez les bonnes pratiques existantes, les normes de conception, les listes de contrôle, les procédures d'exploitation, les conseils et les exigences de gouvernance entre les équipes. Prévoyez des procédures pour demander des modifications, des ajouts et des exceptions aux normes de conception afin de favoriser l'amélioration et l'innovation. Assurez-vous que les équipes sont au courant du contenu publié. Disposer d'un mécanisme permettant de maintenir les normes de conception au up-to-date fur et à mesure que de nouvelles pratiques exemplaires apparaissent.

Exemple client

AnyCompany Retail dispose d'une équipe d'architecture interfonctionnelle qui crée des modèles d'architecture logicielle. Cette équipe construit l'architecture en y intégrant les aspects de conformité et de gouvernance. Les équipes qui adoptent ces normes communes bénéficient des avantages de la conformité et de la gouvernance intégrées. Elles peuvent rapidement s'appuyer sur la norme de conception. L'équipe d'architecture se réunit tous les trimestres pour évaluer les modèles d'architecture et les mettre à jour si nécessaire.

Étapes d'implémentation

1. Identifiez une équipe interfonctionnelle qui sera chargée de développer et de mettre à jour les normes de conception. Cette équipe travaillera avec les parties prenantes de votre organisation pour élaborer des normes de conception, des procédures d'exploitation, des listes de contrôle, des

conseils et des exigences de gouvernance. Documentez les normes de conception et partagez-les au sein de votre organisation.

   a. AWS Service Catalog permet de créer des portefeuilles représentant les normes de conception en utilisant l'infrastructure en tant que code. Vous pouvez partager des portefeuilles entre plusieurs comptes.

2. Mettre en place un mécanisme pour maintenir les normes de conception au up-to-date fur et à mesure que de nouvelles pratiques exemplaires sont identifiées.
3. Si les normes de conception sont appliquées de manière centralisée, il faut prévoir un processus pour demander des modifications, des mises à jour et des exemptions.

Niveau d'effort du plan d'implémentation : moyen. L'élaboration d'un processus de création et de partage des normes de conception peut nécessiter une coordination et une coopération avec les parties prenantes de votre organisation.

Ressources

Bonnes pratiques associées :

- OPS01-BP03 Évaluer les exigences de gouvernance – Les exigences de gouvernance influencent les normes de conception.
- OPS01-BP04 Évaluer les exigences de conformité – La conformité est un élément essentiel de la création de normes de conception.
- OPS07-BP02 Assurer un examen cohérent de l'état de préparation opérationnelle – Les listes de contrôle de la disponibilité opérationnelle constituent un mécanisme de mise en œuvre des normes de conception lors de la conception de votre charge de travail.
- OPS11-BP01 Disposer d'un processus d'amélioration continue – La mise à jour des normes de conception fait partie de l'amélioration continue.
- OPS11-BP04 Effectuer la gestion des connaissances – Dans le cadre de votre pratique de gestion des connaissances, documentez et partagez les normes de conception.

Documents connexes :

- Automatisez AWS Backup nous avec AWS Service Catalog
- AWS Service Catalog Compte amélioré en usine
- Comment Expedia Group a créé une offre de base de données en tant que service (DBaaS) en utilisant AWS Service Catalog

- Assurer la visibilité sur l'utilisation des modèles d'architecture cloud
- Simplifiez le partage de vos AWS Service Catalog portefeuilles dans une AWS Organizations configuration

Vidéos connexes :

- AWS Service Catalog — Mise en route
- AWS re:Invent 2020 : Gérez vos AWS Service Catalog portefeuilles comme un expert

Exemples connexes :

- AWS Service Catalog Architecture de référence
- AWS Service Catalog Atelier

Services connexes :

- AWS Service Catalog

OPS05-BP07 Mise en œuvre de pratiques visant à améliorer la qualité du code

Mettez en place des pratiques pour améliorer la qualité du code et limiter les failles. Parmi les exemples, citons le développement piloté par les tests, les révisions de code, l'adoption de normes et la programmation en binôme. Incorporez ces pratiques dans votre processus d'intégration et de livraison continues.

Résultat escompté : votre organisation utilise des bonnes pratiques comme les révisions de code ou la programmation en binôme pour améliorer la qualité du code. Les développeurs et les opérateurs adoptent les bonnes pratiques en matière de qualité du code dans le cadre du cycle de vie du développement logiciel.

Anti-modèles courants :

- Vous livrez du code à la branche principale de votre application sans effectuer de révision du code. La modification est automatiquement déployée en production et provoque une panne.
- Une nouvelle application est développée sans aucun test d'unité, de bout en bout ou d'intégration. Il n'y a aucun moyen de tester l'application avant son déploiement.

- Vos équipes procèdent à des modifications manuelles en production pour corriger les défauts. Les modifications ne sont pas soumises à des tests ou à des révisions de code et ne sont pas saisies ou enregistrées dans le cadre des processus d'intégration et de livraison continues.

Avantages liés au respect de cette bonne pratique : en adoptant des pratiques visant à améliorer la qualité du code, vous contribuez à minimiser les problèmes introduits dans la production. La qualité du code facilite l'utilisation des bonnes pratiques telles que la programmation en binôme, les révisions de code et la mise en œuvre d'outils de productivité basés sur l'IA.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Mettez en œuvre des pratiques visant à améliorer la qualité du code afin de minimiser les défauts avant leur déploiement. Utilisez des pratiques telles que le développement piloté par les tests, les révisions de code et la programmation en binôme pour améliorer la qualité de votre développement.

Utilisez la puissance de l'IA générative avec Amazon Q Developer pour améliorer la productivité des développeurs et la qualité du code. Amazon Q Developer comprend la génération de suggestions de code (basées sur de grands modèles de langage), la production de tests unitaires (y compris les conditions limites) et l'amélioration de la sécurité du code par la détection et la correction des vulnérabilités de sécurité.

Exemple client

AnyCompany Retail adopte plusieurs pratiques pour améliorer la qualité du code. La société a adopté le développement piloté par les tests comme norme d'écriture des applications. Pour certaines nouvelles fonctionnalités, elle demande aux développeurs de programmer en binôme pendant un sprint. Chaque demande d'extraction est soumise à une révision du code par un développeur principal avant d'être intégrée et déployée.

Étapes d'implémentation

1. Adoptez des pratiques de qualité du code telles que le développement piloté par les tests, les révisions de code et la programmation en binôme dans votre processus d'intégration et de livraison continues. Utilisez ces techniques pour améliorer la qualité des logiciels.
   a. Utilisez Amazon Q Developer, un outil d'IA générative qui peut vous aider à créer des cas de tests unitaires (y compris des conditions limites), à générer des fonctions à l'aide de code et de commentaires, à implémenter des algorithmes connus, à détecter les violations des

politiques de sécurité et les vulnérabilités dans votre code, à détecter les secrets, à scanner l'infrastructure en tant que code (IaC), à documenter le code et à apprendre plus rapidement des bibliothèques de code tierces.

b. Amazon CodeGuru Reviewer peut fournir des recommandations de programmation pour le code Java et Python en utilisant le machine learning.

Niveau d'effort du plan d'implémentation : moyen. Il existe de nombreuses façons de mettre en œuvre cette bonne pratique, mais il peut être difficile de la faire adopter par l'organisation.

Ressources

Bonnes pratiques associées :

- OPS05-BP02 Test et validation des modifications
- OPS05-BP06 Partage des normes de conception

Documents connexes :

- Adopter une approche de développement piloté par les tests
- Accélération du cycle de développement de vos logiciels avec Amazon Q
- Amazon Q Developer, désormais disponible pour le grand public, inclut des aperçus de nouvelles fonctionnalités destinées à réinventer l'expérience des développeurs
- L'aide-mémoire ultime pour utiliser Amazon Q Developer dans votre environnement de développement intégré
- Shift-Left Workload, tirant parti de l'IA pour la création de tests
- Centre de développement Amazon Q
- 10 façons de créer des applications plus rapidement avec Amazon CodeWhisperer
- Au-delà de la couverture du code avec Amazon CodeWhisperer
- Bonnes pratiques pour une ingénierie de requête avec Amazon CodeWhisperer
- Guide du logiciel Agile
- Mon pipeline CI/CD est mon capitaine de versions
- Automatisez les révisions de code avec Amazon CodeGuru Reviewer
- Adopter une approche de développement piloté par les tests
- Comment DevFactory crée de meilleures applications avec Amazon CodeGuru

- Programmation en binôme
- RENGA Inc. automatise les révisions de code avec Amazon CodeGuru
- L'art du développement agile : le développement piloté par les tests
- Pourquoi les révisions de code sont importantes (et font gagner du temps !)

Vidéos connexes :

- Implement an API with Amazon Q Developer Agent for Software Development
- Installing, Configuring, & Using Amazon Q Developer with JetBrains IDEs (How-to)
- Mastering the art of Amazon CodeWhisperer - YouTube playlist
- AWS re:Invent 2020: Continuous improvement of code quality with Amazon CodeGuru
- AWS Summit ANZ 2021 - Driving a test-first strategy with CDK and test driven development

Services connexes :

- Amazon Q Developer
- Amazon CodeGuru Reviewer
- Amazon CodeGuru Profiler

OPS05-BP08 Utilisation de plusieurs environnements

Utilisez plusieurs environnements pour expérimenter, développer et tester votre charge de travail. Utilisez des niveaux de contrôle croissants lorsque les environnements approchent de la production pour vous assurer que votre charge de travail fonctionnera correctement une fois déployée.

Résultat escompté : vous disposez de plusieurs environnements qui répondent à vos besoins en matière de conformité et de gouvernance. Vous testez et promouvez le code dans les différents environnements jusqu'à la production.

1. Pour ce faire, votre organisation établit une zone de destination, qui assure la gouvernance, les contrôles, l'automatisation des comptes, la mise en réseau, la sécurité et l'observabilité opérationnelle. Gérez ces fonctionnalités de zone de destination en utilisant plusieurs environnements. Un exemple courant est celui d'une organisation d'environnement de test (sandbox) chargée de développer et de tester des modifications apportées à une zone de destination basée sur AWS Control Tower, qui inclut AWS IAM Identity Center et des politiques

telles que les politiques de contrôle des services (SCP). Tous ces éléments peuvent avoir un impact significatif sur l'accès aux Comptes AWS et leur fonctionnement dans la zone de destination.

2. En plus de ces services, vos équipes étendent les capacités des zones de destination avec des solutions publiées par AWS et les partenaires AWS ou des solutions personnalisées développées au sein de votre organisation. Les exemples de solutions publiées par AWS incluent Configurations personnalisées d'AWS Control Tower (CfCT) et AWS Control Tower Account Factory pour Terraform (AFT).
3. Votre organisation applique les mêmes principes en matière de test, de promotion du code et de modification des politiques pour la zone de destination via les environnements sur le chemin de la production. Cette stratégie offre un environnement de zone de destination stable et sécurisé à vos équipes chargées des applications et des charges de travail.

Anti-modèles courants :

- Vous effectuez un développement dans un environnement de développement partagé et un autre développeur remplace vos modifications de code.
- Les contrôles de sécurité restrictifs sur votre environnement de développement partagé vous empêchent d'expérimenter de nouveaux services et fonctionnalités.
- Vous effectuez des tests de charge sur vos systèmes de production et provoquez une panne pour vos utilisateurs.
- Une erreur critique entraînant une perte de données s'est produite en production. Dans votre environnement de production, vous essayez de recréer les conditions qui ont conduit à la perte de données afin de pouvoir identifier comment elle s'est produite et empêcher qu'elle ne se reproduise. Pour éviter toute perte de données supplémentaire pendant les tests, vous devez rendre l'application indisponible aux utilisateurs.
- Vous explorez un service multilocataire et n'êtes pas en mesure de répondre à la demande d'un client pour un environnement dédié.
- Il se peut que vous ne réalisiez pas toujours des tests, mais lorsque vous le faites, vous procédez dans votre environnement de production.
- Vous pensez que la simplicité d'un environnement unique l'emporte sur la portée de l'impact des modifications au sein de l'environnement.
- Vous améliorez une fonctionnalité clé de la zone de destination, mais cette modification réduit la capacité de votre équipe à vendre des comptes pour de nouveaux projets ou pour vos charges de travail existantes.

- Vous appliquez de nouveaux contrôles à vos Comptes AWS, mais la modification a un impact sur la capacité de votre équipe chargée des charges de travail à déployer des modifications dans leurs Comptes AWS.

Avantages liés au respect de cette bonne pratique : lorsque vous déployez plusieurs environnements, vous pouvez prendre en charge simultanément plusieurs environnements de développement, de test et de production sans créer de conflits entre les développeurs ou les communautés d'utilisateurs. Pour les fonctionnalités complexes telles que les zones de destination, cela réduit considérablement le risque de modifications, simplifie le processus d'amélioration et réduit le risque de mises à jour critiques de l'environnement. Les organisations qui utilisent des zones de destination tirent naturellement parti des comptes multiples dans leur environnement AWS, avec les configurations de structure de compte, de gouvernance, de réseau et de sécurité. Au fil du temps, à mesure que votre entreprise grandit, la zone de destination peut évoluer pour sécuriser et organiser vos charges de travail et vos ressources.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Utilisez plusieurs environnements et fournissez aux développeurs des environnements de test (sandbox) avec des contrôles réduits au minimum pour faciliter l'expérimentation. Fournissez des environnements de développement individuels pour permettre le travail en parallèle, ce qui augmente l'agilité du développement. Mettez en œuvre davantage de contrôles rigoureux dans les environnements proches de la production pour offrir aux développeurs la liberté d'innover. Utilisez l'infrastructure en tant que code et les systèmes de gestion de la configuration pour déployer des environnements configurés de manière cohérente par rapport aux contrôles de production pour veiller au bon fonctionnement des systèmes lorsqu'ils sont déployés. Lorsque les environnements ne sont pas en cours d'utilisation, désactivez-les pour éviter les coûts associés à des ressources inutilisées (par exemple, les systèmes de développement en soirée et les week-ends). Déployez des environnements équivalents à la production lors des tests de charge pour accroître les résultats valides.

Les équipes chargées de l'ingénierie des plateformes, de la mise en réseau et des opérations de sécurité gèrent souvent les capacités au niveau de l'organisation avec des exigences distinctes. La séparation des comptes ne suffit pas à fournir et à maintenir des environnements distincts pour l'expérimentation, le développement et les tests. Dans ce type de cas, créez des instances distinctes d'AWS Organizations.

Ressources

Documents connexes :

- Instance Scheduler sur AWS
- Présentation de AWS CloudFormation
- Organisation de votre environnement AWS à l'aide de plusieurs comptes - Organisations multiples - Test des modifications apportées à votre environnement AWS global
- Guide AWS Control Tower

OPS05-BP09 Procéder à des modifications fréquentes, mineures et réversibles

Les modifications fréquentes, légères et réversibles limitent la portée et l'impact d'une modification. Lorsqu'elles sont utilisées conjointement avec des systèmes de gestion des modifications, des systèmes de gestion de configuration et des systèmes de construction et de livraison, les modifications fréquentes, mineures et réversibles limitent la portée et l'impact d'une modification. Cela se traduit par une résolution plus efficace des problèmes et par des corrections plus rapides avec la possibilité d'annuler les modifications effectuées.

Anti-modèles courants :

- Vous déployez une nouvelle version de votre application tous les trimestres avec une fenêtre de modification qui signifie qu'un service principal est désactivé.
- Vous modifiez fréquemment le schéma de votre base de données sans suivre les modifications apportées à vos systèmes de gestion.
- Vous effectuez des mises à jour manuelles sur place, en remplaçant les installations et les configurations existantes, sans aucun plan de restauration clair.

Avantages de l'établissement de cette bonne pratique : les efforts de développement sont plus rapides en déployant fréquemment de petites modifications. Lorsque les changements sont minimes, il est beaucoup plus facile d'identifier s'ils ont des conséquences inattendues et ils sont plus faciles à annuler. Lorsque les changements sont réversibles, les risques de mise en œuvre d'une modification sont minimes, car la récupération est simplifiée. Le processus de modification présente un risque réduit et l'impact de l'échec d'une modification est réduit.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d’implémentation

Ayez recours à des modifications fréquentes, légères et réversibles pour limiter leur portée et leur impact. Cela facilite la résolution des problèmes, contribue à accélérer les corrections et offre la possibilité d’annuler une modification. Cela augmente également la vitesse à laquelle vous pouvez apporter de la valeur à votre entreprise.

Ressources

Bonnes pratiques associées :

- OPS05-BP03 Utiliser des systèmes de gestion de la configuration
- OPS05-BP04 Utiliser des systèmes de gestion du développement et du déploiement
- OPS06-BP04 Automatiser les tests et les annulations

Documents connexes :

- Implémentation de microservices sur AWS
- Microservices – Observabilité

OPS05- BP1 0 Automatisez entièrement l'intégration et le déploiement

Automatisez la création, le déploiement et le test de la charge de travail. Cela permet de réduire les erreurs découlant des processus manuels, ainsi que les efforts nécessaires au déploiement des modifications.

Appliquez des métadonnées à l’aide des balises de ressource et de AWS Resource Groups en suivant une stratégie de balisage cohérente pour permettre l’identification de vos ressources. Balisez vos ressources pour l’organisation, la comptabilité analytique, les contrôles d’accès et le ciblage de l’exécution des activités d’opérations automatisées.

Résultat escompté : les développeurs utilisent des outils pour fournir du code et le promouvoir jusqu’à la production. Les développeurs n'ont pas besoin de se connecter au AWS Management Console pour fournir des mises à jour. Il existe une piste d’audit complète des modifications et de la configuration, répondant aux besoins de gouvernance et de conformité. Les processus sont reproductibles et standardisés entre les équipes. Les développeurs sont libres de se concentrer sur le développement et les envois de code, ce qui augmente la productivité.

Anti-modèles courants :

- Vendredi, vous avez fini de créer le code de votre branche de fonctionnalité. Lundi, après avoir exécuté vos scripts de test de la qualité du code et chacun de vos scripts de tests unitaires, vous vérifiez votre code pour la prochaine version prévue.
- Vous êtes chargé de coder un correctif pour un problème critique affectant un grand nombre de clients en production. Après avoir testé le correctif, vous validez votre code et envoyez un e-mail à l'équipe de gestion des modifications pour demander l'autorisation de le déployer en production.
- En tant que développeur, vous vous connectez AWS Management Console au pour créer un nouvel environnement de développement à l'aide de méthodes et de systèmes non standard.

Avantages liés au respect de cette bonne pratique : en mettant en œuvre des systèmes automatisés de gestion de la création et du déploiement, vous réduisez les erreurs causées par les processus manuels et diminuez l'effort de déploiement des changements, ce qui permet aux membres de votre équipe de se concentrer sur la création de valeur ajoutée. Vous accélérez la vitesse de livraison au fur et à mesure que vous progressez jusqu'à la production.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Utilisez des systèmes de gestion du développement et du déploiement afin de suivre et de mettre en œuvre des modifications, de réduire les erreurs causées par les processus manuels et de réduire le niveau d'efforts. Automatisez entièrement le pipeline d'intégration et de déploiement à partir du code d'enregistrement et par le biais du développement, des tests, du déploiement et de la validation. Cela permet de raccourcir les délais, d'augmenter la fréquence des modifications, de réduire le niveau d'effort, d'accélérer la mise sur le marché, d'augmenter la productivité et de renforcer la sécurité de votre code jusqu'à la production.

Ressources

Bonnes pratiques associées :

- OPS05-BP03 Utiliser des systèmes de gestion de la configuration
- OPS05-BP04 Utiliser des systèmes de gestion du développement et du déploiement

Documents connexes :

- Qu'est-ce que c'est AWS CodeBuild ?

- Qu'est-ce que c'est AWS CodeDeploy ?

Vidéos connexes :

- AWS re \ :Invent 2022 - Les meilleures pratiques de AWS Well-Architected pour DevOps AWS

OPS 6. Comment réduire les risques liés au déploiement ?

Adoptez des approches qui fournissent un retour d'information rapide sur la qualité et permettent une reprise rapide à la suite de changements qui n'offrent pas les résultats escomptés. L'utilisation de ces pratiques diminue l'impact des problèmes découlant du déploiement des modifications.

Bonnes pratiques

- OPS06-BP01 Planifier les modifications infructueuses
- OPS06-BP02 Déploiements de tests
- OPS06-BP03 Adoption de stratégies de déploiement sûres
- OPS06-BP04 Automatiser les tests et les annulations

OPS06-BP01 Planifier les modifications infructueuses

Prévoyez de revenir à un état correct connu ou de remédier à la situation dans l'environnement de production si le déploiement entraîne un résultat indésirable. L'existence d'une politique visant à établir un tel plan aide toutes les équipes à développer des stratégies de récupération en cas d'échec des modifications. Parmi les exemples de stratégies, citons les étapes de déploiement et de restauration, les stratégies de changement, les indicateurs de fonctionnalité, l'isolation du trafic et le déplacement du trafic. Une seule version peut inclure plusieurs modifications de composants connexes. La stratégie doit permettre de résister ou de se remettre d'une défaillance de tout changement de composant.

Résultat escompté : vous avez préparé un plan de reprise détaillé pour votre modification en cas d'échec. En outre, vous avez réduit la taille de votre version afin de minimiser l'impact potentiel sur d'autres composants de la charge de travail. Vous avez ainsi réduit l'impact sur l'entreprise en diminuant le temps d'arrêt potentiel causé par une modification ratée et en augmentant la flexibilité et l'efficacité des temps de récupération.

Anti-modèles courants :

- Vous avez effectué un déploiement et votre application est devenue instable, mais il semble qu'il y ait des utilisateurs actifs sur le système. Vous devez décider entre annuler la modification et avoir un impact sur les utilisateurs actifs et attendre pour annuler la modification en sachant que les utilisateurs peuvent être impactés de toute façon.
- Après avoir modifié la routine, vos nouveaux environnements sont accessibles, mais l'un de vos sous-réseaux est devenu inaccessible. Vous devez décider de tout annuler ou d'essayer de réparer le sous-réseau inaccessible. Pendant cette période de détermination, le sous-réseau reste inaccessible.
- Vos systèmes ne sont pas conçus de manière à pouvoir être mis à jour avec de plus petites versions. Par conséquent, il est difficile d'annuler ces modifications en bloc en cas d'échec du déploiement.
- Vous n'utilisez pas l'infrastructure en tant que code (IaC) et vous avez effectué des mises à jour manuelles de votre infrastructure, ce qui a entraîné une configuration indésirable. Vous n'êtes pas en mesure de suivre et d'annuler efficacement les modifications manuelles.
- Parce que vous n'avez pas mesuré l'augmentation de la fréquence de vos déploiements, votre équipe n'est pas incitée à réduire la taille de ses changements et à améliorer ses plans de restauration pour chaque modification, ce qui entraîne une augmentation des risques et des taux d'échec.
- Vous ne mesurez pas la durée totale d'une panne causée par des modifications infructueuses. Votre équipe n'est pas en mesure d'établir des priorités et d'améliorer l'efficacité de son processus de déploiement et de son plan de reprise.

Avantages de la mise en place de cette meilleure pratique : le fait de disposer d'un plan de reprise après des modifications infructueuses permet de minimiser le temps moyen de restauration (MTTR) et de réduire l'impact sur votre entreprise.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Une stratégie et une pratique cohérentes et documentées, adoptées par les équipes de publication des versions, permettent à une organisation de planifier ce qui doit se passer en cas d'échec des modifications. La politique devrait permettre la correction à l'avance dans des circonstances spécifiques. Dans les deux cas, un plan de correction à l'avance ou de restauration doit être bien documenté et testé avant d'être déployé dans la production réelle, afin de réduire au minimum la durée nécessaire pour restaurer une modification.

Étapes d'implémentation

1. Documentez les stratégies qui exigent des équipes qu'elles disposent de plans efficaces pour restaurer les modifications dans un délai donné.
   a. Les stratégies doivent préciser les cas où une situation de correction à l'avance est autorisée.
   b. Exigez qu'un plan de restauration documenté soit accessible à toutes les personnes concernées.
   c. Précisez les conditions de restauration (par exemple, lorsqu'il s'avère que des modifications non autorisées ont été déployées).
2. Analysez le niveau d'impact de toutes les modifications liées à chaque composante d'une charge de travail.
   a. Autorisez les modifications répétitives à être normalisées, modélisées et préautorisées si elles suivent un flux de travail cohérent qui applique les politiques de modification.
   b. Réduisez l'impact potentiel de toute modification en en réduisant la taille, de sorte que la reprise prenne moins de temps et ait moins d'impact sur l'entreprise.
   c. Veillez à ce que les procédures de restauration ramènent le code à l'état correct connu afin d'éviter les incidents dans la mesure du possible.
3. Intégrez des outils et des flux de travail pour appliquer vos politiques de manière programmée.
4. Faites en sorte que les données relatives aux modifications soient visibles pour les autres propriétaires de charges de travail afin d'améliorer la rapidité du diagnostic en cas de modification défaillante impossible à annuler.
   a. Mesurez le degré de réussite de cette pratique à l'aide de données sur les modifications visibles et identifiez les améliorations itératives.
5. Utilisez des outils de surveillance pour vérifier le succès ou l'échec d'un déploiement afin d'accélérer la prise de décision concernant la restauration.
6. Mesurez la durée de l'interruption lors d'un changement infructueux afin d'améliorer continuellement vos plans de reprise.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS06-BP04 Automatiser les tests et les annulations

Documents connexes :

- AWS Builders Library | Garantir la sécurité des annulations lors des déploiements
- AWS Livre blanc | Gestion du changement dans le cloud

Vidéos connexes :

- re:Invent 2019 | Amazon’s Approach to high-availability deployment

OPS06-BP02 Déploiements de tests

Testez les procédures de mise à disposition en préproduction en utilisant la même configuration de déploiement, les mêmes contrôles de sécurité, les mêmes étapes et les mêmes procédures qu’en production. Confirmez que toutes les étapes du déploiement se sont déroulées comme prévu, par exemple en inspectant les fichiers, les configurations et les services. Testez ensuite toutes les modifications à l’aide de tests fonctionnels, d’intégration et de charge, ainsi que de contrôles tels que les surveillances de l’état. En effectuant ces tests, vous pouvez identifier rapidement les problèmes de déploiement et avoir la possibilité de les planifier et de les atténuer avant la mise en production.

Vous pouvez créer des environnements parallèles temporaires pour tester chaque modification. Automatisez le déploiement des environnements de test à l’aide de l’infrastructure en tant que code (IaC) afin de réduire la quantité de travail nécessaire et d’assurer la stabilité, la cohérence et une livraison plus rapide des fonctionnalités.

Résultat escompté : votre organisation adopte une culture de développement piloté par les tests qui inclut des déploiements de tests. Cela permet de veiller à ce que les équipes se concentrent sur la création de valeur pour l’entreprise plutôt que sur la gestion des versions. Les équipes sont impliquées dès l’identification des risques de déploiement afin de déterminer les mesures d’atténuation appropriées.

Anti-modèles courants :

- Pendant les mises en production, les déploiements non testés entraînent des problèmes fréquents qui nécessitent un dépannage et une remontée.
- Votre version contient une infrastructure en tant que code (IaC) qui met à jour les ressources existantes. Vous n’êtes pas certain que l’IaC s’exécute correctement ou qu’elle a un impact sur les ressources.

- Vous déployez une nouvelle fonctionnalité dans votre application. Elle ne fonctionne pas comme prévu et il n'y a aucune visibilité jusqu'à ce qu'elle soit signalée par les utilisateurs concernés.
- Vous mettez à jour vos certificats. Vous installez accidentellement les certificats sur les mauvais composants, ce qui passe inaperçu et a un impact sur les visiteurs du site Web parce qu'il est impossible d'établir une connexion sécurisée avec le site Web.

Avantages liés au respect de cette bonne pratique : des tests approfondis en préproduction des procédures de déploiement et des modifications qu'elles introduisent minimisent l'impact potentiel sur la production causé par les étapes de déploiement. Cela permet d'accroître la confiance lors de la mise en production et de minimiser l'assistance opérationnelle sans ralentir la vitesse des changements apportés.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Il est tout aussi important de tester votre processus de déploiement que les modifications qui en découlent. Pour ce faire, vous pouvez tester vos étapes de déploiement dans un environnement de préproduction qui reflète le plus fidèlement possible l'environnement de production. Les problèmes courants, tels que les étapes de déploiement incomplètes ou incorrectes, ou les mauvaises configurations, peuvent être détectés avant la mise en production. De plus, vous pouvez tester vos étapes de reprise.

Exemple client

Dans le cadre de son pipeline d'intégration continue et de livraison continue (CI/CD), AnyCompany Retail exécute les étapes définies nécessaires pour publier des mises à jour d'infrastructure et de logiciels pour ses clients dans un environnement de type production. Le pipeline comprend des contrôles préalables pour détecter les altérations (détection des changements apportés aux ressources en dehors de votre IaC) dans les ressources avant le déploiement, ainsi que pour valider les actions que l'IaC entreprend lors de son lancement. Il valide les étapes du déploiement, en vérifiant par exemple que certains fichiers et configurations sont en place, que les services sont en cours d'exécution et qu'ils répondent correctement aux surveillances de l'état sur l'hôte local avant de s'enregistrer à nouveau auprès de l'équilibreur de charge. En outre, toutes les modifications font l'objet d'un certain nombre de tests automatisés, tels que des tests fonctionnels, de sécurité, de régression, d'intégration et de charge.

Étapes d’implémentation

1. Effectuez des contrôles avant l’installation pour reproduire l’environnement de préproduction en production.
   a. Utilisez la détection de dérive pour détecter lorsque les ressources ont été modifiées en dehors de AWS CloudFormation.
   b. Utilisez des ensembles de modifications pour vérifier que l'intention d'une mise à jour de la pile correspond aux actions entreprises lorsque l'ensemble de modifications est initié. AWS CloudFormation
2. Cela déclenche une étape d’approbation manuelle dans AWS CodePipeline pour autoriser le déploiement dans l’environnement de préproduction.
3. Utilisez des configurations de déploiement telles que AWS CodeDeploy AppSpecdes fichiers pour définir les étapes de déploiement et de validation.
4. Le cas échéant, AWS CodeDeploy intégrez-le à d'autres AWS services ou AWS CodeDeploy intégrez-le aux produits et services partenaires.
5. Surveillez les déploiements à l'aide CloudWatch d' AWS CloudTrail Amazon et des notifications d'SNSévénements Amazon.
6. Réalisez des tests automatisés après déploiement, y compris des tests fonctionnels, de sécurité, de régression, d’intégration et de charge.
7. Résolution des problèmes de déploiement
8. La validation réussie des étapes précédentes devrait lancer un mécanisme d’autorisation manuel pour autoriser le déploiement en production.

Niveau d’effort du plan d’implémentation : élevé

Ressources

Bonnes pratiques associées :

- OPS05-BP02 Tester et valider les modifications

Documents connexes :

- AWS Bibliothèque pour les constructeurs | Automatisation des déploiements sûrs et sans intervention directe | Déploiements de test
- AWS Livre blanc | Pratiquer l'intégration et la livraison continues sur AWS

- L'histoire d'Apollo, le moteur de déploiement d'Amazon
- Comment tester et déboguer AWS CodeDeploy localement avant d'expédier votre code
- Intégrer les tests de connectivité réseau au déploiement de l'infrastructure

Vidéos connexes :

- re:Invent 2020 | Testing software and systems at Amazon

Exemples connexes :

- Tutoriel | Déploiement et ECS service Amazon avec un test de validation

OPS06-BP03 Adoption de stratégies de déploiement sûres

Les déploiements de production sécurisés contrôlent le flux des modifications bénéfiques dans le but de minimiser l'impact perçu de ces modifications sur les clients. Les contrôles de sécurité fournissent des mécanismes d'inspection permettant de valider les résultats souhaités et de limiter l'étendue de l'impact des défaillances introduites par les modifications ou des échecs de déploiement. Les déploiements sûrs peuvent inclure des stratégies telles que les indicateurs de fonctionnalités, les déploiements sur un seul hôte, les déploiements continus (versions canary), les déploiements immuables, la division du trafic et les déploiements bleu/vert.

Résultat escompté :votre organisation utilise un système d'intégration continue et de livraison continue (CI/CD) qui permet d'automatiser des déploiements sûrs. Les équipes sont tenues d'utiliser des stratégies de déploiement sûres et appropriées.

Anti-modèles courants :

- Vous déployez une modification infructueuse dans l'ensemble de l'environnement de production en une seule fois. Par conséquent, tous les clients sont touchés simultanément.
- Une défaillance introduite lors d'un déploiement simultané dans tous les systèmes nécessite un lancement d'urgence. La correction pour tous les clients prend plusieurs jours.
- La gestion des versions de production nécessite la planification et la participation de plusieurs équipes. Cela limite votre capacité à mettre fréquemment à jour les fonctionnalités pour vos clients.
- Vous effectuez un déploiement mutable en modifiant vos systèmes existants. Après avoir découvert que la modification n'a pas abouti, vous devez modifier à nouveau les systèmes pour restaurer l'ancienne version, ce qui prolonge votre délai de récupération.

Avantages liés au respect de cette bonne pratique : les déploiements automatisés permettent de concilier la rapidité des déploiements et la cohérence des modifications apportées aux clients. Limiter l'impact permet d'éviter des échecs de déploiement coûteux et de maximiser la capacité des équipes à répondre efficacement aux défaillances.

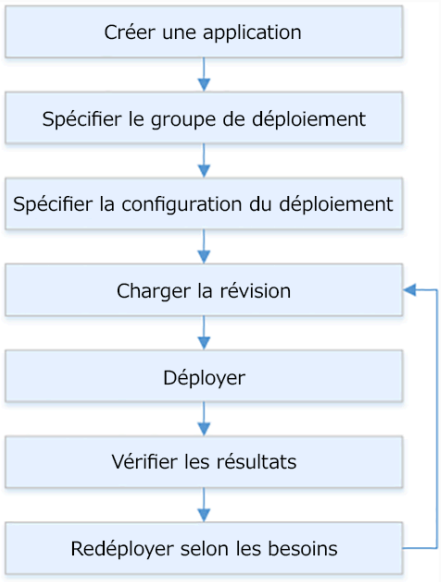
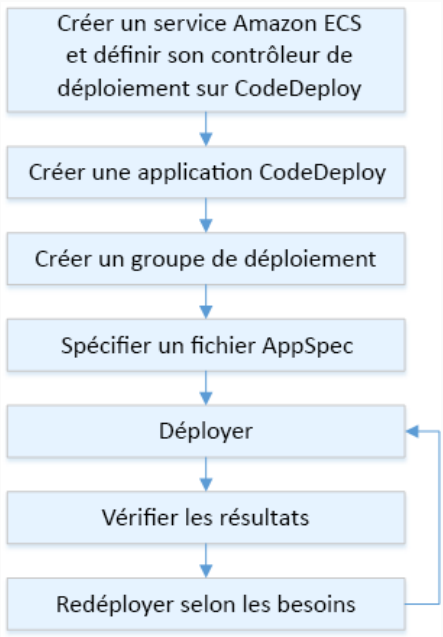
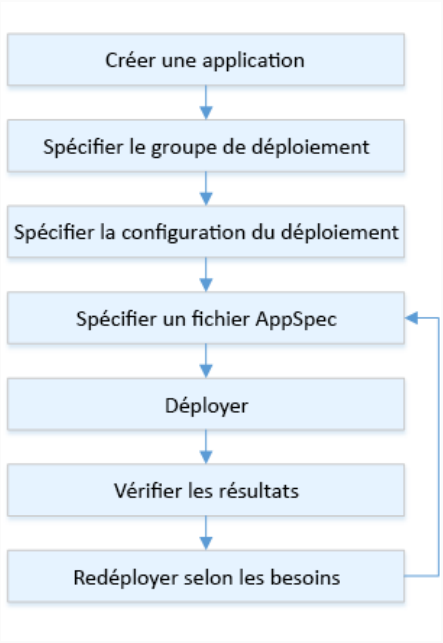
Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les défaillances de la livraison en continu peuvent entraîner une réduction de la disponibilité des services et de mauvaises expériences pour les clients. Pour maximiser le taux de réussite des déploiements, mettez en œuvre des contrôles de sécurité dans le processus de lancement de bout en bout afin de minimiser les erreurs de déploiement ; l'objectif étant de parvenir à zéro échec de déploiement.

Exemple client

AnyCompany Retail a pour mission de réaliser des déploiements avec un temps d'arrêt minimal ou nul, ce qui signifie qu'il n'y a pas d'impact perceptible pour ses utilisateurs pendant les déploiements. Pour ce faire, l'entreprise a établi des modèles de déploiement (voir le diagramme de flux de travail suivant), tels que les déploiements continus et les déploiements bleu/vert. Toutes les équipes adoptent un ou plusieurs de ces modèles dans leur pipeline CI/CD.

Flux de travail CodeDeploy pour Amazon EC2	Flux de travail CodeDeploy pour Amazon ECS	Flux de travail CodeDeploy pour Lambda
		

Étapes d'implémentation

1. Utilisez un flux de travail d'approbation pour lancer la séquence des étapes de déploiement de la production lors de la promotion en production.
2. Utilisez un système de déploiement automatisé comme AWS CodeDeploy. Les options de déploiement d'AWS CodeDeploy comprennent les déploiements sur place pour EC2/sur site et les déploiements bleu/vert pour EC2/sur site, AWS Lambda et Amazon ECS (voir le diagramme de flux de travail précédent).
   a. Le cas échéant, intégrez AWS CodeDeploy à d'autres services AWS ou intégrez AWS CodeDeploy aux produits et services partenaires.
3. Utilisez des déploiements bleu/vert pour les bases de données telles qu'Amazon Aurora et Amazon RDS.
4. Surveillez les déploiements à l'aide d'Amazon CloudWatch, AWS CloudTrail et des notifications d'événements Amazon Simple Notification Service (Amazon SNS).
5. Effectuez des tests automatisés post-déploiement, y compris des tests fonctionnels, de sécurité, de régression, d'intégration et tout test de charge.
6. Résolvez les problèmes de déploiement.

Niveau d’effort du plan d’implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS05-BP02 Tester et valider les modifications
- OPS05-BP09 Procéder à des modifications fréquentes, mineures et réversibles
- OPS05- BP1 0 Automatisez entièrement l'intégration et le déploiement

Documents connexes :

- AWS Builders’ Library | Automatisation de déploiements sécurisés sans intervention | Déploiements en production
- AWS Builders Library | Mon pipeline CI/CD est mon capitaine de versions | Versions de production automatiques et sécurisées
- AWSLivre blanc | Mise en pratique de l’intégration continue et de la livraison continue sur AWS | Méthodes de déploiement
- Guide de l’utilisateur AWS CodeDeploy
- Utilisation des configurations de déploiement dans AWS CodeDeploy
- Configuration d’un déploiement de la version canary API Gateway
- Types de déploiement Amazon ECS
- Déploiements bleu/vert entièrement gérés dans Amazon Aurora et Amazon RDS
- Déploiements bleu/vert avec AWS Elastic Beanstalk

Vidéos connexes :

- re:Invent 2020 | Hands-off: Automating continuous delivery pipelines at Amazon
- re:Invent 2019 | Amazon’s Approach to high-availability deployment

Exemples connexes :

- Essai d’un exemple de déploiement bleu/vert dans AWS CodeDeploy
- Atelier | Création de pipelines CI/CD pour les déploiements canary Lambda à l’aide d’AWS CDK

- Atelier | Création de votre premier pipeline DevOps bleu/vert avec Amazon ECS
- Atelier | Création de votre premier pipeline DevOps bleu/vert avec Amazon EKS
- Atelier | EKS GitOps avec ArgoCD
- Atelier | Atelier CI/CD sur AWS
- Implémentation de pipelines CI/CD entre comptes avec AWS SAM pour les fonctions Lambda basées sur des conteneurs

OPS06-BP04 Automatiser les tests et les annulations

Pour accroître la rapidité, la fiabilité et la confiance de votre processus de déploiement, mettez en place une stratégie de tests automatisés et de restauration dans les environnements de préproduction et de production. Automatisez les tests lors du déploiement en production afin de simuler les interactions entre l'homme et le système et de vérifier les modifications déployées. Automatisez la restauration pour revenir rapidement à un état antérieur sain connu. La restauration doit être déclenchée automatiquement dans des conditions prédéfinies, par exemple lorsque le résultat souhaité de la modification n'est pas atteint ou lorsque le test automatisé échoue. L'automatisation de ces deux activités améliore le taux de réussite de vos déploiements, minimise le temps de reprise et réduit l'impact potentiel sur l'entreprise.

Résultat escompté : vos tests automatisés et vos stratégies de restauration sont intégrés dans votre pipeline d'intégration continue et de livraison continue (CI/CD). Votre surveillance est en mesure de valider vos critères de réussite et de déclencher une restauration automatique en cas d'échec. Cela permet de minimiser l'impact sur les utilisateurs finaux et les clients. Par exemple, lorsque tous les résultats des tests ont été satisfaits, vous transférez votre code dans l'environnement de production où des tests de régression automatisés sont lancés, en utilisant les mêmes cas de test. Si les résultats des tests de régression ne correspondent pas aux attentes, une restauration automatisée est lancée dans le flux de travail du pipeline.

Anti-modèles courants :

- Vos systèmes ne sont pas conçus de manière à pouvoir être mis à jour avec de plus petites versions. Par conséquent, il est difficile d'annuler ces modifications en bloc en cas d'échec du déploiement.
- Votre processus de déploiement consiste en une série d'étapes manuelles. Après avoir apporté des modifications à votre charge de travail, vous commencez les tests de post-déploiement. Après les tests, vous vous rendez compte que votre charge de travail est inopérante et que les clients

sont déconnectés. Vous commencez les opérations pour restaurer la version précédente. Toutes ces étapes manuelles retardent la reprise globale du système et ont un impact prolongé sur vos clients.

- Vous avez passé du temps à développer des cas de tests automatisés pour des fonctionnalités qui ne sont pas fréquemment utilisées dans votre application, minimisant ainsi le retour sur investissement de votre capacité de tests automatisés.
- Votre version est composée de mises à jour d'applications, d'infrastructures, de correctifs et de configurations qui sont indépendantes les unes des autres. Cependant, vous disposez d'un seul pipeline CI/CD qui fournit toutes les modifications en une seule fois. La défaillance d'un composant vous oblige à annuler toutes les modifications, ce qui rend votre restauration complexe et inefficace.
- Votre équipe termine le travail de codage au cours du premier sprint et commence le travail du deuxième sprint, mais votre plan ne prévoyait pas de tests avant le troisième sprint. En conséquence, les tests automatisés ont révélé des défauts du premier sprint qui ont dû être résolus avant que les tests des produits livrables du deuxième sprint puissent commencer et la version entière est retardée, ce qui dévalorise vos tests automatisés.
- Vos tests de régression automatisés pour la version de production sont terminés, mais vous ne surveillez pas l'état de la charge de travail. Comme vous ne savez pas si le service a redémarré ou non, vous ne savez pas si la restauration est nécessaire ou si elle a déjà eu lieu.

Avantages liés au respect de cette bonne pratique : l'automatisation des tests accroît la transparence de votre processus de test et votre capacité à couvrir davantage de fonctionnalités dans un laps de temps plus court. En testant et en validant les modifications en production, vous êtes en mesure d'identifier immédiatement les problèmes. L'amélioration de la cohérence avec les outils de test automatisés permet une meilleure détection des défauts. En restaurant automatiquement la version précédente, vous réduisez l'impact sur vos clients. La restauration automatisée inspire finalement plus de confiance dans vos capacités de déploiement en réduisant l'impact sur l'entreprise. Dans l'ensemble, ces capacités réduisent time-to-delivery tout en garantissant la qualité.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Automatisez le test des environnements déployés pour confirmer les résultats souhaités plus rapidement. Automatisez la restauration du dernier état connu de bonne qualité lorsque les résultats prédéfinis ne sont pas atteints, afin de minimiser les temps de récupération et de réduire les erreurs

causées par les processus manuels. Intégrez des outils de test au flux de travail de votre pipeline afin de tester de manière cohérente et de minimiser les saisies manuelles. Privilégiez l'automatisation des cas de test, tels que ceux qui atténuent les risques les plus importants et qui doivent être testés fréquemment à chaque modification. En outre, vous pouvez automatiser la restauration en fonction de conditions spécifiques prédéfinies dans votre plan de test.

Étapes d'implémentation

1. Établissez un cycle de test pour votre cycle de développement qui définit chaque étape du processus de test, de la planification des exigences au développement des cas de test, en passant par la configuration des outils, les tests automatisés et la clôture des cas de test.
   a. Créez une approche de test spécifique à la charge de travail à partir de votre stratégie de test globale.
   b. Envisagez, le cas échéant, une stratégie de tests continus tout au long du cycle de développement.
2. Choisissez des outils automatisés pour les tests et la restauration en fonction des besoins de votre entreprise et des investissements dans le pipeline.
3. Décidez des cas de test que vous souhaitez automatiser et de ceux qui doivent être exécutés manuellement. Ceux-ci peuvent être définis en fonction de la priorité de la valeur commerciale de la fonctionnalité testée. Alignez tous les membres de l'équipe sur ce plan et vérifiez leur responsabilité en ce qui concerne l'exécution des tests manuels.
   a. Appliquez les capacités de test automatisé à des cas de test spécifiques qui se prêtent à l'automatisation, tels que les cas répétables ou fréquemment exécutés, ceux qui nécessitent des tâches répétitives ou ceux qui sont requis dans plusieurs configurations.
   b. Définissez les scripts d'automatisation des tests ainsi que les critères de réussite dans l'outil d'automatisation afin que l'automatisation continue du flux de travail puisse être lancée lorsque des cas spécifiques échouent.
   c. Définissez des critères d'échec spécifiques pour la restauration automatisée.
4. Donnez la priorité à l'automatisation des tests afin d'obtenir des résultats cohérents grâce à un développement approfondi des cas de test où la complexité et l'interaction humaine présentent un risque d'échec plus élevé.
5. Intégrez vos outils de tests automatisés et de restauration dans votre pipeline CI/CD.
   a. Définissez des critères de réussite clairs pour vos modifications.
   b. Surveillez et observez pour détecter ces critères et annuler automatiquement les modifications lorsque des critères de restauration spécifiques sont remplis.

6. Procédez à différents types de tests de production automatisés, tels que :
   a. des tests A/B pour afficher les résultats par rapport à la version actuelle entre deux groupes d'utilisateurs ;
   b. des tests Canary qui vous permettent de déployer votre modification auprès d'un sous-ensemble d'utilisateurs avant de la diffuser à tous ;
   c. des tests d'indicateur de fonctions qui permettent d'activer et de désactiver une seule fonctionnalité de la nouvelle version depuis l'extérieur de l'application, de sorte que chaque nouvelle fonctionnalité puisse être validée une à la fois ;
   d. des tests de régression pour vérifier les nouvelles fonctionnalités avec les composants interdépendants existants.
7. Contrôlez les aspects opérationnels de l'application, les transactions et les interactions avec d'autres applications et composants. Élaborez des rapports pour illustrer le degré de réussite des modifications en fonction de la charge de travail, afin de pouvoir identifier les parties de l'automatisation et du flux de travail qui peuvent être encore optimisées.
   a. Élaborez des rapports sur les résultats des tests qui vous aideront à prendre des décisions rapides sur le fait d'invoquer ou non les procédures de restauration.
   b. Mettez en œuvre une stratégie permettant une restauration automatisée sur la base de conditions d'échec prédéfinies résultant d'une ou de plusieurs de vos méthodes de test.
8. Développez vos cas de test automatisés pour permettre leur réutilisation dans le cadre de futures modifications répétées.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS06-BP01 Planifier les modifications infructueuses
- OPS06-BP02 Déploiements de tests

Documents connexes :

- AWS Builders Library | Garantir la sécurité des annulations lors des déploiements
- Redéployez et annulez un déploiement avec AWS CodeDeploy
- 8 bonnes pratiques pour automatiser vos déploiements avec AWS CloudFormation

Exemples connexes :

- Test de l'interface utilisateur sans serveur à l'aide de Selenium, AWS LambdaAWS Fargate, et AWS des outils de développement

Vidéos connexes :

- re:Invent 2020 | Hands-off: Automating continuous delivery pipelines at Amazon
- re:Invent 2019 | Amazon’s Approach to high-availability deployment

OPS 7. Comment savoir si vous êtes prêt à assurer une charge de travail ?

Évaluez la disponibilité opérationnelle de votre charge de travail, des processus et des procédures, ainsi que le personnel pour comprendre les risques opérationnels liés à votre charge de travail.

Bonnes pratiques

- OPS07-BP01 Assurer la capacité du personnel
- OPS07-BP02 Assurer un examen cohérent de l'état de préparation opérationnelle
- OPS07-BP03 Utiliser des runbooks pour exécuter des procédures
- OPS07-BP04 Utiliser des playbooks pour étudier les problèmes
- OPS07-BP05 Prendre des décisions éclairées pour déployer des systèmes et apporter des modifications
- OPS07-BP06 Créer des plans de support pour les charges de travail de production

OPS07-BP01 Assurer la capacité du personnel

Prévoyez un mécanisme pour confirmer que vous disposez du nombre approprié de membres du personnel formés pour supporter la charge de travail. Ils doivent être formés à la plateforme et aux services qui constituent votre charge de travail. Donnez-leur les connaissances nécessaires pour exploiter la charge de travail. Vous devez former un nombre suffisant de membres du personnel pour assurer le fonctionnement normal de la charge de travail et résoudre les incidents qui surviennent. Prévoyez suffisamment de personnel pour pouvoir effectuer une rotation pendant les astreintes et les vacances afin d’éviter l’épuisement professionnel.

Résultat escompté :

- Le personnel formé est en nombre suffisant pour faire face à la charge de travail lorsque celle-ci est disponible.
- Vous assurez la formation de votre personnel sur les logiciels et services qui constituent votre charge de travail.

Anti-modèles courants :

- Déploiement d'une charge de travail sans que les membres de l'équipe soient qualifiés pour gérer la plateforme et les services utilisés.
- Ne pas disposer d'un personnel suffisant pour assurer les rotations d'astreinte ou les congés du personnel.

Avantages liés au respect de cette bonne pratique :

- Le fait de disposer de membres d'équipe compétents vous permet de prendre efficacement en charge votre charge de travail.
- Avec un nombre suffisant de membres de l'équipe, vous pouvez prendre en charge la charge de travail et les rotations d'astreinte tout en diminuant le risque d'épuisement professionnel.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Confirmez qu'il y a suffisamment de personnel formé pour soutenir la charge de travail. Vérifiez que vous avez suffisamment de membres de l'équipe pour couvrir les activités opérationnelles normales, y compris les rotations d'astreinte.

Exemple client

AnyCompany Le commerce de détail veille à ce que les équipes chargées de supporter la charge de travail soient correctement dotées en personnel et formées. Elles disposent de suffisamment d'ingénieurs pour assurer une rotation d'astreinte. Le personnel reçoit une formation sur le logiciel et la plateforme sur lesquels repose la charge de travail et est encouragé à obtenir des certifications. Il y a suffisamment de membres du personnel pour que les gens puissent prendre des congés tout en prenant en charge la charge de travail et la rotation des astreintes.

Étapes d'implémentation

1. Affectez un nombre suffisant d'employés à l'exploitation et au soutien de votre charge de travail, y compris aux fonctions d'astreinte.
2. Formez votre personnel aux logiciels et aux plateformes qui composent votre charge de travail.
   a. AWS Training and Certification dispose d'une bibliothèque de cours sur AWS. Le service propose des cours gratuits et payants, en ligne et en personne.
   b. AWS organise des événements et des webinaires où vous pouvez apprendre auprès d' AWS experts.
3. Évaluez régulièrement la taille et les compétences de l'équipe en fonction de l'évolution des conditions d'exploitation et de la charge de travail. Adaptez la taille et les compétences de l'équipe aux besoins opérationnels.

Niveau d'effort du plan d'implémentation : élevé L'embauche et la formation d'une équipe pour soutenir une charge de travail peuvent demander des efforts considérables, mais présentent des avantages importants à long terme.

Ressources

Bonnes pratiques associées :

- OPS11-BP04 Effectuer la gestion des connaissances – Les membres de l'équipe doivent disposer des informations nécessaires au fonctionnement et au soutien de la charge de travail. La gestion des connaissances est la clé pour y parvenir.

Documents connexes :

- AWS Événements et webinaires
- AWS Formation et certification

OPS07-BP02 Assurer un examen cohérent de l'état de préparation opérationnelle

Utilisez les évaluations de préparation opérationnelle (ORRs) pour vérifier que vous pouvez gérer votre charge de travail. ORRest un mécanisme développé par Amazon pour valider que les équipes peuvent gérer leurs charges de travail en toute sécurité. An ORR est un processus de révision et d'inspection utilisant une liste de contrôle des exigences. An ORR est une expérience en libre-service que les équipes utilisent pour certifier leurs charges de travail. ORRsinclure les meilleures pratiques tirées des leçons apprises au cours de nos années passées à créer des logiciels.

Une ORR liste de contrôle est composée de recommandations architecturales, de processus opérationnels, de gestion des événements et de qualité des versions. Notre processus de correction des erreurs (CoE) est l'un des principaux moteurs de ces éléments. Votre propre analyse post-incident doit être le moteur de votre propre ORR évolution. An ne ORR consiste pas seulement à suivre les meilleures pratiques, mais aussi à prévenir la récurrence d'événements que vous avez vus auparavant. Enfin, les exigences de sécurité, de gouvernance et de conformité peuvent également être incluses dans unORR.

Exécuté ORRs avant qu'une charge de travail ne soit mise en disponibilité générale, puis tout au long du cycle de développement du logiciel. L'exécution de la ORR commande avant le lancement augmente votre capacité à gérer la charge de travail en toute sécurité. Réexécutez régulièrement votre charge ORR de travail pour détecter tout écart par rapport aux meilleures pratiques. Vous pouvez avoir des ORR listes de contrôle pour le lancement de nouveaux services et ORRs pour les révisions périodiques. Cela vous permet de vous tenir au courant des nouvelles bonnes pratiques et d'intégrer les leçons tirées de l'analyse après incident. Au fur et à mesure que votre utilisation du cloud évolue, vous pouvez intégrer des ORR exigences par défaut à votre architecture.

Résultat escompté : Vous disposez d'une ORR liste des meilleures pratiques pour votre organisation. ORRssont effectués avant le lancement des charges de travail. ORRssont exécutés périodiquement tout au long du cycle de vie de la charge de travail.

Anti-modèles courants :

- Vous lancez une charge de travail sans savoir si vous pouvez l'utiliser.
- Les exigences en matière de gouvernance et de sécurité ne sont pas incluses dans la certification d'une charge de travail pour le lancement.
- Les charges de travail ne sont pas réévaluées périodiquement.
- Les charges de travail sont lancées sans procédures requises en place.
- Vous voyez la répétition de la même cause première de défaillances dans plusieurs charges de travail.

Avantages liés au respect de cette bonne pratique :

- Vos charges de travail comprennent les bonnes pratiques en matière d'architecture, de processus et de gestion.
- Les leçons apprises sont intégrées à votre ORR processus.
- Les procédures requises sont en place lors du lancement des charges de travail.

- ORRssont exécutés tout au long du cycle de vie logiciel de vos charges de travail.

Niveau de risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

An, ORR c'est deux choses : un processus et une liste de contrôle. Votre ORR processus doit être adopté par votre organisation et soutenu par un sponsor exécutif. Elle ORRs doit au minimum être réalisée avant qu'une charge de travail ne soit mise en disponibilité générale. Exécutez-le ORR tout au long du cycle de développement du logiciel pour le tenir à jour en fonction des meilleures pratiques ou des nouvelles exigences. La ORR liste de contrôle doit inclure les éléments de configuration, les exigences de sécurité et de gouvernance, ainsi que les meilleures pratiques de votre organisation. Au fil du temps, vous pouvez utiliser des services tels que AWS ConfigAWS Security Hub, et AWS Control Tower Guardrails, pour élaborer les meilleures pratiques de A à Z ORR afin de détecter automatiquement les meilleures pratiques.

Exemple client

Après plusieurs incidents de production, AnyCompany Retail a décidé de mettre en œuvre un ORR processus. L’entreprise a élaboré une liste de contrôle composée de bonnes pratiques, d’exigences en matière de gouvernance et de conformité et d’enseignements tirés des pannes. Les nouvelles charges de travail sont exécutées ORRs avant leur lancement. Chaque charge de travail est exécutée chaque année ORR avec un sous-ensemble de meilleures pratiques afin d'intégrer les nouvelles meilleures pratiques et exigences qui sont ajoutées à la ORR liste de contrôle. Au fil du temps, AnyCompany Retail AWS Configavait l'habitude de détecter certaines des meilleures pratiques, accélérant ainsi le ORR processus.

Étapes d’implémentation

Pour en savoir plusORRs, consultez le livre blanc intitulé Operational Readiness Reviews (ORR). Il fournit des informations détaillées sur l'historique du ORR processus, comment créer votre propre ORR cabinet et comment élaborer votre ORR liste de contrôle. Les étapes suivantes sont une version abrégée de ce document. Pour mieux comprendre ce que ORRs sont les vôtres et comment les créer, nous vous recommandons de lire ce livre blanc.

1. Réunissez les parties prenantes clés, notamment les représentants de la sécurité, des opérations et du développement.
2. Demandez à chaque partie prenante de fournir au moins une exigence. Pour la première itération, essayez de limiter le nombre d’éléments à trente ou moins.

- Annexe B : ORR Des exemples de questions tirés du livre blanc intitulé Operational Readiness Reviews (ORR) contiennent des exemples de questions que vous pouvez utiliser pour commencer.

3. Regroupez vos exigences dans une feuille de calcul.
   - Vous pouvez utiliser des objectifs personnalisés AWS Well-Architected Toolpour développer vos objectifs ORR et les partager entre vos comptes et votre AWS organisation.
4. Identifiez une charge de travail ORR à exécuter. Il est recommandé d'utiliser une charge de travail avant le lancement ou une charge de travail interne.
5. Parcourez la ORR liste de contrôle et prenez note de toutes les découvertes faites. Les découvertes peuvent ne pas être acceptables si une mesure d'atténuation est en place. Pour toute découverte qui ne comporte pas de mesures d'atténuation, ajoutez ces dernières à votre liste de tâches en attente et implémentez-les avant le lancement.
6. Continuez à ajouter les meilleures pratiques et les exigences à votre ORR liste de contrôle au fil du temps.

Support les clients bénéficiant du Support aux entreprises peuvent demander l'atelier de révision du niveau de préparation opérationnelle auprès de leur responsable de compte technique. L'atelier est une session interactive de travail à rebours pour développer votre propre ORR liste de contrôle.

Niveau d'effort du plan d'implémentation : élevé L'adoption d'une ORR pratique dans votre organisation nécessite le parrainage de la direction et l'adhésion des parties prenantes. Créez et mettez à jour la liste de contrôle à l'aide des commentaires de l'ensemble de votre organisation.

Ressources

Bonnes pratiques associées :

- OPS01-BP03 Évaluer les exigences de gouvernance— Les exigences en matière de gouvernance conviennent naturellement à une ORR liste de contrôle.
- OPS01-BP04 Évaluer les exigences de conformité— Les exigences de conformité sont parfois incluses dans une ORR liste de contrôle. Parfois, il s'agit d'un processus distinct.
- OPS03-BP07 Ressources appropriées pour les équipes— La capacité d'équipe est un bon candidat pour un ORR besoin.
- OPS06-BP01 Planifier les modifications infructueuses – Un plan de restauration ou de retour en arrière doit être établi avant le lancement de votre charge de travail.

- OPS07-BP01 Assurer la capacité du personnel – Pour gérer une charge de travail, vous devez disposer du personnel requis.
- SEC01-BP03 Identifier et valider les objectifs de contrôle — Les objectifs de contrôle de sécurité constituent d'excellentes ORR exigences.
- REL13-BP01 Définissez des objectifs de restauration en cas d'indisponibilité et de perte de données — Les plans de reprise après sinistre sont une bonne ORR exigence.
- COST02-BP01 Élaborez des politiques basées sur les exigences de votre organisation — Les politiques de gestion des coûts sont bonnes à inclure dans votre ORR liste de contrôle.

Documents connexes :

- AWS Control Tower - Rambardes intégrées AWS Control Tower
- AWS Well-Architected Tool - Verres personnalisés
- Operational Readiness Review Template par Adrian Hornsby
- Livre blanc sur les examens de l'état de préparation opérationnelle (ORR)

Vidéos connexes :

- AWS Support s You | Élaborer un examen de l'état de préparation opérationnelle efficace (ORR)

Exemples connexes :

- Exemple d'examen de l'état de préparation opérationnelle (ORR) Lens

Services connexes :

- AWS Config
- AWS Control Tower
- AWS Security Hub
- AWS Well-Architected Tool

OPS07-BP03 Utiliser des runbooks pour exécuter des procédures

Un runbook est un processus documenté pour atteindre un résultat spécifique. Les runbooks consistent en une série d’étapes permettant à la personne qui les suit d’obtenir des résultats

concrets. L’utilisation des runbooks dans les opérations remonte aux débuts de l’aviation. Dans les opérations de cloud, nous utilisons des runbooks pour réduire les risques et obtenir les résultats souhaités. Dans sa forme la plus simple, un runbook est une liste de contrôle pour exécuter une tâche.

Les runbooks représentent une part essentielle du fonctionnement de votre charge de travail. De l’intégration d’un nouveau membre de l’équipe au déploiement d’une version majeure, les runbooks sont des processus codifiés qui fournissent des résultats cohérents quelle que soit la personne qui les utilise. Les runbooks doivent être publiés dans un emplacement central et mis à jour à mesure que le processus évolue, car la mise à jour des runbooks est un composant essentiel du processus de gestion des changements. Ils doivent également inclure des conseils sur la gestion des erreurs, les outils, les autorisations, les exceptions et les remontées en cas de problème.

À mesure que votre entreprise évolue, commencez à automatiser les runbooks. Prenez tout d’abord les runbooks courts et fréquemment utilisés. Utilisez des langages de scripts pour automatiser les étapes ou les rendre plus faciles. À mesure que vous automatiserez les premiers runbooks, vous consacrerez du temps à l’automatisation de runbooks plus complexes. Au fil du temps, la plupart de vos runbooks seront automatisés d’une certaine façon.

Résultat souhaité : Votre équipe dispose d'un ensemble de step-by-step guides pour effectuer les tâches liées à la charge de travail. Les runbooks contiennent le résultat souhaité, les outils et autorisations nécessaires, ainsi que les instructions pour gérer les erreurs. Ils sont stockés dans un emplacement central (système de contrôle des versions) et mis à jour fréquemment. Par exemple, vos runbooks fournissent à vos équipes des fonctionnalités leur permettant de surveiller, de communiquer et de répondre aux AWS Health événements concernant les comptes critiques lors d'alarmes liées aux applications, de problèmes opérationnels ou d'événements planifiés du cycle de vie.

Anti-modèles courants :

- Utilisation de la mémoire pour exécuter chaque étape d’un processus.
- Déploiement manuel des changements sans liste de contrôle.
- Différents membres de l’équipe exécutant le même processus, mais avec des étapes ou résultats différents.
- Désynchronisation des runbooks avec les changements du système et l’automatisation.

Avantages liés au respect de cette bonne pratique :

- Réduction du taux d'erreur pour les tâches manuelles.
- Exécution cohérente des opérations.
- Exécution plus précoce des tâches par les nouveaux membres de l'équipe.
- Automatisation des runbooks pour diminuer la quantité de travail.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les runbooks peuvent prendre plusieurs formes selon le niveau de maturité de votre entreprise. Ils devraient au minimum consister en un document step-by-step texte. Le résultat souhaité doit être clairement indiqué. Documentez explicitement les autorisations spéciales ou outils nécessaires. Fournissez des conseils sur la gestion des erreurs et les remontées en cas de problème. Recherchez le propriétaire du runbook et publiez-le dans un emplacement central. Une fois votre runbook documenté, validez-le en demandant à un membre de votre équipe de l'exécuter. À mesure que les procédures évoluent, mettez à jour vos runbooks conformément à votre processus de gestion des changements.

Vos runbooks texte doivent être automatisés à mesure que votre entreprise évolue. En utilisant des services tels que les automatisations d'AWS Systems Manager, vous pouvez transformer un fichier texte en automatisations pouvant être exécutées sur votre charge de travail. Ces automatisations peuvent être exécutées en réponse à des événements, réduisant ainsi la charge opérationnelle liée au maintien de votre charge de travail. AWS Systems Manager Automation fournit également une expérience de conception visuelle à faible code pour créer plus facilement des runbooks d'automatisation.

Exemple client

AnyCompany Retail doit effectuer des mises à jour du schéma de base de données lors des déploiements de logiciels. L'équipe en charge des opérations de cloud en collaboration avec l'équipe responsable de l'administration des bases de données a créé un runbook, pour déployer manuellement ces changements. Le runbook répertoriait chacune des étapes du processus sous forme de liste de contrôle. Il comprenait une section sur la gestion des erreurs en cas de problème. Les équipes ont publié le runbook sur leur wiki interne contenant leurs autres runbooks. L'équipe en charge des opérations de cloud envisage d'automatiser le runbook dans un prochain sprint.

Étapes d’implémentation

Si vous ne disposez pas d’un référentiel de documents, un référentiel de contrôle de version est un emplacement idéal pour commencer à créer votre bibliothèque de runbooks. Vous pouvez créer vos runbooks en utilisant le format Markdown. Voici un exemple de modèle de runbook que vous pouvez utiliser pour commencer à créer vos runbooks.

```
# Runbook Title
## Runbook Info
| Runbook ID | Description | Tools Used | Special Permissions | Runbook Author | Last
 Updated | Escalation POC |
|-------|-------|-------|-------|-------|-------|-------|
| RUN001 | What is this runbook for? What is the desired outcome? | Tools | Permissions
 | Your Name | 2022-09-21 | Escalation Name |
## Steps
1. Step one
2. Step two
```

1. Si vous ne possédez pas de référentiel de documentation ou de wiki existant, créez un référentiel de contrôle de version dans votre système de contrôle de version.
2. Identifiez un processus ne possédant pas de runbook. Le processus idéal doit être réalisé de manière semi-régulière, contenir peu d’étapes et avoir des échecs à faible impact.
3. Dans votre référentiel de documents, créer un brouillon au format Markdown en utilisant le modèle. Renseignez le titre du runbook et les champs obligatoires sous Runbook Info (Informations sur le runbook).
4. En commençant par la première étape, remplissez la section Steps (Étapes) du runbook.
5. Donnez le runbook à un membre de l’équipe. Demandez-lui d’utiliser le runbook pour valider les étapes. En cas d’élément manquant ou de besoin de clarification, mettez à jour le runbook.
6. Publiez le runbook sur votre référentiel de documentation interne. Une fois le runbook publié, partagez l’information avec votre équipe et les autres parties prenantes.
7. Au fil du temps, vous créerez une bibliothèque de runbooks. À mesure que cette bibliothèque s’étoffe, commencez à travailler sur l’automatisation des runbooks.

Niveau d’effort du plan d’implémentation : faible La norme minimale pour un runbook est un guide step-by-step textuel. L’automatisation des runbooks peut augmenter l’effort d’implémentation.

Ressources

Bonnes pratiques associées :

- OPS02-BP02 Les processus et les procédures ont identifié les propriétaires
- OPS07-BP04 Utiliser des playbooks pour étudier les problèmes
- OPS10-BP01 Utiliser un processus de gestion des événements, des incidents et des problèmes
- OPS10-BP02 Disposer d'un processus par alerte
- OPS11-BP04 Effectuer la gestion des connaissances

Documents connexes :

- AWS Well-Architected Framework: Concepts: Runbook development
- Achieving Operational Excellence using automated playbook and runbook
- AWS Systems Manager : utilisation des runbooks
- Manuel de migration pour les AWS grandes migrations - Tâche 4 : améliorer vos runbooks de migration
- Utiliser les runbooks AWS Automation pour résoudre des tâches opérationnelles

Vidéos connexes :

- AWS re:Invent 2019 : DIY guide des runbooks, des rapports d'incidents et de la réponse aux incidents
- Comment automatiser les opérations informatiques sur AWS | Amazon Web Services
- Intégrer des scripts dans AWS Systems Manager

Exemples connexes :

- Ateliers Well-Architected : automatisation des opérations avec les playbooks et les runbooks
- AWS Article de blog : Élaborez une pratique d'automatisation du cloud pour l'excellence opérationnelle : les meilleures pratiques de AWS Managed Services
- AWS Systems Manager : procédures détaillées relatives à l'automatisation
- AWS Systems Manager : restauration d'un volume racine à partir du dernier runbook de snapshots

- Création d'un manuel de réponse aux AWS incidents à l'aide des blocs-notes Jupyter et de Lake CloudTrail
- Gitlab – Runbooks
- Rubix – Une bibliothèque Python pour créer des runbooks dans les blocs-notes Jupyter
- Utilisation de Document Builder pour créer un runbook personnalisé

Services connexes :

- AWS Systems Manager Automation

OPS07-BP04 Utiliser des playbooks pour étudier les problèmes

Les playbooks sont des step-by-step guides utilisés pour enquêter sur un incident. Lorsque des incidents se produisent, les playbooks sont utilisés pour analyser, évaluer l'impact et identifier une cause racine. Les playbooks sont utilisés dans le cadre de différents scénarios allant des échecs de déploiement aux incidents de sécurité. Dans la plupart des cas, les playbooks identifient la cause racine qui est atténuée par l'utilisation d'un runbook. Les playbooks sont une composante essentielle des plans de réponse de votre organisation en cas d'incident.

Un playbook efficace comporte plusieurs fonctionnalités clés. Il guide l'utilisateur, étape par étape, dans le processus de découverte. Si vous optez pour un point de vue extérieur, quelles étapes devez-vous suivre pour diagnostiquer un incident ? Définissez clairement dans le playbook si des outils spéciaux ou des autorisations élevées sont nécessaires. Il est essentiel d'élaborer un plan de communication pour informer les parties prenantes du statut de l'analyse. Lorsqu'il est impossible de déterminer la cause racine, le playbook doit comporter un plan de remontée des informations vers la hiérarchie. Si la cause racine est identifiée, le playbook doit faire référence à un runbook décrivant une solution pour la résoudre. Les playbooks doivent être stockés dans un emplacement central et mis à jour régulièrement. Si des playbooks sont utilisés pour des alertes précises, donnez aux membres de votre équipe des indications relatives au playbook dans le cadre de l'alerte.

Au fur et à mesure que votre organisation évolue, automatisez vos playbooks. Commencez par des playbooks qui couvrent les incidents à faible risque. Utilisez des scripts pour automatiser les étapes de découverte. Veillez à créer des runbooks complémentaires destinés à atténuer les causes racines courantes.

Résultat escompté : votre organisation dispose de playbooks pour les incidents courants. Les playbooks sont stockés dans un emplacement central et mis à la disposition des membres de votre

équipe. Les playbooks sont souvent mis à jour. Pour toute cause racine connue, des runbooks complémentaires sont créés.

Anti-modèles courants :

- Il n'existe pas de façon standard d'analyser un incident.
- Les membres de l'équipe comptent sur la mémoire musculaire ou les connaissances institutionnelles pour résoudre un échec de déploiement.
- Les nouveaux membres de l'équipe apprennent à analyser les problèmes par un procédé de tâtonnement.
- Les bonnes pratiques d'analyse des problèmes ne sont pas partagées entre les équipes.

Avantages liés au respect de cette bonne pratique :

- Les playbooks dynamisent les efforts nécessaires pour atténuer les incidents.
- Différents membres de l'équipe peuvent utiliser le même playbook pour identifier une cause racine de façon cohérente.
- Les causes racines connues peuvent être associées à des runbooks développés spécialement pour leur résolution, ce qui permet d'accélérer le délai de récupération.
- Les playbooks permettent aux membres de l'équipe de commencer à apporter leur contribution plus tôt.
- Les équipes peuvent adapter leurs processus à l'aide de playbooks reproductibles.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

La façon dont vous créez et utilisez les playbooks dépend de la maturité de votre organisation. Si vous débutez dans le cloud, créez des playbooks sous forme de texte dans un référentiel de documents centralisé. Au fur et à mesure que votre organisation évolue, les playbooks peuvent devenir semi-automatisés avec des langages de script comme Python. Ces scripts peuvent être exécutés dans un bloc-notes Jupyter afin d'accélérer la découverte. Les organisations avancées ont des playbooks entièrement automatisés pour les problèmes courants qui sont corrigés automatiquement avec des runbooks.

Pour commencer à créer vos playbooks, répertoriez les incidents qui affectent couramment votre charge de travail. Pour commencer, choisissez des playbooks pour les incidents à faible risque dont

la cause racine a été réduite à quelques problèmes. Une fois que vous disposez de playbooks pour des scénarios plus simples, passez aux scénarios à risque élevé ou à ceux dont la cause racine est peu connue.

Vos playbooks sous forme de texte doivent être automatisés à mesure que votre entreprise évolue. Grâce à des services comme AWS Systems Manager Automations, les textes plats peuvent être transformés en automatismes. Ces automatisations peuvent être exécutées en fonction de votre charge de travail pour accélérer les analyses. Ces automatisations peuvent être activées en réponse à des événements, ce qui réduit le temps nécessaire pour découvrir et résoudre les incidents.

Les clients peuvent utiliser AWS Systems Manager Incident Manager pour intervenir en cas d'incidents. Ce service offre une interface unique pour trier les incidents, informer les parties prenantes pendant la découverte et l'atténuation, et collaborer tout au long de l'incident. Il utilise AWS les automatisations de Systems Manager pour accélérer la détection et la restauration.

Exemple client

Un incident de production a eu un impact sur AnyCompany Retail. L'ingénieur d'astreinte a utilisé un playbook pour analyser le problème. À mesure qu'il effectuait les différentes étapes, il a informé les parties prenantes identifiées dans le playbook de l'évolution de la situation. L'ingénieur a identifié que la cause racine était une condition de concurrence dans un service dorsal. À l'aide d'un runbook, l'ingénieur a relancé le service, remettant AnyCompany Retail en ligne.

Étapes d'implémentation

Si vous n'avez pas de référentiel de documents existant, nous vous suggérons de créer un référentiel de contrôle de version pour votre bibliothèque de playbooks. Vous pouvez créer vos playbooks en utilisant Markdown, qui est compatible avec la plupart des systèmes d'automatisation de playbook. Si vous démarrez de zéro, utilisez l'exemple de modèle de playbook suivant.

```
# Playbook Title
## Playbook Info
| Playbook ID | Description | Tools Used | Special Permissions | Playbook Author | Last
 Updated | Escalation POC | Stakeholders | Communication Plan |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| RUN001 | What is this playbook for? What incident is it used for? | Tools |
 Permissions | Your Name | 2022-09-21 | Escalation Name | Stakeholder Name | How will
 updates be communicated during the investigation? |
## Steps
1. Step one
2. Step two
```

1. Si vous ne possédez pas de référentiel de documents ni de wiki existant, créez un référentiel de contrôle de version pour vos playbooks dans votre système de contrôle de version.
2. Identifiez un problème courant qui doit être analysé. Il doit s'agir d'un scénario où la cause racine se limite à quelques problèmes et où la résolution présente peu de risques.
3. À l'aide du modèle Markdown, remplissez la section Playbook Name (Nom du playbook) et les champs sous Playbook Info (Informations sur le playbook).
4. Remplissez les étapes de résolution du problème. Soyez aussi clair que possible sur les actions à effectuer ou les domaines à analyser.
5. Remettez le playbook à un membre de l'équipe et demandez-lui de le passer en revue afin de le valider. S'il manque quelque chose ou si un point n'est pas clair, mettez à jour le playbook.
6. Publiez le playbook dans votre référentiel de documents et informez votre équipe et les parties prenantes.
7. Cette bibliothèque de playbooks s'enrichira à mesure que vous ajouterez d'autres playbooks. Une fois que vous avez plusieurs playbooks, commencez à les automatiser à l'aide d'outils tels que AWS Systems Manager Automations pour synchroniser automatisation et playbooks.

Niveau d'effort du plan d'implémentation : faible Vos playbooks doivent être des documents texte stockés dans un emplacement central. Les organisations plus avancées évolueront vers l'automatisation des playbooks.

Ressources

Bonnes pratiques associées :

- OPS02-BP02 Les processus et les procédures ont identifié les propriétaires
- OPS07-BP03 Utiliser des runbooks pour exécuter des procédures
- OPS10-BP01 Utiliser un processus de gestion des événements, des incidents et des problèmes
- OPS10-BP02 Disposer d'un processus par alerte
- OPS11-BP04 Effectuer la gestion des connaissances

Documents connexes :

- AWS Well-Architected Framework: Concepts: Playbook development
- Achieving Operational Excellence using automated playbook and runbook

- AWS Systems Manager : utilisation des runbooks
- Utiliser les runbooks AWS Automation pour résoudre des tâches opérationnelles

Vidéos connexes :

- AWS re:Invent 2019 : DIY guide des runbooks, des rapports d'incidents et de la réponse aux incidents (-R1) SEC318
- AWS Systems Manager Incident Manager - Ateliers AWS virtuels
- Intégrer des scripts dans AWS Systems Manager

Exemples connexes :

- AWS Customer Playbook Framework
- AWS Systems Manager : procédures détaillées relatives à l'automatisation
- Création d'un manuel de réponse aux AWS incidents à l'aide des blocs-notes Jupyter et de Lake CloudTrail
- Rubix : une bibliothèque Python pour créer des runbooks dans les blocs-notes Jupyter
- Utilisation de Document Builder pour créer un runbook personnalisé
- Ateliers Well-Architected : automatisation des opérations avec les playbooks et les runbooks
- Ateliers Well-Architected : playbook de réponse aux incidents avec Jupyter

Services connexes :

- AWS Systems Manager Automation
- AWS Systems Manager Gestionnaire des incidents

OPS07-BP05 Prendre des décisions éclairées pour déployer des systèmes et apporter des modifications

Mettez en place des processus pour les modifications réussies et ratées de votre charge de travail. Un pré-mortem est un exercice où une équipe simule un échec pour développer des stratégies d’atténuation. Utilisez des pré-mortems pour anticiper les échecs et créer des procédures le cas échéant. Évaluez les avantages et les risques liés au déploiement de modifications dans votre charge de travail. Vérifiez que toutes les modifications sont conformes à la gouvernance.

Résultat escompté :

- Vous prenez des décisions éclairées lorsque vous déployez des modifications dans votre charge de travail.
- Les modifications sont conformes à la gouvernance.

Anti-modèles courants :

- Déploiement d'une modification dans notre charge de travail sans disposer de processus pour gérer un déploiement raté.
- Modifications apportées à votre environnement de production qui ne sont pas conformes aux exigences de gouvernance.
- Déploiement une nouvelle version de votre charge de travail sans établir une base de référence pour l'utilisation des ressources.

Avantages liés au respect de cette bonne pratique :

- Vous êtes préparé à des modifications ratées de votre charge de travail.
- Les modifications apportées à votre charge de travail sont conformes aux politiques de gouvernance.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Utilisez des pré-mortems pour développer des processus pour les modifications ratées. Documentez vos processus pour les modifications ratées. Veillez à ce que toutes les modifications soient conformes à la gouvernance. Évaluez les avantages et les risques liés au déploiement de modifications dans votre charge de travail.

Exemple client

AnyCompany Le commerce de détail procède régulièrement à des autopsies afin de valider ses processus en cas de modifications infructueuses. La société documente ses processus dans un wiki partagé et le met à jour fréquemment. Toutes les modifications sont conformes aux exigences de gouvernance.

Étapes d'implémentation

1. Prenez des décisions éclairées lorsque vous déployez des modifications dans votre charge de travail. Définissez et révisez les critères d'un déploiement réussi. Développez des scénarios ou des critères qui déclencheraient la restauration d'une modification. Comparez les avantages du déploiement des modifications avec les risques associés à l'échec d'une modification.
2. Vérifiez que toutes les modifications sont conformes aux politiques de gouvernance.
3. Utilisez les pré-mortems pour planifier les modifications ratées et documenter les stratégies d'atténuation. Réalisez un exercice théorique pour modéliser une modification qui n'a pas abouti et valider les procédures de restauration.

Niveau d'effort du plan d'implémentation : modéré La mise en œuvre d'une pratique de pré-mortems nécessite une coordination et des efforts de la part des parties prenantes de votre organisation.

Ressources

Bonnes pratiques associées :

- OPS01-BP03 Évaluer les exigences de gouvernance – Les exigences de gouvernance sont un facteur clé pour déterminer s'il faut déployer une modification.
- OPS06-BP01 Planifier les modifications infructueuses – Établissez des plans pour atténuer les effets d'un déploiement raté et utilisez des pré-mortems pour les valider.
- OPS06-BP02 Déploiements de tests – Chaque modification apportée à un logiciel doit être correctement testée avant le déploiement afin de réduire les défauts en production.
- OPS07-BP01 Assurer la capacité du personnel – Il est essentiel de disposer de suffisamment de membres du personnel formés pour gérer la charge de travail afin de prendre une décision éclairée quant au déploiement d'une modification du système.

Documents connexes :

- Amazon Web Services : risques et conformité
- AWS Modèle de responsabilité partagée
- La gouvernance dans le AWS Cloud : le juste équilibre entre agilité et sécurité

OPS07-BP06 Créer des plans de support pour les charges de travail de production

Activez la prise en charge de tous les logiciels et services sur lesquels repose votre charge de travail de production. Sélectionnez un niveau de support approprié pour répondre à vos besoins

en matière de niveau de service de production. Il convient de prévoir des plans de support pour ces dépendances en cas d'interruption de service ou de problème logiciel. Documentez les plans de support et les procédures de demande de support pour tous les fournisseurs de services et de logiciels. Mettez en œuvre des mécanismes permettant de vérifier que les points de contact du support sont tenus à jour.

Résultat escompté :

- Mettez en œuvre des plans de support pour les logiciels et les services sur lesquels reposent les charges de travail de production.
- Choisissez une formule de support appropriée en fonction des besoins du niveau de service.
- Documentez les formules de support, les niveaux de support et les procédures de demande de support.

Anti-modèles courants :

- Vous n'avez pas de plan de support pour un fournisseur de logiciels critiques. Votre charge de travail en est affectée et vous ne pouvez rien faire pour accélérer la mise en place d'une solution ou obtenir des mises à jour en temps voulu de la part du fournisseur.
- Un développeur qui était le principal point de contact d'un fournisseur de logiciels a quitté l'entreprise. Vous n'arrivez pas à joindre directement le support du fournisseur. Vous devez passer du temps à rechercher et à naviguer dans des systèmes de contact génériques, ce qui augmente le temps nécessaire pour répondre en cas de besoin.
- Un fournisseur de logiciels connaît un arrêt de production. Il n'existe pas de documentation sur la manière de déposer un dossier de support.

Avantages liés au respect de cette bonne pratique :

- En adoptant le niveau de support approprié, vous êtes en mesure d'obtenir une réponse dans le délai nécessaire pour répondre aux besoins du niveau de service.
- En tant que client bénéficiant du support, vous pouvez faire remonter les problèmes de production.
- Les fournisseurs de logiciels et de services peuvent contribuer au dépannage pendant un incident.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d’implémentation

Activez les plans de support pour tous les fournisseurs de logiciels et de services sur lesquels repose votre charge de travail de production. Mettez en place des plans de support appropriés pour répondre aux besoins du niveau de service. Pour les AWS clients, cela signifie activer le AWS Business Support ou une version ultérieure sur tous les comptes où vous avez des charges de travail de production. Rencontrez régulièrement les fournisseurs de services de support afin d’obtenir des informations actualisées sur les offres de support, les processus et les contacts. Documentez les procédures de demande de support auprès des fournisseurs de logiciels et de services, y compris la manière de faire remonter les informations en cas de panne. Mettez en œuvre des mécanismes permettant de tenir à jour les contacts du support.

Exemple client

Chez AnyCompany Retail, toutes les dépendances liées aux logiciels et services commerciaux sont assorties de plans de support. Par exemple, le Support aux AWS entreprises est activé sur tous les comptes comportant des charges de travail de production. Tout développeur peut créer une demande de support en cas de problème. Il existe une page wiki contenant des informations sur la manière de demander de l’aide, sur les personnes à prévenir et sur les bonnes pratiques pour accélérer le traitement d’un incident.

Étapes d’implémentation

1. Travaillez avec les parties prenantes de votre organisation pour identifier les fournisseurs de logiciels et de services sur lesquels repose votre charge de travail. Documentez ces dépendances.
2. Déterminez les besoins en matière de niveau de service pour votre charge de travail. Sélectionnez un plan de support qui leur corresponde.
3. Pour les logiciels et services commerciaux, mettez en place une formule de support avec les fournisseurs.
    a. L'abonnement au Support aux AWS entreprises ou à une version ultérieure pour tous les comptes de production permet de réduire le temps de réponse de la part de AWS Support la société, ce qui est vivement recommandé. Si vous ne bénéficiez pas d'un support premium, vous devez disposer d'un plan d'action pour résoudre les problèmes nécessitant l'aide de AWS Support. AWS Support fournit une combinaison d'outils et de technologies, de personnes et de programmes conçus pour vous aider de manière proactive à optimiser les performances, à réduire les coûts et à innover plus rapidement. AWS Business Support offre des avantages supplémentaires, notamment l'accès au AWS Trusted Advisor AWS Personal Health Dashboard et des temps de réponse plus courts.

4. Documentez le plan de support dans votre outil de gestion des connaissances. Il s'agit notamment de savoir comment demander de l'aide, qui avertir en cas de demande de support et comment faire remonter l'information pendant un incident. Un wiki constitue un bon mécanisme pour permettre à quiconque d'apporter les mises à jour nécessaires à la documentation lorsqu'il prend connaissance de changements dans les processus ou les contacts de support.

Niveau d'effort du plan d'implémentation : faible La plupart des fournisseurs de logiciels et de services proposent des plans de support à l'inscription. La documentation et le partage des bonnes pratiques en matière de support sur votre système de gestion des connaissances permettent de vérifier que votre équipe sait ce qu'il faut faire en cas d'incident de production.

Ressources

Bonnes pratiques associées :

- OPS02-BP02 Les processus et procédures ont des propriétaires identifiés

Documents connexes :

- AWS Support Plans

Services connexes :

- AWS Support aux entreprises
- AWS Support aux entreprises

Exploitation

Questions

- OPS 8. Comment exploiter l'observabilité de la charge de travail dans l'organisation ?
- OPS 9. Comment comprendre l'état de vos opérations ?
- OPS 10. Comment gérer les événements relatifs à la charge de travail et aux opérations ?

OPS 8. Comment exploiter l’observabilité de la charge de travail dans l’organisation ?

Garantissez un état optimal de la charge de travail en tirant parti de l’observabilité. Utilisez des métriques, des journaux et des données de suivi pertinents pour obtenir une vue complète des performances de votre charge de travail et résoudre les problèmes de manière efficace.

Bonnes pratiques

- OPS08-BP01 Analyser les métriques de charge de travail
- OPS08-BP02 Analyser les journaux de charge de travail
- OPS08-BP03 Analyser les traces de charge de travail
- OPS08-BP04 Créez des alertes exploitables
- OPS08-BP05 Création de tableaux de bord

OPS08-BP01 Analyser les métriques de charge de travail

Après avoir implémenté la télémétrie des applications, analysez régulièrement les métriques collectées. Bien que la latence, les requêtes, les erreurs et la capacité (ou les quotas) fournissent des informations sur les performances du système, il est essentiel de donner la priorité à l’examen des métriques liées aux résultats commerciaux. Vous vous assurez ainsi de prendre des décisions basées sur des données conformes aux objectifs de votre entreprise.

Résultat escompté : informations précises sur les performances des charges de travail afin de prendre des décisions éclairées par les données, garantissant ainsi l’alignement avec les objectifs de votre entreprise.

Anti-modèles courants :

- Analyse des métriques de manière isolée sans tenir compte de leur impact sur les résultats commerciaux.
- Se fier de manière excessive aux métriques techniques tout en mettant de côté les métriques commerciales.
- Examen rare des métriques, ce qui vous fait passer à côté de possibilités de prise de décision en temps réel.

Avantages liés au respect de cette bonne pratique :

- Meilleure compréhension de la corrélation entre les performances techniques et les résultats commerciaux.
- Processus décisionnel amélioré grâce à des données en temps réel.
- Identification et atténuation proactives des problèmes avant qu'ils n'affectent les résultats commerciaux.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Tirez parti d'outils tels CloudWatch qu'Amazon pour effectuer des analyses métriques. AWS des services tels que la détection des CloudWatch anomalies et Amazon DevOps Guru peuvent être utilisés pour détecter des anomalies, en particulier lorsque les seuils statiques sont inconnus ou lorsque les modèles de comportement sont plus adaptés à la détection d'anomalies.

Étapes d'implémentation

1. Analyser et revoir : examinez et interprétez régulièrement les données relatives à votre charge de travail.
   a. Donnez la priorité aux métriques liées aux résultats commerciaux par rapport aux métriques purement techniques.
   b. Comprenez l'importance des pics, des baisses ou des tendances dans vos données.
2. Utilisez Amazon CloudWatch : utilisez Amazon CloudWatch pour une vue centralisée et une analyse approfondie.
   a. Configurez CloudWatch des tableaux de bord pour visualiser vos indicateurs et les comparer au fil du temps.
   b. Utilisez les percentiles CloudWatch pour avoir une vision claire de la distribution métrique, ce qui peut aider à définir SLAs et à comprendre les valeurs aberrantes.
   c. Configurez la détection des CloudWatch anomalies pour identifier les modèles inhabituels sans vous fier à des seuils statiques.
   d. Mettez en CloudWatch œuvre l'observabilité entre comptes pour surveiller et dépanner les applications qui couvrent plusieurs comptes au sein d'une même région.
   e. Utilisez CloudWatch Metric Insights pour interroger et analyser les données métriques de différents comptes et régions, afin d'identifier les tendances et les anomalies.
   f. Appliquez les mathématiques CloudWatch métriques pour transformer, agréger ou effectuer des calculs sur vos indicateurs afin d'obtenir des informations plus approfondies.

3. Utilisez Amazon DevOps Guru : intégrez Amazon DevOps Guru pour sa détection des anomalies améliorée par le machine learning afin d'identifier les premiers signes de problèmes opérationnels pour vos applications sans serveur et de les corriger avant qu'ils n'affectent vos clients.
4. Optimisation sur la base des informations recueillies : prenez des décisions éclairées grâce à l'analyse de vos métriques afin d'ajuster et d'améliorer vos charges de travail.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identifier les indicateurs de performance clés
- OPS04-BP02 Implémenter la télémétrie des applications

Documents connexes :

- The Wheel Blog : souligner l'importance de revoir continuellement les métriques
- Importance des centiles
- En utilisant AWS Cost Anomaly Detection
- CloudWatch observabilité entre comptes
- Interrogez vos indicateurs avec CloudWatch Metrics Insights

Vidéos connexes :

- Activer l'observabilité entre comptes sur Amazon CloudWatch
- Présentation d'Amazon DevOps Guru
- Analysez continuellement les métriques à l'aide de AWS Cost Anomaly Detection

Exemples connexes :

- Un atelier sur l'observabilité
- Obtenir des informations sur les opérations AIOps grâce à Amazon DevOps Guru

OPS08-BP02 Analyser les journaux de charge de travail

L’analyse régulière des journaux de charge de travail est essentielle pour mieux comprendre les aspects opérationnels de votre application. En analysant, en visualisant et en interprétant efficacement les données des journaux, vous pouvez optimiser en permanence les performances et la sécurité des applications.

Résultat escompté : informations détaillées sur le comportement et le fonctionnement des applications grâce à une analyse approfondie des journaux, garantissant une détection et une atténuation proactives des problèmes.

Anti-modèles courants :

- Négliger l’analyse des journaux jusqu’à ce qu’un problème critique survienne.
- Ne pas utiliser la suite complète d’outils disponibles pour l’analyse des journaux, ce qui fait passer à côté d’informations critiques.
- Se fier uniquement à l’examen manuel des journaux sans tirer parti des fonctionnalités d’automatisation et de requête.

Avantages liés au respect de cette bonne pratique :

- Identification proactive des goulots d’étranglement opérationnels, des menaces de sécurité et d’autres problèmes potentiels.
- Utilisation efficace des données de journal pour une optimisation continue des applications.
- Meilleure compréhension du comportement des applications, ce qui aide au débogage et au dépannage.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Amazon CloudWatch Logs est un puissant outil d'analyse des journaux. Des fonctionnalités intégrées telles que CloudWatch Logs Insights et Contributor Insights rendent le processus d'obtention d'informations pertinentes à partir des journaux intuitif et efficace.

Étapes d’implémentation

1. Configuration CloudWatch des journaux : configurez les applications et les services pour envoyer les journaux aux CloudWatch journaux.

2. Utilisez la détection des anomalies dans les journaux : utilisez la détection des anomalies d'Amazon CloudWatch Logs pour identifier automatiquement les modèles de journalisation inhabituels et vous avertir en cas d'anomalie. Cet outil vous permet de gérer de manière proactive les anomalies dans vos journaux et de détecter rapidement les problèmes potentiels.
3. Configurer CloudWatch Logs Insights : utilisez CloudWatch Logs Insights pour rechercher et analyser de manière interactive les données de vos journaux.
    a. Créez des requêtes pour extraire des modèles, visualiser les données des journaux et obtenir des informations exploitables.
    b. Utilisez l'analyse des modèles de CloudWatch Logs Insights pour analyser et visualiser les modèles de journaux fréquents. Cette fonctionnalité vous permet de comprendre les tendances opérationnelles courantes et les valeurs aberrantes potentielles dans les données de vos journaux.
    c. Utilisez CloudWatch Logs compare (diff) pour effectuer une analyse différentielle entre différentes périodes ou entre différents groupes de journaux. Utilisez cette fonctionnalité pour identifier les changements et évaluer leur impact sur les performances ou le comportement de votre système.
4. Surveillez les journaux en temps réel avec Live Tail : utilisez Amazon CloudWatch Logs Live Tail pour consulter les données des journaux en temps réel. Vous pouvez surveiller activement les activités opérationnelles de votre application au fur et à mesure qu'elles se produisent, ce qui fournit une visibilité immédiate sur les performances du système et les problèmes potentiels.
5. Tirez parti des informations sur les CloudWatchcontributeurs : utilisez les informations sur les contributeurs pour identifier les meilleurs intervenants dans des domaines à forte cardinalité, tels que les adresses IP ou les agents utilisateurs.
6. CloudWatch Implémenter les filtres métriques CloudWatch des journaux : configurez les filtres métriques des journaux pour convertir les données des journaux en indicateurs exploitables. Cela vous permettra de définir des alarmes ou d'analyser davantage les modèles.
7. Mettez en œuvre l'observabilité CloudWatch entre comptes : surveillez et dépannez les applications qui couvrent plusieurs comptes au sein d'une région.
8. Révision et perfectionnement réguliers : passez régulièrement en revue vos stratégies d'analyse des journaux afin de recueillir toutes les informations pertinentes et d'optimiser en permanence les performances des applications.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identifier les indicateurs de performance clés
- OPS04-BP02 Implémenter la télémétrie des applications
- OPS08-BP01 Analyser les métriques de charge de travail

Documents connexes :

- Analyse des données de journal avec CloudWatch Logs Insights
- Utilisation de CloudWatch Contributor Insights
- Création et gestion de filtres CloudWatch Log Metric

Vidéos connexes :

- Analysez les données des CloudWatch journaux avec Logs Insights
- Utilisez CloudWatch Contributor Insights pour analyser les données à haute cardinalité

Exemples connexes :

- CloudWatch Enregistre les exemples de requêtes
- Un atelier sur l'observabilité

OPS08-BP03 Analyser les traces de charge de travail

L'analyse des données de suivi est essentielle pour obtenir une vue complète du parcours opérationnel d'une application. En visualisant et en comprenant les interactions entre les différents composants, il est possible d'affiner les performances, d'identifier les goulots d'étranglement et d'améliorer l'expérience utilisateur.

Résultat escompté : vous bénéficiez d'une visibilité claire sur les opérations distribuées de votre application, ce qui permet de résoudre les problèmes plus rapidement et d'améliorer l'expérience utilisateur.

Anti-modèles courants :

- Négliger les données de suivi, en s'appuyant uniquement sur les journaux et les métriques.
- Aucune corrélation entre les données de suivi et les journaux associés.
- Ignorer les métriques dérivées des données de suivi, telles que la latence et les taux de défaillance.

Avantages liés au respect de cette bonne pratique :

- Améliorez le dépannage et réduisez le délai moyen de résolution (MTTR).
- Obtenez des informations exploitables sur les dépendances et leur impact.
- Accélérez l'identification et la résolution des problèmes de performance.
- Tirez parti des métriques dérivées des données de suivi pour une prise de décision éclairée.
- Améliorez les expériences utilisateur grâce à des interactions optimisées entre les composants.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

AWS X-Ray propose une suite complète pour l'analyse des données de suivi. Il fournit une vue globale des interactions entre les services, surveille les activités des utilisateurs et détecte les problèmes de performance. Des fonctionnalités telles que ServiceLens X-Ray Insights, X-Ray Analytics et Amazon DevOps Guru améliorent la profondeur des informations exploitables dérivées des données de trace.

Étapes d'implémentation

Les étapes suivantes proposent une approche structurée pour mettre en œuvre efficacement l'analyse des données de trace à l'aide de AWS services :

1. Intégrer AWS X-Ray : assurez-vous que X-Ray est intégré à vos applications pour capturer les données de suivi.
2. Analyse des métriques X-Ray : explorez les métriques dérivées des traces X-Ray, telles que la latence, les taux de demandes, les taux d'erreur et la distribution des temps de réponse, en utilisant la carte des services pour surveiller l'état de santé des applications.
3. Utilisation ServiceLens : Tirez parti de la ServiceLenscarte pour améliorer l'observabilité de vos services et applications. Cela permet une visualisation intégrée des données de suivi, des métriques, des journaux, des alarmes et d'autres informations liées à l'état.

4. Activation de X-Ray Insights :
   a. Activez X-Ray Insights pour détecter automatiquement les anomalies dans les traces.
   b. Examinez les informations pour identifier les tendances et en déterminer les causes racines, telles que l'augmentation des taux de défaillance ou des latences.
   c. Consultez la chronologie des informations pour une analyse temporelle des problèmes détectés.
5. Utilisation de X-Ray Analytics : X-Ray Analytics vous permet d'explorer en profondeur les données de trace, d'identifier des modèles et d'en extraire des informations.
6. Utilisation de groupes dans X-Ray : créez des groupes dans X-Ray pour filtrer les données de suivi en fonction de critères tels qu'une latence élevée, afin de permettre une analyse plus ciblée.
7. Intégrez Amazon DevOps Guru : faites appel à Amazon DevOps Guru pour tirer parti des modèles d'apprentissage automatique qui détectent les anomalies opérationnelles dans les traces.
8. Utilisez CloudWatch des synthetics : utilisez des synthetics pour créer des CloudWatchcanaris afin de surveiller en permanence vos points de terminaison et vos flux de travail. Ces scripts canary peuvent s'intégrer à X-Ray pour fournir des données de suivi permettant une analyse approfondie des applications testées.
9. Utilisez Real User Monitoring (RUM) : avec AWS X-Ray et CloudWatch RUM, vous pouvez analyser et déboguer le chemin de la demande en commençant par les utilisateurs finaux de votre application via les services AWS gérés en aval. Cela vous permet d'identifier les tendances de latence et les erreurs qui ont un impact sur les utilisateurs finaux.
10. Corrélation avec les journaux : corrélez les données de suivi avec les journaux associés dans la vue de suivi de X-Ray pour obtenir une perspective détaillée du comportement des applications. Cela vous permet de visualiser les événements de journal directement associés aux transactions suivies.
11. Mettez en œuvre l'observabilité CloudWatch entre comptes : surveillez et dépannez les applications qui couvrent plusieurs comptes au sein d'une même région.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS08-BP01 Analyser les métriques de charge de travail
- OPS08-BP02 Analyser les journaux de charge de travail

Documents connexes :

- Utilisation ServiceLens pour surveiller l'état de santé des applications
- Exploration des données de suivi grâce à X-Ray Analytics
- Détection des anomalies dans les données de suivi grâce à X-Ray Insights
- Surveillance continue avec CloudWatch Synthetics

Vidéos connexes :

- Analysez et déboguez des applications à l'aide d'Amazon CloudWatch Synthetics & AWS X-Ray
- Use AWS X-Ray Insights

Exemples connexes :

- Un atelier sur l’observabilité
- Implémentation de X-Ray avec AWS Lambda
- CloudWatchModèles Synthetics Canary

OPS08-BP04 Créez des alertes exploitables

Il est crucial de détecter rapidement les écarts de comportement de votre application et d’y réagir rapidement. Il est particulièrement important de savoir quand les résultats basés sur des indicateurs de performance clés (KPIs) sont menacés ou lorsque des anomalies inattendues surviennent. La base des alertes KPIs garantit que les signaux que vous recevez sont directement liés à l'impact commercial ou opérationnel. Cette approche des alertes exploitables favorise les réponses proactives et contribue à maintenir les performances et la fiabilité du système.

Résultat souhaité : Recevez des alertes opportunes, pertinentes et exploitables pour identifier et atténuer rapidement les problèmes potentiels, en particulier lorsque KPI les résultats sont menacés.

Anti-modèles courants :

- Configurer un trop grand nombre d’alertes non critiques, ce qui entraîne de la lassitude.
- Ne pas hiérarchiser les alertes en fonction de KPIs celles-ci, ce qui complique la compréhension de l'impact commercial des problèmes.
- Négliger de traiter les causes profondes, ce qui entraîne des alertes répétitives pour le même problème.

Avantages liés au respect de cette bonne pratique :

- Réduction de la lassitude liée aux alertes grâce à des alertes pertinentes et exploitables.
- Disponibilité et fiabilité du système améliorées grâce à la détection et à l'atténuation proactives des problèmes.
- Collaboration d'équipe améliorée et résolution plus rapide des problèmes grâce à l'intégration à des outils connus d'alerte et de communication.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Pour créer un mécanisme d'alerte efficace, il est essentiel d'utiliser des métriques, des journaux et des données de suivi qui signalent les cas où les résultats KPIs sont menacés ou lorsque des anomalies sont détectées.

Étapes d'implémentation

1. Déterminer les indicateurs de performance clés (KPIs) : Identifiez ceux de votre applicationKPIs. Les alertes doivent être liées à celles-ci KPIs afin de refléter avec précision l'impact commercial.
2. Mise en œuvre de la détection des anomalies :
   - Utilisez la détection des CloudWatch anomalies Amazon : configurez la détection des CloudWatch anomalies Amazon pour détecter automatiquement les modèles inhabituels, ce qui vous permet de générer des alertes uniquement pour les anomalies authentiques.
   - Utilisez AWS X-Ray Insights :
     a. Configurez X-Ray Insights pour détecter les anomalies dans les données de trace.
     b. Configurez les notifications pour que X-Ray Insights soit alerté des problèmes détectés.
   - Intégrez Amazon DevOps Guru :
     a. Tirez parti d'Amazon DevOps Guru pour ses capacités d'apprentissage automatique permettant de détecter les anomalies opérationnelles avec les données existantes.
     b. Accédez aux paramètres de notification dans DevOps Guru pour configurer des alertes d'anomalie.
3. Mise en place d'alertes exploitables : concevez des alertes qui fournissent des informations adéquates pour une action immédiate.
   1. Surveillez AWS Health les événements EventBridge selon les règles d'Amazon ou intégrez-les par programmation AWS Health API pour automatiser les actions lorsque vous recevez

des AWS Health événements. Il peut s'agir d'actions générales, telles que l'envoi de tous les messages relatifs aux événements du cycle de vie planifiés vers une interface de discussion, ou d'actions spécifiques, telles que le lancement d'un flux de travail dans un outil de gestion des services informatiques.

4. Réduction de la fatigue liée aux alertes : minimisez les alertes non critiques. Lorsque les équipes sont submergées par de nombreuses alertes insignifiantes, elles peuvent finir par ignorer des problèmes critiques, ce qui diminue l'efficacité globale du mécanisme d'alerte.
5. Configurez des alarmes composites : utilisez les alarmes CloudWatch composites Amazon pour consolider plusieurs alarmes.
6. Intégration aux outils d'alerte : Incorporez des outils tels que Ops Genie et PagerDuty.
7. Engagez-vous AWS Chatbot : intégrez AWS Chatbotpour relayer les alertes vers Amazon Chime, Microsoft Teams et Slack.
8. Alerte basée sur les journaux : utilisez des filtres métriques de journalisation CloudWatch pour créer des alarmes basées sur des événements de journal spécifiques.
9. Révision et itération : révisez et affinez régulièrement les configurations des alertes.

Niveau d'effort du plan d'implémentation : moyen

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identifier les indicateurs de performance clés
- OPS04-BP02 Implémenter la télémétrie des applications
- OPS04-BP03 Implémenter la télémétrie de l'expérience utilisateur
- OPS04-BP04 Implémenter la télémétrie des dépendances
- OPS04-BP05 Mettre en œuvre le traçage distribué
- OPS08-BP01 Analyser les métriques de charge de travail
- OPS08-BP02 Analyser les journaux de charge de travail
- OPS08-BP03 Analyser les traces de charge de travail

Documents connexes :

- Utilisation des CloudWatch alarmes Amazon
- Création d'une alerte composite

- Création d'une CloudWatch alarme basée sur la détection d'anomalies
- DevOpsNotifications du guru
- Notifications relatives aux rayons X
- Surveillez, gérez et dépannez vos AWS ressources grâce à des outils interactifs ChatOps
- Guide CloudWatch d'intégration Amazon | PagerDuty
- Intégrez Opsgenie à Amazon CloudWatch

Vidéos connexes :

- Création d'alarmes composites dans Amazon CloudWatch
- AWS Chatbot Présentation
- AWS Sur Air ft. Commandes mutatives dans AWS Chatbot

Exemples connexes :

- Alarmes, gestion des incidents et résolution dans le cloud avec Amazon CloudWatch
- Tutoriel : Création d'une EventBridge règle Amazon qui envoie des notifications à AWS Chatbot
- Un atelier sur l’observabilité

OPS08-BP05 Création de tableaux de bord

Les tableaux de bord offrent une vue centrée sur l’humain des données télémétriques de vos charges de travail. Bien qu’ils fournissent une interface visuelle essentielle, ils ne doivent pas remplacer les mécanismes d’alerte, mais les compléter. Lorsqu’ils sont conçus avec soin, ils peuvent non seulement fournir des informations rapides sur l’état et les performances du système, mais ils peuvent également présenter aux parties prenantes des informations en temps réel sur les résultats commerciaux et l’impact des problèmes.

Résultat escompté :

Informations claires et exploitables sur l’état du système et de l’entreprise à l’aide de représentations visuelles.

Anti-modèles courants :

- Tableaux de bord trop compliqués avec trop de métriques.

- Utilisation de tableaux de bord sans alertes pour détecter les anomalies.
- Pas de mise à jour des tableaux de bord à mesure que les charges de travail évoluent.

Avantages liés au respect de cette bonne pratique :

- Visibilité immédiate sur les indicateurs critiques du système etKPIs.
- Amélioration de la communication et de la compréhension avec les parties prenantes.
- Aperçu de l’impact des problèmes opérationnels.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Tableaux de bord centrés sur l’entreprise

Les tableaux de bord adaptés aux besoins des entreprises KPIs mobilisent un plus large éventail de parties prenantes. Bien que ces personnes ne soient pas intéressées par les métriques du système, elles souhaitent comprendre les implications commerciales de ces chiffres. Un tableau de bord centré sur l’entreprise garantit que toutes les métriques techniques et opérationnelles surveillées et analysées sont synchronisées avec les objectifs globaux de l’entreprise. Cet alignement apporte de la clarté et garantit que tout le monde est d’accord sur ce qui est essentiel et sur ce qui ne l’est pas. En outre, les tableaux de bord qui mettent en avant les activités KPIs ont tendance à être plus exploitables. Les parties prenantes peuvent rapidement comprendre l’état des opérations, les domaines nécessitant une attention particulière et l’impact potentiel sur les résultats commerciaux.

Dans cette optique, lors de la création de vos tableaux de bord, assurez-vous qu'il existe un équilibre entre les indicateurs techniques et les activités commercialesKPIs. Les deux sont essentiels, mais ils s’adressent à des publics différents. Idéalement, vous devriez disposer de tableaux de bord offrant une vue globale de l’état et des performances du système tout en mettant l’accent sur les principaux résultats commerciaux et leurs implications.

Les CloudWatch tableaux de bord Amazon sont des pages d'accueil personnalisables dans la CloudWatch console que vous pouvez utiliser pour surveiller vos ressources dans une seule vue, même celles qui sont réparties entre différents comptes Régions AWS et.

Étapes d’implémentation

1. Création d'un tableau de bord de base : créez un nouveau tableau de bord dans CloudWatch, en lui donnant un nom descriptif.

2. Utilisez les widgets Markdown : avant de vous plonger dans les statistiques, utilisez les widgets Markdown pour ajouter du contexte textuel en haut de votre tableau de bord. Expliquez ce que couvre le tableau de bord et l'importance des métriques représentées, et ajoutez éventuellement des liens vers d'autres tableaux de bord et outils de résolution des problèmes.
3. Création de variables de tableau de bord : incorporez des variables de tableau de bord le cas échéant pour permettre des vues de tableau de bord dynamiques et flexibles.
4. Créer des widgets de mesure : ajoutez des widgets de mesure pour visualiser les différentes métriques émises par votre application, en personnalisant ces widgets pour représenter efficacement l'état du système et les résultats commerciaux.
5. Requêtes Log Insights : utilisez CloudWatchLog Insights pour obtenir des indicateurs exploitables à partir de vos journaux et afficher ces informations sur votre tableau de bord.
6. Configurez les alarmes : intégrez les CloudWatchalarmes à votre tableau de bord pour avoir un aperçu rapide de toutes les métriques dépassant leurs seuils.
7. Utilisez les informations sur les CloudWatch contributeurs : intégrez les informations sur les contributeurs pour analyser les champs à forte cardinalité et mieux comprendre les principaux contributeurs de votre ressource.
8. Concevez des widgets personnalisés : pour des besoins spécifiques qui ne sont pas satisfaits par les widgets standard, pensez à créer des widgets personnalisés. Ils peuvent être extraits de différentes sources de données ou représenter les données de manière unique.
9. Utilisation AWS Health Dashboard : AWS Health Dashboardà utiliser pour obtenir des informations plus détaillées sur l'état de votre compte, les événements et les modifications à venir susceptibles d'affecter vos services et ressources. Vous pouvez également obtenir une vue centralisée des événements d'état dans votre tableau de bord AWS Organizations ou créer vos propres tableaux de bord personnalisés (pour plus de détails, voir Exemples connexes).
10. Répéter et affiner : au fur et à mesure que votre application évolue, revoyez régulièrement votre tableau de bord pour vous assurer de sa pertinence.

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identifier les indicateurs de performance clés
- OPS08-BP01 Analyser les métriques de charge de travail
- OPS08-BP02 Analyser les journaux de charge de travail
- OPS08-BP03 Analyser les traces de charge de travail

- OPS08-BP04 Créez des alertes exploitables

Documents connexes :

- Création de tableaux de bord pour une visibilité opérationnelle
- Utilisation des tableaux de CloudWatch bord Amazon

Vidéos connexes :

- Créez des CloudWatch tableaux de bord multicomptes et interrégionaux
- AWS re:Invent 2021 - Gain enterprise visibility with AWS Cloud operation dashboards)

Exemples connexes :

- Un atelier sur l'observabilité
- Surveillance des applications avec Amazon CloudWatch
- AWS Health Tableaux de bord et informations sur l'intelligence événementielle
- Visualisez les événements AWS Health à l'aide d'Amazon Managed Grafana

OPS 9. Comment comprendre l'état de vos opérations ?

Définissez, capturez et analysez les métriques des opérations pour obtenir une visibilité sur les événements opérationnels afin de pouvoir prendre des mesures appropriées.

Bonnes pratiques

- OPS09-BP01 Mesurer les objectifs opérationnels et les KPI à l'aide de métriques
- OPS09-BP02 Communiquer l'état et les tendances pour garantir la visibilité des opérations
- OPS09-BP03 Examiner les indicateurs des opérations et prioriser les améliorations

OPS09-BP01 Mesurer les objectifs opérationnels et les KPI à l'aide de métriques

Obtenez des objectifs et des indicateurs de performance clés qui définissent le succès des opérations de votre organisation et déterminez les métriques qui les reflètent. Définissez des points de référence et réévaluez-les régulièrement. Développez des mécanismes permettant de recueillir ces métriques auprès des équipes à des fins d'évaluation. Les métriques DevOps Research and Assessment

(DORA) constituent une méthode populaire pour mesurer les progrès accomplis dans la mise en œuvre des pratiques DevOps en matière de fourniture de logiciels.

Résultat escompté :

- L'organisation publie et partage les objectifs et les KPI des équipes opérationnelles.
- Vous établissez des métriques qui reflètent ces KPI. Exemples :
  - Profondeur de la file d'attente ou âge moyen des tickets
  - Nombre de tickets regroupés par type de problème
  - Temps passé à résoudre les problèmes avec ou sans procédure opérationnelle normalisée (SOP)
  - Délai de récupération après un échec d'envoi de code
  - Volume d'appels

Anti-modèles courants :

- Les délais de déploiement ne sont pas respectés, car les développeurs sont contraints d'effectuer des tâches de dépannage. Les équipes de développement plaident en faveur d'une augmentation du personnel, mais ne peuvent pas quantifier le nombre de collaborateurs dont elles ont besoin, car le temps perdu ne peut pas être mesuré.
- Un bureau de niveau 1 a été mis en place pour traiter les appels des utilisateurs. Au fil du temps, de nouvelles charges de travail ont été ajoutées, mais aucun effectif n'a été affecté au bureau de niveau 1. La satisfaction des clients en pâtit, car les temps d'appel augmentent et la résolution des problèmes ralentit, mais la direction n'en voit aucun signe, ce qui empêche toute action.
- Une charge de travail problématique a été confiée à une équipe opérationnelle distincte pour entretien. Contrairement aux autres charges de travail, cette nouvelle charge de travail n'a pas été fournie avec la documentation et les runbooks appropriés. Les équipes consacrent donc plus de temps au dépannage et à la résolution des défaillances. Cependant, aucune métrique ne permet de documenter ces efforts, ce qui empêche les équipes de rendre compte de la situation.

Avantages liés au respect de cette bonne pratique : lorsque la surveillance de la charge de travail indique l'état de nos applications et services, les équipes chargées de la surveillance des opérations fournissent aux propriétaires un aperçu des changements survenus chez les consommateurs de ces charges de travail, tels que l'évolution des besoins commerciaux. Mesurez l'efficacité de ces équipes et évaluez-les par rapport aux objectifs commerciaux en créant des métriques qui reflètent l'état des

opérations. Ces métriques peuvent mettre en évidence les problèmes de support ou identifier les cas où des écarts se produisent par rapport à une cible de niveau de service.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Planifiez du temps avec les responsables et les parties prenantes afin de déterminer les objectifs généraux du service. Déterminez quelles devraient être les tâches des différentes équipes opérationnelles et quels défis elles pourraient rencontrer. Sur la base de ces informations, réfléchissez à des indicateurs de rendement clés (KPI) susceptibles de refléter ces objectifs opérationnels. Il peut s'agir de la satisfaction des clients, du délai entre la conception des fonctionnalités et leur déploiement, du temps moyen de résolution des problèmes ou de la rentabilité.

À partir de ces KPI, identifiez les métriques et les sources de données qui pourraient mieux refléter ces objectifs. La satisfaction des clients peut être une combinaison de diverses métriques telles que les temps d'attente ou de réponse aux appels, les scores de satisfaction et les types de problèmes soulevés. Les temps de déploiement peuvent être la somme du temps nécessaire aux tests et au déploiement, plus les correctifs à ajouter après le déploiement lui-même. Les statistiques indiquant le temps consacré à différents types de problèmes (ou le nombre de ces problèmes) peuvent fournir un aperçu des domaines dans lesquels des efforts ciblés sont nécessaires.

Ressources

Documents connexes :

- Amazon QuickSight - Utilisation des KPI
- Amazon CloudWatch : utilisation des métriques
- Création de tableaux de bord
- Comment suivre vos KPI en matière d'optimisation des coûts avec le tableau de bord des KPI
- Guide AWS DevOps

Exemples connexes :

- Surveillance des performances de votre livraison de logiciels à l'aide d'outils AWS natifs de surveillance et d'observabilité
- Équilibrage de la vitesse de déploiement et de la stabilité à l'aide des métriques DORA
- Exemples de métriques opérationnelles MLOps dans le secteur des services financiers

- Suivi des KPI d'optimisation des coûts avec KPI Dashboard

OPS09-BP02 Communiquer l'état et les tendances pour garantir la visibilité des opérations

Il est nécessaire de connaître l'état de vos opérations et leurs tendances pour identifier les cas où les résultats peuvent être menacés, pour déterminer si des efforts supplémentaires sont justifiés ou non, ou pour identifier les effets des modifications sur vos équipes. Lors d'événements opérationnels, la possession de pages d'état auxquelles les utilisateurs et les équipes opérationnelles peuvent se référer pour obtenir des informations peut réduire la pression sur les canaux de communication et à diffuser les informations de manière proactive.

Résultat escompté :

- Les responsables des opérations ont un aperçu rapide des volumes d'appels auxquels leurs équipes sont confrontées et des initiatives en cours, telles que les déploiements.
- Des alertes sont diffusées aux parties prenantes et aux communautés d'utilisateurs lorsque des répercussions sur les opérations normales se produisent.
- La direction de l'organisation et les parties prenantes peuvent consulter une page d'état en réponse à une alerte ou à un impact, et obtenir des informations concernant un événement opérationnel, telles que les points de contact, des informations sur les tickets et les délais de reprise estimés.
- Des rapports sont mis à la disposition de la direction et des autres parties prenantes pour présenter des statistiques opérationnelles telles que le volume d'appels sur une période donnée, les scores de satisfaction des utilisateurs, le nombre de tickets en attente et leur ancienneté.

Anti-modèles courants :

- Une charge de travail tombe en panne, ce qui rend un service indisponible. Les volumes d'appels atteignent un pic lorsque les utilisateurs demandent à savoir ce qui se passe. Les responsables ajoutent au volume en demandant à savoir qui est à l'origine du problème. Les différentes équipes opérationnelles redoublent leurs efforts pour tenter d'identifier la cause première.
- Pour répondre à un nouveau besoin, plusieurs membres du personnel sont réaffectés à un effort d'ingénierie. Les postes vacants ne sont pas pourvus, et les délais de résolution des problèmes augmentent. Ces informations ne sont pas capturées, et ce n'est qu'après plusieurs semaines et après avoir reçu des commentaires insatisfaits des utilisateurs que les dirigeants prennent conscience du problème.

Avantages liés au respect de cette bonne pratique : lors d'événements opérationnels affectant l'entreprise, beaucoup de temps et d'énergie peuvent être gaspillés à demander des informations aux différentes équipes qui tentent de comprendre la situation. En mettant en place des pages d'état et des tableaux de bord largement diffusés, les parties prenantes peuvent rapidement se procurer les informations nécessaires et déterminer, par exemple, si un problème a été détecté ou non, qui est responsable du problème ou quand un retour à une activité normale est attendu. Cela évite aux membres de l'équipe d'avoir à passer trop de temps à communiquer la situation aux autres. Ils peuvent ainsi consacrer plus de temps à la résolution des problèmes.

En outre, les tableaux de bord et les rapports peuvent fournir des informations aux décideurs et aux parties prenantes pour voir comment les équipes opérationnelles sont en mesure de répondre aux besoins de l'entreprise et comment leurs ressources sont allouées. Ces informations sont cruciales pour déterminer si des ressources adéquates sont en place pour soutenir l'entreprise.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Créez des tableaux de bord qui présentent les métriques clés actuelles pour vos équipes opérationnelles et mettez-les à disposition des responsables des opérations et de la direction.

Créez des pages d'état qui peuvent être mises à jour rapidement pour indiquer quand un incident ou un événement se produit, qui en est le responsable et qui coordonne la réponse. Partagez sur cette page les étapes ou les solutions que les utilisateurs doivent prendre en compte et diffusez largement l'emplacement. Encouragez les utilisateurs à vérifier d'abord cet emplacement lorsqu'ils sont confrontés à un problème inconnu.

Collectez et fournissez des rapports qui présentent l'état des opérations au fil du temps, et distribuez-les aux dirigeants et aux décideurs pour illustrer le travail des opérations ainsi que les défis et les besoins.

Partagez entre les équipes les métriques et rapports qui reflètent au mieux les objectifs et les KPI, ainsi que les domaines où ils ont contribué au changement. Consacrez du temps à ces activités afin de renforcer l'importance des opérations au sein des équipes et entre elles.

Ressources

Bonnes pratiques associées :

- OPS09-BP01 Mesurer les objectifs opérationnels et les KPI à l'aide de métriques

Documents connexes :

- Mesurer les progrès
- Création de tableaux de bord pour une visibilité opérationnelle

Exemples connexes :

- Opérations de données
- Comment suivre vos KPI en matière d'optimisation des coûts avec le tableau de bord des KPI
- L'importance des indicateurs de performance clés (KPI) pour les migrations cloud à grande échelle

OPS09-BP03 Examiner les indicateurs des opérations et prioriser les améliorations

Le fait de consacrer du temps et des ressources à l'examen de l'état des opérations garantit que servir day-to-day le secteur d'activité demeure une priorité. Réunissez les responsables des opérations et les parties prenantes pour vérifier régulièrement les métriques, réaffirmer ou modifier les objectifs et prioriser les améliorations.

Résultat escompté :

- Les responsables des opérations et le personnel se rencontrent régulièrement pour vérifier les métriques au cours d'une période de référence donnée. Les défis sont communiqués, les victoires sont célébrées et les leçons tirées sont partagées.
- Les parties prenantes et les chefs d'entreprise sont régulièrement informés de l'état des opérations et sollicités pour leur contribution concernant les objectifs et les KPIs initiatives futures. Les compromis entre la prestation de services, les opérations et la maintenance font l'objet de discussions et sont mis en contexte.

Anti-modèles courants :

- Un nouveau produit est lancé, mais les équipes opérationnelles de niveau 1 et de niveau 2 ne sont pas suffisamment formées pour fournir l'assistance nécessaire ou n'ont pas de personnel supplémentaire. Les métriques qui montrent une dégradation des délais de résolution des demandes d'assistance et l'augmentation du volume d'incidents ne sont pas pris en compte par les dirigeants. Des mesures sont prises des semaines plus tard lorsque le nombre d'abonnements commence à baisser alors que les utilisateurs mécontents quittent la plateforme.

- Un processus manuel pour effectuer la maintenance d'une charge de travail est en place depuis longtemps. Bien que le désir d'automatiser soit présent, il n'était pas prioritaire compte tenu de la faible importance du système. Cependant, au fil du temps, le système gagne de l'importance et ces processus manuels occupent désormais la majeure partie du temps des opérations. Aucune ressource n'est prévue pour assister les opérations, ce qui entraîne un épuisement du personnel à mesure que la charge de travail augmente. La direction n'en prend conscience que lorsqu'on lui signale que le personnel démissionne pour aller travailler pour d'autres concurrents.

Avantages liés au respect de cette bonne pratique : dans certaines organisations, il peut être difficile de consacrer le même temps et la même attention à la prestation de services et aux nouveaux produits ou offres. Le cas échéant, le secteur d'activité peut en pâtir, car le niveau de service attendu se détériore lentement. En effet, les opérations ne changent pas et n'évoluent pas avec la croissance de l'entreprise, et peuvent se retrouver à la traîne. En l'absence d'un examen régulier des informations recueillies par les opérations, le risque pour l'entreprise peut ne devenir visible que lorsqu'il sera trop tard. En allouant du temps à l'examen des métriques et des procédures à la fois au sein des équipes opérationnelles et auprès de la direction, le rôle crucial joué par les opérations reste visible, et les risques peuvent être identifiés bien avant qu'ils n'atteignent des niveaux critiques. Les équipes opérationnelles ont une meilleure idée des changements et initiatives commerciaux imminents, ce qui permet de lancer des initiatives proactives. La visibilité qu'ont les dirigeants sur les métriques opérationnelles met en évidence le rôle que jouent ces équipes dans la satisfaction des clients, à la fois en interne et en externe. Elle leur permet également de mieux évaluer les choix en fonction des priorités, ou de s'assurer que les opérations disposent du temps et des ressources nécessaires pour changer et évoluer avec de nouvelles initiatives stratégiques et de charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Consacrez du temps à la vérification des métriques opérationnelles entre les parties prenantes et les équipes opérationnelles et à l'examen des données des rapports. Placez ces rapports dans le contexte des objectifs de l'organisation afin de déterminer s'ils sont atteints. Identifiez les sources d'ambiguïté lorsque les objectifs ne sont pas clairs ou lorsque l'offre ne correspond pas à la demande.

Identifiez les domaines dans lesquels de meilleurs résultats opérationnels peuvent être obtenus avec du temps, du personnel et des outils disponibles. Déterminez quel impact KPIs cela pourrait avoir et quels devraient être les objectifs de réussite. Révisez-les régulièrement pour vous assurer que les opérations disposent de ressources suffisantes pour soutenir le secteur d'activité.

Ressources

Documents connexes :

- Amazon Athena
- Référence CloudWatch des métriques et dimensions Amazon
- Amazon QuickSight
- AWS Glue
- AWS Glue Data Catalog
- Collectez des métriques et des journaux à partir d'EC2instances Amazon et de serveurs sur site avec l'agent Amazon CloudWatch
- Utilisation des CloudWatch métriques Amazon

OPS 10. Comment gérer les événements relatifs à la charge de travail et aux opérations ?

Préparez et validez des procédures de réponse aux événements afin de réduire leur effet disruptif sur votre charge de travail.

Bonnes pratiques

- OPS10-BP01 Utiliser un processus de gestion des événements, des incidents et des problèmes
- OPS10-BP02 Disposer d’un processus par alerte
- OPS10-BP03 Prioriser les événements opérationnels en fonction de leur impact commercial
- OPS10-BP04 Définir les chemins d'escalade
- OPS10-BP05 Définir un plan de communication client pour les événements ayant un impact sur le service
- OPS10-BP06 Communiquer le statut par le biais de tableaux de bord
- OPS10-BP07 Automatiser les réponses aux événements

OPS10-BP01 Utiliser un processus de gestion des événements, des incidents et des problèmes

La capacité à gérer efficacement les événements, les incidents et les problèmes est essentielle pour préserver l’intégrité et les performances de la charge de travail. Il est essentiel de reconnaître et de comprendre les différences entre ces éléments pour développer une stratégie de réponse et de

résolution efficace. La mise en place et le suivi d'un processus bien défini pour chaque aspect aident votre équipe à relever rapidement et efficacement tous les défis opérationnels qui se présentent.

Résultat escompté : votre organisation gère efficacement les événements opérationnels, les incidents et les problèmes grâce à des processus bien documentés et stockés de manière centralisée. Ces processus sont constamment mis à jour pour refléter les changements, rationaliser la gestion et préserver une fiabilité de service et des performances de charge de travail élevées.

Anti-modèles courants :

- Vous êtes réactif et non proactif face aux événements.
- Des approches incohérentes sont adoptées à l'égard de différents types d'événements ou d'incidents.
- Votre organisation n'analyse pas les incidents et n'en tire pas les leçons nécessaires pour éviter qu'ils se reproduisent à l'avenir.

Avantages liés au respect de cette bonne pratique :

- Processus de réponse rationalisés et standardisés.
- Réduction de l'impact des incidents sur les services et les clients.
- Résolution accélérée des problèmes.
- Amélioration continue des processus opérationnels.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élévé

Directives d'implémentation

Le respect de cette bonne pratique signifie que vous suivez les événements de charge de travail. Vous disposez de processus pour gérer les incidents et les problèmes. Les processus sont documentés, partagés et mis à jour fréquemment. Les problèmes sont identifiés, hiérarchisés et résolus.

Comprendre les événements, les incidents et les problèmes

- Événement : un événement est une observation d'action, d'occurrence ou de modification d'un état. Les événements peuvent être planifiés ou imprévus et peuvent avoir une origine interne ou externe à la charge de travail.

- Incidents : les incidents sont des événements qui nécessitent une réponse. Il peut notamment s'agir d'interruptions imprévues ou de dégradations de la qualité du service. Les incidents sont des perturbations qui nécessitent une attention immédiate pour rétablir le fonctionnement normal de la charge de travail.
- Problèmes : les problèmes sont les causes sous-jacentes d'un ou de plusieurs incidents. L'identification et la résolution des problèmes impliquent d'étudier plus en profondeur les incidents afin d'éviter qu'ils se reproduisent.

Étapes d'implémentation

Événements

1. Surveiller des événements :
   - Mettez en œuvre l'observabilité et utilisez l'observabilité de la charge de travail.
   - Les actions de surveillance effectuées par un utilisateur, un rôle ou un AWS service sont enregistrées sous forme d'événements dans AWS CloudTrail.
   - Répondez aux changements opérationnels de vos applications en temps réel avec Amazon EventBridge.
   - Évaluez, surveillez et enregistrez en permanence les modifications de configuration des ressources avec AWS Config.
2. Créez des processus :
   - Élaborez un processus pour évaluer quels événements sont importants et nécessitent une surveillance. Pour ce faire, il faut fixer des seuils et des paramètres pour les activités normales et anormales.
   - Déterminez les critères permettant de transformer un événement en incident. Cette évaluation peut être basée sur la gravité, l'impact sur les utilisateurs ou un écart par rapport au comportement attendu.
   - Passez régulièrement en revue les processus de surveillance et de réponse aux événements. Il s'agit notamment d'analyser les incidents passés, d'ajuster les seuils et d'affiner les mécanismes d'alerte.

Incidents

1. Intervenir en cas d'incident :

- Utilisez les informations issues des outils d'observabilité pour identifier rapidement les incidents et y répondre.
- Mettre en place un centre d'opérations AWS Systems Manager pour regrouper, organiser et hiérarchiser les éléments opérationnels et les incidents.
- Utilisez des services tels qu'Amazon AWS X-Raypour une analyse CloudWatch et un dépannage approfondis.
- Envisagez AWS Managed Services (AMS) pour une meilleure gestion des incidents, en tirant parti de ses capacités proactives, préventives et de détection. AMSétend le support opérationnel avec des services tels que la surveillance, la détection et la réponse aux incidents, ainsi que la gestion de la sécurité.
- Les clients du support aux entreprises peuvent utiliser la détection et la réponse aux incidents AWS, qui fournissent une surveillance proactive continue et une gestion des incidents pour les charges de travail de production.

2. Créez un processus de gestion des incidents :
   - Établissez un processus structuré de gestion des incidents, comprenant des rôles clairs, des protocoles de communication et des étapes de résolution.
   - Intégrez la gestion des incidents à des outils tels que AWS Chatbot pour une réponse et une coordination efficaces.
   - Classez les incidents par ordre de gravité, avec des plans d'intervention en cas d'incidents prédéfinis pour chaque catégorie.
3. Apprenez et améliorez vos processus :
   - Effectuez une analyse post-incident pour comprendre les causes profondes et l'efficacité de l'intervention.
   - Mettez à jour et améliorez en continu les plans de réponse en fonction des examens et de l'évolution des pratiques.
   - Documentez et partagez les leçons apprises entre les équipes afin d'améliorer la résilience opérationnelle.
   - Les clients du support aux entreprises peuvent demander l'atelier de gestion des incidents auprès de leur responsable de compte technique. Le présent atelier guidé vous permet d'évaluer votre plan d'intervention en cas d'incident et d'identifier les points à améliorer.

Problèmes

1. Identifiez les problèmes :

- Utilisez les données relatives aux incidents précédents pour identifier des modèles récurrents susceptibles d’indiquer des problèmes systémiques plus profonds.
- Tirez parti d'outils tels AWS CloudTrail CloudWatchqu'Amazon pour analyser les tendances et découvrir les problèmes sous-jacents.
- Mobilisez des équipes interfonctionnelles, y compris les services des opérations et du développement, ainsi que les unités commerciales, afin d’obtenir des points de vue diversifiés sur les causes profondes.

2. Créez un processus de gestion des problèmes :
   - Développez un processus structuré pour la gestion des problèmes, en mettant l’accent sur des solutions à long terme plutôt que sur des correctifs rapides.
   - Intégrez des techniques d'analyse des causes profondes (RCA) pour étudier et comprendre les causes sous-jacentes des incidents.
   - Mettez à jour les politiques, les procédures et l’infrastructure opérationnelles en fonction des résultats pour éviter tout incident.
3. Continuez à améliorer vos processus :
   - Favorisez une culture d’apprentissage et d’amélioration continus, en incitant les équipes à identifier et à résoudre les problèmes potentiels de manière proactive.
   - Passez régulièrement en revue et révisez les processus et les outils de gestion des problèmes afin de les aligner sur l’évolution des environnements commerciaux et technologiques.
   - Partagez des informations et des bonnes pratiques au sein de l’organisation afin de créer un environnement opérationnel plus résilient et plus efficace.
4. Engagez-vous AWS Support :
   - Utilisez des ressources d' AWS assistance AWS Trusted Advisor, telles que des conseils proactifs et des recommandations d'optimisation.
   - Les clients du support aux entreprises peuvent accéder à des programmes spécialisés tels que AWS Countdown pour obtenir une assistance lors d’événements critiques.

Niveau d’effort du plan d’implémentation : faible

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identifier les indicateurs de performance clés

- OPS04-BP02 Implémenter la télémétrie des applications
- OPS07-BP03 Utiliser des runbooks pour exécuter des procédures
- OPS07-BP04 Utiliser des playbooks pour étudier les problèmes
- OPS08-BP01 Analyser les métriques de charge de travail
- OPS11-BP02 Réaliser une analyse post-incident

Documents connexes :

- Guide d’intervention en cas d’incident de sécuritéAWS
- AWS Détection et réponse aux incidents
- AWS Cadre d'adoption du cloud : point de vue des opérations - Gestion des incidents et des problèmes
- La gestion des incidents à l'ère de DevOps et SRE
- PagerDuty - Qu'est-ce que la gestion des incidents ?

Vidéos connexes :

- Les meilleurs conseils de réponse aux incidents de AWS
- AWS re:Invent 2022 - The Amazon Builders' Library : 25 ans d'excellence opérationnelle d'Amazon
- AWS re:Invent 2022 - Détection et réponse aux AWS incidents (01) SUP2
- Présentation d'Incident Manager de AWS Systems Manager

Exemples connexes :

- AWS Services proactifs — Atelier sur la gestion des incidents
- Comment automatiser la réponse aux incidents avec PagerDuty et AWS Systems Manager Incident Manager
- Impliquez les intervenants en cas d'incident grâce aux horaires d'appel dans AWS Systems Manager Incident Manager
- Améliorez la visibilité et la collaboration lors de la gestion des incidents dans AWS Systems Manager Incident Manager
- Rapports d'incidents et demandes de service dans AMS

Services connexes :

- Amazon EventBridge

OPS10-BP02 Disposer d'un processus par alerte

Il est essentiel d'établir un processus clair et défini pour chaque alerte de votre système afin de garantir une gestion efficace et efficiente des incidents. Cette pratique garantit que chaque alerte entraîne une réponse spécifique et exploitable, améliorant ainsi la fiabilité et la réactivité de vos opérations.

Résultat escompté : chaque alerte déclenche un plan de réponse spécifique et bien défini. Dans la mesure du possible, les réponses sont automatisées, avec une propriété clairement établie et une procédure de remontée définie. Les alertes sont liées à une base de connaissances actualisée afin que chaque opérateur puisse réagir de manière cohérente et efficace. Les réponses sont rapides et uniformes à tous les niveaux, ce qui améliore l'efficacité et la fiabilité opérationnelles.

Anti-modèles courants :

- Les alertes n'ont pas de processus de réponse prédéfini, ce qui entraîne des résolutions improvisées et différées.
- En raison de la surcharge d'alertes, celles qui sont importantes sont ignorées.
- Les alertes ne sont pas traitées de manière cohérente en raison de l'absence de définition claire de la propriété et des responsabilités.

Avantages liés au respect de cette bonne pratique :

- Réduction de la lassitude liée aux alertes en ne déclenchant que des alertes exploitables.
- Diminution du délai moyen de résolution (MTTR) des problèmes opérationnels.
- Diminution du délai moyen d'investigation (MTTI), ce qui contribue à réduire le MTTR.
- Capacité accrue à mettre à l'échelle les réponses opérationnelles.
- Amélioration de la cohérence et de la fiabilité dans la gestion des événements opérationnels.

Par exemple, vous disposez d'un processus défini pour les événements AWS Health pour les comptes critiques, y compris les alarmes d'application, les problèmes opérationnels et les événements planifiés du cycle de vie (comme la mise à jour des versions d'Amazon EKS avant

la mise à jour automatique des clusters), et vous donnez à vos équipes la possibilité de surveiller activement ces événements, de les communiquer et d'y répondre. Ces actions vous aident à prévenir les interruptions de service causées par des modifications côté AWS ou à les atténuer plus rapidement en cas de problèmes inattendus.

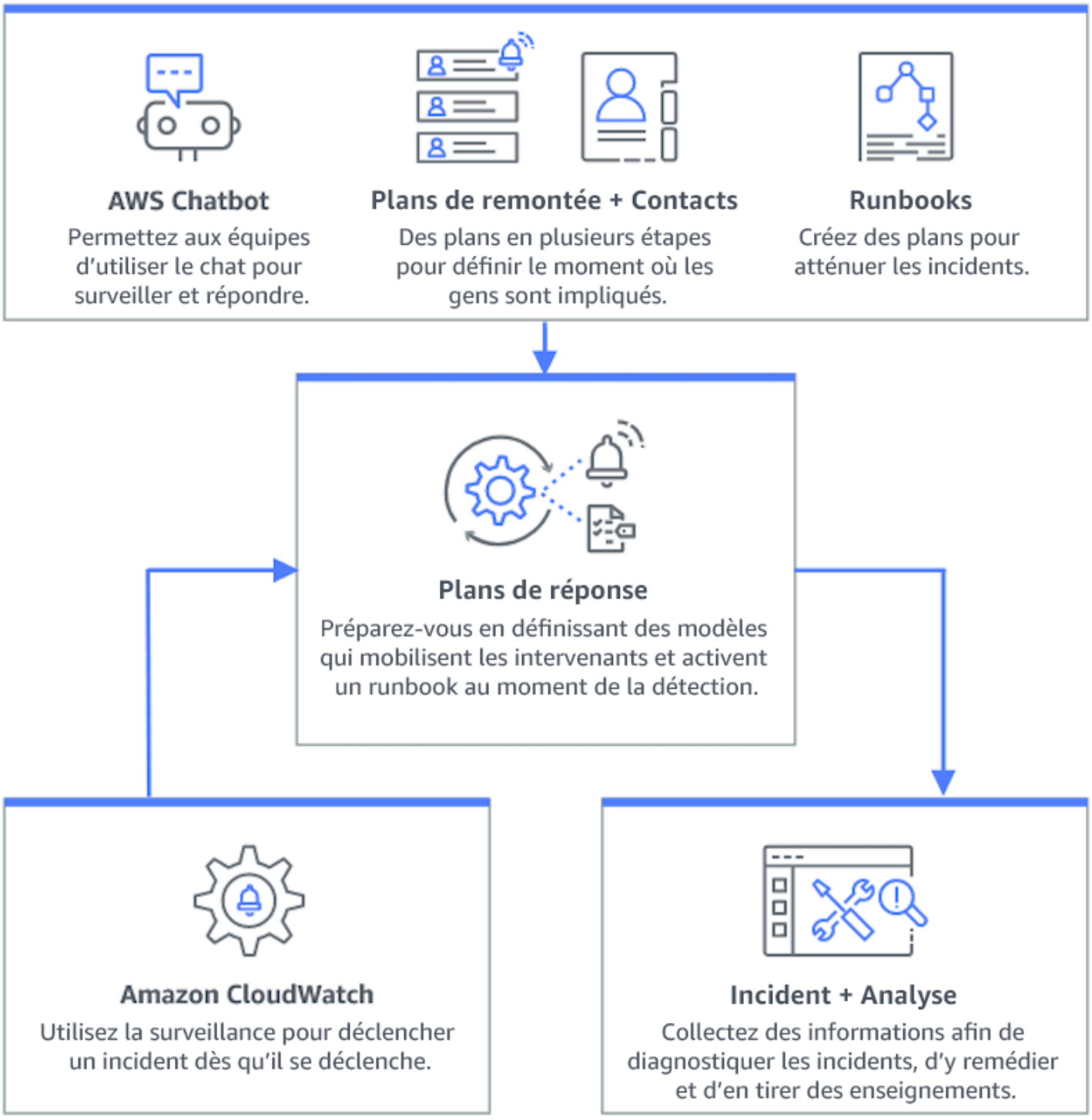
Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Pour disposer d'un processus par alerte, il est nécessaire d'établir un plan de réponse clair pour chaque alerte, d'automatiser les réponses dans la mesure du possible et d'améliorer continuellement ces processus en fonction des commentaires opérationnels et de l'évolution des exigences.

Étapes d'implémentation

Le schéma suivant illustre le flux de travail de gestion des incidents dans AWS Systems Manager Incident Manager. Il est conçu pour répondre rapidement aux problèmes opérationnels en créant automatiquement des incidents en réponse à des événements spécifiques provenant d'Amazon CloudWatch ou d'Amazon EventBridge. Lorsqu'un incident est créé, automatiquement ou manuellement, Incident Manager centralise la gestion de l'incident, organise les informations pertinentes sur les ressources AWS et lance des plans de réponse prédéfinis. Il s'agit entre autres de l'exécution de runbooks Automation pour une action immédiate, ainsi que de la création d'un élément de travail opérationnel parent dans OpsCenter afin de suivre les tâches et les analyses associées. Ce processus rationalisé accélère et coordonne la réponse aux incidents dans l'ensemble de votre environnement AWS.



1. Utiliser des alarmes composites : créez des alarmes composites dans CloudWatch pour regrouper les alarmes associées, réduire le bruit et permettre des réponses plus pertinentes.
2. Surveiller les événements AWS Health à l'aide des règles Amazon EventBridge : surveillez ou intégrez-les par programmation à l'API AWS Health pour automatiser les actions lorsque vous

recevez des événements AWS Health. Il peut s'agir d'actions générales, telles que l'envoi de tous les messages relatifs aux événements du cycle de vie planifiés vers une interface de discussion, ou d'actions spécifiques, telles que le lancement d'un flux de travail dans un outil de gestion des services informatiques.

   a. Configuration des notifications utilisateur AWS pour AWS Health

3. Intégrer les alarmes Amazon CloudWatch avec Incident Manager : configurez les alarmes CloudWatch pour créer automatiquement des incidents dans AWS Systems Manager Incident Manager.
4. Intégrer Amazon EventBridge à Incident Manager : créez des règles EventBridge pour réagir aux événements et créer des incidents à l'aide de plans d'intervention définis.
5. Préparez-vous aux incidents dans Incident Manager :
   - Établissez des plans d'intervention détaillés dans Incident Manager pour chaque type d'alerte.
   - Établissez des canaux de discussion par le biais de AWS Chatbot connecté aux plans d'intervention dans Incident Manager, afin de faciliter la communication en temps réel lors d'incidents sur des plateformes telles que Slack, Microsoft Teams et Amazon Chime.
   - Intégrez les runbooks d'automatisation de la gestion des systèmes dans Incident Manager pour générer des interventions automatisées en cas d'incidents.

Ressources

Bonnes pratiques associées :

- OPS04-BP01 Identifier les indicateurs de performance clés
- OPS08-BP04 Créez des alertes exploitables

Documents connexes :

- AWS Cloud Adoption Framework : Operations Perspective – Gestion des incidents et des problèmes
- Utilisation d'alarmes Amazon CloudWatch
- Configuration de AWS Systems Manager Incident Manager
- Préparation aux incidents dans Incident Manager :

Vidéos connexes :

- Les meilleurs conseils de AWS en matière d'intervention en cas d'incident
- re:Invent 2023 | Manage resource lifecycle events at scale with AWS Health

Exemples connexes :

- AWS Ateliers – AWS Systems Manager Incident Manager – Automatiser les réponses aux événements de sécurité

OPS10-BP03 Prioriser les événements opérationnels en fonction de leur impact commercial

Il est essentiel de réagir rapidement aux événements opérationnels, mais tous les événements ne sont pas identiques. Lorsque vous établissez des priorités en fonction de l'impact sur l'entreprise, vous donnez également la priorité aux événements susceptibles d'avoir des conséquences importantes. Ces événements peuvent être liés à la sécurité, à des pertes financières, à des violations de la réglementation ou à des atteintes à la réputation.

Résultat escompté : les réponses aux événements opérationnels sont classées par ordre de priorité en fonction de leur impact potentiel sur les opérations et les objectifs de l'entreprise. Des réponses efficientes et efficaces peuvent ainsi être mises en place.

Anti-modèles courants :

- Chaque événement est traité avec le même niveau d'urgence, ce qui entraîne de la confusion et des retards dans la résolution des problèmes critiques.
- Vous ne faites pas la distinction entre les événements à fort et à faible impact, ce qui entraîne une mauvaise allocation des ressources.
- Votre organisation ne dispose pas d'un cadre de priorisation clair, ce qui entraîne des réponses incohérentes aux événements opérationnels.
- Les événements sont priorisés en fonction de leur ordre de signalement, plutôt que de leur impact sur les résultats de l'entreprise.

Avantages liés au respect de cette bonne pratique :

- Garantit que les fonctions critiques de l'entreprise sont traitées en premier lieu, minimisant ainsi les dommages potentiels.
- Améliore l'allocation des ressources lors de plusieurs événements simultanés.

- Améliore la capacité de l’organisation à préserver la confiance et à répondre aux exigences réglementaires.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Lorsque plusieurs événements opérationnels ont lieu simultanément, il est essentiel d’adopter une approche structurée de la priorisation en fonction de l’impact et de l’urgence. Cette approche vous permet de prendre des décisions éclairées, d’orienter les efforts vers les domaines où ils sont le plus nécessaires et d’atténuer les risques pour la continuité des activités.

Étapes d’implémentation

1. Mesurer l’impact : élaborez un système de classification pour évaluer la gravité des événements en fonction de leur impact potentiel sur les opérations et les objectifs de l’entreprise. L’exemple suivant illustre les différentes catégories d’impact :

Niveau d’impact	Description
Élevé	Affecte de nombreux employés ou clients, a un impact financier élevé, porte atteinte à la réputation de façon importante ou est lié à des blessures.
Moyen	Affecte un groupe d’employés ou de clients, a un impact financier modéré ou porte atteinte à la réputation de façon modérée.
Faible	Affecte le personnel ou les clients, a un impact financier limité ou porte peu atteinte à la réputation de l’entreprise.

2. Évaluer l'urgence : définissez les niveaux d'urgence pour déterminer la rapidité avec laquelle un événement nécessite une réponse, en tenant compte de facteurs tels que la sécurité, les implications financières et les accords de niveau de service ()SLAs. L’exemple suivant illustre les catégories d’urgence :

Niveau d’urgence	Description
Élevé	Augmentation exponentielle des dégâts, impact sur le travail urgent, escalade imminente ou VIP utilisateurs ou groupes affectés.
Moyen	Les dégâts augmentent au fil du temps, ou un seul VIP utilisateur ou un groupe est affecté.
Faible	Les dommages marginaux augmentent au fil du temps ou les non-time-sensitive travaux sont affectés.

3. Créez une matrice de priorisation :
    - Utilisez une matrice pour associer l’impact et l’urgence, en attribuant des niveaux de priorité à différentes combinaisons.
    - Rendez la matrice accessible et compréhensible par tous les membres de l’équipe responsables des réponses aux événements opérationnels.
    - L’exemple de matrice suivant affiche la gravité des incidents en fonction de leur urgence et de leur impact :

Urgence et impact	Élevé	Moyen	Faible
Élevé	Critique	Urgent	Élevé
Moyen	Urgent	Élevé	Normal
Faible	Élevé	Normal	Faible

4. Former et communiquer : formez les équipes de réponse à la matrice de priorisation et à l’importance de la suivre lors d’un événement. Communiquez le processus de priorisation à toutes les parties prenantes afin de définir des attentes claires.
5. Intégrez la matrice à la gestion des réponses aux incidents :
    - Intégrez la matrice de priorisation à vos plans et outils de réponse aux incidents.

- Automatisez la classification et la hiérarchisation des événements dans la mesure du possible afin d'accélérer les temps de réponse.
- Les clients du support aux entreprises peuvent utiliser la détection et la réponse aux incidents AWS, qui fournissent une surveillance proactive continue et une gestion des incidents pour les charges de travail de production.

6. Examiner et adapter : passez régulièrement en revue l'efficacité du processus de priorisation et apportez des ajustements en fonction des commentaires et de l'évolution de l'environnement métier.

Ressources

Bonnes pratiques associées :

- OPS03-BP03 L'escalade est encouragée
- OPS08-BP04 Créez des alertes exploitables
- OPS09-BP01 Mesurer les objectifs opérationnels et les KPI à l'aide de métriques

Documents connexes :

- Atlassian – Understanding incident severity levels
- IT Process Map - Checklist Incident Priority

OPS10-BP04 Définir les chemins d'escalade

Définissez des procédures de remontée claires dans vos protocoles de réponse aux incidents afin de faciliter une action rapide et efficace. Cela inclut la spécification des invites d'escalade, le détail du processus d'escalade et l'approbation préalable des actions pour accélérer la prise de décision et réduire le délai moyen de résolution (). MTTR

Résultat escompté : un processus structuré et efficace qui transmet les incidents au personnel approprié, minimisant ainsi les temps de réponse et l'impact.

Anti-modèles courants :

- Le manque de clarté des procédures de récupération entraîne des interventions improvisées lors d'incidents critiques.

- L’absence d’autorisations et de propriétaires définis entraîne des retards lorsqu’une action urgente est nécessaire.
- Les parties prenantes et les clients ne sont pas informés conformément aux attentes.
- Les décisions importantes sont reportées.

Avantages liés au respect de cette bonne pratique :

- Réponse rationalisée aux incidents grâce à des procédures de remontée prédéfinies.
- Réduction des temps d’arrêt grâce à des actions préapprouvées et à la définition claire d’un propriétaire.
- Meilleure allocation des ressources et ajustements du niveau d’assistance en fonction de la gravité de l’incident.
- Meilleure communication avec les parties prenantes et les clients.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Des voies d'escalade correctement définies sont cruciales pour une réponse rapide aux incidents. AWS Systems Manager Incident Manager prend en charge la mise en place de plans d'escalade structurés et de calendriers d'astreinte, qui alertent le personnel approprié afin qu'il soit prêt à agir en cas d'incident.

Étapes d’implémentation

1. Configurer des invites d'escalade : configurez des CloudWatch alarmes pour créer un incident dans AWS Systems Manager Incident Manager.
2. Configurez des horaires d’astreinte : créez des horaires d’astreinte dans Incident Manager qui correspondent à vos trajectoires d’escalade. Dotez le personnel d’astreinte des autorisations et des outils nécessaires afin de lui permettre d’agir rapidement.
3. Détaillez les procédures de remontée :
   - Déterminez les conditions spécifiques dans lesquelles un incident doit faire l’objet d’une remontée.
   - Créez des plans d’escalade dans Incident Manager.
   - Les canaux de remontée doivent inclure un contact ou un calendrier d’astreinte.

- Définissez les rôles et les responsabilités de l'équipe à chaque niveau de la remontée.

4. Approuver au préalable les mesures d'atténuation : collaborez avec les décisionnaires pour approuver au préalable les actions associées aux scénarios prévus. Utilisez les runbooks Systems Manager Automation intégrés à Incident Manager pour accélérer la résolution des incidents.
5. Préciser la propriété : identifiez clairement les propriétaires internes pour chaque étape de la procédure de remontée.
6. Détaillez les remontées par des tiers :
   - Documentez les accords de niveau de service tiers (SLAs) et alignez-les sur les objectifs internes.
   - Définissez des protocoles clairs pour la communication avec les fournisseurs lors d'incidents.
   - Intégrez les contacts des fournisseurs dans les outils de gestion des incidents pour un accès direct.
   - Effectuez régulièrement des exercices qui incluent des scénarios de réponse par des tiers.
   - Documentez les informations relatives à la remontée fournisseurs et veillez à ce qu'elles soient facilement accessibles.
7. Former et répéter les plans d'escalade : formez votre équipe à la procédure de remontée et organisez régulièrement des exercices de réponse aux incidents ou des journées de jeu. Les clients du support aux entreprises peuvent demander l'atelier de gestion des incidents auprès de leur responsable de compte technique.
8. Améliorer sans cesse : vérifiez régulièrement l'efficacité de vos procédures de remontée. Mettez à jour vos procédures en fonction des leçons tirées des analyses post-mortem des incidents et des commentaires fournis en continu.

Niveau d'effort du plan d'implémentation : modéré

Ressources

Bonnes pratiques associées :

- OPS08-BP04 Créez des alertes exploitables
- OPS10-BP02 Disposer d'un processus par alerte
- OPS11-BP02 Réaliser une analyse post-incident

Documents connexes :

- AWS Systems Manager Incident Manager Plans d'escalade
- Utilisation des horaires d’astreinte dans Incident Manager
- Création et gestion des Runbooks
- Gestion temporaire des accès surélevés avec AWS IAM Identity Center
- Atlassian - Politiques d’escalade pour une gestion efficace des incidents

OPS10-BP05 Définir un plan de communication client pour les événements ayant un impact sur le service

Il est essentiel de mettre en place une communication efficace lors d’événements ayant un impact sur le service afin de préserver la confiance des clients et la transparence dont vous faites preuve à leur égard. Un plan de communication bien défini permet à votre organisation de partager rapidement et clairement des informations, à la fois en interne et en externe, lors d’incidents.

Résultat souhaité :

- Un plan de communication robuste qui informe efficacement les clients et les parties prenantes lors d’événements ayant un impact sur le service.
- Transparence dans la communication pour renforcer la confiance et réduire l’anxiété des clients.
- Minimiser l’impact des événements ayant un impact sur le service du point de vue de l’expérience client et des opérations métier.

Anti-modèles courants :

- Une communication inadéquate ou retardée entraîne de la confusion et de l’insatisfaction chez les clients.
- Les messages trop techniques ou trop vagues ne reflètent pas l’impact réel sur les utilisateurs.
- Il n’existe pas de stratégie de communication prédéfinie, ce qui entraîne des messages incohérents et réactifs.

Avantages liés au respect de cette bonne pratique :

- Confiance et satisfaction accrues des clients grâce à une communication proactive et claire.
- Réduction de la charge de travail des équipes d’assistance en répondant de manière préventive aux préoccupations des clients.

- Amélioration de la capacité à gérer les incidents et à récupérer de manière efficace.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

La création d’un plan de communication complet pour les événements ayant un impact sur les services implique plusieurs facettes, du choix des canaux appropriés à l’élaboration du message et du ton adéquats. Le plan doit être adaptable, doté d’une capacité de mise à l’échelle et pouvoir répondre à différents scénarios de panne.

Étapes d’implémentation

1. Définissez les rôles et les responsabilités :
   - Désignez un responsable des incidents majeurs qui sera chargé de superviser les activités de réponse aux incidents.
   - Désignez un responsable des communications chargé de coordonner toutes les communications externes et internes.
   - Incluez le responsable de l’assistance afin d’établir une communication cohérente par le biais de tickets d’assistance.
2. Identifiez les canaux de communication : sélectionnez des canaux tels que le chat sur le lieu de travail, le courrier électroniqueSMS, les réseaux sociaux, les notifications intégrées à l'application et les pages de statut. Ces canaux doivent être résilients et capables de fonctionner de manière indépendante lors d’événements ayant un impact sur le service.
3. Communiquez rapidement, clairement et régulièrement avec les clients :
   - Élaborez des modèles pour divers scénarios de détérioration des services, en mettant l’accent sur la simplicité et les détails essentiels. Incluez des informations sur la perturbation du service, le délai de résolution prévu et l’impact.
   - Amazon Pinpoint vous permet d’alerter les clients à l’aide de notifications push, de notifications in-app, d’e-mails, de messages texte, de messages vocaux et de messages sur des canaux personnalisés.
   - Utilisez Amazon Simple Notification Service (AmazonSNS) pour alerter les abonnés par programme ou par e-mail, notifications push mobiles et SMS.
   - Communiquez le statut par le biais de tableaux de bord en partageant publiquement un CloudWatch tableau de bord Amazon.
   - Encouragez l’engagement sur les réseaux sociaux :

- Surveillez activement les réseaux sociaux pour comprendre le sentiment des clients.
- Publiez sur les plateformes de réseaux sociaux pour les mises à jour publiques et un engagement communautaire.
- Préparez des modèles pour une communication cohérente et claire sur les réseaux sociaux.

4. Coordonner la communication interne : mettez en œuvre des protocoles internes à l'aide d'outils tels que AWS Chatbot la coordination et la communication des équipes. Utilisez des CloudWatch tableaux de bord pour communiquer le statut.
5. Orchestrez la communication à l’aide d’outils et de services dédiés :
   - Utilisez AWS Systems Manager Incident Manager with AWS Chatbot pour configurer des canaux de discussion dédiés pour une communication interne et une coordination en temps réel lors d'incidents.
   - Utilisez AWS Systems Manager Incident Manager des runbooks pour automatiser les notifications aux clients via Amazon Pinpoint, SNS Amazon ou des outils tiers tels que les plateformes de réseaux sociaux lors d'incidents.
   - Intégrez des flux de travail d’approbation dans les runbooks pour, si nécessaire, examiner et autoriser toutes les communications externes avant leur envoi.
6. Entraînez-vous et améliorez les processus :
   - Organisez une formation sur l’utilisation des outils et des stratégies de communication. Donnez aux équipes les moyens de prendre des décisions rapidement en cas d’incident.
   - Testez le plan de communication lors d’exercices réguliers ou de journées de jeu. Utilisez ces tests pour affiner les messages et évaluer l’efficacité des canaux.
   - Mettez en œuvre des mécanismes de commentaires pour évaluer l’efficacité de la communication lors d’incidents. Faites évoluer continuellement le plan de communication en fonction des commentaires et de l’évolution des besoins.

Niveau d’effort du plan d’implémentation : élevé

Ressources

Bonnes pratiques associées :

- OPS07-BP03 Utiliser des runbooks pour exécuter des procédures
- OPS10-BP06 Communiquer le statut par le biais de tableaux de bord
- OPS11-BP02 Réaliser une analyse post-incident

Documents connexes :

- Atlassian – Bonnes pratiques en matière de communication sur les incidents
- Atlassian – Comment rédiger une bonne mise à jour de statut
- PagerDuty - Guide de communication en cas d'incident

Vidéos connexes :

- Atlassian – Créez votre propre plan de communication en cas d’incident : modèles d’incidents

Exemples connexes :

- AWS Health Tableau de bord
- Exemples de mises à jour AWS de statut

OPS10-BP06 Communiquer le statut par le biais de tableaux de bord

Utilisez les tableaux de bord comme outil stratégique pour communiquer l’état opérationnel en temps réel et les métriques clés à différents publics, y compris aux équipes techniques internes, à la direction et aux clients. Ces tableaux de bord offrent une représentation visuelle centralisée de l’intégrité du système et des performances de l’entreprise, améliorant ainsi la transparence et l’efficacité de la prise de décision.

Résultat escompté :

- Vos tableaux de bord fournissent une vue complète des métriques système et métier pour les différentes parties prenantes.
- Les parties prenantes peuvent accéder de manière proactive aux informations opérationnelles, ce qui réduit la nécessité d’effectuer fréquemment des demandes de statut.
- La prise de décision en temps réel est améliorée pendant les opérations normales et les incidents.

Anti-modèles courants :

- Les ingénieurs participant à un appel de gestion des incidents ont besoin de mises à jour du statut pour être opérationnels.
- Faire confiance à des rapports manuels pour la gestion, ce qui entraîne des retards et des inexactitudes potentielles.

- Les équipes opérationnelles sont fréquemment interrompues pour des mises à jour de statut lors d’incidents.

Avantages liés au respect de cette bonne pratique :

- Donne aux parties prenantes un accès immédiat aux informations critiques, favorisant ainsi la prise de décisions réfléchies.
- Réduit les inefficacités opérationnelles en minimisant les rapports manuels et en limitant la fréquence des demandes de statut.
- Améliore la transparence et la confiance grâce à une visibilité en temps réel des performances du système et des métriques métier.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Les tableaux de bord communiquent efficacement le statut de vos métriques système et métier, et peuvent être adaptés aux besoins des différents groupes d’audience. Des outils tels que CloudWatch les tableaux de bord Amazon et Amazon vous QuickSight aident à créer des tableaux de bord interactifs en temps réel pour la surveillance du système et l'informatique décisionnelle.

Étapes d’implémentation

1. Identification des besoins des parties prenantes : déterminez les besoins d’informations spécifiques des différents groupes d’audience, tels que les équipes techniques, la direction et les clients.
2. Choisissez les bons outils : sélectionnez les outils appropriés, tels que les CloudWatch tableaux de bord Amazon pour la surveillance du système et Amazon QuickSight pour les informations commerciales interactives.
3. Conception de tableaux de bord efficaces :
    - Concevez des tableaux de bord afin de présenter clairement les indicateurs pertinents et KPIs de vous assurer qu'ils sont compréhensibles et exploitables.
    - Intégrez des vues aux niveaux du système et de l’entreprise selon les besoins.
    - Incluez des tableaux de bord globaux (pour les vues d’ensemble) et détaillés (pour une analyse approfondie).

- Intégrez des alarmes automatisées dans les tableaux de bord pour mettre en évidence les problèmes critiques.
- Annotez les tableaux de bord avec des métriques, des seuils et des objectifs importants pour une visibilité immédiate.

4. Intégration des sources de données :
   - Utilisez Amazon CloudWatch pour agréger et afficher les métriques de différents AWS services et interroger les métriques provenant d'autres sources de données, afin de créer une vue unifiée de l'état de santé de votre système et des indicateurs commerciaux.
   - Utilisez des fonctionnalités telles que CloudWatch Logs Insights pour interroger et visualiser les données des journaux provenant de différents services et applications.
5. Fourniture d’un accès en libre-service :
   - Partagez CloudWatch des tableaux de bord avec les parties prenantes concernées pour accéder aux informations en libre-service à l'aide des fonctionnalités de partage de tableaux de bord.
   - Assurez-vous que les tableaux de bord sont facilement accessibles et fournissent des up-to-date informations en temps réel.
6. Mise à jour et affinage réguliers :
   - Mettez à jour et affinez continuellement les tableaux de bord pour les adapter à l’évolution des besoins de l’entreprise et aux commentaires des parties prenantes.
   - Passez régulièrement en revue les tableaux de bord afin qu’ils restent pertinents et efficaces pour transmettre les informations nécessaires.

Ressources

Bonnes pratiques associées :

- OPS08-BP05 Création de tableaux de bord

Documents connexes :

- Création de tableaux de bord pour une visibilité opérationnelle
- Utilisation des tableaux de CloudWatch bord Amazon
- Création de tableaux de bord flexibles avec des variables de tableau de bord
- Partage de CloudWatch tableaux de bord

- Interrogation de métriques d'autres sources de données
- Ajouter un widget personnalisé à un CloudWatch tableau de bord

Exemples connexes :

- Un atelier sur l'observabilité – Tableaux de bord

OPS10-BP07 Automatiser les réponses aux événements

L'automatisation des réponses aux événements est essentielle pour une gestion opérationnelle rapide, cohérente et sans erreur. Créez des processus rationalisés et utilisez des outils pour gérer et répondre automatiquement aux événements, en minimisant les interventions manuelles et en améliorant l'efficacité opérationnelle.

Résultat escompté :

- Réduction des erreurs humaines et accélération des temps de résolution grâce à l'automatisation.
- Gestion cohérente et fiable des événements opérationnels.
- Amélioration de l'efficacité opérationnelle et de la fiabilité du système.

Anti-modèles courants :

- La gestion manuelle des événements entraîne des retards et des erreurs.
- L'automatisation est négligée pour les tâches critiques et répétitives.
- Les tâches manuelles répétitives entraînent une lassitude liée aux alertes et peuvent nuire à la détection de problèmes critiques.

Avantages liés au respect de cette bonne pratique :

- Réponses accélérées aux événements, réduisant ainsi les temps d'arrêt du système.
- Des opérations fiables avec une gestion automatisée et cohérente des événements.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d’implémentation

Intégrez l’automatisation pour créer des flux de travail opérationnels efficaces et minimiser les interventions manuelles.

Étapes d’implémentation

1. Identification des opportunités d’automatisation : déterminez les tâches répétitives à automatiser, telles que la résolution des problèmes, l’enrichissement des tickets, la gestion des capacités, la mise à l’échelle, les déploiements et les tests.
2. Identification des invites d’automatisation :
   - Évaluez et définissez des conditions ou des métriques spécifiques qui déclenchent des réponses automatisées à l'aide des actions CloudWatch d'alarme Amazon.
   - Utilisez Amazon EventBridge pour répondre aux événements liés aux AWS services, aux charges de travail personnalisées et aux applications SaaS.
   - Tenez compte des événements d'initiation tels que des entrées de journal spécifiques, des seuils de mesures de performance ou des changements d'état des AWS ressources.
3. Mise en œuvre d’une automatisation pilotée par les événements :
   - Utilisez les runbooks AWS Systems Manager d'automatisation pour simplifier les tâches de maintenance, de déploiement et de correction.
   - La création d'incidents dans Incident Manager permet de collecter et d'ajouter automatiquement des informations sur les AWS ressources impliquées dans l'incident.
   - Surveillez les quotas de manière proactive à l’aide de Quota Monitor pour AWS.
   - Ajustez automatiquement la capacité avec AWS Auto Scaling pour maintenir la disponibilité et les performances.
   - Automatisez les pipelines de développement avec Amazon CodeCatalyst.
   - Testez la fumée ou surveillez en permanence les terminaux à APIs l'aide d'une surveillance synthétique.
4. Atténuation des risques grâce à l’automatisation :
   - Utilisez des réponses de sécurité automatisées pour gérer rapidement les risques.
   - Utilisez AWS Systems Manager State Manager pour réduire la dérive de configuration.
   - Corrigez les ressources non conformes avec. AWS Config Rules

~~Niveau d’effort du plan d’implémentation : élevé~~

Ressources

Bonnes pratiques associées :

- OPS08-BP04 Créez des alertes exploitables
- OPS10-BP02 Disposer d’un processus par alerte

Documents connexes :

- Utilisation des runbooks d’automatisation Systems Manager avec Incident Manager
- Création d’incidents dans Incident Manager
- AWS quotas de service
- Contrôle de l’utilisation des ressources et envoi de notifications lorsque les quotas sont atteints
- AWS Auto Scaling
- Qu'est-ce qu'Amazon CodeCatalyst ?
- Utilisation des CloudWatch alarmes Amazon
- Utilisation des actions CloudWatch d'alarme Amazon
- Corriger les ressources non conformes avec AWS Config Rules
- Création de métriques à partir d’événements du journal à l’aide de filtres
- Gestionnaire d’états AWS Systems Manager

Vidéos connexes :

- Créez des runbooks d'automatisation avec AWS Systems Manager
- Comment automatiser les opérations informatiques sur AWS
- AWS Security Hub règles d'automatisation
- Démarrez rapidement votre projet logiciel avec les CodeCatalyst plans Amazon

Exemples connexes :

- CodeCatalyst Tutoriel Amazon : Création d'un projet avec le plan d'application Web moderne à trois niveaux
- Un atelier sur l’observabilité
- Réaction aux incidents à l’aide d’Incident Manager

Évolution

Question

- OPS 11. Comment faire évoluer vos opérations ?

OPS 11. Comment faire évoluer vos opérations ?

Consacrez du temps et des ressources à l’amélioration incrémentielle presque continue pour contribuer à l’évolution de l’efficacité et de l’efficience de vos opérations.

Bonnes pratiques

- OPS11-BP01 Disposer d'un processus d'amélioration continue
- OPS11-BP02 Réaliser une analyse post-incident
- OPS11-BP03 Implémenter des boucles de rétroaction
- OPS11-BP04 Effectuer la gestion des connaissances
- OPS11-BP05 Définir les moteurs d'amélioration
- OPS11-BP06 Valider les informations
- OPS11-BP07 Réaliser des examens des métriques opérationnelles
- OPS11-BP08 Documenter et partager les enseignements
- OPS11-BP09 Allouez du temps pour apporter des améliorations

OPS11-BP01 Disposer d'un processus d'amélioration continue

Évaluez votre charge de travail par rapport aux bonnes pratiques d’architecture internes et externes. Réalisez des examens fréquents et intentionnels de la charge de travail. Priorisez les opportunités d’amélioration dans la cadence de développement de votre logiciel.

Résultat escompté :

- Vous analysez fréquemment votre charge de travail par rapport aux bonnes pratiques d’architecture.
- Vous accordez aux opportunités d’amélioration une priorité égale aux fonctionnalités dans votre processus de développement logiciel.

Anti-modèles courants :

- Vous n’avez pas vérifié l’architecture de votre charge de travail depuis qu’elle a été déployée il y a plusieurs années.
- Vous accordez une moindre priorité aux opportunités d’amélioration. Par rapport aux nouvelles fonctionnalités, ces opportunités restent en suspens.
- Il n’existe aucune norme pour mettre en œuvre des modifications issues des bonnes pratiques pour l’organisation.

Avantages liés au respect de cette bonne pratique :

- Votre charge de travail est limitée up-to-date aux meilleures pratiques en matière d'architecture.
- Vous faites évoluer votre charge de travail de manière intentionnelle.
- Vous pouvez tirer profit des bonnes pratiques de l’organisation pour améliorer toutes les charges de travail.
- Vous réalisez des gains marginaux qui ont un impact cumulatif, ce qui permet de gagner en efficacité.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Réalisez fréquemment un examen architectural de votre charge de travail. Utilisez les bonnes pratiques internes et externes, évaluez votre charge de travail et identifiez les opportunités d’amélioration. Priorisez les opportunités d’amélioration dans la cadence de développement de votre logiciel.

Étapes d’implémentation

1. Réalisez des examens périodiques de l’architecture de votre charge de travail de production à une fréquence convenue. Utilisez une norme architecturale documentée qui inclut les meilleures pratiques AWS spécifiques.
   a. Utilisez vos normes définies en interne pour ces évaluations. Si vous n’avez pas de norme interne, utilisez AWS Well-Architected Framework.
   b. Utilisez-le AWS Well-Architected Tool pour créer un aperçu personnalisé de vos meilleures pratiques internes et effectuer votre révision de l'architecture.
   c. Contactez votre architecte de AWS solution ou votre responsable de compte technique pour effectuer un examen guidé de votre charge de travail par Well-Architected Framework.

2. Priorisez les opportunités d’amélioration identifiées pendant la vérification au sein de votre processus de développement logiciel.

Niveau d’effort du plan d’implémentation : faible Vous pouvez utiliser le AWS Well-Architected Framework pour effectuer votre révision annuelle de l'architecture.

Ressources

Bonnes pratiques associées :

- OPS11-BP02 Réaliser une analyse post-incident
- OPS11-BP08 Documenter et partager les leçons apprises
- OPS04 Mettre en œuvre l'observabilité

Documents connexes :

- AWS Well-Architected Tool - Verres personnalisés
- AWS Livre blanc Well-Architected – Le processus de révision
- Personnalisez les critiques de Well-Architected à l'aide de lentilles personnalisées et du AWS Well-Architected Tool
- Mettre en œuvre le cycle de AWS vie de Well-Architected Custom Lens dans votre organisation

Vidéos connexes :

- Well-Architected Labs - Niveau 100 : objectifs personnalisés sur AWS Well-Architected Tool
- AWS re:INVENT 2023 - Appliquer les meilleures pratiques AWS Well-Architected au sein de votre organisation

Exemples connexes :

- AWS Well-Architected Tool

OPS11-BP02 Réaliser une analyse post-incident

Examinez les événements ayant un impact sur les clients et identifiez les facteurs contributifs et les actions préventives. Utilisez ces informations pour développer des mesures d’atténuation afin de

limiter ou d'empêcher la récurrence. Développez des procédures pour fournir des réponses rapides et efficaces. Publiez, le cas échéant, les facteurs adjuvants et les mesures correctives adaptées au public ciblé.

Résultat escompté :

- Vous avez mis en place des processus de gestion des incidents qui incluent une analyse post-incident.
- Vous avez mis en place des plans d'observabilité pour collecter des données sur les événements.
- Grâce à ces données, vous comprenez et vous collectez des métriques qui soutiennent votre processus d'analyse post-incident.
- Vous tirez des leçons des incidents pour améliorer les résultats futurs.

Anti-modèles courants :

- Vous administrez un serveur d'applications. Toutes vos séances actives sont interrompues toutes les 23 heures et 55 minutes environ. Vous avez essayé d'identifier le problème sur votre serveur d'applications. Vous pensez qu'il pourrait s'agir d'un problème de réseau, mais vous ne pouvez pas obtenir la coopération de l'équipe réseau, car elle est trop occupée pour vous aider. Vous n'avez pas de processus prédéfini à suivre pour obtenir de l'aide et collecter les informations nécessaires pour déterminer ce qui se passe.
- Vous avez subi une perte de données au sein de votre charge de travail. C'est la première fois que cela se produit et la cause n'est pas évidente. Vous décidez que ce n'est pas important, car vous pouvez recréer les données. La perte de données se reproduit plus fréquemment en affectant vos clients. Vous devez également faire face à une charge opérationnelle supplémentaire lorsque vous restaurez les données manquantes.

Avantages liés au respect de cette bonne pratique :

- Vous disposez d'un processus prédéfini pour déterminer les composants, les conditions, les actions et les événements qui ont contribué à un incident, ce qui vous permet d'identifier les possibilités d'amélioration.
- Vous utilisez les données issues de l'analyse post-incident pour apporter des améliorations.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d’implémentation

Utilisez un processus pour déterminer les facteurs adjuvants. Passez en revue tous les incidents ayant un impact sur le client. Dotez-vous d’un processus pour identifier et documenter les facteurs contributifs d’un incident afin de pouvoir mettre au point des mesures d’atténuation pour limiter ou empêcher la récurrence, et élaborez des procédures pour fournir des réponses rapides et efficaces. Communiquez les causes profondes des incidents, le cas échéant, et adaptez la communication à votre public cible. Partagez ouvertement les apprentissages au sein de votre organisation.

Étapes d’implémentation

1. Collectez des métriques telles que le changement de déploiement, le changement de configuration, l’heure de début de l’incident, l’heure d’alarme, l’heure d’engagement, l’heure de début de l’atténuation et l’heure de résolution de l’incident.
2. Décrivez les principaux moments de la chronologie pour comprendre les événements de l’incident.
3. Posez les questions suivantes :
    a. Pourriez-vous améliorer le délai de détection ?
    b. Existe-t-il des mises à jour des métriques et des alarmes qui permettraient de détecter l’incident plus rapidement ?
    c. Pouvez-vous améliorer le délai de diagnostic ?
    d. Existe-t-il des mises à jour de vos plans de réponse ou de vos plans d’escalade qui permettraient d’impliquer plus rapidement les bons intervenants ?
    e. Pouvez-vous améliorer le délai d’atténuation ?
    f. Existe-t-il des étapes du runbook ou du playbook que vous pourriez ajouter ou améliorer ?
    g. Pouvez-vous éviter que de futurs incidents se produisent ?
4. Créez des listes de contrôle et des actions. Suivez et mettez en œuvre toutes les actions.

Niveau d’effort du plan d’implémentation : faible

Ressources

Bonnes pratiques associées :

- OPS11-BP01 Disposer d'un processus d'amélioration continue
- OPS4 - Mettre en œuvre l'observabilité

Documents connexes :

- Performing a post-incident analysis in Incident Manager
- Examen de l'état de préparation opérationnelle

OPS11-BP03 Implémenter des boucles de rétroaction

Les boucles de commentaires fournissent des informations exploitables qui orientent la prise de décision. Créez des boucles de commentaires dans vos procédures et vos charges de travail. Elles vous permettent d'identifier les problèmes et les points à améliorer. Elles valident également les investissements dans les améliorations. Ces boucles de commentaires sont à la base de l'amélioration continue de votre charge de travail.

Les boucles de commentaires se répartissent en deux catégories : les commentaires immédiats et l'analyse rétrospective. Les commentaires immédiats sont collectés via l'examen des performances et des résultats des activités opérationnelles. Ces commentaires proviennent des membres de l'équipe, des clients ou de la sortie automatisée de l'activité. Les commentaires immédiats proviennent notamment de tests A/B et de la mise à disposition de nouvelles fonctionnalités, et sont essentiels à l'interruption immédiate.

Les analyses rétrospectives doivent être effectuées régulièrement pour recueillir des rétroactions concernant l'évaluation des métriques et des résultats opérationnels au fil du temps. Ces analyses rétrospectives se déroulent à la fin d'un sprint, sur une cadence, ou après des versions ou des événements majeurs. Ce type de boucle de rétroaction valide les investissements dans les opérations ou votre charge de travail. Il vous permet de mesurer la réussite et valide votre stratégie.

Résultat escompté : les commentaires immédiats et les analyses rétrospectives permettent d'apporter des améliorations. Il existe un mécanisme pour recueillir les commentaires des utilisateurs et des membres de l'équipe. Les analyses rétrospectives sont utilisées pour déterminer les tendances qui entraînent des améliorations.

Anti-modèles courants :

- Vous lancez une nouvelle fonctionnalité, mais vous n'avez aucun moyen de recevoir les commentaires des clients à ce sujet.
- Après avoir investi dans des améliorations opérationnelles, vous n'effectuez pas d'analyse rétrospective pour les valider.
- Vous recueillez les commentaires des clients, mais ne les examinez pas régulièrement.

- Les boucles de commentaires mènent à des mesures de suivi proposées, mais elles ne sont pas incluses dans le processus de développement de logiciels.
- Les clients ne reçoivent pas de commentaires sur les améliorations qu'ils ont proposées.

Avantages liés au respect de cette bonne pratique :

- Vous pouvez travailler à rebours en partant du client pour générer de nouvelles fonctionnalités.
- Votre culture organisationnelle peut réagir plus rapidement face aux changements.
- Les tendances sont utilisées afin d'identifier des possibilités d'amélioration.
- Les analyses rétrospectives valident les investissements effectués dans votre charge de travail et vos opérations.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

L'implémentation de cette bonne pratique signifie que vous utilisez à la fois les commentaires immédiats et les analyses rétrospectives. Ces boucles de commentaires stimulent les améliorations. Il existe de nombreux mécanismes de commentaires immédiats, notamment des enquêtes, des sondages auprès des clients ou des formulaires de commentaires. Votre organisation utilise également des analyses rétrospectives afin d'identifier les possibilités d'amélioration et de valider les initiatives.

Exemple client

AnyCompany Retail a créé un formulaire Web dans lequel les clients peuvent donner leur avis ou signaler des problèmes. Au cours de la mêlée hebdomadaire, les commentaires des utilisateurs sont évalués par l'équipe de développement logiciel. Les commentaires sont régulièrement utilisés pour orienter l'évolution de la plateforme de l'entreprise. Les utilisateurs effectuent une analyse rétrospective à la fin de chaque sprint afin d'identifier les éléments qu'elle souhaite améliorer.

Étapes d'implémentation

1. Commentaires immédiats
    - Vous avez besoin d'un mécanisme pour recevoir les commentaires des clients et des membres de l'équipe. Vos activités opérationnelles peuvent également être configurées de façon à fournir des commentaires automatisés.

- Votre organisation a besoin d'un processus pour examiner ces commentaires, déterminer ce qui doit être amélioré et planifier l'amélioration.
- Les commentaires doivent être ajoutés à votre processus de développement logiciel.
- Lorsque vous apportez des améliorations, effectuez un suivi auprès de l'auteur des commentaires.
  - Vous pouvez l'utiliser AWS Systems Manager OpsCenterpour créer et suivre ces améliorations en tant que OpsItems.

2. Analyse rétrospective
   - Effectuez des analyses rétrospectives à la fin d'un cycle de développement, sur une cadence définie ou après une version majeure.
   - Réunissez les parties prenantes impliquées dans la charge de travail pour une réunion rétrospective.
   - Créez trois colonnes sur un tableau blanc ou une feuille de calcul : Arrêter, Commencer et Conserver.
     - La colonne Arrêter comportera tout ce que votre équipe doit arrêter de faire.
     - La colonne Commencer comportera tout ce que votre équipe doit commencer à faire.
     - La colonne Conserver comportera tout ce que vous souhaitez continuer à faire.
   - Faites le tour de la salle et recueillez les commentaires des parties prenantes.
   - Privilégiez les commentaires. Attribuez les actions et les parties prenantes aux points que vous souhaitez commencer ou conserver.
   - Ajoutez les actions à votre processus de développement logiciel et communiquez les mises à jour de statut aux parties prenantes à mesure que vous apportez les améliorations.

Niveau d'effort du plan d'implémentation : moyen. Pour implémenter cette bonne pratique, vous avez besoin d'une solution pour recevoir des commentaires immédiats et effectuer une analyse. En outre, vous devez établir un processus d'analyse rétrospective.

Ressources

Bonnes pratiques associées :

- OPS01-BP01 Évaluer les besoins des clients externes : les boucles de commentaires sont un mécanisme qui permet de recueillir les besoins des clients externes.

- OPS01-BP02 Évaluer les besoins des clients internes : les parties prenantes internes peuvent utiliser les boucles de rétroaction afin de communiquer les besoins et les exigences.
- OPS11-BP02 Réaliser une analyse post-incident : les analyses post-incident sont une forme importante d'analyse rétrospective menée après les incidents.
- OPS11-BP07 Réaliser des examens des métriques opérationnelles : les examens des métriques opérationnelles permettent d'identifier les tendances et les points à améliorer.

Documents connexes :

- 7 pièges à éviter lors de la construction d'un CCOE
- Atlassian Team Playbook – Retrospectives
- Email Definitions: Feedback Loops
- Établissement de boucles de rétroaction basées sur la révision du AWS cadre Well-Architected
- IBMMéthodologie du garage - Organisez une rétrospective
- Investopedia — Le cycle PDCS
- Maximizing Developer Effectiveness by Tim Cochran
- Livre blanc sur les examens de l'état de préparation des opérations (ORR) - Itération
- ITILCSI- Amélioration continue du service
- When Toyota met e-commerce: Lean at Amazon

Vidéos connexes :

- Building Effective Customer Feedback Loops

Exemples connexes :

- Astuto - Open source customer feedback tool
- AWS Solutions - Q nABot on AWS
- Fider - A platform to organize customer feedback

Services connexes :

- AWS Systems Manager OpsCenter

OPS11-BP04 Effectuer la gestion des connaissances

La gestion des connaissances aide les membres de l'équipe à trouver les informations nécessaires à l'accomplissement de leur tâche. Dans les organisations qui fonctionnent selon le principe de l'apprentissage, les informations sont librement partagées, ce qui donne du pouvoir aux individus. Les informations peuvent être découvertes ou recherchées. Les informations sont exactes et à jour. Il existe des mécanismes permettant de générer de nouvelles informations, de mettre à jour les informations existantes et d'archiver les informations obsolètes. L'exemple le plus courant de plateforme de gestion des connaissances est un système de gestion de contenu comme un wiki.

Résultat escompté :

- Les membres de l'équipe ont accès à des informations précises et opportunes.
- Les informations sont consultables.
- Il existe des mécanismes pour ajouter, mettre à jour et archiver des informations.

Anti-modèles courants :

- Il n'y a pas de stockage centralisé des connaissances. Les membres de l'équipe gèrent leurs propres notes sur leurs machines locales.
- Vous disposez d'un wiki auto-hébergé mais ne disposez d'aucun mécanisme de gestion des informations, ce qui se traduit par des informations obsolètes.
- Quelqu'un identifie des informations manquantes mais il n'existe aucun processus pour demander leur ajout dans le wiki de l'équipe. Cette personne l'ajoute elle-même mais manque une étape clé, ce qui entraîne une panne.

Avantages liés au respect de cette bonne pratique :

- Les membres de l'équipe sont responsabilisés, car les informations sont partagées librement.
- Les nouveaux membres de l'équipe sont intégrés plus rapidement, car la documentation est à jour et consultable.
- Les informations sont opportunes, précises et exploitables.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

La gestion des connaissances est une facette importante des organisations qui fonctionnent selon le principe de l'apprentissage. Pour commencer, vous avez besoin d'un référentiel central pour stocker vos connaissances (par exemple, un wiki auto-hébergé). Vous devez développer des processus pour ajouter, mettre à jour et archiver les connaissances. Développez des normes pour ce qui doit être documenté et laissez chacune et chacun contribuer.

Exemple client

AnyCompany Retail héberge un wiki interne où toutes les connaissances sont stockées. Les membres de l'équipe sont encouragés à enrichir la base de connaissances dans l'exercice de leurs fonctions quotidiennes. Chaque trimestre, une équipe interfonctionnelle évalue les pages les moins mises à jour et détermine si elles doivent être archivées ou mises à jour.

Étapes d'implémentation

1. Commencez par identifier le système de gestion de contenu dans lequel les connaissances seront stockées. Obtenez l'accord des parties prenantes de votre organisation.
   a. Si vous ne disposez pas d'un système de gestion de contenu, envisagez d'utiliser un wiki hébergé par vos soins ou un référentiel de contrôle de version comme point de départ.
2. Développez des runbooks pour l'ajout, la mise à jour et l'archivage des informations. Formez votre équipe à ces processus.
3. Identifiez les connaissances qui doivent être stockées dans le système de gestion de contenu. Commencez par les activités quotidiennes (runbooks et playbooks) que les membres de l'équipe effectuent. Travaillez avec les parties prenantes pour prioriser les connaissances à ajouter.
4. Travaillez périodiquement avec les parties prenantes pour identifier les out-of-date informations et les archiver ou les mettre à jour.

Niveau d'effort du plan d'implémentation : moyen. Si vous ne disposez pas d'un système de gestion de contenu, vous pouvez mettre en place un wiki auto-hébergé ou un référentiel de documents contrôlé par version.

Ressources

Bonnes pratiques associées :

- OPS11-BP08 Documenter et partager les enseignements : la gestion des connaissances facilite le partage des informations sur les enseignements tirés.

Documents connexes :

- Atlassian – Gestion des connaissances

Exemples connexes :

- DokuWiki
- Gollum
- MediaWiki
- Wiki.js

OPS11-BP05 Définir les moteurs d'amélioration

Identifiez les facteurs d’amélioration pour vous aider à évaluer et à hiérarchiser les opportunités en fonction des données et des boucles de rétroaction. Explorez les opportunités d’amélioration de vos systèmes et processus, et procédez à l’automatisation le cas échéant.

Résultat escompté :

- Vous suivez les données provenant de l’ensemble de votre environnement.
- Vous mettez en corrélation les événements et les activités avec les résultats commerciaux.
- Vous pouvez comparer et contraster les environnements et les systèmes.
- Vous conservez un historique détaillé des activités de vos déploiements et de leurs résultats.
- Vous collectez des données pour renforcer votre niveau de sécurité.

Anti-modèles courants :

- Vous collectez des données provenant de l’ensemble de votre environnement, mais vous ne mettez pas en corrélation les événements et les activités.
- Vous collectez des données détaillées sur l'ensemble de votre patrimoine, et cela stimule Amazon, CloudWatch son AWS CloudTrail activité et ses coûts. Cependant, vous n’utilisez pas ces données de manière significative.
- Vous ne tenez pas compte des résultats commerciaux lorsque vous définissez les facteurs d’amélioration.
- Vous ne mesurez pas les effets des nouvelles fonctionnalités.

Avantages liés au respect de cette bonne pratique :

- Vous minimisez l'impact des motivations liées aux événements ou de l'investissement émotionnel en déterminant des critères d'amélioration.
- Vous répondez à des événements commerciaux, et pas seulement à des événements techniques.
- Vous mesurez votre environnement pour identifier les domaines à améliorer.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

- Comprendre les moteurs de l'amélioration : avant d'apporter des modifications à un système, il faut s'assurer que le résultat souhaité est bien pris en charge par celui-ci.
  - Fonctionnalités souhaitées : évaluez les fonctionnalités souhaitées lorsque vous étudiez les possibilités d'amélioration.
    - Quelles sont les nouveautés avec AWS
  - Problèmes inadmissibles : évaluez les problèmes inadmissibles, les bogues et les vulnérabilités lorsque vous étudiez les possibilités d'amélioration. Suivez les options de dimensionnement et recherchez les opportunités d'optimisation.
    - Derniers bulletins de sécuritéAWS
    - AWS Trusted Advisor
    - Cloud Intelligence Dashboards
  - Exigences de conformité : évaluez les mises à jour et les changements nécessaires pour assurer la conformité avec la réglementation ou une politique, ou pour continuer à bénéficier du soutien d'un tiers, lors de l'examen des possibilités d'amélioration.
    - ConformitéAWS
    - Programmes de conformitéAWS
    - Dernières actualités sur la conformitéAWS

Ressources

Bonnes pratiques associées :

- OPS01 Priorités de l'organisation
- OPS02 Relations et propriétés

- OPS04-BP01 Identifier les indicateurs de performance clés
- OPS08 Utilisation de l'observabilité de la charge de travail
- OPS09 Comprendre la santé opérationnelle
- OPS11-BP03 Implémenter des boucles de rétroaction

Documents connexes :

- Amazon Athena
- Amazon QuickSight
- ConformitéAWS
- Dernières actualités sur la conformitéAWS
- Programmes de conformitéAWS
- AWS Glue
- Derniers bulletins de sécuritéAWS
- AWS Trusted Advisor
- Exporter les données du journal vers Amazon S3
- Nouveautés d' AWS
- Les impératifs de l'innovation centrée sur le client
- Transformation numérique : hype ou nécessité stratégique ?

Vidéos connexes

- AWS re:Invent 2023 - Améliorez l'efficacité opérationnelle et la résilience avec Support (0) SUP31

OPS11-BP06 Valider les informations

Vérifiez vos résultats d'analyse et les réponses avec les équipes interfonctionnelles et les responsables métier. Utilisez ces analyses pour établir la compréhension, identifier des impacts supplémentaires et déterminer des lignes de conduite. Ajustez les réponses si nécessaire.

Résultat escompté :

- Vous passez régulièrement en revue les informations avec les responsables métier. Les propriétaires d'entreprise fournissent un contexte supplémentaire aux nouvelles connaissances.

- Vous examinez des informations et demandez le retour de vos pairs techniques, et vous partagez vos connaissances avec les équipes.
- Vous publiez des données et des informations pour que d’autres équipes techniques et commerciales puissent les examiner. Vous tenez compte de ce que vous avez appris des nouvelles pratiques d’autres départements.
- Résumez et examinez les nouvelles idées avec les hauts responsables. Les hauts responsables utilisent de nouvelles connaissances pour définir leur stratégie.

Anti-modèles courants :

- Vous publiez une nouvelle fonctionnalité. Cette fonctionnalité modifie certains comportements de vos clients. Votre observabilité ne tient pas compte de ces changements. Vous ne quantifiez pas les avantages de ces changements.
- Vous lancez une nouvelle mise à jour et négligez d'actualiser votreCDN. Le CDN cache n'est plus compatible avec la dernière version. Vous mesurez le pourcentage de demandes comportant des erreurs. Tous vos utilisateurs signalent HTTP 400 erreurs lorsqu'ils communiquent avec les serveurs principaux. Vous examinez les erreurs du client et vous constatez que vous avez perdu votre temps parce que vous avez mesuré la mauvaise dimension.
- Votre contrat de niveau de service stipule une disponibilité de 99,9 % et votre objectif de point de restauration est de quatre heures. Le responsable du service affirme que le système ne connaît aucun temps d’arrêt. Vous implémentez une solution de réplication coûteuse et complexe, ce qui représente une perte de temps et d’argent.

Avantages liés au respect de cette bonne pratique :

- Lorsque vous validez les informations avec les responsables métier et les experts du domaine, vous pouvez établir une compréhension commune et orienter plus efficacement les améliorations.
- Vous découvrez des problèmes cachés et vous en tenez compte dans vos décisions futures.
- Vous vous concentrez davantage sur les résultats commerciaux que sur les résultats techniques.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

- Validation des informations : collaborez avec les propriétaires d’entreprise et les experts du domaine pour vous assurer qu’il existe une compréhension et un accord communs sur la

signification des données que vous avez recueillies. Identifiez les autres préoccupations, les impacts potentiels et déterminez les mesures à prendre.

Ressources

Bonnes pratiques associées :

- OPS01-BP06 Évaluer les compromis tout en gérant les avantages et les risques
- OPS02-BP06 Les responsabilités entre les équipes sont prédéfinies ou négociées
- OPS11-BP03 Implémenter des boucles de rétroaction

Documents connexes :

- Conception d'un centre d'excellence dans le cloud (CCOE)

Vidéos connexes :

- Building observability to increase resiliency

OPS11-BP07 Réaliser des examens des métriques opérationnelles

Régulièrement, faites des analyses rétrospectives des métriques opérationnelles avec des intervenants provenant de différents services de l'entreprise. Utilisez ces examens pour identifier les possibilités d'amélioration, les pistes d'action potentielles et pour partager les enseignements tirés. Recherchez des opportunités d'amélioration dans l'ensemble de vos environnements (par exemple, le développement, le test et la production).

Résultat escompté :

- Vous passez fréquemment en revue les métriques qui ont une incidence sur l'activité.
- Vous détectez et examinez les anomalies grâce à vos fonctionnalités d'observabilité.
- Vous utilisez les données pour soutenir les résultats et les objectifs de l'entreprise.

Anti-modèles courants :

- Votre fenêtre de maintenance interrompt une importante promotion de vente au détail. L'entreprise continue d'ignorer qu'il existe une fenêtre de maintenance standard qui peut être retardée si d'autres événements ont un impact sur l'activité.
- Vous avez subi une panne prolongée parce que vous utilisez fréquemment une bibliothèque obsolète dans votre organisation. Vous avez depuis migré vers une bibliothèque prise en charge. Les autres équipes de votre organisation ne savent pas qu'elles sont exposées à des risques.
- Vous ne passez pas régulièrement en revue le nombre de clients SLAs atteints. Vous avez tendance à ne pas rencontrer votre clientSLAs. Le fait de ne pas rencontrer votre client entraîne des pénalités financièresSLAs.

Avantages liés au respect de cette bonne pratique :

- Lorsque vous vous réunissez régulièrement pour examiner les métriques opérationnelles, les événements et les incidents, vous maintenez une compréhension commune entre les équipes.
- Votre équipe se réunit régulièrement pour examiner les indicateurs et les incidents, ce qui vous permet de prendre des mesures en cas de risque et de reconnaître le clientSLAs.
- Vous partagez les leçons apprises, qui fournissent des données permettant de hiérarchiser les priorités et d'améliorer de manière ciblée les résultats commerciaux.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

- Régulièrement, faites des analyses rétrospectives des métriques opérationnelles avec des intervenants provenant de différents services de l'entreprise.
- Faites appel à différents intervenants, y compris des membres de l'équipe commerciale, de l'équipe de développement et de l'équipe opérationnelle, pour qu'ils valident vos résultats par l'intermédiaire de rétroactions immédiates et d'analyses rétrospectives et pour partager les leçons apprises.
- Utilisez leurs informations pour identifier les possibilités d'amélioration et les plans d'action possibles.

Ressources

Bonnes pratiques associées :

- OPS08-BP05 Création de tableaux de bord

- OPS09-BP03 Examiner les indicateurs des opérations et prioriser les améliorations
- OPS10-BP01 Utiliser un processus de gestion des événements, des incidents et des problèmes

Documents connexes :

- Amazon CloudWatch
- Référence CloudWatch des métriques et dimensions d'Amazon
- Publication de métriques personnalisées
- Utilisation des CloudWatch métriques Amazon
- Tableaux de bord et visualisations avec CloudWatch

OPS11-BP08 Documenter et partager les enseignements

Documentation et partage d’enseignements : documentez et partagez les enseignements que vous tirez des activités opérationnelles afin de pouvoir les utiliser en interne et entre les équipes. Vous devez partager les enseignements tirés par vos équipes afin d’en retirer un bénéfice accru pour toute votre organisation. Partagez des informations et des ressources pour éviter les erreurs évitables et faciliter les efforts de développement, et concentrez-vous sur la livraison des fonctionnalités souhaitées.

Utilisez AWS Identity and Access Management (IAM) pour définir les autorisations permettant de contrôler l’accès aux ressources que vous souhaitez partager au sein des comptes et entre les comptes.

Résultat escompté :

- Vous utilisez des référentiels dont les versions sont contrôlées pour partager des bibliothèques d’application, des procédures scriptées, de la documentation de procédure et d’autres documentations système.
- Vous partagez vos normes d’infrastructure sous forme de modèles AWS CloudFormation dont les versions sont contrôlées.
- Vous passez en revue les leçons apprises par les équipes.

Anti-modèles courants :

- Vous avez subi une panne prolongée, car votre organisation utilise couramment une bibliothèque défaillante. Depuis, vous avez migré vers une bibliothèque fiable. Les autres équipes de votre organisation ne savent pas qu'elles sont exposées à des risques. Personne ne documente et ne partage l'expérience vécue avec cette bibliothèque, et personne n'est conscient des risques.
- Vous avez identifié un cas limite dans un microservice partagé en interne qui entraîne l'abandon des sessions. Vous avez mis à jour vos appels au service pour éviter ce cas limite. Les autres équipes de votre organisation ne savent pas qu'elles sont exposées à des risques.
- Vous avez trouvé un moyen de réduire considérablement les besoins d'utilisation du processeur pour l'un de vos microservices. Vous ne savez pas si d'autres équipes peuvent tirer parti de cette technique.

Avantages liés au respect de cette bonne pratique : partagez les enseignements que vous avez tirés pour soutenir l'amélioration et pour optimiser les bénéfices de l'expérience.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

- Documentation et partage d'enseignements : mettez en place des procédures pour documenter les enseignements que vous tirez de l'exécution des activités opérationnelles et des analyses rétrospectives, afin que d'autres équipes puissent les utiliser.
- Partage des enseignements : imaginez des procédures permettant de partager ces enseignements, ainsi que les artefacts qui y sont associés, avec les autres équipes. Partagez par exemple les mises à jour concernant les procédures, les conseils, la gouvernance et les bonnes pratiques par l'intermédiaire d'un wiki accessible. Partagez des scripts, du code et des bibliothèques grâce à un référentiel commun.
  - Tirez parti d'AWS re:Post Private en tant que service de connaissances pour rationaliser la collaboration et le partage des connaissances au sein de votre organisation.

Ressources

Bonnes pratiques associées :

- OPS02-BP06 Les responsabilités entre les équipes sont prédéfinies ou négociées
- OPS05-BP01 Utiliser le contrôle de version
- OPS05-BP06 Partager les normes de conception

- OPS11-BP03 Mise en œuvre de boucles de commentaires
- OPS11-BP07 Examens des métriques des opérations

Documents connexes :

- Amélioration de la collaboration et partage sécurisé des connaissances relatives au cloud avec AWS re:Post Private
- Réduction des délais liés aux projets grâce à une solution Docs-as-Code

Vidéos connexes :

- AWS re:Invent 2023 - Collaborate within your company and with AWS using AWS re:Post Private
- Supports You | Exploring the Incident Management Tabletop Exercise

OPS11-BP09 Allouez du temps pour apporter des améliorations

Consacrez du temps et des ressources à vos processus pour permettre des améliorations progressives continues.

Résultat escompté :

- Vous créez des copies temporaires d'environnements, ce qui réduit les risques, les efforts et les coûts liés à l'expérimentation et aux tests.
- Ces copies d'environnements peuvent être utilisées pour tester les conclusions de votre analyse, expérimenter, et développer et tester des améliorations planifiées.
- Vous organisez des journées de jeu et vous utilisez Fault Injection Service (FIS) pour fournir les commandes et les garde-fous dont les équipes ont besoin pour mener des expériences dans un environnement similaire à celui de la production.

Anti-modèles courants :

- Il existe un problème de performances connu sur votre serveur d'applications. Il s'ajoute au retard accumulé dans la mise en œuvre de chaque fonctionnalité planifiée. Si le rythme d'ajout des fonctionnalités prévues reste constant, la question des performances ne sera jamais abordée.

- Pour permettre l'amélioration continue, vous autorisez les administrateurs et les développeurs à utiliser tout leur temps supplémentaire pour sélectionner et mettre en œuvre les améliorations. Aucune amélioration n'est effectuée.
- L'acceptation opérationnelle est terminée et vous ne testez plus les pratiques opérationnelles.

Avantages liés au respect de cette bonne pratique : ainsi, vous permettez d'apporter des améliorations progressives continues.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

- Allouez du temps aux améliorations : dédiez une partie des ressources et du temps consacrés à vos processus pour apporter des améliorations incrémentielles continues.
- Mettez en œuvre des modifications afin d'améliorer et d'évaluer les résultats, mais également de déterminer le taux de réussite qu'ils représentent.
- Si les résultats sont en deçà des objectifs et que l'amélioration constitue toujours une priorité, exécutez d'autres plans d'action.
- Simulez les charges de travail de production pendant les journées de simulation et utilisez les enseignements tirés de ces simulations pour apporter des améliorations.

Ressources

Bonnes pratiques associées :

- OPS05-BP08 Utiliser plusieurs environnements

Vidéos connexes :

- AWS re:Invent 2023 - Améliorez la résilience des applications grâce au service d'injection de AWS défauts

Sécurité

Le pilier Sécurité présente la capacité de protéger les données ainsi que les systèmes et les ressources pour tirer parti des technologies du cloud et améliorer votre sécurité. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Sécurité.

Domaines de bonnes pratiques

- Bases de la sécurité
- Gestion des identités et des accès
- Détection
- Protection de l'infrastructure
- Protection des données
- Intervention en cas d'incidents
- Sécurité des applications

Bases de la sécurité

Question

- SÉC 1. Comment gérer votre charge de travail en toute sécurité ?

SÉC 1. Comment gérer votre charge de travail en toute sécurité ?

Pour gérer votre charge de travail en toute sécurité, vous devez appliquer les bonnes pratiques générales à tous les domaines de sécurité. Prenez les exigences et les processus que vous avez définis dans le cadre de l'excellence opérationnelle au niveau de l'organisation et de la charge de travail, et appliquez-les à tous les domaines. En restant informé des recommandations AWS et du secteur, ainsi que des renseignements sur les menaces, vous pouvez faire évoluer votre modèle de menace et vos objectifs de contrôle. L'automatisation des processus de sécurité, les tests et la validation vous permettent de mettre à l'échelle vos opérations de sécurité.

Bonnes pratiques

- SEC01-BP01 Séparer les charges de travail à l'aide de comptes
- SEC01-BP02 Utilisateur root et propriétés du compte sécurisé
- SEC01-BP03 Identifier et valider les objectifs de contrôle
- SEC01-BP04 Restez au courant des menaces de sécurité et des recommandations
- SEC01-BP05 Réduire le périmètre de gestion de la sécurité
- SEC01-BP06 Automatiser le déploiement des contrôles de sécurité standard
- SEC01-BP07 Identifier les menaces et hiérarchiser les atténuations à l'aide d'un modèle de menaces

- SEC01-BP08 Évaluer et implémenter régulièrement de nouveaux services et fonctionnalités de sécurité

SEC01-BP01 Séparer les charges de travail à l'aide de comptes

Établissez des barrières de protection et un isolement communs entre les environnements (par exemple, production, développement et test) et les charges de travail grâce à une stratégie multicompte. La séparation au niveau des comptes est vivement recommandée, car elle fournit une solide limite d’isolement pour la sécurité, la facturation et les accès.

Résultat souhaité : une structure de compte qui isole les opérations cloud, les charges de travail indépendantes et les environnements dans des comptes distincts, renforçant ainsi la sécurité de l’infrastructure cloud.

Anti-modèles courants :

- Placer plusieurs charges de travail non liées avec différents niveaux de sensibilité des données dans le même compte.
- Structure d’unité d’organisation mal définie.

Avantages liés au respect de cette bonne pratique :

- Réduction de la portée des répercussions si un utilisateur accède à une charge de travail par inadvertance.
- Gouvernance centrale de l'accès aux AWS services, aux ressources et aux régions.
- Maintien de la sécurité de l’infrastructure cloud avec des politiques et une administration centralisée des services de sécurité.
- Processus automatisé de création et de gestion des comptes.
- Audit centralisé de votre infrastructure pour les exigences en matière de conformité et de réglementation.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Comptes AWS fournir une limite d'isolation de sécurité entre les charges de travail ou les ressources qui fonctionnent à différents niveaux de sensibilité. AWS fournit des outils pour gérer vos charges de

travail dans le cloud à grande échelle grâce à une stratégie multi-comptes afin de tirer parti de cette limite d'isolation. Pour obtenir des conseils sur les concepts, les modèles et la mise en œuvre d'une stratégie multi-comptes AWS, voir Organisation de votre AWS environnement à l'aide de plusieurs comptes.

Lorsque vous en avez plusieurs Comptes AWS sous gestion centralisée, vos comptes doivent être organisés selon une hiérarchie définie par des couches d'unités organisationnelles (OUs). Les contrôles de sécurité peuvent ensuite être organisés et appliqués aux comptes des membres, établissant OUs ainsi des contrôles préventifs cohérents sur les comptes des membres de l'organisation. Les contrôles de sécurité sont hérités, vous pouvez donc filtrer les autorisations disponibles pour les comptes membres situés aux niveaux inférieurs d'une hiérarchie d'unités d'organisation. Une bonne conception tire parti de cet héritage pour réduire le nombre et la complexité des politiques de sécurité nécessaires afin de mettre en place les contrôles de sécurité souhaités pour chaque compte membre.

AWS Organizationset AWS Control Towersont deux services que vous pouvez utiliser pour implémenter et gérer cette structure multi-comptes dans votre AWS environnement. AWS Organizations vous permet d'organiser les comptes selon une hiérarchie définie par une ou plusieurs couches deOUs, chaque unité d'organisation contenant un certain nombre de comptes membres. Les politiques de contrôle des services (SCPs) permettent à l'administrateur de l'organisation d'établir des contrôles préventifs précis sur les comptes des membres et AWS Configpeuvent être utilisées pour établir des contrôles proactifs et détectifs sur les comptes des membres. De nombreux AWS services s'intègrent AWS Organizations pour fournir des contrôles administratifs délégués et effectuer des tâches spécifiques aux services sur tous les comptes membres de l'organisation.

En plus de cela AWS Organizations, AWS Control Towerfournit une configuration des meilleures pratiques en un clic pour un AWS environnement multi-comptes avec une zone de landing zone. La zone de destination est le point d'entrée de l'environnement multicompte établi par Control Tower. Control Tower offre plusieurs avantages par rapport à AWS Organizations. Les trois avantages qui permettent d'améliorer la gouvernance des comptes sont les suivants :

- Des contrôles de sécurité obligatoires intégrés qui sont automatiquement appliquées aux comptes admis dans l'organisation.
- Commandes facultatives qui peuvent être activées ou désactivées pour un ensemble donné deOUs.
- AWS Control Tower Account Factory permet le déploiement automatique de comptes contenant des lignes de base et des options de configuration préapprouvées au sein de votre organisation.

Étapes d’implémentation

1. Conception d’une structure d’unité organisationnelle : une structure d’unité organisationnelle correctement conçue réduit la charge de gestion requise pour créer et maintenir des politiques de contrôle des services et d’autres contrôles de sécurité. La structure de votre unité organisationnelle doit être alignée sur les besoins de votre entreprise, la sensibilité des données et la structure de la charge de travail.
2. Créez une zone de destination pour votre environnement multicomptes : une zone de destination fournit une base de sécurité et d’infrastructure cohérente à partir de laquelle votre organisation peut rapidement développer, lancer et déployer des charges de travail. Vous pouvez utiliser une zone de destination personnalisée ou AWS Control Tower pour orchestrer votre environnement.
3. Établissez des barrières de protection : mettez en place des barrières de protection de sécurité cohérentes pour votre environnement dans toute votre zone de destination. AWS Control Tower fournit une liste de contrôles obligatoires et facultatifs qui peuvent être déployés. Les contrôles obligatoires sont déployés automatiquement lors de l’implémentation de Control Tower. Passez en revue la liste des contrôles hautement recommandés et facultatifs, puis implémentez les contrôles adaptés à vos besoins.
4. Restreindre l'accès aux régions récemment ajoutées : pour les nouvelles régions Régions AWS, les IAM ressources telles que les utilisateurs et les rôles ne sont propagées qu'aux régions que vous spécifiez. Cette action peut être effectuée via la console lorsque vous utilisez Control Tower, ou en ajustant les politiques IAM d'autorisation dans AWS Organizations.
5. Pensez StackSets à AWS CloudFormation StackSets: vous aider à déployer des ressources, notamment des IAM politiques, des rôles et des groupes, dans différentes Comptes AWS régions à partir d'un modèle approuvé.

Ressources

Bonnes pratiques associées :

- SEC02-BP04 S’appuyer sur un fournisseur d’identité centralisé

Documents connexes :

- AWS Control Tower
- Consignes pour les audits de sécuritéAWS
- IAMBonnes pratiques

- CloudFormation StackSets À utiliser pour provisionner des ressources sur plusieurs Comptes AWS régions
- Organisations FAQ
- AWS Organizations terminologie et concepts
- Meilleures pratiques en matière de politiques de contrôle des services dans un AWS Organizations environnement multi-comptes
- Guide de référence sur la gestion des comptes AWS
- Organisation de votre AWS environnement à l'aide de plusieurs comptes

Vidéos connexes :

- Permettre l’adoption d’ AWS à grande échelle grâce à l’automatisation et à la gouvernance
- Bonnes pratiques de sécurité : une approche Well-Architected
- Création et gestion de plusieurs comptes à l'aide AWS Control Tower
- Activer Control Tower pour les organisations existantes

Ateliers connexes :

- Journée d’immersion dans la Control Tower

SEC01-BP02 Utilisateur root et propriétés du compte sécurisé

L'utilisateur root est l'utilisateur le plus privilégié d'un compte Compte AWS, avec un accès administratif complet à toutes les ressources du compte et, dans certains cas, il ne peut pas être limité par des politiques de sécurité. Si vous désactivez l’accès par programmation pour l’utilisateur racine, établissez des contrôles appropriés pour l’utilisateur racine et évitez l’utilisation de routine de l’utilisateur racine, vous réduirez le risque d’exposition accidentelle des informations d’identification racine et de compromission ultérieure de l’environnement cloud.

Résultat souhaité : la sécurisation de l’utilisateur racine permet de réduire les risques de dommages accidentels ou intentionnels dus à une mauvaise utilisation des informations d’identification de l’utilisateur racine. La mise en place de contrôles de détection permet également d’alerter le personnel approprié lorsque des mesures sont prises à l’aide de l’utilisateur racine.

Anti-modèles courants :

- Se servir de l’utilisateur racine pour des tâches autres que celles nécessitant des informations d’identification de l’utilisateur racine.
- Omettre de tester régulièrement des plans d’urgence pour vérifier le fonctionnement de l’infrastructure, des processus et du personnel essentiels dans les situations d’urgence.
- Ne tenir compte que du flux de connexion type du compte et omettre d’envisager ou de tester d’autres méthodes de récupération de compte.
- Ne pas gérer DNS les serveurs de messagerie et les fournisseurs de téléphonie dans le cadre du périmètre de sécurité critique, car ils sont utilisés dans le flux de récupération des comptes.

Avantages du respect de cette bonne pratique : la sécurisation de l’accès à l’utilisateur racine permet de s’assurer que les actions de votre compte sont contrôlées et auditées.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

AWS propose de nombreux outils pour sécuriser votre compte. Toutefois, étant donné que certaines de ces mesures ne sont pas activées par défaut, vous devez intervenir directement pour les implémenter. Considérez ces recommandations comme des étapes fondamentales pour sécuriser votre Compte AWS. À mesure que vous mettez en œuvre ces étapes, il est important d’établir un processus permettant d’évaluer et de surveiller continuellement les contrôles de sécurité.

Lorsque vous créez un Compte AWS, vous commencez par une identité offrant un accès complet à tous les AWS services et ressources du compte. Cette identité est appelée utilisateur Compte AWS root. Vous pouvez vous connecter en tant qu’utilisateur racine avec l’adresse e-mail et le mot de passe que vous avez utilisés pour créer le compte. En raison de l'accès élevé accordé à l'utilisateur AWS root, vous devez limiter l'utilisation de l'utilisateur AWS root pour effectuer des tâches qui l'exigent spécifiquement. Les identifiants de connexion de l'utilisateur root doivent être étroitement protégés, et l'authentification multifactorielle (MFA) doit toujours être utilisée pour l'utilisateur Compte AWS root.

Outre le flux d'authentification normal pour vous connecter à votre utilisateur root à l'aide d'un nom d'utilisateur, d'un mot de passe et d'un dispositif d'authentification multifactorielle (MFA), il existe des flux de récupération de compte permettant de se connecter à votre utilisateur Compte AWS root ayant accès à l'adresse e-mail et au numéro de téléphone associés à votre compte. Par conséquent, il est tout aussi important de sécuriser le compte de messagerie de l’utilisateur racine là où l’e-mail de récupération est envoyé, ainsi que le numéro de téléphone associé au compte. Tenez également

compte des dépendances circulaires potentielles dans lesquelles l'adresse e-mail associée à l'utilisateur root est hébergée sur des serveurs de messagerie ou sur des ressources de service de noms de domaine (DNS) de ces serveurs Compte AWS.

Lors de l'utilisation AWS Organizations, il y en a plusieurs, Comptes AWS chacun ayant un utilisateur root. Un compte est désigné comme compte de gestion et plusieurs couches de comptes membres peuvent alors être ajoutées sous le compte de gestion. Privilégiez la sécurisation de l'utilisateur racine de votre compte de gestion, puis occupez-vous des utilisateurs racine des comptes membres. La stratégie de sécurisation de l'utilisateur racine de votre compte de gestion peut différer de celle des utilisateurs racine des comptes membres et vous pouvez placer des contrôles de sécurité préventifs sur les utilisateurs racine des comptes membres.

Étapes d'implémentation

Les étapes d'implémentation suivantes sont recommandées afin d'établir des contrôles pour l'utilisateur racine. Le cas échéant, les recommandations sont recoupées avec la version 1.4.0 du benchmark de CIS AWS Foundations. Outre ces étapes, consultez les directives relatives aux AWS meilleures pratiques pour sécuriser vos ressources Compte AWS et vos ressources.

Contrôles préventifs

1. Configurez des informations de contact précises pour le compte.
    a. Ces informations sont utilisées pour le flux de récupération du mot de passe perdu, le flux de récupération du compte d'MFAappareil perdu et pour les communications critiques liées à la sécurité avec votre équipe.
    b. Utilisez une adresse e-mail hébergée par votre domaine d'entreprise, de préférence une liste de distribution, comme adresse e-mail de l'utilisateur racine. L'utilisation d'une liste de distribution plutôt que d'un compte de messagerie individuel fournit une redondance et une continuité supplémentaires pour l'accès au compte racine sur de longues périodes.
    c. Le numéro de téléphone indiqué pour les coordonnées doit correspondre à un téléphone dédié et sécurisé à cette fin. Ce numéro de téléphone ne doit figurer sur aucune liste ni être communiqué à personne.
2. Ne créez pas de clés d'accès pour l'utilisateur racine. Si des clés d'accès existent, supprimez-les (CIS1.4).
    a. Éliminez les informations d'identification par programmation de longue durée (clés d'accès et secrètes) pour l'utilisateur racine.

   b. Si des clés d'accès utilisateur root existent déjà, vous devez transférer les processus utilisant ces clés pour utiliser les clés d'accès temporaires d'un rôle AWS Identity and Access Management (IAM), puis supprimer les clés d'accès utilisateur root.

3. Déterminez si vous devez stocker les informations d'identification de l'utilisateur racine.
   a. Si vous créez AWS Organizations de nouveaux comptes membres, le mot de passe initial de l'utilisateur root sur les nouveaux comptes membres est défini sur une valeur aléatoire qui ne vous est pas révélée. Envisagez d'utiliser le flux de réinitialisation du mot de passe du compte de gestion de votre AWS organisation pour accéder au compte membre si nécessaire.
   b. Dans le cas d'un compte autonome Comptes AWS ou d'un compte d' AWS organisation de gestion, pensez à créer et à stocker en toute sécurité les informations d'identification de l'utilisateur root. À utiliser MFA pour l'utilisateur root.
4. Utilisez des contrôles préventifs pour les utilisateurs root des comptes membres dans les environnements AWS multi-comptes.
   a. Envisagez d'utiliser la barrière de sécurité préventive Interdire la création de clés d'accès racine pour l'utilisateur racine pour les comptes des membres.
   b. Envisagez d'utiliser la barrière de sécurité préventive Désactiver les actions en tant qu'utilisateur racine pour les comptes des membres.
5. Si vous avez besoin d'informations d'identification pour l'utilisateur racine :
   a. Utilisez un mot de passe complexe.
   b. Activez l'authentification multifactorielle (MFA) pour l'utilisateur root, en particulier pour les comptes AWS Organizations de gestion (payeur) (CIS1.5).
   c. Pensez aux MFA périphériques matériels pour des raisons de résilience et de sécurité, car les appareils à usage unique peuvent réduire les chances que les appareils contenant vos MFA codes soient réutilisés à d'autres fins. Vérifiez que MFA les périphériques matériels alimentés par une batterie sont remplacés régulièrement. (CIS1,6)
      - Pour configurer MFA pour l'utilisateur root, suivez les instructions de création d'un MFApériphérique virtuel MFA ou matériel.
   d. Envisagez d'inscrire plusieurs MFA appareils à des fins de sauvegarde. Jusqu'à 8 MFA appareils sont autorisés par compte.
      - Notez que l'inscription de plusieurs MFA appareils pour l'utilisateur root désactive automatiquement le processus de récupération de votre compte en cas de perte de l'MFAappareil.

e. Stockez le mot de passe en sécurité et tenez compte des dépendances circulaires si vous le stockez électroniquement. Ne stockez pas le mot de passe d'une manière qui nécessiterait d'y accéder Compte AWS pour l'obtenir.

6. Facultatif : envisagez d’établir un calendrier périodique de rotation des mots de passe pour l’utilisateur racine.
   - Les bonnes pratiques relatives à la gestion des informations d’identification dépendent de vos exigences en matière de réglementation et de politiques. Les utilisateurs root protégés par ne MFA comptent pas sur le mot de passe comme seul facteur d'authentification.
   - La modification périodique du mot de passe de l’utilisateur racine réduit le risque d’utilisation abusive d’un mot de passe exposé par inadvertance.

Contrôles de détection

- Créez des alarmes pour détecter l'utilisation des informations d'identification root (CIS1.7). Amazon GuardDuty peut surveiller l'utilisation des informations d'APIidentification de l'utilisateur root et émettre des alertes à ce sujet grâce à cette RootCredentialUsagerecherche.
- Évaluez et mettez en œuvre les contrôles de détection inclus dans le AWS pack de conformité Well-Architected Security Pillar AWS Config pour, ou si vous AWS Control Tower utilisez, les contrôles fortement recommandés disponibles dans Control Tower.

Conseils opérationnels

- Déterminez qui, au sein de l’organisation, doit avoir accès aux informations d’identification de l’utilisateur racine.
  - Utilisez une règle à deux personnes afin qu'aucune personne n'ait accès à toutes les informations d'identification nécessaires et MFA pour obtenir l'accès de l'utilisateur root.
  - Vérifiez que l'organisation, et non une seule personne, garde le contrôle sur le numéro de téléphone et l'alias e-mail associés au compte (qui sont utilisés pour la réinitialisation du mot de passe et le flux de MFA réinitialisation).
- Utiliser l'utilisateur root uniquement par exception (CIS1.7).
  - L'utilisateur AWS root ne doit pas être utilisé pour les tâches quotidiennes, même administratives. Connectez-vous uniquement en tant qu’utilisateur racine pour effectuer des tâches AWS nécessitant un utilisateur racine. Toutes les autres actions doivent être effectuées par d’autres utilisateurs assumant les rôles appropriés.

- Vérifiez régulièrement que l'accès à l'utilisateur racine fonctionne afin que les procédures soient testées avant une situation d'urgence nécessitant l'utilisation des informations d'identification de l'utilisateur racine.
- Vérifiez régulièrement que l'adresse e-mail associée au compte et celles répertoriées sous Contacts alternatifs fonctionnent. Vérifiez dans ces boîtes de réception si vous avez reçu des notifications de sécurité de la part de <abuse@amazon.com>. Assurez-vous également que les numéros de téléphone associés au compte fonctionnent.
- Préparez les procédures d'intervention en cas d'incident pour réagir face à une utilisation inappropriée du compte racine. Consultez le guide de réponse aux incidents de sécuritéAWS et les bonnes pratiques décrites dans la section Réponse aux incidents du livre blanc sur le pilier Sécurité pour plus d'informations sur l'élaboration d'une stratégie de réponse aux incidents adaptée à votre Compte AWS.

Ressources

Bonnes pratiques associées :

- SEC01-BP01 Séparer les charges de travail à l'aide de comptes
- SEC02-BP01 Utiliser de solides mécanismes d'authentification
- SEC03-BP02 Accorder un accès selon le principe du moindre privilège
- SEC03-BP03 Établir un processus d'accès d'urgence
- SEC10-BP05 Préallouer les accès

Documents connexes :

- AWS Control Tower
- Consignes pour les audits de sécuritéAWS
- IAMBonnes pratiques
- Amazon GuardDuty — Alerte d'utilisation des informations d'identification root
- Un tep-by-step guide sur la surveillance de l'utilisation des informations d'identification root via CloudTrail
- MFAjetons approuvés pour une utilisation avec AWS
- Mise en œuvre de l'accès aux bris de verre sur AWS
- Les 10 meilleurs éléments de sécurité à améliorer dans votre Compte AWS

- Que faire si je remarque une activité non autorisée dans mon Compte AWS ?

Vidéos connexes :

- Permettre l’adoption d’ AWS à grande échelle grâce à l’automatisation et à la gouvernance
- Bonnes pratiques de sécurité : une approche Well-Architected
- Limitation de l'utilisation des informations d'identification AWS root depuis AWS re:inforce 2022 — Meilleures pratiques de sécurité avec AWS IAM

Exemples et ateliers connexes :

- Atelier : Compte AWS configuration et utilisateur root

SEC01-BP03 Identifier et valider les objectifs de contrôle

Fixez et validez les objectifs de contrôle et les contrôles que vous devez appliquer à votre charge de travail en fonction de vos exigences de conformité et des risques identifiés à partir de votre modèle de menace. La validation continue des objectifs de contrôle et des contrôles permet de mesurer l’efficacité de l’atténuation des risques.

Résultat souhaité : les objectifs de contrôle de sécurité de votre entreprise sont bien définis et conformes à vos exigences de conformité. Des contrôles sont mis en œuvre et appliqués par le biais de l’automatisation et des politiques. Leur efficacité dans le cadre de la réalisation de vos objectifs est évaluée en continu. Les preuves de l’efficacité à un moment donné et au cours d’une période spécifique peuvent être facilement transmises aux auditeurs.

Anti-modèles courants :

- Les exigences réglementaires, les attentes du marché et les normes du secteur en matière de sécurité assurable ne sont pas bien comprises pour votre entreprise
- Vos cadres de cybersécurité et vos objectifs de contrôle ne sont pas adaptés aux exigences de votre entreprise
- La mise en œuvre des contrôles n’est pas étroitement liée à vos objectifs de contrôle de manière mesurable
- Vous n’utilisez pas l’automatisation pour rendre compte de l’efficacité de vos contrôles

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

De nombreux cadres de cybersécurité courants peuvent constituer la base de vos objectifs en matière de contrôle de sécurité. Tenez compte des exigences réglementaires, des attentes du marché et des normes du secteur pour votre entreprise afin de déterminer les cadres les plus adaptés à vos besoins. Les exemples incluent AICPASOC2 HITRUST, PCI- DSS, ISO27001 et NISTSP 800-53.

En ce qui concerne les objectifs de contrôle que vous identifiez, comprenez comment les AWS services que vous consommez vous aident à atteindre ces objectifs. AWS ArtifactÀ utiliser pour trouver de la documentation et des rapports conformes à vos cadres cibles qui décrivent l'étendue des responsabilités couvertes AWS et des conseils pour le champ d'application restant sous votre responsabilité. Pour obtenir des conseils supplémentaires spécifiques aux services, dans la mesure où ils s’alignent sur les différentes déclarations de contrôle du cadre, consultez les guides de conformité destinés aux clients AWS.

Lorsque vous définissez les contrôles destinés à vous permettre d’atteindre vos objectifs, codifiez l’application à l’aide de contrôles préventifs et automatisez les mesures d’atténuation à l’aide de contrôles de détection. Aidez à prévenir les configurations de ressources et les actions non conformes dans l'ensemble de vos activités à AWS Organizations l'aide de politiques de contrôle des services (SCP). Mettez en œuvre des règles dans AWS Config pour surveiller et signaler les ressources non conformes, puis basculez les règles vers un modèle d’application une fois que vous êtes sûr de leur comportement. Pour déployer des ensembles de règles prédéfinies et gérées qui s’alignent sur vos cadres de cybersécurité, évaluez l’utilisation des normes AWS Security Hub comme première option. La norme AWS Foundational Service Best Practices (FSBP) et le CIS AWS Foundations Benchmark constituent de bons points de départ avec des contrôles qui s'alignent sur de nombreux objectifs partagés entre plusieurs cadres standard. Lorsque Security Hub ne dispose pas intrinsèquement des détections de contrôle souhaitées, il peut être complété par des packs de conformitéAWS Config.

Utilisez les packs de APN partenaires recommandés par l'équipe AWS Global Security and Compliance Acceleration (GSCA) pour obtenir l'assistance de conseillers en sécurité, d'agences de conseil, de systèmes de collecte de preuves et de reporting, d'auditeurs et d'autres services complémentaires en cas de besoin.

Étapes d’implémentation

1. Évaluez les cadres de cybersécurité courants et alignez vos objectifs de contrôle sur ceux que vous aurez choisis.

2. Obtenez la documentation pertinente sur les conseils et les responsabilités liés à l'utilisation de votre framework AWS Artifact. Identifiez les aspects de la conformité qui relèvent AWS du modèle de responsabilité partagée et ceux qui relèvent de votre responsabilité.
3. UtilisationSCPs, politiques de ressources, politiques de confiance dans les rôles et autres mesures de protection visant à empêcher les configurations et actions de ressources non conformes.
4. Évaluez le déploiement des normes et des packs de AWS Config conformité du Security Hub conformes à vos objectifs de contrôle.

Ressources

Bonnes pratiques associées :

- SEC03-BP01 Définir les exigences d'accès
- SEC04-BP01 Configuration de la journalisation des services et des applications
- SEC07-BP01 Comprenez votre schéma de classification des données
- OPS01-BP03 Évaluer les exigences de gouvernance
- OPS01-BP04 Évaluer les exigences de conformité
- PERF01-BP05 Politiques d'utilisation et architectures de référence
- COST02-BP01 Élaborez des politiques basées sur les exigences de votre organisation

Documents connexes :

- Guides de conformité pour les clients AWS

Outils associés :

- AWS Artifact

SEC01-BP04 Restez au courant des menaces de sécurité et des recommandations

Restez au fait des dernières menaces et mesures d’atténuation en surveillant les publications et les flux de données sur les menaces dans le secteur afin de connaître les données les plus à jour. Évaluez les offres de services gérés qui sont automatiquement mises à jour en fonction des données les plus récentes sur les menaces.

Résultat souhaité : vous restez informé au fur et à mesure que les publications du secteur sont mises à jour avec les dernières menaces et recommandations. Vous utilisez l'automatisation pour détecter les vulnérabilités et les expositions potentielles au fur et à mesure que vous identifiez de nouvelles menaces. Vous prenez des mesures pour atténuer ces menaces. Vous adoptez AWS des services qui se mettent automatiquement à jour en fonction des informations les plus récentes sur les menaces.

Anti-modèles courants :

- Ne pas disposer d'un mécanisme fiable et reproductible pour connaître les informations les plus récentes sur les menaces.
- Gérer un inventaire manuel de votre portefeuille technologique, de vos charges de travail et de vos dépendances qui nécessitent un examen humain pour détecter les vulnérabilités et les expositions potentielles.
- Ne pas avoir mis en place de mécanismes pour mettre à jour vos charges de travail et vos dépendances avec les dernières versions disponibles qui fournissent des mesures d'atténuation des menaces connues.

Avantages du respect de cette bonne pratique : l'utilisation de sources d'information sur les menaces pour rester à jour réduit le risque de passer à côté de changements importants du paysage des menaces susceptibles d'avoir un impact sur votre entreprise. La mise en place d'une automatisation pour analyser, détecter et corriger les vulnérabilités ou les expositions potentielles présentes dans vos charges de travail et leurs dépendances peut vous aider à atténuer les risques de manière rapide et prévisible, par rapport à des solutions manuelles. Cela permet de contrôler le temps et les coûts liés à l'atténuation des vulnérabilités.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Consultez des publications fiables comportant des informations sur les menaces pour rester au fait de l'évolution de ces dernières. Consultez la base de connaissances MITREATT&CK pour obtenir de la documentation sur les tactiques, techniques et procédures antagonistes connues ()TTPs. Consultez MITRE la liste des vulnérabilités et expositions courantes (CVE) pour rester informé des vulnérabilités connues des produits sur lesquels vous comptez. Comprenez les risques critiques auxquels sont exposées les applications Web grâce au célèbre projet OWASPTop 10 de l'Open Worldwide Application Security Project (OWASP).

Tenez-vous au courant des événements AWS de sécurité et des mesures correctives recommandées grâce aux bulletins AWS de sécurité pourCVEs.

Pour réduire les efforts et les frais généraux liés à la mise à jour, pensez à utiliser AWS des services qui intègrent automatiquement les nouvelles informations sur les menaces au fil du temps. Par exemple, Amazon se GuardDuty tient au courant des informations sur les menaces du secteur pour détecter les comportements anormaux et les signatures de menaces au sein de vos comptes. Amazon Inspector met automatiquement à jour une base de données contenant les informations CVEs qu'il utilise pour ses fonctionnalités de numérisation continue. AWS WAF et AWS Shield Advanced fournissent tous deux des groupes de règles gérés qui sont mis à jour automatiquement à mesure que de nouvelles menaces apparaissent.

Passez en revue le pilier Excellence opérationnelle Well-Architected pour la gestion automatisée des flottes et l'application de correctifs.

Étapes d'implémentation

- Abonnez-vous aux mises à jour des publications comportant des informations sur les menaces pertinentes pour votre entreprise et votre secteur d'activité. Abonnez-vous aux bulletins de sécurité AWS .
- Envisagez d'adopter des services qui intègrent automatiquement les nouvelles informations sur les menaces, tels qu'Amazon GuardDuty et Amazon Inspector.
- Déployez une stratégie de gestion de flotte et de correctifs conforme aux bonnes pratiques du pilier Excellence opérationnelle Well-Architected.

Ressources

Bonnes pratiques associées :

- SEC01-BP07 Identifier les menaces et prioriser les mesures d'atténuation à l'aide d'un modèle de menace
- OPS01-BP05 Évaluer le paysage des menaces
- OPS11-BP01 Disposer d'un processus d'amélioration continue

SEC01-BP05 Réduire le périmètre de gestion de la sécurité

Déterminez si vous pouvez réduire votre périmètre de sécurité en utilisant AWS des services qui transfèrent la gestion de certains contrôles vers AWS (services gérés). Ces services peuvent

vous aider à réduire vos tâches de maintenance de sécurité, telles que le provisionnement de l'infrastructure, la configuration logicielle, l'application de correctifs ou les sauvegardes.

Résultat escompté : Vous tenez compte de l'étendue de votre gestion de la sécurité lorsque vous sélectionnez les AWS services adaptés à votre charge de travail. Le coût des frais généraux de gestion et des tâches de maintenance (le coût total de possession, ouTCO) est mis en balance avec le coût des services que vous sélectionnez, en plus des autres considérations relatives à Well-Architected. Vous intégrez la documentation AWS de contrôle et de conformité dans vos procédures d'évaluation et de vérification des contrôles.

Anti-modèles courants :

- Déployer des charges de travail sans bien comprendre le modèle de responsabilité partagée pour les services que vous sélectionnez.
- Héberger des bases de données et d'autres technologies sur des machines virtuelles sans avoir évalué un équivalent de service géré.
- Ne pas inclure les tâches de gestion de la sécurité dans le coût total de possession des technologies d'hébergement sur les machines virtuelles par rapport aux options de services gérés.

Avantages du respect de cette bonne pratique : l'utilisation de services gérés peut réduire la charge globale que représente la gestion des contrôles de sécurité opérationnels, ce qui peut réduire les risques de sécurité et le coût total de possession. Le temps qui serait autrement consacré à certaines tâches de sécurité peut être réinvesti dans des tâches qui apportent plus de valeur à votre entreprise. Les services gérés peuvent également réduire la portée de vos exigences de conformité en transférant certaines exigences de contrôle à AWS.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Vous pouvez intégrer les composants de votre charge de travail sur AWS de plusieurs manières. Pour installer et exécuter des technologies sur EC2 des instances Amazon, vous devez souvent assumer la plus grande part de la responsabilité globale en matière de sécurité. Pour vous aider à réduire le fardeau lié à l'exploitation de certains contrôles, identifiez les services AWS gérés qui réduisent la portée de votre part du modèle de responsabilité partagée et comprenez comment vous pouvez les utiliser dans votre architecture existante. Les exemples incluent l'utilisation d'Amazon Relational Database Service (RDSAmazon) pour déployer des bases de données, d'Amazon Elastic Kubernetes Service (Amazon) ou d'Amazon Elastic Container ECS Service EKS (Amazon) pour

orchestrer des conteneurs, ou d'utiliser des options sans serveur. Lorsque vous créez de nouvelles applications, réfléchissez aux services qui peuvent vous aider à réduire les délais et les coûts liés à la mise en œuvre et à la gestion des contrôles de sécurité.

Les exigences de conformité peuvent également entrer en ligne de compte lors de la sélection des services. Les services gérés peuvent transférer la conformité de certaines exigences vers AWS. Discutez avec votre équipe de conformité de la mesure dans laquelle elle est à l'aise pour auditer les aspects des services que vous exploitez et gérez et pour accepter les déclarations de contrôle dans les rapports AWS d'audit pertinents. Vous pouvez fournir les artefacts d'audit trouvés AWS Artifactà vos auditeurs ou régulateurs comme preuve des contrôles de AWS sécurité. Vous pouvez également utiliser les conseils de responsabilité fournis par certains artefacts AWS d'audit pour concevoir votre architecture, ainsi que les guides de conformitéAWS client. Ces conseils aident à déterminer les contrôles de sécurité supplémentaires à mettre en place afin de prendre en charge les cas d'utilisation spécifiques de votre système.

Lorsque vous utilisez des services gérés, familiarisez-vous avec le processus de mise à jour de leurs ressources vers des versions plus récentes (par exemple, mise à jour de la version d'une base de données gérée par Amazon RDS ou d'un environnement d'exécution de langage de programmation pour une AWS Lambda fonction). Bien que le service géré puisse effectuer cette opération pour vous, il vous incombe de configurer le calendrier de la mise à jour et de comprendre son impact sur vos opérations. Des outils comme AWS Health peuvent vous aider à suivre et à gérer ces mises à jour dans l'ensemble de vos environnements.

Étapes d'implémentation

1. Évaluez les composants de votre charge de travail qui peuvent être remplacés par un service géré.
    a. Si vous migrez une charge de travail vers AWS, tenez compte de la réduction de la gestion (temps et dépenses) et de la réduction des risques lorsque vous déterminez si vous devez réhéberger, refactoriser, reprogrammer, reconstruire ou remplacer votre charge de travail. Dans certains cas, des investissements supplémentaires au début d'une migration peuvent permettre de réaliser des économies importantes à long terme.
2. Envisagez de mettre en œuvre des services gérésRDS, comme Amazon, au lieu d'installer et de gérer vos propres déploiements technologiques.
3. Utilisez les directives relatives à la responsabilité AWS Artifact pour déterminer les contrôles de sécurité que vous devez mettre en place pour votre charge de travail.
4. Tenez un inventaire des ressources utilisées et tenez-vous au courant up-to-date des nouveaux services et approches afin d'identifier de nouvelles opportunités de réduction de la portée.

Ressources

Bonnes pratiques associées :

- PERF02-BP01 Sélectionnez les meilleures options de calcul pour votre charge de travail
- PERF03-BP01 Utilisez un magasin de données spécialement conçu pour répondre au mieux à vos besoins en matière d'accès aux données et de stockage
- SUS05-BP03 Utiliser des services gérés

Documents connexes :

- Événements du cycle de vie planifiés pour AWS Health

Outils associés :

- AWS Health
- AWS Artifact
- Guides de conformité pour les clients AWS

Vidéos connexes :

- Comment migrer vers une instance Amazon RDS ou Aurora My SQL DB à l'aide de AWS DMS ?
- AWS re:Invent 2023 - Gérez les événements du cycle de vie des ressources à grande échelle avec AWS Health

SEC01-BP06 Automatiser le déploiement des contrôles de sécurité standard

Appliquez DevOps des pratiques modernes lorsque vous développez et déployez des contrôles de sécurité standard dans tous vos AWS environnements. Définissez des contrôles et des configurations de sécurité standard à l'aide de modèles d'infrastructure en tant que code (IaC), capturez les modifications dans un système de contrôle de version, testez les modifications dans le cadre d'un pipeline CI/CD et automatisez le déploiement des modifications dans vos AWS environnements.

Résultat souhaité : les modèles IaC capturent des contrôles de sécurité standardisés et les transmettent à un système de contrôle de version. Les pipelines CI/CD sont situés à des endroits qui détectent les changements et automatisent les tests et le déploiement de vos AWS environnements.

Des barrières de protection sont en place pour détecter et signaler les erreurs de configuration des modèles avant de procéder au déploiement. Les charges de travail sont déployées dans des environnements où des contrôles standard sont en place. Les équipes peuvent déployer des configurations de service approuvées via un mécanisme en libre-service. Des stratégies de sauvegarde et de restauration sécurisées sont en place pour contrôler les configurations, les scripts et les données associées.

Anti-modèles courants :

- Apporter des modifications à vos contrôles de sécurité standard manuellement, via une console Web ou une interface de ligne de commande.
- S'appuyer sur des équipes chargées de la charge de travail individuelles pour mettre en œuvre manuellement les contrôles définis par une équipe centrale.
- S'appuyer sur une équipe de sécurité centrale pour déployer des contrôles au niveau de la charge de travail à la demande d'une équipe responsable d'une charge de travail.
- Permettre aux mêmes personnes ou équipes de développer, de tester et de déployer des scripts d'automatisation des contrôles de sécurité sans séparation appropriée des tâches ni freins et contrepoids.

Avantages du respect de cette bonne pratique : l'utilisation de modèles pour définir vos contrôles de sécurité standard vous permet de suivre et de comparer les modifications au fil du temps à l'aide d'un système de contrôle de version. L'utilisation de l'automatisation pour tester et déployer les modifications crée de la standardisation et de la prévisibilité, ce qui augmente les chances de réussite du déploiement et réduit les tâches manuelles répétitives. La fourniture d'un mécanisme en libre-service permettant aux équipes responsables de la charge de travail de déployer des services et des configurations approuvés réduit le risque d'erreurs et de mauvaise utilisation de la configuration. Cela leur permet également d'intégrer des contrôles plus tôt dans le processus de développement.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Si vous suivez les pratiques décrites dans SEC01-BP01 Séparez les charges de travail à l'aide de comptes, vous vous retrouverez avec plusieurs charges de travail Comptes AWS pour différents environnements que vous gérez à l'aide de comptes. AWS Organizations Bien que ces environnements et charges de travail puissent nécessiter des contrôles de sécurité distincts, vous pouvez standardiser certains contrôles de sécurité au sein de votre organisation. Cela concerne

notamment l'intégration de fournisseurs d'identité centralisés, la définition de réseaux et de pare-feux, et la configuration d'emplacements standard pour le stockage et l'analyse des journaux. De la même manière que vous pouvez utiliser l'infrastructure en tant que code (IaC) pour appliquer la même rigueur de développement de code d'application au provisionnement de l'infrastructure, vous pouvez également utiliser l'IaC pour définir et déployer vos contrôles de sécurité standard.

Dans la mesure du possible, définissez vos contrôles de sécurité de manière déclarative, par exemple dans AWS CloudFormation, et stockez-les dans un système de contrôle source. Utilisez DevOps des pratiques pour automatiser le déploiement de vos contrôles pour des versions plus prévisibles, des tests automatisés à l'aide d'outils tels que AWS CloudFormation Guard, et la détection des écarts entre les contrôles déployés et la configuration souhaitée. Vous pouvez utiliser des services tels que AWS CodePipeline, AWS CodeBuild et AWS CodeDeploy pour construire un pipeline CI/CD. Consultez les instructions de la section Organisation de votre AWS environnement à l'aide de plusieurs comptes pour configurer ces services dans leurs propres comptes, distincts des autres pipelines de déploiement.

Vous pouvez également définir des modèles pour normaliser la définition et le déploiement Comptes AWS, les services et les configurations. Cette technique permet à une équipe de sécurité centrale de gérer ces définitions et de les fournir aux équipes responsables de la charge de travail via une approche en libre-service. L'un des moyens d'y parvenir consiste à utiliser Service Catalog, dans lequel vous pouvez publier des modèles sous forme de produits que les équipes chargées des charges de travail peuvent intégrer dans leurs propres déploiements de pipeline. Si vous en utilisez AWS Control Tower, certains modèles et contrôles sont disponibles comme point de départ. Control Tower fournit également la fonctionnalité Account Factory, qui permet aux équipes chargées de la charge de travail de créer de nouveaux Comptes AWS en utilisant les normes que vous définissez. Cette fonctionnalité permet de supprimer les dépendances vis-à-vis d'une équipe centrale chargée d'approuver et de créer de nouveaux comptes lorsqu'ils sont identifiés comme nécessaires par vos équipes responsables des charges de travail. Vous pouvez avoir besoin de ces comptes pour isoler les différents composants de la charge de travail en fonction de raisons telles que la fonction qu'ils remplissent, la sensibilité des données traitées ou leur comportement.

Étapes d'implémentation

1. Déterminez comment vous allez stocker et gérer vos modèles dans un système de contrôle de version.
2. Créez des pipelines CI/CD pour tester et déployer vos modèles. Définissez des tests pour vérifier les erreurs de configuration et vérifier que les modèles sont conformes aux normes de votre entreprise.

3. Créez un catalogue de modèles standardisés à déployer par les équipes chargées de la charge de travail Comptes AWS et de services adaptés à vos besoins.
4. Mettez en œuvre des stratégies de sauvegarde et de restauration sécurisées pour vos configurations de contrôle, vos scripts et les données associées.

Ressources

Bonnes pratiques associées :

- OPS05-BP01 Utiliser le contrôle de version
- OPS05-BP04 Utiliser des systèmes de gestion de construction et de déploiement
- REL08-BP05 Déployer les modifications grâce à l'automatisation
- SUS06-BP01 Adopter des méthodes permettant d'introduire rapidement des améliorations en matière de durabilité

Documents connexes :

- Organisation de votre AWS environnement à l'aide de plusieurs comptes

Exemples connexes :

- Automatisez la création de comptes et le provisionnement des ressources à l'aide de Service Catalog AWS Organizations, et AWS Lambda
- Renforcez le DevOps pipeline et protégez les données avec AWS Secrets ManagerAWS KMS, et AWS Certificate Manager

Outils associés :

- AWS CloudFormation Guard
- Accélérateur de zone d'atterrissage activé AWS

SEC01-BP07 Identifier les menaces et hiérarchiser les atténuations à l’aide d’un modèle de menaces

Effectuez une modélisation des menaces pour identifier et gérer un registre actualisé des menaces potentielles et des mesures d’atténuation connexes pour votre charge de travail. Hiérarchisez vos menaces et adaptez vos atténuations des contrôles de sécurité pour les prévenir, les détecter et y

répondre. Retenez et maintenez ces mesures en fonction de votre charge de travail et de l'évolution de l'environnement de sécurité.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Qu'est-ce que la modélisation des menaces ?

« La modélisation des menaces permet d'identifier, de communiquer et de comprendre les menaces et les mesures d'atténuation dans le contexte de la protection de quelque chose de valeur. » — Modélisation des menaces liées aux applications dans le cadre de l'Open Web Application Security Project (OWASP)

Pourquoi devriez-vous modéliser les menaces ?

Les systèmes sont complexes et deviennent de plus en plus complexes et compétents au fil du temps, offrant plus de valeur opérationnelle, ainsi qu'une satisfaction et un engagement client accrus. Cela signifie que les décisions de conception informatique doivent tenir compte d'un nombre toujours croissant de cas d'utilisation. Cette complexité et ce nombre de permutations des cas d'utilisation nuisent généralement à l'efficacité des approches non structurées pour trouver et atténuer les menaces. Dans ces conditions, il est préférable d'adopter une approche systématique pour recenser les menaces potentielles qui pèsent sur le système, de concevoir les atténuations et d'établir la priorité de ces atténuations afin de veiller à ce que les ressources limitées de votre organisation aient un impact maximal sur l'amélioration de la posture de sécurité globale du système.

La modélisation des menaces est conçue pour fournir cette approche systématique, dans le but de trouver et de régler les problèmes au début du processus de conception, lorsque les atténuations impliquent un coût relatif et des efforts limités par rapport à plus tard dans le cycle de vie. Cette approche est conforme au principe industriel de la sécurité basée sur le décalage à gauche. Au final, la modélisation des menaces s'intègre au processus de gestion des risques d'une organisation et aide à prendre des décisions sur les contrôles à mettre en œuvre en utilisant une approche axée sur les menaces.

Quand faut-il modéliser les menaces ?

Commencez la modélisation des menaces le plus tôt possible dans le cycle de vie de votre charge de travail, afin de bénéficier de plus de flexibilité pour la gestion des menaces identifiées. Comme pour les bogues logiciels, plus vous identifiez les menaces rapidement, plus leur résolution est économique. Un modèle de menace est un document évolutif et il doit continuer à évoluer avec

vos charges de travail. Retenez vos modèles de menaces au fil du temps, y compris lorsqu'il y a un changement majeur, une évolution du contexte des menaces ou lorsque vous adoptez une nouvelle fonctionnalité ou un nouveau service.

Étapes d'implémentation

Comment modéliser les menaces ?

Il existe de nombreuses façons de modéliser les menaces. Comme pour les langages de programmation, chaque méthode a ses avantages et ses inconvénients. À vous de choisir celle qui fonctionne le mieux pour votre organisation. L'une des approches consiste à commencer par le cadre à 4 questions de Shostack pour la modélisation des menaces, qui pose des questions ouvertes afin de structurer votre exercice de modélisation des menaces :

1. Sur quoi travaillons-nous ?

   Le but de cette question est de vous aider à comprendre et à vous mettre d'accord sur le système que vous créez et les détails associés qui sont pertinents pour la sécurité. La création d'un modèle ou d'un diagramme est le moyen le plus courant de répondre à cette question, car elle vous permet de visualiser ce que vous construisez, par exemple à l'aide d'un diagramme de flux de données. Le fait de noter les hypothèses et les détails importants sur votre système vous aide également à définir ce qui est inclus dans le champ d'application. Cela permet à tous ceux qui contribuent au modèle de menaces de se concentrer sur la même chose et d'éviter les détours fastidieux pour étudier des sujets qui ne rentrent pas dans le champ d'application (y compris les versions obsolètes de votre système). Par exemple, si vous créez une application Web, il n'est probablement pas intéressant de consacrer du temps à la modélisation de la séquence de démarrage autorisé du système d'exploitation pour les clients du navigateur, car vous ne pouvez pas avoir un impact sur ce point avec votre conception.
2. Quels problèmes pouvons-nous rencontrer ?

   C'est là que vous identifiez les menaces qui pèsent sur votre système. Les menaces sont des actions ou des événements accidentels ou intentionnels qui ont des impacts indésirables et pourraient affecter la sécurité de votre système. Sans une compréhension claire de ce qui pourrait poser un problème, vous n'avez aucun moyen de faire quoi que ce soit.

   Il n'existe pas de liste standard des problèmes potentiels. La création de cette liste nécessite un brainstorming et une collaboration entre tous les membres de votre équipe et les personnes concernées impliquées dans l'exercice de modélisation des menaces. Vous pouvez faciliter votre brainstorming en utilisant un modèle d'identification des menaces, tel que STRIDE, qui suggère

différentes catégories à évaluer : usurpation d'identité, falsification, répudiation, divulgation d'informations, déni de service et élévation de privilèges. En outre, vous pouvez faciliter le brainstorming en consultant les listes existantes et les recherches pour vous inspirer, notamment le Top 10 de l'OWASP, le catalogue des menaces HiTrust et le catalogue des menaces de votre organisation.

3. Qu'allons-nous faire à ce sujet ?

   Comme pour la question précédente, il n'existe pas de liste standard avec toutes les atténuations possibles. Lors de cette étape, les informations utilisées sont les menaces, les acteurs et les domaines d'amélioration identifiés par rapport à l'étape précédente.

   La sécurité et la conformité sont la responsabilité partagée d'AWS et de vous. Il est important de comprendre que lorsque vous demandez « Qu'allons-nous faire à ce sujet ? », vous demandez également qui est responsable de ce qui doit être fait. En comprenant l'équilibre des responsabilités entre vous-même et AWS, vous pouvez évaluer votre exercice de modélisation des menaces en fonction des atténuations qui sont sous votre contrôle, c'est-à-dire, en règle générale, une combinaison des options de configuration du service AWS et vos propres atténuations spécifiques au système.

   En ce qui concerne la partie AWS de la responsabilité partagée, vous constaterez que les services AWS sont couverts par de nombreux programmes de conformité. Ces programmes vous aident à comprendre les contrôles rigoureux en place chez AWS afin de garantir la sécurité et la conformité du cloud. Les rapports d'audit de ces programmes peuvent être téléchargés par les clients AWS sur AWS Artifact.

   Quels que soient les services AWS utilisés, il y a toujours un élément de responsabilité client et les atténuations correspondant à ces responsabilités doivent être incluses dans votre modèle de menaces. En ce qui concerne les atténuations en matière de contrôle de sécurité pour les services AWS eux-mêmes, envisagez l'implémentation de contrôles de sécurité dans tous les domaines, y compris la gestion des identités et des accès (authentification et autorisation), la protection des données (au repos et en transit), la sécurité de l'infrastructure, la journalisation et la surveillance. La documentation de chaque service AWS comporte un chapitre dédié à la sécurité qui fournit des conseils sur les contrôles de sécurité à considérer comme des mesures d'atténuation. Il est surtout important de réfléchir au code que vous écrivez et à ses dépendances, ainsi que de penser aux contrôles que vous pourriez mettre en place pour résoudre ces menaces. Ces contrôles peuvent être des éléments tels que la validation des entrées, la gestion des sessions et la gestion des limites. La plupart des vulnérabilités sont souvent introduites dans le code personnalisé, c'est pourquoi il est important de se concentrer sur ce domaine.

4. Avons-nous fait du bon travail ?

   L'objectif est que votre équipe et votre organisation améliorent la qualité des modèles de menaces et la vitesse à laquelle vous effectuez la modélisation des menaces au fil du temps. Ces améliorations découlent d'une combinaison de pratique, d'apprentissage, d'enseignement et de révision. Pour aller plus loin et vous familiariser avec le sujet, il est recommandé que vous et votre équipe suiviez le cours sur la bonne modélisation des menaces pour les constructeurs ou l'atelier sur ce sujet. En outre, si vous recherchez des conseils sur la manière d'intégrer la modélisation des menaces dans le cycle de développement des applications de votre entreprise, consultez l'article Comment aborder la modélisation des menaces sur le blog de sécurité AWS.

Threat Composer

Pour vous aider et vous guider dans la modélisation des menaces, pensez à utiliser l'outil Threat Composer, qui vise à réduire le délai de rentabilisation lors de la modélisation des menaces. L'outil vous permet d'effectuer les opérations suivantes :

- Rédiger des déclarations de menace utiles, alignées sur la grammaire des menaces, qui fonctionnent dans un flux de travail naturel non linéaire
- Générer un modèle de menaces lisible par l'homme
- Générer un modèle de menaces lisible par machine pour vous permettre de traiter les modèles de menaces comme du code
- Identifier rapidement les domaines dans lesquels la qualité et la couverture peuvent être améliorées à l'aide du tableau de bord

Pour de plus amples informations, rendez-vous sur Threat Composer et passez à l'exemple d'espace de travail défini par le système.

Ressources

Bonnes pratiques associées :

- SEC01-BP03 Identifier et valider les objectifs de contrôle
- SEC01-BP04 Restez au courant des menaces de sécurité et des recommandations
- SEC01-BP05 Réduire le périmètre de gestion de la sécurité
- SEC01-BP08 Évaluer et implémenter régulièrement de nouveaux services et fonctionnalités de sécurité

Documents connexes :

- Comment aborder la modélisation des menaces (blog sur la sécurité AWS)
- NIST : Guide de modélisation des menaces système centrées sur les données

Vidéos connexes :

- AWS Summit ANZ 2021 - How to approach threat modelling
- AWS Summit ANZ 2022 - Scaling security – Optimise for fast and secure delivery

Formations associées :

- Modéliser les menaces de la bonne manière pour les constructeurs : formation virtuelle à suivre à leur rythme avec AWS Skill Builder
- Modéliser les menaces de la bonne manière pour les constructeurs — Atelier AWS

Outils associés :

- Threat Composer

SEC01-BP08 Évaluer et implémenter régulièrement de nouveaux services et fonctionnalités de sécurité

Évaluez et mettez en œuvre les services et fonctionnalités de sécurité fournis par AWS et par les AWS partenaires qui vous aident à faire évoluer le niveau de sécurité de votre charge de travail.

Résultat souhaité : Vous avez mis en place une pratique standard qui vous informe des nouvelles fonctionnalités et des nouveaux services publiés par AWS et par les AWS partenaires. Vous évaluez l'influence de ces nouvelles fonctionnalités sur la conception des contrôles actuels et nouveaux pour vos environnements et vos charges de travail.

Anti-modèles courants :

- Vous ne vous abonnez pas aux AWS blogs et aux RSS fils d'actualités pour découvrir rapidement les nouvelles fonctionnalités et services pertinents
- Vous vous fiez aux actualités et aux mises à jour concernant les services et fonctionnalités de sécurité provenant de sources secondaires.

- Vous n'encouragez pas AWS les utilisateurs de votre organisation à se tenir informés des dernières mises à jour

Avantages liés au respect de cette bonne pratique : lorsque vous restez au fait des nouveaux services et fonctionnalités de sécurité, vous pouvez prendre des décisions éclairées concernant la mise en œuvre des contrôles dans vos environnements cloud et vos charges de travail. Ces sources contribuent à sensibiliser le public à l'évolution du paysage de la sécurité et à la manière dont les AWS services peuvent être utilisés pour se protéger contre les menaces nouvelles et émergentes.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d’implémentation

AWS informe les clients des nouveaux services et fonctionnalités de sécurité par le biais de plusieurs canaux :

- AWS Quoi de neuf
- AWS Blog d'actualités
- Blog de sécuritéAWS
- Bulletins de sécuritéAWS
- Aperçu de la documentation AWS

Vous pouvez vous abonner à une rubrique consacrée aux mises à jour AWS quotidiennes des fonctionnalités via Amazon Simple Notification Service (AmazonSNS) pour obtenir un résumé quotidien complet des mises à jour. Certains services de sécurité, tels qu'Amazon GuardDuty et Amazon AWS Security Hub, proposent leurs propres SNS rubriques pour rester informés des nouvelles normes, des découvertes et des autres mises à jour relatives à ces services en particulier.

Les nouveaux services et fonctionnalités sont également annoncés et décrits en détail lors de conférences, d’événements et de webinaires organisés chaque année dans le monde entier. Il convient de noter en particulier la conférence annuelle sur la sécurité AWS re:Inforce et la conférence plus générale AWS re:Invent. Les chaînes d' AWS information mentionnées précédemment partagent ces annonces de conférence sur la sécurité et d'autres services, et vous pouvez suivre des sessions de formation approfondies en petits groupes en ligne sur la AWS chaîne Events on. YouTube

Vous pouvez également demander à votre équipe Compte AWS les dernières mises à jour et recommandations des services de sécurité. Vous pouvez contacter votre équipe via le formulaire

de Support commercial si vous ne disposez pas de ses coordonnées directes. De même, si vous êtes abonné au Support aux AWS entreprises, vous recevrez des mises à jour hebdomadaires de la part de votre responsable de compte technique (TAM) et pourrez planifier une réunion de révision régulière avec lui.

Étapes d'implémentation

1. Abonnez-vous aux différents blogs et bulletins avec votre RSS lecteur préféré ou à la SNS rubrique Mises à jour quotidiennes des fonctionnalités.
2. Évaluez les AWS événements auxquels vous devez assister pour découvrir de première main les nouvelles fonctionnalités et les nouveaux services.
3. Organisez des réunions avec votre Compte AWS équipe pour toute question concernant la mise à jour des services et fonctionnalités de sécurité.
4. Envisagez de vous abonner au Support aux entreprises pour consulter régulièrement un responsable de compte technique (TAM).

Ressources

Bonnes pratiques associées :

- PERF01-BP01 Découvrez et comprenez les services et fonctionnalités cloud disponibles
- COST01-BP07 Tenez-vous au courant up-to-date des nouvelles versions de service

Gestion des identités et des accès

Questions

- SÉC 2. Comment gérer l'authentification des personnes et des machines ?
- SÉC 3. Comment gérer les autorisations des personnes et des machines ?

SÉC 2. Comment gérer l'authentification des personnes et des machines ?

Il existe deux types d'identités que vous devez gérer dans le cadre de l'exploitation de charges de travail AWS sécurisées.

- Identités humaines : les identités humaines qui nécessitent l'accès à vos environnements et applications AWS peuvent être classées en trois groupes : employés, tiers et utilisateurs.

Le groupe des employés comprend les administrateurs, les développeurs et les opérateurs qui font partie de votre organisation. Ils ont besoin d'un accès pour gérer, créer et exploiter vos ressources AWS.

Les tiers sont les collaborateurs externes, tels que les sous-traitants, les fournisseurs et les partenaires. Ils interagissent avec vos ressources AWS dans le cadre de leur engagement auprès de votre organisation.

Les utilisateurs sont les consommateurs de vos applications. Ils accèdent à vos ressources AWS via des navigateurs Web, des applications client, des applications mobiles ou des outils de ligne de commande interactifs.

- Identités des machines : les applications, les outils opérationnels et les composants de votre charge de travail ont besoin d'une identité pour adresser des demandes aux services AWS, telles que la lecture de données. Ces identités incluent également les machines qui s'exécutent dans votre environnement AWS, comme les instances Amazon EC2 ou les fonctions AWS Lambda. Vous pouvez également gérer les identités des machines pour des parties externes ou des machines en dehors d'AWS, qui ont besoin d'accéder à votre environnement AWS.

Bonnes pratiques

- SEC02-BP01 Utiliser de solides mécanismes d'authentification
- SEC02-BP02 Utiliser des informations d'identification temporaires
- SEC02-BP03 Stocker et utiliser des secrets en toute sécurité
- SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé
- SEC02-BP05 Contrôler et effectuer régulièrement une rotation des informations d'identification
- SEC02-BP06 Utiliser des groupes d'utilisateurs et des attributs

SEC02-BP01 Utiliser de solides mécanismes d'authentification

Les connexions (authentification au moyen d'informations d'identification de connexion) peuvent présenter des risques lorsque l'on n'utilise pas des mécanismes tels que l'authentification multifactorielle (MFA), surtout dans les situations où les informations d'identification de connexion ont été divulguées par inadvertance ou peuvent être devinées facilement. Vous devez utiliser de solides mécanismes d'authentification pour réduire ces risques en exigeant l'authentification multifactorielle (MFA) et des politiques strictes de gestion des mots de passe.

Résultat escompté : réduisez les risques d'accès involontaire aux informations d'identification dans AWS en utilisant des mécanismes de connexion robustes pour les utilisateurs AWS Identity and Access Management (IAM), l'utilisateur racine du Compte AWS, AWS IAM Identity Center et les fournisseurs d'identité tiers. Cela signifie que vous devez exiger une authentification multifactorielle, appliquer des politiques strictes de gestion des mots de passe et détecter les comportements de connexion anormaux.

Anti-modèles courants :

- Ne pas appliquer de politique stricte de gestion des mots de passe pour vos identités, notamment des mots de passe complexes et l'authentification multifactorielle (MFA).
- Utiliser les mêmes informations d'identification pour différents utilisateurs.
- Ne pas utiliser de contrôles de détection pour les connexions suspectes.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les identités humaines peuvent se connecter à AWS de plusieurs façons. Une bonne pratique AWS consiste à s'appuyer sur un fournisseur d'identité centralisé en utilisant la fédération (fédération SAML 2.0 directe entre AWS IAM et le fournisseur d'identité centralisé ou utilisant AWS IAM Identity Center) lors de l'authentification auprès d'AWS. Dans ce cas, établissez un processus de connexion sécurisée avec votre fournisseur d'identité ou Microsoft Active Directory.

Lorsque vous ouvrez pour la première fois un Compte AWS, vous commencez avec un utilisateur racine de Compte AWS. Vous ne devez utiliser le compte utilisateur racine que pour configurer l'accès de vos utilisateurs (et pour les tâches nécessitant l'utilisateur racine). Il est important d'activer l'authentification multifactorielle (MFA) pour l'utilisateur racine du compte immédiatement après l'ouverture de votre Compte AWS et de sécuriser l'utilisateur racine à l'aide du guide des bonnes pratiques AWS.

AWS IAM Identity Center est conçu pour les utilisateurs en interne et vous pouvez créer et gérer des identités utilisateur au sein du service et sécuriser le processus d'authentification avec la MFA. AWS Cognito, quant à lui, est conçu pour la gestion de l'identité et de l'accès des clients (CIAM), qui fournit des groupes d'utilisateurs et des fournisseurs d'identité pour les identités des utilisateurs externes dans vos applications.

Si vous créez des utilisateurs dans AWS IAM Identity Center, sécurisez le processus d'authentification dans ce service et activez la MFA. Pour les identités des utilisateurs externes

dans vos applications, vous pouvez utiliser les groupes d'utilisateurs Amazon Cognito et sécuriser le processus d'authentification dans ce service ou utiliser l'un des fournisseurs d'identité pris en charge dans les groupes d'utilisateurs Amazon Cognito.

En outre, pour les utilisateurs dans AWS IAM Identity Center, vous pouvez utiliser Accès vérifié par AWS pour fournir un niveau de sécurité supplémentaire en vérifiant l'identité de l'utilisateur et la position de l'appareil avant d'accorder l'accès aux ressources AWS.

Si vous utilisez des utilisateurs AWS Identity and Access Management (IAM), sécurisez le processus d'authentification à l'aide d'IAM.

Vous pouvez utiliser AWS IAM Identity Center et la fédération IAM directe simultanément pour gérer l'accès à AWS. Vous pouvez utiliser la fédération IAM pour gérer l'accès à la AWS Management Console et aux services et IAM Identity Center pour gérer l'accès aux applications professionnelles telles qu'Amazon QuickSight ou Amazon Q Business.

Quelle que soit la méthode de connexion, il est essentiel d'appliquer une politique de connexion rigoureuse.

Étapes d'implémentation

Voici des recommandations générales en matière de connexion. Les paramètres réels que vous configurez doivent être définis par la politique de votre entreprise ou utiliser une norme telle que NIST 800-63.

- Require MFA (Demander l'authentification MFA). L'une des bonnes pratiques IAM consiste à exiger l'authentification multifactorielle pour les identités humaines et les charges de travail. L'activation de l'authentification multifactorielle fournit une couche de sécurité supplémentaire en exigeant que les utilisateurs fournissent des informations d'identification et un mot de passe unique (OTP) ou une chaîne de caractères générée et vérifiée cryptographiquement à partir d'un appareil physique.
- Mettez en place une longueur de mot de passe minimale, il s'agit d'un facteur essentiel pour garantir la force du mot de passe.
- Appliquez la complexité des mots de passe pour les rendre plus difficiles à deviner.
- Autorisez les utilisateurs à modifier leurs propres mots de passe.
- Créez des identités individuelles plutôt que des informations d'identification partagées. En créant des identités individuelles, vous pouvez attribuer à chaque utilisateur un ensemble unique d'informations d'identification de sécurité. Les utilisateurs individuels offrent la possibilité d'auditer l'activité de chaque utilisateur.

Recommandations IAM Identity Center :

- IAM Identity Center fournit une politique de mot de passe prédéfinie lors de l'utilisation du répertoire par défaut qui définit la longueur, la complexité et les exigences de réutilisation des mots de passe.
- Activez l'authentification multifactorielle et configurez le paramètre contextuel ou permanent pour l'authentification multifactorielle lorsque la source d'identité est le répertoire par défaut, AWS Managed Microsoft AD ou AD Connector.
- Permettez aux utilisateurs d'enregistrer leurs propres appareils d'authentification multifactorielle.

Recommandations d'annuaire des groupes d'utilisateurs Amazon Cognito :

- Configurez les paramètres de sécurité du mot de passe.
- Exigez l'authentification multifactorielle pour les utilisateurs.
- Utilisez les paramètres de sécurité avancés des groupes d'utilisateurs Amazon Cognito pour des fonctionnalités telles que l'authentification adaptative qui permet de bloquer les connexions suspectes.

Recommandations pour les utilisateurs IAM :

- Idéalement, vous utilisez IAM Identity Center ou la fédération directe. Cependant, vous aurez peut-être besoin d'utilisateurs IAM. Dans ce cas, définissez une politique de mot de passe pour les utilisateurs IAM. Vous pouvez utiliser la politique de gestion des mots de passe pour définir des exigences telles que la longueur minimale ou la nécessité d'utiliser des caractères non alphabétiques.
- Créez une politique IAM pour appliquer la connexion MFA afin que les utilisateurs soient autorisés à gérer leurs propres mots de passe et appareils d'authentification multifactorielle.

Ressources

Bonnes pratiques associées :

- SEC02-BP03 Stocker et utiliser des secrets en toute sécurité
- SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé
- SEC03-BP08 Partager des ressources en toute sécurité au sein de votre organisation

Documents connexes :

- Politique de mot de passe AWS IAM Identity Center
- Politique de mot de passe de l'utilisateur IAM
- Définition du mot de passe de l'utilisateur racine du Compte AWS
- Politique de mot de passe Amazon Cognito
- Informations d'identification AWS
- Bonnes pratiques de sécurité IAM

Vidéos connexes :

- Gestion des autorisations des utilisateurs à grande échelle avec IAM Identity Center AWS
- Maîtrise des identités dans chaque couche

SEC02-BP02 Utiliser des informations d'identification temporaires

Lors de tout type d'authentification, il est préférable d'utiliser des informations d'identification temporaires plutôt que des informations d'identification à long terme afin de réduire ou d'éliminer les risques, tels que la divulgation, le partage ou le vol des informations d'identification par inadvertance.

Résultat escompté : pour réduire le risque d'informations d'identification à long terme, utilisez des informations d'identification temporaires dans la mesure du possible pour les identités humaines et les identités machine. Les informations d'identification à long terme créent de nombreux risques, tels que l'exposition par le biais de chargements dans des référentiels publics. En utilisant des informations d'identification temporaires, vous réduisez considérablement les risques de compromission de ces informations d'identification.

Anti-modèles courants :

- Les développeurs utilisent des clés d'accès à long terme issues des utilisateurs IAM au lieu d'obtenir des informations d'identification temporaires de la CLI à l'aide de la fédération.
- Les développeurs intègrent des clés d'accès à long terme dans leur code et téléchargent ce code dans des référentiels Git publics.
- Les développeurs intègrent des clés d'accès à long terme dans les applications mobiles qui sont ensuite disponibles dans les boutiques d'applications.
- Les utilisateurs partagent des clés d'accès à long terme avec d'autres utilisateurs ou des employés quittent l'entreprise avec des clés d'accès à long terme toujours en leur possession.

- Utilisation des clés d'accès à long terme pour les identités machine lorsque des informations d'identification temporaires peuvent être utilisées.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Utilisez des informations d'identification de sécurité temporaires plutôt que des informations d'identification à long terme pour toutes les demandes d'API et de CLI AWS. Les demandes d'API et de CLI adressées aux services AWS doivent, dans presque tous les cas, être signées à l'aide de clés d'accès AWS. Ces demandes peuvent être signées avec des informations d'identification temporaires ou à long terme. Vous ne devez utiliser des informations d'identification à long terme, également appelées clés d'accès à long terme, que si vous utilisez un utilisateur IAM ou un utilisateur racine Compte AWS. Lorsque vous fédérez vers AWS ou que vous assumez un rôle IAM par le biais d'autres méthodes, des informations d'identification temporaires sont générées. Même lorsque vous accédez à la AWS Management Console à l'aide des informations d'identification de connexion, des informations d'identification temporaires sont générées pour vous permettre d'appeler les services AWS. Vous avez rarement besoin d'informations d'identification à long terme et vous pouvez accomplir presque toutes les tâches en utilisant des informations d'identification temporaires.

Privilégiez les informations d'identification temporaires plutôt que les informations d'identification à long terme et, parallèlement, mettez en place une stratégie de réduction des utilisateurs IAM au profit de la fédération et des rôles IAM. Bien que les utilisateurs IAM aient été employés pour les identités humaines et machine dans le passé, nous recommandons désormais de ne plus procéder ainsi afin d'éviter les risques liés à l'utilisation de clés d'accès à long terme.

Étapes d'implémentation

Identités humaines

Pour les identités du personnel comme les employés, les administrateurs, les développeurs et les opérateurs :

- Vous devez vous appuyer sur un fournisseur d'identité centralisé et exiger des utilisateurs humains qu'ils se joignent à un fournisseur d'identité pour accéder à AWS en utilisant des informations d'identification temporaires. La fédération de vos utilisateurs peut se faire soit par une fédération directe à chaque Compte AWS, soit en utilisant AWS IAM Identity Center et le fournisseur d'identité de votre choix. La fédération offre un certain nombre d'avantages par rapport aux utilisateurs IAM, outre l'élimination des informations d'identification à long terme. Vos utilisateurs peuvent

également demander des informations d'identification temporaires depuis la ligne de commande pour une fédération directe ou en utilisant IAM Identity Center. Cela signifie que peu de cas d'utilisation nécessitent des utilisateurs IAM ou des informations d'identification à long terme pour vos utilisateurs.

Pour les identités tierces :

- Lorsque vous accordez à des tiers, tels que des fournisseurs de logiciels en tant que service (SaaS), l'accès aux ressources de votre Compte AWS, vous pouvez utiliser des rôles entre comptes et des politiques basées sur les ressources. En outre, vous pouvez utiliser le flux d'informations d'identification client d'octroi OAuth 2.0 Amazon Cognito pour les partenaires et clients SaaS B2B.

Pour les identités utilisateur qui accèdent à vos ressources AWS via des navigateurs Web, des applications client, des applications mobiles ou des outils de ligne de commande interactifs :

- Si vous devez autoriser des demandes permettant à des consommateurs ou à des clients d'accéder à vos ressources AWS, vous pouvez utiliser les groupes d'identités Amazon Cognito ou les groupes d'utilisateurs Amazon Cognito pour fournir des informations d'identification temporaires. Les autorisations pour ces informations d'identification sont configurées via des rôles IAM. Vous pouvez également définir un rôle IAM distinct avec des autorisations limitées pour les utilisateurs invités qui ne sont pas authentifiés.

Identités de machines

Pour les identités machine, vous devrez peut-être utiliser des informations d'identification à long terme. Dans ces cas, vous devez exiger que les charges de travail utilisent des informations d'identification temporaires avec des rôles IAM pour accéder à AWS.

- Pour Amazon Elastic Compute Cloud (Amazon EC2), vous pouvez utiliser des rôles pour Amazon EC2.
- AWS Lambda vous permet de configurer un rôle d'exécution Lambda pour accorder au service les autorisations nécessaires pour effectuer des actions AWS à l'aide d'informations d'identification temporaires. Il existe de nombreux modèles similaires pour permettre aux services AWS d'octroyer des informations d'identification temporaires à l'aide des rôles IAM.
- Pour les appareils IoT, vous pouvez utiliser le fournisseur d'informations d'identification AWS IoT Core pour demander des informations d'identification temporaires.

- Pour les systèmes sur site ou ceux qui s'exécutent en dehors de AWS qui nécessitent un accès aux ressources AWS, vous pouvez utiliser Rôles Anywhere IAM.

Dans certains cas, les identifiants temporaires ne sont pas pris en charge et vous devez utiliser des informations d'identification à long terme. Dans ces cas, contrôlez et effectuez régulièrement une rotation de ces informations d'identification et effectuez régulièrement une rotation des clés d'accès. Pour les clés d'accès d'utilisateurs IAM à l'accès très restreint, envisagez d'utiliser les mesures de sécurité supplémentaires suivantes :

- Accordez des autorisations très restreintes :
  - Optez pour le principe du moindre privilège (soyez précis quant aux actions, aux ressources et aux conditions).
  - Envisagez d'accorder à l'utilisateur IAM uniquement l'opération AssumeRole pour un seul rôle spécifique. En fonction de l'architecture sur site, cette approche permet d'isoler et de sécuriser les informations d'identification IAM à long terme.
- Limitez les sources réseau et les adresses IP autorisées dans la politique d'approbation de rôle IAM.
- Surveillez l'utilisation et configurez des alertes pour des autorisations non utilisées ou abusives (à l'aide des filtres de métriques et des alarmes AWS CloudWatch Logs).
- Appliquez des limites d'autorisation (les politiques de contrôle des services (SCP) et les limites d'autorisation se complètent : les SCP sont grossières, tandis que les limites d'autorisation sont détaillées).
- Mettez en œuvre un processus pour provisionner et stocker en toute sécurité (dans un coffre-fort sur site) les informations d'identification.

Autres options pour les scénarios nécessitant des informations d'identification à long terme :

- Créez votre propre API de distribution de jetons (à l'aide d'Amazon API Gateway).
- Dans les scénarios où vous devez utiliser des informations d'identification à long terme ou des informations d'identification autres que des clés d'accès AWS (telles que les connexions à la base de données), vous pouvez utiliser un service conçu pour gérer la gestion des secrets, tel qu'AWS Secrets Manager. Secrets Manager simplifie la gestion, la rotation et le stockage sécurisé des secrets chiffrés. De nombreux services AWS prennent en charge une intégration directe avec Secrets Manager.

- Pour les intégrations multicloud, vous pouvez utiliser la fédération d'identité en fonction des informations d'identification de votre fournisseur de services d'informations d'identification (CSP) source (voir AWS STS AssumeRoleWithWebIdentity).

Pour plus d'informations sur la rotation des informations d'identification à long terme, veuillez consulter Rotation des clés d'accès.

Ressources

Bonnes pratiques associées :

- SEC02-BP03 Stocker et utiliser des secrets en toute sécurité
- SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé
- SEC03-BP08 Partager des ressources en toute sécurité au sein de votre organisation

Documents connexes :

- Informations d'identification de sécurité temporaires
- AWS Informations d'identification
- Bonnes pratiques de sécurité IAM
- Rôles IAM
- IAM Identity Center
- Fournisseurs d'identité et fédération
- Rotation des clés d'accès
- Solutions partenaires de sécurité : accès et contrôle d'accès
- Utilisateur racine d'un compte AWS
- Accès à AWS via une identité de charge de travail native de Google Cloud Platform
- Comment accéder aux ressources AWS à partir de locataires Microsoft Entra ID à l'aide d'AWS Security Token Service

Vidéos connexes :

- Gestion des autorisations des utilisateurs à grande échelle avec IAM Identity Center AWS
- Maîtrise des identités dans chaque couche

SEC02-BP03 Stocker et utiliser des secrets en toute sécurité

Une charge de travail nécessite une capacité automatisée pour prouver son identité aux bases de données, aux ressources et aux services tiers. Cela se fait à l'aide d'informations d'identification d'accès secrets, tels que des clés d'accès à l'API, des mots de passe et des jetons OAuth. L'utilisation d'un service spécialement conçu pour stocker, gérer et faire tourner ces informations d'identification permet de réduire les risques de compromission de ces informations d'identification.

Résultat escompté : mise en œuvre d'un mécanisme de gestion sécurisée des informations d'identification des applications permettant d'atteindre les objectifs suivants :

- Identification des secrets nécessaires pour la charge de travail.
- Réduction du nombre d'informations d'identification à long terme requis, en les remplaçant par des informations d'identification à court terme dans la mesure du possible.
- Établissement d'un stockage sécurisé et d'une rotation automatisée des informations d'identification à long terme restantes.
- Audit de l'accès aux secrets qui existent dans la charge de travail.
- Surveillance continue pour vérifier qu'aucun secret n'est intégré dans le code source pendant le processus de développement.
- Réduction des risques de divulgation des informations d'identification par inadvertance.

Anti-modèles courants :

- Aucune rotation des informations d'identification.
- Stockage des informations d'identification à long terme dans le code source ou les fichiers de configuration.
- Stockage des informations d'identification au repos non chiffrées.

Avantages liés au respect de cette bonne pratique :

- Les secrets sont chiffrés au repos et en transit.
- L'accès aux informations d'identification est bloqué par le biais d'une API (considérez-la comme un distributeur automatique d'informations d'identification).
- L'accès à une information d'identification (en lecture et en écriture) est audité et consigné.
- Séparation des préoccupations : la rotation des informations d'identification est effectuée par un composant distinct, qui peut être séparé du reste de l'architecture.

- Les secrets sont distribués automatiquement à la demande aux composants logiciels et la rotation se produit dans un emplacement central.
- L'accès aux informations d'identification peut être contrôlé de façon précise.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Dans le passé, les informations d'identification utilisées pour s'authentifier auprès des bases de données, des API tierces, des jetons et d'autres secrets pouvaient être intégrées dans du code source ou des fichiers d'environnement. AWS fournit plusieurs mécanismes pour stocker ces informations d'identification en toute sécurité, en effectuer la rotation automatiquement et vérifier leur utilisation.

Pour gérer les secrets de façon optimale, la meilleure solution consiste à suivre les directives de suppression, de remplacement et de rotation. Les informations d'identification les plus sûres sont celles que vous n'avez pas à stocker, gérer ou manipuler. Certaines informations d'identification qui ne sont plus nécessaires au fonctionnement de la charge de travail peuvent être supprimées en toute sécurité.

Pour les informations d'identification qui restent nécessaires au bon fonctionnement de la charge de travail, il peut être possible d'opter pour une solution temporaire ou à court terme au lieu d'utiliser des informations d'identification à long terme. Par exemple, au lieu du codage en dur une AWS clé d'accès secrète , envisagez de remplacer les informations d'identification à long terme par des informations d'identification temporaires à l'aide de rôles IAM.

Certains secrets de longue durée ne peuvent pas être supprimés ni remplacés. Ces secrets peuvent être stockés dans un service tel que AWS Secrets Manager, où ils peuvent être stockés, gérés et subir une rotation de manière centralisée et régulière.

Un audit du code source et des fichiers de configuration de la charge de travail peut révéler de nombreux types d'informations d'identification. Le tableau suivant résume les stratégies de traitement des types courants d'informations d'identification :

Type d'informations d'identification	Description	Stratégie suggérée
Clés d'accès IAM	AWS Accès IAM et clés secrètes utilisés pour assumer	Remplacer : utilisez plutôt les rôles IAM attribués aux

Type d'informations d'identification	Description	Stratégie suggérée
	des rôles IAM au sein d'une charge de travail	instances de calcul (telles que Amazon EC2 ou AWS Lambda). Pour l'interopérabilité avec des tiers qui ont besoin d'accéder aux ressources de votre compteCompte AWS, demandez-leur s'ils proposent AWS l'accès intercompte. Pour les applications mobiles, pensez à utiliser des informations d'identification temporaires via les groupes d'identités Amazon Cognito (identités fédérées). Pour les charges de travail exécutées en dehors de AWS, pensez à Rôles Anywhere IAM ou à AWS Systems Manager Hybrid Activations. Pour les conteneurs, consultez le rôle IAM de la tâche Amazon ECS ou le rôle IAM du nœud Amazon EKS.
Clés SSH	Clés privées Secure shell utilisées pour se connecter aux instances Linux EC2, manuellement ou dans le cadre d'un processus automatisé	Remplacer : utilisez AWS Systems Manager ou EC2 Instance Connect pour fournir un accès programmatique et humain aux instances EC2 à l'aide de rôles IAM.

Type d'informations d'identification	Description	Stratégie suggérée
Informations d'identification d'application et base de données	Mots de passe — chaîne de texte brut	Rotation : stockez les informations d'identification dans AWS Secrets Manager et établissez une rotation automatique si possible.
Informations d'identification d'administration Amazon RDS et Amazon Aurora	Mots de passe — chaîne de texte brut	Remplacer : utilisez l'intégration de Secrets Manager avec Amazon RDS ou Amazon Aurora. En outre, certains types de bases de données RDS peuvent utiliser des rôles IAM au lieu de mots de passe dans certains cas d'utilisation (pour plus de détails, voir Authentification de base de données IAM).
Jetons OAuth	Jetons secrets — chaîne de texte brut	Rotation : stockez les jetons dans AWS Secrets Manager et configurez la rotation automatique.
Jetons et clés API	Jetons secrets — chaîne de texte brut	Rotation : stockez dans AWS Secrets Manager et établissez une rotation automatique si possible.

Parmi les anti-modèles courants, citons l'intégration des clés d'accès IAM dans le code source, les fichiers de configuration ou les applications mobiles. Lorsqu'une clé d'accès IAM est requise pour communiquer avec un AWS service, utilisez des informations d'identification de sécurité temporaires (à court terme). Ces informations d'identification à court terme peuvent être fournies via des rôles IAM pour les instances EC2, des rôles d'exécution pour les fonctions Lambda, des rôles IAM Cognito pour l'accès des utilisateurs mobiles et des politiques IoT Core pour les appareils IoT. Lorsque vous

interagissez avec des tiers, préférez la délégation de l'accès à un rôle IAM disposant de l'accès requis aux ressources de votre compte à la configuration d'un utilisateur IAM et à l'envoi au tiers de la clé d'accès secrète de cet utilisateur.

Dans de nombreux cas, la charge de travail nécessite le stockage des secrets nécessaires à l'interopérabilité avec d'autres services et ressources. AWS Secrets Manager est spécialement conçu pour gérer en toute sécurité ces informations d'identification, ainsi que le stockage, l'utilisation et la rotation des jetons API, des mots de passe et d'autres informations d'identification.

AWS Secrets Manager fournit cinq fonctionnalités clés pour garantir le stockage et le traitement sécurisés des informations d'identification sensibles : le chiffrement au repos, le chiffrement en transit, l'audit complet, le contrôle d'accès précis et la rotation extensible des informations d'identification. D'autres services de gestion des secrets créés par des AWS partenaires ou des solutions développées localement qui offrent des capacités et des assurances similaires sont également acceptables.

Lorsque vous récupérez un secret, vous pouvez utiliser les composants de mise en cache côté client de Secrets Manager pour le mettre en cache en vue d'une utilisation future. Il est plus rapide de récupérer un secret mis en cache que de le récupérer à partir de Secrets Manager. De plus, étant donné que l'appel des API Secrets Manager implique des coûts, l'utilisation d'un cache peut réduire vos coûts. Pour connaître toutes les manières dont vous pouvez récupérer des secrets, consultez Obtenir des secrets.

> **Note**
>
> Certains langages peuvent vous obliger à implémenter votre propre chiffrement en mémoire pour la mise en cache côté client.

Étapes d'implémentation

1. Identifiez les chemins de code contenant des informations d'identification codées en dur à l'aide d'outils automatisés tels qu'Amazon CodeGuru.
   a. Utilisez Amazon CodeGuru pour analyser vos référentiels de code. Une fois l'examen terminé, filtrez sur Type=Secrets dans CodeGuru pour trouver les lignes de code problématiques.
2. Identifiez les informations d'identification qui peuvent être supprimées ou remplacées.
   a. Identifiez les informations d'identification qui ne sont plus nécessaires et marquez-les en vue de leur suppression.

   b. Pour les clés secrètes AWS qui sont intégrées au code source, remplacez-les par des rôles IAM associés aux ressources nécessaires. Si une partie de votre charge de travail est externe à AWS mais nécessite des informations d'identification IAM pour accéder aux ressources AWS, envisagez d'utiliser Rôles Anywhere IAM ou des activations hybrides AWS Systems Manager.

3. Pour les autres secrets tiers de longue durée qui nécessitent l'utilisation de la stratégie de rotation, intégrez Secrets Manager dans votre code afin d'extraire les secrets tiers au moment de l'exécution.
   a. La console CodeGuru peut créer automatiquement un secret dans Secrets Manager à l'aide des informations d'identification découvertes.
   b. Intégrez l'extraction des secrets à partir de Secrets Manager dans votre code d'application.
      i. Les fonctions Lambda sans serveur peuvent utiliser une extension Lambda indépendante du langage.
      ii. Pour les instances ou les conteneurs EC2, AWS fournit un exemple de code côté client permettant de récupérer des secrets à partir de Secrets Manager dans plusieurs langages de programmation courants.
4. Examinez régulièrement votre base de code et effectuez une nouvelle analyse afin de vérifier qu'aucun nouveau secret n'a été ajouté au code.
   a. Envisagez l'utilisation d'un outil tel que git-secrets pour éviter l'ajout de nouveaux secrets à votre dépôt de code source.
5. Surveillez l'activité de Secrets Manager pour détecter tout signe d'utilisation inattendue, d'accès secret inapproprié ou de tentative de suppression de secrets.
6. Réduisez l'exposition humaine aux informations d'identification. Limitez l'accès à la lecture, à l'écriture et à la modification des informations d'identification à un rôle IAM dédié à cette fin et fournissez un accès uniquement pour assumer le rôle à un petit sous-ensemble d'utilisateurs opérationnels.

Ressources

Bonnes pratiques associées :

- SEC02-BP02 Utiliser des informations d'identification temporaires
- SEC02-BP05 Contrôler et effectuer régulièrement une rotation des informations d'identification

Documents connexes :

- Mise en route avec AWS Secrets Manager
- Fournisseurs d'identité et fédération
- Amazon CodeGuru présente un détecteur de secrets
- Comment AWS Secrets Manager utilise-t-il AWS Key Management Service ?
- Chiffrement et déchiffrement de secret dans Secrets Manager
- Entrées du journal de Secrets Manager
- Amazon RDS annonce l'intégration avec AWS Secrets Manager

Vidéos connexes :

- Bonnes pratiques de gestion, d'extraction et de renouvellement des secrets à grande échelle
- Trouvez des secrets codés en dur à l'aide du détecteur de secrets Amazon CodeGuru
- Sécurisation des secrets des charges de travail hybrides à l'aide de AWS Secrets Manager

Ateliers connexes :

- Stockez, récupérez et gérez les informations d'identification sensibles dans AWS Secrets Manager
- Activations hybrides AWS Systems Manager

SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé

Pour les identités du personnel (employés et sous-traitants), faites confiance à un fournisseur d'identité qui vous permet de gérer les identités de manière centralisée. Cela facilite la gestion de l'accès entre plusieurs applications et systèmes, car vous créez, attribuez, gérez, révoquez et auditez l'accès depuis un seul emplacement.

Résultat escompté : Vous disposez d'un fournisseur d'identité centralisé dans lequel vous gérez de manière centralisée les utilisateurs faisant partie du personnel, les politiques d'authentification (telles que l'exigence d'authentification multifactorielle (MFA)) et les autorisations accordées aux systèmes et aux applications (telles que l'attribution de l'accès en fonction de l'appartenance à un groupe ou des attributs d'un utilisateur). Les utilisateurs en interne se connectent au fournisseur d'identité central et se fédèrent (authentification unique) avec les applications internes et externes, ce qui leur évite d'avoir à mémoriser différentes informations d'identification. Votre fournisseur d'identité est intégré à vos systèmes de ressources humaines (RH) afin que les changements de personnel soient automatiquement synchronisés avec lui. Par exemple, si quelqu'un quitte votre organisation,

vous pouvez automatiquement révoquer l'accès aux applications et systèmes fédérés (y compris AWS). Vous avez activé la journalisation détaillée des audits dans votre fournisseur d'identité et vous surveillez ces journaux pour détecter tout comportement inhabituel des utilisateurs.

Anti-modèles courants :

- Vous n'utilisez pas la fédération ni l'authentification unique. Les utilisateurs en interne créent des comptes utilisateur et des informations d'identification distincts dans plusieurs applications et systèmes.
- Vous n'avez pas automatisé le cycle de vie des identités pour les utilisateurs en interne, par exemple en intégrant votre fournisseur d'identité à vos systèmes RH. Lorsqu'un utilisateur quitte votre organisation ou change de rôle, vous suivez un processus manuel pour supprimer ou mettre à jour ses enregistrements dans plusieurs applications et systèmes.

Avantages liés au respect de cette bonne pratique : en utilisant un fournisseur d'identité centralisé, vous disposez d'un emplacement unique pour gérer les identités et les politiques des utilisateurs en interne, de la possibilité d'attribuer l'accès aux applications, aux utilisateurs et aux groupes, et de la capacité de surveiller l'activité de connexion des utilisateurs. Grâce à l'intégration du fournisseur d'identité dans vos systèmes de ressources humaines (RH), lorsqu'un utilisateur change de rôle, ces modifications sont synchronisées avec le fournisseur d'identité et mettent automatiquement à jour les applications et les autorisations qui lui ont été attribuées. Lorsqu'un utilisateur quitte votre organisation, son identité est automatiquement désactivée dans le fournisseur d'identité, révoquant ainsi son accès aux applications et systèmes fédérés.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Conseils pour les utilisateurs en interne accédant à AWS Les utilisateurs en interne, tels que les employés et les sous-traitants de votre organisation, peuvent avoir besoin d'accéder à AWS avec la AWS Management Console ou l'AWS Command Line Interface (AWS CLI) pour exécuter leurs tâches. Vous pouvez accorder l'accès AWS aux utilisateurs en interne en les fédérant avec AWS à deux niveaux à partir de votre fournisseur d'identité centralisé : fédération directe vers chaque Compte AWS ou fédération vers plusieurs comptes dans votre organisation AWS.

Pour fédérer les utilisateurs en interne directement avec chaque Compte AWS, vous pouvez utiliser un fournisseur d'identité centralisé afin de les fédérer à AWS Identity and Access Management dans ce compte. La flexibilité d'IAM vous permet d'activer un fournisseur d'identité SAML 2.0 ou

Open ID Connect (OIDC) distinct pour chaque Compte AWS et d'utiliser des attributs d'utilisateur fédéré pour le contrôle d'accès. Les utilisateurs en interne utiliseront leur navigateur Web pour se connecter au fournisseur d'identité en indiquant leurs informations d'identification (telles que des mots de passe et des codes de jeton MFA). Le fournisseur d'identité enverra à son navigateur une assertion SAML soumise à l'URL de connexion de la AWS Management Console pour permettre à l'utilisateur de s'authentifier de manière unique auprès de la AWS Management Console en assumant un rôle IAM. Vos utilisateurs peuvent également obtenir des informations d'identification d'API AWS temporaires à utiliser dans le AWS CLI ou AWS les SDK à partir de AWS STS en assumant le rôle IAM à l'aide d'une assertion SAML du fournisseur d'identité.

Pour fédérer vos utilisateurs en interne avec plusieurs comptes dans votre organisation AWS, vous pouvez utiliser AWS IAM Identity Center pour gérer de manière centralisée l'accès des utilisateurs en interne aux Comptes AWS et aux applications. Vous activez Identity Center pour votre organisation et configurez votre source d'identité. IAM Identity Center fournit un répertoire de sources d'identité par défaut que vous pouvez utiliser pour gérer vos utilisateurs et vos groupes. Vous pouvez également choisir une source d'identité externe en vous connectant à votre fournisseur d'identité externe à l'aide de SAML 2.0 et en provisionnant automatiquement les utilisateurs et les groupes à l'aide de SCIM, ou en vous connectant à votre annuaire Microsoft AD à l'aide de AWS Directory Service. Une fois qu'une source d'identité est configurée, vous pouvez attribuer aux utilisateurs et aux groupes l'accès aux Comptes AWS en définissant des politiques de moindre privilège dans vos ensembles d'autorisation. Les utilisateurs de votre personnel peuvent s'authentifier par l'intermédiaire de votre fournisseur d'identité central pour se connecter au portail d'accès AWS et ouvrir une session unique dans le Comptes AWS et les applications cloud qui leur sont attribuées. Vos utilisateurs peuvent configurer la AWS CLI v2 pour s'authentifier auprès d'Identity Center et obtenir des informations d'identification pour exécuter des commandes AWS CLI. Identity Center permet également l'accès par authentification unique aux applications AWS telles qu'Amazon SageMaker AI Studio et les portails AWS IoT Sitewise Monitor.

Une fois que vous aurez suivi les instructions précédentes, vos utilisateurs en interne n'auront plus besoin d'utiliser des utilisateur IAM et des groupes pour les opérations normales lors de la gestion des charges de travail sur AWS. Au lieu de cela, vos utilisateurs et vos groupes sont gérés en dehors de AWS et les utilisateurs peuvent accéder aux ressources AWS sous la forme d'une identité fédérée. Les identités fédérées utilisent les groupes définis par votre fournisseur d'identité centralisé. Vous devez identifier et supprimer les groupes IAM, les utilisateur IAM et les informations d'identification utilisateur de longue durée (mots de passe et clés d'accès) dont vous n'avez plus besoin dans vos Comptes AWS. Vous pouvez trouver les informations d'identification non utilisées à l'aide de rapports d'informations d'identification IAM, supprimer les utilisateurs IAM correspondants

et supprimer les groupes IAM. Vous pouvez appliquer une Politique de contrôle des services (SCP) à votre organisation afin d'empêcher la création de nouveaux utilisateurs IAM et groupes, en renforçant cet accès à AWS via des identités fédérées.

> **Note**
>
> Vous êtes responsable de la gestion de la rotation des jetons d'accès SCIM, comme décrit dans la documentation relative au provisionnement automatique. En outre, vous êtes responsable de la rotation des certificats prenant en charge votre fédération d'identité.

Conseils pour les utilisateurs de vos applications Vous pouvez gérer les identités des utilisateurs de vos applications, telles qu'une application mobile, en utilisant Amazon Cognito comme fournisseur d'identité centralisé. Amazon Cognito assure l'authentification, l'autorisation et la gestion des utilisateurs pour vos applications Web et mobiles. Amazon Cognito fournit une banque d'identités adaptée à des millions d'utilisateurs, prend en charge la fédération d'identité sociale de l'entreprise et propose des fonctionnalités de sécurité avancées pour protéger vos utilisateurs et votre entreprise. Vous pouvez intégrer votre application Web ou mobile personnalisée avec Amazon Cognito pour ajouter l'authentification des utilisateurs et le contrôle d'accès à vos applications en quelques minutes. Fondé sur des normes d'identité ouvertes telles que SAML et OpenID Connect (OIDC), Amazon Cognito prend en charge diverses réglementations de conformité et s'intègre aux ressources de développement front-end et dorsal.

Étapes d'implémentation

Étapes à suivre pour permettre aux utilisateurs en interne d'accéder à AWS

- Fédérez les utilisateurs en interne à AWS à l'aide d'un fournisseur d'identité centralisé en utilisant l'une des approches suivantes :
  - Utilisez IAM Identity Center pour activer l'authentification unique à plusieurs Comptes AWS dans votre organisation AWS en vous fédérant avec votre fournisseur d'identité.
  - Utilisez IAM pour connecter votre fournisseur d'identité directement à chaque Compte AWS, afin de permettre un accès fédéré précis.
- Identifiez et supprimez les utilisateurs IAM et les groupes qui seront remplacés par des identités fédérées.

Étapes à suivre pour les utilisateurs de vos applications

- Utilisez Amazon Cognito comme fournisseur d'identité centralisé pour vos applications.
- Intégrez vos applications personnalisées à Amazon Cognito à l'aide d'OpenID Connect et d'OAuth. Vous pouvez développer vos applications personnalisées à l'aide des bibliothèques Amplify qui fournissent des interfaces simples à intégrer à divers services AWS, tels que Amazon Cognito pour l'authentification.

Ressources

Bonnes pratiques associées :

- SEC02-BP06 Utiliser des groupes d'utilisateurs et des attributs
- SEC03-BP02 Accorder un accès selon le principe du moindre privilège
- SEC03-BP06 Gérer l'accès en fonction du cycle de vie

Documents connexes :

- Fédération d'identité dans AWS
- Bonnes pratiques de sécurité dans IAM
- Bonnes pratiques AWS Identity and Access Management
- Mise en route avec l'administration déléguée d'IAM Identity Center
- Comment utiliser les politiques gérées par le client dans IAM Identity Center pour les cas d'utilisation avancés
- AWS CLI v2 : fournisseur d'informations d'identification IAM Identity Center

Vidéos connexes :

- AWS re:Inforce 2022 - AWS Identity and Access Management (IAM) deep dive
- AWS re:Invent 2022 - Simplify your existing workforce access with IAM Identity Center
- AWS re:Invent 2018: Mastering Identity at Every Layer of the Cake

Exemples associés :

- Atelier : utilisation d'AWS IAM Identity Center pour assurer une gestion forte des identités
- Atelier : identité sans serveur

Outils associés :

- AWS Partenaires disposant de la compétence Sécurité : gestion des identités et des accès
- saml2aws

SEC02-BP05 Contrôler et effectuer régulièrement une rotation des informations d'identification

Contrôlez et effectuez régulièrement une rotation des informations d'identification afin de limiter leur durée d'utilisation pour accéder à vos ressources. Les informations d'identification à long terme créent de nombreux risques, et ces risques peuvent être réduits par une rotation régulière de ces informations.

Résultat escompté : mettre en œuvre la rotation des informations d'identification afin de réduire les risques associés à l'utilisation à long terme des informations d'identification. Auditez et corrigez régulièrement toute non-conformité avec les politiques de rotation des informations d'identification.

Anti-modèles courants :

- Ne pas auditer l'utilisation des informations d'identification.
- Utiliser inutilement des informations d'identification à long terme.
- Utiliser des informations d'identification à long terme et ne pas effectuer de rotation régulièrement.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Lorsque vous ne pouvez pas compter sur des informations d'identification temporaires et que vous avez besoin d'informations d'identification à long terme, auditez les informations d'identification pour vous assurer que les contrôles définis tels que l'authentification multifactorielle (MFA) sont appliqués, qu'ils font l'objet d'une rotation régulière et qu'ils ont le niveau d'accès approprié.

La validation régulière, de préférence via un outil automatisé, est nécessaire pour vérifier que les contrôles corrects sont appliqués. Pour les identités humaines, vous devez obliger les utilisateurs à modifier leurs mots de passe régulièrement et à mettre hors service les clés d'accès au profit d'informations d'identification temporaires. Lorsque vous passez des AWS Identity and Access Management utilisateurs (IAM) aux identités centralisées, vous pouvez générer un rapport d'informations d'identification pour auditer vos utilisateurs.

Nous vous recommandons également d'appliquer les paramètres d'authentification multifactorielle dans votre fournisseur d'identité. Vous pouvez configurer AWS Config Rules, ou utiliser des normes de sécurité AWS Security Hub pour vérifier si les utilisateurs ont configuré l'authentification multifactorielle (MFA). Envisagez d'utiliser Rôles Anywhere IAM afin de fournir des informations d'identification temporaires pour les identités des machines. Lorsque l'utilisation de rôles IAM et d'informations d'identification temporaires n'est pas possible, il est nécessaire de réaliser fréquemment des audits et la rotation des clés d'accès.

Étapes d'implémentation

- Auditer fréquemment des informations d'identification : l'audit des identités configurées dans votre fournisseur d'identités et dans IAM aide à garantir que seules les identités autorisées ont accès à votre charge de travail. Ces identités peuvent inclure, sans s'y limiter, des utilisateurs IAM, des utilisateurs AWS IAM Identity Center, des utilisateurs Active Directory ou des utilisateurs dans un autre fournisseur d'identité en amont. Par exemple, supprimez les personnes qui quittent l'organisation et supprimez les rôles intercomptes qui ne sont plus requis. Mettez en place un processus pour auditer périodiquement les autorisations aux services auxquels accède une entité IAM. Cela vous permet d'identifier les politiques à modifier afin de supprimer les autorisations inutilisées. Utilisez les rapports d'informations d'identification et AWS Identity and Access Management Access Analyzer pour auditer les informations d'identification et les autorisations IAM. Vous pouvez utiliser Amazon CloudWatch pour configurer des alarmes pour des appels d'API spécifiques appelés dans votre environnement AWS. Amazon GuardDuty peut également vous avertir en cas d'activité inattendue, qui peut indiquer un accès trop permissif ou un accès involontaire aux informations d'identification IAM.
- Effectuer la rotation régulière des informations d'identification : lorsque vous ne pouvez pas utiliser d'informations d'identification temporaires, alternez régulièrement les clés d'accès IAM à long terme (maximum tous les 90 jours). Si une clé d'accès est divulguée involontairement à votre insu, cela limite la durée pendant laquelle les informations d'identification peuvent être utilisées pour accéder à vos ressources. Pour plus d'informations sur la rotation des clés d'accès pour les utilisateurs IAM, consultez la rubrique Rotation des clés d'accès.
- Examinez les autorisations IAM : pour améliorer la sécurité de votre Compte AWS, examinez et surveillez régulièrement chacune de vos politiques IAM. Vérifiez que les politiques respectent le principe du moindre privilège.
- Envisagez d'automatiser la création et la mise à jour des ressources IAM : IAM Identity Center automatise de nombreuses tâches IAM, telles que la gestion des rôles et des politiques. Sinon, AWS CloudFormation peut être utilisé afin d'automatiser le déploiement des ressources IAM, y

compris les rôles et les politiques, afin de réduire le risque d'erreur humaine, car les modèles peuvent être vérifiés et la version contrôlée.

- Utilisez Rôles Anywhere IAM pour remplacer les utilisateurs IAM par des identités de machines : Rôles Anywhere IAM vous permet d'utiliser des rôles dans des domaines que vous ne pouviez pas utiliser auparavant, tels que les serveurs sur site. Rôles Anywhere IAM utilise un certificat X.509 approuvé afin de s'authentifier auprès d'AWS et de recevoir des informations d'identification temporaires. L'utilisation de Rôles Anywhere IAM vous évite d'avoir à effectuer des rotations de ces informations d'identification, car les informations d'identification à long terme ne sont plus stockées dans votre environnement sur site. Veuillez noter que vous devrez surveiller et faire tourner le certificat X.509 à l'approche de son expiration.

Ressources

Bonnes pratiques associées :

- SEC02-BP02 Utiliser des informations d'identification temporaires
- SEC02-BP03 Stocker et utiliser des secrets en toute sécurité

Documents connexes :

- Mise en route avec AWS Secrets Manager
- Bonnes pratiques IAM
- Fournisseurs d'identité et fédération
- Solutions partenaires de sécurité : accès et contrôle d'accès
- Informations d'identification de sécurité temporaires
- Obtention de rapports d'informations d'identification pour votre Compte AWS

Vidéos connexes :

- Bonnes pratiques de gestion, d'extraction et de renouvellement des secrets à grande échelle
- Gestion des autorisations des utilisateurs à grande échelle avec IAM Identity Center AWS
- Maîtrise des identités dans chaque couche

Exemples connexes :

- Atelier Well-Architected – Nettoyage automatique des utilisateurs IAM
- Atelier Well-Architected – Déploiement automatisé de groupes et de rôles IAM

SEC02-BP06 Utiliser des groupes d'utilisateurs et des attributs

La définition des autorisations en fonction des groupes d'utilisateurs et des attributs contribue à réduire le nombre et la complexité des politiques, ce qui simplifie la mise en œuvre du principe du moindre privilège. Vous pouvez utiliser des groupes d'utilisateurs pour gérer les autorisations de nombreuses personnes en un seul endroit selon la fonction qu'elles occupent au sein de votre organisation. Les attributs, tels que le service, le projet ou l'emplacement, peuvent fournir une couche supplémentaire de portée des autorisations lorsque des personnes occupent une fonction similaire, mais pour des sous-ensembles de ressources différents.

Résultat escompté : vous pouvez appliquer des modifications aux autorisations selon la fonction de tous les utilisateurs qui exécutent cette fonction. L'appartenance aux groupes et les attributs régissent les autorisations des utilisateurs, ce qui réduit la nécessité de gérer les autorisations au niveau de chaque utilisateur. Les groupes et les attributs que vous définissez dans votre fournisseur d'identité (IdP) sont propagés automatiquement à vos environnements AWS.

Anti-modèles courants :

- Gestion des autorisations pour les utilisateurs individuels et duplication entre de nombreux utilisateurs.
- Définition de groupes à un niveau trop élevé, autorisations trop étendues accordées.
- Définition de groupes à un niveau trop détaillé, ce qui crée des duplications et de la confusion quant à l'appartenance.
- Utilisation de groupes avec des autorisations dupliquées sur des sous-ensembles de ressources lorsque des attributs peuvent être utilisés à la place.
- Aucune gestion de groupes, d'attributs et d'appartenances par le biais d'un fournisseur d'identité standardisé intégré à vos environnements AWS.
- Utilisation du chaînage des rôles lors de l'utilisation de sessions AWS IAM Identity Center

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les autorisations AWS sont définies dans des documents appelés politiques, qui sont associés à un principal, tel qu'un utilisateur, un groupe, un rôle ou une ressource. Vous pouvez mettre à l'échelle la gestion des autorisations en organisant les attributions d'autorisations (groupe, autorisations, compte) en fonction de la fonction, de la charge de travail et de l'environnement SDLC. En ce qui concerne le personnel, cela vous permet de définir des groupes selon la fonction occupée par les utilisateurs au sein de votre organisation, plutôt que selon les ressources auxquelles ils accèdent. Par exemple, un groupe WebAppDeveloper peut être associé à une politique pour configurer des services tels qu'Amazon CloudFront au sein d'un compte de développement. Un groupe `AutomationDeveloper` peut avoir des autorisations qui se chevauchent avec le groupe `WebAppDeveloper`. Ces autorisations communes peuvent être saisies dans une politique distincte et associées aux deux groupes, au lieu de faire en sorte que les utilisateurs des deux fonctions appartiennent à un groupe `CloudFrontAccess`.

Outre les groupes, vous pouvez utiliser des attributs pour élargir l'accès. Par exemple, vous pouvez avoir un attribut de projet permettant aux utilisateurs de votre groupe `WebAppDeveloper` de définir l'accès aux ressources spécifiques à leur projet. L'utilisation de cette technique élimine la nécessité de créer différents groupes pour les développeurs d'applications qui travaillent sur différents projets si leurs autorisations sont par ailleurs les mêmes. La façon dont vous faites référence aux attributs dans les politiques d'autorisation dépend de leur source, qu'ils soient définis dans le cadre de votre protocole de fédération (tel que SAML, OIDC ou SCIM), en tant qu'assertions SAML personnalisées ou définis dans le cadre d'IAM Identity Center.

Étapes d'implémentation

1. Déterminez où vous allez définir les groupes et les attributs :
    a. En suivant les instructions fournies dans SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé, vous pouvez déterminer si vous devez définir des groupes et des attributs au sein de votre fournisseur d'identité, dans IAM Identity Center ou utiliser des groupes d'utilisateurs IAM dans un compte spécifique.
2. Définissez des groupes :
    a. Déterminez vos groupes selon la fonction et la portée de l'accès requis. Envisagez d'utiliser une structure hiérarchique ou des conventions de dénomination pour organiser efficacement les groupes.
    b. Si vous optez pour une définition au sein d'IAM Identity Center, créez des groupes et associez le niveau d'accès souhaité à l'aide d'ensembles d'autorisations.

c. Si vous optez pour une définition au sein d'un fournisseur d'identité externe, déterminez si le fournisseur prend en charge le protocole SCIM et envisagez d'activer le provisionnement automatique au sein d'IAM Identity Center. Cette capacité synchronise la création, l'appartenance et la suppression de groupes entre votre fournisseur et IAM Identity Center.

3. Définissez des attributs :

a. Si vous utilisez un fournisseur d'identité externe, les protocoles SCIM et SAML 2.0 fournissent certains attributs par défaut. Des attributs supplémentaires peuvent être définis et transmis à l'aide d'assertions SAML utilisant le nom de l'attribut `https://aws.amazon.com/SAML/Attributes/PrincipalTag`. Consultez la documentation de votre fournisseur d'identité pour obtenir des recommandations quant à la définition et la configuration d'attributs personnalisés.

b. Si vous définissez des rôles dans IAM Identity Center, activez la fonctionnalité de contrôle d'accès par attributs (ABAC) et définissez les attributs comme vous le souhaitez. Tenez compte des attributs qui correspondent à la stratégie de balisage des ressources ou à la structure de votre organisation.

Si vous avez besoin d'un chaînage des rôles IAM à partir des rôles IAM assumés via IAM Identity Center, les valeurs telles que `source-identity` et `principal-tags` ne se propagent pas. Pour plus de détails, consultez Activer et configurer les attributs pour le contrôle d'accès.

1. Déterminez la portée des autorisations en fonction des groupes et des attributs :

a. Envisagez d'inclure dans vos politiques d'autorisation des conditions qui comparent les attributs de votre principal à ceux des ressources auxquelles vous accédez. Par exemple, vous pouvez définir une condition pour autoriser l'accès à une ressource uniquement si la valeur d'une clé de condition `PrincipalTag` correspond à la valeur d'une clé `ResourceTag` du même nom.

b. Lorsque vous définissez des politiques ABAC, suivez les instructions figurant dans les bonnes pratiques et les exemples relatifs à l'autorisation ABAC.

c. Passez régulièrement en revue et mettez à jour la structure de votre groupe et de vos attributs au fur et à mesure de l'évolution des besoins de votre organisation afin de garantir une gestion optimale des autorisations.

Ressources

Bonnes pratiques associées :

- SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé

- SEC03-BP02 Accorder un accès selon le principe du moindre privilège
- COST02-BP04 Mettre en œuvre des groupes et des rôles

Documents connexes :

- Bonnes pratiques IAM
- Gestion des identités dans IAM Identity Center
- Qu'est-ce que le contrôle d'accès par attributs (ABAC) pour AWS ?
- Contrôle d'accès par attributs (ABAC) dans IAM Identity Center
- Exemples de politique ABAC

Vidéos connexes :

- Gestion des autorisations des utilisateurs à grande échelle avec IAM Identity Center AWS
- Maîtrise des identités dans chaque couche

SÉC 3. Comment gérer les autorisations des personnes et des machines ?

Gérez les autorisations des identités humaines et machines qui nécessitent un accès à AWS ainsi qu'à votre charge de travail. Les autorisations vous permettent de contrôler qui peut accéder à quoi et dans quelles conditions. En définissant des autorisations pour des identités humaines et des identités de machines spécifiques, vous leur donnez accès à des actions de service spécifiques sur des ressources spécifiques. En outre, vous pouvez spécifier les conditions qui doivent être remplies pour que l'accès soit accordé.

Bonnes pratiques

- SEC03-BP01 Définir les conditions d'accès
- SEC03-BP02 Accorder un accès selon le principe du moindre privilège
- SEC03-BP03 Établir un processus d'accès d'urgence
- SEC03-BP04 Limiter les autorisations au minimum requis en permanence
- SEC03-BP05 Définir des garde-fous des autorisations pour votre organisation
- SEC03-BP06 Gérer l'accès en fonction du cycle de vie
- SEC03-BP07 Analyser l'accès public et intercompte
- SEC03-BP08 Partager des ressources en toute sécurité au sein de votre organisation

- SEC03-BP09 Partager des ressources en toute sécurité avec un tiers

SEC03-BP01 Définir les conditions d'accès

Chaque composant ou ressource de votre charge de travail doit être accessible aux administrateurs, aux utilisateurs finaux ou à d'autres composants. Définissez clairement qui ou quoi doit avoir accès à chaque composant, choisissez le type d'identité approprié et la méthode d'authentification et d'autorisation.

Anti-modèles courants :

- Codage en dur ou stockage de secrets dans votre application.
- Octroi d'autorisations personnalisées à chaque utilisateur.
- Utilisation d'informations d'identification durables.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Chaque composant ou ressource de votre charge de travail doit être accessible aux administrateurs, aux utilisateurs finaux ou à d'autres composants. Définissez clairement qui ou quoi doit avoir accès à chaque composant, choisissez le type d'identité approprié et la méthode d'authentification et d'autorisation.

L'accès régulier aux Comptes AWS au sein d'une organisation doit être assuré par un accès fédéré ou un fournisseur d'identité centralisé. Vous devez également centraliser la gestion des identités et vous assurer qu'il existe une pratique établie pour intégrer l'accès à AWS au cycle de vie de l'accès des employés. Par exemple, lorsqu'un employé change de poste et de niveau d'accès, son appartenance à un groupe doit également évoluer de façon à refléter les nouvelles conditions d'accès qui lui sont associées.

Lorsque vous définissez des conditions d'accès pour des identités non humaines, déterminez quels applications et composants ont besoin d'un accès et comment les autorisations sont accordées. Dans cette optique, il est recommandé d'utiliser les rôles IAM créés avec le modèle d'accès du moindre privilège. AWS Les politiques gérées établissent des politiques IAM prédéfinies qui couvrent les cas d'utilisation les plus courants.

Les services AWS, tels que AWS Secrets Manager et AWS Systems Manager Parameter Store, peuvent aider à dissocier les secrets de l'application ou de la charge de travail en toute sécurité dans

les cas où il n'est pas possible d'utiliser des rôles IAM. Dans Secrets Manager, vous pouvez établir une rotation automatique de vos informations d'identification. Vous pouvez utiliser Secrets Manager de façon à référencer les paramètres dans vos scripts, commandes, documents SSM, configuration et flux de travail d'automatisation en utilisant le nom unique que vous avez spécifié lors de la création du paramètre.

Vous pouvez utiliser Rôles Anywhere AWS IAM pour obtenir des informations d'identification de sécurité temporaires dans IAM pour les charges de travail qui s'exécutent en dehors d'AWS. Vos charges de travail peuvent utiliser les mêmes politiques et rôles IAM que ceux que vous utilisez avec les applications AWS pour accéder aux ressources AWS.

Dans la mesure du possible, privilégiez les informations d'identification temporaires à court terme plutôt que les informations d'identification statiques à long terme. Pour les scénarios dans lesquels vous avez besoin d'utilisateurs disposant d'un accès programmatique et d'informations d'identification à long terme, utilisez les dernières informations utilisées concernant les clés d'accès pour faire pivoter et supprimer les clés d'accès.

Les utilisateurs ont besoin d'un accès programmatique s'ils souhaitent interagir avec AWS en dehors de la AWS Management Console. La manière d'octroyer un accès par programmation dépend du type d'utilisateur qui accède à AWS.

Pour accorder aux utilisateurs un accès programmatique, choisissez l'une des options suivantes.

Quel utilisateur a besoin d'un accès programmatique ?	Pour	Par
Identité de la main-d'œuvre (Utilisateurs gérés dans IAM Identity Center)	Utilisez des informations d'identification temporaires pour signer des demandes par programmation destinées à la AWS CLI, aux AWS SDK ou aux API AWS.	Suivez les instructions de l'interface que vous souhaitez utiliser. • Pour l'AWS CLI, consultez la rubrique Configuration de l'AWS CLI pour l'utilisation d'AWS IAM Identity Center dans le Guide de l'utilisateur AWS Command Line Interface. • Pour les kits SDK et les outils AWS ainsi que les API

Quel utilisateur a besoin d'un accès programmatique ?	Pour	Par
		AWS, consultez la rubrique Authentification IAM Identity Center dans le Guide de référence des kits SDK et des outils AWS.
IAM	Utilisez des informations d'identification temporaires pour signer des demandes par programmation destinées à la AWS CLI, aux AWS SDK ou aux API AWS.	Suivez les instructions de la section Utilisation d'informations d'identification temporaires avec des ressources AWS dans le Guide de l'utilisateur IAM.

Quel utilisateur a besoin d'un accès programmatique ?	Pour	Par
IAM	(Non recommandé) Utilisez des informations d'identification à long terme pour signer des demandes par programmation destinées à la AWS CLI, aux AWS SDK ou aux API AWS.	Suivez les instructions de l'interface que vous souhaitez utiliser. • Pour l'AWS CLI, consultez la rubrique Authentification à l'aide des informations d'identification d'utilisateur IAM dans le Guide de l'utilisateur AWS Command Line Interface. • Pour les kits SDK et les outils AWS, consultez la rubrique Authentification à l'aide d'informations d'identification à long terme dans le Guide de référence des kits SDK et des outils AWS. • Pour les API AWS, consultez la rubrique Gestion des clés d'accès pour les utilisateurs IAM dans le Guide de l'utilisateur IAM.

Ressources

Documents connexes :

- Contrôle d'accès par attributs (ABAC)
- AWS IAM Identity Center
- Rôles Anywhere IAM
- Politiques gérées AWS pour IAM Identity Center

- Conditions des politiques AWS IAM
- Cas d'utilisation d'IAM
- Supprimer les informations d'identification inutiles
- Utilisation de stratégies
- How to control access to AWS resources based on Compte AWS, OU, or organization
- Identify, arrange, and manage secrets easily using enhanced search in AWS Secrets Manager

Vidéos connexes :

- Become an IAM Policy Master in 60 Minutes or Less
- Separation of Duties, Least Privilege, Delegation, and CI/CD
- Streamlining identity and access management for innovation

SEC03-BP02 Accorder un accès selon le principe du moindre privilège

Accordez uniquement l'accès dont les utilisateurs ont besoin pour effectuer des actions spécifiques sur des ressources spécifiques dans des conditions spécifiques. Faites appel à des groupes et des attributs d'identité pour définir de façon dynamique des autorisations à grande échelle, plutôt que pour des utilisateurs individuels. Par exemple, vous pouvez autoriser un groupe de développeurs à gérer uniquement les ressources de leur projet. Ainsi, si un développeur quitte le projet, son accès est automatiquement révoqué sans que les stratégies d'accès sous-jacentes soient modifiées.

Résultat escompté : les utilisateurs ne disposent que des autorisations minimales requises pour leurs fonctions professionnelles spécifiques. Vous utilisez des Comptes AWS séparés pour isoler les développeurs des environnements de production. Lorsque les développeurs ont besoin d'accéder à des environnements de production pour des tâches spécifiques, un accès limité et contrôlé leur est accordé seulement pour la durée de ces tâches. Leur accès en production est immédiatement révoqué une fois qu'ils ont terminé les travaux nécessaires. Vous révisez régulièrement les autorisations et vous les révoquez rapidement lorsqu'elles ne sont plus nécessaires, par exemple lorsqu'un utilisateur change de rôle ou quitte l'organisation. Vous limitez les privilèges d'administrateur à un petit groupe de confiance afin de réduire l'exposition aux risques. Vous accordez aux comptes de machines ou de systèmes uniquement les autorisations minimales requises pour effectuer les tâches prévues.

Anti-modèles courants :

- Par défaut, vous accordez des autorisations d'administrateur aux utilisateurs.
- Vous utilisez le compte d'utilisateur racine pour les activités quotidiennes.
- Vous créez des politiques trop permissives sans les délimiter correctement.
- Vos révisions d'autorisations sont peu fréquentes, ce qui entraîne des dérives.
- Vous vous appuyez uniquement sur un contrôle d'accès basé sur les attributs pour l'isolation de l'environnement ou la gestion des autorisations.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Le principe du moindre privilège stipule que les identités ne devraient être autorisées à effectuer que le plus petit ensemble d'actions nécessaires pour accomplir une tâche spécifique. Il permet d'atteindre un équilibre entre la convivialité, l'efficacité et la sécurité. Le respect de ce principe permet de limiter l'accès non intentionnel et de déterminer qui a accès aux ressources. Par défaut, les utilisateurs et les rôles IAM ne disposent d'aucune autorisation. L'utilisateur racine dispose d'un accès complet par défaut et doit être étroitement contrôlé, surveillé et utilisé uniquement pour les tâches nécessitant un accès racine.

Les politiques IAM sont utilisées pour octroyer explicitement des autorisations aux rôles IAM ou à des ressources spécifiques. Par exemple, les politiques basées sur l'identité peuvent être attachées à des groupes IAM, tandis que les compartiments S3 peuvent être contrôlés par des politiques basées sur les ressources.

Lorsque vous créez une politique IAM, vous pouvez spécifier les actions de service, les ressources et les conditions qui doivent être remplies pour qu'AWS autorise ou refuse l'accès. AWS prend en charge diverses conditions pour vous aider à limiter l'accès. Par exemple, en utilisant la clé de condition PrincipalOrgID, vous pouvez refuser des actions si le demandeur ne fait pas partie de votre organisation AWS.

Vous pouvez également contrôler les demandes que les services AWS effectuent en votre nom, comme AWS CloudFormation qui crée une fonction AWS Lambda, à l'aide de la clé de condition CalledVia. Vous pouvez superposer différents types de politiques pour établir une défense en profondeur et limiter les autorisations globales de vos utilisateurs. Vous pouvez également limiter les autorisations qui peuvent être accordées et sous quelles conditions. Par exemple, vous pouvez autoriser vos équipes responsables de la charge de travail à créer leurs propres politiques IAM pour les systèmes qu'elles créent, mais seulement si elles appliquent une limite des autorisations afin de limiter le nombre maximal d'autorisations qu'elles peuvent accorder.

Étapes d'implémentation

- Implémentez des politiques de moindre privilège : attribuez des stratégies d'accès avec le moindre privilège aux groupes et rôles IAM pour rester cohérent avec le rôle ou la fonction de l'utilisateur que vous avez défini.
- Isolez les environnements de développement et de production via des Comptes AWS séparés : utilisez des Comptes AWS séparés pour les environnements de développement et de production, et contrôlez l'accès entre eux à l'aide de politiques de contrôle des services, de politiques de ressources et de politiques d'identité.
- Politiques de base relatives à l'utilisation des API : l'un des moyens de déterminer les autorisations nécessaires consiste à consulter les journaux AWS CloudTrail. Vous pouvez utiliser cette révision pour créer des autorisations adaptées aux actions effectivement réalisées par l'utilisateur dans AWS. IAM Access Analyzer peut générer automatiquement une politique IAM basée sur l'activité d'accès. Vous pouvez utiliser IAM Access Advisor au niveau de l'organisation ou du compte pour suivre les dernières informations consultées pour une stratégie donnée.
- Envisagez d'utiliser des politiques gérées par AWS pour les fonctions professionnelles : lorsque vous commencez à créer des politiques d'autorisations détaillées, il peut être utile d'utiliser des politiques gérées par AWS pour les rôles professionnels courants, tels que la facturation, les administrateurs de base de données et les scientifiques des données. Ces politiques peuvent permettre de restreindre l'accès des utilisateurs en déterminant comment mettre en œuvre les politiques de moindre privilège.
- Supprimez les autorisations inutiles : détectez et supprimez les entités, les informations d'identification et les autorisations IAM inutilisées afin d'appliquer le principe du moindre privilège. Vous pouvez utiliser l'Analyseur d'accès IAM pour identifier les accès externes et non utilisés, et la génération de politiques de l'Analyseur d'accès IAM peut aider à optimiser les politiques d'autorisation.
- S'assurer que les utilisateurs ont un accès limité aux environnements de production : les utilisateurs ne doivent avoir accès qu'aux environnements de production présentant un cas d'utilisation valide. Une fois que l'utilisateur a effectué les tâches précises qui nécessitent un accès en production, cet accès doit être révoqué. Le fait de limiter l'accès aux environnements de production permet de prévenir les événements imprévus ayant une incidence sur la production et réduit la portée des répercussions de l'accès involontaire.
- Envisagez les limites des autorisations : une limite des autorisations est une fonctionnalité permettant d'utiliser une politique gérée qui définit les autorisations maximales qu'une politique basée sur l'identité peut accorder à une entité IAM. La limite des autorisations d'une entité lui

permet d'effectuer uniquement les actions autorisées à la fois par ses politiques basées sur l'identité et ses limites des autorisations.

- Affinez l'accès à l'aide du contrôle d'accès basé sur les attributs et des balises de ressources : un contrôle d'accès basé sur les attributs (ABAC) utilisant des balises de ressources peut être utilisé pour affiner les autorisations lorsqu'il est pris en charge. Vous pouvez utiliser un modèle ABAC qui compare les balises principales aux balises de ressources pour affiner l'accès en fonction des dimensions personnalisées que vous définissez. Cette approche permet de simplifier et de réduire le nombre de politiques d'autorisation au sein de votre organisation.
  - Il est recommandé d'utiliser uniquement ABAC pour le contrôle d'accès lorsque les principaux et les ressources appartiennent à votre organisation AWS. Les parties externes peuvent utiliser les mêmes noms de balises et les mêmes valeurs que votre organisation pour leurs propres principaux et ressources. Si vous vous appuyez uniquement sur ces paires nom-valeur pour accorder l'accès à des principaux ou à des ressources externes, vous pouvez fournir des autorisations involontaires.
- Utilisez des politiques de contrôle des services pour AWS Organizations : les politiques de contrôle des services contrôlent de façon centralisée les autorisations disponibles maximales pour les comptes membres de votre organisation. Il est important de noter que vous pouvez utiliser les politiques de contrôle des services pour limiter les autorisations des utilisateurs racine dans les comptes membres. Envisagez également d'utiliser AWS Control Tower, qui fournit des contrôles gérés normatifs permettant d'enrichir AWS Organizations. Vous pouvez également définir vos propres contrôles dans Control Tower.
- Établissez une politique de cycle de vie des utilisateurs pour votre organisation : les politiques de cycle de vie des utilisateurs définissent les tâches à effectuer lorsque les utilisateurs sont intégrés à AWS, changent de rôle ou de champ d'activité, ou n'ont plus besoin d'accéder à AWS. Examinez les autorisations à chaque étape du cycle de vie d'un utilisateur pour vous assurer qu'elles sont suffisamment restrictives et éviter les dérives.
- Établissez un calendrier régulier pour vérifier les autorisations et supprimer les autorisations inutiles : vous devez régulièrement vérifier l'accès des utilisateurs pour vérifier qu'il n'est pas trop permissif. AWS Config et l'Analyseur d'accès IAM peuvent aider lors des audits des autorisations des utilisateurs.
- Établissez une matrice des rôles de tâches : une matrice des rôles de tâches permet de visualiser les différents rôles et niveaux d'accès requis au sein de votre couverture AWS. À l'aide d'une matrice des fonctions professionnelles, vous pouvez définir et séparer les autorisations en fonction des responsabilités des utilisateurs au sein de votre organisation. Utilisez des groupes au lieu d'appliquer des autorisations directement à des utilisateurs ou à des rôles individuels.

Ressources

Documents connexes :

- Accorder le moindre privilège
- Limites des autorisations pour les entités IAM
- Techniques for writing least privilege IAM policies
- IAM Access Analyzer facilite la mise en œuvre d'autorisations de moindre privilège en générant des politiques IAM en fonction de l'activité d'accès
- Déléguer la gestion des autorisations aux développeurs en utilisant les limites des autorisations IAM
- Ajustement des autorisations à l'aide des dernières informations consultées
- IAM policy types and when to use them
- Test des politiques IAM avec le simulateur de politiques IAM
- Barrières de protection dans AWS Control Tower
- Zero Trust architectures: An AWS perspective
- How to implement the principle of least privilege with CloudFormation StackSets
- Contrôle d'accès par attributs (ABAC)
- Réduction de la portée de la stratégie en affichant l'activité des utilisateurs
- Afficher l'accès au rôle
- Use Tagging to Organize Your Environment and Drive Accountability
- Stratégies de balisage AWS
- Balisage de ressources AWS

Vidéos connexes :

- Next-generation permissions management
- Zero Trust: An AWS perspective

Exemples connexes :

- Atelier : Limites des autorisations IAM et délégation de la création de rôles
- Atelier : Contrôle d'accès basé sur les balises IAM pour EC2

SEC03-BP03 Établir un processus d’accès d’urgence

Élaborez un processus permettant un accès d’urgence à vos charges de travail dans le cas peu probable où un problème avec votre fournisseur d’identité centralisé surviendrait.

Vous devez concevoir des processus pour les différents modes de défaillance susceptibles de provoquer un événement d’urgence. Par exemple, dans des circonstances normales, les utilisateurs en interne se fédèrent au cloud à l’aide d’un fournisseur d’identité centralisé (SEC02-BP04) pour gérer leurs charges de travail. Toutefois, si votre fournisseur d’identité centralisé échoue ou si la configuration de la fédération dans le cloud est modifiée, les utilisateurs en interne risquent de ne pas parvenir à se fédérer dans le cloud. Un processus d’accès d’urgence permet aux administrateurs autorisés d’accéder à vos ressources cloud par d’autres moyens (tels qu’une autre forme de fédération ou un accès utilisateur direct) afin de résoudre les problèmes liés à la configuration de la fédération ou à vos charges de travail. Le processus d’accès d’urgence est utilisé jusqu’à ce que le mécanisme de fédération normal soit rétabli.

Résultat escompté :

- Vous avez défini et documenté les modes de défaillance considérés comme une urgence : envisagez les circonstances habituelles et les systèmes dont dépendent vos utilisateurs pour gérer leurs charges de travail. Réfléchissez à la façon dont chacune de ces dépendances peut échouer et provoquer une situation d’urgence. Les questions et les bonnes pratiques du pilier Fiabilité peuvent vous être utiles pour identifier les modes de défaillance et concevoir des systèmes plus résilients afin de minimiser le risque de défaillance.
- Vous avez documenté les étapes à suivre pour confirmer qu’une défaillance est une urgence. Par exemple, vous pouvez demander aux administrateurs d’identité de vérifier l’état des fournisseurs d’identité principal et secondaire et, si les deux ne sont pas disponibles, de déclarer un événement d’urgence pour cause de défaillance du fournisseur d’identité.
- Vous avez défini un processus d’accès d’urgence spécifique à chaque type d’urgence ou de mode de défaillance. En étant aussi précis que possible, vous éviterez que les utilisateurs abusent d’un processus général pour tous les types d’urgence. Vos processus d’accès d’urgence décrivent les circonstances dans lesquelles chaque processus doit être utilisé, et inversement les situations dans lesquelles le processus ne doit pas être utilisé et renvoie à d’autres processus qui peuvent s’appliquer.
- Vos processus sont bien documentés avec des instructions détaillées et des playbooks qui peuvent être suivis rapidement et efficacement. N’oubliez pas qu’un événement d’urgence peut être stressant pour vos utilisateurs et qu’ils peuvent être soumis à des contraintes de temps extrêmes. Concevez donc votre processus de manière à ce qu’il soit aussi simple que possible.

Anti-modèles courants :

- Vous ne disposez pas de processus d'accès d'urgence bien documentés et bien testés. Vos utilisateurs ne sont pas préparés à une situation d'urgence et suivent des processus improvisés lorsqu'une situation d'urgence survient.
- Vos processus d'accès d'urgence dépendent des mêmes systèmes (tels qu'un fournisseur d'identité centralisé) que vos mécanismes d'accès habituels. Autrement dit, la défaillance d'un système de ce type peut avoir un impact à la fois sur vos mécanismes d'accès habituels et sur les mécanismes d'accès d'urgence, et nuire à votre capacité à vous remettre de la panne.
- Vos processus d'accès d'urgence sont utilisés dans des situations non urgentes. Par exemple, vos utilisateurs utilisent fréquemment à mauvais escient les processus d'accès d'urgence, car ils trouvent qu'il est plus facile d'apporter des modifications directement que de les soumettre par le biais d'un pipeline.
- Vos processus d'accès d'urgence ne génèrent pas suffisamment de journaux pour auditer les processus, ou les journaux ne sont pas surveillés pour signaler une éventuelle utilisation inappropriée des processus.

Avantages liés au respect de cette bonne pratique :

- En disposant de processus d'accès d'urgence bien documentés et bien testés, vous réduisez le temps nécessaire à vos utilisateurs pour répondre à un événement d'urgence et le résoudre. Cela peut se traduire par une réduction des temps d'arrêt et une meilleure disponibilité des services que vous offrez à vos clients.
- Vous pouvez suivre chaque demande d'accès d'urgence, détecter les tentatives non autorisées d'utilisation abusive du processus pour des événements non urgents et les signaler.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Cette section fournit des conseils pour créer des processus d'accès d'urgence pour plusieurs modes de défaillance liés aux charges de travail déployées sur AWS, en commençant par des conseils communs applicables à tous les modes de défaillance, suivis de directives spécifiques basées sur le type de mode de défaillance.

Conseils communs pour tous les modes de défaillance

Envisagez les points suivants lorsque vous concevez un processus d'accès d'urgence pour un mode de défaillance :

- Documentez les conditions préalables et les hypothèses du processus : situations dans lesquelles le processus doit être utilisé et situations dans lesquelles il ne doit pas être utilisé. Il est utile de détailler le mode de défaillance et de documenter les hypothèses, telles que l'état d'autres systèmes connexes. Par exemple, le processus du mode de défaillance 2 suppose que le fournisseur d'identité est disponible, mais que la configuration sur AWS est modifiée ou a expiré.
- Créez au préalable les ressources nécessaires au processus d'accès d'urgence (SEC10-BP05). Par exemple, créez au préalable le Compte AWS d'accès d'urgence avec les rôles et utilisateurs IAM, ainsi que les rôles IAM entre comptes dans tous les comptes de la charge de travail. Vous pourrez ainsi vérifier que ces ressources sont prêtes et disponibles en cas d'urgence. En créant des ressources au préalable, vous n'êtes pas tributaire des API du plan de contrôle AWS (utilisées pour créer et modifier des ressources AWS) qui peuvent ne pas être disponibles en cas d'urgence. De plus, en créant au préalable les ressources IAM, vous n'avez pas besoin de prendre en compte les retards potentiels dus à une éventuelle cohérence.
- Incluez les processus d'accès d'urgence dans vos plans de gestion des incidents (SEC10-BP02). Documentez la manière dont les événements d'urgence sont suivis et communiqués aux autres membres de votre organisation (tels que vos pairs et la direction) et, le cas échéant, à vos clients et partenaires commerciaux.
- Définissez le processus de demande d'accès d'urgence dans votre système de flux de travail des demandes de service existant, si vous en avez un. Généralement, ces systèmes de flux de travail vous permettent de créer des formulaires de réception pour collecter des informations sur la demande, de suivre la demande à chaque étape du flux de travail et d'ajouter des étapes d'approbation automatisées et manuelles. Associez chaque demande à un événement d'urgence correspondant suivi dans votre système de gestion des incidents. Le fait de disposer d'un système uniforme pour les accès d'urgence vous permet de suivre ces demandes dans un seul système, d'analyser les tendances d'utilisation et d'améliorer vos processus.
- Vérifiez que vos processus d'accès d'urgence ne peuvent être initiés que par des utilisateurs autorisés et nécessitent l'approbation de pairs ou de la direction de l'utilisateur, le cas échéant. Le processus d'approbation doit fonctionner efficacement pendant les heures de bureau et au-delà. Définissez comment les demandes d'approbation autorisent les approbateurs secondaires si les approbateurs principaux ne sont pas disponibles et comment elles remontent dans la chaîne de gestion jusqu'à ce qu'elles soient approuvées.
- Mettez en œuvre des mécanismes robustes de journalisation, de surveillance et d'alerte pour le processus et les mécanismes d'accès d'urgence. Générez des journaux d'audit détaillés pour

toutes les tentatives réussies et échouées d'obtenir un accès d'urgence. Établissez une corrélation entre l'activité et les événements d'urgence en cours à partir de votre système de gestion des incidents, et lancez des alertes lorsque des actions se produisent en dehors des périodes prévues ou lorsque le compte d'accès d'urgence est utilisé pendant les opérations normales. Le compte d'accès d'urgence ne doit être accessible qu'en cas d'urgence, car les procédures de bris de glace peuvent être considérées comme une porte dérobée. Intégrez-le à votre outil de gestion des informations et des événements de sécurité (SIEM) ou à AWS Security Hubpour signaler et auditer toutes les activités pendant la période d'accès d'urgence. À la reprise des activités normales, effectuez une rotation automatique des informations d'identification d'accès d'urgence et informez les équipes concernées.

- Testez régulièrement les processus d'accès d'urgence pour vérifier que les étapes sont claires et accorder le niveau d'accès approprié rapidement et efficacement. Vos processus d'accès d'urgence doivent être testés dans le cadre de simulations de réponse aux incidents (SEC10-BP07) et de tests de reprise après sinistre (REL13-BP03).

Mode de défaillance 1 : le fournisseur d'identité utilisé pour la fédération à AWS n'est pas disponible

Comme décrit dans SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé, nous vous recommandons de faire appel à un fournisseur d'identité centralisé pour fédérer les utilisateurs en interne et accorder l'accès aux Comptes AWS. Vous pouvez fédérer les utilisateurs à plusieurs Comptes AWS au sein de votre organisation AWS à l'aide d'IAM Identity Center, ou vous pouvez les fédérer à des Comptes AWS individuels avec IAM. Dans les deux cas, les utilisateurs en interne s'authentifient auprès de votre fournisseur d'identité centralisé avant d'être redirigés vers un point de terminaison de connexion AWS pour l'authentification unique.

Dans le cas peu probable où votre fournisseur d'identité centralisé ne serait pas disponible, les utilisateurs en interne ne pourraient pas se fédérer aux Comptes AWS ni gérer leurs charges de travail. Dans ce cas d'urgence, vous pouvez fournir un processus d'accès d'urgence permettant à un petit groupe d'administrateurs d'accéder aux Comptes AWS pour effectuer des tâches critiques qui ne peuvent pas attendre que vos fournisseurs d'identité centralisés soient de nouveau disponibles. Par exemple, votre fournisseur d'identité n'est pas disponible pendant 4 heures et, durant cette période, vous devez modifier les limites supérieures d'un groupe Amazon EC2 Auto Scaling dans un compte de production pour faire face à un pic inattendu du trafic client. Vos administrateurs d'urgence doivent suivre le processus d'accès d'urgence pour accéder au Compte AWS de production spécifique et apporter les modifications nécessaires.

Le processus d'accès d'urgence repose sur un Compte AWS d'accès d'urgence créé au préalable, qui est utilisé uniquement pour l'accès d'urgence et dispose de ressources AWS (comme les rôles IAM et les utilisateurs IAM) pour soutenir le processus d'accès d'urgence. Pendant les opérations normales, personne ne doit accéder au compte d'accès d'urgence et vous devez surveiller et signaler tout cas d'utilisation abusive de ce compte (pour plus de détails, consultez la section précédente consacrée aux conseils communs).

Le compte d'accès d'urgence possède des rôles IAM d'accès d'urgence autorisés à endosser des rôles entre comptes dans les Comptes AWS nécessitant un accès d'urgence. Ces rôles IAM sont créés au préalable et configurés avec des politiques d'approbation qui assurent la validité des rôles IAM du compte d'urgence.

Le processus d'accès d'urgence peut utiliser l'une des approches suivantes :

- Vous pouvez créer au préalable un ensemble d'utilisateurs IAM pour vos administrateurs d'urgence dans le compte d'accès d'urgence avec des mots de passe forts et des jetons MFA associés. Ces utilisateurs IAM seront autorisés à endosser les rôles IAM qui autoriseront ensuite l'accès intercompte au Compte AWS où un accès d'urgence est requis. Nous vous recommandons de créer le moins d'utilisateurs possible et d'affecter chaque utilisateur à un seul administrateur d'urgence. En cas d'urgence, un utilisateur administrateur d'urgence se connecte au compte d'accès d'urgence à l'aide de son mot de passe et de son code de jeton MFA, passe au rôle IAM d'accès d'urgence dans le compte d'urgence, puis passe au rôle IAM d'accès d'urgence dans le compte de la charge de travail pour effectuer l'action de modification d'urgence. L'avantage de cette approche est que chaque utilisateur IAM est associé à un seul administrateur d'urgence et que vous pouvez savoir quel utilisateur s'est connecté en consultant les événements CloudTrail. L'inconvénient est que vous devez gérer plusieurs utilisateurs IAM avec leurs mots de passe de longue durée de vie et leurs jetons MFA associés.
- Vous pouvez utiliser l'utilisateur racine Compte AWS d'accès d'urgence pour vous connecter au compte d'accès d'urgence, endosser le rôle IAM d'accès d'urgence et endosser le rôle entre comptes dans le compte de la charge de travail. Nous recommandons de définir un mot de passe fort et plusieurs jetons MFA pour l'utilisateur racine. Nous conseillons également de stocker le mot de passe et les jetons MFA dans un coffre-fort d'informations d'identification d'entreprise sécurisé qui applique des mécanismes solides d'authentification et d'autorisation. Vous devez sécuriser les facteurs de réinitialisation des mots de passe et des jetons MFA : configurez l'adresse e-mail du compte sur une liste de distribution surveillée par vos administrateurs de sécurité cloud, et le numéro de téléphone du compte doit être un numéro partagé également surveillé par ces administrateurs. L'avantage de cette approche est qu'il n'existe qu'un seul ensemble d'informations d'identification d'utilisateur racine à gérer. L'inconvénient est qu'étant donné qu'il

s’agit d’un utilisateur partagé, plusieurs administrateurs ont la possibilité de se connecter en tant qu’utilisateur racine. Vous devez auditer les événements de journal de votre coffre-fort d’entreprise pour identifier quel administrateur a extrait le mot de passe de l’utilisateur racine.

Mode de défaillance 2 : la configuration du fournisseur d’identité sur AWS est modifiée ou a expiré

Pour permettre aux utilisateurs en interne de se fédérer aux Comptes AWS, vous pouvez configurer l’IAM Identity Center auprès d’un fournisseur d’identité externe ou créer un fournisseur d’identité IAM (SEC02-BP04). Généralement, vous les configurez en important un document XML de métadonnées SAML fourni par votre fournisseur d’identité. Ce document XML de métadonnées inclut un certificat X.509 correspondant à une clé privée que le fournisseur d’identité utilise pour signer ses assertions SAML.

Ces configurations côté AWS peuvent être modifiées ou supprimées par erreur par un administrateur. Dans un autre scénario, le certificat X.509 importé dans AWS peut expirer, et aucun nouveau fichier XML de métadonnées contenant un nouveau certificat n’a encore été importé dans AWS. Ces deux scénarios peuvent désactiver la fédération des utilisateurs en interne à AWS, ce qui peut entraîner une situation d’urgence.

Dans un tel cas d’urgence, vous pouvez fournir à vos administrateurs d’identité un accès à AWS pour résoudre les problèmes de fédération. Par exemple, votre administrateur d’identité utilisera le processus d’accès d’urgence pour se connecter au Compte AWS d’accès d’urgence, passera à un rôle dans le compte administrateur d’Identity Center et mettra à jour la configuration du fournisseur d’identité externe en important le dernier document XML de métadonnées SAML de votre fournisseur d’identité afin de réactiver la fédération. Une fois la fédération rétablie, les utilisateurs en interne pourront continuer à utiliser le processus d’exploitation habituel pour se fédérer aux comptes de leur charge de travail.

Vous pouvez suivre les approches détaillées dans le précédent mode de défaillance 1 pour créer un processus d’accès d’urgence. Vous pouvez accorder des autorisations de moindre privilège aux administrateurs d’identité pour qu’ils ne puissent accéder qu’au compte administrateur d’Identity Center et effectuer des actions sur Identity Center dans ce compte uniquement.

Mode de défaillance 3 : interruption d’Identity Center

Dans le cas peu probable où un IAM Identity Center ou une Région AWS serait interrompue, nous vous recommandons de créer une configuration que vous pourrez utiliser pour assurer un accès temporaire à la AWS Management Console.

Le processus d'accès d'urgence utilise une fédération directe entre votre fournisseur d'identité et IAM dans un compte d'urgence. Pour plus de détails sur le processus et les considérations de conception, voir Configurer l'accès d'urgence à la AWS Management Console.

Étapes d'implémentation

Étapes communes pour tous les modes de défaillance

- Créez un Compte AWS dédié aux processus d'accès d'urgence. Créez au préalable les ressources IAM nécessaires dans le compte, telles que les rôles IAM ou les utilisateurs IAM et, éventuellement, les fournisseurs d'identité IAM. En outre, créez au préalable des rôles IAM entre comptes dans les Comptes AWS de la charge de travail avec des relations d'approbation avec les rôles IAM correspondants dans le compte d'accès d'urgence. Vous pouvez utiliser AWS CloudFormation StackSets avec AWS Organizations pour créer ces ressources dans les comptes membres de votre organisation.

- Créez des politiques de contrôle des services (SCP) AWS Organizations pour refuser la suppression et la modification des rôles IAM entre comptes dans les Comptes AWS membres.
- Activez CloudTrail pour le Compte AWS d'accès d'urgence et envoyez les événements de suivi vers un compartiment S3 central du Compte AWS de collecte de journaux. Si vous utilisez AWS Control Tower pour configurer et gérer votre environnement AWS multi-comptes, chaque compte que vous créez avec AWS Control Tower ou que vous inscrivez dans AWS Control Tower est activé pour CloudTrail par défaut et envoyé vers un compartiment S3 dans un Compte AWS d'archive de journal dédié.
- Surveillez l'activité du compte d'accès d'urgence en créant des règles EventBridge qui correspondent lors de la connexion à la console et de l'activité de l'API en fonction des rôles IAM d'urgence. Envoyez des notifications à votre centre des opérations de sécurité lorsque des activités se produisent en dehors d'un événement d'urgence en cours suivi dans votre système de gestion des incidents.

Étapes supplémentaires pour le mode de défaillance 1 : le fournisseur d'identité utilisé pour la fédération à AWS n'est pas disponible et pour le mode de défaillance 2 : la configuration du fournisseur d'identité sur AWS est modifiée ou a expiré

- Créez des ressources au préalable en fonction du mécanisme que vous avez choisi pour l'accès d'urgence :

- Utilisation d'utilisateurs IAM : créez au préalable les utilisateurs IAM avec des mots de passe forts et les dispositifs MFA associés.
- Utilisation de l'utilisateur racine du compte d'urgence : configurez l'utilisateur racine avec un mot de passe fort et stockez ce mot de passe dans le coffre-fort d'informations d'identification de votre entreprise. Associez plusieurs appareils MFA physiques à l'utilisateur racine et stockez-les à des emplacements auxquels les membres de votre équipe d'administrateurs d'urgence peuvent accéder rapidement.

Étapes supplémentaires pour le mode de défaillance 3 : interruption d'Identity Center

- Comme décrit dans Définir l'accès d'urgence à la AWS Management Console, dans le Compte AWS d'accès d'urgence, créez un fournisseur d'identité IAM pour activer la fédération SAML directe à partir de votre fournisseur d'identité.
- Créez des groupes d'opérations d'urgence dans votre fournisseur d'identité sans aucun membre.
- Créez des rôles IAM correspondant aux groupes d'opérations d'urgence dans le compte d'accès d'urgence.

Ressources

Bonnes pratiques Well-Architected connexes :

- SEC02-BP04 S'appuyer sur un fournisseur d'identité centralisé
- SEC03-BP02 Accorder un accès selon le principe du moindre privilège
- SEC10-BP02 Développer des plans de gestion des incidents
- SEC10-BP07 Organiser des jeux de rôle

Documents connexes :

- Configurer un accès d'urgence à la AWS Management Console
- Activation de l'accès des utilisateurs fédérés SAML 2.0 à la AWS Management Console
- Accès en mode « bris de glace »

Vidéos connexes :

- AWS re:Invent 2022 - Simplify your existing workforce access with IAM Identity Center

- AWS re:Inforce 2022 - AWS Identity and Access Management (IAM) deep dive

Exemples connexes :

- AWS Break Glass Role
- AWS customer playbook framework
- AWS incident response playbook samples

SEC03-BP04 Limiter les autorisations au minimum requis en permanence

Au fur et à mesure que vos équipes déterminent les accès nécessaires, supprimez les autorisations inutiles et mettez en place des processus de révision afin d'obtenir des autorisations de moindre privilège. Surveillez et supprimez en permanence les identités et autorisations inutilisées pour les accès humains et machines.

Résultat escompté : les politiques d'autorisation doivent respecter le principe du moindre privilège. Au fur et à mesure que les tâches et les rôles sont mieux définis, vos politiques d'autorisation doivent être revues de façon à supprimer les autorisations inutiles. Cette approche réduit l'impact si les informations d'identification sont exposées par inadvertance ou autrement consultées sans autorisation.

Anti-modèles courants :

- Octroi par défaut des autorisations d'administrateur aux utilisateurs.
- Création de politiques trop permissives, mais sans tous les privilèges d'administrateur.
- Maintien des politiques d'autorisation une fois qu'elles ne sont plus nécessaires.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Lorsque les équipes et les projets ne font que commencer, des politiques d'autorisation permissives peuvent être utilisées pour favoriser l'innovation et l'agilité. Par exemple, dans un environnement de développement ou de test, les développeurs peuvent se voir octroyer un accès à un large éventail de services AWS. Nous vous recommandons d'évaluer l'accès en continu et de restreindre l'accès aux services et aux actions de service nécessaires pour effectuer le travail en cours. Nous recommandons cette évaluation pour les identités humaines et machine. Les identités machine, parfois appelées comptes de système ou de service, donnent un accès AWS aux applications ou aux

serveurs. Cet accès est particulièrement important dans un environnement de production, où des autorisations trop permissives peuvent avoir un impact important et exposer les données des clients.

AWS fournit plusieurs méthodes pour identifier les utilisateurs, les rôles, les autorisations et les informations d'identification inutilisés. AWS peut également faciliter l'analyse de l'activité d'accès des utilisateurs et rôles IAM, notamment des clés d'accès associées, ainsi que l'accès aux ressources AWS telles que les objets dans les compartiments Amazon S3. La génération de politiques AWS Identity and Access Management Access Analyzer peut vous aider à créer des politiques d'autorisations restrictives basées sur les services et les actions réels avec lesquels un principal interagit. Le contrôle d'accès par attributs (ABAC) peut contribuer à simplifier la gestion des autorisations, car vous pouvez fournir des autorisations aux utilisateurs en utilisant leurs attributs au lieu d'associer des politiques d'autorisations directement à chaque utilisateur.

Étapes d'implémentation

- Utilisez AWS Identity and Access Management Access Analyzer : l'Analyseur d'accès IAM vous aide à identifier les ressources de votre organisation et de vos comptes, comme les compartiments Amazon Simple Storage Service (Amazon S3) ou les rôles IAM, qui sont partagées avec une entité externe.
- Utilisez la génération de politiques de l'Analyseur d'accès IAM : la génération de politiques de l'Analyseur d'accès IAM vous permet de créer des politiques d'autorisation précises reposant sur l'activité d'accès d'un utilisateur ou d'un rôle IAM.
- Testez les autorisations dans les environnements inférieurs avant la production : commencez par utiliser les environnements de développement et de test (sandbox) les moins critiques pour tester les autorisations requises pour les différentes fonctions professionnelles à l'aide de l'Analyseur d'accès IAM. Ensuite, renforcez progressivement et validez ces autorisations dans les environnements de test, d'assurance qualité et intermédiaires avant de les appliquer en production. Les environnements inférieurs peuvent avoir des autorisations plus souples au départ, car les politiques de contrôle des services (SCP) constituent un garde-fou en limitant le nombre maximal d'autorisations accordées.
- Déterminez un délai et une politique d'utilisation acceptables pour les utilisateurs et les rôles IAM : utilisez le dernier horodatage consulté pour identifier les utilisateurs et les rôles non utilisés et les supprimer. Consultez les informations relatives aux services et actions consultées en dernier afin d'identifier et de délimiter les autorisations à des utilisateurs et des rôles spécifiques. Par exemple, vous pouvez utiliser les dernières informations consultées pour identifier les actions Amazon S3 spécifiques dont votre rôle d'application a besoin et limiter l'accès du rôle à celles-ci uniquement. Ces fonctionnalités relatives aux informations sur les derniers accès sont disponibles dans la AWS

Management Console et par programmation pour vous permettre de les intégrer dans vos flux de travail d'infrastructure et vos outils automatisés.

- Envisagez de consigner les événements de données dans AWS CloudTrail : par défaut, CloudTrail n'enregistre pas les événements de données tels que l'activité au niveau des objets Amazon S3 (par exemple, `GetObject` et `DeleteObject`) ou les activités des tables Amazon DynamoDB (par exemple, `PutItem` et `DeleteItem`). Envisagez d'utiliser la journalisation de ces événements afin de déterminer quels utilisateurs et rôles ont besoin d'accéder à des objets Amazon S3 ou des éléments de table DynamoDB spécifiques.

Ressources

Documents connexes :

- Accorder le moindre privilège
- Supprimer les informations d'identification inutiles
- Présentation de AWS CloudTrail
- Utilisation de stratégies
- Journalisation et surveillance dans DynamoDB
- Utilisation de la journalisation des événements CloudTrail pour vos compartiments et objets Amazon S3
- Obtention de rapports d'informations d'identification pour votre Compte AWS

Vidéos connexes :

- Become an IAM Policy Master in 60 Minutes or Less
- Separation of Duties, Least Privilege, Delegation, and CI/CD
- AWS re:Inforce 2022 - AWS Identity and Access Management (IAM) deep dive

SEC03-BP05 Définir des garde-fous des autorisations pour votre organisation

Utilisez des barrières de protection pour réduire la portée des autorisations disponibles qui peuvent être accordées aux principaux. La chaîne d'évaluation des politiques d'autorisation inclut vos barrières de protection permettant de déterminer les autorisations effectives d'un principal lorsqu'il prend des décisions en matière d'autorisation. Vous pouvez définir des garde-fous à l'aide d'une approche par couches. Appliquez certaines barrières de protection de manière générale

à l'ensemble de votre organisation et appliquez-en d'autres de manière granulaire aux sessions d'accès temporaires.

Résultat escompté : vous isolez clairement les environnements en utilisant des Comptes AWS distincts. Les politiques de contrôle des services (SCP) sont utilisées pour définir des garde-fous des autorisations à l'échelle de l'organisation. Des barrières de protection plus étendues sont définies aux niveaux hiérarchiques les plus proches de la racine de votre organisation, tandis que des barrières de protection plus strictes sont définies plus près du niveau des comptes individuels.

Lorsqu'elles sont prises en charge, les politiques de ressources définissent les conditions qu'un principal doit remplir pour accéder à une ressource. Les politiques relatives aux ressources définissent également l'ensemble des actions autorisées, le cas échéant. Les limites des autorisations sont placées sur les principaux qui gèrent les autorisations de charge de travail, en déléguant la gestion des autorisations aux propriétaires individuels de la charge de travail.

Anti-modèles courants :

- Créer un membre Comptes AWS au sein d'une organisation AWS, mais ne pas utiliser les SCP pour restreindre l'utilisation et les autorisations accordées à leurs informations d'identification racine.
- Attribuer des autorisations en fonction du principe du moindre privilège, mais ne pas placer de barrières de protection sur l'ensemble maximum d'autorisations pouvant être accordées.
- S'appuyer sur le principe de refus implicite d'AWS IAM pour restreindre les autorisations, en étant sûr que les politiques n'accorderont pas d'autorisation explicite indésirable.
- Exécuter plusieurs environnements de charge de travail dans le même Compte AWS, puis s'appuyer sur des mécanismes tels que des VPC, des balises ou des politiques de ressources pour appliquer les limites des autorisations.

Avantages du respect de cette bonne pratique : les barrières de protection des autorisations contribuent à garantir que des autorisations indésirables ne peuvent pas être accordées, même lorsqu'une politique d'autorisation tente de le faire. Cela peut simplifier la définition et la gestion des autorisations en réduisant la portée maximale des autorisations à prendre en compte.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Nous vous recommandons d'utiliser une approche par couches afin de définir les barrières de protection des autorisations pour votre organisation. Cette approche réduit systématiquement

l'ensemble maximum d'autorisations possibles à mesure que des couches supplémentaires sont appliquées. Cela vous permet d'accorder l'accès sur la base du principe du moindre privilège, réduisant ainsi le risque d'accès involontaire dû à une mauvaise configuration des politiques.

La première étape pour établir des barrières de protection des autorisations consiste à isoler vos charges de travail et vos environnements dans des Comptes AWS séparés. Les principaux d'un compte ne peuvent pas accéder aux ressources d'un autre compte sans autorisation explicite, même lorsque les deux comptes appartiennent à la même organisation AWS ou à la même unité organisationnelle. Vous pouvez utiliser les unités organisationnelles pour regrouper les comptes que vous souhaitez administrer en tant qu'unité unique.

L'étape suivante consiste à réduire le nombre maximum d'autorisations que vous pouvez accorder aux principaux au sein des comptes membres de votre organisation. Vous pouvez utiliser des politiques de contrôle des services (SCP) à cette fin, que vous pouvez appliquer à une unité d'organisation ou à un compte. Les SCP peuvent appliquer des contrôles d'accès courants, tels que la restriction de l'accès à des Régions AWS spécifiques, la protection contre la suppression des ressources ou la désactivation d'actions de service potentiellement risquées. Les SCP que vous appliquez à la racine de votre organisation n'affectent que ses comptes membres, pas le compte de gestion. Les SCP régissent uniquement les principaux au sein de votre organisation. Vos SCP ne régissent pas les principaux extérieurs à votre organisation qui accèdent à vos ressources.

Si vous utilisez AWS Control Tower, vous pouvez tirer parti de ses contrôles et de ses zones de destination comme base de vos garde-fous d'autorisations et de votre environnement multi-compte. Les zones de destination fournissent un environnement de base sécurisé préconfiguré avec des comptes distincts pour différentes charges de travail et applications. Les garde-fous appliquent des contrôles obligatoires en matière de sécurité, d'exploitation et de conformité par le biais d'une combinaison de politiques de contrôle des services (SCP), de règles AWS Config et d'autres configurations. Toutefois, lorsque vous utilisez les garde-fous et les zones de destination Control Tower en plus des politiques SCP personnalisées de l'organisation, il est essentiel de suivre les bonnes pratiques décrites dans la documentation AWS afin d'éviter les conflits et de garantir une bonne gouvernance. Reportez-vous aux recommandations AWS Control Tower pour AWS Organizations afin d'obtenir des recommandations détaillées sur la gestion des politiques SCP, des comptes et des unités organisationnelles (OU) dans un environnement Control Tower.

En respectant ces directives, vous pouvez tirer parti efficacement des garde-fous, des zones de destination et des politiques SCP personnalisées de Control Tower tout en atténuant les conflits potentiels et en garantissant une gouvernance et un contrôle appropriés de votre environnement AWS multi-compte.

Une autre étape consiste à utiliser les politiques de ressources IAM pour définir les actions disponibles que vous pouvez entreprendre sur les ressources qu'elles régissent, ainsi que les conditions que le principal temporaire doit respecter. Cela peut être aussi large que d'autoriser toutes les actions tant que le principal fait partie de votre organisation (en utilisant la clé de condition PrincipalOrgID), ou aussi granulaire que de n'autoriser que des actions spécifiques par un rôle IAM spécifique. Vous pouvez adopter une approche similaire avec des conditions dans les politiques de confiance des rôles IAM. Si une politique d'approbation en matière de ressources ou de rôles désigne explicitement un principal dans le même compte que le rôle ou la ressource qu'elle gère, ce principal n'a pas besoin de politique IAM associée qui accorde les mêmes autorisations. Si le principal se trouve dans un compte différent de celui de la ressource, il a besoin d'une politique IAM associée qui accorde ces autorisations.

Souvent, une équipe responsable d'une charge de travail souhaite gérer les autorisations requises pour sa charge de travail. Cela peut l'obliger à créer de nouveaux rôles et politiques d'autorisation IAM. Vous pouvez saisir l'étendue maximale des autorisations que l'équipe est autorisée à accorder dans une limite des autorisations IAM, et associer ce document à un rôle IAM que l'équipe peut ensuite utiliser pour gérer ses rôles et autorisations IAM. Cette approche peut lui donner la liberté de terminer son travail, tout en limitant les risques liés au fait de disposer d'un accès administratif IAM.

Une étape plus précise consiste à mettre en œuvre des techniques de gestion des accès privilégiés (PAM) et de gestion des accès élevés temporaires (TEAM). Un exemple de PAM consiste à obliger les principaux à effectuer une authentification multifactorielle avant de réaliser des actions avec privilège. Pour plus d'informations, consultez Configuration de l'accès aux API protégé par MFA. TEAM a besoin d'une solution qui gère l'approbation et le délai pendant lequel un principal est autorisé à bénéficier d'un accès élevé. Une approche consiste à ajouter temporairement le principal à la politique d'approbation des rôles pour un rôle IAM dont l'accès est élevé. Une autre approche consiste, dans le cadre d'un fonctionnement normal, à réduire les autorisations accordées à un principal par un rôle IAM à l'aide d'une politique de session, puis à lever temporairement cette restriction pendant la période approuvée. Pour en savoir plus sur les solutions validées par AWS et des partenaires sélectionnés, consultez la section Accès élevé temporaire.

Étapes d'implémentation

1. Isolez vos charges de travail et vos environnements dans des Comptes AWS distincts.
2. Utilisez les SCP pour réduire le nombre maximum d'autorisations pouvant être accordées aux principaux sur les comptes membres de votre organisation.
   a. Lorsque vous définissez des politiques SCP pour réduire le nombre maximal d'autorisations pouvant être accordées aux principaux sur les comptes membres de votre organisation, vous

pouvez choisir entre une approche avec liste d'autorisations ou liste de refus. La stratégie avec liste d'autorisations spécifie explicitement les accès autorisés et bloque implicitement tous les autres accès. La stratégie avec liste de refus spécifie explicitement les accès qui sont refusés et autorise par défaut tous les autres accès. Ces deux stratégies ont leurs avantages et leurs inconvénients, et le choix approprié dépend des exigences spécifiques et du modèle de risque de votre organisation. Pour plus de détails, consultez Stratégie d'utilisation des politiques SCP.

   b. En outre, passez en revue les exemples de politiques de contrôle des services pour comprendre comment élaborer efficacement des politiques SCP.

3. Utilisez les politiques relatives aux ressources IAM pour réduire la portée et spécifier les conditions des actions autorisées sur les ressources. Utilisez des conditions dans les politiques de confiance relatives aux rôles IAM pour créer des restrictions quant à l'attribution des rôles.
4. Attribuez des limites des autorisations IAM aux rôles IAM que les équipes de la charge de travail peuvent ensuite utiliser afin de gérer leurs propres rôles et autorisations IAM en matière de charge de travail.
5. Évaluez les solutions PAM et TEAM en fonction de vos besoins.

Ressources

Documents connexes :

- Périmètres de données sur AWS
- Établir des barrières de protection d'autorisations à l'aide de périmètres de données
- Logique d'évaluation de stratégies

Exemples connexes :

- Exemples de politiques de contrôle des services

Outils associés :

- Solution AWS : gestion des accès élevés temporaires
- Solutions de partenaires de sécurité validées pour TEAM

SEC03-BP06 Gérer l’accès en fonction du cycle de vie

Surveillez et ajustez les autorisations accordées à vos principaux (utilisateurs, rôles et groupes) tout au long de leur cycle de vie au sein de votre organisation. Ajustez les appartenances aux groupes lorsque les utilisateurs changent de rôle et supprimez l’accès lorsqu’un utilisateur quitte l’organisation.

Résultat escompté : vous surveillez et ajustez les autorisations tout au long du cycle de vie des principaux au sein de l’organisation, réduisant ainsi le risque de privilèges inutiles. Vous accordez l’accès approprié lorsque vous créez un utilisateur. Vous modifiez l’accès en fonction de l’évolution des responsabilités de l’utilisateur et vous supprimez l’accès lorsque l’utilisateur n’est plus actif ou s’il a quitté l’organisation. Vous gérez de manière centralisée les modifications apportées à vos utilisateurs, rôles et groupes. Vous utilisez l’automatisation pour propager les modifications à vos environnements AWS.

Anti-modèles courants :

- Accorder en amont des privilèges d’accès excessifs ou étendus aux identités, au-delà de ce qui est initialement requis.
- Ne pas examiner et ne pas ajuster les privilèges d’accès au fur et à mesure de l’évolution des rôles et des responsabilités des identités.
- Laisser des identités inactives ou résiliées avec des privilèges d’accès actifs. Cela augmente le risque d’accès non autorisé.
- Ne pas tirer parti de l’automatisation pour gérer le cycle de vie des identités.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Gérez et ajustez avec soin les privilèges d’accès que vous accordez aux identités (telles que les utilisateurs, les rôles, les groupes) tout au long de leur cycle de vie. Ce cycle de vie comprend la phase initiale d’intégration, les changements continus des rôles et des responsabilités, et le départ ou la résiliation éventuels. Gérez l’accès de manière proactive en fonction de l’étape du cycle de vie afin de maintenir un niveau d’accès approprié. Optez pour le principe du moindre privilège afin de réduire le risque de privilèges d’accès excessifs ou inutiles.

Vous pouvez gérer le cycle de vie des utilisateurs IAM directement dans le Compte AWS ou par le biais d’une fédération entre le fournisseur d’identité de votre personnel et AWS IAM Identity Center. Pour les utilisateurs IAM, vous pouvez créer, modifier et supprimer des utilisateurs et leurs

autorisations associées dans le Compte AWS. Pour les utilisateurs fédérés, vous pouvez utiliser IAM Identity Center afin de gérer leur cycle de vie en synchronisant les informations relatives aux utilisateurs et aux groupes provenant du fournisseur d'identité de votre organisation à l'aide du protocole System for Cross-domain Identity Management (SCIM).

SCIM est un protocole standard ouvert pour le provisionnement et le déprovisionnement automatisés des identités des utilisateurs sur différents systèmes. En intégrant votre fournisseur d'identité avec IAM Identity Center à l'aide de SCIM, vous pouvez synchroniser automatiquement les informations des utilisateurs et des groupes, ce qui aide à valider que les privilèges d'accès sont accordés, modifiés ou révoqués en fonction des modifications apportées à la source d'identité officielle de votre organisation.

Ajustez les privilèges en fonction de l'évolution des rôles et responsabilités des employés au sein de votre organisation. Vous pouvez utiliser les ensembles d'autorisations d'IAM Identity Center pour définir différents rôles ou responsabilités professionnels et les associer aux politiques et autorisations IAM Identity Center appropriées. Lorsque le rôle d'un employé change, vous pouvez mettre à jour l'ensemble d'autorisations qui lui a été attribué de façon à refléter ses nouvelles responsabilités. Vérifiez qu'il dispose de l'accès nécessaire tout en respectant le principe du moindre privilège.

Étapes d'implémentation

1. Définissez et documentez un processus de gestion des accès tout au long du cycle de vie, y compris les procédures relatives à l'octroi de l'accès initial, aux révisions périodiques et à la révocation de l'accès.
2. Mettez en œuvre les rôles, groupes et limites des autorisations IAM pour gérer l'accès collectivement et appliquer des niveaux d'accès maximaux autorisés.
3. Intégrez un fournisseur d'identité fédéré (tel que Microsoft Active Directory, Okta, Ping Identity) en tant que source officielle d'informations sur les utilisateurs et les groupes utilisant IAM Identity Center.
4. Utilisez le protocole SCIM pour synchroniser les informations sur les utilisateurs et les groupes provenant du fournisseur d'identité dans l'Identity Store d'IAM Identity Center.
5. Dans IAM Identity Center, créez des ensembles d'autorisations qui représentent les différents rôles ou responsabilités au sein de votre organisation. Définissez les politiques et les autorisations IAM appropriées pour chaque ensemble d'autorisations.
6. Mettez en œuvre des contrôles d'accès réguliers, une révocation rapide des accès et une amélioration continue du processus de gestion du cycle de vie des accès.
7. Formez et sensibilisez les employés aux bonnes pratiques de gestion des accès.

Ressources

Bonnes pratiques associées :

- SEC02-BP04 S’appuyer sur un fournisseur d’identité centralisé

Documents connexes :

- Gérer votre source d’identité
- Gestion des identités dans IAM Identity Center
- Utilisation de AWS Identity and Access Management Access Analyzer
- Génération d’une politique de l’Analyseur d’accès IAM

Vidéos connexes :

- AWS re:Inforce 2023 - Manage temporary elevated access with AWS IAM Identity Center
- AWS re:Invent 2022 - Simplify your existing workforce access with IAM Identity Center
- AWS re:Invent 2022 - Harness power of IAM policies & rein in permissions w/Access Analyzer

SEC03-BP07 Analyser l’accès public et intercompte

Surveillez en continu les résultats qui mettent en évidence l’accès public et intercompte. Limitez l’accès public et l’accès intercompte aux seules ressources spécifiques qui nécessitent cet accès.

Résultat escompté : sachez quelles ressources AWS sont partagées et avec qui. Surveillez et auditez continuellement vos ressources partagées afin de vérifier qu’elles ne sont partagées qu’avec les principaux autorisés.

Anti-modèles courants :

- Ne pas tenir un inventaire des ressources partagées.
- Ne pas suivre de processus pour régir l’accès intercompte et public aux ressources.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d'implémentation

Si votre compte est dans AWS Organizations, vous pouvez accorder l'accès aux ressources à l'ensemble de l'organisation, à des unités d'organisation spécifiques ou à des comptes individuels. Si votre compte n'est pas membre d'une organisation, vous pouvez partager des ressources avec des comptes individuels. Vous pouvez accorder un accès intercompte direct à l'aide de politiques basées sur les ressources, par exemple les politiques de compartiment Amazon Simple Storage Service (Amazon S3), ou en autorisant le principal d'un autre compte à assumer un rôle IAM dans votre compte. Lorsque vous utilisez des politiques de ressources, vérifiez que l'accès n'est accordé qu'aux principaux autorisés. Définir un processus d'approbation de toutes les ressources qui doivent être accessibles au public.

AWS Identity and Access Management Access Analyzer utilise la sécurité prouvable pour identifier tous les chemins d'accès à une ressource en dehors de son compte. Il passe en revue les stratégies de ressources en continu et présente les résultats d'accès public et intercompte pour vous permettre d'analyser facilement un accès potentiellement étendu. Envisagez de configurer l'Analyseur d'accès IAM avec AWS Organizations afin de vérifier que vous avez une visibilité sur tous vos comptes. IAM Access Analyzer vous permet également de prévisualiser les résultats avant de déployer les autorisations relatives aux ressources. Vous pouvez ainsi vérifier que vos modifications de politique n'accordent que l'accès public et intercompte prévu à vos ressources. Lorsque vous concevez un accès multicompte, vous pouvez utiliser des politiques de confiance pour contrôler dans quels cas un rôle peut être assumé. Par exemple, vous pouvez utiliser la clé de condition `PrincipalOrgId` pour refuser une tentative d'assumer un rôle en dehors de votre AWS Organizations.

AWS Config peut signaler les ressources mal configurées et, par le biais de vérifications des politiques AWS Config, détecter les ressources dont l'accès public est configuré. Les services comme AWS Control Tower et AWS Security Hub simplifient le déploiement de la détection et des barrières de protection sur AWS Organizations afin d'identifier et de corriger les ressources publiquement exposées. Par exemple, AWS Control Tower dispose d'une barrière de protection gérée qui peut détecter si des instantanés Amazon EBS peuvent être restaurés par Comptes AWS.

Étapes d'implémentation

- Envisagez d'utiliser AWS Config pour AWS Organizations : AWS Config vous permet d'agréger les résultats de plusieurs comptes au sein d'un AWS Organizations vers un compte d'administrateur délégué. Cela fournit une vue complète et vous permet de déployer AWS Config Rules sur plusieurs comptes afin de détecter les ressources accessibles au public.

- Configurez AWS Identity and Access Management Access Analyzer : l'Analyseur d'accès IAM vous aide à identifier les ressources dans votre organisation et vos comptes, telles que les compartiments Amazon S3 ou les rôles IAM, qui sont partagés avec une entité externe.
- Utilisez la correction automatique dans AWS Config pour répondre aux modifications de la configuration de l'accès public des compartiments Amazon S3 : vous pouvez activer automatiquement les paramètres de blocage de l'accès public pour les compartiments Amazon S3.
- Mettez en œuvre la surveillance et les alertes pour déterminer si les compartiments Amazon S3 sont devenus publics : vous devez mettre en place un système de surveillance et d'alerte pour identifier quand le blocage de l'accès public Amazon S3 est désactivé et si les compartiments Amazon S3 deviennent publics. En outre, si vous utilisez AWS Organizations, vous pouvez créer une politique de contrôle des services qui empêche toute modification des stratégies d'accès public d'Amazon S3. AWS Trusted Advisor vérifie les compartiments Amazon S3 dont les autorisations permettent un libre accès. Les autorisations de compartiment qui accordent à tous un accès au chargement ou à la suppression créent des problèmes de sécurité potentiels, en permettant à quiconque d'ajouter, de modifier ou de supprimer les éléments d'un compartiment. La vérification Trusted Advisor examine les autorisations explicites de compartiment et les politiques associées de compartiment susceptibles de remplacer les autorisations de compartiment. Vous pouvez également utiliser AWS Config pour surveiller l'accès public de vos compartiments Amazon S3. Pour obtenir plus d'informations, consultez Comment utiliser AWS Config pour surveiller et gérer les compartiments Amazon S3 autorisant l'accès public.

Lorsque vous examinez les contrôles d'accès pour les compartiments Amazon S3, il est important de prendre en compte la nature des données qui y sont stockées. Amazon Macie est un service conçu pour vous aider à découvrir et à protéger les données sensibles, telles que les données d'identification personnelle (PII), les informations protégées sur la santé (PHI) et les informations d'identification telles que les clés privées ou les clés d'accès AWS.

Ressources

Documents connexes :

- Utilisation de AWS Identity and Access Management Access Analyzer
- Bibliothèque de contrôles AWS Control Tower
- Norme concernant les bonnes pratiques de sécurité de base AWS
- Règles AWS Config gérées
- Référence de la vérification AWS Trusted Advisor

- Surveillance des résultats des vérifications AWS Trusted Advisor avec Amazon EventBridge
- Gestion des règles AWS Config pour tous les comptes de votre organisation
- AWS Config et AWS Organizations
- Mise à disposition de votre AMI au public pour son utilisation dans Amazon EC2

Vidéos connexes :

- Best Practices for securing your multi-account environment
- Dive Deep into IAM Access Analyzer

SEC03-BP08 Partager des ressources en toute sécurité au sein de votre organisation

À mesure que le nombre de charges de travail augmente, vous devrez peut-être partager l'accès aux ressources de ces charges de travail ou fournir les ressources plusieurs fois pour plusieurs comptes. Vous pouvez utiliser des constructions pour compartimenter votre environnement, par exemple des environnements de développement, de test et de production. Cependant, le fait d'avoir des constructions distinctes ne vous empêche pas de partager en toute sécurité. En partageant des composants qui se chevauchent, vous pouvez réduire les frais d'exploitation et offrir une expérience cohérente sans avoir à deviner ce que vous avez pu manquer en créant la même ressource plusieurs fois.

Résultat escompté : minimisez les accès involontaires en utilisant des méthodes sécurisées pour partager les ressources au sein de votre organisation et contribuer à votre initiative de prévention des pertes de données. Réduisez vos frais généraux opérationnels par rapport à la gestion de composants individuels, réduisez les erreurs liées à la création manuelle du même composant plusieurs fois et augmentez la capacité de mise à l'échelle de vos charges de travail. Vous pouvez bénéficier d'une réduction du délai de résolution dans les scénarios de défaillance multipoints et augmenter votre confiance dans l'évaluation du moment où un composant n'est plus nécessaire. Pour obtenir des conseils prescriptifs sur l'analyse des ressources partagées en externe, voir SEC03-BP07 Analyser l'accès public et intercompte.

Anti-modèles courants :

- Manque de processus pour surveiller continuellement et alerter automatiquement sur un partage externe inattendu.
- Manque de référence sur ce qui doit être partagé et ce qui ne doit pas l'être.

- Adoption par défaut d'une politique largement ouverte au lieu de la partager explicitement lorsque c'est nécessaire.
- Création manuelle des ressources de base qui se chevauchent si nécessaire.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Concevez vos contrôles et modèles d'accès de façon à régir la consommation de ressources partagées en toute sécurité et uniquement avec des entités approuvées. Surveillez les ressources partagées et examinez l'accès aux ressources partagées en permanence, et soyez alerté sur les partages inappropriés ou inattendus. Consultez Analyser l'accès public et intercompte pour vous aider à établir une gouvernance visant à réduire l'accès externe aux seules ressources qui en ont besoin, et à établir un processus de surveillance continue et d'alerte automatique.

Le partage entre comptes au sein de AWS Organizations est pris en charge par un certain nombre de services AWS, tels que AWS Security Hub, Amazon GuardDuty et AWS Backup. Ces services permettent de partager les données vers un compte central, de rendre les données accessibles à partir d'un compte central ou de gérer les ressources et les données à partir d'un compte central. Par exemple, AWS Security Hub peut transférer les résultats des comptes individuels vers un compte central où vous pouvez voir tous ces résultats. AWS Backup peut prendre une sauvegarde pour une ressource et la partager entre les comptes. Vous pouvez utiliser AWS Resource Access Manager (AWS RAM) pour partager d'autres ressources communes, telles que des sous-réseaux VPC et des attachements de la passerelle de transit, AWS Network Firewall ou des pipelines d'IA Amazon SageMaker.

Pour empêcher votre compte de partager uniquement les ressources au sein de votre organisation, utilisez des politiques de contrôle des services (SCP) pour empêcher l'accès aux principaux externes. Lorsque vous partagez des ressources, combinez les contrôles basés sur l'identité et les contrôles réseau pour créer un périmètre de données pour votre organisation afin de la protéger contre tout accès non intentionnel. Un périmètre de données est un ensemble de barrières de protection préventives qui vous permettent de vous assurer que seules les identités approuvées accèdent aux ressources approuvées à partir des réseaux attendus. Ces contrôles doivent placer des limites appropriées pour les ressources susceptibles d'être partagées et empêcher le partage ou l'exposition de ressources qui ne doivent pas être autorisées. Par exemple, dans le cadre de votre périmètre de données, vous pouvez utiliser les politiques de point de terminaison d'un VPC et la condition `AWS:PrincipalOrgId` pour garantir que les identités accédant à vos compartiments Amazon S3

appartiennent à votre organisation. Il est important de noter que les SCP ne s'appliquent pas aux rôles liés aux services ou aux principaux de service AWS.

Lorsque vous utilisez Amazon S3, désactivez les ACL pour votre compartiment Amazon S3 et utilisez les politiques IAM pour définir le contrôle d'accès. Pour restreindre l'accès à une origine Amazon S3 à partir d'Amazon CloudFront, migrez l'identité d'accès d'origine (OAI) vers le contrôle d'accès d'origine (OAC) qui prend en charge des fonctionnalités supplémentaires, dont le chiffrement côté serveur avec AWS Key Management Service.

Dans certains cas, vous pouvez autoriser le partage des ressources à l'extérieur de votre organisation ou accorder à un tiers l'accès à vos ressources. Pour obtenir des conseils prescriptifs sur la gestion des autorisations de partage de ressources en externe, consultez la section Gestion des autorisations.

Étapes d'implémentation

1. Utilisez AWS Organizations : AWS Organizations est un service de gestion de comptes qui vous permet de consolider plusieurs Comptes AWS en une organisation que vous créez et gérez de façon centralisée. Vous pouvez regrouper vos comptes en unités d'organisation (OU) et joindre différentes politiques à chacune d'entre elles afin de vous aider à répondre à vos besoins en matière de budget, de sécurité et de conformité. Vous pouvez également contrôler la façon dont l'intelligence artificielle (IA) et le machine learning (ML) d'AWS peuvent collecter et stocker des données, et utiliser la gestion multicompte des services AWS intégrés à Organizations.
2. Intégrez AWS Organizations aux services AWS : lorsque vous utilisez un service AWS pour exécuter des tâches en votre nom dans les comptes membres de votre organisation, AWS Organizations crée un rôle lié à un service IAM (SLR) pour ce service dans chaque compte membre. Gérez l'accès approuvé à l'aide de la AWS Management Console, des API AWS ou de la AWS CLI. Pour obtenir des conseils prescriptifs sur l'activation de l'accès sécurisé, voir Utilisation d'AWS Organizations avec d'autres services AWS et services AWS que vous pouvez utiliser avec Organizations.
3. Établissez un périmètre de données : un périmètre de données fournit une délimitation claire de la confiance et de la propriété. Sur AWS, il est généralement représenté comme votre organisation AWS gérée par AWS Organizations, ainsi que par tous les réseaux ou systèmes sur site qui accèdent à vos ressources AWS. L'objectif du périmètre de données est de vérifier que l'accès est autorisé si l'identité est approuvée, si la ressource est approuvée et si le réseau est attendu. Toutefois, l'établissement d'un périmètre de données n'est pas une approche universelle. Évaluez et adoptez les objectifs de contrôle décrits dans le livre blanc Construire un périmètre sur AWS sur la base de vos modèles de risques de sécurité et de vos exigences spécifiques. Vous devez

examiner attentivement votre position unique en matière de risque et mettre en œuvre les contrôles périmétriques adaptés à vos besoins en matière de sécurité.

4. Utilisez le partage des ressources dans les services AWS et limitez en conséquence : de nombreux services AWS vous permettent de partager des ressources avec un autre compte ou de cibler une ressource d'un autre compte, comme Amazon Machine Images (AMI) et AWS Resource Access Manager (AWS RAM). Limitez l'API `ModifyImageAttribute` pour spécifier les comptes fiables avec lesquels partager l'AMI. Spécifiez la condition `ram:RequestedAllowsExternalPrincipals` lors de l'utilisation de AWS RAM pour limiter le partage à votre organisation uniquement, afin d'empêcher l'accès par des identités non fiables. Pour des conseils et des considérations prescriptifs, voir Partage des ressources et cibles externes.
5. Utilisez AWS RAM pour partager en toute sécurité dans un compte ou avec d'autres Comptes AWS : AWS RAM vous aide à partager en toute sécurité les ressources que vous avez créées avec les rôles et les utilisateurs de votre compte et avec d'autres Comptes AWS. Dans un environnement multicompte, AWS RAM vous permet de créer une ressource une fois et de la partager avec d'autres comptes. Cette approche permet de réduire vos frais généraux opérationnels tout en assurant la cohérence, la visibilité et l'auditabilité grâce à des intégrations avec Amazon CloudWatch et AWS CloudTrail, que vous ne recevez pas lorsque vous utilisez l'accès intercompte.

   Si vous avez déjà partagé des ressources à l'aide d'une politique basée sur les ressources, vous pouvez utiliser l'API `PromoteResourceShareCreatedFromPolicy` ou un équivalent pour transformer le partage de ressources en partage de ressources AWS RAM complet.

   Dans certains cas, vous devrez peut-être prendre des mesures supplémentaires pour partager les ressources. Par exemple, pour partager un instantané chiffré, vous devez partager une clé AWS KMS.

Ressources

Bonnes pratiques associées :

- SEC03-BP07 Analyser l'accès public et intercompte
- SEC03-BP09 Partager des ressources en toute sécurité avec un tiers
- SEC05-BP01 Création de couches réseau

Documents connexes :

- Propriétaire d'un compartiment accordant des autorisations entre comptes à des objets qu'il ne possède pas
- How to use Trust Policies with IAM
- Building Data Perimeter on AWS
- Procédure d'utilisation d'un ID externe lorsque vous accordez l'accès à vos ressources AWS à un tiers
- Services AWS que vous pouvez utiliser avec AWS Organizations
- Établissement d'un périmètre de données sur AWS : autoriser uniquement les identités fiables à accéder aux données de l'entreprise

Vidéos connexes :

- Granular Access with AWS Resource Access Manager
- Securing your data perimeter with VPC endpoints
- Establishing a data perimeter on AWS

Outils associés :

- Exemples de stratégies relatives au périmètre des données

SEC03-BP09 Partager des ressources en toute sécurité avec un tiers

La sécurité de votre environnement cloud ne s'arrête pas à votre organisation. Votre organisation peut faire appel à un tiers pour gérer une partie de vos données. La gestion des autorisations pour le système géré par un tiers doit suivre la pratique de l'accès juste à temps en utilisant le principe du moindre privilège avec des informations d'identification temporaires. En travaillant en étroite collaboration avec un tiers, vous pouvez réduire ensemble l'étendue de l'impact et le risque d'accès involontaire.

Résultat escompté : vous évitez d'utiliser des informations d'identification à long terme AWS Identity and Access Management (IAM) telles que des clés d'accès et des clés secrètes, car elles présentent un risque de sécurité en cas d'utilisation abusive. Vous utilisez plutôt des rôles IAM et des informations d'identification temporaires pour améliorer votre niveau de sécurité et minimiser les frais opérationnels liés à la gestion des informations d'identification à long terme. Lorsque vous accordez

l'accès à un tiers, vous utilisez un identifiant unique universel (UUID) comme ID externe dans la politique d'approbation IAM et vous maintenez sous votre contrôle les politiques IAM attachées au rôle afin de garantir un accès sur la base du moindre privilège. Pour obtenir des conseils prescriptifs sur l'analyse des ressources partagées en externe, consultez SEC03-BP07 Analyser l'accès public et intercompte.

Anti-modèles courants :

- Utilisation de la politique d'approbation IAM sans aucune condition.
- Utilisation d'informations d'identification IAM et de clés d'accès à long terme.
- Réutilisation des ID externes.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Vous pouvez autoriser le partage des ressources en dehors d'AWS Organizations ou accorder un accès tiers à votre compte. Par exemple, un tiers peut fournir une solution de surveillance qui doit accéder aux ressources de votre compte. Dans ces cas de figure, vous devez créer un rôle intercompte IAM en lui attribuant uniquement les privilèges requis par le tiers. Définissez également une politique d'approbation à l'aide de la condition d'ID externe. Lorsque vous utilisez un identifiant externe, vous pouvez (ou le tiers peut) générer un identifiant unique pour chaque client, tiers ou location. L'ID unique ne doit être contrôlé que par vous après sa création. Le tiers doit implémenter un processus pour relier l'ID externe au client de manière sécurisée, auditable et reproductible.

Vous pouvez également utiliser Rôles Anywhere IAM pour gérer les rôles IAM pour des applications hors AWS qui utilisent des API AWS.

Si le tiers n'a plus besoin d'accéder à votre environnement, supprimez le rôle. Évitez de fournir à des tiers des informations d'identification à long terme. Gardez un œil sur les autres services AWS qui prennent en charge le partage, comme l'AWS Well-Architected Tool qui permet le partage d'une charge de travail avec d'autres Comptes AWS et AWS Resource Access Manager qui vous aide à partager en toute sécurité une ressource AWS que vous possédez avec d'autres comptes.

Étapes d'implémentation

1. Utilisez des rôles intercomptes pour fournir l'accès aux comptes externes. Les rôles intercomptes réduisent la quantité d'informations sensibles stockées par les comptes externes et les tiers pour servir leurs clients. Les rôles intercomptes vous permettent d'accorder l'accès aux ressources

AWS de votre compte en toute sécurité à un tiers, par exemple aux partenaires AWS ou à d'autres comptes de votre organisation, tout en préservant la capacité de gérer et d'auditer cet accès. Il se peut que le tiers vous fournisse des services à partir d'une infrastructure hybride ou qu'il extraie des données hors site pour les transférer dans un emplacement hors site. Rôles Anywhere IAM vous permet d'autoriser des charges de travail tierces à interagir en toute sécurité avec vos charges de travail AWS et de réduire encore le besoin d'informations d'identification à long terme.

Vous ne devez pas utiliser d'informations d'identification à long terme ni de clés d'accès associées aux utilisateurs pour fournir un accès à un compte externe. Utilisez plutôt les rôles intercomptes pour fournir l'accès intercompte.

2. Faites preuve d'une diligence raisonnable et garantissez un accès sécurisé aux fournisseurs SaaS tiers. Lorsque vous partagez des ressources avec des fournisseurs SaaS tiers, faites preuve de toute la diligence appropriée pour vous assurer qu'ils adoptent une approche sûre et responsable de l'accès à vos ressources AWS. Évaluez leur modèle de responsabilité partagée pour comprendre les mesures de sécurité qu'ils fournissent et celles qui relèvent de votre responsabilité. Assurez-vous que le fournisseur SaaS dispose d'un processus sécurisé et auditable pour accéder à vos ressources, y compris l'utilisation d'identifiants externes et les principes d'accès sur la base du moindre privilège. L'utilisation d'identifiants externes permet de résoudre le problème de l'adjoint confus.

   Mettez en œuvre des contrôles de sécurité pour garantir un accès sécurisé et le respect du principe du moindre privilège lorsque vous accordez l'accès à des fournisseurs SaaS tiers. Cela peut inclure l'utilisation d'identifiants externes, d'identifiants uniques universels (UUID) et de politiques d'approbation IAM qui limitent l'accès au strict nécessaire. Travaillez en étroite collaboration avec le fournisseur SaaS pour établir des mécanismes d'accès sécurisés, passez régulièrement en revue son accès à vos ressources AWS et effectuez des audits pour garantir le respect de vos exigences de sécurité.
3. Rendez obsolètes les informations d'identification à long terme fournies par le client. Rendez obsolète l'utilisation d'informations d'identification à long terme et utilisez des rôles intercomptes ou Rôles Anywhere IAM. Si vous devez utiliser des informations d'identification à long terme, établissez un plan pour migrer vers un accès basé sur les rôles. Pour plus de détails sur la gestion des clés, consultez Gestion des identités. Collaborez également avec votre équipe Compte AWS et le tiers pour établir un dossier d¦exploitation d'atténuation des risques. Pour obtenir des conseils prescriptifs sur la réponse à un incident de sécurité et l'atténuation de son impact potentiel, consultez Réponse aux incidents.
4. Vérifiez que la configuration est guidée par des instructions prescriptives ou qu'elle est automatisée. L'ID externe n'est pas traité comme un secret, mais il ne doit pas être facile à

deviner, comme un numéro de téléphone, un nom ou un numéro de compte. Faites de l'ID externe un champ en lecture seule afin qu'il ne puisse pas être modifié dans le but de se faire passer pour la configuration.

Le tiers ou vous-même pouvez générer l'ID externe. Définissez un processus pour déterminer qui est responsable de la génération de l'ID. Quelle que soit l'entité qui crée l'ID externe, le tiers applique l'unicité et les formats de façon uniforme parmi les clients.

La politique créée pour l'accès intercompte à vos comptes doit respecter le principe du moindre privilège. Le tiers doit fournir un document de politique de rôle ou un mécanisme de configuration automatisé qui utilise un modèle AWS CloudFormation ou un équivalent pour vous. Cela réduit le risque d'erreurs associées à la création manuelle de politiques et offre une piste auditable. Pour plus d'informations sur l'utilisation d'un modèle AWS CloudFormation pour créer des rôles intercomptes, consultez Rôles intercomptes.

Le tiers doit fournir un mécanisme de configuration automatisé et auditable. Cependant, si vous utilisez le document de politique de rôle décrivant l'accès nécessaire, vous devez automatiser la configuration du rôle. Si vous utilisez un modèle AWS CloudFormation ou un équivalent, vous devez surveiller les changements via une détection des dérives dans le cadre de la pratique d'audit.

5. Tenez compte des modifications. Votre structure de compte, la nécessité de faire appel à un tiers, ou son offre de service peuvent changer. Vous devez anticiper les changements et les défaillances, et planifier en conséquence avec les personnes, processus et technologies appropriés. Auditez régulièrement le niveau d'accès que vous fournissez et implémentez des méthodes de détection pour vous avertir des changements imprévus. Surveillez et auditez l'utilisation du rôle et de l'entrepôt de données des ID externes. Vous devez être prêt à révoquer l'accès tiers, de façon temporaire ou permanente, en raison de changements ou de tendances d'accès imprévus. De plus, mesurez l'impact sur votre opération de révocation, y compris le temps nécessaire pour l'exécution, les personnes impliquées, le coût et l'impact sur d'autres ressources.

   Pour obtenir des conseils prescriptifs sur les méthodes de détection, consultez les bonnes pratiques en matière de détection.

Ressources

Bonnes pratiques associées :

- SEC02-BP02 Utiliser des informations d'identification temporaires

- SEC03-BP05 Définir des garde-fous des autorisations pour votre organisation
- SEC03-BP06 Gérer l'accès en fonction du cycle de vie
- SEC03-BP07 Analyser l'accès public et intercompte
- SEC04 Détection

Documents connexes :

- Propriétaire d'un compartiment accordant des autorisations entre comptes à des objets qu'il ne possède pas
- Comment utiliser des politiques d'approbation avec les rôles IAM
- Déléguer l'accès entre des Comptes AWS à l'aide des rôles IAM
- Comment accéder aux ressources d'un autre Compte AWS à l'aide d'IAM ?
- Bonnes pratiques de sécurité dans IAM
- Logique d'évaluation des politiques intercomptes
- Procédure d'utilisation d'un ID externe lorsque vous accordez l'accès à vos ressources AWS à un tiers
- Collecte d'informations à partir de ressources AWS CloudFormation créées dans des comptes externes avec des ressources personnalisées
- Utilisation sécurisée d'identifiants externes pour accéder à des comptes AWS détenus par d'autres
- Extension des rôles IAM à des charges de travail situées en dehors d'IAM avec Rôles Anywhere IAM

Vidéos connexes :

- Comment accorder à des utilisateurs ou des rôles situés dans un Compte AWS distinct l'accès à mon Compte AWS ?
- AWS re:Invent 2018: Become an IAM Policy Master in 60 Minutes or Less
- AWS Knowledge Center Live : Bonnes pratiques IAM et décisions de conception

Exemples connexes :

- Atelier Well-Architected – Assomption du rôle IAM intercompte Lambda (niveau 300)
- Configuration de l'accès intercompte à Amazon DynamoDB

- Outil de requête réseau AWS STS

Détection

Question

- SÉC 4. Comment détecter les événements de sécurité et comment y répondre ?

SÉC 4. Comment détecter les événements de sécurité et comment y répondre ?

Capturez et analysez les événements à partir des journaux et des métriques pour gagner en visibilité. Prenez des mesures en cas d'événements de sécurité et de menaces potentielles afin de sécuriser votre charge de travail.

Bonnes pratiques

- SEC04-BP01 Configurer une journalisation de service et d'application
- SEC04-BP02 Capturer les journaux, les résultats et les métriques dans des emplacements standardisés
- SEC04-BP03 Corréler et enrichir les alertes de sécurité
- SEC04-BP04 Lancer la correction pour les ressources non conformes

SEC04-BP01 Configurer une journalisation de service et d'application

Conservez les journaux des événements de sécurité générés par les services et les applications. Il s'agit d'un principe de sécurité fondamental pour les cas d'audit, d'enquête et d'utilisation opérationnelle, et d'une exigence de sécurité commune dictée par les normes, politiques et procédures de gouvernance, de risque et de conformité (GRC).

Résultat escompté : une organisation doit pouvoir récupérer de manière fiable et cohérente les journaux des événements de sécurité provenant des services AWS et des applications en temps voulu, lorsqu'il est nécessaire de répondre à un processus interne ou à une obligation, comme une réponse à un incident de sécurité. Envisagez de centraliser les journaux pour obtenir de meilleurs résultats opérationnels.

Anti-modèles courants :

- Les journaux sont conservés indéfiniment ou supprimés trop tôt.
- Tout le monde peut accéder aux journaux.

- Se fier entièrement aux processus manuels pour la gouvernance et l'utilisation des journaux.
- Stocke de tous types de journaux, même si leur utilisation n'est pas garantie.
- Vérification de l'intégrité des journaux uniquement lorsque cela s'avère nécessaire.

Avantages de la mise en place de cette bonne pratique : mettez en œuvre un mécanisme d'analyse de cause racine (RCA) pour les incidents de sécurité et une source de preuves pour vos obligations en matière de gouvernance, de risque et de conformité.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Au cours d'une enquête de sécurité ou d'autres cas d'utilisation en fonction de vos besoins, vous devez être en mesure d'examiner les journaux pertinents pour consigner et comprendre la portée et la chronologie complètes de l'incident. Des journaux sont également requis pour la génération d'alertes, indiquant que certaines actions intéressantes ont eu lieu. Il est essentiel de sélectionner, d'activer, de stocker et de configurer les mécanismes d'interrogation et de récupération ainsi que les alertes.

Étapes d'implémentation

- Sélectionnez et utilisez les sources de journaux. Avant une enquête de sécurité, vous devez saisir les journaux pertinents pour reconstruire rétroactivement l'activité dans un Compte AWS. Sélectionnez les sources de journaux pertinentes pour vos charges de travail.

  Les critères de sélection des sources de journaux doivent être fondés sur les cas d'utilisation requis par votre entreprise. Mettez en place une piste pour chaque Compte AWS en utilisant AWS CloudTrail ou une piste AWS Organizations, puis configurez un compartiment Amazon S3 pour cette piste.

  AWS CloudTrail est un service de journalisation qui suit les appels API sur un Compte AWS pour capturer l'activité de service AWS. Il est activé par défaut avec une rétention de 90 jours des événements de gestion qui peuvent être récupérés via l'historique des événements CloudTrail à l'aide AWS Management Console, de AWS CLI, ou d'un SDK AWS. Pour une rétention plus longue et une meilleure visibilité des événements de données, créez une piste CloudTrail et associez-la à un compartiment Amazon S3, et éventuellement à un groupe de journaux Amazon CloudWatch. Vous pouvez également créer un CloudTrail Lake, qui conserve les journaux CloudTrail pendant une période maximale de sept ans et fournit une fonction de requête basée sur SQL

AWS recommande aux clients utilisant un VPC d'activer le trafic réseau et les journaux DNS à l'aide des journaux de flux VPC et des journaux de requêtes du résolveur Amazon Route 53, respectivement, et de les diffuser vers un compartiment Amazon S3 ou un groupe de journaux CloudWatch. Vous pouvez créer un journal de flux VPC pour un VPC, un sous-réseau ou une interface réseau. Pour les journaux de flux VPC, vous pouvez choisir la façon dont et l'endroit où vous les utilisez pour réduire les coûts.

Les journaux AWS CloudTrail, les journaux de flux VPC et les journaux de requêtes du résolveur Route 53 sont les sources de journalisation de base qui soutiennent les enquêtes de sécurité dans AWS. Vous pouvez également utiliser Amazon Security Lake pour collecter, normaliser et stocker ces données de journal au format Apache Parquet et au format Open Cybersecurity Schema Framework (OCSF), qui est prêt à être interrogé. Security Lake prend également en charge d'autres journaux AWS et des journaux de sources tierces.

Les services AWS peuvent générer des journaux non capturés par les sources de journaux de base, comme les journaux Elastic Load Balancing, les journaux AWS WAF, les journaux de l'enregistreur AWS Config, les résultats Amazon GuardDuty, les journaux d'audit Amazon Elastic Kubernetes Service (Amazon EKS) et les journaux d'application et de système d'exploitation des instances Amazon EC2. Pour une liste complète des options de journalisation et de surveillance, consultez l'annexe A : Définitions des fonctionnalités cloud : journalisation et événements du Guide de réponse aux incidents de sécurité AWS.

- Recherchez des capacités de journalisation pour chaque service et application AWS : chaque service et application AWS vous propose des options de stockage des journaux, chacune ayant ses propres capacités de conservation et de cycle de vie. Amazon Simple Storage Service (Amazon S3) et Amazon CloudWatch sont les deux services de stockage de journaux les plus courants. Pour de longues périodes de conservation, il est recommandé d'utiliser Amazon S3 pour sa rentabilité et ses capacités de cycle de vie flexibles. Si l'option de journalisation principale est Journaux Amazon CloudWatch Logs, en tant qu'option, vous devez envisager d'archiver les journaux les moins consultés dans Amazon S3.
- Sélectionnez le stockage des journaux : le choix du stockage des journaux dépend généralement de l'outil de requête que vous utilisez, des capacités de conservation, de la familiarité et du coût. Les options principales du stockage de journaux sont un compartiment Amazon S3 ou un groupe de journaux CloudWatch.

Un compartiment Amazon S3 offre un stockage durable et rentable avec une politique de cycle de vie facultative. Les journaux stockés dans des compartiments Amazon S3 peuvent être interrogés à l'aide de services tels qu'Amazon Athena.

Un groupe de journaux CloudWatch offre un stockage durable et une installation de requête intégrée via CloudWatch Logs Insights.

- Identifiez la rétention appropriée des journaux : lorsque vous utilisez un compartiment Amazon S3 ou un groupe de journaux CloudWatch pour stocker des journaux, vous devez établir des cycles de vie adéquats pour chaque source de journaux afin d'optimiser les coûts de stockage et de récupération. Les clients ont généralement entre trois mois et un an de journaux facilement disponibles pour la recherche, avec une conservation de sept ans maximum. Le choix de la disponibilité et de la conservation doit correspondre à vos exigences en matière de sécurité et à un ensemble d'obligations statutaires, réglementaires et opérationnelles.
- Utilisez la journalisation pour chaque service et application AWS avec des politiques de conservation et de cycle de vie appropriées : pour chaque service ou application AWS de votre organisation, consultez les instructions de configuration de journalisation spécifiques :
  - Configurer AWS CloudTrail Trail
  - Configurer des journaux de flux VPC
  - Configurer Amazon GuardDuty Finding Export
  - Configurer l'enregistrement AWS Config
  - Configurer le trafic ACL AWS WAF Web
  - Configurer les journaux de trafic réseau AWS Network Firewall
  - Journaux d'accès Elastic Load Balancing
  - Configurer les journaux de requête Amazon Route 53 Resolver
  - Configurer les journaux Amazon RDS
  - Configurer les journaux du plan de contrôle Amazon EKS
  - Configurer l'agent Amazon CloudWatch pour les instances Amazon EC2 et les serveurs sur site
- Sélectionnez et implémentez des mécanismes d'interrogation pour les journaux : pour les interrogations de journal, vous pouvez utiliser CloudWatch Logs Insights pour les données stockées dans les groupes de journaux CloudWatch, et Amazon Athena et Amazon OpenSearch Service pour les données stockées dans Amazon S3. Vous pouvez également utiliser des outils d'interrogation tiers tels qu'un service de gestion des informations de sécurité et des événements (SIEM).

Le processus de sélection d'un outil d'interrogation de journaux doit tenir compte des aspects humains, technologiques et de processus de vos opérations de sécurité. Choisissez un outil qui répond aux exigences opérationnelles, métier et de sécurité, tout en étant accessible et gérable à long terme. Gardez à l'esprit que les outils d'interrogation de journaux fonctionnent de manière optimale lorsque le nombre de journaux à analyser est maintenu dans les limites de l'outil. Il n'est pas rare d'avoir plusieurs outils d'interrogation en raison de contraintes de coût ou techniques.

Par exemple, vous pouvez utiliser un outil de gestion des événements et des informations de sécurité tiers pour effectuer des requêtes sur les 90 derniers jours de données, mais utiliser Athena pour effectuer des requêtes au-delà de 90 jours en raison du coût d'ingestion du journal d'un SIEM. Quelle que soit l'implémentation choisie, assurez-vous que votre approche réduit au minimum le nombre d'outils requis pour maximiser l'efficacité opérationnelle, en particulier pendant une enquête sur un événement de sécurité.

- Utiliser les journaux pour les alertes : AWS fournit des alertes par le biais de plusieurs services de sécurité :
  - AWS Config surveille et enregistre les configurations de vos ressources AWS et permet d'automatiser l'évaluation et la remédiation par rapport aux configurations souhaitées.
  - Amazon GuardDuty est un service de détection des menaces qui surveille en permanence les activités malveillantes et les comportements non autorisés afin de protéger vos Comptes AWS et vos charges de travail. GuardDuty ingère, agrège et analyse les informations provenant de sources telles que les événements de gestion et de données AWS CloudTrail, les journaux DNS, les journaux de flux VPC et les journaux d'audit Amazon EKS. GuardDuty extrait des flux de données indépendants directement depuis CloudTrail, les journaux de flux VPC, les journaux de requêtes DNS et Amazon EKS. Vous n'avez pas besoin de gérer les politiques de compartiment Amazon S3 ni de modifier la façon dont vous collectez et stockez les journaux. Il est toujours recommandé de conserver ces journaux à des fins d'enquête et de conformité.
  - AWS Security Hub fournit un emplacement unique qui regroupe, organise et priorise vos alertes de sécurité ou vos résultats provenant de plusieurs services AWS et de produits tiers en option pour vous donner une vue complète des alertes de sécurité et du statut de conformité.

  Vous pouvez également utiliser des moteurs de génération d'alertes personnalisés pour les alertes de sécurité non couvertes par ces services ou pour les alertes spécifiques pertinentes à votre environnement. Pour plus d'informations sur la création de ces alertes et détections, consultez la section Détection dans le Guide de réponse aux incidents de sécurité AWS.

Ressources

Bonnes pratiques associées :

- SEC04-BP02 Capturer les journaux, les résultats et les métriques dans des emplacements standardisés
- SEC07-BP04 Définir la gestion évolutive du cycle de vie des données
- SEC10-BP06 Outils de pré-déploiement

Documents connexes :

- Guide d'intervention en cas d'incident de sécurité AWS
- Premiers pas avec Amazon Security Lake
- Démarrer : Amazon CloudWatch Logs

Vidéos connexes :

- AWS re:Invent 2022 - Introducing Amazon Security Lake

Exemples connexes :

- Assistant Log Enabler pour AWS
- Exportation historique des résultats AWS Security Hub

SEC04-BP02 Capturer les journaux, les résultats et les métriques dans des emplacements standardisés

Les équipes de sécurité s'appuient sur les journaux et les résultats pour analyser les événements susceptibles d'indiquer une activité non autorisée ou des modifications involontaires. Afin de simplifier cette analyse, capturez les journaux de sécurité et les résultats dans des emplacements standardisés. Vous pourrez ainsi rendre disponibles les points de données intéressants pour la corrélation et simplifier les intégrations d'outils.

Résultat escompté : vous disposez d'une approche standardisée pour collecter, analyser et visualiser les données de journal, les résultats et les métriques. Les équipes de sécurité peuvent corréler, analyser et visualiser efficacement les données de sécurité provenant de systèmes disparates afin de découvrir les événements de sécurité potentiels et d'identifier les anomalies. Des systèmes

de gestion des informations et des événements de sécurité (SIEM) ou d'autres mécanismes sont intégrés pour interroger et analyser les données des journaux afin de répondre rapidement, de suivre et de faire remonter les événements de sécurité.

Anti-modèles courants :

- Les équipes possèdent et gèrent indépendamment la journalisation et la collecte de métriques qui ne sont pas conformes à la stratégie de journalisation de l'organisation.
- Les équipes ne disposent pas de contrôles d'accès adéquats pour restreindre la visibilité et la modification des données collectées.
- Les équipes ne gèrent pas leurs journaux de sécurité, leurs résultats et leurs métriques dans le cadre de leur politique de classification des données.
- Les équipes négligent les exigences de souveraineté et de localisation des données lors de la configuration des collections de données.

Avantages liés au respect de cette bonne pratique : une solution de journalisation standardisée pour collecter et interroger les données et les événements des journaux améliore les informations dérivées des informations qu'ils contiennent. La configuration d'un cycle de vie automatisé pour les données de journal collectées peut permettre de réduire les coûts liés au stockage des journaux. Vous pouvez créer un contrôle d'accès précis pour les informations de journal collectées en fonction de la sensibilité des données et des modèles d'accès nécessaires à vos équipes. Vous pouvez intégrer des outils pour corréler, visualiser et déduire des informations à partir des données.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

La croissance de l'utilisation d'AWS au sein d'une organisation entraîne une augmentation du nombre de charges de travail et d'environnements distribués. Dans la mesure où chaque charge de travail et environnement génère des données sur l'activité qui s'y déroule, la capture et le stockage de ces données localement constituent un défi pour les opérations de sécurité. Les équipes de sécurité utilisent des outils tels que les systèmes de gestion des informations et des événements de sécurité (SIEM) pour collecter des données à partir de sources distribuées et effectuer des flux de travail de corrélation, d'analyse et de réponse. Ces opérations requièrent la gestion d'un ensemble complexe d'autorisations pour accéder aux différentes sources de données et impliquent des frais supplémentaires liés à l'exploitation des processus d'extraction, de transformation et de chargement (ETL).

Pour surmonter ces difficultés, pensez à agréger toutes les sources pertinentes de données des journaux de sécurité dans un compte Log Archive, comme décrit dans la section Organisation de votre environnement AWS à l'aide de plusieurs comptes. Cela inclut toutes les données liées à la sécurité provenant de votre charge de travail et les journaux générés par les services AWS, tels que AWS CloudTrail, AWS WAF, Elastic Load Balancing et Amazon Route 53. La saisie de ces données dans des emplacements standardisés dans un Compte AWS avec des autorisations intercomptes appropriées présente plusieurs avantages. Cette pratique permet d'empêcher l'altération des journaux dans les charges de travail et les environnements compromis, fournit un point d'intégration unique pour des outils supplémentaires et propose un modèle plus simplifié pour configurer la conservation et le cycle de vie des données. Évaluez les impacts de la souveraineté des données, des périmètres de conformité et d'autres réglementations afin de déterminer si plusieurs emplacements de stockage des données de sécurité et périodes de conservation sont nécessaires.

Pour faciliter la capture et la standardisation des journaux et des résultats, évaluez Amazon Security Lake dans votre compte d'archivage des journaux. Vous pouvez configurer Security Lake pour ingérer automatiquement les données provenant de sources courantes telles que CloudTrail, Route 53, Amazon EKS et les journaux de flux VPC. Vous pouvez également configurer AWS Security Hub en tant que source de données dans Security Lake, ce qui vous permet de corréler les résultats d'autres services AWS, tels qu'Amazon GuardDuty et Amazon Inspector, avec les données de vos journaux. Vous pouvez également utiliser des intégrations de sources de données tierces ou configurer des sources de données personnalisées. Toutes les intégrations normalisent vos données au format OCSF (Open Cybersecurity Schema Framework) et sont stockées dans des compartiments Amazon S3 sous forme de fichiers Parquet, éliminant ainsi le besoin de traitement ETL.

Le stockage des données de sécurité dans des emplacements standardisés fournit des fonctionnalités d'analyse avancées. AWS vous recommande de déployer des outils d'analyse de sécurité qui fonctionnent dans un environnement AWS dans un compte Security Tooling distinct de votre compte d'archivage des journaux. Cette approche vous permet de mettre en œuvre des contrôles approfondis afin de protéger l'intégrité et la disponibilité des journaux et du processus de gestion des journaux, indépendamment des outils qui y accèdent. Envisagez d'utiliser des services tels qu'Amazon Athena pour exécuter des requêtes à la demande qui mettent en corrélation plusieurs sources de données. Vous pouvez également intégrer des outils de visualisation, tels qu'Amazon QuickSight. Les solutions optimisées par l'IA sont de plus en plus disponibles et peuvent exécuter des fonctions telles que la conversion des résultats en résumés lisibles par l'homme et l'interaction en langage naturel. L'intégration de ces solutions est généralement plus simple si vous disposez d'un emplacement de stockage de données standardisé pour les requêtes.

Étapes d'implémentation

1. Création des comptes d'archivage des journaux et d'outils de sécurité
   a. À l'aide d'AWS Organizations, créez les comptes d'archivage des journaux et d'outils de sécurité dans une unité organisationnelle de sécurité. Si vous utilisez AWS Control Tower pour gérer votre organisation, les comptes d'archivage des journaux et d'outils de sécurité sont créés automatiquement pour vous. Configurez les rôles et les autorisations pour accéder à ces comptes et les administrer selon les besoins.
2. Configurer vos emplacements de données de sécurité standardisés
   a. Déterminez votre stratégie pour créer des emplacements de données de sécurité standardisés. Vous pouvez y parvenir grâce à des options telles que des approches d'architecture de lac de données courantes, des produits de données tiers ou Amazon Security Lake. AWS vous recommande de capturer les données de sécurité à partir des Régions AWS qui sont activées pour vos comptes, même lorsque vous ne les utilisez pas activement.
3. Configurer la publication des sources de données dans vos emplacements standardisés
   a. Identifiez les sources de vos données de sécurité et configurez-les pour les publier dans vos emplacements standardisés. Évaluez les options permettant d'exporter automatiquement les données dans le format souhaité, par opposition à celles nécessitant le développement de processus ETL. Avec Amazon Security Lake, vous pouvez collecter des données à partir de sources AWS prises en charge et de systèmes tiers intégrés.
4. Configurer des outils pour accéder à vos emplacements standardisés
   a. Configurez des outils tels qu'Amazon Athena, Amazon QuickSight, ou des solutions tierces pour disposer de l'accès requis à vos emplacements standardisés. Configurez ces outils de façon à ce qu'ils fonctionnent à partir du compte d'outils de sécurité avec un accès en lecture intercompte au compte d'archivage des journaux, le cas échéant. Créez des abonnés dans Amazon Security Lake pour permettre à ces outils d'accéder à vos données.

Ressources

Bonnes pratiques associées :

- SEC01-BP01 Séparer les charges de travail à l'aide de comptes
- SEC07-BP04 Définir la gestion du cycle de vie des données
- SEC08-BP04 Appliquer le contrôle d'accès
- OPS08-BP02 Analyse des journaux de charge de travail

Documents connexes :

- Livres blancs AWS : organisation de votre environnement AWS à l'aide de comptes multiple
- Conseils prescriptifs AWS : architecture de référence de sécurité AWS (SRA AWS)
- Conseils prescriptifs AWS : guide journalisation et surveillance pour les propriétaires d'applications

Exemples connexes :

- Agrégation, recherche et visualisation des données de journal provenant de sources distribuées avec Amazon Athena et Amazon QuickSight
- Comment visualiser les résultats d'Amazon Security Lake avec Amazon QuickSight
- Génération d'informations optimisées par l'IA pour Amazon Security Lake à l'aide d'Amazon SageMaker AI Studio et d'Amazon Bedrock
- Identification des anomalies de cybersécurité dans vos données Amazon Security Lake à l'aide de l'IA Amazon SageMaker
- Ingestion, transformation et diffusion des événements publiés par Amazon Security Lake sur Amazon OpenSearch Service
- Simplification de l'analyse des journaux AWS CloudTrail via la génération de requêtes en langage naturel dans CloudTrail Lake

Outils associés :

- Amazon Security Lake
- Intégrations des partenaires Amazon Security Lake
- Cadre de schéma de cybersécurité ouvert (OCSF)
- Amazon Athena
- Amazon QuickSight
- Amazon Bedrock

SEC04-BP03 Corréler et enrichir les alertes de sécurité

Les activités inattendues peuvent générer plusieurs alertes de sécurité provenant de différentes sources, ce qui nécessite une corrélation et un enrichissement supplémentaires pour comprendre le contexte complet. Mettez en œuvre une corrélation et un enrichissement automatisés des alertes de sécurité afin de permettre une identification et une réponse plus précises aux incidents.

Résultat escompté : au fur et à mesure que l'activité génère différentes alertes au sein de vos charges de travail et de vos environnements, des mécanismes automatisés mettent en corrélation les données et les enrichissent avec des informations supplémentaires. Ce prétraitement permet de mieux comprendre l'événement, ce qui aide vos enquêteurs à déterminer sa criticité et s'il s'agit d'un incident qui requiert une réponse officielle. Ce processus réduit la charge de travail de vos équipes de surveillance et d'enquête.

Anti-modèles courants :

- Différents groupes de personnes étudient les résultats et les alertes générés par différents systèmes, sauf si les exigences relatives à la séparation des tâches en disposent autrement.
- Votre organisation achemine tous les résultats de sécurité et toutes les données d'alerte vers des emplacements standard, mais demande aux enquêteurs d'effectuer une corrélation et un enrichissement manuels.
- Vous vous fiez uniquement à l'intelligence des systèmes de détection des menaces pour rendre compte des résultats et établir la criticité.

Avantages du respect de cette bonne pratique : la corrélation et l'enrichissement automatisés des alertes contribuent à réduire la charge cognitive globale et la préparation manuelle des données requises par vos enquêteurs. Cette pratique permet de réduire le temps nécessaire pour déterminer si l'événement représente un incident et lancer une réponse officielle. Un contexte supplémentaire vous permet également d'évaluer avec précision la gravité réelle d'un événement, car celle-ci peut être supérieure ou inférieure à ce que suggère une alerte.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Les alertes de sécurité peuvent provenir de nombreuses sources différentes dans AWS, y compris :

- Des services tels qu'Amazon GuardDuty, AWS Security Hub, Amazon Macie, Amazon Inspector, AWS Config, AWS Identity and Access Management Access Analyzer et Analyseur d'accès réseau
- Alertes issues de l'analyse automatique des journaux de service AWS, d'infrastructure et d'application, telles que celles issues de Security Analytics pour Amazon OpenSearch Service.
- Alarmes en réponse à des modifications de votre activité de facturation provenant de sources telles qu'Amazon CloudWatch, Amazon EventBridge ou AWS Budgets.
- Des sources tierces, telles que les flux de renseignements sur les menaces et les solutions des partenaires de sécurité d'AWS Partner Network.

- Contact par AWS Trust & Safety ou par d'autres sources, telles que des clients ou des employés internes.

Dans leur forme la plus fondamentale, les alertes contiennent des informations sur qui (le principal ou l'identité) fait quoi (les mesures prises) à quoi (les ressources concernées). Pour chacune de ces sources, déterminez s'il existe des moyens de créer des mappages entre les identifiants de ces identités, actions et ressources afin d'effectuer une corrélation. À cet effet, vous pouvez notamment intégrer des sources d'alerte à un outil de gestion des informations et des événements de sécurité (SIEM) afin d'effectuer une corrélation automatique, créer vos propres pipelines et traitements de données, ou mettre en place une combinaison de ces deux solutions.

Amazon Detective est un exemple de service capable d'effectuer une corrélation pour vous. Detective ingère en permanence les alertes provenant de différentes sources AWS et tierces. Il utilise différentes formes d'informations pour créer un graphique visuel de leurs relations afin de faciliter les enquêtes.

Alors que la criticité initiale d'une alerte facilite l'établissement des priorités, le contexte dans lequel l'alerte s'est produite détermine sa véritable criticité. Par exemple, Amazon GuardDuty peut vous avertir qu'une instance Amazon EC2 de votre charge de travail demande un nom de domaine inattendu. GuardDuty peut attribuer à elle seule une faible criticité à cette alerte. Cependant, une corrélation automatique avec d'autres activités au moment de l'alerte peut révéler que plusieurs centaines d'instances EC2 ont été déployées sous la même identité, ce qui augmente les coûts d'exploitation globaux. Dans ce cas, le contexte de l'événement corrélé justifierait une nouvelle alerte de sécurité et la criticité pourrait être portée à un niveau élevé, ce qui accélérerait la mise en place d'une réponse.

Étapes d'implémentation

1. Identifiez les sources d'informations relatives aux alertes de sécurité. Comprenez comment les alertes de ces systèmes représentent l'identité, l'action et les ressources afin de déterminer où une corrélation est possible.
2. Mettez en place un mécanisme permettant de capturer les alertes provenant de différentes sources. À cette fin, pensez à des services tels que Security Hub, EventBridge et CloudWatch.
3. Identifiez les sources pour la corrélation et l'enrichissement des données. Les exemples de sources incluent AWS CloudTrail, les journaux de flux VPC, les journaux de Route 53 Resolver et les journaux d'infrastructure et d'application. Tout ou partie de ces journaux peuvent être consommés par le biais d'une seule intégration avec Amazon Security Lake.

4. Intégrez les alertes à vos sources de corrélation et d'enrichissement des données pour créer des contextes d'événements de sécurité plus détaillés et établir leur criticité.
   a. Amazon Detective, les outils SIEM ou d'autres solutions tierces peuvent effectuer automatiquement un certain niveau d'ingestion, de corrélation et d'enrichissement.
   b. Vous pouvez également utiliser des services AWS pour créer le vôtre. Par exemple, vous pouvez invoquer une fonction AWS Lambda pour exécuter une requête Amazon Athena par rapport à AWS CloudTrail ou Amazon Security Lake, et publier les résultats dans EventBridge.

Ressources

Bonnes pratiques associées :

- SEC10-BP03 Préparer les fonctionnalités d'analyse poussée
- OPS08-BP04 Création d'alertes exploitables
- REL06-BP03 Envoyer des notifications (traitement et alarmes en temps réel)

Documents connexes :

- Guide d'intervention en cas d'incident de sécurité AWS

Exemples associés :

- Comment enrichir les résultats AWS Security Hub grâce aux métadonnées des comptes

Outils associés :

- Amazon Detective
- Amazon EventBridge
- AWS Lambda
- Amazon Athena

SEC04-BP04 Lancer la correction pour les ressources non conformes

Vos contrôles de détection peuvent signaler la présence de ressources non conformes à vos exigences de configuration. Vous pouvez lancer des mesures correctives définies par programme, manuellement ou automatiquement, afin de corriger ces ressources et de minimiser les impacts

potentiels. Lorsque vous définissez des mesures correctives par programmation, vous pouvez agir rapidement et avec cohérence.

Bien que l'automatisation puisse améliorer les opérations de sécurité, vous devez la mettre en œuvre et la gérer avec soin. Mettez en place des mécanismes de supervision et de contrôle appropriés pour vérifier que les réponses automatisées sont efficaces, précises et conformes aux politiques organisationnelles et à la propension au risque.

Résultat escompté : vous définissez les normes de configuration des ressources ainsi que les étapes à suivre pour corriger les ressources détectées comme étant non conformes. Dans la mesure du possible, vous avez défini les mesures correctives par programmation afin qu'elles puissent être lancées manuellement ou automatiquement. Des systèmes de détection sont en place pour identifier les ressources non conformes et publier des alertes dans des outils centralisés surveillés par votre personnel de sécurité. Ces outils vous permettent d'exécuter vos corrections programmatiques, manuellement ou automatiquement. Les mesures correctives automatiques sont dotées de mécanismes de supervision et de contrôle appropriés pour régir leur utilisation.

Anti-modèles courants :

- Vous mettez en œuvre l'automatisation, mais vous ne parvenez pas à tester ni à valider de manière approfondie les mesures correctives. Cela peut avoir des conséquences imprévues, telles que la perturbation des opérations commerciales légitimes ou l'instabilité du système.
- Vous améliorez les temps de réponse et les procédures grâce à l'automatisation, mais sans surveillance appropriée et sans mécanismes permettant une intervention et un discernement humains en cas de besoin.
- Vous vous fiez uniquement aux mesures correctives, au lieu de les intégrer dans le cadre d'un programme plus large de réponse aux incidents et de reprise.

Avantages du respect de cette bonne pratique : les corrections automatiques peuvent répondre aux erreurs de configuration plus rapidement que les processus manuels, ce qui vous permet de minimiser les impacts commerciaux potentiels et de réduire les opportunités d'utilisation involontaire. Lorsque vous définissez des mesures correctives de manière programmatique, elles sont appliquées de manière cohérente, ce qui réduit le risque d'erreur humaine. L'automatisation peut également gérer simultanément un plus grand volume d'alertes, ce qui est particulièrement important dans les environnements fonctionnant à grande échelle.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Comme décrit dans SEC01-BP03 Identifier et valider les objectifs de contrôle, des services tels que AWS Config et AWS Security Hub peuvent vous aider à surveiller la configuration des ressources de vos comptes afin de garantir leur conformité à vos exigences. Lorsque des ressources non conformes sont détectées, des services tels que AWS Security Hub peuvent aider à acheminer les alertes de manière appropriée et à prendre des mesures correctives. Ces solutions fournissent à vos enquêteurs de sécurité un emplacement central qui leur permet de surveiller les problèmes et de prendre des mesures correctives.

Alors que certaines situations de ressources non conformes sont uniques et requièrent un discernement humain pour être résolues, d'autres situations ont besoin d'une réponse standard que vous pouvez définir par programmation. Par exemple, une solution standard à un problème de configuration du groupe de sécurité VPC peut consister à supprimer les règles d'interdiction et à en informer le propriétaire. Les réponses peuvent être définies dans les fonctions AWS Lambda, les documents AWS Systems Manager Automation ou via d'autres environnements de code que vous préférez. Assurez-vous que l'environnement est capable de s'authentifier auprès d'AWS à l'aide d'un rôle IAM avec le moins d'autorisations nécessaires pour prendre des mesures correctives.

Une fois que vous avez défini la correction souhaitée, vous pouvez ensuite déterminer le moyen par lequel vous préférez la lancer. AWS Config peut initier des mesures correctives pour vous. Si vous utilisez Security Hub, vous pouvez le faire par le biais d'actions personnalisées, qui publient les informations de résultat sur Amazon EventBridge. Une règle EventBridge peut ensuite lancer votre correction. Vous pouvez configurer les corrections via Security Hub pour qu'elles s'exécutent automatiquement ou manuellement.

Pour la correction programmatique, nous vous recommandons de disposer de journaux et d'audits complets des actions entreprises, ainsi que de leurs résultats. Passez en revue et analysez ces journaux pour évaluer l'efficacité des processus automatisés et identifier les domaines à améliorer. Capturez les journaux dans Amazon CloudWatch Logs et les résultats des mesures correctives sous forme de notes de résultat dans Security Hub.

Comme point de départ, pensez à Automated Security Response on AWS, qui propose des correctifs prédéfinis pour résoudre les erreurs de configuration de sécurité courantes.

Étapes d'implémentation

1. Analysez et hiérarchisez les alertes.

   a. Regroupez les alertes de sécurité provenant de différents services AWS dans Security Hub pour centraliser la visibilité, la hiérarchisation et les mesures correctives.

2. Élaborez des mesures correctives.
   a. Utilisez des services tels que Systems Manager et AWS Lambda pour exécuter des corrections programmatiques.
3. Configurez la façon dont les mesures correctives sont lancées.
   a. À l'aide de Systems Manager, définissez des actions personnalisées qui publient les résultats dans EventBridge. Configurez ces actions de façon à ce qu'elles soient lancées manuellement ou automatiquement.
   b. Vous pouvez également utiliser Amazon Simple Notification Service (SNS) pour envoyer des notifications et des alertes aux parties prenantes concernées (comme l'équipe de sécurité ou les équipes de réponse aux incidents) pour une intervention manuelle ou une escalade, si nécessaire.
4. Passez en revue et analysez les journaux de correction afin d'en vérifier l'efficacité et de les améliorer.
   a. Envoyez une sortie de journal à CloudWatch Logs. Enregistrez les résultats sous forme de notes de résultats dans Security Hub.

Ressources

Bonnes pratiques associées :

- SEC06-BP03 Réduire la gestion manuelle et l'accès interactif

Documents connexes :

- Guide d'intervention en cas d'incident de sécurité AWS - Détection

Exemples connexes :

- Automated Security Response on AWS
- Surveiller les paires de clés d'instance EC2 à l'aide de AWS Config
- Créer des règles AWS Config personnalisées à l'aide de stratégies AWS CloudFormation Guard
- Corriger automatiquement les instances et clusters de base de données Amazon RDS non chiffrés

Outils associés :

- AWS Systems Manager Automation
- Automated Security Response on AWS

Protection de l’infrastructure

Questions

- SÉC 5. Comment protéger vos ressources réseau ?
- SÉC 6. Comment protéger vos ressources informatiques ?

SÉC 5. Comment protéger vos ressources réseau ?

Pour toute charge de travail ayant une forme quelconque de connectivité réseau, qu’il s’agisse d’Internet ou d’un réseau privé, plusieurs couches de défense sont nécessaires pour vous protéger contre les menaces externes et internes basées sur le réseau.

Bonnes pratiques

- SEC05-BP01 Création de couches réseau
- SEC05-BP02 Contrôlez le flux de trafic au sein des couches de votre réseau
- SEC05-BP03 Mettre en œuvre une protection basée sur l'inspection
- SEC05-BP04 Automatiser la protection du réseau

SEC05-BP01 Création de couches réseau

Segmentez la topologie de votre réseau en différentes couches en procédant à des regroupements logiques des composants de votre charge de travail en fonction de la sensibilité des données et de leurs exigences en matière d’accès. Faites la distinction entre les composants qui requièrent un accès entrant depuis Internet, comme les points de terminaison Web publics, et ceux qui requièrent uniquement un accès interne, comme les bases de données.

Résultat souhaité : Les couches de votre réseau font partie d'une defense-in-depth approche intégrale de la sécurité qui complète la stratégie d'authentification et d'autorisation des identités de vos charges de travail. Les couches sont positionnées en fonction de la sensibilité des données et des exigences d’accès, avec des mécanismes de flux de trafic et de contrôle appropriés.

Anti-modèles courants :

- Vous créez toutes les ressources dans un seul VPC ou un sous-réseau.
- Vous construisez vos couches réseau sans tenir compte des exigences de sensibilité des données, du comportement des composants ou des fonctionnalités.
- Vous utilisez VPCs des sous-réseaux et par défaut pour toutes les considérations relatives à la couche réseau, sans tenir compte de l'influence des services AWS gérés sur votre topologie.

Avantages du respect de cette bonne pratique : la mise en place de couches réseau est la première étape pour limiter les chemins inutiles à travers le réseau, en particulier ceux qui mènent à des systèmes et à des données critiques. Il est donc plus difficile pour les acteurs non autorisés d'accéder à votre réseau et aux ressources supplémentaires qu'il contient. Les couches réseau individuelles présentent l'avantage de réduire la portée de l'analyse des systèmes d'inspection, par exemple pour la détection des intrusions ou la prévention des programmes malveillants. Cela réduit le risque de faux positifs et les frais de traitement inutiles.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Lors de la conception d'une architecture de charge de travail, il est courant de séparer les composants en différentes couches en fonction de leur responsabilité. Par exemple, une application Web peut comporter une couche de présentation, une couche d'application et une couche de données. Vous pouvez adopter une approche similaire lors de la conception de la topologie de votre réseau. Les contrôles réseau sous-jacents peuvent vous aider à faire respecter les exigences d'accès aux données de votre charge de travail. Par exemple, dans une architecture d'application Web à trois niveaux, vous pouvez stocker vos fichiers de couche de présentation statiques sur Amazon S3 et les diffuser à partir d'un réseau de diffusion de contenu (CDN), tel qu'Amazon CloudFront. La couche application peut comporter des points de terminaison publics qu'un Application Load Balancer ALB () dessert dans un sous-réseau public VPCAmazon (similaire à une zone démilitarisée, DMZ ou), avec des services principaux déployés dans des sous-réseaux privés. La couche de données, qui héberge des ressources telles que des bases de données et des systèmes de fichiers partagés, peut résider dans des sous-réseaux privés différents des ressources de votre couche d'application. À chacune de ces limites de couche (sous-réseau publicCDN, sous-réseau privé), vous pouvez déployer des contrôles permettant uniquement au trafic autorisé de franchir ces limites.

Comme pour la modélisation des couches réseau en fonction de l'objectif fonctionnel des composants de votre charge de travail, tenez également compte de la sensibilité des données traitées. Dans l'exemple de l'application Web, bien que tous vos services de charge de travail puissent résider dans la couche d'application, différents services peuvent traiter des données avec des niveaux de sensibilité différents. Dans ce cas, il peut être approprié de diviser la couche d'application en utilisant plusieurs sous-réseaux privés Compte AWS, différents VPCs dans le même cas, ou même différents Comptes AWS pour chaque niveau de sensibilité des données, VPCs en fonction de vos exigences de contrôle.

La cohérence du comportement des composants de votre charge de travail est un autre facteur à prendre en compte pour les couches réseau. Si nous poursuivons avec le même exemple, dans la couche d'application, vous pouvez avoir des services qui acceptent des entrées provenant d'utilisateurs finaux ou d'intégrations de systèmes externes qui sont intrinsèquement plus risquées que les entrées d'autres services. Il peut notamment s'agir du téléchargement de fichiers, de scripts de code à exécuter, de l'analyse d'e-mails, etc. Le fait de placer ces services dans leur propre couche réseau contribue à créer une limite d'isolation plus forte autour d'eux et peut empêcher leur comportement unique de créer des alertes faussement positives dans les systèmes d'inspection.

Dans le cadre de votre conception, réfléchissez à l'influence de l'utilisation des services AWS gérés sur la topologie de votre réseau. Découvrez comment des services tels qu'Amazon VPC Lattice peuvent faciliter l'interopérabilité des composants de votre charge de travail entre les couches réseau. Lors de l'utilisation AWS Lambda, déployez dans vos VPC sous-réseaux, sauf pour des raisons spécifiques. Déterminez où se trouvent les VPC terminaux et AWS PrivateLinkpouvez simplifier le respect des politiques de sécurité qui limitent l'accès aux passerelles Internet.

Étapes d'implémentation

1. Passez en revue l'architecture de votre charge de travail. Regroupez logiquement les composants et les services selon les fonctions qu'ils remplissent, la sensibilité des données traitées et leur comportement.
2. En ce qui concerne les composants qui répondent à des demandes provenant d'Internet, pensez à utiliser des équilibreurs de charge ou d'autres proxys pour fournir des points de terminaison publics. Explorez l'évolution des contrôles de sécurité en utilisant des services gérés, tels qu'Amazon API Gateway CloudFront, Elastic Load Balancing, et AWS Amplifypour héberger des points de terminaison publics.
3. Pour les composants exécutés dans des environnements informatiques, tels que les EC2 instances Amazon, les AWS Fargateconteneurs ou les fonctions Lambda, déployez-les dans des sous-réseaux privés en fonction de vos groupes dès la première étape.

4. Pour les AWS services entièrement gérés, tels qu'Amazon DynamoDB, Amazon Kinesis ou SQSAmazon, envisagez d'utiliser des points de terminaison par défaut pour VPC l'accès via des adresses IP privées.

Ressources

Bonnes pratiques associées :

- REL02 Planifiez la topologie de votre réseau
- PERF04-BP01 Comprendre l'impact du réseau sur les performances

Vidéos connexes :

- AWS re:Invent 2023 - AWS mise en réseau des fondations

Exemples connexes :

- VPCexemples
- Accédez aux applications de conteneur en privé sur Amazon en ECS utilisant AWS FargateAWS PrivateLink, et un Network Load Balancer
- Diffusez du contenu statique dans un compartiment Amazon S3 via un VPC en utilisant Amazon CloudFront

SEC05-BP02 Contrôlez le flux de trafic au sein des couches de votre réseau

Au sein des couches de votre réseau, segmentez davantage pour limiter le trafic uniquement aux flux nécessaires à chaque charge de travail. Tout d'abord, concentrez-vous sur le contrôle du trafic entre Internet ou d'autres systèmes externes vers une charge de travail et votre environnement (trafic nord-sud). Ensuite, examinez les flux entre les différents composants et systèmes (trafic est-ouest).

Résultat souhaité : vous autorisez uniquement les flux réseau nécessaires aux composants de vos charges de travail pour communiquer entre eux, avec leurs clients et avec tout autre service dont ils dépendent. Votre conception prend en compte des facteurs tels que les entrées et sorties publiques par rapport aux entrées et sorties privées, la classification des données, les réglementations régionales et les exigences en matière de protocole. Dans la mesure du possible, vous privilégiez point-to-point les flux par rapport au peering réseau dans le cadre d'une conception fondée sur le principe du moindre privilège.

Anti-modèles courants :

- Vous adoptez une approche de la sécurité du réseau basée sur le périmètre et vous ne contrôlez le flux de trafic qu'à la limite des couches de votre réseau.
- Vous supposez que tout le trafic au sein d'une couche réseau est authentifié et autorisé.
- Vous appliquez des contrôles à votre trafic d'entrée ou de sortie, mais pas aux deux.
- Vous vous fiez uniquement aux composants de votre charge de travail et aux contrôles réseau pour authentifier et autoriser le trafic.

Avantages du respect de cette bonne pratique : cette pratique permet de réduire le risque de mouvements non autorisés au sein de votre réseau et ajoute une couche d'autorisation supplémentaire à vos charges de travail. En contrôlant le flux de trafic, vous pouvez limiter l'ampleur de l'impact d'un incident de sécurité et accélérer la détection et la réponse.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Bien que les couches réseau aident à définir les limites des composants de votre charge de travail qui remplissent une fonction, un niveau de sensibilité des données et un comportement similaires, vous pouvez créer un niveau de contrôle du trafic beaucoup plus précis en utilisant des techniques permettant de segmenter davantage les composants au sein de ces couches conformément au principe du moindre privilège. À l'intérieur AWS, les couches réseau sont principalement définies à l'aide de sous-réseaux en fonction des plages d'adresses IP au sein d'un AmazonVPC. Les couches peuvent également être définies à l'aide de différentes VPCs méthodes, par exemple pour regrouper des environnements de microservices par domaine d'activité. Lorsque vous utilisez plusieursVPCs, négociez le routage à l'aide d'un AWS Transit Gateway. Bien que cela permette de contrôler le trafic au niveau de la couche 4 (adresses IP et plages de ports) à l'aide de groupes de sécurité et de tables de routage, vous pouvez renforcer le contrôle grâce à des services supplémentaires AWS PrivateLink, tels que Amazon Route 53 Resolver DNS Firewall et AWS WAF. AWS Network Firewall

Comprenez et inventoriez le flux de données et les exigences de communication de vos charges de travail en termes de parties initiatrices de connexion, de ports, de protocoles et de couches réseau. Évaluez les protocoles disponibles pour établir des connexions et transmettre des données afin de sélectionner ceux qui répondent à vos exigences de protection (par exemple, HTTPS plutôt queHTTP). Capturez ces exigences à la fois aux limites de vos réseaux et au sein de chaque couche. Une fois ces exigences identifiées, explorez les options permettant d'autoriser uniquement

le trafic requis à circuler à chaque point de connexion. Un bon point de départ consiste à utiliser des groupes de sécurité au sein de votre entrepriseVPC, car ils peuvent être associés à des ressources utilisant une interface réseau élastique (ENI), telles que des EC2 instances Amazon, des ECS tâches Amazon, EKS des pods Amazon ou des RDS bases de données Amazon. Contrairement à un pare-feu de couche 4, un groupe de sécurité peut avoir une règle qui autorise le trafic provenant d'un autre groupe de sécurité en fonction de son identifiant, minimisant ainsi les mises à jour à mesure que les ressources du groupe changent au fil du temps. Vous pouvez également filtrer le trafic à l'aide de règles entrantes et sortantes à l'aide de groupes de sécurité.

Lorsque le trafic se déplace entre les deuxVPCs, il est courant d'utiliser le VPC peering pour un routage simple ou AWS Transit Gateway pour un routage complexe. Grâce à ces approches, vous facilitez les flux de trafic entre la plage d'adresses IP des réseaux source et de destination. Toutefois, si votre charge de travail ne nécessite que des flux de trafic entre des composants spécifiques de différentes manièresVPCs, pensez à utiliser une point-to-point connexion utilisant AWS PrivateLink. Pour ce faire, identifiez quel service doit agir en tant que producteur et lequel doit agir en tant que consommateur. Déployez un équilibreur de charge compatible pour le producteur, PrivateLink activez-le en conséquence, puis acceptez une demande de connexion du consommateur. Le service du producteur se voit ensuite attribuer une adresse IP privée provenant de VPC celle du consommateur, que celui-ci peut utiliser pour effectuer des demandes ultérieures. Cette approche réduit le besoin d'appairage entre les réseaux. Incluez les coûts de traitement des données et d'équilibrage de charge dans le cadre de l'évaluation PrivateLink.

Bien que les groupes de sécurité PrivateLink aident à contrôler le flux entre les composants de vos charges de travail, une autre considération majeure est de savoir comment contrôler les DNS domaines auxquels vos ressources sont autorisées à accéder (le cas échéant). En fonction de la DHCP configuration de votreVPCs, vous pouvez envisager deux AWS services différents à cette fin. La plupart des clients utilisent le DNS service Route 53 Resolver par défaut (également appelé DNS serveur Amazon ou AmazonProvidedDNS) disponible VPCs à l'adresse +2 de sa CIDR plage. Grâce à cette approche, vous pouvez créer des règles de DNS pare-feu et les associer à vos règles VPC afin de déterminer les actions à entreprendre pour les listes de domaines que vous fournissez.

Si vous n'utilisez pas le Route 53 Resolver, ou si vous souhaitez le compléter par des fonctionnalités d'inspection et de contrôle de flux plus approfondies allant au-delà du filtrage de domaine, envisagez de déployer un AWS Network Firewall. Ce service inspecte les paquets individuels en utilisant des règles sans ou avec état afin de déterminer s'il est nécessaire de refuser ou d'autoriser le trafic. Vous pouvez adopter une approche similaire pour filtrer le trafic Web entrant vers vos points de terminaison publics à l'aide de AWS WAF. Pour plus d'informations sur ces services, voir SEC05-BP03 Mettre en œuvre une protection basée sur l'inspection.

Étapes d’implémentation

1. Identifiez les flux de données requis entre les composants de vos charges de travail.
2. Appliquez plusieurs contrôles selon une defense-in-depth approche à la fois pour le trafic entrant et sortant, notamment en utilisant des groupes de sécurité et des tables de routage.
3. Utilisez des pare-feux pour définir un contrôle précis du trafic réseau entrant, sortant et traversant votre réseauVPCs, comme le DNS pare-feu Route 53 Resolver, et. AWS Network Firewall AWS WAF Envisagez d’utiliser le AWS Firewall Manager pour configurer et gérer de manière centralisée les règles de pare-feu au sein de votre organisation.

Ressources

Bonnes pratiques associées :

- REL03-BP01 Choisissez comment segmenter votre charge de travail
- SEC09-BP02 Appliquer le chiffrement en transit

Documents connexes :

- Bonnes pratiques en matière de sécurité pour votre VPC
- Conseils d’optimisation du réseau AWS
- Conseils pour la sécurité du réseau sur AWS
- Sécurisez VPC le trafic réseau sortant dans AWS Cloud

Outils associés :

- AWS Firewall Manager

Vidéos connexes :

- AWS Transit Gateway architectures de référence pour de nombreuses VPCs
- Accélération et protection des applications avec Amazon CloudFront AWS WAF, et AWS Shield
- AWS re:Inforce 2023: Firewalls and where to put them

Exemples connexes :

- Atelier : CloudFront pour les applications Web

SEC05-BP03 Mettre en œuvre une protection basée sur l'inspection

Configurez des points d'inspection du trafic entre les couches de votre réseau afin de vous assurer que les données en transit correspondent aux catégories et aux modèles attendus. Analysez les flux de trafic, les métadonnées et les modèles pour identifier et détecter les événements, et y répondre plus efficacement.

Résultat souhaité : le trafic qui passe d'une couche à l'autre de votre réseau est inspecté et autorisé. Les décisions d'autorisation et de refus sont basées sur des règles explicites, des informations sur les menaces et des écarts par rapport aux comportements de base. Les protections deviennent plus strictes à mesure que le trafic se rapproche des données sensibles.

Anti-modèles courants :

- S'appuyer uniquement sur les règles de pare-feu basées sur les ports et les protocoles. Ne pas tirer parti des systèmes intelligents.
- Créer des règles de pare-feu basées sur des modèles de menaces actuels spécifiques susceptibles de changer.
- Inspecter uniquement le trafic transitant des sous-réseaux privés vers des sous-réseaux publics, ou des sous-réseaux publics vers Internet.
- Ne pas disposer d'une vue de base de votre trafic réseau à utiliser à titre de comparaison afin de détecter les anomalies de comportement.

Avantages du respect de cette bonne pratique : les systèmes d'inspection vous permettent de créer des règles intelligentes, telles que l'autorisation ou le refus du trafic uniquement lorsque certaines conditions relatives aux données de trafic existent. Bénéficiez d'ensembles de règles gérés par les partenaires AWS et basés sur les informations les plus récentes sur les menaces, à mesure que le paysage des menaces évolue au fil du temps. Cela réduit les frais liés à la mise à jour des règles et à la recherche d'indicateurs de compromis, réduisant ainsi le risque de faux positifs.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Contrôlez avec précision votre trafic réseau dynamique et apatride à l'aide AWS Network Firewall d'autres pare-feux et systèmes de prévention des intrusions (IPS) AWS Marketplace que vous

pouvez déployer derrière un Gateway Load Balancer (). GWLB AWS Network Firewall prend en charge les IPS spécifications open source compatibles avec Suricata pour protéger votre charge de travail.

Les solutions AWS Network Firewall des fournisseurs qui utilisent un GWLB prennent en charge différents modèles de déploiement de l'inspection en ligne. Par exemple, vous pouvez effectuer une inspection sur une VPC base individuelle, la centraliser dans le cadre d'une inspection VPC ou la déployer dans un modèle hybride dans lequel le trafic est-ouest passe par une inspection VPC et où les entrées Internet sont inspectées par inspection. VPC Une autre considération est de savoir si la solution prend en charge le déballage de Transport Layer Security (TLS), permettant une inspection approfondie des paquets pour les flux de trafic initiés dans les deux sens. Pour plus d'informations et des détails détaillés sur ces configurations, consultez le guide des bonnes pratiques AWS Network Firewall.

Si vous utilisez des solutions qui effectuent des out-of-band inspections, telles que l'analyse pcap des données par paquets provenant d'interfaces réseau fonctionnant en mode promiscuité, vous pouvez configurer VPC la mise en miroir du trafic. Le trafic en miroir est pris en compte dans la bande passante disponible de vos interfaces et il est soumis aux mêmes frais de transfert de données que le trafic non mis en miroir. Vous pouvez voir si des versions virtuelles de ces appliances sont disponibles sur le AWS Marketplace, qui peut prendre en charge le déploiement en ligne derrière unGWLB.

Pour les composants qui effectuent des transactions via des protocoles HTTP basés, protégez votre application contre les menaces courantes à l'aide d'un pare-feu pour applications Web (WAF). AWS WAFest un pare-feu d'applications Web qui vous permet de surveiller et de bloquer les demandes HTTP (S) conformes à vos règles configurables avant de les envoyer à Amazon API Gateway CloudFront, Amazon AWS AppSync ou à un Application Load Balancer. Envisagez une inspection approfondie des paquets lorsque vous évaluez le déploiement de votre pare-feu d'applications Web, car certains nécessitent que vous vous arrêtiez TLS avant d'inspecter le trafic. Pour commencer AWS WAF, vous pouvez l'utiliser AWS Managed Rulesen combinaison avec les vôtres ou utiliser les intégrations de partenaires existantes.

Vous pouvez gérer de manière centralisée AWS WAF, AWS Shield Advanced AWS Network Firewall, et les groupes VPC de sécurité Amazon au sein de votre AWS organisation avec AWS Firewall Manager.

Étapes d’implémentation

1. Déterminez si vous pouvez élargir la portée des règles d'inspection, par exemple par le biais d'une inspectionVPC, ou si vous avez besoin d'une approche plus précise par VPC approche.
2. Pour les solutions d’inspection en ligne :
   a. Si vous l'utilisez AWS Network Firewall, créez des règles, des politiques de pare-feu et le pare-feu lui-même. Une fois ceux-ci configurés, vous pouvez acheminer le trafic vers le point de terminaison du pare-feu pour permettre l’inspection.
   b. Si vous utilisez une appliance tierce avec un Gateway Load Balancer (GWLB), déployez et configurez votre appliance dans une ou plusieurs zones de disponibilité. Créez ensuite votreGWLB, le service de point de terminaison, le point de terminaison et configurez le routage de votre trafic.
3. Pour les solutions out-of-band d'inspection :
   1. Activez la mise en miroir VPC du trafic sur les interfaces où le trafic entrant et sortant doit être reflété. Vous pouvez utiliser EventBridge les règles Amazon pour appeler une AWS Lambda fonction afin d'activer la mise en miroir du trafic sur les interfaces lorsque de nouvelles ressources sont créées. Dirigez les sessions de mise en miroir du trafic vers le Network Load Balancer situé devant votre appareil qui traite le trafic.
4. Pour les solutions de trafic Web entrant :
   a. Pour configurer AWS WAF, commencez par configurer une liste de contrôle d'accès Web (WebACL). Le Web ACL est un ensemble de règles comportant une action par défaut (ALLOWouDENY) traitée en série qui définit la manière dont vous gérez le WAF trafic. Vous pouvez créer vos propres règles et groupes ou utiliser des groupes de règles AWS gérés sur votre site WebACL.
   b. Une fois votre site Web ACL configuré, associez-le à une AWS ressource (Application Load Balancer, API Gateway REST API ou CloudFront distribution, par exemple) pour commencer à protéger le trafic Web. ACL

Ressources

Documents connexes :

- Qu’est-ce que le Traffic Mirroring ?
- Mise en œuvre de l’inspection du trafic en ligne à l’aide d’appareils de sécurité tiers
- AWS Network Firewall exemples d'architectures avec routage

- Architecture d'inspection centralisée avec AWS Gateway Load Balancer et AWS Transit Gateway

Exemples connexes :

- Bonnes pratiques pour le déploiement de Gateway Load Balancer
- TLSconfiguration d'inspection pour le trafic de sortie crypté et AWS Network Firewall

Outils associés :

- AWS Marketplace IDS/IPS

SEC05-BP04 Automatiser la protection du réseau

Automatisez le déploiement des protections de votre réseau à l'aide de DevOps pratiques telles que l'infrastructure sous forme de code (IaC) et les pipelines CI/CD. Ces pratiques peuvent vous aider à suivre les modifications apportées aux protections de votre réseau via un système de contrôle de version, à réduire le temps nécessaire au déploiement des modifications et à détecter si les protections de votre réseau s'écartent de la configuration souhaitée.

Résultat souhaité : vous définissez les protections du réseau à l'aide de modèles et vous les validez dans un système de contrôle de version. Les pipelines automatisés sont lancés lorsque de nouvelles modifications sont apportées pour orchestrer les tests et le déploiement. Des vérifications des politiques et d'autres tests statiques sont en place pour valider les modifications avant le déploiement. Vous déployez les modifications dans un environnement intermédiaire afin de vérifier que les contrôles fonctionnent comme prévu. Le déploiement dans vos environnements de production est également effectué automatiquement une fois les contrôles approuvés.

Anti-modèles courants :

- Attendre de chaque équipe responsable de la charge de travail qu'elle définisse individuellement sa pile réseau complète, ses protections et ses automatisations. Ne pas publier les aspects standard de la pile réseau et des protections de manière centralisée pour que les équipes chargées de la charge de travail puissent les utiliser.
- S'appuyer sur une équipe réseau centrale pour définir tous les aspects du réseau, les protections et les automatisations. Ne pas déléguer les aspects spécifiques à la charge de travail de la pile réseau et des protections à l'équipe responsable de cette charge de travail.

- Trouver le juste équilibre entre la centralisation et la délégation entre une équipe réseau et les équipes responsables des charges de travail, sans appliquer des normes de test et de déploiement cohérentes à vos modèles IaC et à vos pipelines CI/CD. Ne pas capturer les configurations requises dans les outils qui vérifient la conformité de vos modèles.

Avantages du respect de cette bonne pratique : l'utilisation de modèles pour définir les protections de votre réseau vous permet de suivre et de comparer les modifications au fil du temps avec un système de contrôle de version. L'utilisation de l'automatisation pour tester et déployer les modifications crée de la standardisation et de la prévisibilité, ce qui augmente les chances de réussite du déploiement et réduit les configurations manuelles répétitives.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Un certain nombre de contrôles de protection réseau décrits dans SEC05-BP02 Contrôlez les flux de trafic au sein de vos couches réseau et SEC05-BP03 Mettre en œuvre une protection basée sur l'inspection s'accompagnent de systèmes de règles gérés qui peuvent être mis à jour automatiquement en fonction des dernières informations sur les menaces. Les exemples de protection de vos points de terminaison Web incluent les règles AWS WAF gérées et l'DDoSatténuation AWS Shield Advanced automatique de la couche d'application. Utilisez des groupes de règles gérées AWS Network Firewall pour vous tenir au courant des listes de domaines de mauvaise réputation et des signatures de menaces.

Au-delà des règles gérées, nous vous recommandons d'utiliser DevOps des pratiques pour automatiser le déploiement des ressources de votre réseau, des protections et des règles que vous spécifiez. Vous pouvez capturer ces définitions dans AWS CloudFormation ou dans un autre outil d'infrastructure en tant que code (IaC) de votre choix, les valider dans un système de contrôle de version et les déployer à l'aide de pipelines CI/CD. Utilisez cette approche DevOps pour bénéficier des avantages traditionnels de la gestion des contrôles de votre réseau, tels que des versions plus prévisibles, des tests automatisés à l'aide d'outils tels que AWS CloudFormation Guard, et la détection des écarts entre votre environnement déployé et la configuration souhaitée.

Sur la base des décisions que vous avez prises dans le cadre de SEC05-BP01 Create network layers, vous pouvez avoir adopté une approche de gestion centralisée pour créer VPCs des couches dédiées aux flux d'entrée, de sortie et d'inspection. Comme décrit dans l'architecture AWS de référence de sécurité (AWS SRA), vous pouvez les définir VPCs dans un compte d'infrastructure réseau dédié. Vous pouvez utiliser des techniques similaires pour définir de manière centralisée

les charges de travail VPCs utilisées dans d'autres comptes, leurs groupes de sécurité, leurs AWS Network Firewall déploiements, les règles du résolveur Route 53 et les configurations de DNS pare-feu, ainsi que d'autres ressources réseau. Vous pouvez partager ces ressources avec vos autres comptes grâce à AWS Resource Access Manager. Cette approche vous permet de simplifier les tests automatisés et le déploiement de vos contrôles réseau sur le compte Réseau, en ne gérant qu'une seule destination. Vous pouvez le faire dans un modèle hybride, dans lequel vous déployez et partagez certains contrôles de manière centralisée et déléguez d'autres contrôles aux différentes équipes responsables des charges de travail et à leurs comptes respectifs.

Étapes d'implémentation

1. Déterminez quels aspects du réseau et des protections sont définis de manière centralisée et quels aspects peuvent être gérés par vos équipes qui s'occupent des charges de travail.
2. Créez des environnements pour tester et déployer les modifications apportées à votre réseau et à ses protections. Par exemple, utilisez un compte de test réseau et un compte de production réseau.
3. Déterminez comment vous allez stocker et gérer vos modèles dans un système de contrôle de version. Stockez les modèles centraux dans un référentiel distinct des référentiels de charge de travail, tandis que les modèles de charge de travail peuvent être stockés dans des référentiels spécifiques à cette charge de travail.
4. Créez des pipelines CI/CD pour tester et déployer des modèles. Définissez des tests pour vérifier les erreurs de configuration et vérifier que les modèles sont conformes aux normes de votre entreprise.

Ressources

Bonnes pratiques associées :

- SEC01-BP06 Automatiser le déploiement des contrôles de sécurité standard

Documents connexes :

- AWS Security Reference Architecture - Network account

Exemples connexes :

- Architecture de référence des pipelines de déploiement d'AWS

- NetDevSecOpspour moderniser les déploiements AWS réseau
- Intégration des tests AWS CloudFormation de sécurité AWS Security Hub et AWS CodeBuild des rapports

Outils associés :

- AWS CloudFormation
- AWS CloudFormation Guard
- cfn_nag

SÉC 6. Comment protéger vos ressources informatiques ?

Les ressources informatiques de votre charge de travail nécessitent plusieurs couches de protection contre les menaces externes et internes. Les ressources de calcul incluent les instances EC2, les conteneurs, les fonctions AWS Lambda, les services de bases de données, les appareils IoT, et bien plus encore.

Bonnes pratiques

- SEC06-BP01 Gérer les vulnérabilités
- SEC06-BP02 Provisionner des calculs à partir d'images renforcées
- SEC06-BP03 Réduire la gestion manuelle et l'accès interactif
- SEC06-BP04 Valider l'intégrité du logiciel
- SEC06-BP05 Automatiser la protection informatique

SEC06-BP01 Gérer les vulnérabilités

Analysez et éliminez fréquemment les vulnérabilités dans votre code, vos dépendances et votre infrastructure afin de vous protéger contre les nouvelles menaces.

Résultat escompté : vous disposez d’une solution qui analyse en permanence votre charge de travail pour détecter les vulnérabilités logicielles, les défauts potentiels et l’exposition involontaire au réseau. Vous avez établi des processus et des procédures pour identifier, hiérarchiser et corriger ces vulnérabilités en fonction de critères d’évaluation des risques. En outre, vous avez mis en place une gestion automatisée des correctifs pour vos instances de calcul. Votre programme de gestion des vulnérabilités est intégré au cycle de vie de votre développement logiciel, avec des solutions pour analyser votre code source dans le cadre du pipeline CI/CD.

Anti-modèles courants :

- Absence de programme de gestion des vulnérabilités.
- Application de correctifs système sans tenir compte de la gravité ni de l'évitement des risques.
- Utilisation d'un logiciel dont la date de fin de vie (EOL) a été dépassée.
- Déploiement du code en production avant de l'analyser afin de détecter tout problème de sécurité.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

La gestion des vulnérabilités est essentielle au maintien d'un environnement cloud sécurisé et robuste. Elle implique un processus complet qui inclut des analyses de sécurité, l'identification et la hiérarchisation des problèmes, ainsi que des opérations d'application de correctifs pour résoudre les vulnérabilités identifiées. L'automatisation joue un rôle essentiel dans ce processus car elle facilite l'analyse continue des charges de travail pour détecter d'éventuels problèmes et une exposition involontaire du réseau, ainsi que les efforts de correction.

Le modèle de responsabilité partagée AWS est un concept fondamental qui sous-tend la gestion des vulnérabilités. Selon ce modèle, AWS est responsable de la sécurisation de l'infrastructure sous-jacente, y compris du matériel, des logiciels, de la mise en réseau et des installations exécutant les services AWS. À l'inverse, vous êtes responsable de la sécurisation de vos données, des configurations de sécurité et des tâches de gestion associées aux services tels que les instances Amazon EC2 et les objets Amazon S3.

AWS propose une gamme de services pour soutenir les programmes de gestion des vulnérabilités. Amazon Inspector analyse en permanence les charges de travail AWS pour détecter les vulnérabilités logicielles et les accès réseau involontaires, tandis que le Gestionnaire de correctifs d'AWS Systems Manager permet de gérer les correctifs sur l'ensemble des instances Amazon EC2. Ces services peuvent être intégrés à AWS Security Hub, un service de gestion de la posture de sécurité dans le cloud qui automatise les contrôles de sécurité AWS, centralise les alertes de sécurité et fournit une vue complète de la posture de sécurité d'une organisation. De plus, la sécurité Amazon CodeGuru utilise l'analyse du code statique pour identifier les problèmes potentiels dans les applications Java et Python pendant la phase de développement.

En incorporant les pratiques de gestion des vulnérabilités au cycle de vie du développement logiciel, vous pouvez traiter les vulnérabilités de manière proactive avant qu'elles soient introduites dans les

environnements de production, ce qui réduit le risque d'événements de sécurité et minimise l'impact potentiel des vulnérabilités.

Étapes d'implémentation

1. Comprendre le modèle de responsabilité partagée : passez en revue le modèle de responsabilité partagée AWS pour comprendre vos responsabilités en matière de sécurisation de vos charges de travail et de vos données dans le cloud. AWS est responsable de la sécurisation de l'infrastructure cloud sous-jacente, tandis que vous êtes responsable de la sécurisation de vos applications, de vos données et des services que vous utilisez.
2. Mettre en œuvre une analyse des vulnérabilités : configurez un service d'analyse des vulnérabilités, tel qu'Amazon Inspector, pour analyser automatiquement vos instances de calcul (par exemple, les machines virtuelles, les conteneurs ou les fonctions sans serveur) afin de détecter les vulnérabilités logicielles, les défauts potentiels et l'exposition involontaire du réseau.
3. Établir des processus de gestion des vulnérabilités : définissez des processus et des procédures pour identifier, hiérarchiser et corriger les vulnérabilités. Cela peut inclure la mise en place de programmes réguliers d'analyse des vulnérabilités, l'établissement de critères d'évaluation des risques et la définition de délais de correction en fonction de la gravité de la vulnérabilité.
4. Configurer la gestion des correctifs : utilisez un service de gestion des correctifs pour automatiser le processus d'application des correctifs à vos instances de calcul, tant pour les systèmes d'exploitation que pour les applications. Vous pouvez configurer le service pour rechercher les correctifs manquants dans les instances et les installer automatiquement selon un calendrier. Envisagez d'utiliser le Gestionnaire de correctifs d'AWS Systems Manager pour fournir cette fonctionnalité.
5. Configurer une protection contre les programmes malveillants : implémentez des mécanismes pour détecter les logiciels malveillants dans votre environnement. Par exemple, vous pouvez utiliser des outils tels qu'Amazon GuardDuty pour analyser, détecter et signaler les logiciels malveillants dans les volumes EC2 et EBS. GuardDuty peut également analyser les objets récemment chargés sur Amazon S3 pour détecter d'éventuels logiciels malveillants ou virus et prendre des mesures pour les isoler avant qu'ils ne soient ingérés dans les processus en aval.
6. Intégrer l'analyse des vulnérabilités dans les pipelines CI/CD : si vous utilisez un pipeline CI/CD pour le déploiement de votre application, intégrez des outils d'analyse des vulnérabilités dans votre pipeline. Des outils tels que la sécurité Amazon CodeGuru et des options open source peuvent analyser votre code source, vos dépendances et vos artefacts pour détecter d'éventuels problèmes de sécurité.

7. Configurer un service de surveillance de la sécurité : configurez un service de surveillance de la sécurité, tel qu'AWS Security Hub, pour obtenir une vue complète de votre posture de sécurité sur plusieurs services cloud. Le service doit collecter les résultats de sécurité provenant de diverses sources et les présenter dans un format normalisé pour faciliter leur hiérarchisation et leur correction.
8. Mettre en œuvre un test de pénétration des applications Web : si votre application est une application Web et que votre organisation possède les compétences nécessaires ou peut engager une assistance extérieure, envisagez de mettre en œuvre un test de pénétration des applications Web afin d'identifier les vulnérabilités potentielles de votre application.
9. Automatiser avec une infrastructure en tant que code : utilisez des outils d'infrastructure en tant que code (IaC), tels qu'AWS CloudFormation, pour automatiser le déploiement et la configuration de vos ressources, y compris les services de sécurité mentionnés précédemment. Cette pratique vous aide à créer une architecture de ressources plus cohérente et standardisée sur plusieurs comptes et environnements.
10. Surveiller et améliorer continuellement : surveillez en permanence l'efficacité de votre programme de gestion des vulnérabilités et apportez les améliorations nécessaires. Passez en revue les résultats de sécurité, évaluez l'efficacité de vos efforts de correction et ajustez vos processus et outils en conséquence.

Ressources

Documents connexes :

- AWS Systems Manager
- Présentation de la sécurité de AWS Lambda
- Amazon CodeGuru
- Gestion automatisée et améliorée des vulnérabilités pour les charges de travail dans le cloud grâce au nouvel Amazon Inspector
- Automatiser la gestion et la correction des vulnérabilités dans AWS à l'aide d'Amazon Inspector et AWS Systems Manager — Partie 1

Vidéos connexes :

- Sécurisation des services sans serveur et de conteneur
- Security best practices for the Amazon EC2 instance metadata service

SEC06-BP02 Provisionner des calculs à partir d'images renforcées

Réduisez les possibilités d’accès involontaire à vos environnements d’exécution en les déployant à partir d’images renforcées. Acquérez uniquement les dépendances d’exécution, telles que les images de conteneurs et les bibliothèques d’applications, à partir de registres fiables et vérifiez leurs signatures. Créez vos propres registres privés pour stocker des images et des bibliothèques fiables à utiliser dans vos processus de création et de déploiement.

Résultat souhaité : vos ressources de calcul sont provisionnées à partir d’images de référence renforcées. Vous récupérez les dépendances externes, telles que les images de conteneurs et les bibliothèques d’applications, uniquement à partir de registres fiables et vous vérifiez leurs signatures. Celles-ci sont stockées dans des registres privés à des fins de référence pour vos processus de création et de déploiement. Vous analysez et mettez à jour régulièrement les images et les dépendances pour vous protéger contre les vulnérabilités récemment découvertes.

Anti-modèles courants :

- Acquérir des images et des bibliothèques à partir de registres fiables, mais sans vérifier leur signature ni effectuer d’analyses de vulnérabilité avant de les utiliser.
- Renforcer les images, mais ne pas les tester régulièrement pour détecter de nouvelles vulnérabilités ou ne pas les mettre à jour vers la dernière version.
- Installer ou ne pas supprimer les packages logiciels qui ne sont pas nécessaires pendant le cycle de vie prévu de l’image.
- S’appuyer uniquement sur l’application de correctifs pour maintenir à jour les ressources de calcul de production. L’application de correctifs à elle seule peut encore entraîner une dérive des ressources de calcul par rapport à la norme renforcée au fil du temps. Il est également possible que l’application de correctifs ne parvienne pas à supprimer les programmes malveillants qui ont pu être installés par un acteur malveillant lors d’un événement de sécurité.

Avantages du respect de cette bonne pratique : le renforcement des images permet de réduire le nombre de chemins disponibles dans votre environnement d’exécution susceptibles de permettre un accès involontaire à des utilisateurs ou à des services non autorisés. Cela peut également réduire l’ampleur de l’impact en cas d’accès involontaire.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Pour renforcer vos systèmes, utilisez les dernières versions des systèmes d’exploitation, des images de conteneurs et des bibliothèques d’applications. Appliquez des correctifs aux problèmes connus. Minimisez le système en supprimant la totalité des applications, services, pilotes d’appareils, utilisateurs par défaut et autres informations d’identification inutiles. Prenez toutes les autres mesures nécessaires, telles que la désactivation des ports pour créer un environnement disposant uniquement des ressources et des fonctionnalités requises pour vos charges de travail. À partir de cette situation de référence, vous pouvez ensuite installer les logiciels, les agents ou les autres processus dont vous avez besoin à des fins telles que la surveillance de la charge de travail ou la gestion des vulnérabilités.

Vous pouvez réduire le fardeau que représente le renforcement des systèmes en utilisant les conseils fournis par des sources fiables, tels que les guides de mise en œuvre technique de sécurité du Center for Internet Security (CISDISA) et de la Defense Information Systems Agency (STIGs) ().
Nous vous recommandons de commencer par une Amazon Machine Image (AMI) publiée par un partenaire AWS ou par un APN partenaire, et d'utiliser AWS EC2Image Builder pour automatiser la configuration en fonction d'une combinaison appropriée de STIG contrôles CIS et de commandes.

Bien qu'il existe des images renforcées et des recettes EC2 Image Builder qui appliquent les CIS DISA STIG recommandations du mode opératoire, il se peut que leur configuration empêche le bon fonctionnement de votre logiciel. Dans ce cas, vous pouvez partir d'une image de base non renforcée, installer votre logiciel, puis appliquer progressivement CIS des contrôles pour tester leur impact. Pour tout CIS contrôle qui empêche l'exécution de votre logiciel, testez si vous pouvez plutôt mettre en œuvre les recommandations de renforcement plus précises. DISA Gardez une trace des différents CIS contrôles et DISA STIG configurations que vous êtes en mesure d'appliquer avec succès. Utilisez-les pour définir vos recettes de durcissement d'image dans EC2 Image Builder en conséquence.

Pour les charges de travail conteneurisées, les images renforcées de Docker sont disponibles dans le référentiel public Amazon Elastic Container Registry () ECR. Vous pouvez utiliser EC2 Image Builder pour renforcer les images des conteneurs en parallèleAMIs.

Comme pour les systèmes d'exploitation et les images de conteneurs, vous pouvez obtenir des packages de code (ou des bibliothèques) à partir de référentiels publics, via des outils tels que pip, npm, Maven et. NuGet Nous vous recommandons de gérer les packages de code en intégrant des référentiels privés, comme dans AWS CodeArtifact, à des référentiels publics fiables. Cette intégration peut gérer la récupération, le stockage et la conservation des packages up-to-date pour vous. Les processus de création de votre application peuvent ensuite obtenir et tester la dernière

version de ces packages en même temps que votre application, à l'aide de techniques telles que l'analyse de la composition logicielle (SCA), les tests statiques de sécurité des applications (SAST) et les tests dynamiques de sécurité des applications (DAST).

Pour les charges de travail sans serveur qui l'utilisent AWS Lambda, simplifiez la gestion des dépendances des packages à l'aide des couches Lambda. Utilisez des couches Lambda afin de configurer un ensemble de dépendances standard qui sont partagées entre différentes fonctions dans une archive autonome. Vous pouvez créer et gérer des couches par le biais de leur propre processus de création, ce qui permet à vos fonctions de rester centralisées up-to-date.

Étapes d’implémentation

- Renforcez les systèmes d’exploitation. Utilisez des images de base provenant de sources fiables comme base pour développer votre système renforcéAMIs. Utilisez EC2Image Builder pour personnaliser le logiciel installé sur vos images.
- Renforcez les ressources conteneurisées. Configurez les ressources conteneurisées de manière à respecter les bonnes pratiques en matière de sécurité. Lorsque vous utilisez des conteneurs, implémentez la numérisation d'ECRimages dans votre pipeline de génération et régulièrement par rapport à votre référentiel d'images pour rechercher CVEs dans vos conteneurs.
- Lorsque vous utilisez une implémentation sans serveur avec AWS Lambda, utilisez des couches Lambda pour séparer le code des fonctions de l'application et les bibliothèques dépendantes partagées. La signature de code pour Lambda permet de s’assurer que seul du code approuvé s’exécute dans vos fonctions Lambda.

Ressources

Bonnes pratiques associées :

- OPS05-BP05 Effectuer la gestion des correctifs

Vidéos connexes :

- Une plongée approfondie dans le domaine de AWS Lambda la sécurité

Exemples connexes :

- Créez rapidement une version STIG conforme à AMI l'aide d'EC2Image Builder
- Création de meilleures images de conteneurs

- Utilisation de couches Lambda pour simplifier votre processus de développement
- Développez et déployez des AWS Lambda couches à l'aide d'un framework sans serveur
- Création d'un pipeline end-to-end AWS DevSecOps CI/CD avec des outils et des logiciels open source SCA SAST DAST

SEC06-BP03 Réduire la gestion manuelle et l'accès interactif

Utilisez l’automatisation pour effectuer des tâches de déploiement, de configuration, de maintenance et d’investigation dans la mesure du possible. Envisagez l’accès manuel aux ressources de calcul en cas de procédures d’urgence ou dans des environnements sécurisés (environnement de test [sandbox]), lorsque l’automatisation n’est pas disponible.

Résultat souhaité : les scripts programmatiques et les documents d’automatisation (runbooks) capturent les actions autorisées sur vos ressources informatiques. Ces runbooks sont lancés soit automatiquement, par le biais de systèmes de détection des changements, soit manuellement, lorsque le jugement humain est requis. L’accès direct aux ressources de calcul n’est disponible qu’en cas d’urgence, lorsque l’automatisation n’est pas disponible. Toutes les activités manuelles sont enregistrées et intégrées à un processus de révision afin d’améliorer continuellement vos capacités d’automatisation.

Anti-modèles courants :

- Accès interactif aux EC2 instances Amazon avec des protocoles tels que SSH ouRDP.
- Gestion des connexions utilisateur individuelles, telles que celles des utilisateurs locaux `/etc/passwd` ou Windows.
- Partage d’un mot de passe ou d’une clé privée pour accéder à une instance entre plusieurs utilisateurs.
- Installation manuelle des logiciels et création ou mise à jour des fichiers de configuration.
- Mise à jour ou correction manuelle des logiciels.
- Connexion à une instance pour résoudre les problèmes.

Avantages du respect de cette bonne pratique : la réalisation d’actions automatisées vous aide à réduire le risque opérationnel lié à des modifications involontaires et à des erreurs de configuration. La suppression de l'utilisation de Secure Shell (SSH) et du Remote Desktop Protocol (RDP) pour l'accès interactif réduit l'étendue de l'accès à vos ressources informatiques. Cette opération supprime un chemin commun pour les actions non autorisées. La saisie de vos tâches de gestion des

ressources de calcul dans des documents d'automatisation et des scripts programmatiques fournit un mécanisme permettant de définir et d'auditer l'ensemble des activités autorisées à un niveau de détail précis.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

La connexion à une instance est une approche classique de l'administration du système. Après avoir installé le système d'exploitation du serveur, les utilisateurs se connectent généralement manuellement pour configurer le système et installer les logiciels souhaités. Pendant la durée de vie du serveur, les utilisateurs peuvent se connecter pour effectuer des mises à jour logicielles, appliquer des correctifs, modifier des configurations et résoudre des problèmes.

L'accès manuel présente toutefois un certain nombre de risques. Cela nécessite un serveur qui écoute les demandes, telles qu'un RDP service SSH ou, qui peuvent fournir un chemin potentiel vers un accès non autorisé. Cela augmente également le risque d'erreur humaine associée à l'exécution d'étapes manuelles. Ces étapes peuvent entraîner des incidents liés à la charge de travail, une corruption ou une destruction de données, ou d'autres problèmes de sécurité. L'accès humain requiert également des protections contre le partage d'informations d'identification, ce qui entraîne des frais de gestion supplémentaires.

Pour atténuer ces risques, vous pouvez implémenter une solution d'accès à distance basée sur des agents, telle que AWS Systems Manager. AWS Systems Manager L'agent (SSMagent) initie un canal crypté et ne repose donc pas sur l'écoute des demandes initiées de l'extérieur. Envisagez de configurer SSM l'agent pour établir ce canal sur un VPC point de terminaison.

Systems Manager vous permet de contrôler avec précision la manière dont vous pouvez interagir avec vos instances gérées. Vous définissez les automatisations à exécuter, qui peut les exécuter et quand elles peuvent être exécutées. Systems Manager peut appliquer des correctifs, installer des logiciels et apporter des modifications de configuration sans accès interactif à l'instance. Systems Manager peut également fournir un accès à un shell distant et enregistrer chaque commande invoquée, ainsi que sa sortie, pendant la session dans les journaux et Amazon S3. AWS CloudTrailenregistre les appels de Systems Manager à des APIs fins d'inspection.

Étapes d'implémentation

1. Installez l'agent AWS Systems Manager (SSMagent) sur vos EC2 instances Amazon. Vérifiez si l'SSMagent est inclus et démarré automatiquement dans le cadre de votre AMI configuration de base.

2. Vérifiez que les IAM rôles associés à vos profils d'EC2instance incluent la IAMpolitique `AmazonSSMManagedInstanceCore` gérée.
3. Désactivez SSH et RDP les autres services d'accès à distance exécutés sur vos instances. Vous pouvez le faire en exécutant des scripts configurés dans la section des données utilisateur de vos modèles de lancement ou en les personnalisant à l'AMIsaide d'outils tels qu'EC2Image Builder.
4. Vérifiez que les règles d'entrée du groupe de sécurité applicables à vos EC2 instances n'autorisent pas l'accès sur le port 22/tcp (SSH) ou le port 3389/tcp (). RDP Mettez en œuvre la détection et l'alerte en cas de groupes de sécurité mal configurés à l'aide de services tels que AWS Config.
5. Définissez les automatisations, les runbooks et les commandes d'exécution appropriés dans Systems Manager. Utilisez IAM des politiques pour définir qui peut effectuer ces actions et les conditions dans lesquelles elles sont autorisées. Testez minutieusement ces automatisations dans un environnement hors production. Invoquez ces automatisations si nécessaire, au lieu d'accéder à l'instance de manière interactive.
6. Utilisez AWS Systems Manager Session Manager pour fournir un accès interactif aux instances lorsque cela est nécessaire. Activez la journalisation des activités de session pour conserver une piste d'audit dans Amazon CloudWatch Logs ou Amazon S3.

Ressources

Bonnes pratiques associées :

- REL08-BP04 Déploiement à l'aide d'une infrastructure immuable

Exemples connexes :

- Remplacement de SSH l'accès pour réduire les frais de gestion et de sécurité par AWS Systems Manager

Outils associés :

- AWS Systems Manager

Vidéos connexes :

- Contrôle de l'accès des sessions utilisateur aux instances dans le gestionnaire de AWS Systems Manager session

SEC06-BP04 Valider l'intégrité du logiciel

Utilisez la vérification cryptographique pour valider l’intégrité des artefacts logiciels (y compris les images) utilisés par votre charge de travail. Signez vos logiciels de manière cryptographique afin de vous protéger contre les modifications non autorisées exécutées dans vos environnements de calcul.

Résultat souhaité : tous les artefacts proviennent de sources fiables. Les certificats des sites Web des fournisseurs sont validés. Les artefacts téléchargés sont vérifiés cryptographiquement par leur signature. Vos propres logiciels sont signés cryptographiquement et vérifiés par vos environnements informatiques.

Anti-modèles courants :

- Faire confiance aux sites Web de fournisseurs réputés pour obtenir des artefacts logiciels, mais ignorer les avis d’expiration des certificats. Procéder aux téléchargements sans confirmer la validité des certificats.
- Valider les certificats des sites Web des fournisseurs, mais sans vérifier cryptographiquement les artefacts téléchargés depuis ces sites Web.
- S’appuyer uniquement sur des condensés ou des hachages pour valider l’intégrité des logiciels. Les hachages établissent que les artefacts n’ont pas été modifiés par rapport à leur version d’origine, mais ils ne valident pas leur source.
- Ne pas signer vos propres logiciels, codes ou bibliothèques, même s’ils ne sont utilisés que dans le cadre de vos propres déploiements.

Avantages du respect de cette bonne pratique : la validation de l’intégrité des artefacts dont dépend votre charge de travail permet d’empêcher les logiciels malveillants de pénétrer dans vos environnements informatiques. La signature de vos logiciels contribue à vous protéger contre toute exécution non autorisée dans vos environnements de calcul. Sécurisez votre chaîne d’approvisionnement logicielle en signant et en vérifiant le code.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Les images du système d’exploitation, les images de conteneurs et les artefacts de code sont souvent distribués avec des contrôles d’intégrité disponibles, par exemple via un condensé ou un hachage. Cela permet aux clients de vérifier l’intégrité en calculant leur propre hachage de la charge utile et en s’assurant qu’il est identique à celui publié. Bien que ces vérifications permettent de vérifier que la charge utile n’a pas été falsifiée, elles ne permettent pas de valider que la charge

utile provient de la source d'origine (sa provenance). La vérification de la provenance nécessite un certificat délivré par une autorité de confiance pour signer numériquement l'artefact.

Si vous utilisez un logiciel téléchargé ou des artefacts dans votre charge de travail, vérifiez si le fournisseur fournit une clé publique pour la vérification des signatures numériques. Voici quelques exemples de la façon dont AWS fournit une clé publique et des instructions de vérification pour les logiciels que nous publions :

- EC2Image Builder : vérifier la signature du téléchargement de l' AWS TOEinstallation
- AWS Systems Manager: Vérification de la signature de l'SSMagent
- Amazon CloudWatch : vérification de la signature du package d' CloudWatch agent

Intégrez la vérification des signatures numériques dans les processus que vous utilisez pour obtenir et renforcer les images, comme indiqué dans SEC06-BP02 Provisionner le calcul à partir d'images renforcées.

Vous pouvez utiliser AWS Signer pour gérer la vérification des signatures, ainsi que votre propre cycle de vie de signature de code pour vos propres logiciels et artefacts. AWS Lambda et Amazon Elastic Container Registry proposent des intégrations avec Signer pour vérifier les signatures de votre code et de vos images. À l'aide des exemples de la section Ressources, vous pouvez intégrer Signer à vos pipelines d'intégration et de diffusion continues (CI/CD) afin d'automatiser la vérification des signatures et la signature de vos propres codes et images.

Ressources

Documents connexes :

- Signature cryptographique pour les conteneurs
- Meilleures pratiques pour sécuriser votre pipeline de création d'images de conteneur en utilisant AWS Signer
- Annonce de la signature d'images de conteneurs avec AWS Signer et Amazon EKS
- Configuration de la signature de code pour AWS Lambda
- Bonnes pratiques et modèles avancés pour la signature de code Lambda
- Signature de code à l'aide d' AWS Certificate Manager une autorité de certification privée et de AWS Key Management Service clés asymétriques

Exemples connexes :

- Automatisez la signature du code Lambda avec Amazon et CodeCatalyst AWS Signer
- Signature et validation des OCI artefacts avec AWS Signer

Outils associés :

- AWS Lambda
- AWS Signer
- AWS Certificate Manager
- AWS Key Management Service
- AWS CodeArtifact

SEC06-BP05 Automatiser la protection informatique

Automatisez les opérations de protection informatique afin de réduire le besoin d'intervention humaine. Utilisez l'analyse automatique pour détecter les problèmes potentiels au sein de vos ressources informatiques et y remédier grâce à des réponses programmatiques automatisées ou à des opérations de gestion de flotte. Intégrez l'automatisation à vos processus CI/CD pour déployer des charges de travail fiables avec dépendances. up-to-date

Résultat escompté : les systèmes automatisés effectuent toutes les analyses et tous les correctifs des ressources informatiques. Vous utilisez la vérification automatique pour vérifier que les images logicielles et les dépendances proviennent de sources fiables et qu'elles n'ont pas été falsifiées. Les charges de travail sont automatiquement vérifiées pour détecter up-to-date les dépendances et sont signées pour garantir la fiabilité des environnements informatiques. AWS Des mesures correctives automatisées sont lancées lorsque des ressources non conformes sont détectées.

Anti-modèles courants :

- Suivre la pratique d'une infrastructure immuable, mais ne pas avoir de solution en place pour l'application de correctifs d'urgence ou le remplacement des systèmes de production.
- Utiliser l'automatisation pour corriger les ressources mal configurées, mais ne pas mettre en place de mécanisme de remplacement manuel. Dans certains cas, vous devrez ajuster les exigences et suspendre les automatisations jusqu'à ce que vous apportiez ces modifications.

Avantages liés au respect de cette bonne pratique : l'automatisation peut réduire le risque d'accès et d'utilisation non autorisés de vos ressources informatiques. Elle contribue à éviter que les erreurs de

configuration soient transférées dans les environnements de production, et à détecter et corriger ces erreurs le cas échéant. L'automatisation facilite également la détection des accès et utilisations non autorisés des ressources de calcul afin de réduire le temps de réponse. Vous pouvez ainsi réduire la portée globale de l'impact du problème.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Vous pouvez appliquer les automatisations décrites dans les pratiques du pilier de sécurité pour protéger vos ressources de calcul. SEC06-BP01 Perform Vulnerability management décrit comment vous pouvez utiliser Amazon Inspector à la fois dans vos pipelines CI/CD et pour analyser en permanence vos environnements d'exécution à la recherche de vulnérabilités et d'expositions courantes connues (). CVEs Vous pouvez utiliser AWS Systems Manager pour appliquer des correctifs ou effectuer des redéploiements à partir d'images récentes via des runbooks automatisés afin de maintenir votre parc informatique à jour avec les derniers logiciels et bibliothèques. Utilisez ces techniques pour limiter la nécessité de mettre en place des processus manuels et des accès interactifs à vos ressources de calcul. Voir SEC06-BP03 Réduisez la gestion manuelle et l'accès interactif pour en savoir plus.

L'automatisation joue également un rôle dans le déploiement de charges de travail fiables, comme décrit dans SEC06-BP02 Provisionner le calcul à partir d'images renforcées et SEC06-BP04 Valider l'intégrité du logiciel. Vous pouvez utiliser des services tels qu'EC2Image Builder et Amazon Elastic Container Registry (ECR) pour télécharger, vérifier, créer et stocker des images et des dépendances de code renforcées et approuvées. AWS SignerAWS CodeArtifact Outre Inspector, chacun de ces éléments peut jouer un rôle dans votre processus CI/CD, de sorte que votre charge de travail n'est transmise à la production que lorsqu'il est confirmé que ses dépendances existent up-to-date et qu'elles proviennent de sources fiables. Votre charge de travail est également signée afin que les environnements de AWS calcul, tels qu'AWS LambdaAmazon Elastic Kubernetes Service EKS (), puissent vérifier qu'elle n'a pas été modifiée avant de l'autoriser à s'exécuter.

Au-delà de ces contrôles préventifs, vous pouvez également utiliser l'automatisation dans vos contrôles de détection pour vos ressources de calcul. À titre d'exemple, les AWS Security Huboffres de la norme NIST800-53 Rev. 5 qui inclut des contrôles tels que [EC2.8] les EC2 instances doivent utiliser le service de métadonnées d'instance version 2 (). IMDSv2 IMDSv2utilise les techniques d'authentification de session, de blocage des demandes contenant un X-Forwarded-For HTTP en-tête et d'un réseau TTL de 1 pour arrêter le trafic provenant de sources externes afin de récupérer des informations sur l'EC2instance. Ce check in Security Hub permet de détecter le moment où les

EC2 instances utilisent IMDSv1 et de lancer des mesures correctives automatisées. Pour en savoir plus sur la détection et les corrections automatisées, consultez le document SEC04-BP04 Initiez la correction des ressources non conformes.

Étapes d'implémentation

1. Automatisez la création sécurisée, conforme et renforcée AMIs avec EC2Image Builder. Vous pouvez produire des images intégrant des contrôles issus des standards du Center for Internet Security (CIS) ou des normes du Security Technical Implementation Guide (STIG) à partir d'images de base AWS et de APN partenaires.
2. Automatisez la gestion de la configuration. Appliquez et validez des configurations sécurisées dans vos ressources de calcul automatiquement à l'aide d'un service ou d'un outil de gestion de la configuration.
   a. Gestion de configuration automatisée à l'aide de AWS Config
   b. Gestion automatisée de la posture de sécurité et de conformité à l'aide de AWS Security Hub
3. Automatisez l'application de correctifs ou le remplacement des instances Amazon Elastic Compute Cloud (AmazonEC2). AWS Systems Manager Patch Manager automatise le processus d'application des correctifs aux instances gérées avec des mises à jour liées à la sécurité et d'autres types de mises à jour. Vous pouvez utiliser le Gestionnaire de correctifs pour appliquer des correctifs pour les systèmes d'exploitation et les applications.
   a. AWS Systems Manager Patch Manager
4. Automatisez l'analyse des ressources informatiques pour détecter les vulnérabilités et les risques courants (CVEs), et intégrez des solutions d'analyse de sécurité dans votre pipeline de création.
   a. Amazon Inspector
   b. ECRNumérisation d'images
5. Pensez à Amazon GuardDuty pour la détection automatique des malwares et des menaces afin de protéger les ressources informatiques. GuardDuty peut également identifier les problèmes potentiels lorsqu'une AWS Lambdafonction est invoquée dans votre AWS environnement.
   a. Amazon GuardDuty
6. Envisagez des solutions AWS partenaires. AWS Les partenaires proposent des produits de pointe équivalents, identiques ou intégrés aux contrôles existants dans vos environnements sur site. Ces produits complètent les services AWS existants pour vous permettre de déployer une architecture de sécurité exhaustive et une expérience plus homogène dans vos environnements cloud et sur site.
   a. Sécurité de l'infrastructure

Ressources

Bonnes pratiques associées :

- SEC01-BP06 Automatiser le déploiement des contrôles de sécurité standard

Documents connexes :

- Profitez de tous les avantages de votre infrastructure IMDSv2 et désactivez-la IMDSv1 sur l'ensemble de votre AWS infrastructure

Vidéos connexes :

- Bonnes pratiques de sécurité pour le service de métadonnées d'EC2instance Amazon

Protection des données

Questions

- SÉC 7. Comment classer vos données ?
- SÉC 8. Comment protéger vos données au repos ?
- SÉC 9. Comment protéger vos données en transit ?

SÉC 7. Comment classer vos données ?

La classification des données fournit un moyen de classer les données en fonction de leur importance et de leur sensibilité afin de vous aider à déterminer les contrôles de protection et de conservation appropriés.

Bonnes pratiques

- SEC07-BP01 Comprendre votre schéma de classification des données
- SEC07-BP02 Appliquer des contrôles de protection des données en fonction de la sensibilité des données
- SEC07-BP03 Automatiser l’identification et la classification
- SEC07-BP04 Définir la gestion évolutive du cycle de vie des données

SEC07-BP01 Comprendre votre schéma de classification des données

Comprenez la classification des données traitées par votre charge de travail, ses exigences en matière de traitement, les processus métier associés, l'endroit où les données sont stockées et qui en est le propriétaire. Votre système de classification et de traitement des données doit tenir compte des exigences légales et de conformité applicables à votre charge de travail, ainsi que des contrôles de données nécessaires. La compréhension des données est la première étape du processus de classification des données.

Résultat escompté : les types de données présents dans votre charge de travail sont bien compris et documentés. Des contrôles appropriés sont en place pour protéger les données sensibles en fonction de leur classification. Ces contrôles régissent les considérations suivantes : qui est autorisé à accéder aux données et dans quel but, où les données sont stockées, la politique de chiffrement de ces données et la manière dont les clés de chiffrement sont gérées, le cycle de vie des données et leurs exigences de conservation, les processus de destruction appropriés, les processus de sauvegarde et de restauration mis en place et l'audit de l'accès.

Anti-modèles courants :

- Ne pas établir de politique officielle de classification des données pour définir les niveaux de sensibilité des données et leurs exigences de traitement.
- Ne pas bien comprendre les niveaux de sensibilité des données de votre charge de travail et ne pas saisir ces informations dans la documentation relative à l'architecture et aux opérations.
- Ne pas appliquer les contrôles appropriés à vos données en fonction de leur sensibilité et de leurs exigences, comme indiqué dans votre politique de classification et de traitement des données.
- Ne pas fournir de rétroaction sur les exigences de classification et de traitement des données aux propriétaires des politiques.

Avantages liés au respect de cette bonne pratique : cette pratique élimine toute ambiguïté quant au traitement approprié des données dans le cadre de votre charge de travail. L'application d'une politique officielle qui définit les niveaux de sensibilité des données de votre organisation et les protections requises peut vous aider à vous conformer aux réglementations légales et aux autres attestations et certifications de cybersécurité. Les propriétaires de la charge de travail peuvent savoir en toute confiance où sont stockées les données sensibles et quels contrôles de protection sont en place. La consignation de ces informations dans la documentation aide les nouveaux membres de l'équipe à mieux les comprendre et à gérer les contrôles dès qu'ils commencent à occuper

leur fonction. Ces pratiques peuvent également contribuer à réduire les coûts en dimensionnant correctement les contrôles pour chaque type de données.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Lors de la conception d'une charge de travail, vous pouvez réfléchir aux moyens de protéger les données sensibles de manière intuitive. Par exemple, dans une application multilocataire, il est intuitif de considérer les données de chaque locataire comme sensibles et de mettre en place des protections afin qu'un locataire ne puisse pas accéder aux données d'un autre. De même, vous pouvez concevoir des contrôles d'accès intuitifs de telle sorte que seuls les administrateurs puissent modifier les données, tandis que les autres utilisateurs ne bénéficient que d'un accès en lecture, voire d'aucun accès.

En définissant et en saisissant ces niveaux de sensibilité des données dans les politiques, ainsi que leurs exigences en matière de protection des données, vous pouvez identifier formellement quelles données se trouvent dans votre charge de travail. Vous pouvez ensuite déterminer si les contrôles appropriés sont en place, s'ils peuvent être audités et quelles réponses sont pertinentes en cas de mauvaise gestion des données.

Pour mieux identifier l'emplacement de données sensibles dans votre charge de travail, envisagez d'utiliser un catalogue de données. Un catalogue de données est une base de données qui cartographie les données de votre organisation, leur emplacement, leur niveau de sensibilité et les contrôles mis en place pour protéger ces données. En outre, envisagez d'utiliser des balises de ressources le cas échéant. Par exemple, vous pouvez appliquer une balise possédant une clé de balise de `Classification` et une valeur de balise de `PHI` pour les informations de santé protégées (PHI), et une autre balise possédant une clé de balise de `Sensitivity` et une valeur de balise de `High`. Les services tels que AWS Config peuvent ensuite être utilisés pour surveiller les modifications apportées à ces ressources et vous avertir si elles font l'objet d'une modification les rendant non conformes à vos exigences de protection (telles que la modification des paramètres de chiffrement). Vous pouvez capturer la définition standard de vos clés de balise et les valeurs acceptables à l'aide des politiques de balises, une fonctionnalité d'AWS Organizations. Il est déconseillé d'avoir une clé ou une valeur de balise qui contient des données privées ou sensibles.

Étapes d'implémentation

1. Comprenez le schéma de classification des données et les exigences de protection de votre organisation.

2. Identifiez les types de données sensibles traitées par vos charges de travail.
3. Capturez les données dans un catalogue de données qui fournit une vue unique de l'emplacement des données dans l'organisation et du niveau de sensibilité de ces données.
4. Envisagez d'utiliser le balisage au niveau des ressources et des données, le cas échéant, pour baliser les données en fonction de leur niveau de sensibilité et d'autres métadonnées opérationnelles susceptibles de faciliter la surveillance et la réponse aux incidents.
   a. Les politiques de balisage d'AWS Organizations peuvent être utilisées pour appliquer les normes de balisage.

Ressources

Bonnes pratiques associées :

- SUS04-BP01 Mettre en œuvre une politique de classification des données

Documents connexes :

- Livre blanc sur la classification des données
- Bonnes pratiques en matière de balisage des ressources AWS

Exemples connexes :

- Syntaxe d'une stratégie de balise AWS Organizations et exemples

Outils associés

- AWS Tag Editor

SEC07-BP02 Appliquer des contrôles de protection des données en fonction de la sensibilité des données

Appliquez des contrôles de protection des données qui fournissent un niveau de contrôle approprié pour chaque classe de données définie dans votre politique de classification. Cette pratique peut vous permettre de protéger les données sensibles contre tout accès et toute utilisation non autorisés, tout en préservant la disponibilité et l'utilisation des données.

Résultat escompté : vous disposez d'une politique de classification qui définit les différents niveaux de sensibilité des données au sein de votre organisation. Pour chacun de ces niveaux de sensibilité, vous avez publié des directives claires concernant les services et sites de stockage et de traitement approuvés, ainsi que leur configuration requise. Vous mettez en œuvre les contrôles pour chaque niveau en fonction du niveau de protection requis et des coûts associés. Vous avez mis en place une surveillance et des alertes afin de détecter si des données sont présentes dans des emplacements non autorisés, traitées dans des environnements non autorisés, consultées par des acteurs non autorisés ou si la configuration des services associés devient non conforme.

Anti-modèles courants :

- Appliquer le même niveau de contrôles de protection à toutes les données. Cela peut entraîner un surprovisionnement des contrôles de sécurité pour les données peu sensibles ou une protection insuffisante des données hautement sensibles.
- Ne pas impliquer les parties prenantes concernées des équipes chargées de la sécurité, de la conformité et des opérations lors de la définition des contrôles de protection des données.
- Surveiller les frais opérationnels et les coûts associés à la mise en œuvre et à la maintenance des contrôles de protection des données.
- Ne pas procéder à des examens périodiques des contrôles de protection des données pour préserver l'alignement avec les politiques de classification.
- Ne pas disposer d'un inventaire complet des emplacements des données au repos et en transit.

Avantages liés au respect de cette bonne pratique : en alignant vos contrôles sur le niveau de classification de vos données, votre organisation peut investir dans des niveaux de contrôle plus élevés si nécessaire. Cela peut inclure l'augmentation des ressources consacrées à la sécurisation, à la surveillance, à la mesure, à la correction et à la production de rapports. Lorsque moins de contrôles sont nécessaires, vous pouvez améliorer l'accessibilité et l'exhaustivité des données pour votre personnel, vos clients ou vos administrés. Cette approche offre à votre organisation une grande flexibilité en matière d'utilisation des données, tout en respectant les exigences de protection des données.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

La mise en œuvre de contrôles de protection des données basés sur les niveaux de sensibilité des données implique plusieurs étapes clés. Tout d'abord, identifiez les différents niveaux de sensibilité

des données au sein de votre architecture de charge de travail (public, interne, confidentiel et restreint) et évaluez où vous stockez et traitez ces données. Définissez ensuite les limites d'isolement autour des données en fonction de leur niveau de sensibilité. Nous vous recommandons de séparer les données en différents Comptes AWS, en utilisant des politiques de contrôle des services (SCP) pour restreindre les services et les actions autorisés pour chaque niveau de sensibilité des données. Vous pouvez ainsi créer des limites d'isolement solides et appliquer le principe du moindre privilège.

Après avoir défini les limites d'isolement, implémentez les contrôles de protection appropriés en fonction des niveaux de sensibilité des données. Consultez les bonnes pratiques en matière de protection des données au repos et de protection des données en transit pour mettre en œuvre des contrôles pertinents tels que le chiffrement, les contrôles d'accès et l'audit. Envisagez des techniques telles que la création de jeton ou l'anonymisation pour réduire le niveau de sensibilité de vos données. Simplifiez l'application de politiques de données cohérentes dans l'ensemble de votre entreprise grâce à un système centralisé de création/décréation de jeton.

Surveillez et testez en permanence l'efficacité des contrôles mis en œuvre. Régulièrement, passez en revue et mettez à jour le schéma de classification des données, les évaluations des risques et les contrôles de protection à mesure que le paysage des données et les menaces de votre organisation évoluent. Alignez les contrôles de protection des données mis en œuvre avec les réglementations, normes et exigences légales pertinentes du secteur. En outre, sensibilisez et formez les employés à la sécurité pour les aider à comprendre le système de classification des données et leurs responsabilités en matière de traitement et de protection des données sensibles.

Étapes d'implémentation

1. Identifiez la classification et les niveaux de sensibilité des données au sein de votre charge de travail.
2. Définissez des limites d'isolement pour chaque niveau et déterminez une stratégie d'application.
3. Évaluez les contrôles que vous définissez pour régir l'accès, le chiffrement, l'audit, la conservation et les autres éléments requis par votre politique de classification des données.
4. Évaluez les options permettant de réduire le niveau de sensibilité des données le cas échéant, par exemple en utilisant la création de jeton ou l'anonymisation.
5. Vérifiez vos contrôles à l'aide de tests et de contrôles automatisés de vos ressources configurées.

Ressources

Bonnes pratiques associées :

- PERF03-BP01 Utiliser un magasin de données dédié le mieux adapté à vos besoins en matière de stockage des données et d'accès aux données
- COST04-BP05 Appliquer les politiques de conservation des données

Documents connexes :

- Livre blanc sur la classification des données
- Bonnes pratiques en matière de sécurité, d'identité et de conformité
- Meilleures pratiques pour AWS KMS.
- Bonnes pratiques et fonctionnalités de chiffrement pour les services AWS

Exemples connexes :

- Création d'une solution de création de jeton sans serveur pour masquer les données sensibles
- Comment utiliser la création de jeton pour améliorer la sécurité des données et réduire la portée de l'audit

Outils associés :

- AWS Key Management Service (AWS KMS)
- AWS CloudHSM
- AWS Organizations

SEC07-BP03 Automatiser l'identification et la classification

L'automatisation de l'identification et de la classification des données peut vous aider à mettre en œuvre les contrôles appropriés. L'utilisation de l'automatisation pour améliorer la détermination manuelle réduit le risque d'erreur humaine et d'exposition.

Résultat escompté : vous êtes en mesure de vérifier si les contrôles appropriés sont en place en fonction de votre politique de classification et de manutention. Les outils et services automatisés vous aident à identifier et à classer le niveau de sensibilité de vos données. L'automatisation vous aide également à surveiller en permanence vos environnements afin de détecter et d'alerter si des données sont stockées ou traitées de manière non autorisée, pour que des mesures correctives puissent être prises rapidement.

Anti-modèles courants :

- S'appuyer uniquement sur des processus manuels pour l'identification et la classification des données, ce qui peut être source d'erreur et prendre beaucoup de temps. Cela peut entraîner une classification des données inefficace et incohérente, en particulier lorsque les volumes de données augmentent.
- Ne pas disposer de mécanismes pour suivre et gérer les ressources de données dans l'ensemble de l'organisation.
- Perdre de vue la nécessité d'une surveillance et d'une classification continues des données au fur et à mesure de leur évolution au sein de l'organisation.

Avantages liés au respect de cette bonne pratique : l'automatisation de l'identification et de la classification des données peut permettre une application plus cohérente et plus précise des contrôles de protection des données, réduisant ainsi le risque d'erreur humaine. L'automatisation peut également fournir une visibilité sur l'accès et le mouvement des données sensibles, ce qui vous permet de détecter les manipulations non autorisées et de prendre des mesures correctives.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Bien que le discernement humain soit souvent utilisé pour classer les données pendant les phases initiales de conception d'une charge de travail, envisagez de mettre en place des systèmes qui automatisent l'identification et la classification des données de test à titre de contrôle préventif. Par exemple, les développeurs peuvent disposer d'un outil ou d'un service leur permettant d'analyser des données représentatives afin de déterminer leur sensibilité. Au sein d'AWS, vous pouvez télécharger des ensembles de données dans Amazon S3 et les analyser à l'aide d'Amazon Macie, Amazon Comprehend ou Amazon Comprehend Medical. De même, envisagez d'analyser les données dans le cadre de tests unitaires et d'intégration afin de détecter les endroits où des données sensibles ne sont pas attendues. Les alertes sur les données sensibles à ce stade peuvent mettre en évidence les lacunes en matière de protection avant le déploiement en production. D'autres fonctionnalités, telles que la détection des données sensibles dans AWS Glue, Amazon SNS et Amazon CloudWatch, peuvent également être utilisées pour détecter les informations personnelles et prendre des mesures d'atténuation. Pour tout outil ou service automatisé, comprenez comment il définit les données sensibles et enrichissez-le avec d'autres solutions humaines ou automatisées pour combler les lacunes si nécessaire.

À des fins de détection, utilisez une surveillance continue de vos environnements pour identifier si des données sensibles sont stockées de manière non conforme. Cela peut permettre de détecter des situations telles que l'émission de données sensibles dans des fichiers journaux ou leur copie dans un environnement d'analytique des données sans anonymisation ou suppression appropriée. Les données stockées dans Amazon S3 peuvent être surveillées en permanence afin de détecter les données sensibles à l'aide d'Amazon Macie.

Étapes d'implémentation

1. Passez en revue le schéma de classification des données au sein de votre organisation, décrit dans le document SEC07-BP01.
   a. En comprenant le schéma de classification des données de votre organisation, vous pouvez établir des processus précis d'identification et de classification automatisées conformes aux politiques de votre entreprise.
2. Effectuez une analyse initiale de vos environnements pour une identification et une classification automatisées.
   a. Une analyse initiale complète de vos données peut vous aider à comprendre de manière exhaustive où se trouvent les données sensibles dans vos environnements. Lorsqu'une analyse complète n'est pas requise au départ ou ne peut pas être réalisée en amont pour des raisons de coût, évaluez si les techniques d'échantillonnage des données sont appropriées pour obtenir vos résultats. Par exemple, Amazon Macie peut être configuré de façon à effectuer une vaste opération automatisée de découverte des données sensibles dans vos compartiments S3. Cette capacité utilise des techniques d'échantillonnage pour effectuer de manière rentable une analyse préliminaire de l'emplacement des données sensibles. Une analyse plus approfondie des compartiments S3 peut ensuite être réalisée à l'aide d'une tâche de découverte des données sensibles. D'autres magasins de données peuvent également être exportés vers S3 en vue de leur analyse par Macie.
   b. Établissez le contrôle d'accès défini dans le document SEC07-BP02 pour vos ressources de stockage de données identifiées lors de votre analyse.
3. Configurez des analyses continues de vos environnements.
   a. La capacité de découverte automatique des données sensibles de Macie peut être utilisée afin d'effectuer des analyses continues de vos environnements. Les compartiments S3 connus qui sont autorisés à stocker des données sensibles peuvent être exclus à l'aide d'une liste d'autorisation dans Macie.
4. Intégrez l'identification et la classification à vos processus de construction et de test.

   a. Identifiez les outils que les développeurs peuvent utiliser pour analyser la sensibilité des données pendant le développement des charges de travail. Utilisez ces outils dans le cadre des tests d'intégration pour vous avertir lorsque des données sensibles sont inattendues et empêcher tout déploiement ultérieur.

5. Mettez en œuvre un système ou un runbook pour agir lorsque des données sensibles sont détectées dans des emplacements non autorisés.

   a. Limitez l'accès aux données utilisant la correction automatique. Par exemple, vous pouvez déplacer ces données vers un compartiment S3 à accès restreint ou baliser l'objet si vous utilisez le contrôle d'accès par attributs (ABAC). Envisagez également de masquer les données lorsqu'elles sont détectées.

   b. Alertez vos équipes de protection des données et de réponse aux incidents pour qu'elles étudient la cause racine de l'incident. Tous les enseignements qu'elles identifient peuvent aider à prévenir de futurs incidents.

Ressources

Documents connexes :

- AWS Glue : Détecter et traiter les données sensibles
- Utilisation des identifiants de données gérés dans Amazon SNS
- Amazon CloudWatch Logs : Aider à protéger les données sensibles des journaux grâce au masquage

Exemples connexes :

- Activation de la classification des données pour la base de données Amazon RDS avec Macie
- Détection de données sensibles dans DynamoDB avec Macie

Outils associés :

- Amazon Macie
- Amazon Comprehend
- Amazon Comprehend Medical
- AWS Glue

SEC07-BP04 Définir la gestion évolutive du cycle de vie des données

Comprenez vos exigences relatives au cycle de vie des données en fonction de vos différents niveaux de classification et de traitement des données. Cela peut inclure la manière dont les données sont traitées lorsqu'elles entrent pour la première fois dans votre environnement, la manière dont les données sont transformées, ainsi que les règles relatives à leur destruction. Tenez compte de facteurs tels que les périodes de conservation, l'accès, l'audit et le suivi de la provenance.

Résultat escompté : vous classez les données le plus près possible du point et de l'heure d'ingestion. Lorsque la classification des données requiert un masquage, une création de jeton ou d'autres processus réduisant le niveau de sensibilité, vous effectuez ces actions le plus près possible du point et de l'heure de l'ingestion.

Vous supprimez les données conformément à votre politique lorsqu'il n'est plus approprié de les conserver, en fonction de leur classification.

Anti-modèles courants :

- Mettre en œuvre une approche universelle de la gestion du cycle de vie des données, sans tenir compte des différents niveaux de sensibilité et des exigences d'accès.
- Envisager la gestion du cycle de vie uniquement du point de vue des données utilisables ou des données sauvegardées, mais pas des deux.
- Supposer que les données entrées dans votre charge de travail sont valides, sans établir leur valeur ni leur provenance.
- S'appuyer sur la durabilité des données pour remplacer la sauvegarde et la protection des données.
- Conserver les données au-delà de leur utilité et de la période de conservation requise.

Avantages du respect de cette bonne pratique : une stratégie de gestion du cycle de vie des données bien définie et évolutive permet de maintenir la conformité réglementaire, d'améliorer la sécurité des données, d'optimiser les coûts de stockage et de permettre un accès et un partage efficaces des données tout en maintenant les contrôles appropriés.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les données d'une charge de travail sont souvent dynamiques. La forme qu'elles prennent lors de leur entrée dans votre environnement de charge de travail peut être différente de celle prise pour

le stockage ou l'utilisation dans la logique métier, les rapports, l'analytique ou le machine learning. De plus, la valeur des données peut évoluer au fil du temps. Certaines données sont de nature temporelle et perdent de la valeur à mesure qu'elles vieillissent. Réfléchissez à l'impact de ces modifications sur l'évaluation de vos données dans le cadre de votre système de classification des données et des contrôles associés. Dans la mesure du possible, utilisez un mécanisme de cycle de vie automatisé, tel que les stratégies de cycle de vie d'Amazon S3 et le gestionnaire de cycle de vie Amazon Data, pour configurer vos processus de conservation, d'archivage et d'expiration des données. Pour les données stockées dans DynamoDB, vous pouvez utiliser la fonctionnalité Time To Live (TTL) pour définir un horodatage d'expiration par élément.

Faites la distinction entre les données qui peuvent être utilisées et celles qui sont stockées en tant que sauvegarde. Envisagez d'utiliser AWS Backup pour automatiser la sauvegarde des données entre les services AWS. Les instantanés Amazon EBS permettent de copier un volume EBS et de le stocker à l'aide des fonctionnalités S3, notamment le cycle de vie, la protection des données et l'accès aux mécanismes de protection. Deux de ces mécanismes sont le verrouillage d'objet S3 et AWS Backup Vault Lock, qui peuvent vous apporter une sécurité et un contrôle supplémentaires sur vos sauvegardes. Gérez une séparation claire des tâches et des accès pour les sauvegardes. Isolez les sauvegardes au niveau du compte afin de préserver la séparation avec l'environnement affecté lors d'un événement.

Un autre aspect de la gestion du cycle de vie consiste à enregistrer l'historique des données au fur et à mesure de leur progression dans votre charge de travail, ce que l'on appelle le suivi de la provenance des données. Vous avez ainsi l'assurance de savoir d'où viennent les données, si des transformations ont été effectuées, quel propriétaire ou processus a appliqué ces modifications et quand. Le fait de disposer de cet historique contribue à résoudre les problèmes et à réaliser des enquêtes lors d'événements de sécurité potentiels. Par exemple, vous pouvez journaliser les métadonnées relatives aux transformations dans une table Amazon DynamoDB. Au sein d'un lac de données, vous pouvez conserver des copies des données transformées dans différents compartiments S3 pour chaque étape du pipeline de données. Stockez les informations relatives au schéma et à l'horodatage dans un AWS Glue Data Catalog. Quelle que soit la solution adoptée, tenez compte des exigences de vos utilisateurs finaux afin de déterminer l'outillage approprié dont vous avez besoin pour établir des rapports sur la provenance de vos données. Cela vous aidera à déterminer la meilleure façon de suivre la provenance des données.

Étapes d'implémentation

1. Analysez les types de données, les niveaux de sensibilité et les exigences d'accès de la charge de travail pour classer les données et définir des stratégies de gestion du cycle de vie appropriées.

2. Concevez et mettez en œuvre des politiques de conservation des données et des processus de destruction automatisés conformes aux exigences légales, réglementaires et organisationnelles.
3. Établissez des processus et une automatisation pour une surveillance, un audit et un ajustement continus des stratégies, contrôles et politiques de gestion du cycle de vie des données en fonction de l'évolution des exigences et des réglementations en matière de charge de travail.
   a. Détectez les ressources pour lesquelles la gestion automatique du cycle de vie n'est pas activée avec AWS Config.

Ressources

Bonnes pratiques associées :

- COST04-BP05 Appliquer les politiques de conservation des données
- SUS04-BP03 Utiliser des politiques pour gérer le cycle de vie de vos jeux de données

Documents connexes :

- Livre blanc sur la classification des données
- AWS Blueprint for Ransomware Defense
- Guide DevOps : Améliorer la traçabilité grâce au suivi de la provenance des données

Exemples connexes :

- Comment protéger les données sensibles pendant tout leur cycle de vie dans AWS
- Créer un lignage de données pour les lacs de données à l'aide de AWS Glue, d'Amazon Neptune et de Spline

Outils associés :

- AWS Backup
- Amazon Data Lifecycle Manager
- AWS Identity and Access Management Access Analyzer

SÉC 8. Comment protéger vos données au repos ?

Protégez vos données au repos en mettant en place plusieurs contrôles, afin de réduire le risque d'accès non autorisé ou de mauvaise gestion.

Bonnes pratiques

- SEC08-BP01 Mise en œuvre de la gestion sécurisée des clés
- SEC08-BP02 Appliquer le chiffrement au repos
- SEC08-BP03 Automatiser la protection des données au repos
- SEC08-BP04 Appliquer le contrôle d'accès

SEC08-BP01 Mise en œuvre de la gestion sécurisée des clés

La gestion sécurisée des clés inclut le stockage, la rotation, le contrôle d'accès et la surveillance des informations sur les clés nécessaires pour sécuriser les données au repos adaptées à votre charge de travail.

Résultat escompté : vous disposez d'un mécanisme de gestion de clés évolutif, reproductible et automatisé. Ce mécanisme applique un accès sur la base du moindre privilège aux éléments de clé et fournit le juste équilibre entre la disponibilité, la confidentialité et l'intégrité des clés. Vous surveillez l'accès aux clés et, si une rotation des éléments de clé est requise, vous effectuez leur rotation à l'aide d'un processus automatisé. Vous ne permettez pas à des opérateurs humains d'accéder aux éléments de clé.

Anti-modèles courants :

- Accès humain à des informations sur les clés non chiffrées.
- Création d'algorithmes cryptographiques personnalisés.
- Autorisations trop larges pour accéder aux informations sur les clés.

Avantages du respect de cette bonne pratique : en établissant un mécanisme sécurisé de gestion des clés pour votre charge de travail, vous contribuez à protéger votre contenu contre tout accès non autorisé. En outre, vous pouvez être soumis à des exigences réglementaires en matière de chiffrement de vos données. Une solution efficace de gestion des clés peut fournir des mécanismes techniques conformes à ces réglementations afin de protéger les informations sur les clés.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Le chiffrement des données au repos est un contrôle de sécurité fondamental. Pour mettre en œuvre ce contrôle, votre charge de travail a besoin d'un mécanisme permettant de stocker et de gérer en toute sécurité les éléments de clé utilisés pour chiffrer vos données au repos.

AWS propose AWS Key Management Service (AWS KMS) pour fournir un stockage durable, sécurisé et redondant pour les clés AWS KMS. De nombreux services AWS s'intègrent à AWS KMS pour prendre en charge le chiffrement de vos données. AWS KMS utilise des modules de sécurité matériels validés FIPS 140-2 niveau 3 pour protéger vos clés. Il n'existe aucun mécanisme permettant d'exporter les clés AWS KMS en texte brut.

Lorsque vous déployez des charges de travail à l'aide d'une stratégie multi-compte, vous devez conserver les clés AWS KMS dans le même compte que la charge de travail qui les utilise. Ce modèle distribué délègue la responsabilité de la gestion des clés AWS KMS à votre équipe. Dans d'autres cas d'utilisation, votre organisation peut choisir de stocker les clés AWS KMS dans un compte centralisé. Cette structure centralisée nécessite des politiques supplémentaires pour permettre l'accès intercompte requis afin que le compte de la charge de travail puisse accéder aux clés stockées dans le compte centralisé, mais elle s'applique peut-être plus aux cas d'utilisation où une seule clé est partagée entre plusieurs Comptes AWS.

Quel que soit l'endroit où les éléments de clé sont stockés, vous devez contrôler étroitement l'accès aux clés en utilisant des stratégies de clé et des politiques IAM. Les stratégies de clé constituent le principal moyen de contrôler l'accès à une clé AWS KMS. En outre, les octrois de clés AWS KMS peuvent fournir un accès aux services AWS pour chiffrer et déchiffrer les données en votre nom. Passez en revue les recommandations en matière de contrôle d'accès à vos clés AWS KMS.

Vous devez surveiller l'utilisation des clés de chiffrement afin de détecter les modèles d'accès inhabituels. Les opérations effectuées à l'aide de clés gérées par AWS et de clés gérées par le client stockées dans AWS KMS peuvent être journalisées dans AWS CloudTrail et doivent être examinées périodiquement. Portez une attention particulière à la surveillance des événements de destruction des clés. Pour limiter la destruction accidentelle ou malveillante des informations sur les clés, les événements de destruction des clés ne suppriment pas immédiatement ces informations. Les tentatives de suppression de clés dans AWS KMS sont soumises à un délai d'attente, qui est de 30 jours par défaut et de 7 jours au minimum, ce qui donne aux administrateurs le temps d'examiner ces actions et d'annuler la demande si nécessaire.

La plupart des services AWS utilisent AWS KMS de manière transparente pour vous. Vous n'avez qu'à décider si vous souhaitez utiliser une clé gérée par AWS ou une clé gérée par le client. Si votre

charge de travail nécessite l'utilisation directe de AWS KMS pour chiffrer ou déchiffrer des données, vous devez utiliser le chiffrement d'enveloppe pour protéger vos données. AWS Encryption SDK peut fournir à vos applications des primitives de chiffrement côté client pour implémenter le chiffrement d'enveloppe et l'intégrer à AWS KMS.

Étapes d'implémentation

1. Déterminez les options de gestion des clés appropriées (gérées par AWS ou gérées par le client) pour la clé.
   a. Pour faciliter l'utilisation, AWS propose des clés AWS qui appartiennent au client et des clés gérées par AWS pour la plupart des services. Elles fournissent une fonctionnalité de chiffrement au repos sans qu'il soit nécessaire de gérer les informations sur les clés ou les stratégies les concernant.
   b. Lorsque vous utilisez des clés gérées par le client, pensez au magasin de clé par défaut afin de trouver le meilleur équilibre entre agilité, sécurité, souveraineté des données et disponibilité. D'autres cas d'utilisation peuvent nécessiter l'utilisation de magasins de clés personnalisés avec AWS CloudHSM ou le magasin de clés externe.
2. Consultez la liste des services que vous utilisez pour votre charge de travail afin de comprendre comment AWS KMS s'y intègre. Par exemple, les instances EC2 peuvent utiliser des volumes EBS chiffrés. Elles vérifient ainsi que les instantanés Amazon EBS créés à partir de ces volumes sont également chiffrés à l'aide d'une clé gérée par le client et limitent la divulgation accidentelle des données instantanées non chiffrées.
   a. Comment les services AWS utilisent AWS KMS
   b. Pour plus d'informations sur les options de chiffrement proposées par un service AWS, consultez la rubrique Chiffrement au repos dans le guide de l'utilisateur ou le guide du développeur du service.
3. Mettez en œuvre AWS KMS : AWS KMS simplifie la création et la gestion des clés et le contrôle de l'utilisation du chiffrement dans un large éventail de services AWS et dans vos applications.
   a. Premiers pas : AWS Key Management Service (AWS KMS)
   b. Passez en revue les bonnes pratiques en matière de contrôle d'accès à vos clés AWS KMS.
4. Envisagez d'utiliser AWS Encryption SDK : utilisez AWS Encryption SDK avec l'intégration de AWS KMS lorsque votre application doit chiffrer des données côté client.
   a. AWS Encryption SDK
5. Activez l'Analyseur d'accès IAM pour examiner et envoyer automatiquement des notifications si les stratégies de clés AWS KMS sont trop génériques.

   a. Envisagez d'utiliser des contrôles de politique personnalisés pour vérifier qu'une mise à jour de la politique de ressources n'accorde pas un accès public aux clés KMS.

6. Activez Security Hub pour recevoir des notifications en cas de mauvaise configuration des stratégies de clés, de clés dont la suppression est prévue ou de clés dont la rotation automatique est activée.
7. Déterminez le niveau de journalisation approprié pour vos clés AWS KMS. Étant donné que les appels à AWS KMS, y compris les événements en lecture seule, sont journalisés, les journaux CloudTrail associés à AWS KMS peuvent devenir volumineux.
   a. Certaines organisations préfèrent séparer les activités de journalisation AWS KMS à un emplacement distinct. Pour plus de détails, consultez la section Journalisation des appels d'API AWS KMS avec CloudTrail du guide du développeur AWS KMS.

Ressources

Documents connexes :

- AWS Key Management Service
- Services et outils cryptographiques AWS
- Protection des données Amazon S3 à l'aide du chiffrement
- Chiffrement d'enveloppe
- Engagement de souveraineté numérique
- Démystifier les opérations de clés AWS KMS, apporter votre propre clé, magasin de clés personnalisé et portabilité du texte chiffré
- Détails cryptographiques AWS Key Management Service

Vidéos connexes :

- How Encryption Works in AWS
- Securing Your Block Storage on AWS
- AWS data protection: Using locks, keys, signatures, and certificates

Exemples connexes :

- Mettre en œuvre des mécanismes de contrôle d'accès avancés avec AWS KMS

SEC08-BP02 Appliquer le chiffrement au repos

Chiffrez les données privées au repos pour préserver leur confidentialité et offrir une couche de protection supplémentaire contre la divulgation ou l'exfiltration involontaire des données. Le chiffrement protège les données de manière à ce qu'elles ne puissent pas être lues ou consultées sans être préalablement déchiffrées. Inventoriez et contrôlez les données non chiffrées afin d'atténuer les risques associés à l'exposition des données.

Résultat escompté : vous disposez de mécanismes qui chiffrent les données privées par défaut lorsqu'elles sont au repos. Ces mécanismes contribuent à préserver la confidentialité des données et offre une couche de protection supplémentaire contre la divulgation ou l'exfiltration involontaires des données. Vous maintenez un inventaire des données non chiffrées et vous comprenez les contrôles mis en place pour les protéger.

Anti-modèles courants :

- Ne pas utiliser les configurations chiffrées par défaut.
- Fournir un accès trop permissif aux clés de déchiffrement.
- Ne pas surveiller l'utilisation des clés de chiffrement et de déchiffrement.
- Stocker des données non chiffrées.
- Utiliser la même clé de chiffrement pour toutes les données, quels que soient l'utilisation, le type et la classification des données.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Mappez les clés de chiffrement aux classifications de données dans vos charges de travail. Cette approche permet de se protéger contre un accès trop permissif lorsque vous utilisez une seule clé de chiffrement ou un très petit nombre de clés de chiffrement pour vos données (voir SEC07-BP01 Comprendre votre schéma de classification des données).

AWS Key Management Service (AWS KMS) s'intègre à de nombreux services AWS afin de faciliter le chiffrement des données au repos. Par exemple, dans Amazon Elastic Compute Cloud (Amazon EC2), vous pouvez définir un chiffrement par défaut sur les comptes pour que les nouveaux volumes EBS soient chiffrés automatiquement. Lorsque vous utilisez AWS KMS, tenez compte du degré de restriction des données. Les clés AWS KMS par défaut et contrôlées par le service sont gérées et utilisées en votre nom par AWS. Pour les données sensibles qui nécessitent un

accès précis à la clé de chiffrement sous-jacente, envisagez les clés gérées par le client (CMK). Vous disposez d'un contrôle total sur les CMK, y compris la rotation et la gestion des accès grâce à l'utilisation de stratégies de clés.

En outre, des services tels qu'Amazon Simple Storage Service (Amazon S3) chiffrent désormais tous les nouveaux objets par défaut. Cette implémentation offre une sécurité renforcée sans aucun impact sur les performances.

D'autres services, comme Amazon Elastic Compute Cloud (Amazon EC2) ou Amazon Elastic File System (Amazon EFS), prennent en charge les paramètres de chiffrement par défaut. Vous pouvez également utiliser AWS Config Rules pour vérifier automatiquement que vous utilisez le chiffrement pour les volumes Amazon Elastic Block Store (Amazon EBS), les instances Amazon Relational Database Service (Amazon RDS), les compartiments Amazon S3 et d'autres services au sein de votre organisation.

AWS fournit également des options de chiffrement côté client, ce qui vous permet de chiffrer les données avant de les télécharger dans le cloud. AWS Encryption SDK fournit un moyen de chiffrer vos données à l'aide du chiffrement d'enveloppe. Vous fournissez la clé de wrapping et AWS Encryption SDK génère une clé de données unique pour chaque objet de données qu'il chiffre. Envisagez AWS CloudHSM si vous avez besoin d'un module de sécurité du matériel géré à un seul locataire (HSM). AWS CloudHSM vous permet de générer, d'importer et de gérer des clés de chiffrement sur un HSM validé FIPS 140-2 de niveau 3. Certains cas d'utilisation pour AWS CloudHSM incluent la protection des clés privées pour l'émission d'une autorité de certification (CA) et le chiffrement transparent des données (TDE) pour les bases de données Oracle. Le kit SDK client AWS CloudHSM fournit un logiciel qui vous permet de chiffrer des données côté client à l'aide de clés stockées dans AWS CloudHSM avant de télécharger vos données dans AWS. Le client de chiffrement Amazon DynamoDB vous permet également de chiffrer et de signer les éléments avant de les télécharger dans une table DynamoDB.

Étapes d'implémentation

- Configurer le chiffrement par défaut pour les nouveaux volumes Amazon EBS : indiquez que vous souhaitez que tous les nouveaux volumes Amazon EBS soient créés sous forme chiffrée, avec la possibilité d'utiliser la clé par défaut fournie par AWS ou une clé que vous créez.
- Configurer des Amazon Machine Images (AMI) chiffrées : la copie d'une AMI existante avec le chiffrement configuré chiffrera automatiquement les volumes racine et les instantanés.
- Configurer le chiffrement Amazon RDS : configurez le chiffrement pour vos clusters de base de données Amazon RDS et vos instantanés au repos en activant l'option de chiffrement.

- Créer et configurer des clés AWS KMS avec des stratégies qui limitent l'accès aux principaux appropriés pour chaque classification de données : par exemple, créez une clé AWS KMS pour chiffrer les données de production et une clé différente pour chiffrer les données de développement ou de test. Vous pouvez également fournir un accès de clé à d'autres Comptes AWS. Envisagez d'avoir différents comptes pour vos environnements de développement et de production. Si votre environnement de production a besoin de déchiffrer des artefacts dans le compte de développement, vous pouvez modifier la politique de CMK utilisée pour chiffrer les artefacts de développement afin de permettre au compte de production de déchiffrer ces artefacts. L'environnement de production peut ensuite ingérer les données déchiffrées afin de les utiliser en production.
- Configurer le chiffrement dans des services AWS supplémentaires : pour les autres services AWS que vous utilisez, passez en revue la documentation de sécurité de ce service afin d'en déterminer les options de chiffrement.

Ressources

Documents connexes :

- Outils de chiffrement AWS
- AWS Encryption SDK
- Livre blanc sur les informations cryptographiques AWS KMS
- AWS Key Management Service
- Services et outils cryptographiques AWS
- Chiffrement Amazon EBS
- Chiffrement par défaut pour les volumes Amazon EBS
- Chiffrement des ressources Amazon RDS
- Comment activer le chiffrement par défaut pour un compartiment Amazon S3 ?
- Protection des données Amazon S3 à l'aide du chiffrement

Vidéos connexes :

- How Encryption Works in AWS
- Securing Your Block Storage on AWS

SEC08-BP03 Automatiser la protection des données au repos

Utilisez l'automatisation pour valider et appliquer les contrôles des données au repos. Utilisez l'analyse automatique pour détecter les erreurs de configuration de vos solutions de stockage de données et effectuez des corrections par le biais d'une réponse programmatique automatisée dans la mesure du possible. Intégrez l'automatisation à vos processus de CI/CD afin de détecter les erreurs de configuration du stockage de données avant leur déploiement en production.

Résultat escompté : les systèmes automatisés analysent et surveillent les emplacements de stockage de données pour détecter les erreurs de configuration des commandes, les accès non autorisés et les utilisations inattendues. La détection d'emplacements de stockage mal configurés déclenche des mesures correctives automatisées. Les processus automatisés créent des sauvegardes de données et stockent des copies immuables en dehors de l'environnement d'origine.

Anti-modèles courants :

- Ne pas prendre en compte les options permettant d'activer des paramètres de chiffrement par défaut, lorsque le chiffrement est pris en charge.
- Ne pas prendre en compte les événements de sécurité, en plus des événements opérationnels, lors de la formulation d'une stratégie de sauvegarde et de restauration automatisée.
- Ne pas appliquer les paramètres d'accès public pour les services de stockage.
- Ne pas surveiller ni auditer vos contrôles pour protéger les données au repos.

Avantages du respect de cette bonne pratique : l'automatisation permet de prévenir le risque de mauvaise configuration de vos emplacements de stockage de données. Cela permet d'éviter que des erreurs de configuration ne pénètrent dans vos environnements de production. Grâce à cette bonne pratique, vous pouvez également détecter et corriger les erreurs de configuration, le cas échéant.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

L'automatisation est un thème récurrent dans les pratiques de protection de vos données au repos. SEC01-BP06 Automatiser le déploiement de contrôles de sécurité standard décrit comment vous pouvez capturer la configuration de vos ressources à l'aide de modèles d'infrastructure en tant que code (IaC), tels que AWS CloudFormation. Ces modèles sont validés dans un système de contrôle de version et sont utilisés pour déployer des ressources sur AWS via un pipeline CI/CD. Ces techniques s'appliquent également à l'automatisation de la configuration de vos solutions de stockage de données, telles que les paramètres de chiffrement des compartiments Amazon S3.

Vous pouvez vérifier si les paramètres que vous définissez dans les modèles IaC ont été configurés correctement dans vos pipelines CI/CD à l'aide de règles dans AWS CloudFormation Guard. Vous pouvez surveiller les paramètres qui ne sont pas encore disponibles dans CloudFormation ou dans d'autres outils IaC pour détecter toute mauvaise configuration avec AWS Config. Les alertes générées par Config en cas d'erreur de configuration peuvent être corrigées automatiquement, comme décrit dans SEC04-BP04 Initier les mesures de correction pour les ressources non conformes.

L'utilisation de l'automatisation dans le cadre de votre stratégie de gestion des autorisations fait également partie intégrante des protections automatisées des données. SEC03-BP02 Accorder l'accès au moindre privilège et SEC03-BP04 Réduire les autorisations en continu décrivent la configuration de stratégies d'accès au moindre privilège qui sont surveillées en permanence par AWS Identity and Access Management Access Analyzer pour générer des résultats lorsque les autorisations peuvent être réduites. Au-delà de l'automatisation des autorisations de surveillance, vous pouvez configurer Amazon GuardDuty pour détecter tout comportement anormal d'accès aux données pour vos volumes EBS (via une instance EC2), vos compartiments S3 et les bases de données Amazon Relational Database Service prises en charge.

L'automatisation joue également un rôle dans la détection des données sensibles stockées dans des emplacements non autorisés. SEC07-BP03 Automatiser l'identification et la classification décrit comment Amazon Macie peut surveiller vos compartiments S3 pour détecter les données sensibles inattendues et générer des alertes susceptibles de déclencher une réponse automatique.

Suivez les pratiques décrites dans REL09 Données de sauvegarde pour développer une stratégie automatisée de sauvegarde et de restauration des données. La sauvegarde et la restauration des données sont aussi importantes pour la restauration après des événements de sécurité que pour des événements opérationnels.

Étapes d'implémentation

1. Capturez la configuration du stockage de données dans des modèles IaC. Utilisez des contrôles automatisés dans vos pipelines CI/CD pour détecter les erreurs de configuration.
   a. Vous pouvez utiliser vos modèles d'infrastructure en tant que code pour AWS CloudFormation et utiliser AWS CloudFormation Guard pour vérifier que les modèles sont bien configurés.
   b. Utilisez AWS Config pour exécuter des règles dans un mode d'évaluation proactif. Utilisez ce paramètre pour vérifier la conformité d'une ressource en tant qu'étape de votre pipeline CI/CD avant de la créer.
2. Surveillez les ressources pour détecter les erreurs de configuration du stockage de données.

   a. Paramétrez AWS Config pour qu'il surveille les ressources de stockage de données afin de détecter les modifications apportées aux configurations de contrôle et pour générer des alertes afin d'invoquer des mesures correctives lorsqu'il détecte une mauvaise configuration.

   b. Consultez SEC04-BP04 Lancer la correction des ressources non conformes pour plus de conseils sur les corrections automatisées.

3. Surveillez et réduisez continuellement les autorisations d'accès aux données grâce à l'automatisation.

   a. IAM Access Analyzer peut fonctionner en continu pour générer des alertes lorsque les autorisations sont susceptibles d'être réduites.

4. Surveillez et signalez les comportements anormaux en matière d'accès aux données.

   a. GuardDuty surveille à la fois les signatures de menaces connues et les écarts par rapport aux comportements d'accès de base pour les ressources de stockage de données telles que les volumes EBS, les compartiments S3 et les bases de données RDS.

5. Surveillez les données sensibles et donnez l'alerte si certaines d'entre elles sont stockées dans des endroits inattendus.

   a. Utilisez Amazon Macie pour analyser en permanence vos compartiments S3 à la recherche de données sensibles.

6. Automatisez les sauvegardes sécurisées et chiffrées de vos données.

   a. AWS Backup est un service géré qui crée des sauvegardes chiffrées et sécurisées de diverses sources de données sur AWS. Elastic Disaster Recovery vous permet de copier les charges de travail complètes du serveur et de maintenir une protection continue des données avec un objectif de point de reprise (RPO) mesuré en secondes. Vous pouvez configurer les deux services de façon à ce qu'ils fonctionnent ensemble pour automatiser la création de sauvegardes de données et leur copie vers des emplacements de basculement. Vous pouvez ainsi garantir la disponibilité de vos données lorsqu'elles sont touchées par des événements opérationnels ou de sécurité.

Ressources

Bonnes pratiques associées :

- SEC01-BP06 Automatiser le déploiement des contrôles de sécurité standard
- SEC03-BP02 Accorder un accès selon le principe du moindre privilège
- SEC03-BP04 Limiter les autorisations au minimum requis en permanence

- SEC04-BP04 Lancer la correction pour les ressources non conformes
- SEC07-BP03 Automatiser l'identification et la classification
- REL09-BP02 Sécuriser et chiffrer les sauvegardes
- REL09-BP03 Effectuer automatiquement la sauvegarde des données

Documents connexes :

- Conseils prescriptifs AWS : chiffrer automatiquement les volumes Amazon EBS existants et nouveaux
- Gestion des risques liés aux rançongiciels sur AWS à l'aide du NIST Cybersecurity Framework (CSF)

Exemples connexes :

- Comment utiliser des règles proactives AWS Config et des hooks AWS CloudFormation pour empêcher la création de ressources cloud non conformes
- Automatiser et gérer de manière centralisée la protection des données pour Amazon S3 avec AWS Backup
- AWS re:Invent 2023 - Implement proactive data protection using Amazon EBS snapshots
- AWS re:Invent 2022 - Build and automate for resilience with modern data protection

Outils associés :

- AWS CloudFormation Guard
- Registre des règles AWS CloudFormation Guard
- IAM Access Analyzer
- Amazon Macie
- AWS Backup
- Reprise après sinistre Elastic

SEC08-BP04 Appliquer le contrôle d’accès

Pour vous aider à protéger vos données au repos, appliquez le contrôle d’accès à l’aide de mécanismes tels que l’isolement et la gestion des versions. Appliquez les contrôles d’accès conditionnel et de moindre privilège. Empêchez l’octroi d’un accès public à vos données.

Résultat escompté : vous vérifiez que seuls les utilisateurs autorisés peuvent accéder aux données lorsqu’ils en ont besoin. Vous protégez vos données avec des sauvegardes régulières et la gestion des versions pour éviter toute modification ou suppression intentionnelle ou involontaire des données. Vous isolez les données critiques des autres données afin de protéger leur confidentialité et leur intégrité.

Anti-modèles courants :

- Stocker ensemble des données ayant différentes exigences en termes de sensibilité ou de classification.
- Utiliser des autorisations trop permissives sur les clés de déchiffrement.
- Classer les données de façon incorrecte.
- Ne pas conserver les sauvegardes détaillées des données importantes.
- Fournir un accès permanent aux données de production.
- Ne pas auditer l’accès aux données ni examiner régulièrement les autorisations.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Il est important de protéger les données au repos pour préserver leur intégrité, leur confidentialité et leur conformité aux exigences réglementaires. Vous pouvez mettre en œuvre plusieurs contrôles pour y parvenir, notamment le contrôle d’accès, l’isolation, l’accès conditionnel et la gestion des versions.

Vous pouvez appliquer le contrôle d’accès selon le principe du moindre privilège, qui fournit uniquement les autorisations nécessaires aux utilisateurs et aux services pour qu’ils effectuent leurs tâches. Cela inclut l’accès aux clés de chiffrement. Passez en revue vos politiques AWS Key Management Service (AWS KMS) pour vous assurer que le niveau d’accès que vous accordez est approprié et que les conditions appropriées s’appliquent.

Vous pouvez séparer les données en fonction de différents niveaux de classification en utilisant des Comptes AWS distincts pour chaque niveau, et gérer ces comptes à l’aide d’AWS Organizations.

Cette isolation contribue à empêcher tout accès non autorisé et minimise le risque d'exposition des données.

Examinez régulièrement le niveau d'accès accordé dans les politiques de compartiment Amazon S3. Évitez d'utiliser des compartiments publiquement accessibles en lecture ou en écriture à moins que cela ne soit absolument nécessaire. Envisagez d'utiliser AWS Config pour détecter les compartiments publiquement disponibles et Amazon CloudFront pour diffuser du contenu à partir d'Amazon S3. Vérifiez que les compartiments qui ne doivent pas autoriser l'accès public sont configurés correctement pour l'empêcher.

Mettez en œuvre des mécanismes de gestion des versions et de verrouillage d'objets pour les données critiques stockées dans Amazon S3. La gestion des versions d'Amazon S3 préserve les versions précédentes des objets pour permettre de récupérer les données en cas de suppression ou de remplacement accidentels. Le verrouillage d'objet Amazon S3 fournit un contrôle d'accès obligatoire pour les objets, ce qui empêche leur suppression ou leur remplacement, même par l'utilisateur racine, jusqu'à l'expiration du verrou. En outre, le verrouillage de coffre Amazon S3 Glacier propose une fonctionnalité similaire pour les archives stockées dans Amazon S3 Glacier.

Étapes d'implémentation

1. Appliquez un contrôle d'accès selon le principe du moindre privilège :
   - Passez en revue les autorisations d'accès accordées aux utilisateurs et aux services, et vérifiez que ces derniers ne disposent que des autorisations nécessaires à l'exécution de leurs tâches.
   - Passez en revue l'accès aux clés de chiffrement en vérifiant les politiques AWS Key Management Service (AWS KMS).
2. Séparez les données en fonction des différents niveaux de classification :
   - Utilisez des Comptes AWS distincts pour chaque niveau de classification des données.
   - Gérez ces comptes avec AWS Organizations.
3. Passez en revue les autorisations relatives aux objets et aux compartiments Amazon S3 :
   - Examinez régulièrement le niveau d'accès accordé dans les politiques de compartiment Amazon S3.
   - Évitez d'utiliser des compartiments publiquement accessibles en lecture ou en écriture à moins que cela ne soit absolument nécessaire.
   - Envisagez d'utiliser AWS Config pour détecter les compartiments disponibles publiquement.
   - Utilisez Amazon CloudFront pour diffuser du contenu à partir d'Amazon S3.

- Vérifiez que les compartiments qui ne doivent pas autoriser l'accès public sont configurés correctement pour l'empêcher.
- Vous pouvez appliquer le même processus de révision aux bases de données et à toutes les autres sources de données qui utilisent l'authentification IAM, telles que SQS ou les entrepôts de données tiers.

4. Utilisez l'Analyseur d'accès AWS IAM :
   - Vous pouvez configurer l'Analyseur d'accès AWS IAM pour analyser des compartiments Amazon S3 et générer des résultats lorsqu'une politique S3 accorde l'accès à une entité externe.
5. Implémentez des mécanismes de gestion des versions et de verrouillage d'objet :
   - Utilisez la gestion des versions d'Amazon S3 pour préserver les versions précédentes des objets et permettre de récupérer les données en cas de suppression ou de remplacement accidentels.
   - Utilisez le verrouillage d'objet Amazon S3 pour fournir un contrôle d'accès obligatoire pour les objets, ce qui empêche leur suppression ou leur remplacement, même par l'utilisateur racine, jusqu'à l'expiration du verrou.
   - Utilisez le verrouillage de coffre Amazon S3 Glacier pour les archives stockées dans Amazon S3 Glacier.
6. Utilisez l'inventaire Amazon S3 :
   - Vous pouvez utiliser l'inventaire Amazon S3 pour auditer et signaler le statut de réplication et de chiffrement de vos objets S3.
7. Vérifiez les autorisations de partage Amazon EBS et d'AMI :
   - Passez en revue vos autorisations de partage pour Amazon EBS et le partage d'AMI afin de vous assurer que vos images et volumes ne sont pas partagés avec des Comptes AWS en dehors de votre charge de travail.
8. Passez régulièrement en revue les partages d'AWS Resource Access Manager :
   - Vous pouvez utiliser AWS Resource Access Manager pour partager des ressources, telles que des politiques AWS Network Firewall, des règles d'Amazon Route 53 Resolver et des sous-réseaux, au sein de vos VPC Amazon.
   - Auditez régulièrement les ressources partagées et cessez de partager les ressources qui n'ont plus besoin de l'être.

Ressources

Bonnes pratiques associées :

- SEC03-BP01 Définir les conditions d’accès
- SEC03-BP02 Accorder un accès selon le principe du moindre privilège

Documents connexes :

- Livre blanc sur les informations cryptographiques AWS KMS
- Introduction à la gestion des autorisations d’accès à vos ressources Amazon S3
- Présentation de la gestion de l’accès à vos ressources AWS KMS
- AWS Config Rules
- Amazon S3 + Amazon CloudFront : une combinaison parfaite dans le cloud
- Utilisation de la gestion des versions
- Verrouillage d’objets avec la fonctionnalité de verrouillage d’objet Amazon S3
- Partager un instantané Amazon EBS
- AMI partagées
- Hébergement d’une application d’une seule page sur Amazon S3
- Clés de condition globale AWS
- Créération d’un périmètre des données sur AWS

Vidéos connexes :

- Securing Your Block Storage on AWS

SÉC 9. Comment protéger vos données en transit ?

Protégez vos données en transit en mettant en place plusieurs contrôles, afin de réduire le risque d’accès non autorisé ou de perte.

Bonnes pratiques

- SEC09-BP01 Implémenter la gestion sécurisée des clés et des certificats
- SEC09-BP02 Application du chiffrement en transit
- SEC09-BP03 Authentifier les communications réseau

SEC09-BP01 Implémenter la gestion sécurisée des clés et des certificats

Les certificats du protocole TLS (Transport Layer Security) permettent de sécuriser les communications réseau et établir l'identité des sites Web, des ressources et des charges de travail sur Internet, ainsi que sur les réseaux privés.

Résultat escompté : un système de gestion des certificats sécurisé qui peut provisionner, déployer, stocker et renouveler des certificats dans une infrastructure à clé publique (PKI). Un mécanisme sécurisé de gestion des clés et des certificats empêche la divulgation de la clé privée du certificat et renouvelle automatiquement et périodiquement le certificat. Il s'intègre également à d'autres services pour fournir des communications réseau et une identité sécurisées pour les ressources de la machine au sein de votre charge de travail. Les clés ne doivent jamais être accessibles aux identités humaines.

Anti-modèles courants :

- Exécuter des étapes manuelles au cours des processus de déploiement ou de renouvellement des certificats.
- Ne pas accorder suffisamment d'attention à la hiérarchie de l'autorité de certification (AC) lors de la conception d'une AC privée.
- Utiliser des certificats auto-signés pour les ressources publiques.

Avantages liés au respect de cette bonne pratique :

- Simplifiez la gestion des certificats en automatisant leur déploiement et leur renouvellement
- Encouragez le chiffrage des données en transit à l'aide de certificats TLS
- Amélioration de la sécurité et de l'auditabilité des actions de certification entreprises par l'autorité de certification
- Organisation des tâches de gestion à différents niveaux de la hiérarchie de l'AC

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les charges de travail modernes font un usage intensif des communications réseau chiffrées à l'aide de protocoles PKI tels que le protocole TLS. La gestion des certificats PKI peut être complexe, mais le provisionnement, le déploiement et le renouvellement automatisés des certificats peuvent réduire les inconvénients liés à la gestion des certificats.

AWS fournit deux services pour gérer les certificats PKI à usage général : AWS Certificate Manager et AWS Private Certificate Authority (AWS Private CA). ACM est le principal service que les clients utilisent pour provisionner, gérer et déployer des certificats destinés à être utilisés dans des charges de travail AWS publiques et privées. ACM émet des certificats privés en utilisant AWS Private CA et s'intègre à de nombreux autres services gérés par AWS pour fournir des certificats TLS sécurisés pour les charges de travail. ACM peut également délivrer des certificats reconnus publiquement à partir d'Amazon Trust Services. Les certificats publics d'ACM peuvent être utilisés sur des charges de travail destinées au public, car les navigateurs et les systèmes d'exploitation modernes approuvent par défaut ces certificats.

AWS Private CA vous permet d'établir votre propre autorité de certification racine ou subordonnée et d'émettre des certificats TLS par l'intermédiaire d'une API. Vous pouvez utiliser ce type de certificats dans des scénarios où vous contrôlez et gérez la chaîne de confiance du côté client de la connexion TLS. En plus des cas d'utilisation TLS, AWS Private CA peut émettre des certificats à des pods Kubernetes, des attestations produits pour appareils Matter, une signature de code et d'autres cas d'utilisation avec un modèle personnalisé. Vous pouvez également utiliser Rôles Anywhere IAM pour fournir des informations d'identification IAM temporaires aux charges de travail sur site qui ont reçu des certificats X.509 signés par votre autorité de certification privée.

Outre ACM et AWS Private CA, AWS IoT Core fournit un support spécialisé pour le provisionnement, la gestion et le déploiement de certificats PKI sur les appareils IoT. AWS IoT Core fournit des mécanismes spécialisés pour intégrer des appareils IoT dans votre infrastructure à clé publique à grande échelle.

Certains services AWS, tels qu'Amazon API Gateway et Elastic Load Balancing, proposent leurs propres fonctionnalités d'utilisation de certificats pour sécuriser les connexions des applications. Par exemple, API Gateway et Application Load Balancer (ALB) prennent en charge le protocole TLS mutuel (mTLS) à l'aide de certificats client que vous créez et exportez à l'aide de la AWS Management Console, de CLI ou des API.

Considérations relatives à l'établissement d'une hiérarchie d'autorité de certification privée

Lorsque vous devez établir une autorité de certification privée, il est important de prendre soin de concevoir correctement la hiérarchie de l'autorité de certification dès le départ. La bonne pratique consiste à déployer chaque niveau de votre hiérarchie d'autorité de certification dans des Comptes AWS distincts lorsque vous créez une hiérarchie d'autorité de certification privée. Cette étape intentionnelle réduit la surface de chaque niveau de la hiérarchie de l'autorité de certification, ce qui facilite la découverte d'anomalies dans les données de journalisation CloudTrail et réduit

l'étendue de l'accès ou l'impact en cas d'accès non autorisé à l'un des comptes. L'autorité de certification racine doit résider dans son propre compte et ne doit être utilisée que pour émettre un ou plusieurs certificats d'autorité de certification intermédiaire.

Créez ensuite une ou plusieurs autorités de certification intermédiaires dans des comptes distincts du compte de l'autorité de certification racine afin d'émettre des certificats pour les utilisateurs finaux, les appareils ou d'autres charges de travail. Enfin, émettez des certificats à partir de votre autorité de certification racine vers les autorités de certification intermédiaires, qui émettront à leur tour des certificats vers vos utilisateurs finaux ou vos appareils. Pour plus d'informations sur la planification du déploiement des AC et la conception de la hiérarchie des AC, y compris la planification de la résilience, la réplication interrégionale, le partage des AC au sein de votre organisation et plus encore, voir Planifier votre déploiement AWS Private CA.

Étapes d'implémentation

1. Déterminez les services AWS pertinents requis pour votre cas d'utilisation :
   - De nombreux cas d'utilisation peuvent s'appuyer sur l'infrastructure de clés publiques existante d'AWS à l'aide d'AWS Certificate Manager. ACM peut déployer des certificats TLS pour les serveurs Web, les équilibreurs de charge ou d'autres utilisations pour des certificats publiquement approuvés.
   - Envisagez AWS Private CA si vous devez établir votre propre hiérarchie d'autorité de certification privée ou si vous avez besoin d'accéder à des certificats exportables. ACM peut ensuite être utilisé pour émettre de nombreux types de certificats d'entité finale à l'aide du AWS Private CA.
   - Pour les cas d'utilisation où les certificats doivent être provisionnés à grande échelle pour les appareils de l'Internet des objets (IoT) embarqués, envisagez AWS IoT Core.
   - Envisagez d'utiliser les fonctionnalités mTLS natives dans des services tels qu'Amazon API Gateway ou Application Load Balancer.
2. Mettez en œuvre le renouvellement automatisé des certificats dans la mesure du possible :
   - Utilisez le renouvellement géré par ACM pour les certificats émis par ACM, ainsi que les services intégrés gérés par AWS.
3. Établissez des journaux et des pistes d'audit :
   - Activez les journaux CloudTrail pour suivre l'accès aux comptes détenant des autorités de certification. Envisagez de configurer la validation de l'intégrité des fichiers journaux dans CloudTrail pour vérifier l'authenticité des données du journal.

- Vous pouvez générer des rapports d'audit qui répertorient les certificats émis et révoqués par votre autorité de certification privée. Ces rapports peuvent être exportés vers un compartiment S3.
- Lors du déploiement d'une autorité de certification privée, vous devrez également créer un compartiment S3 pour stocker la liste de révocation des certificats (CRL). Pour obtenir des conseils sur la configuration de ce compartiment S3 en fonction des exigences de votre charge de travail, voir Planification d'une liste de révocation de certificats (CRL).

Ressources

Bonnes pratiques associées :

- SEC02-BP02 Utiliser des informations d'identification temporaires
- SEC08-BP01 Mise en œuvre de la gestion sécurisée des clés
- SEC09-BP03 Authentifier les communications réseau

Documents connexes :

- Comment héberger et gérer une infrastructure complète de certificats privés dans AWS
- Comment garantir une hiérarchie d'autorités de certification privées ACM à l'échelle de l'entreprise pour l'automobile et la fabrication
- Bonnes pratiques en matière d'AC privée
- Comment utiliser AWS RAM pour partager votre compte croisé Private CA

Vidéos connexes :

- Activer Private CA AWS Certificate Manager (atelier)

Exemples connexes :

- Atelier sur les autorités de certification privées
- Atelier sur la gestion des appareils IoT (y compris le provisionnement des appareils)

Outils associés :

- Plugin pour Kubernetes cert-manager pour utiliser AWS Private CA

SEC09-BP02 Application du chiffrement en transit

Appliquez vos exigences de chiffrement définies en fonction des politiques, des obligations réglementaires et des normes de votre entreprise afin de répondre aux exigences organisationnelles, juridiques et de conformité. Utilisez uniquement les protocoles avec chiffrement lors de la transmission de données sensibles en dehors de votre cloud privé virtuel (VPC). Le chiffrement permet de préserver la confidentialité des données, même lorsque celles-ci transitent par des réseaux non fiables.

Résultat escompté : vous chiffrez le trafic réseau entre vos ressources et Internet afin de limiter l'accès non autorisé aux données. Vous chiffrez le trafic réseau au sein de votre environnement AWS interne en fonction de vos exigences de sécurité. Vous chiffrez les données en transit à l'aide des protocoles TLS sécurisés et de suites de chiffrement.

Anti-modèles courants :

- Utiliser des versions obsolètes de composants SSL, TLS et de suite de chiffrement (par exemple, SSL v3.0, clés RSA 1024 bits et chiffrement RC4).
- Autoriser le trafic non chiffré (HTTP) vers ou depuis des ressources publiques.
- Ne pas surveiller et ne pas remplacer les certificats X.509 avant leur expiration.
- Utiliser des certificats X.509 auto-signés pour TLS.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les services AWS fournissent des points de terminaison HTTPS utilisant TLS pour la communication, ce qui assure le chiffrement en transit lors de la communication avec les API AWS. Les protocoles HTTP non sécurisés peuvent être audités et bloqués dans un cloud privé virtuel (VPC) dans le cadre de l'utilisation de groupes de sécurité. Les requêtes HTTP peuvent être également redirigées automatiquement vers HTTPS dans Amazon CloudFront ou sur un Application Load Balancer. Vous pouvez utiliser une politique de compartiment Amazon Simple Storage Service (Amazon S3) pour restreindre la possibilité de charger des objets via HTTP, en imposant l'utilisation du protocole HTTPS pour les chargements d'objets vers votre ou vos compartiments. Vous disposez d'un contrôle total sur vos ressources de calcul pour mettre en œuvre le chiffrement en transit dans l'ensemble

de vos services. De plus, vous pouvez utiliser la connectivité VPN dans votre VPC à partir d'un réseau externe ou d'AWS Direct Connect pour faciliter le chiffrement du trafic. Vérifiez que vos clients appellent les API AWS en utilisant au moins le protocole TLS 1.2, car AWS a rendu en février 2024 obsolète l'utilisation des versions de TLS antérieures. Nous vous recommandons d'utiliser TLS 1.3. Si vous avez des exigences particulières en matière de chiffrement en transit, vous pouvez trouver des solutions tierces dans AWS Marketplace.

Étapes d'implémentation

- Appliquer le chiffrement en transit : vos exigences en matière de chiffrement doivent être définies selon les dernières normes et bonnes pratiques en matière de sécurité, et doivent autoriser uniquement des protocoles sécurisés. Par exemple, configurez un groupe de sécurité afin d'autoriser uniquement le protocole HTTPS pour un Application Load Balancer ou une instance Amazon EC2.
- Configurez des protocoles sécurisés dans les services de périphérie : configurez le protocole HTTPS avec Amazon CloudFront et utilisez un profil de sécurité adapté à votre posture de sécurité et à votre cas d'utilisation.
- Utiliser un VPN pour la connectivité externe : envisagez d'utiliser un VPN IPsec pour sécuriser les connexions point à point ou réseau à réseau afin d'assurer à la fois la confidentialité et l'intégrité des données.
- Configurer des protocoles sécurisés dans les équilibreurs de charge : sélectionnez une politique de sécurité fournissant les suites de chiffrement les plus puissantes prises en charge par les clients qui se connecteront à l'écouteur. Créez un écouteur HTTPS pour votre Application Load Balancer.
- Configurer des protocoles sécurisés dans Amazon Redshift : configurez votre cluster pour exiger une connexion SSL (Secure Socket Layer) ou TLS (Transport Layer Security).
- Configurer des protocoles sécurisés : consultez la documentation de service AWS pour déterminer les capacités de chiffrement en transit.
- Configurez un accès sécurisé lors du téléchargement vers des compartiments Amazon S3 : utilisez les contrôles de stratégie de compartiment Amazon S3 pour garantir un accès sécurisé aux données.
- Envisagez d'utiliser AWS Certificate Manager : ACM vous permet de provisionner, de gérer et de déployer des certificats TLS publics à utiliser avec des services AWS.
- Envisagez d'utiliser AWS Private Certificate Authority pour les besoins du PKI privé : AWS Private CA vous permet de créer des hiérarchies d'autorités de certification (AC) privées pour délivrer des certificats X.509 d'entité finale qui peuvent être utilisés pour créer des canaux TLS cryptés.

Ressources

Documents connexes :

- Utilisation du protocole HTTPS avec CloudFront
- Connexion de votre VPC à des réseaux distants utilisant AWS Virtual Private Network
- Création d'un écouteur HTTPS pour votre Application Load Balancer
- Didacticiel : Configurer SSL/TLS sur Amazon Linux 2
- Utilisation de SSL/TLS pour chiffrer une connexion à une instance de base de données
- Configuration des options de sécurité des connexions

SEC09-BP03 Authentifier les communications réseau

Vérifiez l'identité des communications à l'aide de protocoles comme TLS (Transport Layer Security) ou IPsec qui prennent en charge l'authentification.

Concevez votre charge de travail de manière à utiliser des protocoles réseau sécurisés et authentifiés lors de la communication entre les services, les applications ou avec les utilisateurs. L'utilisation de protocoles réseau qui prennent en charge l'authentification et l'autorisation permet de mieux contrôler les flux du réseau et de réduire l'impact des accès non autorisés.

Résultat escompté : une charge de travail avec un plan de données et un plan de contrôle bien définis circulent entre les services. Les flux de trafic utilisent des protocoles réseau authentifiés et chiffrés lorsque cela est techniquement possible.

Anti-modèles courants :

- Flux de trafic non chiffrés ou non authentifiés au sein de votre charge de travail.
- Réutilisation des informations d'authentification par plusieurs utilisateurs ou entités.
- S'appuyer uniquement sur les contrôles réseau pour contrôler les accès.
- Créer un mécanisme d'authentification personnalisé au lieu d'utiliser des mécanismes d'authentification standard.
- Flux de trafic trop permissifs entre les composants des services ou d'autres ressources dans le VPC.

Avantages liés au respect de cette bonne pratique :

- Limite l'impact des accès non autorisés à une partie de la charge de travail.
- Offre la garantie que les actions ne sont effectuées que par des entités authentifiées.
- Améliore le découplage des services en définissant clairement et en appliquant les interfaces de transfert de données prévues.
- Améliore la surveillance, la journalisation et la réponse aux incidents grâce à l'attribution des demandes et à des interfaces de communication bien définies.
- Assure une défense approfondie de vos charges de travail en combinant des contrôles réseau avec des contrôles d'authentification et d'autorisation.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Les modèles de trafic réseau de votre charge de travail peuvent être classés en deux catégories :

- Le trafic est-ouest représente les flux de trafic entre les services qui constituent une charge de travail.
- Le trafic nord-sud représente les flux de trafic entre votre charge de travail et les consommateurs.

Le chiffrement du trafic nord-sud est courant, mais la sécurisation du trafic est-ouest à l'aide de protocoles authentifiés l'est moins. Les pratiques modernes de sécurité recommandent que la conception du réseau ne permette pas à elle seule d'établir une relation de confiance entre deux entités. Lorsque deux services peuvent résider dans les limites d'un réseau commun, il est toujours recommandé de chiffrer, d'authentifier et d'autoriser les communications entre ces services.

Par exemple, les API de service AWS utilisent le protocole de signature AWS Signature Version 4 (SigV4) pour authentifier l'appelant, quel que soit le réseau d'où provient la demande. Cette authentification garantit que les API AWS peuvent vérifier l'identité de la personne qui a demandé l'action, et cette identité peut ensuite être combinée avec des stratégies pour décider si l'action doit être autorisée ou non.

Des services tels qu'Amazon VPC Lattice et Amazon API Gateway vous permettent d'utiliser le même protocole de signature SigV4 pour ajouter une authentification et une autorisation au trafic est-ouest dans vos propres charges de travail. Si des ressources extérieures à votre environnement AWS ont besoin de communiquer avec des services qui nécessitent une authentification et une autorisation basées sur le protocole SIGv4, vous pouvez utiliser AWS Identity and Access

Management Rôles Anywhere (IAM) sur la ressource hors AWS pour obtenir des informations d'identification AWS temporaires. Ces informations d'identification peuvent être utilisées pour signer les demandes de services utilisant SigV4 pour autoriser l'accès.

L'authentification mutuelle TLS (mTLS) est un autre mécanisme courant pour authentifier le trafic est-ouest. De nombreuses applications IoT (Internet des objets) et B2B, ainsi que des microservices utilisent mTLS pour valider l'identité des deux côtés d'une communication TLS à l'aide de certificats X.509 côté client et côté serveur. Ces certificats peuvent être émis par AWS Private Certificate Authority (AWS Private CA). Vous pouvez utiliser des services tels qu'Amazon API Gateway pour fournir une authentification mTLS pour les communications inter-charges de travail ou intra-charge de travail. Application Load Balancer prend également en charge mTLS pour les charges de travail orientées côté interne ou externe. mTLS fournit des informations d'authentification pour les deux côtés d'une communication TLS, mais elle ne fournit pas de mécanisme d'autorisation.

Enfin, OAuth 2.0 et OpenID Connect (OIDC) sont deux protocoles généralement utilisés pour contrôler l'accès aux services par les utilisateurs, mais ils sont également de plus en plus populaires pour le trafic de service à service. API Gateway fournit un autorisateur JSON Web Token (JWT) permettant aux charges de travail de restreindre l'accès aux routes d'API à l'aide des JWT émis par des fournisseurs d'identité OIDC ou OAuth 2.0. Les champs d'application OAuth2 peuvent être utilisés comme source pour les décisions d'autorisation de base, mais les contrôles d'autorisation doivent encore être mis en œuvre dans la couche applicative, et les champs d'application OAuth2 ne peuvent pas à eux seuls répondre à des besoins d'autorisation plus complexes.

Étapes d'implémentation

- Définissez et documentez les flux de votre réseau de charge de travail : la première étape de la mise en œuvre d'une stratégie de défense en profondeur consiste à définir les flux de trafic de votre charge de travail.
  - Créez un diagramme de flux de données qui définit clairement la transmission des données entre les différents services qui constituent votre charge de travail. Ce schéma constitue la première étape de l'application de ces flux par le biais de réseaux authentifiés.
  - Instrumentez votre charge de travail lors des phases de développement et de test pour vérifier que le diagramme de flux de données reflète avec précision le comportement de la charge de travail lors de l'exécution.
  - Un diagramme de flux de données peut également être utile lors de l'exécution d'un exercice de modélisation des menaces, comme décrit dans SEC01-BP07 Identifier les menaces et hiérarchiser les mesures d'atténuation à l'aide d'un modèle de menace.

- Établissez des contrôles réseau : tenez compte des capacités AWS permettant d'établir des contrôles réseau alignés sur vos flux de données. Les limites du réseau ne doivent pas représenter le seul contrôle de sécurité, mais elles constituent une couche de la stratégie de défense en profondeur visant à protéger votre charge de travail.
  - Utilisez des groupes de sécurité pour établir, définir et limiter les flux de données entre les ressources.
  - Envisagez d'utiliser AWS PrivateLink pour communiquer à la fois avec AWS et les services tiers qui prennent en charge AWS PrivateLink. Les données envoyées via un point de terminaison d'interface AWS PrivateLink restent dans le réseau AWS et ne transitent pas par l'Internet public.
- Mettez en œuvre l'authentification et l'autorisation pour tous les services de votre charge de travail : choisissez l'ensemble de services AWS le plus approprié pour fournir des flux de trafic authentifiés et cryptés dans votre charge de travail.
  - Pensez à Amazon VPC Lattice pour sécuriser les communications entre services. VPC Lattice peut utiliser l'authentification SigV4 combinée à des politiques d'authentification pour contrôler l'accès de service à service.
  - Pour la communication de service à service à l'aide de mTLS, envisagez d'utiliser API Gateway ou Application Load Balancer. AWS Private CA peut être utilisé pour établir une hiérarchie d'autorité de certification privée capable d'émettre des certificats à utiliser avec mTLS.
  - Lors de l'intégration à des services utilisant OAuth 2.0 ou OIDC, envisagez d'utiliser API Gateway avec l'autorisateur JWT.
  - Pour la communication entre votre charge de travail et les appareils IoT, envisagez AWS IoT Core, qui propose plusieurs options pour le chiffrement et l'authentification du trafic réseau.
- Surveillez les accès non autorisés : surveillez en permanence les canaux de communication imprévus, les principaux non autorisés qui tentent d'accéder à des ressources protégées et les autres modèles d'accès inappropriés.
  - Si vous utilisez VPC Lattice pour gérer l'accès à vos services, pensez à activer et à surveiller les journaux d'accès VPC Lattice. Ces journaux contiennent des informations sur le demandeur et le réseau, notamment le VPC source et de destination, et les métadonnées des demandes.
  - Envisagez d'activer les journaux de flux VPC pour capturer des métadonnées sur les flux réseau et vérifier régulièrement la présence d'anomalies.
  - Reportez-vous au Guide de réponse aux incidents de sécurité AWS et à la section Réponse aux incidents du pilier Sécurité AWS Well-Architected Framework pour plus de conseils sur la planification, la simulation et la réponse aux incidents de sécurité.

Ressources

Bonnes pratiques associées :

- SEC03-BP07 Analyser l’accès public et intercompte
- SEC02-BP02 Utiliser des informations d’identification temporaires
- SEC01-BP07 Identifier les menaces et hiérarchiser les atténuations à l’aide d’un modèle de menaces

Documents connexes :

- Évaluation des méthodes de contrôle d’accès pour sécuriser les API Amazon API Gateway
- Configuration de l’authentification TLS mutuelle pour une API REST
- Comment sécuriser les points de terminaison HTTP API Gateway avec l’autorisateur JWT
- Autoriser les appels directs vers les services AWS à l’aide du fournisseur d’informations d’identification AWS IoT Core
- Guide d’intervention en cas d’incident de sécurité AWS

Vidéos connexes :

- AWS re:invent 2022: Introducing VPC Lattice
- AWS re:invent 2020: Serverless API authentication for HTTP APIs on AWS

Exemples connexes :

- Atelier Amazon VPC Lattice
- Épisode 1 de Zero-Trust — L’atelier Phantom Service Perimeter

Intervention en cas d’incidents

Question

- SÉC 10. Comment anticiper les incidents, y répondre et effectuer une reprise après incident ?

SÉC 10. Comment anticiper les incidents, y répondre et effectuer une reprise après incident ?

Même avec des contrôles préventifs et de détection matures, votre organisation doit mettre en place des mécanismes pour répondre aux incidents de sécurité et en atténuer l'impact potentiel. Votre préparation affectera fortement la capacité de vos équipes à opérer efficacement lors d'un incident, à analyser, isoler et contenir les problèmes, et à rétablir les opérations à un état de fonctionnement correct. La mise en place des outils et des accès avant un incident de sécurité, puis la pratique régulière de la réponse aux incidents pendant des exercices de simulation, vous permettent de rétablir les opérations tout en minimisant les interruptions d'activité.

Bonnes pratiques

- SEC10-BP01 Identifier les postes clés internes ainsi que les principales ressources externes
- SEC10-BP02 Développer des plans de gestion des incidents
- SEC10-BP03 Préparer les capacités de criminalistique
- SEC10-BP04 Développer et tester des playbooks de réponse aux incidents de sécurité
- SEC10-BP05 Préallouer les accès
- SEC10-BP06 Outils de pré-déploiement
- SEC10-BP07 Exécuter des simulations
- SEC10-BP08 Mettre en place un cadre pour tirer les leçons des incidents

SEC10-BP01 Identifier les postes clés internes ainsi que les principales ressources externes

Identifiez les postes clés internes et externes, les ressources et les obligations légales qui aideront votre organisation à réagir en cas d'incident.

Résultat escompté : vous disposez d'une liste des principaux membres du personnel, de leurs coordonnées et des rôles qu'ils jouent lorsqu'ils répondent à un événement de sécurité. Vous consultez régulièrement ces informations et vous les mettez à jour de façon à refléter les changements de personnel du point de vue des outils internes et externes. Lorsque vous documentez ces informations, vous tenez compte de tous les fournisseurs de services et fournisseurs tiers, y compris les partenaires de sécurité, les fournisseurs de cloud et les applications de logiciel en tant que service (SaaS). Lors d'un événement de sécurité, le personnel est disponible avec le niveau de responsabilité, de contexte et d'accès approprié pour être en mesure de réagir et de récupérer.

Anti-modèles courants :

- Ne pas gérer une liste actualisée des principaux membres du personnel avec leurs coordonnées, leurs rôles et leurs responsabilités lorsqu'ils répondent à des événements de sécurité.
- Supposer que tout le monde comprend les personnes, les dépendances, l'infrastructure et les solutions lors de la réponse et de la reprise après un événement.
- Ne pas disposer d'un document ou d'un référentiel de connaissances représentant la conception d'une infrastructure ou d'une application clé.
- Ne pas disposer de processus d'intégration appropriés permettant aux nouveaux employés de contribuer efficacement à la réponse à un événement de sécurité, par exemple en effectuant des simulations d'événements.
- Aucune procédure de remontée n'est en place lorsque le personnel clé est temporairement indisponible ou s'il ne répond pas lors d'événements de sécurité.

Avantages liés au respect de cette bonne pratique : cette pratique réduit le temps de triage et de réponse consacré à l'identification du personnel approprié et de ses rôles lors d'un événement. Minimisez le temps perdu lors d'un événement en tenant à jour une liste des principaux membres du personnel et de leurs rôles afin de pouvoir faire le tri des personnes appropriées et de les aider à récupérer après un événement.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Identifier les postes clés au sein de votre organisation : tenez à jour une liste des employés au sein de votre organisation que vous devez impliquer. Passez régulièrement en revue et mettez à jour ces informations en cas de changements au sein du personnel, par exemple des changements organisationnels, des promotions et des changements d'équipe. Cette étape est particulièrement importante pour les rôles clés tels que les gestionnaires d'incidents, les intervenants en cas d'incident et le responsable des communications.

- Gestionnaire d'incidents : les gestionnaires d'incidents ont une autorité globale lors de la réponse à l'événement.
- Intervenants en cas d'incidents : les intervenants en cas d'incidents sont responsables des activités d'enquête et de correction. Ces personnes peuvent varier en fonction du type d'événement, mais il s'agit généralement de développeurs et d'équipes opérationnelles responsables de l'application concernée.

- Responsable des communications : le responsable des communications est responsable des communications internes et externes, en particulier avec les agences publiques, les régulateurs et les clients.
- Processus d'intégration : formez et intégrez régulièrement les nouveaux employés afin de les doter des compétences et des connaissances nécessaires pour contribuer efficacement aux efforts de réponse aux incidents. Incorporez des simulations et des exercices pratiques dans le cadre du processus d'intégration afin de faciliter leur préparation.
- Experts en la matière (PME) : dans le cas d'équipes distribuées et autonomes, nous vous recommandons d'identifier une PME pour les charges de travail critiques. Ils fournissent des informations sur le fonctionnement et la classification des données des charges de travail critiques impliquées dans l'événement.

Exemple de format de table :

```
  | Role | Name | Contact Information | Responsibilities |
1 | ——- | ——- | ——- | ——- |
2 | Incident Manager | Jane Doe| jane.doe@example.com | Overall authority during
 response |
3 | Incident Responder | John Smith | john.smith@example.com | Investigation and
 remediation |
4 | Communications Lead | Emily Johnson | emily.johnson@example.com | Internal and
 external communications |
5 | Communications Lead | Michael Brown | michael.brown@example.com | Insights on
 critical workloads |
```

Envisagez d'utiliser la fonctionnalité AWSSystems Manager Incident Manager pour capturer les contacts clés, définir un plan d'intervention, automatiser les plannings d'astreinte et définir des plans d'escalade. Automatisez et alternez tout le personnel grâce à un calendrier d'astreinte, de sorte que la responsabilité de la charge de travail soit partagée entre ses propriétaires. Cela favorise les bonnes pratiques, telles que l'émission de métriques et de journaux pertinents, ainsi que la définition de seuils d'alarme importants pour la charge de travail.

Identifier les partenaires externes : les entreprises utilisent des outils conçus par des fournisseurs indépendants de logiciels (ISV), des partenaires et des sous-traitants pour créer des solutions différenciantes pour leurs clients. Impliquez le personnel clé de ces parties qui peut vous aider à répondre à un incident et à récupérer après un incident. Nous vous recommandons de vous inscrire au niveau approprié d'Support afin d'accéder rapidement à des experts AWS en la matière par le biais d'une demande d'assistance. Envisagez des agencements similaires avec tous les fournisseurs

de solutions critiques pour les charges de travail. Certains événements de sécurité obligent les entreprises cotées en bourse à informer les agences publiques et les régulateurs concernés de l'événement et de ses impacts. Dressez une liste avec les coordonnées des départements concernés et des personnes responsables, et veillez à ce qu'elle reste à jour.

Étapes d'implémentation

1. Mettez en place une solution de gestion des incidents.
   a. Envisagez de déployer Incident Manager dans votre compte d'outils de sécurité.
2. Définissez les contacts dans votre solution de gestion des incidents.
   a. Définissez au moins deux types de canaux de communication pour chaque contact (SMS, téléphone ou e-mail, par exemple), afin de pouvoir avec certitude joindre les personnes concernées lors d'un incident.
3. Définissez un plan d'intervention.
   a. Identifiez les contacts les plus appropriés à impliquer lors d'un incident. Définissez des plans de remontée en fonction des rôles du personnel à impliquer, plutôt que des contacts individuels. Envisagez d'inclure des contacts susceptibles d'être chargés d'informer les entités externes, même s'ils ne sont pas directement impliqués dans la résolution de l'incident.

Ressources

Bonnes pratiques associées :

- OPS02-BP03 Les activités opérationnelles ont des propriétaires identifiés responsables de leurs performances

Documents connexes :

- Guide d'intervention en cas d'incident de sécurité AWS

Exemples connexes :

- Cadre du playbook du client AWS
- Prepare for and respond to security incidents in your environment AWS

Outils associés :

- AWS Systems Manager Incident Manager

Vidéos connexes :

- L’approche d’Amazon en matière de sécurité pendant le développement

SEC10-BP02 Développer des plans de gestion des incidents

Le premier document à élaborer pour la réponse aux incidents est le plan d’intervention en cas d’incident. Le plan d’intervention en cas d’incident est conçu pour servir de base à votre programme et à votre stratégie de réponse aux incidents.

Avantages liés au respect de cette bonne pratique : le développement de processus de réponse aux incidents complets et clairement définis est essentiel à la réussite d’un programme de réponse aux incidents évolutif. Lorsqu’un incident de sécurité se produit, des étapes et des flux de travail clairs peuvent vous aider à réagir rapidement. Vous disposez peut-être déjà de processus de réponse aux incidents. Quel que soit votre état actuel, il est important de mettre à jour, d’itérer et de tester régulièrement vos processus de réponse aux incidents.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Un plan de gestion des incidents est essentiel pour réagir, atténuer et se remettre des répercussions potentielles des incidents de sécurité. Un plan de gestion des incidents est un processus structuré qui permet d’identifier les incidents de sécurité, d’y remédier et d’y répondre rapidement.

Le cloud comporte un grand nombre de rôles et exigences opérationnels identiques à ceux d’un environnement sur site. Lorsque vous créez un plan de gestion des incidents, il est important de tenir compte des stratégies d’intervention et de récupération qui correspondent le mieux aux résultats métier et aux exigences de conformité. Par exemple, si vous exécutez des charges de travail dans AWS qui sont conformes à FedRAMP aux États-Unis, suivez les recommandations fournies dans le document NIST SP 800-61 Guide relatif à la gestion de la sécurité informatique. De la même manière, lorsque vous exécutez des charges de travail qui stockent des données d’identification personnelle (PII), réfléchissez à la façon de protéger ces données et de résoudre les problèmes liés à la résidence et à l’utilisation des données.

Lorsque vous élaborez un plan de gestion des incidents pour vos charges de travail dans AWS, commencez par le modèle de responsabilité partagée AWS pour élaborer une approche de défense

approfondie en matière d'intervention en cas d'incidents. Dans le cadre de ce modèle, AWS gère la sécurité du cloud et vous êtes responsable de la sécurité dans le cloud. Cela signifie que vous conservez le contrôle et que vous êtes responsable des contrôles de sécurité que vous choisissez d'implémenter. Le Guide sur les interventions en cas d'incident de sécurité AWS détaille les concepts clés et les conseils de base pour l'élaboration d'un plan de gestion des incidents axé sur le cloud.

Un plan de gestion des incidents efficace doit être répété constamment, tout en poursuivant votre objectif d'opérations dans le cloud. Envisagez d'utiliser les plans d'implémentation décrits ci-dessous pour créer et faire évoluer votre plan de gestion des incidents.

Étapes d'implémentation

1. Définissez les rôles et les responsabilités au sein de votre organisation pour gérer les événements de sécurité. Cela devrait impliquer des représentants de différents départements, notamment :
   - Ressources humaines (RH)
   - Équipe de direction
   - Département juridique
   - Propriétaires et développeurs d'applications (experts spécifiques ou SME)
2. Désignez clairement les intervenants responsables, redevables, consultés et informés (RACI, Responsible, Accountable, Consulted, Informed) lors d'un incident. Créez une matrice RACI pour faciliter une communication rapide et directe, et désignez clairement le leadership aux différentes étapes d'un événement.
3. Impliquez les propriétaires et les développeurs d'applications (SME) lors d'un incident, car ils peuvent fournir des informations et un contexte précieux pour la mesure de l'impact. Établissez des relations avec ces SME et mettez en pratique des scénarios de réponse aux incidents avec elles avant qu'un véritable incident se produise.
4. Impliquez des partenaires de confiance ou des experts externes dans le processus d'enquête ou de réponse, car ils peuvent apporter une expertise et un point de vue supplémentaires.
5. Alignez vos rôles et plans de gestion des incidents sur les réglementations locales ou les exigences de conformité qui régissent votre organisation.
6. Mettez en pratique et testez régulièrement vos plans de réponse aux incidents, et impliquez tous les rôles et responsabilités définis. Cela contribue à rationaliser le processus et à vérifier que vous disposez d'une réponse coordonnée et efficace aux incidents de sécurité.
7. Passez en revue et mettez à jour les rôles, les responsabilités et la matrice RACI régulièrement ou à mesure que votre structure organisationnelle ou vos exigences changent.

Comprenez les équipes d'intervention et le support AWS

- AWS Support
  - Support propose un large choix de formules qui vous permettent d'accéder aux outils et aux compétences nécessaires pour garantir la réussite et la santé opérationnelle de vos solutions AWS. Si vous avez besoin d'un support technique et de ressources supplémentaires pour planifier, déployer et optimiser votre environnement AWS, vous pouvez sélectionner le plan de support le plus adapté à votre cas d'utilisation AWS.
  - Considérez le Centre d'assistance dans AWS Management Console (connexion requise) en tant que point de contact central pour obtenir de l'aide en cas de problèmes affectant vos ressources AWS. L'accès à Support est contrôlé par AWS Identity and Access Management. Pour plus d'informations sur l'accès aux fonctionnalités Support, consultez Démarrer avec Support.
- Équipe de réponse aux incidents clients (CIRT) AWS
  - L'équipe de réponse aux incidents clients (CIRT) AWS est une équipe AWS internationale spécialisée et disponible 24 heures sur 24, 7 jours sur 7, qui fournit une assistance aux clients lors d'événements de sécurité actifs côté client du Modèle de responsabilité partagée AWS.
  - Lorsque la CIRT AWS vous accompagne, elle fournit une assistance en matière de triage et de récupération en cas d'événement de sécurité actif sur AWS. Elle peut vous aider à analyser les causes profondes à l'aide des journaux de service AWS et vous fournir des recommandations pour la récupération. Elle peut également fournir des recommandations de sécurité et des bonnes pratiques pour vous aider à éviter des incidents de sécurité à l'avenir.
  - Les clients AWS peuvent contacter la CIRT AWS par le biais d'un cas Support.
- Support de réponse aux attaques DDoS
  - AWS propose AWS Shield, qui est un service de protection DDoS (Distributed Denial of Service) géré qui protège les applications Web s'exécutant sous AWS. Shield assure une détection continue et une atténuation automatique des risques pour minimiser les temps d'arrêt et la latence des applications, afin qu'il ne soit pas nécessaire d'avoir recours à Support pour bénéficier de la protection DDoS. Il existe deux niveaux de Shield : AWS Shield Standard et AWS Shield Advanced. Pour en savoir plus sur les différences entre ces deux niveaux, consultez la documentation relative aux fonctionnalités Shield.
- AWS Managed Services (AMS)
  - AWS Managed Services (AMS) fournit une gestion continue de votre infrastructure AWS afin que vous puissiez vous concentrer sur vos applications. En appliquant les bonnes pratiques pour gérer votre infrastructure, AMS permet de réduire les coûts et risques de fonctionnement. AMS automatise les activités courantes telles que les demandes de modification, la surveillance, la

gestion des correctifs, la sécurité et les services de sauvegarde, et fournit des services pour l'intégralité du cycle de vie pour mettre en service, exécuter et soutenir votre infrastructure.

- AMS prend la responsabilité de déployer une suite de contrôles de sécurité et fournit une réponse de première ligne 24 heures sur 24, 7 jours sur 7 aux alertes. Lorsqu'une alerte est déclenchée, AMS suit un ensemble standard de playbooks automatisés et manuels pour vérifier une réponse cohérente. Ces playbooks sont partagés avec les clients AMS lors de l'intégration afin qu'ils puissent développer et coordonner une réponse avec AMS.

Élaborez le plan d'intervention en cas d'incident

Le plan d'intervention en cas d'incident est conçu pour servir de base à votre programme et à votre stratégie de réponse aux incidents. Le plan d'intervention en cas d'incident doit figurer dans un document formel. Un plan d'intervention en cas d'incident comprend généralement les sections suivantes :

- Présentation de l'équipe d'intervention en cas d'incidents : décrit les objectifs et les fonctions de l'équipe de réponse aux incidents.
- Rôles et responsabilités : répertorie les parties prenantes de la réponse aux incidents et détaille leurs rôles en cas d'incident.
- Plan de communication : détaille les coordonnées et la manière dont vous communiquez lors d'un incident.
- Méthodes de communication relative à la sauvegarde : il est recommandé d'utiliser une communication hors bande comme solution de secours pour les communications en cas d'incident. Un exemple d'application qui fournit un canal de communication hors bande sécurisé est AWS Wickr.
- Phases de l'intervention en cas d'incident et mesures à prendre : énumère les phases de la réponse aux incidents (par exemple, détection, analyse, éradication, maîtrise et récupération), y compris les mesures de haut niveau à prendre au cours de ces phases.
- Définitions de la gravité et de la priorité des incidents : décrit en détail comment classer la gravité d'un incident, comment hiérarchiser l'incident, puis comment les définitions de gravité affectent les procédures de remontée.

Bien que ces sections soient communes à des entreprises de tailles et de secteurs différents, le plan d'intervention en cas d'incident de chaque organisation est unique. Vous devez élaborer un plan d'intervention en cas d'incident qui convient le mieux à votre organisation.

Ressources

Bonnes pratiques associées :

- SEC04 Détection

Documents connexes :

- Guide d’intervention en cas d’incident de sécurité AWS
- NIST: Computer Security Incident Handling Guide

SEC10-BP03 Préparer les capacités de criminalistique

Pour anticiper un incident de sécurité, envisagez de développer des fonctionnalités d’analyse poussée pour faciliter les enquêtes sur les événements de sécurité.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Les concepts issus de la criminalistique traditionnelle sur site s'appliquent à. AWS Pour obtenir des informations clés permettant de commencer à renforcer les capacités de criminalistique dans le AWS Cloud, voir Stratégies relatives à l'environnement d'investigation judiciaire dans le. AWS Cloud

Une fois que vous avez configuré votre environnement et votre Compte AWS structure pour la criminalistique, définissez les technologies nécessaires pour appliquer efficacement des méthodologies médico-légales fiables au cours des quatre phases :

- Collecte : collectez les AWS journaux pertinents AWS CloudTrail AWS Config, tels que les journaux de VPC flux et les journaux au niveau de l'hôte. Collectez des instantanés, des sauvegardes et des vidages de mémoire des AWS ressources concernées, le cas échéant.
- Examen : examinez les données collectées en extrayant et en évaluant les informations pertinentes.
- Analyse : analysez les données collectées afin de comprendre l’incident et d’en tirer des conclusions.
- Production de rapports : présentez les informations issues de la phase d’analyse.

Étapes d’implémentation

Préparer votre environnement d’analyse poussée

AWS Organizationsvous permet de gérer et de gouverner de manière centralisée un AWS environnement à mesure que vous développez et adaptez vos AWS ressources. Une AWS organisation consolide les vôtres Comptes AWS afin que vous puissiez les administrer en tant qu'unité unique. Vous pouvez utiliser les unités organisationnelles (OUs) pour regrouper les comptes afin de les administrer en tant qu'unité unique.

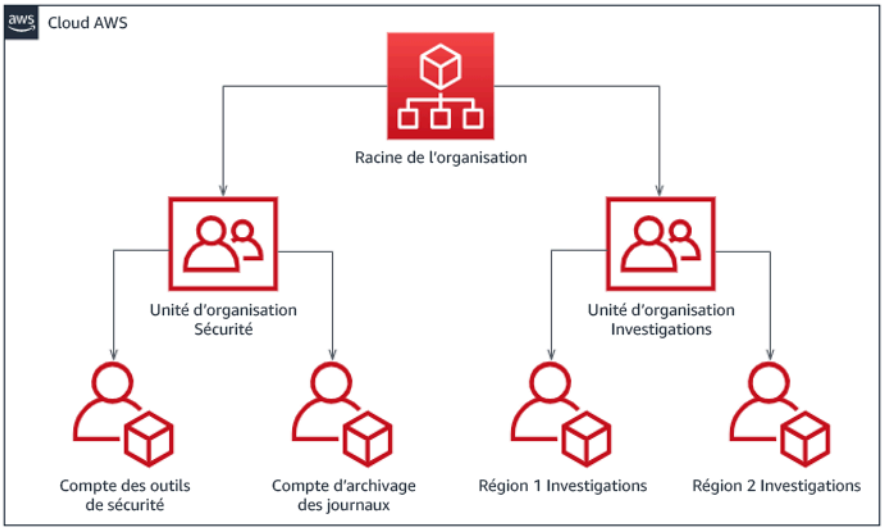
Pour la réponse aux incidents, il est utile de disposer d'une Compte AWS structure prenant en charge les fonctions de réponse aux incidents, qui comprend une unité d'organisation de sécurité et une unité d'organisation médico-légale. Au sein de l'unité d'organisation de sécurité, vous devez disposer de comptes pour :

- Archivage des journaux : regroupez les journaux dans une archive de journaux Compte AWS avec des autorisations limitées.
- Outils de sécurité : centralisez les services de sécurité dans un outil Compte AWS de sécurité. Ce compte joue le rôle d'administrateur délégué pour les services de sécurité.

Au sein de l'unité d'organisation d'analyse poussée, vous avez la possibilité de mettre en place un ou plusieurs comptes d'analyse poussée pour chaque région dans laquelle vous opérez, selon ce qui convient le mieux à votre entreprise et à votre modèle opérationnel. Si vous créez un compte médico-légal par région, vous pouvez bloquer la création de AWS ressources en dehors de cette région et réduire le risque que des ressources soient copiées vers une région non prévue. Par exemple, si vous opérez uniquement dans les régions USA Est (Virginie du Nord) (`us-east-1`) et USA Ouest (Oregon) (`us-west-2`), vous aurez deux comptes dans l'unité d'organisation médico-légale : un pour `us-east-1` et un pour `us-west-2`.

Vous pouvez créer une enquête Compte AWS pour plusieurs régions. Vous devez faire preuve de prudence lorsque vous copiez AWS des ressources sur ce compte afin de vérifier que vous respectez vos exigences en matière de souveraineté des données. Étant donné que la mise en place de nouveaux comptes prend du temps, il est impératif de créer et d'instrumenter les comptes d'analyse poussée bien avant un incident afin que les intervenants puissent être prêts à les utiliser efficacement pour intervenir.

Le diagramme suivant présente un exemple de structure de compte, y compris une unité d'organisation d'analyse poussée avec des comptes d'analyse poussée par région :



Structure de compte par région pour la réponse aux incidents

Conservez les sauvegardes et les instantanés

La configuration de sauvegardes des systèmes et des bases de données clés s'avère essentielle pour récupérer d'un incident de sécurité et à des fins d'analyse poussée. Une fois les sauvegardes en place, vous pouvez restaurer vos systèmes à leur état stable antérieur. AWS Activé, vous pouvez prendre des instantanés de différentes ressources. Les instantanés vous fournissent des point-in-time copies de sauvegarde de ces ressources. De nombreux AWS services peuvent vous aider en matière de sauvegarde et de restauration. Pour en savoir plus sur ces services et approches de la sauvegarde et de la récupération, reportez-vous au Recommandation en matière de sauvegarde et de récupération et à Utiliser les sauvegardes pour récupérer après un incident de sécurité.

Il est essentiel que vos sauvegardes soient bien protégées, en particulier dans le cas de rançongiciels. Pour obtenir des conseils sur la sécurisation de vos sauvegardes, reportez-vous aux 10 meilleures pratiques de sécurité en matière de sauvegarde dans AWS. Outre la sécurisation de vos sauvegardes, vous devez régulièrement tester vos processus de sauvegarde et de restauration pour vérifier que la technologie et les processus que vous avez mis en place fonctionnent comme prévu.

Automatisez la criminalistique

Lors d'un événement de sécurité, votre équipe de réponse aux incidents doit être en mesure de collecter et d'analyser des preuves rapidement tout en préservant la précision pendant la période entourant l'événement (par exemple en capturant les journaux relatifs à un événement ou à une ressource spécifique ou en collectant un fichier mémoire d'une EC2 instance Amazon). Il est à la fois

difficile et fastidieux pour l'équipe de réponse aux incidents de collecter manuellement les preuves pertinentes, en particulier sur un grand nombre d'instances et de comptes. De plus, la collecte manuelle peut faire l'objet d'erreurs humaines. Pour ces raisons, vous devez développer et mettre en œuvre autant que possible l'automatisation de l'analyse poussée.

AWS propose un certain nombre de ressources d'automatisation pour la criminalistique, qui sont répertoriées dans la section Ressources suivante. Ces ressources sont des exemples de modèles d'analyse poussée que nous avons développés et que les clients ont mis en œuvre. Bien qu'elles puissent constituer une architecture de référence utile au départ, envisagez de les modifier ou de créer de nouveaux modèles d'automatisation de l'analyse poussée en fonction de votre environnement, de vos exigences, de vos outils et de vos processus d'analyse poussée.

Ressources

Documents connexes :

- AWS Guide de réponse aux incidents de sécurité - Développez des capacités de criminalistique
- AWS Guide de réponse aux incidents de sécurité - Ressources médico-légales
- Stratégies relatives à l'environnement d'investigation médico-légale dans le AWS Cloud
- Comment automatiser la collecte médico-légale de disques dans AWS
- AWS Conseils prescriptifs - Automatisez la réponse aux incidents et la criminalistique

Vidéos connexes :

- Automatisation de la réponse aux incidents et investigations

Exemples connexes :

- Cadre de réponse automatique aux incidents et de criminalistique
- Orchestrateur de criminalistique automatisé pour Amazon EC2

SEC10-BP04 Développer et tester des playbooks de réponse aux incidents de sécurité

L'élaboration de playbooks est une étape clé de la préparation de vos processus de réponse aux incidents. Les playbooks de réponse aux incidents fournissent des recommandations et les étapes à suivre en cas d'événement de sécurité. Le fait de disposer d'une structure et d'étapes claires simplifie la réponse et réduit le risque d'erreur humaine.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Il est recommandé de créer des playbooks dans les scénarios d'incidents suivants :

- Incidents prévus : créez des playbooks pour les incidents que vous anticipez. Cela inclut des menaces telles que le déni de service (DoS), les rançongiciels et la compromission des informations d'identification.
- Constatations ou alertes de sécurité connues : créez des playbooks pour vos résultats et alertes de sécurité connus, tels que les résultats d'Amazon GuardDuty. Lorsque vous recevez un résultat de GuardDuty, le playbook doit indiquer des étapes claires pour éviter de mal gérer ou d'ignorer l'alerte. Pour plus de détails et de conseils de correction, consultez Correction des problèmes de sécurité découverts par GuardDuty.

Les playbooks doivent contenir les étapes techniques qu'un analyste de sécurité doit suivre afin d'enquêter de manière adéquate et de répondre à un éventuel incident de sécurité.

Étapes d'implémentation

Les éléments à inclure dans un playbook incluent :

- Présentation du Playbook : quel scénario de risque ou d'incident ce playbook aborde-t-il ? Quel est l'objectif du playbook ?
- Prérequis : quels journaux, mécanismes de détection et outils automatisés sont requis pour ce scénario d'incident ? Quelle est la notification attendue ?
- Informations de communication et d'escalade : qui est impliqué et quelles sont ses coordonnées ? Quelles sont les responsabilités de chacune des parties prenantes ?
- Étapes d'intervention : quelles sont les mesures tactiques à prendre au cours des différentes phases de la réponse à un incident ? Quelles requêtes un analyste doit-il exécuter ? Quel code doit être exécuté pour obtenir le résultat souhaité ?
  - Détection : comment l'incident sera-t-il détecté ?
  - Analyse : comment l'étendue de l'impact sera-t-elle déterminée ?
  - Contenu : comment l'incident sera-t-il isolé pour en limiter la portée ?
  - Éradication : comment éliminer la menace de l'environnement ?
  - Remise : comment le système ou la ressource concernés seront-ils remis en production ?

- Résultats escomptés : une fois les requêtes et le code exécutés, quel est le résultat attendu du playbook ?

Ressources

Bonnes pratiques Well-Architected connexes :

- SEC10-BP02 – Développer des plans de gestion des incidents

Documents connexes :

- Cadre pour les playbooks d'intervention en cas d'incident
- Élaborer vos propres playbooks d'intervention en cas d'incident
- Modèles de guides d'intervention en cas d'incident
- Création d'un runbook de réponse aux incidents AWS à l'aide de playbooks Jupyter et Lake (langue française non garantie)

SEC10-BP05 Préallouer les accès

Vérifiez que les intervenants en cas d'incident disposent du bon accès préalablement alloué dans AWS afin de réduire le temps d'investigation jusqu'à la reprise.

Anti-modèles courants :

- Utilisation du compte racine pour la réponse aux incidents.
- Modification des comptes existants.
- Manipulation des autorisations IAM directement lors de la fourniture d'une élévation de privilèges juste-à-temps.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

AWS recommande de réduire ou de supprimer l'utilisation des informations d'identification durables dans la mesure du possible et de privilégier les informations d'identification temporaire à la place, ainsi que des mécanismes d'escalade des privilèges juste à temps. Les informations d'identification

durables sont sujettes aux risques de sécurité et augmentent les frais généraux opérationnels. Pour la plupart des tâches de gestion, ainsi que pour les tâches de réponse aux incidents, nous vous recommandons de mettre en œuvre la fédération d'identités et l'escalade temporaire pour l'accès administratif. Dans le cadre de ce modèle, un utilisateur demande une élévation à un niveau de privilège plus élevé (par exemple un rôle de réponse aux incidents) et, si l'utilisateur est admissible à cette élévation, une demande est envoyée à un approbateur. Si la demande est approuvée, l'utilisateur reçoit un ensemble d'informations d'identification AWS temporaires qui peuvent être utilisées pour effectuer ses tâches. Une fois que ces informations d'identification ont expiré, l'utilisateur doit soumettre une nouvelle demande d'élévation.

Nous vous recommandons d'utiliser une élévation temporaire des privilèges dans la plupart des cas de réponse aux incidents. La bonne façon de procéder consiste à utiliser AWS Security Token Service et les politiques de session pour délimiter l'accès.

Dans certains cas, les identités fédérées ne sont pas disponibles, par exemple :

- Panne liée à la compromission d'un fournisseur d'identité (IdP).
- Mauvaise configuration ou erreur humaine entraînant la panne d'un système de gestion d'accès fédéré.
- Activité malveillante, par exemple un déni de service distribué (DDoS) ou une indisponibilité du système.

Dans les cas précédents, un accès d'urgence aux bris de verre doit être configuré pour permettre une enquête et une résolution rapide des incidents. Nous vous recommandons d'utiliser un utilisateur, un groupe ou un rôle doté des autorisations appropriées pour effectuer des tâches et accéder aux ressources AWS. Utiliser l'utilisateur racine uniquement pour les tâches qui nécessitent des informations d'identification. Pour vérifier que les intervenants en cas d'incident disposent d'un niveau d'accès approprié à AWS et aux autres systèmes pertinents, nous vous recommandons de pré-allouer des comptes dédiés. Les comptes requièrent un accès privilégié et doivent être étroitement contrôlés et surveillés. Les comptes doivent être créés avec le moins de privilèges requis pour effectuer les tâches nécessaires et le niveau d'accès doit être basé sur les playbooks créés dans le cadre du plan de gestion des incidents.

Utilisez des utilisateurs et des rôles spécialement conçus et dédiés au titre de bonne pratique. L'élévation temporaire de l'accès des utilisateurs ou des rôles via l'ajout de politiques IAM ne permet pas de savoir clairement de quel type d'accès bénéficiaient les utilisateurs pendant l'incident et peut empêcher la révocation des privilèges élevés au niveau supérieur.

Il est important de supprimer autant de dépendances que possible afin de vérifier que l'accès peut être obtenu dans le plus grand nombre possible de scénarios de défaillance. Afin de vous faciliter la tâche, créez un playbook permettant de vérifier que les utilisateurs chargés des réponses en cas d'incident ont été créés en tant qu'utilisateurs dans un compte de sécurité dédié et qu'ils ne sont pas gérés via une solution d'authentification unique ou de fédération existante. Chaque intervenant en cas d'incident doit avoir son propre compte nommé. La configuration du compte doit appliquer des stratégies de mot de passe d'un niveau de sécurité élevé à l'authentification multifactorielle (MFA). Si les playbooks de réponse aux incidents ne nécessitent qu'un accès à la AWS Management Console, l'utilisateur ne doit pas avoir de clés d'accès configurées et il doit lui être explicitement interdit de créer des clés d'accès. Cela peut être configuré avec des politiques IAM ou des politiques de contrôle des services (SCP), comme mentionné dans les bonnes pratiques de sécurité AWS pour les AWS Organizations SCP. Les utilisateurs ne doivent pas avoir d'autres privilèges que la capacité d'assumer des rôles de réponse aux incidents dans d'autres comptes.

Pendant un incident, il peut être nécessaire d'accorder l'accès à d'autres personnes internes ou externes afin de prendre en charge les activités d'analyse, de correction ou de reprise. Dans ce cas, utilisez le mécanisme de playbook mentionné précédemment. Celui-ci doit comporter un processus permettant de s'assurer que tout accès supplémentaire est révoqué immédiatement après l'incident.

Pour s'assurer que l'utilisation des rôles de réponse aux incidents peut être correctement surveillée et vérifiée, il est essentiel que les comptes utilisateur IAM créés à cette fin ne soient pas partagés entre les personnes et que l'Utilisateur racine d'un compte AWS ne soit pas utilisé, à moins qu'ils ne soient nécessaires pour une tâche spécifique. Si l'utilisateur root est requis (par exemple, l'accès IAM à un compte spécifique n'est pas disponible), utilisez un processus distinct avec un playbook disponible afin de vérifier la disponibilité des informations d'identification de l'utilisateur racine et du jeton d'authentification multifactorielle.

Pour configurer les politiques IAM pour les rôles de réponse aux incidents, pensez à utiliser IAM Access Analyzer pour générer des politiques basées sur les journaux AWS CloudTrail. Pour cela, accordez à l'administrateur l'accès au rôle de réponse aux incidents sur un compte hors production et exécutez vos playbooks. Une fois que vous aurez terminé, vous pourrez créer une politique autorisant uniquement les mesures prises. Cette politique peut ensuite être appliquée à tous les rôles de réponse aux incidents dans tous les comptes. Vous pouvez éventuellement créer une politique IAM distincte pour chaque playbook afin de faciliter la gestion et la vérification. Les exemples de playbooks peuvent comprendre des plans d'intervention pour les rançongiciels, les atteintes à la protection des données, la perte d'accès à la production et d'autres scénarios.

Utilisez les comptes de réponse aux incidents pour assumer des rôles IAM d'intervention en cas d'incident dans d'autres Comptes AWS. Ces rôles doivent être configurés de façon à pouvoir être assumés uniquement par les utilisateurs du compte de sécurité et la relation de confiance doit exiger que le principal appelant ait été authentifié au moyen de l'authentification multifactorielle. Les rôles doivent utiliser des politiques IAM à portée limitée afin de contrôler l'accès. Veillez à ce que toutes les demandes de `AssumeRole` pour ces rôles soient enregistrées dans CloudTrail et fassent l'objet d'une alerte, et à ce que toutes les actions effectuées à l'aide de ces rôles soient enregistrées.

Il est vivement recommandé de nommer les comptes utilisateur et rôles IAM afin d'en faciliter la recherche dans les journaux CloudTrail. Un exemple serait de nommer les comptes IAM *<USER_ID>*`-BREAK-GLASS` et les rôles IAM `BREAK-GLASS-ROLE`.

CloudTrail est utilisé pour enregistrer l'activité des API dans vos comptes AWS et doit être utilisé pour configurer les alertes relatives à l'utilisation des rôles d'ntervention en cas d'incidents. Consultez la publication de blog sur la configuration des alertes lorsque les clés racine sont utilisées. Les instructions peuvent être modifiées pour configurer le filtre métrique Amazon CloudWatch afin de filtrer les événements `AssumeRole` liés au rôle IAM de réponse aux incidents :

```
{ $.eventName = "AssumeRole" && $.requestParameters.roleArn =
 "<INCIDENT_RESPONSE_ROLE_ARN>" && $.userIdentity.invokedBy NOT EXISTS && $.eventType !
= "AwsServiceEvent" }
```

Dans la mesure où les rôles de réponse aux incidents sont susceptibles d'avoir un niveau d'accès élevé, il est important que ces alertes soient transmises à un vaste groupe et qui y donnera suite rapidement.

Lors d'un incident, il est possible qu'un intervenant ait besoin d'accéder à des systèmes qui ne sont pas sécurisés directement par . Celle-ci peut inclure des instances Amazon Elastic Compute Cloud, des bases de données Amazon Relational Database Service ou des plateformes de logiciel en tant que service (SaaS). Il est fortement recommandé d'utiliser AWS Systems Manager Session Manager pour tous les accès administratifs aux instances Amazon EC2 plutôt que d'utiliser les protocoles natifs tels que SSH ou RDP. Cet accès peut être contrôlé à l'aide d'IAM, qui est sécurisé et vérifié. Il est également possible d'automatiser certaines parties de vos playbooks à l'aide des documents AWS Systems Manager Run Command, qui peuvent réduire les erreurs des utilisateurs et accélérer le temps de restauration. Pour accéder aux bases de données et aux outils tiers, nous recommandons de stocker les informations d'identification dans AWS Secrets Manager et d'accorder l'accès aux rôles des intervenants en cas d'incident.

Enfin, la gestion des comptes IAM de réponse aux incidents doit être ajoutée à vos processus Joiners, Movers et Leavers et revue et testée périodiquement pour vérifier que seul l'accès prévu est autorisé.

Ressources

Documents connexes :

- Gérer les accès temporaires à votre environnement AWS
- Guide d'intervention en cas d'incident de sécurité AWS
- AWS Elastic Disaster Recovery
- AWS Systems Manager Incident Manager
- Définition d'une politique de mot de passe du compte pour les utilisateurs IAM
- Utilisation de l'authentification multifactorielle (MFA) dans l'interface AWS
- Configuration d'accès inter-compte pour MFA
- Utilisation de l'analyseur d'accès IAM pour générer des politiques IAM
- Best Practices for AWS Organizations Service Control Policies in a Multi-Account Environment
- Comment recevoir des notifications lorsque les clés d'accès racine de votre AWS sont utilisées
- Create fine-grained session permissions using IAM managed policies
- Accès en mode « bris de glace »

Vidéos connexes :

- Automatisation de la réponse aux incidents et de l'analyse poussée dans AWS
- DIY guide to runbooks, incident reports, and incident response
- Prepare for and respond to security incidents in your AWS environment

Exemples connexes :

- Atelier : AWS Account Setup and Root User
- Atelier : Incident Response with AWS Console and CLI

SEC10-BP06 Outils de pré-déploiement

Vérifiez que le personnel de sécurité dispose des outils appropriés préalablement déployés pour accélérer l'enquête jusqu'à la récupération.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Pour automatiser les fonctions opérationnelles et de réponse en matière de sécurité, vous pouvez utiliser un ensemble complet APIs d'outils issus de AWS. Vous pouvez automatiser entièrement la gestion des identités, la sécurité des réseaux, la protection des données et les fonctionnalités de surveillance, et les mettre en œuvre en utilisant les méthodes de développement de logiciel les plus courantes que vous avez déjà mises en place. Lorsque vous automatisez la sécurité, votre système peut surveiller, examiner et déclencher une réponse, plutôt que d'avoir à demander à des personnes de surveiller votre niveau de sécurité et de réagir manuellement aux événements.

Si vos équipes de réponse aux incidents continuent de répondre aux alertes de la même manière, elles risquent de se lasser des alertes. Au fil du temps, l'équipe peut faire moins attention aux alertes et soit faire des erreurs en gérant des situations ordinaires, soit manquer des alertes inhabituelles. L'automatisation permet d'éliminer la lassitude liée aux alertes en utilisant des fonctions qui traitent les alertes répétitives et ordinaires, laissant aux personnes le soin de gérer les incidents sensibles et uniques. L'intégration de systèmes de détection des anomalies, tels qu'Amazon GuardDuty, AWS CloudTrail Insights et Amazon CloudWatch Anomaly Detection, peut réduire le fardeau des alertes courantes basées sur des seuils.

Vous pouvez améliorer les processus manuels en automatisant par programmation les étapes du processus. Une fois que vous avez défini le modèle de correction d'un événement, vous pouvez le décomposer en logique exploitable et écrire le code pour exécuter cette logique. Les intervenants peuvent ensuite exécuter ce code pour corriger le problème. Au fil du temps, vous pouvez automatiser un nombre croissant d'étapes et, enfin, gérer automatiquement des catégories entières d'incidents courants.

Au cours d'une enquête de sécurité, vous devez être en mesure d'examiner les journaux pertinents pour consigner et comprendre la portée et la chronologie complètes de l'incident. Des journaux sont également requis pour la génération d'alertes, indiquant que certaines actions intéressantes ont eu lieu. Il est essentiel de sélectionner, d'activer, de stocker et de configurer les mécanismes d'interrogation et de récupération et de configurer les alertes. En outre, une solution efficace qui fournit des outils de recherche dans les données des journaux est Amazon Detective.

AWS propose plus de 200 services cloud et des milliers de fonctionnalités. Nous vous recommandons de passer en revue les services susceptibles de prendre en charge et de simplifier votre stratégie de réponse aux incidents.

Outre la journalisation, vous devez développer et mettre en œuvre une stratégie de balisage. Le balisage peut aider à mettre en contexte l'objectif d'une AWS ressource. Le balisage peut également être utilisé à des fins d’automatisation.

Étapes d’implémentation

Sélection et configuration de journaux à des fins d’analyse et d’alerte

Consultez la documentation suivante relative à la configuration de la journalisation pour la réponse aux incidents :

- Stratégies de journalisation pour l’intervention en cas d’incidents de sécurité
- SEC04-BP01 Configurer une journalisation de service et d’application

Permettre aux services de sécurité de prendre en charge la détection et l’intervention

AWS fournit des fonctionnalités natives de détection, de prévention et de réactivité, et d'autres services peuvent être utilisés pour concevoir des solutions de sécurité personnalisées. Pour obtenir la liste des services les plus pertinents en matière de réponse aux incidents de sécurité, consultez Définitions des fonctionnalités du cloud.

Élaboration et mise en œuvre d’une stratégie de marquage

Il peut être difficile d'obtenir des informations contextuelles sur le cas d'utilisation métier et les parties prenantes internes concernées par une AWS ressource. Pour ce faire, vous pouvez notamment utiliser des balises, qui attribuent des métadonnées à vos AWS ressources et consistent en une clé et une valeur définies par l'utilisateur. Vous pouvez créer des balises pour classer les ressources par objectif, propriétaire, environnement, type de données traitées et d’autres critères de votre choix.

Une stratégie de balisage cohérente peut accélérer les temps de réponse et minimiser le temps consacré au contexte organisationnel en vous permettant d'identifier et de discerner rapidement les informations contextuelles relatives à une ressource. AWS Les balises peuvent également servir de mécanisme pour initier l’automatisation des réponses. Pour plus de détails sur les éléments à étiqueter, consultez la section Marquage de vos AWS ressources. Vous devez d’abord définir les balises que vous souhaitez implémenter dans votre organisation. Ensuite, vous mettez en œuvre et appliquez cette stratégie de balisage. Pour plus de détails sur la mise en œuvre et l'application, voir

Implémenter une stratégie de balisage AWS des ressources à l'aide des politiques de AWS balises et des politiques de contrôle des services (SCPs).

Ressources

Bonnes pratiques Well-Architected connexes :

- SEC04-BP01 Configurer une journalisation de service et d'application
- SEC04-BP02 Capturer les journaux, les résultats et les métriques dans des emplacements standardisés

Documents connexes :

- Stratégies de journalisation pour l'intervention en cas d'incidents de sécurité
- Définitions des fonctionnalités cloud de réponse aux incidents

Exemples connexes :

- Détection des menaces et réponse avec Amazon GuardDuty et Amazon Detective
- Atelier Security Hub
- Gestion des vulnérabilités avec Amazon Inspector

SEC10-BP07 Exécuter des simulations

À mesure que les organisations se développent et évoluent au fil du temps, le paysage des menaces change. Il est donc important de revoir en permanence vos capacités de réponse aux incidents. L'organisation de simulations (également appelées « tests de simulation de panne ») est une méthode qui peut être utilisée pour effectuer cette évaluation. Les simulations utilisent des scénarios d'événements de sécurité réels conçus pour imiter les tactiques, techniques et procédures (TTP) d'un acteur de la menace et permettre à une organisation d'exercer et d'évaluer ses capacités de réponse aux incidents en réagissant à ces cyberévénements fictifs tels qu'ils peuvent se produire dans la réalité.

Avantages liés au respect de cette bonne pratique : les simulations présentent de nombreux avantages :

- Validation de l'état de préparation à la cybersécurité et renforcement de la confiance de vos intervenants en cas d'incident.

- Test de la précision et de l’efficacité des outils et des flux de travail.
- Amélioration des méthodes de communication et de remontées en fonction de votre plan d’intervention en cas d’incident.
- Possibilité de répondre à des vecteurs moins courants.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Il existe trois principaux types de simulations :

- Exercices sur table : l’approche théorique des simulations est une session basée sur des discussions auxquelles participent les différentes parties prenantes de la réponse aux incidents afin de mettre en pratique leurs rôles et leurs responsabilités et d’utiliser des outils de communication et des manuels établis. L’animation d’exercices peut généralement être réalisée en une journée complète dans un lieu virtuel, un lieu physique ou une combinaison des deux. Dans la mesure où il repose sur la discussion, l’exercice théorique met l’accent sur les processus, les personnes et la collaboration. La technologie fait partie intégrante de la discussion, mais l’utilisation effective d’outils ou de scripts de réponse aux incidents ne fait généralement pas partie de l’exercice théorique.
- Exercices de l’équipe violette : les exercices de l’équipe violette augmentent le niveau de collaboration entre les intervenants en cas d’incident (équipe bleue) et les acteurs de menaces simulées (équipe rouge). L’équipe bleue est composée de membres du centre des opérations de sécurité (SOC), mais peut également inclure d’autres parties prenantes qui seraient impliquées lors d’un véritable cyberévénement. L’équipe rouge est composée d’une équipe de tests de pénétration ou de parties prenantes clés formées à la sécurité offensive. L’équipe rouge travaille en collaboration avec les animateurs de l’exercice lors de la conception d’un scénario afin que celui-ci soit précis et réalisable. Lors des exercices de l’équipe violette, l’accent est principalement mis sur les mécanismes de détection, les outils et les procédures opérationnelles standard (SOP) qui soutiennent les efforts de réponse aux incidents.
- Exercces de l’équipe rouge : au cours d’un exercice de l’équipe rouge, l’attaque (l’équipe rouge) effectue une simulation pour atteindre un objectif donné ou un ensemble d’objectifs à partir d’une portée prédéterminée. Les défenseurs (équipe bleue) ne seront pas nécessairement au courant de la portée ni de la durée de l’exercice, ce qui permet d’évaluer de manière plus réaliste la manière dont ils réagiraient en cas d’incident réel. Étant donné que les exercices de l’équipe rouge peuvent être des tests invasifs, soyez prudent et mettez en œuvre des contrôles pour vérifier que l’exercice ne cause pas de dommages réels à votre environnement.

Envisagez d'animer des simulations cybernétiques à intervalles réguliers. Chaque type d'exercice peut apporter des avantages uniques aux participants et à l'organisation dans son ensemble. Vous pouvez donc choisir de commencer par des types de simulation moins complexes (tels que des exercices théoriques) et de passer ensuite à des types de simulation plus complexes (exercices de l'équipe rouge). Vous devez sélectionner un type de simulation en fonction de la maturité de votre sécurité, de vos ressources et des résultats souhaités. Certains clients peuvent décider de ne pas effectuer les exercices de l'équipe rouge en raison de leur complexité et de leur coût.

Étapes d'implémentation

Quel que soit le type de simulation que vous choisissez, les simulations suivent généralement les étapes de mise en œuvre suivantes :

1. Définir les éléments essentiels de l'exercice : définissez le scénario de simulation et les objectifs de la simulation. Les deux doivent être acceptés par les dirigeants.
2. Identifier les principales parties prenantes : un exercice nécessite au minimum des animateurs et des participants. Selon le scénario, d'autres parties prenantes telles que les services juridiques, l'équipe de communication ou la direction, peuvent être impliquées.
3. Concevoir et tester le scénario : le scénario devra peut-être être redéfini au fur et à mesure de sa création si des éléments spécifiques ne sont pas réalisables. Un scénario finalisé est attendu à l'issue de cette étape.
4. Faciliter la simulation : le type de simulation détermine l'animation utilisée (un scénario papier par rapport à un scénario simulé hautement technique). Les animateurs doivent adapter leurs tactiques d'animation aux objectifs de l'exercice et impliquer tous les participants dans l'exercice dans la mesure du possible afin d'en tirer le meilleur parti.
5. Élaborer le rapport après action (AAR) : identifier les domaines qui se sont bien déroulés, ceux qui peuvent être améliorés et les lacunes potentielles. L'AAR doit mesurer l'efficacité de la simulation ainsi que la réponse de l'équipe à l'événement simulé afin que les progrès puissent être suivis au fil du temps lors de futures simulations.

Ressources

Documents connexes :

- Réponse aux incidents dans AWS

Vidéos connexes :

- AWS GameDay - Security Edition
- Exécution de simulations de réponses efficaces aux incidents de sécurité

SEC10-BP08 Mettre en place un cadre pour tirer les leçons des incidents

La mise en œuvre d'un cadre de leçons apprises et d'une capacité d'analyse des causes profondes permettra non seulement d'améliorer les capacités de réponse aux incidents, mais aussi d'éviter que l'incident ne se reproduise. En tirant les leçons de chaque incident, vous pouvez éviter de répéter les mêmes erreurs, expositions ou erreurs de configuration, non seulement en améliorant votre posture de sécurité, mais également en réduisant le temps perdu dans des situations évitables.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Il est important de mettre en œuvre un cadre des leçons apprises qui établit et atteint, à un niveau élevé, les points suivants :

- Quand se déroule un processus des enseignements tirés ?
- En quoi consiste le processus des enseignements tirés ?
- Comment se déroule un processus des enseignements tirés ?
- Qui est impliqué dans le processus et comment ?
- Comment les domaines à améliorer seront-ils identifiés ?
- Comment allez-vous vérifier que les améliorations sont suivies et mises en œuvre de manière efficace ?

Le cadre ne doit pas se concentrer sur les individus ni les blâmer, mais doit plutôt se concentrer sur l'amélioration des outils et des processus.

Étapes d'implémentation

Outre les résultats de haut niveau énumérés ci-dessus, il est important de poser les bonnes questions afin de tirer le meilleur parti (informations menant à des améliorations réalisables) du processus. Posez-vous les questions suivantes pour commencer à développer vos discussions sur les enseignements tirés :

- Quel a été l'incident ?

- Quand l’incident a-t-il été identifié pour la première fois ?
- Comment a-t-il été identifié ?
- Quels systèmes ont alerté sur l’activité ?
- Quels systèmes, services et données étaient concernés ?
- Que s’est-il passé précisément ?
- Qu’est-ce qui a bien fonctionné ?
- Qu’est-ce qui n’a pas bien fonctionné ?
- Quels processus ou procédures ont échoué ou n’ont pas pu être mis à l’échelle pour répondre à l’incident ?
- Qu’est-ce qui peut être amélioré dans les domaines suivants :
  - Personnes
    - Les personnes à contacter étaient-elles réellement disponibles et la liste de contacts était-elle à jour ?
    - Les personnes manquaient-elles de formation ou n’avaient-elles pas les capacités nécessaires pour intervenir et enquêter efficacement sur l’incident ?
    - Les ressources appropriées étaient-elles prêtes et disponibles ?
  - Processus
    - Les processus et procédures ont-ils été suivis ?
    - Les processus et procédures étaient-ils documentés et disponibles pour cet incident ou ce type d’incident ?
    - Les processus et procédures requis étaient-ils absents ?
    - Les intervenants ont-ils pu accéder en temps opportun aux informations requises pour répondre au problème ?
  - Technologie
    - Les systèmes d’alerte existants ont-ils identifié l’activité et ont-ils envoyé des alertes efficaces ?
    - Comment aurions-nous pu le réduire time-to-detection de 50 % ?
    - Les alertes existantes doivent-elles être améliorées ou de nouvelles alertes doivent-elles être créées pour cet incident ou ce type d’incident ?
    - Les outils existants ont-ils permis d’enquêter efficacement (recherche/analyse) sur l’incident ?
    - Que peut-on faire pour identifier cet incident ou ce type d’incident plus rapidement ?
    - Que peut-on faire pour éviter que cet incident ou ce type d’incident ne se reproduise ?

- À qui appartient le plan d'amélioration et comment allez-vous vérifier qu'il a été mis en œuvre ?
- Quel est le calendrier des contrôles et processus de surveillance ou de prévention supplémentaires à mettre en œuvre et à tester ?

Cette liste n'est pas exhaustive, mais vise à servir de point de départ pour identifier les besoins de l'organisation et de l'entreprise et la manière dont vous pouvez les analyser afin de tirer les meilleurs enseignements des incidents et d'améliorer en permanence votre posture de sécurité. Le plus important est de commencer par intégrer les enseignements tirés dans le cadre standard de votre processus de réponse aux incidents, de la documentation et des attentes des parties prenantes.

Ressources

Documents connexes :

- Guide de réponse aux incidents de sécurité – Établir un cadre pour tirer des enseignements des incidents (langue française non garantie)AWS
- NCSCCAFconseils - Leçons apprises

Sécurité des applications

Question

- SÉC 11. Comment intégrer et valider les propriétés de sécurité des applications tout au long du cycle de vie de la conception, du développement et du déploiement ?

SÉC 11. Comment intégrer et valider les propriétés de sécurité des applications tout au long du cycle de vie de la conception, du développement et du déploiement ?

La formation du personnel, le test à l'aide de l'automatisation, la compréhension des dépendances et la validation des propriétés de sécurité des outils et des applications contribuent à réduire la probabilité de problèmes de sécurité dans les charges de travail de production.

Bonnes pratiques

- SEC11-BP01 Formation à la sécurité des applications
- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication
- SEC11-BP03 Effectuer des tests de pénétration réguliers

- SEC11-BP04 Mener des examens de code
- SEC11-BP05 Centralisation des services pour les packages et les dépendances
- SEC11-BP06 Déploiement programmatique de logiciels
- SEC11-BP07 Évaluation régulière des caractéristiques de sécurité des pipelines
- SEC11-BP08 Création d'un programme permettant aux équipes responsables de la charge de travail de s'approprier les mécanismes de sécurité

SEC11-BP01 Formation à la sécurité des applications

Offrez à votre équipe une formation sur les pratiques de développement et d'exploitation sécurisées, afin de l'aider à créer des logiciels sécurisés et de haute qualité. Cette pratique aide votre équipe à prévenir, détecter et corriger les problèmes de sécurité à un stade précoce du cycle de développement. Envisagez une formation qui couvre la modélisation des menaces, les pratiques de codage sécurisé et l'utilisation des services pour des configurations et des opérations sécurisées. Donnez à votre équipe un accès à la formation par le biais de ressources en libre-service et recueillez régulièrement leurs commentaires en vue de son amélioration continue.

Résultat escompté : vous dotez votre équipe des connaissances et des compétences nécessaires pour concevoir et créer des logiciels en tenant compte de la sécurité dès le départ. Grâce à une formation sur la modélisation des menaces et les pratiques de développement sécurisé, votre équipe possède une connaissance approfondie des risques de sécurité potentiels et comprend mieux comment les atténuer au cours du cycle de développement logiciel (SDLC). Cette approche proactive de la sécurité fait partie de la culture de votre équipe et vous permet d'identifier et de résoudre rapidement les problèmes de sécurité potentiels. Par conséquent, votre équipe fournit des logiciels et des fonctionnalités sécurisés et de haute qualité de manière plus efficace, ce qui accélère le délai de livraison global. Vous avez une culture collaborative et inclusive de la sécurité au sein de votre organisation, et la responsabilité de la sécurité est partagée entre tous ses acteurs.

Anti-modèles courants :

- Vous attendez un examen de la sécurité pour tenir compte des propriétés de sécurité d'un système.
- Vous laissez toutes les décisions en matière de sécurité à une équipe de sécurité centrale.
- Vous ne communiquez pas sur la manière dont les décisions prises au cours du cycle de développement logiciel sont liées aux attentes ou aux politiques générales de l'organisation en matière de sécurité.

- Vous effectuez trop tard le processus d'examen de la sécurité.

Avantages liés au respect de cette bonne pratique :

- Meilleure connaissance des exigences organisationnelles en matière de sécurité dès le début du cycle de développement.
- Possibilité d'identifier les problèmes de sécurité potentiels et d'y remédier plus rapidement, ce qui se traduit par une mise à disposition plus rapide des fonctionnalités.
- Amélioration de la qualité des logiciels et des systèmes.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Pour créer des logiciels sécurisés et de haute qualité, formez votre équipe aux pratiques courantes de développement et d'exploitation sécurisés des applications. Cette pratique peut aider votre équipe à prévenir, détecter et corriger les problèmes de sécurité plus tôt dans le cycle de développement, ce qui peut raccourcir le délai de livraison.

Pour mettre en œuvre cette pratique, envisagez de former votre équipe à la modélisation des menaces à l'aide de ressources AWS, telles que l'atelier sur la modélisation des menaces. La modélisation des menaces peut aider votre équipe à comprendre les risques de sécurité potentiels et à concevoir des systèmes en tenant compte de la sécurité dès le départ. En outre, vous pouvez fournir un accès à une formation AWS Training and Certification, du secteur ou destinée aux partenaires AWS sur les pratiques de développement sécurisées. Pour plus de détails sur une approche globale de la conception, du développement, de la sécurisation et de l'exploitation efficace à grande échelle, consultez le Guide AWS DevOps.

Définissez et communiquez clairement le processus d'examen de la sécurité de votre organisation et définissez les responsabilités de votre équipe, de l'équipe chargée de la sécurité et des autres parties prenantes. Publiez des conseils en libre-service, des exemples de code et des modèles illustrant comment répondre à vos exigences en matière de sécurité. Vous pouvez utiliser des services AWS tels que AWS CloudFormation, les constructs AWS Cloud Development Kit (AWS CDK) (AWS CDK) et Service Catalog pour fournir des configurations préapprouvées et sécurisées et réduire les besoins en configurations personnalisées.

Recueillez régulièrement les commentaires de votre équipe sur son expérience du processus d'examen de la sécurité et de la formation, et mettez à profit ces commentaires pour vous améliorer

en permanence. Organisez des tests de simulation de panne ou des campagnes de lutte contre les bogues pour identifier et résoudre les problèmes de sécurité tout en améliorant les compétences de votre équipe.

Étapes d'implémentation

1. Identifier les besoins de formation : évaluez le niveau de compétence actuel et les lacunes en matière de connaissances au sein de votre équipe en ce qui concerne les pratiques de développement sécurisées par le biais d'enquêtes, d'examens de code ou de discussions avec les membres de l'équipe.
2. Planifier la formation : sur la base des besoins identifiés, créez un plan de formation qui couvre des sujets pertinents tels que la modélisation des menaces, les pratiques de codage sécurisé, les tests de sécurité et les pratiques de déploiement sécurisé. Utilisez des ressources telles que l'atelier sur la modélisation des menaces, AWS Training and Certification et les programmes de formation destinés au secteur ou aux partenaires AWS.
3. Planifier et offrir des formations : planifiez des ateliers ou des sessions de formation réguliers pour votre équipe. Ils peuvent être dispensés par un instructeur ou à un rythme personnalisé, selon les préférences et la disponibilité de votre équipe. Encouragez les exercices pratiques et les exemples pratiques pour renforcer l'apprentissage.
4. Définir un processus d'examen de la sécurité : collaborez avec votre équipe de sécurité et les autres parties prenantes pour définir clairement le processus d'examen de la sécurité pour vos applications. Documentez les responsabilités de chaque équipe ou personne impliquée dans ce processus, y compris votre équipe de développement, votre équipe de sécurité et les autres parties prenantes concernées.
5. Créer des ressources en libre-service : développez des recommandations, des exemples de code et des modèles en libre-service pour illustrer comment répondre aux exigences de votre entreprise en matière de sécurité. Envisagez d'utiliser des services AWS tels que CloudFormation, les constructs AWS CDK et Service Catalog pour fournir des configurations préapprouvées et sécurisées et réduire les besoins en configurations personnalisées.
6. Communiquer et partager : communiquez efficacement à votre équipe le processus d'examen de la sécurité et les ressources en libre-service disponibles. Organisez des ateliers ou des sessions de formation pour familiariser votre équipe à ces ressources et vérifiez qu'elle comprend comment les utiliser.
7. Recueillir des commentaires et s'améliore : collectez régulièrement les commentaires de votre équipe sur son expérience du processus d'examen de la sécurité et de la formation. Utilisez ces

commentaires pour identifier les domaines à améliorer et affiner en permanence les supports de formation, les ressources en libre-service et le processus d'examen de la sécurité.

8. Réaliser des exercices de sécurité : organisez des tests de simulation de panne ou des campagnes de lutte contre les bogues pour identifier et résoudre les problèmes de sécurité au sein de vos applications. Ces exercices permettent non seulement de découvrir des vulnérabilités potentielles, mais constituent également des opportunités d'apprentissage pratique pour votre équipe, visant à améliorer ses compétences en matière de développement et d'exploitation sécurisés.
9. Continuer à apprendre et à s'améliorer : encouragez votre équipe à rester au fait des derniers outils, techniques et pratiques de développement sécurisé. Passez en revue et mettez à jour régulièrement vos supports et ressources de formation afin de refléter l'évolution du contexte et des bonnes pratiques de sécurité.

Ressources

Bonnes pratiques associées :

- SEC11-BP08 Création d'un programme permettant aux équipes responsables de la charge de travail de s'approprier les mécanismes de sécurité

Documents connexes :

- AWS Training et la certification
- Comment envisager la gouvernance de la sécurité dans le cloud
- Comment aborder la modélisation des menaces
- Accélérer l'entraînement – The AWS Skills Guild
- Sagas AWS DevOps

Vidéos connexes :

- Proactive security: Considerations and approaches

Exemples connexes :

- Atelier sur la modélisation des menaces
- Sensibilisation des développeurs à l'industrie

Services connexes :

- AWS CloudFormation
- AWS Cloud Development Kit (AWS CDK) (AWS CDK) Constructions
- Service Catalog

SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication

Automatisez les tests des propriétés de sécurité tout au long du cycle de développement et de publication. L'automatisation facilite l'identification systématique et répétée des problèmes potentiels dans les logiciels avant leur diffusion, ce qui réduit le risque de problèmes de sécurité dans les logiciels fournis.

Résultat escompté : l'objectif des tests automatisés est de fournir un moyen programmatique de détecter les problèmes potentiels à un stade précoce et fréquent tout au long du cycle de développement. Lorsque vous automatisez les tests de régression, vous pouvez exécuter à nouveau les tests fonctionnels et non fonctionnels pour vérifier que le logiciel testé précédemment fonctionne toujours comme prévu après une modification. Lorsque vous définissez des tests d'unités de sécurité pour vérifier les erreurs de configuration courantes, telles qu'une authentification défaillante ou manquante, vous pouvez identifier et résoudre ces problèmes dès le début du processus de développement.

L'automatisation des tests utilise des cas de test spécifiques pour la validation de l'application, sur la base des exigences de l'application et de la fonctionnalité souhaitée. Le résultat du test automatisé est basé sur la comparaison entre le résultat du test généré et le résultat attendu, ce qui accélère le cycle de vie global du test. Les méthodologies de test telles que les tests de régression et les suites de tests d'unités sont les mieux adaptées à l'automatisation. L'automatisation des tests des propriétés de sécurité permet aux concepteurs de recevoir des commentaires automatisés sans avoir à attendre un examen de sécurité. Les tests automatisés sous forme d'analyse statique ou dynamique du code peuvent améliorer la qualité du code et aider à détecter les problèmes logiciels potentiels dès le début du cycle de développement.

Anti-modèles courants :

- Ne pas communiquer les cas de test et les résultats des tests automatisés.
- Effectuer uniquement les tests automatisés juste avant la mise en production.
- Automatiser les cas de test avec des exigences qui changent fréquemment.
- Ne pas fournir de recommandations sur la manière de traiter les résultats des tests de sécurité.

Avantages liés au respect de cette bonne pratique :

- Réduction de la dépendance à l'égard des personnes qui évaluent les propriétés de sécurité des systèmes.
- Le fait de disposer de résultats cohérents dans plusieurs domaines de travail améliore la cohérence.
- Réduction de la probabilité d'introduire des problèmes de sécurité dans les logiciels de production.
- Un délai plus court entre la détection et la remédiation grâce à une détection plus précoce des problèmes logiciels.
- Visibilité accrue des comportements systémiques ou répétés dans plusieurs domaines de travail, ce qui peut être utilisé pour apporter des améliorations à l'échelle de l'organisation.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Au fur et à mesure du développement de votre logiciel, adoptez divers mécanismes de test pour vous assurer que vous testez votre application à la fois pour les exigences fonctionnelles, basées sur la logique commerciale de votre application, et pour les exigences non fonctionnelles, qui sont axées sur la fiabilité, la performance et la sécurité de l'application.

Les tests statiques de sécurité des applications (SAST) analysent votre code source à la recherche de schémas de sécurité anormaux et fournissent des indications sur le code sujet aux défauts. Les tests SAST s'appuient sur des données statiques, telles que la documentation (spécifications des exigences, documentation de conception et spécifications de conception) et le code source de l'application, pour tester une série de problèmes de sécurité connus. Les analyseurs de code statique permettent d'accélérer l'analyse de gros volumes de code. NIST Quality Group propose une comparaison des analyseurs de sécurité du code source, qui inclut des outils open source pour les scanners de code d'octets et les scanners de code binaire.

Complétez vos tests statiques par des méthodes de sécurité des applications (DAST), qui consistent à effectuer des tests sur l'application en cours d'exécution afin d'identifier les comportements potentiellement inattendus. Les tests dynamiques peuvent détecter des problèmes potentiels qui ne sont pas détectables par l'analyse statique. Les tests effectués aux stades du référentiel de code, de la build et du pipeline vous permettent de vérifier différents types de problèmes potentiels avant qu'ils ne s'introduisent dans votre code. Amazon Q Developer fournit des recommandations de

code, y compris des analyses de sécurité, dans l'IDE du générateur. La sécurité Amazon CodeGuru peut identifier les problèmes critiques, les problèmes de sécurité et les bogues difficiles à détecter lors du développement d'applications, et fournit des recommandations pour améliorer la qualité du code. L'extraction de la nomenclature logicielle (SBOM) vous permet également d'extraire un enregistrement formel contenant les détails et les relations des différents composants utilisés dans la création de votre logiciel. Cela vous permet d'informer la gestion des vulnérabilités et d'identifier rapidement les dépendances des logiciels ou des composants, et les risques liés à la chaîne d'approvisionnement.

L'atelier Security for Developers utilise des outils de développement AWS, tels que AWS CodeBuild, AWS CodeCommit et AWS CodePipeline, pour l'automatisation du pipeline de versions, qui incluent les méthodologies de test SAST et DAST.

Au fur et à mesure que vous progressez dans votre cycle de développement du logiciel, mettez en place un processus itératif qui comprend des révisions périodiques des applications avec votre équipe de sécurité. Les commentaires recueillis lors de ces examens de sécurité doivent être traités et validés dans le cadre de l'examen de l'état de préparation à la mise en production. Ces examens permettent de définir un solide niveau de sécurité des applications et fournissent aux concepteurs des commentaires exploitables pour résoudre les problèmes potentiels.

Étapes d'implémentation

- Implémentez des outils cohérents d'IDE, d'examen de code et de CI/CD qui incluent des tests de sécurité.
- Réfléchissez à l'étape du cycle de développement du logiciel où il convient de bloquer les pipelines au lieu de simplement avertir les concepteurs que des problèmes doivent être résolus.
- Automated Security Helper (ASH) est un exemple d'outil d'analyse de sécurité de code open source.
- La réalisation de tests ou d'analyses de code à l'aide d'outils automatisés, tels qu'Amazon Q Developer intégré aux IDE pour développeurs et la sécurité Amazon CodeGuru pour l'analyse du code lors de la validation, permet aux créateurs d'obtenir des commentaires au bon moment.
- Lorsque vous créez avec AWS Lambda, Amazon Inspector peut vous permettre de scanner le code de l'application dans vos fonctions.
- Lorsque les tests automatisés sont inclus dans les pipelines CI/CD, vous devez utiliser un système de tickets pour suivre la notification et la résolution des problèmes logiciels.
- Pour les tests de sécurité susceptibles de donner lieu à des conclusions, un lien vers des conseils pour remédier à la situation aide les concepteurs à améliorer la qualité du code.

- Analysez régulièrement les résultats des outils automatisés afin de donner la priorité à la prochaine automatisation, à la formation des concepteurs ou à la campagne de sensibilisation.
- Pour extraire la nomenclature logicielle dans le cadre de vos pipelines CI/CD, utilisez Amazon Inspector SBOM Generator pour produire des nomenclatures logicielles pour les archives, les images de conteneur, les répertoires, les systèmes locaux et les fichiers binaires Go et Rust compilés au format SBOM CycloneDX.

Ressources

Bonnes pratiques associées :

- Guide DevOps : DL.CR.3 Établissement de critères d'achèvement clairs pour les tâches liées au code

Documents connexes :

- Livraison et déploiement continus

- Partenaires disposant de la compétence AWS DevOps
- Partenaires disposant de compétences en sécurité AWS pour la sécurité des applications
- Choisir une approche CI/CD Well-Architected
- Détection de secrets dans la sécurité Amazon CodeGuru
- Bibliothèque de détection de la sécurité Amazon CodeGuru
- Accelerate deployments on AWS with effective governance
- Comment AWS automatise les déploiements en toute sécurité et sans intervention
- Comment la sécurité Amazon CodeGuru vous aide à équilibrer efficacement la sécurité et la rapidité

Vidéos connexes :

- Hands-off: Automating continuous delivery pipelines at Amazon
- Automating cross-account CI/CD pipelines
- Processus de développement logiciel chez Amazon
- Tests des logiciels et des systèmes chez Amazon

Exemples connexes :

- Sensibilisation des développeurs à l'industrie
- Automated Security Helper (ASH)
- Gouvernance AWS CodePipeline – GitHub

SEC11-BP03 Effectuer des tests de pénétration réguliers

Effectuez régulièrement des tests de pénétration de votre logiciel. Ce mécanisme permet d'identifier les problèmes logiciels potentiels impossibles à détecter par des tests automatisés ou une révision manuelle du code. Il peut également vous permettre de comprendre l'efficacité de vos contrôles de détection. Les tests de pénétration doivent tenter de déterminer si le logiciel peut être amené à fonctionner de manière inattendue, par exemple en exposant des données qui devraient être protégées ou en accordant des autorisations plus étendues que prévu.

Résultat escompté : les tests de pénétration sont utilisés pour détecter, corriger et valider les propriétés de sécurité de votre application. Des tests de pénétration réguliers et planifiés doivent être effectués dans le cadre du cycle de développement du logiciel (SDLC). Les résultats des tests de pénétration doivent être pris en compte avant le lancement du logiciel. Vous devez analyser les résultats des tests de pénétration pour déterminer s'il existe des problèmes qui pourraient être détectés grâce à l'automatisation. Le fait de disposer d'un processus de test de pénétration régulier et reproductible, qui comprend un mécanisme de commentaires actif, permet d'éclairer les conseils donnés aux concepteurs et d'améliorer la qualité des logiciels.

Anti-modèles courants :

- Les tests de pénétration ne concernent que les problèmes de sécurité connus ou répandus.
- Tests de pénétration d'applications sans outils et bibliothèques tiers dépendants.
- Uniquement des tests de pénétration pour les problèmes de sécurité des packages, et non l'évaluation de la logique métier implémentée.

Avantages liés au respect de cette bonne pratique :

- Confiance accrue dans les propriétés de sécurité du logiciel avant sa diffusion.
- Possibilité d'identifier des modèles d'application privilégiés, ce qui permet d'améliorer la qualité des logiciels.

- Une boucle de rétroaction permettant d'identifier plus tôt dans le cycle de développement où l'automatisation ou une formation supplémentaire peuvent améliorer les propriétés de sécurité des logiciels.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Le test de pénétration est un exercice de test de sécurité structuré dans lequel vous exécutez des scénarios de faille de sécurité planifiés afin de détecter des problèmes, d'y remédier et de valider les contrôles de sécurité. Les tests de pénétration commencent par une reconnaissance, au cours de laquelle des données sont recueillies sur la base de la conception actuelle de l'application et de ses dépendances. Une liste de scénarios de test spécifiques à la sécurité est élaborée et exécutée. L'objectif principal de ces tests est de découvrir les problèmes de sécurité de votre application, qui pourraient être exploités pour obtenir un accès involontaire à votre environnement ou un accès non autorisé aux données. Vous devez effectuer des tests de pénétration lorsque vous lancez de nouvelles fonctionnalités, ou chaque fois que votre application a subi des changements majeurs en matière de fonction ou d'implémentation technique.

Vous devez identifier l'étape la plus appropriée du cycle de développement pour effectuer des tests de pénétration. Ces tests doivent avoir lieu suffisamment tard pour que la fonctionnalité du système soit proche de l'état final prévu, mais avec suffisamment de temps pour remédier aux éventuels problèmes.

Étapes d'implémentation

- Disposez d'un processus structuré pour définir le périmètre des tests de pénétration. Baser ce processus sur le modèle de menace est un bon moyen de maintenir le contexte.
- Identifiez l'endroit approprié dans le cycle de développement pour effectuer des tests de pénétration. Ce délai doit être respecté lorsque les changements attendus dans l'application sont minimes, mais qu'il reste suffisamment de temps pour mettre en œuvre des mesures correctives.
- Formez vos créateurs sur ce qu'il faut attendre des résultats des tests de pénétration et sur la manière d'obtenir des informations sur les mesures correctives.
- Utilisez des outils pour accélérer le processus de test de pénétration en automatisant les tests courants ou reproductibles.
- Analysez les résultats des tests de pénétration afin d'identifier les problèmes de sécurité systémiques et utilisez ces données pour effectuer des tests automatisés supplémentaires et former en permanence les créateurs.

Ressources

Bonnes pratiques associées :

- SEC11-BP01 Formation à la sécurité des applications
- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication

Documents connexes :

- AWS Les tests de pénétration fournissent des conseils détaillés pour les tests de pénétration sur AWS
- Accélérez les déploiements AWS grâce à une gouvernance efficace
- AWS Security Competency Partners
- Modernisez votre architecture de tests d'intrusion sur AWS Fargate
- AWS Simulateur d'injection de défauts

Exemples connexes :

- Automatisez API les tests avec AWS CodePipeline (GitHub)
- Assistant de sécurité automatisé () GitHub

SEC11-BP04 Mener des examens de code

Mettez en œuvre des examens de code pour vérifier la qualité et la sécurité d'un logiciel en cours de développement. Les examens de code impliquent que des membres de l'équipe autres que l'auteur du code d'origine examinent le code pour détecter les problèmes et les vulnérabilités potentiels et vérifier le respect des normes et des bonnes pratiques de codage. Ce processus permet de détecter les erreurs, les incohérences et les failles de sécurité qui auraient pu être omises par le développeur d'origine. Utilisez des outils automatisés pour faciliter les examens de code.

Résultat escompté : vous incluez des examens de code pendant le développement afin d'améliorer la qualité du logiciel en cours d'écriture. Vous perfectionnez les membres moins expérimentés de l'équipe grâce aux enseignements identifiés lors de l'examen de code. Vous identifiez les opportunités d'automatisation et soutenez le processus d'examen de code à l'aide d'outils et de tests automatisés.

Anti-modèles courants :

- Vous n’effectuez pas d’examen de code avant le déploiement.
- La même personne écrit et examine le code.
- Vous n’utilisez pas d’automatisation ni d’outils pour assister ou orchestrer les examens de code.
- Vous ne formez pas les concepteurs à la sécurité des applications avant qu’ils procèdent à l’examen du code.

Avantages liés au respect de cette bonne pratique :

- Amélioration de la qualité du code.
- Amélioration de la cohérence de développement du code grâce à la réutilisation d’approches communes.
- Réduction du nombre de problèmes découverts lors des tests de pénétration et des étapes ultérieures.
- Amélioration du transfert de connaissances au sein de l’équipe.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Les examens de code permettent de vérifier la qualité et la sécurité du logiciel au cours de son développement. Les examens manuels impliquent qu’un membres de l’équipe autre que l’auteur du code d’origine examine le code pour détecter les problèmes et les vulnérabilités potentiels et vérifie le respect des normes et des bonnes pratiques de codage. Ce processus permet de détecter les erreurs, les incohérences et les failles de sécurité qui auraient pu être omises par le développeur d’origine.

Envisagez d’utiliser la sécurité Amazon CodeGuru pour vous aider à effectuer des examens de code automatisés. La sécurité CodeGuru utilise le machine learning et un raisonnement automatisé pour analyser votre code et identifier les vulnérabilités de sécurité et les problèmes de codage potentiels. Intégrez des examens de code automatisés à vos référentiels de code existants et à vos pipelines d’intégration continue/de déploiement continu (CI/CD).

Étapes d’implémentation

1. Établissez un processus d’examen de code :
   - Définissez à quel moment les examens de code doivent avoir lieu, par exemple avant de fusionner le code dans la branche principale ou avant de le déployer en production.

- Déterminez qui doit participer au processus d'examen de code, par exemple les membres de l'équipe, les développeurs senior et les experts en sécurité.
- Décidez de la méthodologie d'examen de code, y compris du processus et des outils à utiliser.

2. Configurez les outils d'examen de code :
   - Évaluez et sélectionnez les outils d'examen de code qui répondent aux besoins de votre équipe, tels que les demandes d'extraction GitHub ou la sécurité CodeGuru.
   - Intégrez les outils choisis à vos référentiels de code et à vos pipelines CI/CD existants.
   - Configurez ces outils pour appliquer les exigences d'examen de code, telles que le nombre minimal de réviseurs et les règles d'approbation.
3. Définissez une liste de contrôle et des directives d'examen de code :
   - Créez une liste de contrôle ou des directives d'examen de code explicitant ce qui doit être examiné. Tenez compte de facteurs tels que la qualité du code, les vulnérabilités de sécurité, le respect des normes de codage et les performances.
   - Partagez la liste de contrôle ou les directives avec l'équipe de développement et vérifiez que tout le monde comprend les attentes.
4. Formez les développeurs aux bonnes pratiques d'examen de code :
   - Offrez à votre équipe une formation sur la manière de mener des examens de code efficaces.
   - Sensibilisez votre équipe aux principes de sécurité des applications et aux vulnérabilités courantes à rechercher lors des examens.
   - Encouragez le partage des connaissances et les sessions de programmation en binômes pour perfectionner les membres moins expérimentés de l'équipe.
5. Mettez en œuvre le processus d'examen de code :
   - Intégrez l'étape d'examen de code dans votre flux de travail de développement, par exemple en créant une demande d'extraction et en affectant des réviseurs.
   - Exigez que les modifications de code fassent l'objet d'un examen de code avant la fusion ou le déploiement.
   - Encouragez une communication ouverte et des commentaires constructifs pendant le processus d'examen.
6. Surveillez et améliorez :
   - Vérifiez régulièrement l'efficacité de votre processus d'examen de code et recueillez les commentaires de l'équipe.

- Identifiez les opportunités d'automatisation ou d'amélioration des outils afin de rationaliser le processus d'examen de code.
- Mettez à jour et affinez en permanence la liste de contrôle ou les directives d'examen de code en fonction des enseignements tirés et des bonnes pratiques du secteur.

7. Favorisez une culture d'examen de code :
   - Soulignez l'importance des examens de code pour maintenir la qualité et la sécurité du code.
   - Célébrez les réussites et les enseignements tirés du processus d'examen de code.
   - Favorisez un environnement de collaboration et de soutien dans lequel les développeurs se sentent à l'aise pour effectuer et recevoir des commentaires.

Ressources

Bonnes pratiques associées :

- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication

Documents connexes :

- Guide DevOps : DL.CR.2 Réalisation d'un examen par les pairs des modifications de code
- À propos des pull requests dans GitHub

Exemples connexes :

- Automatisation des examens de code avec la sécurité Amazon CodeGuru
- Automatisation de la détection des vulnérabilités de sécurité et des bogues dans les pipelines CI/CD à l'aide de la CLI de sécurité Amazon CodeGuru

Vidéos connexes :

- Amélioration continue de la qualité du code avec la sécurité Amazon CodeGuru

SEC11-BP05 Centralisation des services pour les packages et les dépendances

Fournissez des services centralisés pour permettre à vos équipes d'obtenir des packages logiciels et d'autres dépendances. Les packages peuvent ainsi être validés avant d'être inclus dans le logiciel

que vous écrivez. De plus, une source de données est disponible pour l'analyse des logiciels utilisés dans votre organisation.

Résultat escompté : vous créez votre charge de travail à partir de packages logiciels externes en plus du code que vous écrivez. Cela simplifie la mise en œuvre de fonctionnalités utilisées de manière répétée, telles qu'un analyseur JSON ou une bibliothèque de chiffrement. Vous centralisez les sources de ces packages et dépendances afin que votre équipe de sécurité puisse les valider avant leur utilisation. Vous utilisez cette approche en conjonction avec les flux de tests manuels et automatisés pour accroître la confiance dans la qualité du logiciel que vous développez.

Anti-modèles courants :

- Vous extrayez des packages de référentiels arbitraires sur Internet.
- Vous ne testez pas les nouveaux packages avant de les mettre à la disposition des créateurs.

Avantages liés au respect de cette bonne pratique :

- Meilleure compréhension des packages utilisés dans le logiciel en cours de création.
- Possibilité d'informer les équipes responsables de la charge de travail lorsqu'un package doit être mis à jour en fonction de la compréhension de qui utilise quoi.
- Réduire le risque qu'un package présentant des problèmes soit inclus dans votre logiciel.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Fournissez des services centralisés pour les packages et les dépendances d'une manière simple à utiliser pour les créateurs. Les services centralisés sont logiquement centraux plutôt que d'être implémentés sous la forme d'un système monolithique. Cette approche vous permet de fournir des services de manière à répondre aux besoins de vos concepteurs. Vous devez mettre en œuvre un moyen efficace d'ajouter des packages au référentiel lorsque des mises à jour sont effectuées ou que de nouvelles exigences apparaissent. Les services AWS tels que AWS CodeArtifact ou des solutions partenaires AWS similaires offrent cette fonctionnalité.

Étapes d'implémentation

- Implémentez un service de référentiel centralisé et logique, disponible dans tous les environnements où des logiciels sont développés.

- Prévoir l'accès au référentiel dans le cadre de la procédure d'attribution du Compte AWS.
- Concevez une automatisation pour tester les packages avant qu'ils ne soient publiés dans un référentiel.
- Conservez des métriques concernant les packages, les langages et les équipes les plus couramment utilisés et ayant subi le plus grand nombre de changements.
- Prévoyez un mécanisme automatisé permettant aux équipes de créateurs de demander de nouveaux packages et de fournir des commentaires.
- Analysez régulièrement les packages de votre référentiel afin d'identifier l'impact potentiel des problèmes récemment découverts.

Ressources

Bonnes pratiques associées :

- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication

Documents connexes :

- Guide DevOps : DL.CS.2 Signature des artefacts de code après chaque build
- Niveaux de la chaîne d'approvisionnement pour les artefacts logiciels (SLSA)

Exemples connexes :

- Accelerate deployments on AWS with effective governance
- Renforcez la sécurité de vos packages avec le kit d'outils CodeArtifact Package Origin Control
- Pipeline de publication de packages multi-régions (GitHub)
- Publication de modules Node.js sur AWS CodeArtifact à l'aide de AWS CodePipeline (GitHub)
- Exemple de pipeline Java CodeArtifact AWS CDK (GitHub)
- Distribuer des packages .NET NuGet privés avec CodeArtifact AWS (GitHub)

Vidéos connexes :

- Proactive security: Considerations and approaches
- The AWS Philosophy of Security (re:Invent 2017)

- When security, safety, and urgency all matter: Handling Log4Shell

SEC11-BP06 Déploiement programmatique de logiciels

Dans la mesure du possible, procédez à des déploiements de logiciels par programme. Cette approche réduit la probabilité qu'un déploiement échoue ou qu'une erreur humaine entraîne un problème inattendu.

Résultat escompté : la version de votre charge de travail que vous testez est la version que vous déployez, et le déploiement est effectué de manière cohérente à chaque fois. Vous externalisez la configuration de votre charge de travail, ce qui vous permet de déployer dans différents environnements sans modification. Vous utilisez la signature cryptographique de vos packages logiciels pour vérifier que rien ne change d'un environnement à l'autre.

Anti-modèles courants :

- Déploiement manuel d'un logiciel en production.
- Modification manuelle d'un logiciel pour l'adapter à des environnements différents.

Avantages liés au respect de cette bonne pratique :

- Confiance accrue dans le processus de lancement des logiciels.
- Réduction du risque que l'échec d'une modification affecte l'entreprise.
- Augmentation de la cadence de lancement en raison de la diminution du risque de changement.
- Capacité de restauration automatique en cas d'événements inattendus au cours du déploiement.
- Capacité à prouver par chiffrement que le logiciel testé est celui qui est déployé.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Pour maintenir une infrastructure d'applications robuste et fiable, mettez en œuvre des pratiques de déploiement sécurisées et automatisées. Cette pratique implique de supprimer l'accès humain persistant aux environnements de production, d'utiliser des outils CI/CD pour les déploiements et d'externaliser les données de configuration spécifiques à l'environnement. En suivant cette approche, vous pouvez améliorer la sécurité, réduire le risque d'erreurs humaines et rationaliser le processus de déploiement.

Vous pouvez créer votre structure de Compte AWS pour supprimer l'accès humain persistant aux environnements de production. Cette pratique minimise le risque de modifications non autorisées ou accidentelles, ce qui améliore l'intégrité de vos systèmes de production. Au lieu d'un accès humain direct, vous pouvez utiliser des outils CI/CD tels que AWS CodeBuild et AWS CodePipeline pour effectuer des déploiements. Vous pouvez utiliser ces services pour automatiser les processus de création, de test et de déploiement, ce qui réduit les interventions manuelles et améliore la cohérence.

Pour renforcer encore la sécurité et la traçabilité, vous pouvez signer vos packages d'applications après les avoir testés et valider ces signatures lors du déploiement. Pour ce faire, utilisez des outils cryptographiques tels que AWS Signer ou AWS Key Management Service (AWS KMS). En signant et en vérifiant les packages, vous pouvez vous assurer que vous ne déployez que du code autorisé et validé dans vos environnements.

En outre, votre équipe peut concevoir votre charge de travail pour obtenir des données de configuration spécifiques à l'environnement à partir d'une source externe, telle que AWS Systems Manager Parameter Store. Cette pratique sépare le code d'application des données de configuration, ce qui vous permet de gérer et de mettre à jour les configurations indépendamment sans modifier le code d'application lui-même.

Pour rationaliser le provisionnement et la gestion de l'infrastructure, envisagez d'utiliser des outils d'infrastructure en tant que code (IaC) tels qu'AWS CloudFormation ou AWS CDK. Vous pouvez utiliser ces outils pour définir votre infrastructure en tant que code, ce qui améliore la cohérence et la répétabilité des déploiements dans différents environnements.

Envisagez d'utiliser des déploiements canary pour valider la réussite du déploiement de votre logiciel. Les déploiements canary impliquent le déploiement de modifications sur un sous-ensemble d'instances ou d'utilisateurs avant leur déploiement dans l'environnement de production tout entier. Vous pouvez ainsi surveiller l'impact des modifications et revenir en arrière si nécessaire, ce qui minimise le risque de problèmes généralisés.

Suivez les recommandations décrites dans le livre blanc Organisation de votre environnement AWS à l'aide de comptes multiples. Ce livre blanc fournit des conseils sur la manière de séparer les environnements (par exemple de développement, intermédiaire et de production) dans des Comptes AWS distincts, ce qui améliore encore la sécurité et l'isolation.

Étapes d'implémentation

1. Configurez la structure de Compte AWS :

- Suivez les instructions du livre blanc Organisation de votre environnement AWS à l'aide de comptes multiples pour créer des Comptes AWS distincts pour les différents environnements (par exemple, de développement, intermédiaire et de production).
- Configurez les contrôles d'accès et les autorisations appropriés pour chaque compte afin de limiter l'accès humain direct aux environnements de production.

2. Implémentez un pipeline CI/CD :
   - Configurez un pipeline CI/CD à l'aide de services tels qu'AWS CodeBuild et AWS CodePipeline.
   - Configurez le pipeline pour créer, tester et déployer automatiquement votre code d'application dans les environnements respectifs.
   - Intégrez des référentiels de code au pipeline CI/CD pour le contrôle des versions et la gestion du code.
3. Signez et vérifiez les packages d'applications :
   - Utilisez AWS Signer ou AWS Key Management Service (AWS KMS) pour signer vos packages d'application une fois qu'ils ont été testés et validés.
   - Configurez le processus de déploiement pour vérifier les signatures des packages d'applications avant de les déployer dans les environnements cibles.
4. Externalisez les données de configuration :
   - Stockez les données de configuration spécifiques à l'environnement dans AWS Systems Manager Parameter Store.
   - Modifiez votre code d'application pour récupérer les données de configuration depuis Parameter Store pendant le déploiement ou l'exécution.
5. Mettez en œuvre une infrastructure en tant que code (IaC) :
   - Utilisez des outils d'infrastructure en tant que code tels qu'AWS CloudFormation ou AWS CDK pour définir et gérer votre infrastructure en tant que code.
   - Créez des modèles CloudFormation ou des scripts CDK pour provisionner et configurer les ressources AWS nécessaires pour votre application.
   - Intégrez l'infrastructure en tant que code à votre pipeline CI/CD pour déployer automatiquement les modifications d'infrastructure en même temps que les modifications du code d'application.
6. Mettez en œuvre des déploiements canary :
   - Configurez votre processus de déploiement pour prendre en charge les déploiements canary, dans lesquels les modifications sont appliquées à un sous-ensemble d'instances ou d'utilisateurs avant d'être déployées dans l'environnement de production tout entier.

- Utilisez des services tels qu'AWS CodeDeploy ou AWS ECS pour gérer les déploiements canary et surveiller l'impact des modifications.
- Mettez en œuvre des mécanismes de restauration pour pouvoir rétablir la version stable précédente si des problèmes sont détectés au cours du déploiement canary.

7. Surveillez et auditez :
   - Configurez des mécanismes de surveillance et de journalisation pour suivre les déploiements, les performances des applications et les modifications de l'infrastructure.
   - Utilisez des services tels qu'Amazon CloudWatch et AWS CloudTrail pour collecter et analyser des journaux et des métriques.
   - Mettez en œuvre des audits et des contrôles de conformité pour vérifier le respect des bonnes pratiques de sécurité et des exigences réglementaires.
8. Améliorez continuellement :
   - Passez en revue et mettez à jour régulièrement vos pratiques de déploiement et incorporez les commentaires et les enseignements tirés des déploiements précédents.
   - Automatisez autant que possible le processus de déploiement afin de réduire les interventions manuelles et les erreurs humaines potentielles.
   - Collaborez avec des équipes interfonctionnelles (par exemple, des opérations ou de la sécurité) pour aligner et améliorer continuellement les pratiques de déploiement.

En suivant ces étapes, vous pouvez mettre en œuvre des pratiques de déploiement sécurisées et automatisées dans votre environnement AWS, ce qui améliore la sécurité, réduit le risque d'erreurs humaines et rationalise le processus de déploiement.

Ressources

Bonnes pratiques associées :

- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication
- DL.CI.2 Déclenchement automatique de la génération lors de modifications du code source

Documents connexes :

- AWS CI/CD Workshop
- Accelerate deployments on AWS with effective governance
- Automating safe, hands-off deployments

- Code signing using AWS Certificate Manager Private CA and AWS Key Management Service asymmetric keys
- Code Signing, a Trust and Integrity Control for AWS Lambda

Vidéos connexes :

- Hands-off: Automating continuous delivery pipelines at Amazon

Exemples connexes :

- Déploiements bleu/vert avec AWS Fargate

SEC11-BP07 Évaluation régulière des caractéristiques de sécurité des pipelines

Appliquez les principes du pilier Sécurité Well-Architected à vos pipelines, en accordant une attention particulière à la séparation des autorisations. Évaluez régulièrement les caractéristiques de sécurité de votre infrastructure de pipelines. Une gestion efficace de la sécurité des pipelines vous permet d'assurer la sécurité des logiciels qui transitent par ces pipelines.

Résultat escompté : les pipelines que vous utilisez pour créer et déployer votre logiciel suivent les mêmes pratiques recommandées que toute autre charge de travail de votre environnement. Les tests que vous implémentez dans vos pipelines ne sont pas modifiables par les équipes qui les utilisent. Vous ne donnez aux pipelines que les autorisations nécessaires aux déploiements qu'ils effectuent à l'aide d'informations d'identification temporaires. Vous mettez en œuvre des protections pour empêcher les pipelines de se déployer dans les mauvais environnements. Vous configurez vos pipelines pour qu'ils émettent un état afin que l'intégrité de vos environnements de génération puisse être validée.

Anti-modèles courants :

- Tests de sécurité qui peuvent être contournés par les créateurs.
- Des autorisations trop larges pour les pipelines de déploiement.
- Les pipelines ne sont pas configurés pour valider les entrées.
- Ne pas passer régulièrement en revue les autorisations associées à votre infrastructure CI/CD.
- Utilisation d'informations d'identification à long terme ou codées en dur.

Avantages liés au respect de cette bonne pratique :

- Une plus grande confiance dans l'intégrité du logiciel conçu et déployé par le biais des pipelines.
- Possibilité d'interrompre un déploiement en cas d'activité suspecte.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Vos pipelines de déploiement constituent un élément essentiel du cycle de développement de votre logiciel et devraient suivre les mêmes principes et pratiques de sécurité que toute autre charge de travail dans votre environnement. Cela inclut la mise en œuvre de contrôles d'accès appropriés, la validation des entrées et l'examen et l'audit réguliers des autorisations associées à votre infrastructure CI/CD.

Vérifiez que les équipes responsables de la création et du déploiement des applications ne sont pas en mesure de modifier ou de contourner les tests et les contrôles de sécurité mis en œuvre dans vos pipelines. Cette séparation des préoccupations permet de préserver l'intégrité de vos processus de création et de déploiement.

Comme point de départ, envisagez d'utiliser l'architecture de référence des pipelines de déploiement AWS. Cette architecture de référence fournit une base sécurisée et évolutive pour la construction de vos pipelines CI/CD sur AWS.

En outre, vous pouvez utiliser des services tels qu'AWS Identity and Access Management Access Analyzer pour générer des politiques IAM de moindre privilège à la fois pour les autorisations de votre pipeline et comme étape de votre pipeline pour vérifier les autorisations de charge de travail. Cela permet de vérifier que vos pipelines et vos charges de travail disposent uniquement des autorisations nécessaires pour leurs fonctions spécifiques, ce qui réduit le risque d'accès ou d'actions non autorisés.

Étapes d'implémentation

- Commencez par l'architecture de référence des pipelines de déploiement AWS.
- Envisagez d'utiliser l'Analyseur d'accès AWS IAM pour générer par programmation des politiques IAM de moindre privilège pour les pipelines.
- Intégrez vos pipelines à la surveillance et aux alertes afin d'être averti en cas d'activité inattendue ou anormale. Pour les services gérés par AWS, Amazon EventBridge vous permet d'acheminer les données vers des cibles telles comme AWS Lambda ou Amazon Simple Notification Service (Amazon SNS).

Ressources

Documents connexes :

- Architecture de référence des pipelines de déploiement d'AWS
- Surveillance de AWS CodePipeline
- Bonnes pratiques de sécurité pour AWS CodePipeline

Exemples connexes :

- Tableau de bord de surveillance DevOps (GitHub)

SEC11-BP08 Création d'un programme permettant aux équipes responsables de la charge de travail de s'approprier les mécanismes de sécurité

Créez un programme ou un mécanisme qui permette aux équipes de créateurs de prendre des décisions en matière de sécurité pour les logiciels qu'ils créent. Votre équipe de sécurité doit toujours valider ces décisions au cours d'un examen, mais le fait de donner la responsabilité de la sécurité aux équipes de concepteurs permet d'élaborer des charges de travail plus rapides et plus sûres. Ce mécanisme favorise également une culture de responsabilisation qui a un impact positif sur le fonctionnement des systèmes que vous construisez.

Résultat escompté : vous avez intégré la prise en charge de la sécurité et la prise de décision dans vos équipes. Vous avez formé vos équipes à la façon de réfléchir à la sécurité ou vous avez renforcé les équipes en y intégrant ou associant des personnes chargées de la sécurité. Vos équipes prennent ainsi des décisions de meilleure qualité en matière de sécurité plus tôt dans le cycle de développement.

Anti-modèles courants :

- Laisser à une équipe de sécurité le soin de prendre toutes les décisions relatives à la conception de la sécurité.
- Ne pas tenir compte des exigences de sécurité suffisamment tôt dans le processus de développement.
- Ne pas recueillir de commentaires des créateurs et des responsables de la sécurité sur le fonctionnement du programme.

Avantages liés au respect de cette bonne pratique :

- Réduction du temps nécessaire à la réalisation des examens de sécurité.
- Réduction des problèmes de sécurité qui ne sont détectés qu'au stade de l'examen de la sécurité.
- Amélioration de la qualité globale du logiciel en cours d'écriture.
- Possibilité d'identifier et de comprendre les problèmes systémiques ou les domaines d'amélioration à forte valeur ajoutée.
- Réduction de la quantité de travail à refaire en raison des conclusions de l'examen de sécurité.
- Amélioration de la perception de la fonction de sécurité.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Commencez par suivre les instructions SEC11-BP01 Formation à la sécurité des applications. Identifiez ensuite le modèle opérationnel du programme qui vous semble le plus adapté à votre organisation. Les deux principaux modèles consistent à former les créateurs ou à intégrer les responsables de la sécurité dans les équipes de créateurs. Une fois que vous avez décidé de l'approche initiale, vous devez mener un projet pilote avec une seule équipe ou un petit groupe d'équipes de charge de travail afin de prouver que le modèle fonctionne pour votre organisation. Le soutien de la direction de l'organisation en matière de construction et de sécurité contribue à la mise en œuvre et à la réussite du programme. Lors de la création de ce programme, il est important de choisir des métriques qui peuvent être utilisées pour montrer la valeur du programme. Apprendre de la manière dont AWS les autres ont abordé ce problème est une bonne expérience d'apprentissage. Cette bonne pratique est très axée sur le changement organisationnel et la culture. Les outils que vous utilisez doivent favoriser la collaboration entre les créateurs et les responsables de la sécurité.

Étapes d'implémentation

- Commencez par former vos créateurs à la cybersécurité des applications.
- Créer une communauté et un programme d'intégration pour former les créateurs.
- Choisissez un nom pour le programme. Les termes « tuteur », « champion » ou « défenseur » sont couramment utilisés.
- Identifier le modèle à utiliser : former des créateurs, intégrer des ingénieurs en sécurité ou avoir des rôles de sécurité connexes.

- Identifier les sponsors du projet parmi les responsables de la sécurité, les créateurs et éventuellement d'autres groupes concernés.
- Suivez les métriques concernant le nombre de personnes impliquées dans le programme, le temps nécessaire aux examens et les commentaires des créateurs et des responsables de la sécurité. Utilisez ces métriques pour apporter des améliorations.

Ressources

Bonnes pratiques associées :

- SEC11-BP01 Formation à la sécurité des applications
- SEC11-BP02 Automatisation des tests tout au long du cycle de développement et de publication

Documents connexes :

- Comment aborder la modélisation des menaces
- Comment envisager la gouvernance de la sécurité dans le cloud
- Création par AWS du programme Security Guardians, un mécanisme de répartition de la prise en charge de la sécurité
- Comment créer un programme Security Guardians pour répartir la prise en charge de la sécurité

Vidéos connexes :

- Proactive security: Considerations and approaches
- Conseils relatifs à l'outillage et à la culture AppSec fournis par AWS et Toyota Motor North America

Fiabilité

Le pilier Fiabilité englobe la capacité d'une charge de travail à exécuter sa fonction de manière correcte et cohérente et ce, en temps utile. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Fiabilité.

Domaines de bonnes pratiques

- Fondations
- Architecture de charge de travail

- Gestion des modifications
- Gestion des défaillances

Fondations

Questions

- FIA 1. Comment gérer les Service Quotas et les contraintes ?
- FIA 2. Comment planifier la topologie de votre réseau ?

FIA 1. Comment gérer les Service Quotas et les contraintes ?

Pour les architectures de charge de travail basées sur le cloud, il existe des Service Quotas (également appelés limites de service). Ces quotas permettent d'éviter de fournir accidentellement plus de ressources que nécessaire et de limiter les taux de demande des opérations d'API afin de protéger les services de tout abus. Il existe également des contraintes de ressources, par exemple la vitesse à laquelle les bits peuvent être transmis par un câble à fibre optique ou la quantité de données stockées sur un disque physique.

Bonnes pratiques

- REL01-BP01 Connaître les quotas et les contraintes de service
- REL01-BP02 Gérer les quotas de service entre les comptes et les régions
- REL01-BP03 Respecter les quotas et les contraintes de service fixes grâce à l'architecture
- REL01-BP04 Surveiller et gérer les quotas
- REL01-BP05 Automatisation de la gestion des quotas
- REL01-BP06 Assurez-vous qu'il existe un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement

REL01-BP01 Connaître les quotas et les contraintes de service

Connaissez vos quotas par défaut et gérez vos demandes d'augmentation de quota pour votre architecture de charge de travail. Connaissez également les contraintes de ressources, comme le disque ou le réseau, qui sont susceptibles d'avoir un impact.

Résultat souhaité : Les clients peuvent empêcher la dégradation ou l'interruption de leurs services en Comptes AWS mettant en œuvre des directives appropriées pour le suivi des indicateurs clés, des

examens de l'infrastructure et des mesures correctives automatisées afin de vérifier que les quotas et les contraintes des services ne sont pas atteints, ce qui pourrait entraîner une dégradation ou une interruption du service.

Anti-modèles courants :

- Déployer une charge de travail sans comprendre les quotas matériels ou logiciels et leurs limites pour les services utilisés.
- Déployer une charge de travail de remplacement sans analyser ni reconfigurer les quotas nécessaires ou contacter d'abord l'assistance.
- Supposer que les services cloud sont sans limite et que les services peuvent être utilisés sans prendre en compte les taux, les limites, les nombres et les quantités.
- Supposer que les quotas augmenteront automatiquement.
- Ne pas connaître le processus et la chronologie des demandes de quotas.
- Supposer que le quota du service cloud pas défaut est le même pour chaque service par rapport à d'autres régions.
- Supposer que les contraintes de service peuvent être enfreintes et que les systèmes se mettront automatiquement à l'échelle ou augmenteront la limite au-delà des contraintes de la ressource.
- Ne pas tester l'application sur des pics de trafic pour tester la résistance de l'utilisation de ces ressources.
- Provisionner les ressources sans analyser la taille de ressource nécessaire.
- Surprovisionner la capacité en choisissant des types de ressources qui vont bien au-delà des besoins réels ou des pics attendus.
- Ne pas évaluer les exigences de capacité pour les nouveaux niveaux de trafic avant un nouvel événement client ou le déploiement d'une nouvelle technologie.

Avantages de l'établissement de cette bonne pratique : la surveillance et la gestion automatisée des quotas de service et des contraintes de ressources peuvent réduire les défaillances de manière proactive. Les changements dans les modèles de trafic d'un service client peuvent entraîner une interruption ou une dégradation si les bonnes pratiques ne sont pas suivies. En surveillant et en gérant les valeurs de quota sur toutes les régions et tous les comptes, les applications peuvent bénéficier d'une meilleure résilience lors d'événements indésirables ou imprévus.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d’implémentation

Service Quotas est un AWS service qui vous permet de gérer vos quotas pour plus de 250 AWS services à partir d'un seul endroit. Outre la recherche des valeurs des quotas, vous pouvez également demander et suivre les augmentations de quotas depuis la console Service Quotas ou en utilisant le AWS SDK. AWS Trusted Advisor propose une vérification des quotas de service qui affiche votre utilisation et vos quotas pour certains aspects de certains services. Les quotas de service par défaut par service figurent également dans la AWS documentation de chaque service (par exemple, voir Amazon VPC Quotas).

Certaines limites de service, telles que les limites de débit en cas de limitation, APIs sont définies dans Amazon API Gateway lui-même en configurant un plan d'utilisation. Certaines limites définies en tant que configuration sur leurs services respectifs incluent ProvisionedIOPS, le RDS stockage Amazon alloué et les allocations de EBS volume Amazon. Amazon Elastic Compute Cloud dispose de son propre tableau de bord des limites de service qui peut vous aider à gérer vos limites d’instances, d’Amazon Elastic Block Store et d’adresses IP élastiques. Si vous avez un cas d'utilisation où les quotas de service ont un impact sur les performances de votre application et ne sont pas ajustables à vos besoins, contactez-nous Support pour voir s'il existe des mesures d'atténuation.

Les quotas de service peuvent être spécifiques à une région ou mondiaux par nature. L'utilisation d'un AWS service qui atteint son quota ne fonctionnera pas comme prévu dans le cadre d'une utilisation normale et peut entraîner une interruption ou une dégradation du service. Par exemple, un quota de service limite le nombre d'EC2instances DL Amazon utilisées dans une région. Cette limite peut être atteinte lors d'un événement de dimensionnement du trafic à l'aide des groupes Auto Scaling (ASG).

L’utilisation des quotas de service pour chaque compte doit être évaluée régulièrement pour déterminer quelles seraient les limites de service appropriées pour ce compte. Ces quotas de service existent en tant que barrières de protection opérationnelles pour empêcher le provisionnement accidentel de plus de ressources que nécessaire. Ils servent également à limiter le taux de demandes sur les API opérations afin de protéger les services contre les abus.

Les contraintes de service sont différentes des quotas de service. Les contraintes de service représentent les limites d’une ressource spécifique, telles que définies par ce type de ressource. Il peut s'agir de la capacité de stockage (par exemple, gp2 a une limite de taille de 1 Go à 16 To) ou du débit du disque. Il est essentiel qu’une contrainte d’un type de ressource soit optimisée et constamment évaluée par rapport à une utilisation qui pourrait atteindre ses limites. Si une contrainte

est atteinte de manière inattendue, les applications ou les services du compte peuvent être dégradés ou interrompus.

S'il existe un cas d'utilisation où les quotas de service ont un impact sur les performances d'une application et ne peuvent pas être ajustés aux besoins requis, contactez-nous Support pour voir s'il existe des mesures d'atténuation. Pour plus de détails sur l'ajustement des quotas fixes, consultez REL01-BP03 Respecter les quotas et les contraintes de service fixes grâce à l'architecture.

Il existe un certain nombre de AWS services et d'outils permettant de surveiller et de gérer les Quotas de Service. Les services et les outils doivent être exploités pour vérifier automatiquement ou manuellement les niveaux de quotas.

- AWS Trusted Advisor propose un contrôle des quotas de service qui affiche votre utilisation et vos quotas pour certains aspects de certains services. Il peut aider à identifier des services proches du quota.
- AWS Management Console fournit des méthodes permettant d'afficher les valeurs des quotas de services, de les gérer, de demander de nouveaux quotas, de surveiller l'état des demandes de quotas et d'afficher l'historique des quotas.
- AWS CLI et CDKs proposent des méthodes programmatiques pour gérer et surveiller automatiquement les niveaux et l'utilisation des quotas de service.

Étapes d'implémentation

Pour Service Quotas :

- Vérifiez les Quotas de AWS Service.
- Pour connaître vos quotas de service existants, déterminez les services (tels qu'IAMAccess Analyzer) utilisés. Il existe environ 250 AWS services contrôlés par des quotas de services. Ensuite, déterminez le nom du quota de service spécifique qui pourrait être utilisé au sein de chaque compte et région. Il existe environ 3 000 noms de quotas de service par région.
- Complétez cette analyse des quotas AWS Config pour trouver toutes les AWS ressources utilisées dans votre Comptes AWS.
- Utilisez AWS CloudFormation les données pour déterminer les AWS ressources que vous utilisez. Examinez les ressources qui ont été créées dans AWS Management Console ou avec la `list-stack-resources` AWS CLI commande. Vous pouvez également voir les ressources configurées pour être déployées directement dans le modèle.

- Déterminez tous les services indispensables à votre charge de travail en prenant en compte le code de déploiement.
- Identifiez les quotas de service pertinents. Utilisez les informations accessibles par programmation provenant de Service Quotas Trusted Advisor et de Service Quotas.
- Établissez une méthode de surveillance automatisée (consultez REL01-BP02 Gérer les quotas de service entre les comptes et les régions et REL01-BP04 Surveiller et gérer les quotas) pour alerter et informer si les quotas de services sont proches de leur limite ou ont atteint leur limite.
- Établissez une méthode automatisée et programmatique pour vérifier si un quota de service a été modifié dans une région mais pas dans d'autres régions du même compte (consultez REL01-BP02 Gérer les quotas de service entre les comptes et les régions etREL01-BP04 Surveiller et gérer les quotas).
- Automatisez l'analyse des journaux et des métriques de l'application pour déterminer s'il existe des erreurs de contraintes de quotas ou de services. Si de telles erreurs existent, envoyez des alertes au système de surveillance.
- Établissez des procédures d'ingénierie pour calculer le changement de quota requis (consultez REL01-BP05 Automatisation de la gestion des quotas) une fois qu'il a été déterminé que des quotas plus importants sont nécessaires pour des services spécifiques.
- Créez un flux de travail de provisionnement et d'approbation pour demander des modifications des quotas de service. Cela doit inclure un flux de travail d'exception en cas de refus d'une demande ou d'une approbation partielle.
- Créez une méthode d'ingénierie pour revoir les quotas de service avant de fournir et d'utiliser de nouveaux AWS services avant de les déployer dans des environnements de production ou chargés. (par exemple, compte de test de charge).

Pour les contraintes de service :

- Établissez des méthodes de surveillance et des métriques pour alerter quand les ressources sont proches de leurs contraintes. Tirez parti CloudWatch le cas échéant des métriques ou de la surveillance des journaux.
- Établissez des seuils d'alertes pour chaque ressource ayant une contrainte importante pour l'application ou le système.
- Créez des procédures de gestion des flux de travail et de l'infrastructure pour changer le type de ressource si la contrainte est proche de l'utilisation. Ce flux de travail doit inclure le test de charge comme une bonne pratique pour vérifier que ce nouveau type est le bon type de ressource avec les nouvelles contraintes.

- Procédez à la migration des ressources identifiées vers le nouveau type de ressource recommandé avec les procédures et les processus existants.

Ressources

Bonnes pratiques associées :

- REL01-BP02 Gérer les quotas de service entre les comptes et les régions
- REL01-BP03 Respecter les quotas et les contraintes de service fixes grâce à l'architecture
- REL01-BP04 Surveiller et gérer les quotas
- REL01-BP05 Automatisation de la gestion des quotas
- REL01-BP06 Assurez-vous qu'il existe un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement
- REL03-BP01 Choisissez comment segmenter votre charge de travail
- REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la guérison sur toutes les couches
- REL12-BP04 Tester la résilience à l’aide de l’ingénierie du chaos

Documents connexes :

- AWS Le pilier de fiabilité du framework Well-Architected : la disponibilité
- AWS Quotas de service (anciennement appelés limites de service)
- AWS Trusted Advisor Contrôles des meilleures pratiques (voir la section Limites de service)
- AWS limitez le nombre de AWS réponses
- Limites EC2 de service Amazon
- Qu’est-ce que Service Quotas ?
- Comment demander une augmentation de quota
- Points de terminaison et quotas de service
- Guide de l’utilisateur de Service Quotas
- Quota Monitor pour AWS
- AWS Limites d'isolation des défauts
- Disponibilité avec redondance

- AWS pour les données
- Qu’est-ce que l’intégration continue ?
- Qu’est-ce que la livraison continue ?
- APNPartenaire : partenaires qui peuvent vous aider à gérer la configuration
- Gestion du cycle de vie des comptes dans les environnements account-per-tenant SaaS sur AWS
- Gestion et surveillance de la API régulation de vos charges de travail
- Consultez les AWS Trusted Advisor recommandations à grande échelle avec AWS Organizations
- Automatiser l'augmentation des limites de service et le support aux entreprises avec AWS Control Tower

Vidéos connexes :

- AWS Live Re:inforce 2019 - Service Quotas
- Afficher et gérer les quotas pour les AWS services à l'aide de quotas de service
- AWS IAMDémo de quotas

Outils associés :

- CodeGuru Réviseur Amazon
- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP02 Gérer les quotas de service entre les comptes et les régions

Si vous utilisez plusieurs comptes ou régions, demandez les quotas appropriés dans tous les environnements où vos charges de travail de production s'exécutent.

Résultat escompté : les services et les applications ne devraient pas être affectés par l'épuisement des quotas de services pour les configurations qui couvrent plusieurs comptes ou régions ou qui ont des conceptions de résilience utilisant le basculement de zone, de région ou de compte.

Anti-modèles courants :

- Laisser l'utilisation des ressources dans une région d'isolement se développer sans aucun mécanisme pour maintenir de la capacité dans les autres zones.
- Définition manuelle de tous les quotas de manière indépendante dans les régions d'isolement.
- Non-prise en considération de l'effet des architectures de résilience (par exemple, actives ou passives) dans les futurs besoins de quotas alors qu'une dégradation est observée dans la région non principale.
- Absence d'évaluation régulière des quotas et des changements qui s'imposent dans chaque région et chaque compte où la charge de travail s'exécute.
- Non-utilisation des modèles de demande de quotas pour demander des augmentations dans plusieurs régions et comptes.
- Absence de mise à jour des quotas de services pensant à tort que l'augmentation de quotas a des répercussions sur les coûts comme les demandes de réservation de capacité de calcul.

Avantages de l'établissement de cette bonne pratique : vérification que vous pouvez gérer votre charge de travail actuelle dans les régions ou les comptes secondaires si les services régionaux ne sont plus disponibles. Cela peut contribuer à limiter le nombre d'erreurs ou les niveaux de dégradations observés lors d'une perte de région.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les quotas de services sont suivis par compte. Sauf indication contraire, chaque quota est Région AWS spécifique. En plus des environnements de production, gérez également les quotas dans tous les autres environnements applicables de façon à ne pas entraver les tests et le développement. Pour maintenir un haut niveau de résilience, il convient d'évaluer constamment les quotas de services (que ce soit de façon automatisée ou manuelle).

Compte tenu de l'augmentation des charges de travail couvrant les régions en raison de la mise en œuvre de conceptions utilisant les approches active/active, active/passive : à chaud, active/passive : à froid et active/passive : veilleuse, il est essentiel de comprendre tous les niveaux de quotas de région et de compte. Les modèles de trafic passés ne permettent pas toujours de déterminer correctement si le quota de service est bien défini.

Tout aussi important, la valeur limite d'un nom de quota de service n'est pas toujours identique d'une région à l'autre. Ainsi, cette valeur peut être égale à cinq dans une région et à dix dans une autre. La gestion de ces quotas doit englober tous les services, comptes et régions identiques pour offrir une résilience cohérente dans des conditions de charge.

Rapprochez toutes les différences de quotas de services entre les différentes régions (région active ou région passive) et créez des processus permettant de rapprocher constamment ces différences. Les plans de test de basculements de régions passives sont rarement mis à l'échelle pour atteindre une capacité active de pointe, ce qui signifie que les exercices de simulation (« game day ») et les exercices de table (« table top ») ne permettent pas nécessairement d'identifier les différences dans les quotas de services entre les régions et donc de maintenir les limites adéquates.

Il est très important de suivre et d'évaluer la dérive des quotas de services, c'est-à-dire la situation dans laquelle les limites des quotas de services pour un quota nommé spécifique sont modifiées dans une seule région, et non dans toutes les régions. Il doit être envisagé de changer le quota dans les régions qui présentent du trafic ou qui pourraient potentiellement en véhiculer.

- Sélectionnez les comptes et les régions appropriés en fonction de vos exigences de service, de latence, de réglementation et de reprise après sinistre (DR).
- Identifiez les quotas de services dans l'ensemble des comptes, régions et zones de disponibilité appropriés. Les limites s'appliquent au compte et à la région. Ces valeurs doivent être comparées pour repérer les différences.

Étapes d'implémentation

- Examinez les valeurs de Service Quotas susceptibles d'avoir transgressé le niveau d'utilisation à risque. AWS Trusted Advisor propose des alertes pour les violations de seuil de 80 % et 90 %.
- Examinez les valeurs de quotas de services dans les régions passives (dans une conception de type actif/passif). Vérifiez que la charge de travail s'exécutera correctement dans les régions secondaires en cas de défaillance dans la région principale.
- Automatisez l'évaluation pour identifier une éventuelle dérive de Service Quota entre des régions d'un même compte et agissez en conséquence pour changer les limites.

- Si la structure des unités d'organisation (UO) est prise en charge, les modèles de quotas de services doivent être mis à jour en fonction des changements apportés aux quotas qui doivent s'appliquer à plusieurs régions et comptes.
  - Créez un modèle et associez les régions au changement de quota.
  - Examinez tous les modèles de quotas de services existants pour y apporter les changements nécessaires (région, limites et comptes).

Ressources

Bonnes pratiques associées :

- REL01-BP01 Connaître les quotas et les contraintes de service
- REL01-BP03 Respecter les quotas et les contraintes de service fixes grâce à l'architecture
- REL01-BP04 Surveiller et gérer les quotas
- REL01-BP05 Automatisation de la gestion des quotas
- REL01-BP06 Assurez-vous qu'il existe un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement
- REL03-BP01 Choisissez comment segmenter votre charge de travail
- REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la guérison sur toutes les couches
- REL12-BP04 Tester la résilience à l'aide de l'ingénierie du chaos

Documents connexes :

- AWS Le pilier de fiabilité du framework Well-Architected : la disponibilité
- AWS Quotas de service (anciennement appelés limites de service)
- AWS Trusted Advisor Contrôles des meilleures pratiques (voir la section Limites de service)
- AWS limitez le nombre de AWS réponses
- Limites EC2 de service Amazon
- Qu'est-ce que Service Quotas ?
- Comment demander une augmentation de quota
- Points de terminaison et quotas de service

- Guide de l’utilisateur de Service Quotas
- Quota Monitor pour AWS
- AWS Limites d'isolation des défauts
- Disponibilité avec redondance
- AWS pour les données
- Qu’est-ce que l’intégration continue ?
- Qu’est-ce que la livraison continue ?
- APNPartenaire : partenaires qui peuvent vous aider à gérer la configuration
- Gestion du cycle de vie des comptes dans les environnements account-per-tenant SaaS sur AWS
- Gestion et surveillance de la API régulation de vos charges de travail
- Consultez les AWS Trusted Advisor recommandations à grande échelle avec AWS Organizations
- Automatiser l'augmentation des limites de service et le support aux entreprises avec AWS Control Tower

Vidéos connexes :

- AWS Live Re:inforce 2019 - Service Quotas
- Afficher et gérer les quotas pour les AWS services à l'aide de quotas de service
- AWS IAMDémo de quotas

Services connexes :

- CodeGuru Réviseur Amazon
- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK

- AWS Systems Manager
- AWS Marketplace

REL01-BP03 Respecter les quotas et les contraintes de service fixes grâce à l'architecture

Ayez conscience des quotas de services non modifiables, des contraintes de service et des limites de ressources physiques. Concevez des architectures pour les applications et les services afin d'éviter que ces limites n'aient un impact sur la fiabilité.

Les exemples incluent la bande passante du réseau, la taille de la charge utile d'invocation de fonctions sans serveur, le débit d'accélération d'une API passerelle et les connexions utilisateur simultanées à une base de données.

Résultat escompté : l'application ou le service fonctionne comme prévu dans des conditions de trafic normales et intenses. Ils ont été conçus pour fonctionner dans les limites des contraintes fixes ou des quotas de services de cette ressource.

Anti-modèles courants :

- Choix d'une conception qui utilise une ressource d'un service, sans savoir qu'il existe des contraintes de conception qui entraîneront l'échec de cette conception au fil des mises à l'échelle.
- Réalisation d'une évaluation comparative qui n'est pas réaliste et qui atteindra les quotas fixés par le service pendant les tests. Par exemple, l'exécution de tests à une limite de débordement mais pendant une durée prolongée.
- Le choix d'une conception qui ne peut pas se mettre à l'échelle ou être modifiée si des quotas de services fixes doivent être dépassés. Par exemple, une SQS charge utile de 256 Ko.
- L'observabilité n'a pas été conçue et mise en œuvre pour surveiller et alerter sur les seuils des quotas de services qui pourraient être compromis lors d'événements à fort trafic

Avantages de l'établissement de cette bonne pratique : vérifier que l'application s'exécutera sous tous les niveaux de charge de service prévus, sans interruption ni dégradation.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Contrairement aux quotas de services logiciels ou aux ressources qui sont remplacées par des unités de plus grande capacité, les quotas fixes AWS des services ne peuvent pas être modifiés. Cela

signifie que tous ces types de AWS services doivent être évalués pour détecter d'éventuelles limites de capacité stricte lorsqu'ils sont utilisés dans le cadre de la conception d'une application.

Les limites strictes sont affichées dans la console Service Quotas. Si les colonnes indiquent `ADJUSTABLE = No`, le service est soumis à une limite stricte. Des limites strictes sont également indiquées dans les pages de configuration de certaines ressources. Par exemple, Lambda possède des limites strictes spécifiques qui ne peuvent pas être ajustées.

À titre d'exemple, lors de la conception d'une application python destinée à être exécutée dans une fonction Lambda, l'application doit être évaluée pour déterminer si Lambda risque de s'exécuter pendant plus de 15 minutes. Si le code peut fonctionner au-delà de cette limite de Service Quota, il faut envisager d'autres technologies ou conceptions. Si cette limite est atteinte après le déploiement de la production, l'application subira une dégradation et des perturbations jusqu'à ce qu'il soit possible d'y remédier. Contrairement aux quotas souples, il n'existe aucune méthode permettant de passer à ces limites, même en cas d'événements d'urgence de gravité 1.

Une fois que l'application a été déployée dans un environnement de test, il convient d'utiliser des stratégies pour déterminer si des limites strictes peuvent être atteintes. Les tests de résistance, les tests de charge et les tests de chaos doivent faire partie du plan de test d'introduction.

Étapes d'implémentation

- Consultez la liste complète des AWS services qui pourraient être utilisés lors de la phase de conception de l'application.
- Examinez les limites de quota flexible et de quota fixe pour tous ces services. Les limites ne sont pas toutes affichées dans la console Service Quotas. Certains services décrivent ces limites dans d'autres lieux.
- Lors de la conception de votre application, examinez les facteurs opérationnels et technologiques de votre charge de travail, tels que les résultats opérationnels, le cas d'utilisation, les systèmes dépendants, les objectifs de disponibilité et les objets de reprise après sinistre. Laissez vos facteurs commerciaux et technologiques guider le processus d'identification du système distribué qui convient à votre charge de travail.
- Analysez la charge de service dans les régions et les comptes. De nombreuses limites strictes se basent sur la région pour les services. Cependant, certaines limites sont basées sur le compte.
- Analysez les architectures de résilience pour l'utilisation des ressources lors d'une panne de zone et d'une panne régionale. Dans la progression des conceptions multirégionales utilisant des approches actives/actives, actives/passives : à chaud, actives/passives : à froid, et actives/

passives : veilleuse, ces cas de panne entraîneront une utilisation plus importante. Cela crée un cas d'utilisation potentiel pour atteindre des limites strictes.

Ressources

Bonnes pratiques associées :

- REL01-BP01 Connaître les quotas et les contraintes de service
- REL01-BP02 Gérer les quotas de service entre les comptes et les régions
- REL01-BP04 Surveiller et gérer les quotas
- REL01-BP05 Automatisation de la gestion des quotas
- REL01-BP06 Assurez-vous qu'il existe un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement
- REL03-BP01 Choisissez comment segmenter votre charge de travail
- REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la guérison sur toutes les couches
- REL12-BP04 Tester la résilience à l'aide de l'ingénierie du chaos

Documents connexes :

- AWS Le pilier de fiabilité du framework Well-Architected : la disponibilité
- AWS Quotas de service (anciennement appelés limites de service)
- AWS Trusted Advisor Contrôles des meilleures pratiques (voir la section Limites de service)
- AWS limitez le nombre de AWS réponses
- Limites EC2 de service Amazon
- Qu'est-ce que Service Quotas ?
- Comment demander une augmentation de quota
- Points de terminaison et quotas de service
- Guide de l'utilisateur de Service Quotas
- Quota Monitor pour AWS
- AWS Limites d'isolation des défauts

- Disponibilité avec redondance
- AWS pour les données
- Qu’est-ce que l’intégration continue ?
- Qu’est-ce que la livraison continue ?
- APNPartenaire : partenaires qui peuvent vous aider à gérer la configuration
- Gestion du cycle de vie des comptes dans les environnements account-per-tenant SaaS sur AWS
- Gestion et surveillance de la API régulation de vos charges de travail
- Consultez les AWS Trusted Advisor recommandations à grande échelle avec AWS Organizations
- Automatiser l'augmentation des limites de service et le support aux entreprises avec AWS Control Tower
- Actions, ressources et clés de condition pour Service Quotas

Vidéos connexes :

- AWS Live Re:inforce 2019 - Service Quotas
- Afficher et gérer les quotas pour les AWS services à l'aide de quotas de service
- AWS IAMDémo de quotas
- AWS re:Invent 2018 : Boucles serrées et ouverture d'esprit : comment prendre le contrôle des systèmes, grands et petits

Outils associés :

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager

- AWS Marketplace

REL01-BP04 Surveiller et gérer les quotas

Évaluez votre utilisation potentielle et augmentez vos quotas de manière appropriée afin d'assurer une croissance planifiée de l'utilisation.

Résultat escompté : des systèmes actifs et automatisés de gestion et de surveillance ont été déployés. Ces solutions opérationnelles permettent de s'assurer que les seuils d'utilisation des quotas sont sur le point d'être atteints. Les changements de quotas demandés permettraient de remédier à ces problèmes de manière proactive.

Anti-modèles courants :

- Absence de configuration de la surveillance pour vérifier les seuils de quotas de services
- Absence de configuration de la surveillance des limites strictes, même si ces valeurs ne peuvent pas être modifiées.
- En supposant que le délai nécessaire pour demander et obtenir un changement de quota souple soit immédiat ou de courte durée.
- Configuration d'alarmes d'approche des quotas de services, mais sans processus sur la façon de répondre à une alerte.
- Configurez les alarmes uniquement pour les services pris en charge par les AWS Service Quotas et ne surveillez pas les autres AWS services.
- Non-prise en compte de la gestion des quotas pour les conceptions de résilience à régions multiples, comme les approches actives/actives, actives/passives : à chaud, actives/passives : à froid et actives/passives : veilleuse..
- Absence d'évaluation des différences de quotas entre les régions.
- Absence d'évaluation des besoins de chaque région pour une demande spécifique d'augmentation de quota.
- Absence d'utilisation de modèles pour la gestion des quotas multirégionaux.

Avantages de l'établissement de cette meilleure pratique : le suivi automatique des Quotas du AWS Service et le suivi de votre utilisation par rapport à ces quotas vous permettront de savoir quand vous approchez d'une limite de quota. Vous pouvez également utiliser ces données de surveillance pour limiter les dégradations dues à l'épuisement des quotas.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Pour les services pris en charge, vous pouvez surveiller vos quotas en configurant différents services qui peuvent évaluer et ensuite envoyer des alertes ou des alarmes. Cela peut aider à surveiller l’utilisation et vous alerter sur l’approche des quotas. Ces alarmes peuvent être invoquées depuis AWS Config les fonctions Lambda CloudWatch, Amazon ou depuis. AWS Trusted Advisor Vous pouvez également utiliser des filtres métriques sur les CloudWatch journaux pour rechercher et extraire des modèles dans les journaux afin de déterminer si l'utilisation approche les seuils de quota.

Étapes d’implémentation

Pour la surveillance :

- Enregistrez la consommation des ressources actuelles (par exemple, les compartiments, ou les instances). Utilisez API les opérations de service, telles que Amazon EC2 `DescribeInstances`API, pour collecter la consommation actuelle de ressources.
- Saisissez vos quotas actuels qui sont essentiels et applicables aux services utilisant ce qui suit :
  - AWS Quotas de service
  - AWS Trusted Advisor
  - AWS documentation
  - AWS pages spécifiques au service
  - AWS Command Line Interface (AWS CLI)
  - AWS Cloud Development Kit (AWS CDK)
- Utilisez AWS Service Quotas, un AWS service qui vous permet de gérer vos quotas pour plus de 250 AWS services à partir d'un seul emplacement.
- Utilisez les limites de Trusted Advisor service pour surveiller vos limites de service actuelles à différents seuils.
- Utilisez l'historique des quotas de service (console ou AWS CLI) pour vérifier les augmentations régionales.
- Comparez les changements de quotas de services dans chaque région et chaque compte pour créer une équivalence, si nécessaire.

Pour la gestion :

- Automatisé : définissez une règle AWS Config personnalisée pour analyser les quotas de service entre les régions et comparer les différences.
- Automatisé : configurez une fonction Lambda programmée pour analyser les quotas de services dans les régions et comparer les différences.
- Manuel : Scannez les quotas de services par le biais AWS CLI de AWS la console ou analysez les quotas de services entre les régions et comparez les différences. API Signalez les différences.
- Si des différences de quotas sont identifiées entre les régions, demandez un changement de quota, si nécessaire.
- Passez en revue le résultat de toutes les demandes.

Ressources

Bonnes pratiques associées :

- REL01-BP01 Connaître les quotas et les contraintes de service
- REL01-BP02 Gérer les quotas de service entre les comptes et les régions
- REL01-BP03 Respecter les quotas et les contraintes de service fixes grâce à l'architecture
- REL01-BP05 Automatisation de la gestion des quotas
- REL01-BP06 Assurez-vous qu'il existe un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement
- REL03-BP01 Choisissez comment segmenter votre charge de travail
- REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la guérison sur toutes les couches
- REL12-BP04 Tester la résilience à l’aide de l’ingénierie du chaos

Documents connexes :

- AWS Le pilier de fiabilité du framework Well-Architected : la disponibilité
- AWS Quotas de service (anciennement appelés limites de service)
- AWS Trusted Advisor Contrôles des meilleures pratiques (voir la section Limites de service)
- AWS limitez le nombre de AWS réponses

- Limites EC2 de service Amazon
- Qu’est-ce que Service Quotas ?
- Comment demander une augmentation de quota
- Points de terminaison et quotas de service
- Guide de l’utilisateur de Service Quotas
- Quota Monitor pour AWS
- AWS Limites d'isolation des défauts
- Disponibilité avec redondance
- AWS pour les données
- Qu’est-ce que l’intégration continue ?
- Qu’est-ce que la livraison continue ?
- APNPartenaire : partenaires qui peuvent vous aider à gérer la configuration
- Gestion du cycle de vie des comptes dans les environnements account-per-tenant SaaS sur AWS
- Gestion et surveillance de la API régulation de vos charges de travail
- Consultez les AWS Trusted Advisor recommandations à grande échelle avec AWS Organizations
- Automatiser l'augmentation des limites de service et le support aux entreprises avec AWS Control Tower
- Actions, ressources et clés de condition pour Service Quotas

Vidéos connexes :

- AWS Live Re:inforce 2019 - Service Quotas
- Afficher et gérer les quotas pour les AWS services à l'aide de quotas de service
- AWS IAMDémo de quotas
- AWS re:Invent 2018 : Boucles serrées et ouverture d'esprit : comment prendre le contrôle des systèmes, grands et petits

Outils associés :

- AWS CodeDeploy
- AWS CloudTrail

- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP05 Automatisation de la gestion des quotas

Les quotas de service, également appelés limites dans les services AWS, sont les valeurs maximales pour les ressources de votre Compte AWS. Chaque service AWS définit un ensemble de quotas et leurs valeurs par défaut. Pour permettre à votre charge de travail d'accéder à toutes les ressources dont elle a besoin, vous devrez peut-être augmenter les valeurs de vos quotas de service.

L'augmentation de la consommation de AWS ressources par la charge de travail peut menacer la stabilité de la charge de travail et avoir un impact sur l'expérience utilisateur en cas de dépassement des quotas. Mettez en œuvre des outils qui vous alerteront quand votre charge de travail approchera des limites et envisagez de créer automatiquement des demandes d'augmentation de quotas.

Résultat escompté : les quotas sont configurés de manière appropriée pour les charges de travail exécutées dans chaque Compte AWS et chaque région.

Anti-modèles courants :

- vous ne tenez pas compte des quotas et ne les ajustez pas de manière appropriée pour répondre aux exigences de charge de travail.
- Vous suivez les quotas et l'utilisation à l'aide de méthodes qui peuvent devenir obsolètes, comme les feuilles de calcul.
- Vous ne mettez à jour les limites de service que lors de planifications périodiques.
- Votre organisation ne dispose pas de processus opérationnels permettant de revoir les quotas existants et de demander des augmentations de quotas de service si nécessaire.

Avantages liés au respect de cette bonne pratique :

- Résilience améliorée de la charge de travail : vous évitez les erreurs causées par le dépassement des quotas de ressources AWS.
- Reprise après sinistre simplifiée : vous pouvez réutiliser les mécanismes de gestion automatique des quotas intégrés dans la région principale lors de la configuration de la reprise après sinistre dans une autre Région AWS.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Consultez les quotas actuels et suivez la consommation continue des quotas via des mécanismes tels que la console AWS Service Quotas, AWS Command Line Interface (AWS CLI) et les kits AWS SDK. Vous pouvez également intégrer vos bases de données de gestion des configurations (CMDB) et vos systèmes de gestion des services informatiques (ITSM) avec les API AWS Service Quotas.

Générez des alertes automatiques si l'utilisation des quotas atteint les seuils que vous avez définis, et définissez un processus pour soumettre les demandes d'augmentation de quotas lorsque vous recevez des alertes. Si la charge de travail sous-jacente est essentielle pour votre entreprise, vous pouvez automatiser les demandes d'augmentation de quotas, mais testez soigneusement l'automatisation pour éviter le risque d'une action incontrôlée, telle qu'une boucle de rétroaction sur la croissance.

Les petites augmentations de quotas sont souvent automatiquement approuvées. Les demandes de quotas plus importants peuvent nécessiter un traitement manuel par AWS Support, et leur examen et leur traitement peuvent prendre plus de temps. Prévoyez du temps supplémentaire pour traiter plusieurs demandes ou des demandes d'augmentation importante.

Étapes d'implémentation

- Mettez en œuvre une surveillance automatisée des quotas de service et émettez des alertes si la croissance de l'utilisation des ressources de votre charge de travail approche de vos limites de quotas. Par exemple, Quota Monitor for AWS peut assurer la surveillance automatisée des quotas de service. Cet outil s'intègre à AWS Organizations et se déploie à l'aide de Cloudformation StackSets, de sorte que les nouveaux comptes sont automatiquement surveillés à leur création.
- Utilisez des fonctionnalités telles que les modèles de demande de quotas de service ou AWS Control Tower pour simplifier la configuration de Service Quotas pour les nouveaux comptes.
- Créez des tableaux de bord de l'utilisation actuelle de vos quotas de service dans tous les Comptes AWS et toutes les régions, et faites-y référence si nécessaire pour éviter de dépasser

vos quotas. Le tableau de bord Trusted Advisor Organizational (TAO), qui fait partie du framework Cloud Intelligence Dashboards, peut vous aider à rapidement prendre en main un tel tableau de bord.

- Effectuez le suivi des demandes d'augmentation de limite de service. La page Consolidated Insights from Multiple Accounts (CIMA) peut fournir une vue de toutes vos demandes au niveau de l'organisation.
- Testez la génération d'alertes et l'automatisation de toute demande d'augmentation de quotas en définissant des seuils de quotas inférieurs dans les comptes hors production. N'effectuez pas ces tests dans un compte de production.

Ressources

Bonnes pratiques associées :

- OPS10-BP07 Automatisation des réponses aux événements

Documents connexes :

- Partenaire APN : partenaires facilitant la gestion de la configuration
- AWS Marketplace : produits CMDB facilitant le suivi des limites
- AWS Service Quotas (anciennement Service Limits)
- Vérifications des bonnes pratiques AWS Trusted Advisor (voir la section Service Limits)
- Solution Quota Monitor sur AWS – Solution AWS
- Qu'est-ce que Service Quotas ?
- Que sont les modèles de demande de quotas de service ?

Vidéos connexes :

- AWS Live re:Inforce 2019 - Service Quotas
- Automatisation de l'augmentation des limites de service et du support aux entreprises avec AWS Control Tower

Outils associés :

- Quota Monitor for AWS

REL01-BP06 Assurez-vous qu'il existe un écart suffisant entre les quotas actuels et l'utilisation maximale pour permettre le basculement

Cet article explique comment maintenir l’espace entre le quota de ressources et votre utilisation, et comment cela peut être bénéfique pour votre organisation. Une fois que vous avez fini d’utiliser une ressource, le quota d’utilisation peut continuer à prendre en compte cette ressource. Cela peut entraîner une défaillance ou une inaccessibilité de la ressource. Empêchez la défaillance de ressource en vérifiant que vos quotas couvrent le chevauchement des ressources inaccessibles et leurs remplacements. Prenez en compte les cas d’utilisation tels que les pannes de réseau, la panne de la zone de disponibilité ou les pannes régionales lorsque vous calculez cet écart.

Résultat escompté : les défaillances mineures ou importantes en matière de ressources ou d’accessibilité des ressources peuvent être couvertes dans les limites des seuils de service actuels. Les pannes de zone, les pannes de réseau, voire les pannes régionales ont été prises en compte dans la planification des ressources.

Anti-modèles courants :

- Définition de quotas de service en fonction des quotas actuels sans tenir compte des scénarios de basculement.
- Non prise en compte des principes de stabilité statique lors du calcul du quota de pointe pour un service.
- Non prise en compte du potentiel des ressources inaccessibles dans le calcul du quota total nécessaire pour chaque région.
- Ne pas tenir compte des limites d'isolation des défaillances de AWS service pour certains services et de leurs modèles d'utilisation anormaux potentiels.

Avantages de l’établissement de cette bonne pratique : lorsque des interruptions de service ont un impact sur la disponibilité des applications, utilisez le cloud pour mettre en œuvre des stratégies de reprise après ces événements. Un exemple de stratégie consiste à créer des ressources supplémentaires pour remplacer les ressources inaccessibles afin de s’adapter aux conditions de basculement sans épuiser votre limite de service.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Lors de l’évaluation d’une limite de quotas, il faut tenir compte des cas de basculement qui pourraient survenir en raison d’une certaine dégradation. Prenez en compte les cas de basculement suivants.

- Interrompue ou inaccessibleVPC.
- Sous-réseau inaccessible.
- Zone de disponibilité dégradée qui a un impact sur l'accessibilité des ressources.
- Des itinéraires de mise en réseau ou des points d'entrée et de sortie sont bloqués ou modifiés.
- Région dégradée qui a un impact sur l'accessibilité des ressources.
- Sous-ensemble de ressources affectées par une défaillance dans une région ou une zone de disponibilité.

La décision de basculement est unique selon la situation, car l'impact sur l'entreprise peut varier. Abordez la planification de la capacité des ressources sur le site de basculement et les quotas des ressources avant de décider de basculer une application ou un service.

Tenez compte des pics d'activité supérieurs à la normale lors de l'examen des quotas pour chaque service. Ces pics peuvent être liés à des ressources inaccessibles en raison de la mise en réseau ou des autorisations, mais toujours actives. Les ressources actives non résiliées comptent dans la limite du quota de service.

Étapes d'implémentation

- Maintenez un espace entre votre quota de service et votre utilisation maximale pour faire face à un basculement ou à une perte d'accessibilité.
- Déterminez vos quotas de service. Tenez compte des modèles de déploiement typiques, des exigences de disponibilité et de la croissance de la consommation.
- Demandez des augmentations de quota si nécessaire. Prévoyez un temps d'attente pour la demande d'augmentation de quota.
- Déterminez vos exigences de fiabilité (également connues sous le nom de « nombre de neuf »).
- Comprenez les scénarios de panne potentiels tels que la perte d'un composant, d'une zone de disponibilité ou d'une région.
- Définissez votre méthodologie de déploiement (par exemple, canary, bleu/vert, rouge/noir et roulement).
- Incluez une mémoire tampon appropriée pour la limite actuelle de quota. Un exemple de tampon pourrait être de 15 %.
- Incluez les calculs de stabilité statique (zonale et régionale), le cas échéant.
- Anticipez la croissance de la consommation et surveillez vos tendances de consommation.

- Songez à l'impact de la stabilité statique pour vos charges de travail les plus critiques. Évaluez les ressources conformes à un système statiquement stable dans toutes les régions et zones de disponibilité.
- Envisagez l'utilisation de réserves de capacité à la demande pour programmer la capacité avant tout basculement. L'implémentation de cette stratégie est utile pour les calendriers d'activité critiques afin de réduire les risques potentiels liés à l'obtention de la bonne quantité et du bon type de ressources lors du basculement.

Ressources

Bonnes pratiques associées :

- REL01-BP01 Connaître les quotas et les contraintes de service
- REL01-BP02 Gérer les quotas de service entre les comptes et les régions
- REL01-BP03 Respecter les quotas et les contraintes de service fixes grâce à l'architecture
- REL01-BP04 Surveiller et gérer les quotas
- REL01-BP05 Automatisation de la gestion des quotas
- REL03-BP01 Choisissez comment segmenter votre charge de travail
- REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la guérison sur toutes les couches
- REL12-BP04 Tester la résilience à l'aide de l'ingénierie du chaos

Documents connexes :

- AWS Le pilier de fiabilité du framework Well-Architected : la disponibilité
- AWS Quotas de service (anciennement appelés limites de service)
- AWS Trusted Advisor Contrôles des meilleures pratiques (voir la section Limites de service)
- AWS limitez le nombre de AWS réponses
- Limites EC2 de service Amazon
- Qu'est-ce que Service Quotas ?
- Comment demander une augmentation de quota
- Points de terminaison et quotas de service

- Guide de l’utilisateur de Service Quotas
- Quota Monitor pour AWS
- AWS Limites d'isolation des défauts
- Disponibilité avec redondance
- AWS pour les données
- Qu’est-ce que l’intégration continue ?
- Qu’est-ce que la livraison continue ?
- APNPartenaire : partenaires qui peuvent vous aider à gérer la configuration
- Gestion du cycle de vie des comptes dans les environnements account-per-tenant SaaS sur AWS
- Gestion et surveillance de la API régulation de vos charges de travail
- Consultez les AWS Trusted Advisor recommandations à grande échelle avec AWS Organizations
- Automatiser l'augmentation des limites de service et le support aux entreprises avec AWS Control Tower
- Actions, ressources et clés de condition pour Service Quotas

Vidéos connexes :

- AWS Live Re:inforce 2019 - Service Quotas
- Afficher et gérer les quotas pour les AWS services à l'aide de quotas de service
- AWS IAMDémo de quotas
- AWS re:Invent 2018 : Boucles serrées et ouverture d'esprit : comment prendre le contrôle des systèmes, grands et petits

Outils associés :

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config

- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

FIA 2. Comment planifier la topologie de votre réseau ?

Les charges de travail existent souvent dans plusieurs environnements. Il s'agit notamment de plusieurs environnements Cloud (accessibles au public et privés) et éventuellement de votre infrastructure de centre de données existante. Les plans doivent inclure des considérations réseau telles que la connectivité intrasystème et intersystème, la gestion des adresses IP publiques, la gestion des adresses IP privées et la résolution des noms de domaine.

Bonnes pratiques

- REL02-BP01 Utiliser une connectivité réseau hautement disponible pour vos points de terminaison publics de charge de travail
- REL02-BP02 Fournir une connectivité redondante entre les réseaux privés dans le cloud et les environnements sur site
- REL02-BP03 Garantir que l'allocation des sous-réseaux IP tient compte de l'extension et de la disponibilité
- REL02-BP04 Préférer les topologies en étoile au maillage « many-to-many »
- REL02-BP05 Appliquer des plages d'adresses IP privées qui ne se chevauchent pas dans tous les espaces d'adressage privés auxquels ils sont connectés

REL02-BP01 Utiliser une connectivité réseau hautement disponible pour vos points de terminaison publics de charge de travail

La mise en place d'une connectivité réseau hautement disponible aux points de terminaison publics de vos charges de travail peut vous aider à réduire les temps d'arrêt dus à la perte de connectivité et à améliorer la disponibilité et le SLA de votre charge de travail. Pour ce faire, utilisez le DNS hautement disponible, les réseaux de diffusion de contenu (CDN), des passerelles API, l'équilibrage de charge ou les proxys inverses.

Résultat escompté : il est essentiel de planifier, de créer et de rendre opérationnelle une connectivité réseau hautement disponible pour vos points de terminaison publics. Si votre charge de travail

devient inaccessible en raison d'une perte de connectivité, même si elle est en cours d'exécution et disponible, vos clients verront votre système comme étant en panne. En combinant une connectivité réseau hautement disponible et résiliente pour les points de terminaison publics de votre charge de travail, ainsi qu'une architecture résiliente pour votre charge de travail elle-même, vous pouvez offrir la meilleure disponibilité et le meilleur niveau de service possible à vos clients.

AWS Global Accelerator, Amazon CloudFront, Amazon API Gateway, les URL de fonction AWS Lambda, les API AWS AppSync et Elastic Load Balancing (ELB) fournissent tous des points de terminaison publics hautement disponibles. Amazon Route 53 fournit un service DNS hautement disponible pour la résolution des noms de domaine afin de vérifier que vos adresses de point de terminaison publiques peuvent être résolues.

Vous pouvez également évaluer des appliances logicielles AWS Marketplace pour l'équilibrage de charge et les proxys.

Anti-modèles courants :

- Concevoir une charge de travail hautement disponible sans planifier le DNS et la connectivité réseau pour la haute disponibilité.
- Utiliser des adresses Internet publiques sur des instances ou des conteneurs individuels et gestion de la connectivité à ces adresses avec le DNS.
- Utiliser des adresses IP au lieu des noms de domaine pour localiser les services.
- Ne pas tester des scénarios où la connectivité à vos points de terminaison publics est perdue.
- Ne pas analyser les besoins en débit du réseau et les modèles de distribution.
- Ne pas tester et planifier des scénarios dans lesquels la connectivité du réseau Internet aux points de terminaison publics de votre charge de travail pourrait être interrompue.
- Fournir du contenu (comme les pages web, les ressources statiques ou les fichiers multimédias) à une grande zone géographique sans utiliser un réseau de diffusion de contenu.
- Ne pas se préparer aux attaques par déni de service distribué (DDoS). Les attaques DDoS risquent d'interrompre le trafic légitime et de réduire la disponibilité pour vos utilisateurs.

Avantages du respect de cette bonne pratique : la conception d'une connectivité réseau hautement disponible et résiliente garantit que votre charge de travail est accessible et disponible pour vos utilisateurs.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Le routage du trafic est au cœur de la mise en place d'une connectivité réseau hautement disponible pour vos points de terminaison publics. Pour vérifier que votre trafic est en mesure d'atteindre les points de terminaison, le DNS doit être capable de résoudre les noms de domaine à leurs adresses IP correspondantes. Utilisez un système de nom de domaine (DNS) hautement disponible et évolutif tel qu'Amazon Route 53 pour gérer les enregistrements DNS de votre domaine. Vous pouvez également utiliser les surveillances de l'état fournies par Amazon Route 53. Les surveillances de l'état permettent de s'assurer que votre application est accessible, disponible et fonctionnelle. Elles peuvent être configurées de manière à imiter le comportement de l'utilisateur, comme la demande d'une page web ou d'une URL spécifique. En cas de panne, Amazon Route 53 répond aux demandes de résolution DNS et dirige uniquement le trafic vers les points de terminaison en bonne santé. Vous pouvez également envisager d'utiliser les fonctionnalités de Geo DNS et de routage basé sur la latence offertes par Amazon Route 53.

Pour vérifier que votre charge de travail elle-même est hautement disponible, utilisez Elastic Load Balancing (ELB). Amazon Route 53 peut être utilisé pour cibler le trafic vers ELB, qui le distribue aux instances de calcul cibles. Vous pouvez également utiliser Amazon API Gateway avec AWS Lambda pour une solution sans serveur. Les clients peuvent également exécuter des charges de travail dans plusieurs Régions AWS. Avec un modèle actif/actif multisite, la charge de travail peut desservir le trafic provenant de plusieurs régions. Avec un schéma actif/passif multisite, la charge de travail sert le trafic provenant de la région active tandis que les données sont répliquées vers la région secondaire et deviennent actives en cas de panne dans la région principale. Les surveillances de l'état de Route 53 peuvent ensuite être utilisées pour contrôler le basculement DNS depuis n'importe quel point de terminaison d'une région principale vers un point de terminaison d'une région secondaire, en vérifiant que votre charge de travail est accessible et disponible pour vos utilisateurs.

Amazon CloudFront fournit une API simple pour distribuer du contenu avec une faible latence et des taux de transfert de données élevés en répondant aux demandes à l'aide d'un réseau d'emplacements périphériques dans le monde entier. Les réseaux de diffusion de contenu (CDN) desservent les clients en proposant un contenu situé ou mis en cache à un endroit proche de l'utilisateur. Cela améliore également la disponibilité de votre application, car la charge de contenu est transférée de vos serveurs vers les emplacements périphériques de CloudFront. Les emplacements périphériques et les caches périphériques régionaux conservent des copies en cache de votre contenu à proximité de vos utilisateurs, ce qui permet une récupération rapide et augmente l'accessibilité et la disponibilité de votre charge de travail.

Pour les charges de travail avec des utilisateurs dispersés géographiquement, AWS Global Accelerator améliore la disponibilité et les performances des applications. AWS Global Accelerator fournit des adresses IP statiques Anycast qui servent de point d'entrée fixe à votre application hébergée dans une ou plusieurs Régions AWS. Cela permet au trafic d'entrer sur le réseau mondial AWS aussi près que possible de vos utilisateurs, améliorant ainsi l'accessibilité et la disponibilité de votre charge de travail. AWS Global Accelerator surveille également l'état de santé des points de terminaison de vos applications en utilisant la surveillance de l'état TCP, HTTP et HTTPS. Toute modification de l'état ou de la configuration de vos points de terminaison permet la redirection du trafic utilisateur vers des points de terminaison sains qui offrent les meilleures performances et la meilleure disponibilité à vos utilisateurs. De plus, AWS Global Accelerator est conçu pour être isolé des pannes et utilise deux adresses IPv4 statiques qui sont desservies par des zones réseau indépendantes, ce qui augmente la disponibilité de vos applications.

Pour aider à protéger les clients contre les attaques DDoS, AWS fournit AWS Shield Standard. Shield Standard est activé automatiquement et protège contre les attaques d'infrastructure courantes (couches 3 et 4) telles que les inondations SYN/UDP et les attaques par réflexion afin de garantir la haute disponibilité de vos applications sur AWS. Pour bénéficier de protections supplémentaires contre des attaques plus sophistiquées et plus importantes (comme les inondations UDP), les attaques par épuisement d'état (comme les inondations TCP SYN), et pour aider à protéger vos applications fonctionnant sur Amazon Elastic Compute Cloud (Amazon EC2), Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator et Route 53, vous pouvez envisager d'utiliser AWS Shield Advanced. Pour se protéger contre les attaques au niveau de la couche application, comme les inondations HTTP POST ou GET, utilisez AWS WAF. AWS WAF peut utiliser les adresses IP, les en-têtes HTTP, le corps HTTP, les chaînes URI, l'injection SQL et les conditions de script intersite pour déterminer si une requête doit être bloquée ou autorisée.

Étapes d'implémentation

1. Définir un DNS hautement disponible : Amazon Route 53 est un service web de système de nom de domaine (DNS) hautement disponible et évolutif. Route 53 connecte les demandes des utilisateurs aux applications Internet exécutées sur AWS ou sur site. Pour plus d'informations, consultez Configuration d'Amazon Route 53 en tant que service DNS.
2. Configurez la surveillance de l'état : lorsque vous utilisez Route 53, vérifiez que seules les cibles saines sont résolubles. Commencez par la création de surveillance de l'état Route 53 et la configuration du basculement DNS. Il est important de tenir compte des aspects suivants lors de la mise en place des surveillances de l'état :
   a. Comment Amazon Route 53 détermine si une surveillance de l'état est saine

   b. Création, mise à jour et suppression de surveillances de l'état
   c. Surveillance du statut de la surveillance de l'état et obtention de notifications
   d. Bonnes pratiques relatives à Amazon Route 53 DNS
3. Connectez votre service DNS à vos points de terminaison.
   a. Lorsque vous utilisez Elastic Load Balancing comme cible pour votre trafic, créez un enregistrement d'alias à l'aide d'Amazon Route 53 qui pointe vers le point de terminaison régional de votre équilibreur de charge. Pendant la création de l'enregistrement de l'alias, réglez l'option « Évaluer l'état de la cible » sur « Oui ».
   b. Pour les charges de travail sans serveur ou les API privées lorsqu'une passerelle API est utilisée, utilisez Route 53 pour diriger le trafic vers la passerelle API.
4. Choisissez un réseau de diffusion de contenu.
   a. Pour diffuser du contenu en utilisant des emplacements périphériques plus proches de l'utilisateur, commencez par comprendre comment CloudFront diffuse le contenu.
   b. Démarrez avec une distribution CloudFront simple. CloudFront comprend alors l'endroit d'où vous souhaitez que le contenu soit diffusé, ainsi que les détails concernant le suivi et la gestion de la diffusion du contenu. Il est important de comprendre et de prendre en compte les aspects suivants lors de la mise en place de la distribution CloudFront :
      i. Fonctionnement de la mise en cache avec les emplacements périphériques CloudFront
      ii. Augmentation de la proportion de demandes servies directement à partir des caches CloudFront (taux d'accès au cache)
      iii. Utilisation d'Amazon CloudFront Origin Shield
      iv. Optimisation de la haute disponibilité avec le basculement d'origine CloudFront
5. Configurez la protection de la couche d'application : AWS WAF vous aide à vous protéger contre les exploits et les bots web courants qui peuvent affecter la disponibilité, compromettre la sécurité ou consommer des ressources excessives. Pour mieux comprendre, découvrez comment AWS WAF fonctionne et quand vous serez prêt à mettre en œuvre des protections contre les inondations HTTP POST AND GET de la couche d'application, consultez Démarrer avec AWS WAF. Vous pouvez également utiliser AWS WAF avec CloudFront. Consultez la documentation sur le fonctionnement de AWS WAF avec des fonctionnalités d'Amazon CloudFront.
6. Configurez une protection DDoS supplémentaire : par défaut, tous les clients AWS bénéficient d'une protection contre les attaques DDoS les plus fréquentes au niveau de la couche réseau et de la couche transport qui ciblent votre site web ou votre application avec AWS Shield Standard, et ce sans frais supplémentaires. Pour une protection supplémentaire des applications connectées

à Internet exécutées sur Amazon EC2, Elastic Load Balancing, Amazon CloudFront, AWS Global Accelerator et Amazon Route 53, vous pouvez prendre en compte AWS Shield Advanced et examiner des exemples d'architectures résilientes aux attaques DDoS. Pour protéger votre charge de travail et vos points de terminaison publics contre les attaques DDoS, consultez Démarrer avec AWS Shield Advanced.

Ressources

Bonnes pratiques associées :

- REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements
- REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle lors de la restauration
- REL11-BP06 Envoyer des notifications lorsque des événements ont un impact sur la disponibilité

Documents connexes :

- Partenaire APN : partenaires pouvant vous aider à planifier votre mise en réseau
- AWS Marketplace pour l'infrastructure réseau
- Présentation de AWS Global Accelerator
- Qu'est-ce qu'Amazon CloudFront ?
- Qu'est-ce qu'Amazon Route 53 ?
- Qu'est-ce qu'Elastic Load Balancing ?
- Capacité de connectivité réseau : établir les bases de votre cloud
- Qu'est-ce qu'Amazon API Gateway ?
- Que sont AWS WAF, AWS Shield et AWS Firewall Manager ?
- Qu'est-ce qu'Amazon Application Recovery Controller ?
- Configurer des surveillances de l'état personnalisées pour le basculement DNS

Vidéos connexes :

- AWS re:Invent 2022 - Improve performance and availability with AWS Global Accelerator
- AWS re:Invent 2020: Global traffic management with Amazon Route 53

- AWS re:Invent 2022 - Operating highly available Multi-AZ applications
- AWS re:Invent 2022 - Dive deep on AWS networking infrastructure
- AWS re:Invent 2022 - Building resilient networks

Exemples associés :

- Reprise après sinistre avec Amazon Application Recovery Controller (ARC)
- Ateliers de fiabilité
- Atelier AWS Global Accelerator

REL02-BP02 Fournir une connectivité redondante entre les réseaux privés dans le cloud et les environnements sur site

Mettez en œuvre une redondance dans vos connexions entre les réseaux privés dans le cloud et les environnements sur site pour obtenir la résilience de la connectivité. Cela peut se faire en déployant au moins deux liens et chemins de trafic, afin de préserver la connectivité en cas de défaillance du réseau.

Anti-modèles courants :

- Vous dépendez d'une seule connexion réseau, ce qui crée un point de défaillance unique.
- Vous n'utilisez qu'un seul VPN tunnel ou plusieurs tunnels qui se terminent dans la même zone de disponibilité.
- Vous comptez sur l'un d'ISPentre eux pour la VPN connectivité, ce qui peut entraîner des pannes complètes en cas de panneISP.
- Ne pas implémenter des protocoles de routage dynamiques tels que BGP ceux qui sont essentiels pour rediriger le trafic en cas de perturbation du réseau.
- Vous ignorez les limites de bande passante des VPN tunnels et vous surestimez leurs capacités de sauvegarde.

Avantages du respect de cette bonne pratique : l'implémentation d'une connectivité redondante entre votre environnement cloud et votre environnement d'entreprise/sur site permet aux services dépendants entre les deux environnements de communiquer de manière fiable.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d’implémentation

Lorsque vous connectez votre réseau local AWS Direct Connect à AWS, vous pouvez obtenir une résilience maximale du réseau (SLAde 99,99 %) en utilisant des connexions distinctes qui se terminent sur des appareils distincts situés dans plusieurs sites sur site et sur plusieurs sites. AWS Direct Connect Cette topologie offre une résilience contre les pannes d’appareils, les problèmes de connectivité et les pannes complètes d’un site. Vous pouvez également obtenir une résilience élevée (SLAde 99,9 %) en utilisant deux connexions individuelles vers plusieurs sites (chaque site sur site étant connecté à un seul site Direct Connect). Cette approche assure une protection contre les interruptions de connectivité causées par des coupures de fibre ou des pannes d’appareils, et contribue à pallier des défaillances complètes d’un site. Le AWS Direct Connect Resiliency Toolkit peut vous aider à concevoir votre AWS Direct Connect topologie.

Vous pouvez également envisager de AWS Site-to-Site VPN mettre fin AWS Transit Gateway à une sauvegarde rentable de votre AWS Direct Connect connexion principale. Cette configuration permet un routage multichemin (ECMP) à coût égal sur plusieurs VPN tunnels, permettant un débit allant jusqu'à 50 Gbit/s, même si chaque VPN tunnel est plafonné à 1,25 Gbit/s. Il est toutefois important de noter que cela AWS Direct Connect reste le choix le plus efficace pour minimiser les perturbations du réseau et fournir une connectivité stable.

Lorsque vous utilisez VPNs Internet pour connecter votre environnement cloud à votre centre de données sur site, configurez deux VPN tunnels dans le cadre d'une site-to-site VPN connexion unique. Chaque tunnel doit mener à une zone de disponibilité différente pour garantir la haute disponibilité, et utiliser du matériel redondant pour éviter une panne d’appareil sur site. En outre, envisagez plusieurs connexions Internet provenant de différents fournisseurs de services Internet (ISPs) sur votre site afin d'éviter toute interruption complète de la VPN connectivité due à une seule ISP panne. La sélection ISPs d'infrastructures et de routages variés, en particulier ceux dotés de chemins physiques distincts vers les AWS points de terminaison, garantit une haute disponibilité de la connectivité.

Outre la redondance physique avec plusieurs AWS Direct Connect connexions et plusieurs VPN tunnels (ou une combinaison des deux), la mise en œuvre du routage dynamique du Border Gateway Protocol (BGP) est également cruciale. Dynamic BGP permet de réacheminer automatiquement le trafic d'un chemin à un autre en fonction des conditions du réseau en temps réel et des politiques configurées. Ce comportement dynamique est particulièrement utile pour maintenir la disponibilité du réseau et la continuité des services en cas de panne de liaison ou de réseau. Il sélectionne rapidement des chemins alternatifs, améliorant ainsi la résilience et la fiabilité du réseau.

Étapes d’implémentation

- Acquérez une connectivité hautement disponible entre AWS et votre environnement sur site.
  - Utilisez plusieurs AWS Direct Connect connexions ou VPN tunnels entre des réseaux privés déployés séparément.
  - Utilisez plusieurs AWS Direct Connect sites pour une haute disponibilité.
  - Si vous en utilisez plusieurs Régions AWS, créez une redondance dans au moins deux d'entre eux.
- Utilisez AWS Transit Gateway, dans la mesure du possible, pour mettre fin à votre VPNconnexion.
- Évaluez les AWS Marketplace appareils auxquels vous VPNs souhaitez mettre fin ou étendre votre SD- WAN AWS. Si vous utilisez des AWS Marketplace appliances, déployez des instances redondantes pour une haute disponibilité dans différentes zones de disponibilité.
- Assurez-vous que vous disposez d’une connexion redondante à votre environnement sur site.
  - Vous pouvez avoir besoin de connexions redondantes à plusieurs Régions AWS pour répondre à vos besoins de disponibilité.
  - Utilisez la boîte à outils de résilience AWS Direct Connect pour démarrer.

Ressources

Documents connexes :

- AWS Direct Connect Recommandations en matière de résilience
- Utilisation de Site-to-Site VPN connexions redondantes pour assurer le basculement
- Politiques de routage et BGP communautés
- Configurations actives/actives et actives/passives dans AWS Direct Connect
- APNPartenaire : partenaires qui peuvent vous aider à planifier votre réseau
- AWS Marketplace pour l'infrastructure réseau
- Livre blanc sur les options de connectivité d’Amazon Virtual Private Cloud
- Création d'une infrastructure VPC AWS multiréseau évolutive et sécurisée
- Utilisation de Site-to-Site VPN connexions redondantes pour assurer le basculement
- Utiliser le AWS Direct Connect Resiliency Toolkit pour démarrer
- VPCTerminaux et services de point de VPC terminaison ()AWS PrivateLink
- Qu'est-ce qu'Amazon VPC ?

- Qu’est-ce qu’une passerelle de transit ?
- Qu'est-ce que c'est AWS Site-to-Site VPN ?
- Utilisation des passerelles Direct Connect

Vidéos connexes :

- AWS re:Invent 2018 : VPC conception avancée et nouvelles fonctionnalités pour Amazon VPC
- AWS re:Invent 2019 : des architectures de AWS Transit Gateway référence pour beaucoup VPCs

REL02-BP03 Garantir que l'allocation des sous-réseaux IP tient compte de l'extension et de la disponibilité

Les plages d'adresses VPC IP Amazon doivent être suffisamment grandes pour répondre aux exigences de charge de travail, notamment en tenant compte de l'expansion future et de l'allocation d'adresses IP aux sous-réseaux des zones de disponibilité. Cela inclut les équilibreurs de charge, les EC2 instances et les applications basées sur des conteneurs.

Lorsque vous planifiez votre topologie de réseau, la première étape consiste à définir l’espace d’adressage IP lui-même. Des plages d'adresses IP privées (conformément aux directives de RFC 1918) doivent être allouées à chacune d'ellesVPC. Vous devez remplir les exigences suivantes dans le cadre de ce processus :

- Autorisez l'espace d'adresses IP pour plusieurs adresses VPC par région.
- Dans unVPC, prévoyez de l'espace pour plusieurs sous-réseaux afin de pouvoir couvrir plusieurs zones de disponibilité.
- Envisagez de laisser l'espace de CIDR bloc inutilisé dans un espace VPC pour une future extension.
- Assurez-vous qu'il existe un espace d'adressage IP adapté aux besoins de tous les flottes transitoires d'EC2instances Amazon que vous pourriez utiliser, tels que les flottes ponctuelles pour le machine learning, les EMR clusters Amazon ou les clusters Amazon Redshift. Une attention similaire doit être accordée aux clusters Kubernetes, tels qu'Amazon Elastic Kubernetes Service EKS (Amazon), car chaque pod Kubernetes se voit attribuer une adresse routable à partir du bloc par défaut. VPC CIDR
- Notez que les quatre premières adresses IP et la dernière adresse IP de chaque CIDR bloc de sous-réseau sont réservées et ne sont pas disponibles pour votre usage.

- Notez que le VPC CIDR bloc initial qui vous est attribué VPC ne peut être ni modifié ni supprimé, mais vous pouvez ajouter des CIDR blocs supplémentaires qui ne se chevauchent pas au. VPC Le sous-réseau IPv4 CIDRs ne peut pas être modifié, mais c'est IPv6 CIDRs possible.
- Le plus grand VPC CIDR bloc possible est un /16, et le plus petit un /28.
- Envisagez d'autres réseaux connectés (VPCsur site ou autres fournisseurs de cloud) et veillez à ce que l'espace d'adresses IP ne se chevauche pas. Pour plus d'informations, voir REL02-BP05 Appliquer des plages d'adresses IP privées qui ne se chevauchent pas dans tous les espaces d'adressage privés auxquels ils sont connectés.

Résultat souhaité : un sous-réseau IP évolutif peut vous aider à faire face à la croissance future et à éviter tout gaspillage inutile.

Anti-modèles courants :

- Si l'on ne tient pas compte de la croissance future, les CIDR blocs sont trop petits et nécessitent une reconfiguration, ce qui peut entraîner des temps d'arrêt.
- Estimation incorrecte du nombre d’adresses IP qu’un Elastic Load Balancer peut utiliser.
- Déploiement de nombreux équilibreurs de charge à trafic élevé dans les mêmes sous-réseaux
- Utilisation de mécanismes de dimensionnement automatisés sans surveiller la consommation d’adresses IP.
- Définir des CIDR plages excessivement étendues bien au-delà des prévisions de croissance futures, ce qui peut entraîner des difficultés d'appairage avec d'autres réseaux dont les plages d'adresses se chevauchent.

Avantages de respecter cette bonne pratique : ces tailles sont la garantie que vous pouvez prendre en charge la croissance de vos charges de travail et continuer à fournir une disponibilité lors de l’augmentation verticale.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Planifiez votre réseau en prévision de votre croissance, de la conformité réglementaire et de son intégration avec d’autres composants. La croissance peut être sous-estimée, la conformité réglementaire peut changer, et les acquisitions ou les connexions à des réseaux privés peuvent être difficiles à implémenter sans une planification appropriée.

- Sélectionnez les régions Comptes AWS et les régions pertinentes en fonction de vos exigences en matière de service, de latence, de réglementation et de reprise après sinistre (DR).
- Identifiez vos besoins en matière de VPC déploiements régionaux.
- Identifiez la taille duVPCs.
  - Déterminez si vous comptez déployer la VPC multiconnectivité.
    - Qu’est-ce qu’une passerelle de transit ?
    - VPCConnectivité multirégionale unique
  - Déterminez si vous avez besoin d’une mise en réseau séparée pour les exigences réglementaires.
  - Fabriquez VPCs avec des CIDR blocs de taille appropriée pour répondre à vos besoins actuels et futurs.
    - Si vous avez des prévisions de croissance inconnues, vous pouvez opter pour des CIDR blocs plus grands afin de réduire le risque de reconfiguration future
  - Envisagez d'utiliser l'IPv6adressage pour les sous-réseaux dans le cadre d'une double VPC pile. IPv6est parfaitement adapté à une utilisation dans des sous-réseaux privés contenant des flottes d'instances éphémères ou des conteneurs qui nécessiteraient autrement un grand nombre d'adresses. IPv4

Ressources

Bonnes pratiques Well-Architected connexes :

- REL02-BP05 Appliquer des plages d'adresses IP privées qui ne se chevauchent pas dans tous les espaces d'adressage privés auxquels ils sont connectés

Documents connexes :

- APNPartenaire : partenaires qui peuvent vous aider à planifier votre réseau
- AWS Marketplace pour l'infrastructure réseau
- Livre blanc sur les options de connectivité d’Amazon Virtual Private Cloud
- Connectivité au réseau à haute disponibilité de plusieurs centres de données
- VPCConnectivité multirégionale unique
- Qu'est-ce qu'Amazon VPC ?
- IPv6sur AWS

- IPv6sur les architectures de référence
- Amazon Elastic Kubernetes Service lance le support IPv6
- Recommandations pour vos VPC équilibreurs de charge classiques
- Sous-réseaux de zone de disponibilité - Équilibreurs de charge des applications
- Zones de disponibilité - Équilibreurs de charge réseau

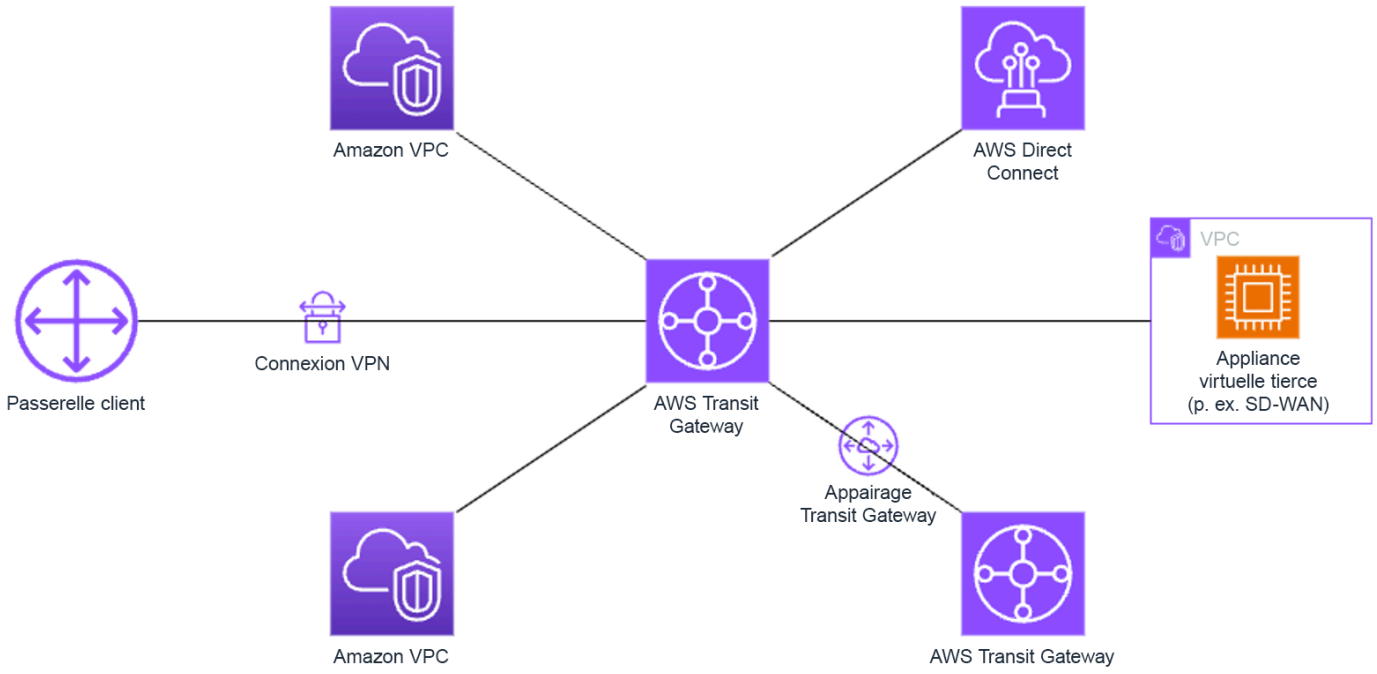
Vidéos connexes :

- AWS re:Invent 2018 : VPC Design avancé et nouvelles fonctionnalités pour Amazon VPC (03) NET3
- AWS re:Invent 2019 : des architectures de AWS Transit Gateway référence pour de nombreuses personnes VPCs (NET406-R1)
- AWS re:INVENT 2023 : AWS Prêts pour la suite ? Conception de réseaux pour la croissance et la flexibilité (NET310)

REL02-BP04 Préférer les topologies en étoile au maillage « many-to-many »

Lorsque vous connectez plusieurs réseaux privés, tels que des clouds privés virtuels (VPC) et des réseaux sur site, optez pour une topologie en étoile plutôt qu'une topologie maillée. Contrairement aux topologies maillées, où chaque réseau se connecte directement aux autres et augmente la complexité et les frais de gestion, l'architecture en étoile centralise les connexions via un hub unique. Cette centralisation simplifie la structure du réseau et améliore son opérabilité, sa capacité de mise à l'échelle et son contrôle.

AWS Transit Gateway est un service géré, évolutif et hautement disponible conçu pour la construction de réseaux en étoile sur AWS. Il fait office de hub central de votre réseau et assure la segmentation du réseau, le routage centralisé et la connexion simplifiée aux environnements cloud et sur site. La figure suivante montre comment utiliser AWS Transit Gateway pour créer une topologie en étoile.



Résultat escompté : vous avez connecté vos clouds privés virtuels (VPC) et vos réseaux sur site via un hub central. Vous configurez vos connexions d'appairage via le hub, qui agit comme un routeur cloud hautement évolutif. Le routage est simplifié car vous n'avez pas à travailler avec des relations d'appairage complexes. Le trafic entre les réseaux est chiffré et vous avez la possibilité d'isoler les réseaux.

Anti-modèles courants :

- Vous établissez des règles d'appairage réseau complexes.
- Vous fournissez des routes entre des réseaux qui ne doivent pas communiquer entre eux (par exemple, des charges de travail distinctes qui n'ont aucune interdépendance).
- La gouvernance de l'instance du hub est inefficace.

Avantages du respect de cette bonne pratique : à mesure que le nombre de réseaux connectés augmente, la gestion et le développement de la connectivité maillée deviennent de plus en plus difficiles. Une architecture maillée introduit des défis supplémentaires, tels que des composants d'infrastructure, des exigences de configuration et des considérations de déploiement supplémentaires. Le maillage introduit également une surcharge supplémentaire pour gérer et surveiller les composants du plan de données et du plan de contrôle. Vous devez réfléchir

à la manière d'assurer la haute disponibilité de l'architecture maillée, de surveiller l'état et les performances du maillage et de gérer les mises à niveau des composants du maillage.

Un modèle en étoile (hub and spoke), quant à lui, établit un routage centralisé du trafic sur plusieurs réseaux. Il fournit une approche plus simple de la gestion et de la surveillance des composants du plan de données et du plan de contrôle.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Créez un compte de services réseau s'il n'en existe pas. Placez le hub dans le compte de services réseau de l'organisation. Cette approche permet au hub d'être géré de manière centralisée par les ingénieurs réseau.

Le hub du modèle en étoile (hub and spoke) agit en tant que routeur virtuel pour le trafic circulant entre vos clouds privés virtuels (VPC) et les réseaux sur site. Cette approche réduit la complexité du réseau et facilite la résolution des problèmes de mise en réseau.

Tenez compte de la conception de votre réseau, notamment des VPC, d'AWS Direct Connect et des connexions VPN de site à site que vous souhaitez interconnecter.

Envisagez d'utiliser un sous-réseau distinct pour chaque attachement de VPC de passerelle de transit. Pour chaque sous-réseau, utilisez un petit CIDR (par exemple /28), afin d'avoir plus d'espace d'adressage pour les ressources de calcul. De plus, créez une liste ACL réseau et associez-la à tous les sous-réseaux qui sont associés au hub. Gardez la liste ACL réseau ouverte dans les directions entrantes et sortantes.

Concevez et implémentez vos tables de routage de telle sorte que les routes ne soient fournies qu'entre les réseaux qui doivent communiquer. Omettez les routes entre des réseaux qui ne doivent pas communiquer entre eux (par exemple, entre des charges de travail distinctes qui n'ont aucune interdépendance).

Étapes d'implémentation

1. Planifiez votre réseau. Déterminez les réseaux que vous souhaitez connecter et vérifiez qu'ils ne partagent pas de plages CIDR superposées.
2. Créez une passerelle AWS Transit Gateway et attachez-lui vos VPC.
3. Si nécessaire, créez des connexions VPN ou des passerelles Direct Connect et associez-les à la passerelle Transit Gateway.

4. Définissez la façon dont le trafic est acheminé entre les VPC connectés et les autres connexions via la configuration de vos tables de routage Transit Gateway.
5. Utilisez Amazon CloudWatch pour surveiller et ajuster les configurations selon les besoins afin d'optimiser les performances et les coûts.

Ressources

Bonnes pratiques associées :

- REL02-BP03 S'assurer que l'allocation des sous-réseaux IP tient compte de l'expansion et de la disponibilité
- REL02-BP05 Appliquer des plages d'adresses IP privées sans chevauchement dans tous les espaces d'adressage privés où elles sont connectées

Documents connexes :

- Qu'est-ce qu'une passerelle de transit ?
- Bonnes pratiques pour la conception de passerelles de transit
- Création d'une infrastructure réseau AWS multi-VPC évolutive et sécurisée
- Création d'un réseau global à l'aide de AWS Transit Gateway Inter-Region peering
- Amazon Virtual Private Cloud Connectivity Options
- Partenaire APN : partenaires pouvant vous aider à planifier votre mise en réseau
- AWS Marketplace pour l'infrastructure réseau

Vidéos connexes :

- AWS re:Invent 2023 - AWS networking foundations
- AWS re:Invent 2023 - Advanced VPC designs and new capabilities

Ateliers connexes :

- Atelier AWS Transit Gateway

REL02-BP05 Appliquer des plages d'adresses IP privées qui ne se chevauchent pas dans tous les espaces d'adressage privés auxquels ils sont connectés

Les plages d'adresses IP de chacun d'entre vous ne VPCs doivent pas se chevaucher lorsque vous êtes connecté, connecté via Transit Gateway ou connecté. VPN Évitez les conflits d'adresses IP entre les environnements a VPC et sur site ou avec les autres fournisseurs de cloud que vous utilisez. Vous devez également disposer d'un moyen d'allouer des plages d'adresses IP privées lorsque cela est nécessaire. Un système de gestion des adresses IP (IPAM) peut aider à automatiser cette opération.

Résultat escompté :

- Aucun conflit de plage d'adresses IP entre VPCs des environnements sur site ou d'autres fournisseurs de cloud.
- Une bonne gestion des adresses IP permet d'adapter plus facilement l'infrastructure réseau à la croissance et à l'évolution des exigences en matière de réseau.

Anti-modèles courants :

- En utilisant la même plage d'adresses IP que celle VPC que vous avez sur site, dans votre réseau d'entreprise ou chez d'autres fournisseurs de cloud
- Ne pas suivre les plages d'adresses IP VPCs utilisées pour déployer vos charges de travail.
- Utilisation de processus manuels de gestion des adresses IP, tels que des feuilles de calcul.
- Surdimensionner ou sous-dimensionner les CIDR blocs, ce qui entraîne un gaspillage d'adresses IP ou un espace d'adressage insuffisant pour votre charge de travail.

Avantages du respect de cette bonne pratique : la planification active de votre réseau garantit que vous n'avez pas plusieurs occurrences de la même adresse IP dans les réseaux interconnectés. Cela empêche les problèmes de routage de se produire dans certaines parties de la charge de travail qui utilisent les différentes applications.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Utilisez unIPAM, tel que le gestionnaire d'adresses VPC IP Amazon, pour surveiller et gérer votre CIDR utilisation. Plusieurs IPAMs sont également disponibles auprès du AWS Marketplace. Évaluez

votre utilisation potentielle AWS, ajoutez des CIDR gammes aux gammes existantes VPCs et créez VPCs pour permettre une croissance planifiée de l'utilisation.

Étapes d’implémentation

- Capturez CIDR la consommation actuelle (par exemple, VPCs et les sous-réseaux).
  - Utilisez les API opérations de service pour collecter CIDR la consommation de courant.
  - Utilisez le gestionnaire d'adresses VPC IP Amazon pour découvrir des ressources.
- Capturez l’utilisation actuelle de votre sous-réseau.
  - Utilisez API les opérations de service pour collecter des sous-réseaux par VPC région.
  - Utilisez le gestionnaire d'adresses VPC IP Amazon pour découvrir des ressources.
- Enregistrez l’utilisation actuelle.
- Déterminez si vous avez créé des plages d’adresses IP se chevauchant.
- Calculez la capacité inutilisée.
- Identifiez les plages d’adresses IP qui se chevauchent. Vous pouvez soit migrer vers une nouvelle plage d'adresses, soit envisager d'utiliser des techniques telles que la NATpasserelle privée ou AWS PrivateLinksi vous devez connecter les plages qui se chevauchent.

Ressources

Bonnes pratiques associées :

- Protection des réseaux

Documents connexes :

- APNPartenaire : partenaires qui peuvent vous aider à planifier votre réseau
- AWS Marketplace pour l'infrastructure réseau
- Livre blanc sur les options de connectivité d’Amazon Virtual Private Cloud
- Connectivité au réseau à haute disponibilité de plusieurs centres de données
- Connexion de réseaux dont les plages d’adresses IP se chevauchent
- Qu'est-ce qu'Amazon VPC ?
- Qu'est-ce que c'est IPAM ?

Vidéos connexes :

- AWS re:Invent 2023 - VPC Designs avancés et nouvelles fonctionnalités
- AWS re:Invent 2019 : des architectures de AWS Transit Gateway référence pour beaucoup VPCs
- AWS re:Invent 2023 - Prêts pour la suite ? Designing networks for growth and flexibility
- AWS re:Invent 2021 - {New Launch} Gérez vos adresses IP à grande échelle sur AWS

Architecture de charge de travail

Questions

- FIA 3. Comment concevez-vous l’architecture de service de votre charge de travail ?
- FIA 4. Comment concevoir des interactions dans un système distribué pour éviter les défaillances ?
- FIA 5. Comment concevoir les interactions dans un système distribué afin d’atténuer les défaillances ou d’y résister ?

FIA 3. Comment concevez-vous l’architecture de service de votre charge de travail ?

Créez des charges de travail hautement évolutives et fiables à l’aide d’une architecture orientée services (SOA) ou d’une architecture de microservices. L’architecture orientée services (SOA) consiste à rendre les composants logiciels réutilisables via les interfaces de service. L’architecture des microservices va plus loin, en particulier en rendant les composants plus petits et plus simples.

Bonnes pratiques

- REL03-BP01 Choisissez comment segmenter votre charge de travail
- REL03-BP02 Créer des services axés sur des domaines commerciaux et des fonctionnalités spécifiques
- REL03-BP03 Fournir des contrats de service par API

REL03-BP01 Choisissez comment segmenter votre charge de travail

La segmentation de la charge de travail est importante lorsqu’il s’agit de déterminer les exigences de résilience de votre application. L’architecture monolithique doit être évitée dans la mesure du possible. À la place, réfléchissez bien aux composants de l’application capables d’être divisés en microservices. Selon les exigences de votre application, il peut s'agir d'une combinaison d'une architecture orientée services (SOA) avec des microservices dans la mesure du possible. Les

charges de travail capables d’absence d’état sont davantage en mesure d’être déployées en tant que microservices.

Résultat désiré : les charges de travail doivent être supportables, évolutives et aussi faiblement couplées que possible.

Lorsque vous choisissez comment segmenter votre charge de travail, comparez les avantages aux complexités. Ce qui convient pour un nouveau produit en course pour un premier lancement est différent de ce dont a besoin une charge de travail conçue pour augmenter d’échelle. Lors de la refactorisation d’une architecture monolithique existante, vous devez évaluer comment l’application prendra en charge une décomposition vers l’absence d’état. La division de services en microservices permet aux petites équipes bien définies de les développer et les gérer. Toutefois, les services plus petits peuvent créer des complexités, dont une latence supérieure, un débogage plus complexe et une charge opérationnelle accrue.

Anti-modèles courants :

- Le microservice Death Star est une situation dans laquelle les composants atomiques deviennent si interdépendants que l’échec de l’un d’entre eux résulte en un échec encore plus important, ce qui rend les composants aussi rigides et fragiles qu’une architecture monolithique.

Avantages liés au respect de cette pratique :

- L’utilisation de segments plus petits permet une plus grande agilité, une plus grande flexibilité organisationnelle et une capacité de mise à l’échelle.
- L’impact réduit des interruptions de service.
- Les composants de l’application peuvent avoir différentes exigences de disponibilité, pouvant être pris en charge par une segmentation plus atomique.
- Des responsabilités bien définies pour les équipes prenant en charge la charge de travail.

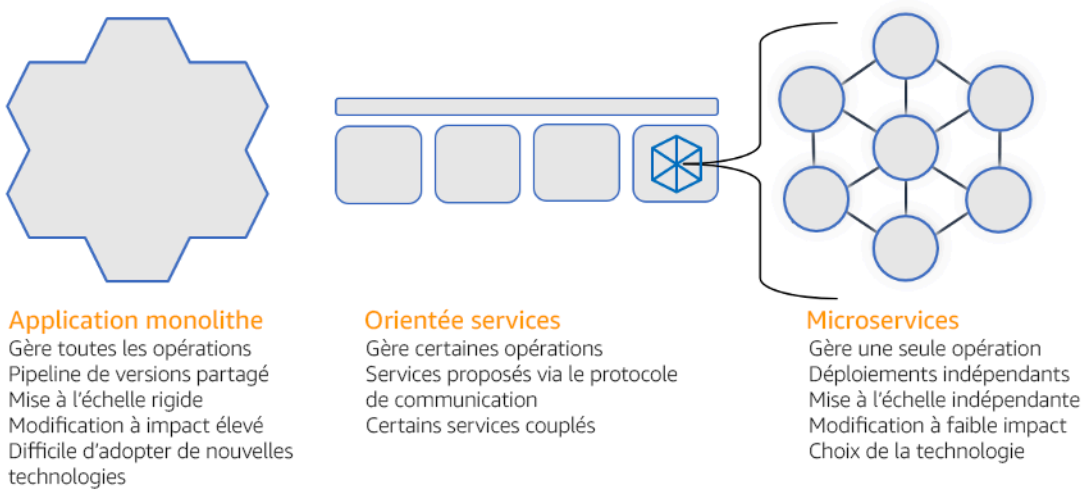
Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Choisissez votre type d’architecture en fonction de la façon dont vous segmenterez votre charge de travail. Choisissez une architecture SOA ou une architecture de microservices (ou dans de rares cas, une architecture monolithique). Même si vous choisissez de commencer par une architecture monolithe, vous devez vous assurer qu'elle est modulaire et qu'elle puisse finalement évoluer

vers des microservices au fur et à SOA mesure que votre produit évolue avec l'adoption par les utilisateurs. SOAet les microservices offrent respectivement une segmentation plus petite, ce qui est préférable en tant qu'architecture moderne, évolutive et fiable, mais certains compromis doivent être pris en compte, en particulier lors du déploiement d'une architecture de microservices.

Le principal compromis est que vous avez maintenant une architecture pour le calcul distribué qui peut compliquer le respect des exigences en matière de latence des utilisateurs et qui complexifie le suivi et le débogage des interactions des utilisateurs. AWS X-Ray peut vous aider à résoudre ce problème. Un autre effet à prendre en compte est la hausse de la complexité opérationnelle à mesure que vous augmentez le nombre d'applications que vous gérez, ce qui nécessite le déploiement de plusieurs composants indépendants.



Architectures monolithique, orientée services et de microservices

Étapes d'implémentation

- Déterminer l'architecture adaptée pour refactoriser ou créer votre application. SOAet les microservices offrent respectivement une segmentation plus petite, ce qui est préférable en tant qu'architecture moderne, évolutive et fiable. SOApeut être un bon compromis pour réduire la segmentation tout en évitant certaines des complexités des microservices. Pour plus de détails, consultez la section Compromis des microservices.
- Si votre charge de travail est appropriée et que votre organisation peut la prendre en charge, vous devez utiliser une architecture de microservices pour obtenir la meilleure agilité et la meilleure fiabilité. Pour plus de détails, voir Implémentation de microservices sur AWS.
- Envisagez de suivre le modèle Strangler Fig pour refactoriser un monolithe en composants plus petits. Cela implique le remplacement progressif de composants spécifiques de l'application par de

nouvelles applications et de nouveaux services. AWS Migration Hub Refactor Spaces sert de point de départ à la refactorisation incrémentielle. Pour plus de détails, consulter Migrer sans interruption vers des charges de travail existantes sur site à l'aide d'un modèle Figuier étrangleur.

- La mise en œuvre de microservices peut nécessiter un mécanisme de découverte de services pour permettre à ces services distribués de communiquer entre eux. AWS App Meshpeut être utilisé avec des architectures orientées services pour permettre une découverte et un accès fiables aux services. AWS Cloud Mappeut également être utilisé pour la découverte de services dynamique DNS basée sur des données.
- Si vous passez d'un monolithe à un bus de service, Amazon SOA MQ peut vous aider à combler cette lacune lors de la refonte d'applications existantes dans le cloud.
- Pour les architectures monolithiques existantes avec une base de données partagée unique, choisissez comment réorganiser les données en segments plus petits. Vous pouvez les réorganiser par unité commerciale, modèle d'accès ou structure de données. À ce stade du processus de refactorisation, vous devez choisir de passer à un type de base de données relationnel ou non relationnel (nonSQL). Pour plus de détails, voir De SQL à Non SQL.

Niveau d'effort du plan d'implémentation : élevé

Ressources

Bonnes pratiques associées :

- REL03-BP02 Créer des services axés sur des domaines commerciaux et des fonctionnalités spécifiques

Documents connexes :

- Amazon API Gateway : Configuration d'une application à REST API l'aide d'Open API
- Qu'est-ce que l'architecture orientée service ?
- Contexte délimité (modèle central dans la conception pilotée par domaine)
- Implémentation de microservices sur AWS
- Compromis des microservices
- Microservices : une définition de ce nouveau terme architectural
- Microservices sur AWS
- Qu'est-ce que c'est AWS App Mesh ?

Exemples connexes :

- Atelier sur la modernisation itérative des applications

Vidéos connexes :

- Garantir l'excellence grâce aux microservices sur AWS

REL03-BP02 Créer des services axés sur des domaines commerciaux et des fonctionnalités spécifiques

Les architectures orientées services (SOA) définissent des services dotés de fonctions bien définies en fonction des besoins de l'entreprise. Les microservices utilisent des modèles de domaine et un contexte limité pour définir les limites des services en fonction des limites du contexte métier. En se concentrant sur les domaines d'activité et les fonctionnalités, les équipes peuvent définir des exigences de fiabilité indépendantes pour leurs services. Les contextes limités isolent et encapsulent la logique métier, ce qui permet aux équipes de mieux raisonner sur la manière de gérer les défaillances.

Résultat désiré : les ingénieurs et les parties prenantes de l'entreprise définissent conjointement des contextes délimités et les utilisent pour concevoir des systèmes en tant que services remplissant des fonctions commerciales spécifiques. Ces équipes utilisent des pratiques établies telles que l'event storming pour définir les exigences. Les nouvelles applications sont conçues comme des services dont les limites sont bien définies et qui possèdent un couplage faible. Les monolithes existants sont décomposés en contextes délimités et les conceptions des systèmes évoluent vers SOA des architectures de microservices. Lorsque les monolithes sont refactorisés, des approches établies telles que les contextes de bulles et les modèles de décomposition des monolithes sont appliquées.

Les services orientés domaine sont exécutés sous la forme d'un ou de plusieurs processus qui ne partagent pas d'état. Ils répondent de manière indépendante aux fluctuations de la demande et gèrent les scénarios de panne à la lumière des exigences spécifiques du domaine.

Anti-modèles courants :

- Les équipes sont constituées autour de domaines techniques spécifiques tels que l'interface utilisateur et l'expérience utilisateur, les intergiciels ou les bases de données plutôt que de domaines commerciaux spécifiques.

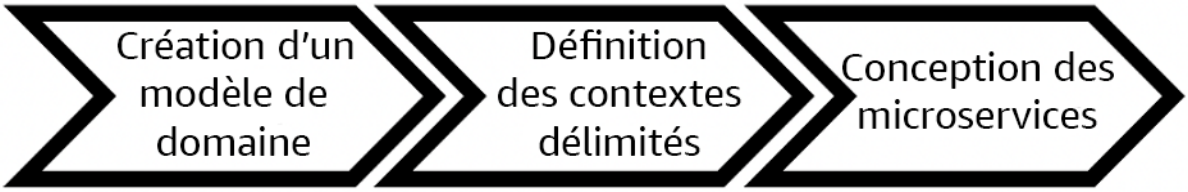
- Les applications couvrent des responsabilités de domaine. Les services qui couvrent des contextes limités peuvent être plus difficiles à gérer, nécessiter des efforts de test plus importants et nécessiter la participation de plusieurs équipes de domaine aux mises à jour logicielles.
- Les dépendances de domaine, telles que les bibliothèques d'entités de domaine, sont partagées entre les services de telle sorte que les modifications apportées à un domaine de service nécessitent des modifications apportées à d'autres domaines de service
- Les contrats de service et la logique métier n'expriment pas les entités dans un langage de domaine commun et cohérent, ce qui crée des couches de traduction qui compliquent les systèmes et augmentent les efforts de débogage.

Avantages du respect de cette bonne pratique : les applications sont conçues comme des services indépendants délimités par domaines d'activité et utilisent un langage métier commun. Les services peuvent être testés et déployés indépendamment. Les services répondent aux exigences de résilience spécifiques au domaine mis en œuvre.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

La conception axée sur le domaine (DDD) est l'approche fondamentale de la conception et de la création de logiciels autour de domaines commerciaux. Il est utile de travailler avec un cadre existant lorsque vous créez des services axés sur des domaines métier. Lorsque vous travaillez avec des applications monolithiques existantes, vous pouvez tirer parti des modèles de décomposition qui fournissent des techniques éprouvées pour moderniser les applications en services.



Conception pilotée par domaine

Étapes d'implémentation

- Les équipes peuvent organiser des ateliers event storming pour identifier rapidement les événements, les commandes, les agrégats et les domaines dans un format léger de notes autocollantes.

- Une fois que les entités et les fonctions de domaine ont été formées dans un contexte de domaine, vous pouvez diviser votre domaine en services à l'aide d'un contexte délimité, dans lequel les entités partageant des caractéristiques et des attributs similaires sont regroupées. La division en contextes permet de faire émerger un modèle de délimitation des microservices.
  - Par exemple, les entités du site Amazon.com peuvent inclure le colis, la livraison, le calendrier, le prix, la remise et la devise.
  - Le colis, la livraison et le calendrier sont regroupés dans le contexte d'expédition, tandis que le prix, la remise et la devise sont regroupés dans le contexte de tarification.
- Décomposition des monolithes en microservices décrit les modèles de refactorisation des microservices. L'utilisation de modèles de décomposition par capacité métier, sous-domaine ou transaction s'inscrit parfaitement dans les approches axées sur le domaine.
- Les techniques tactiques telles que le contexte à bulles vous permettent de les introduire DDD dans des applications existantes ou héritées sans avoir à procéder à des réécritures initiales et à des engagements complets. DDD Dans une approche basée sur un contexte à bulles, un petit contexte délimité est établi à l'aide d'une cartographie et d'une coordination des services, ou couche anticorruption, qui protège le modèle de domaine nouvellement défini des influences extérieures.

Une fois que les équipes ont effectué une analyse de domaine et défini des entités et des contrats de service, elles peuvent tirer parti des AWS services pour mettre en œuvre leur conception axée sur le domaine sous forme de services basés sur le cloud.

- Commencez votre développement en définissant des tests qui appliquent les règles métier de votre domaine. Le développement piloté par les tests (TDD) et le développement piloté par le comportement (BDD) aident les équipes à concentrer leurs services sur la résolution des problèmes commerciaux.
- Sélectionnez les services AWS qui répondent le mieux aux exigences de votre domaine d'activité et à votre architecture de microservices :
  - AWS Le mode Serverless permet à votre équipe de se concentrer sur une logique de domaine spécifique au lieu de se concentrer sur la gestion des serveurs et de l'infrastructure.
  - Les conteneurs AWS simplifient la gestion de votre infrastructure afin que vous puissiez vous concentrer sur les exigences de votre domaine.
  - Les bases de données sur mesure vous permettent d'adapter les exigences de votre domaine au type de base de données le mieux adapté.

- Création d’architectures hexagonales sur AWS décrit un cadre permettant d’intégrer une logique métier à des services en procédant de manière rétroactive à partir d’un domaine métier afin de répondre à des exigences fonctionnelles, puis d’associer des adaptateurs d’intégration. Les modèles qui séparent les détails de l'interface de la logique métier grâce AWS aux services aident les équipes à se concentrer sur les fonctionnalités du domaine et à améliorer la qualité des logiciels.

Ressources

Bonnes pratiques associées :

- REL03-BP01 Choisissez comment segmenter votre charge de travail
- REL03-BP03 Fournir des contrats de service par API

Documents connexes :

- AWS Microservices
- Implémentation de microservices sur AWS
- Comment convertir un monolithe en microservices
- Commencer à fonctionner DDD lorsqu'on est entouré de systèmes existants
- Conception axée sur le domaine : aborder la complexité au cœur du logiciel
- Construire des architectures hexagonales sur AWS
- Décomposition des monolithes en microservices
- Event Storming
- Messages entre contextes limités
- Microservices
- Développement piloté par les tests
- Développement axé sur le comportement

Exemples connexes :

- Conception de microservices cloud natifs sur AWS (à partir deDDD/EventStormingWorkshop)

Outils associés :

- AWS Cloud bases de données
- Sans serveur activé AWS
- Conteneurs chez AWS

REL03-BP03 Fournir des contrats de service par API

Les contrats de service sont des accords documentés entre API producteurs et consommateurs définis dans une définition lisible par API machine. Une stratégie de gestion des versions des contrats permet aux consommateurs de continuer à utiliser l'existant API et de migrer leurs applications vers une version plus récente API lorsqu'elles sont prêtes. Le déploiement du producteur peut avoir lieu à tout moment tant que le contrat est respecté. Les équipes de service peuvent utiliser la pile technologique de leur choix pour exécuter le API contrat.

Résultat escompté : les applications conçues avec des architectures orientées services ou microservices sont capables de fonctionner de manière indépendante tout en intégrant une dépendance à l'environnement d'exécution. Les modifications apportées à un API consommateur ou à un producteur n'interrompent pas la stabilité de l'ensemble du système lorsque les deux parties suivent un API contrat commun. Les composants qui communiquent via le service APIs peuvent exécuter des versions fonctionnelles indépendantes, mettre à niveau les dépendances d'exécution ou basculer vers un site de reprise après sinistre (DR) avec peu ou pas d'impact les uns sur les autres. En outre, les services discrets sont capables d’évoluer de manière indépendante en absorbant la demande de ressources sans que les autres services soient réduits horizontalement à l’unisson.

Anti-modèles courants :

- Création d'un service APIs sans schémas fortement typés. Il en résulte APIs que cela ne peut pas être utilisé pour générer des API liaisons et des charges utiles qui ne peuvent pas être validées par programmation.
- Ne pas adopter de stratégie de gestion des versions, qui oblige API les consommateurs à effectuer des mises à jour et à publier ou à échouer lorsque les contrats de service évoluent.
- Messages d’erreur qui divulguent les détails de l’implémentation du service sous-jacent au lieu de décrire les échecs d’intégration dans le contexte et le langage du domaine.
- Ne pas utiliser de API contrats pour développer des scénarios de test et API des mises en œuvre fictives afin de permettre des tests indépendants des composants du service.

Avantages de l'établissement de cette meilleure pratique : les systèmes distribués composés de composants communiquant par le biais de contrats de API service peuvent améliorer la fiabilité.

Les développeurs peuvent détecter les problèmes potentiels dès le début du processus de développement en vérifiant le type lors de la compilation afin de vérifier que les demandes et les réponses respectent le API contrat et que les champs obligatoires sont présents. APIles contrats fournissent une interface autodocumentée claire APIs et fournissent une meilleure interopérabilité entre les différents systèmes et langages de programmation.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Une fois que vous avez identifié les domaines commerciaux et déterminé la segmentation de votre charge de travail, vous pouvez développer votre serviceAPIs. Définissez d'abord des contrats de service lisibles par machine pourAPIs, puis implémentez une stratégie de gestion des API versions. Lorsque vous êtes prêt à intégrer des services via des protocoles courants tels que REST GraphQL ou des événements asynchrones, vous pouvez intégrer des AWS services dans votre architecture afin d'intégrer vos composants à l'aide de contrats bien typés. API

AWS services pour les API contrats de service

Intégrez AWS des services tels qu'Amazon API Gateway et Amazon EventBridge dans votre architecture pour utiliser des contrats de API service dans votre application. AWS AppSync Amazon API Gateway vous permet d'intégrer directement des AWS services natifs et d'autres services Web. APIGateway prend en charge la APIspécification Open et le versionnement. AWS AppSync est un point de terminaison GraphQL géré que vous configurez en définissant un schéma GraphQL pour définir une interface de service pour les requêtes, les mutations et les abonnements. Amazon EventBridge utilise des schémas d'événements pour définir des événements et générer des liaisons de code pour vos événements.

Étapes d'implémentation

- Tout d'abord, définissez un contrat pour votreAPI. Un contrat exprimera les capacités d'un API et définira des objets de données et des champs fortement typés pour l'APIentrée et la sortie.
- Lorsque vous configurez APIs dans API Gateway, vous pouvez importer et exporter des API spécifications ouvertes pour vos points de terminaison.
  - L'importation d'une API définition ouverte simplifie la création de votre API et peut être intégrée à AWS l'infrastructure sous forme d'outils de code tels que le AWS Serverless Application Modelet AWS Cloud Development Kit (AWS CDK).
  - L'exportation d'une API définition simplifie l'intégration avec API les outils de test et fournit aux consommateurs de services une spécification d'intégration.

- Vous pouvez définir et gérer GraphQL AWS AppSync en définissant APIs un fichier de schéma GraphQL pour générer votre interface de contrat et simplifier l'interaction avec des REST modèles complexes, plusieurs tables de base de données ou des services existants.
- AWS Amplifyles projets intégrés AWS AppSync génèrent des fichiers de JavaScript requêtes fortement typés à utiliser dans votre application, ainsi qu'une bibliothèque cliente AWS AppSync GraphQL pour les tables Amazon DynamoDB.
- Lorsque vous consommez des événements de service d'Amazon EventBridge, les événements adhèrent aux schémas qui existent déjà dans le registre des schémas ou que vous définissez avec l'Open API Spec. Avec un schéma défini dans le registre, vous pouvez également générer des liaisons client à partir du contrat de schéma afin d'intégrer votre code aux événements.
- Extension ou version de votreAPI. L'extension d'une API est une option plus simple lorsque vous ajoutez des champs qui peuvent être configurés avec des champs facultatifs ou des valeurs par défaut pour les champs obligatoires.
  - JSONles contrats basés sur des protocoles tels que REST GraphQL peuvent être une bonne solution pour l'extension de contrat.
  - XMLdes contrats basés sur des protocoles tels que SOAP ceux qui devraient être testés auprès des consommateurs de services afin de déterminer la faisabilité d'une prolongation du contrat.
- Lors du versionnement d'unAPI, envisagez d'implémenter le versionnage par proxy dans le cadre duquel une façade est utilisée pour prendre en charge les versions afin que la logique puisse être maintenue dans une base de code unique.
  - Avec API Gateway, vous pouvez utiliser les mappages de demandes et de réponses pour simplifier l'absorption des modifications de contrat en établissant une façade pour fournir des valeurs par défaut pour les nouveaux champs ou pour supprimer les champs supprimés d'une demande ou d'une réponse. Avec cette approche, le service sous-jacent peut gérer une base de code unique.

Ressources

Bonnes pratiques associées :

- REL03-BP01 Choisissez comment segmenter votre charge de travail
- REL03-BP02 Créer des services axés sur des domaines commerciaux et des fonctionnalités spécifiques
- REL04-BP02 Implémenter des dépendances faiblement couplées
- REL05-BP03 Contrôler et limiter les appels de nouvelle tentative

- REL05-BP05 Définir les délais d'expiration des clients

Documents connexes :

- Qu'est-ce qu'une API (interface de programmation d'applications) ?
- Implémentation de microservices sur AWS
- Compromis des microservices
- Microservices : une définition de ce nouveau terme architectural
- Microservices sur AWS
- Utilisation des extensions de API passerelle pour ouvrir API
- Spécification ouverte API
- GraphQL : schémas et types
- liaisons EventBridge de code Amazon

Exemples connexes :

- Amazon API Gateway : Configuration d'une application à REST API l'aide d'Open API
- Amazon API Gateway vers l'application Amazon CRUD DynamoDB à l'aide d'Open API
- Modèles modernes d'intégration des applications à l'ère du sans serveur : intégration des services de API passerelle
- Implémentation du versionnement des API passerelles basé sur les en-têtes avec Amazon CloudFront
- AWS AppSync : création d’une application client

Vidéos connexes :

- Utilisation d'Open API in AWS SAM pour gérer API Gateway

Outils associés :

- APIPasserelle Amazon
- AWS AppSync
- Amazon EventBridge

FIA 4. Comment concevoir des interactions dans un système distribué pour éviter les défaillances ?

Les systèmes distribués s'appuient sur des réseaux de communication pour interconnecter les composants, comme les serveurs ou les services. Votre charge de travail doit fonctionner de manière fiable malgré la perte de données ou la latence sur ces réseaux. Les composants du système distribué doivent fonctionner de manière à ne pas avoir d'impact négatif sur les autres composants ou sur la charge de travail. Ces bonnes pratiques permettent d'éviter les défaillances et d'améliorer le temps moyen entre défaillances (MTBF).

Bonnes pratiques

- REL04-BP01 Identifier le type de systèmes distribués dont vous dépendez
- REL04-BP02 Implémenter des dépendances faiblement couplées
- REL04-BP03 Faire un travail constant
- REL04-BP04 Rendre les opérations de mutation idempotentes

REL04-BP01 Identifier le type de systèmes distribués dont vous dépendez

Les systèmes distribués peuvent être synchrones, asynchrones ou par lots. Les systèmes synchrones doivent traiter les demandes le plus rapidement possible et communiquer entre eux en effectuant des appels de demande et de réponse synchrones à l'aide des protocoles HTTP/S, REST ou RPC (Remote Procedure Call). Les systèmes asynchrones communiquent entre eux en échangeant des données de manière asynchrone via un service intermédiaire sans coupler des systèmes individuels. Les systèmes par lots reçoivent un volume important de données d'entrée, exécutent des processus de données automatisés sans intervention humaine et génèrent des données de sortie.

Résultat souhaité : concevez une charge de travail qui interagit efficacement avec les dépendances synchrones, asynchrones et par lots.

Anti-modèles courants :

- La charge de travail attend indéfiniment une réponse de la part de ses dépendances, ce qui peut entraîner une expiration des clients de la charge de travail, qui ne savent pas si leur demande a été reçue.
- La charge de travail utilise une chaîne de systèmes dépendants qui s'appellent les uns les autres de manière synchrone. Cela nécessite que chaque système soit disponible et traite correctement

une demande avant que l'ensemble de la chaîne puisse aboutir, ce qui entraîne un comportement et une disponibilité globale potentiellement fragiles.

- La charge de travail communique avec ses dépendances de manière asynchrone et repose sur le concept de livraison garantie unique des messages, alors qu'il est souvent encore possible de recevoir des messages dupliqués.
- La charge de travail n'utilise pas les outils de planification par lots appropriés et permet l'exécution simultanée de la même tâche de traitement par lots.

Avantages du respect de cette bonne pratique : il est courant qu'une charge de travail donnée implémente un ou plusieurs styles de communication entre synchrone, asynchrone et par lots. Cette bonne pratique vous aide à identifier les différents compromis associés à chaque style de communication afin que votre charge de travail soit capable de tolérer les interruptions liées à toutes ses dépendances.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les sections suivantes contiennent des instructions de mise en œuvre générales et spécifiques pour chaque type de dépendance.

General guidance

- Assurez-vous que les objectifs de niveau de service (SLO) offerts par vos dépendances en matière de performance et de fiabilité répondent aux exigences de performance et de fiabilité de votre charge de travail.
- Utilisez les services d'observabilité AWS pour surveiller les temps de réponse et les taux d'erreur afin de vous assurer que votre dépendance fournit des services aux niveaux requis par votre charge de travail.
- Identifiez les défis potentiels auxquels votre charge de travail peut être confrontée lors de la communication avec ses dépendances. Les systèmes distribués présentent un large éventail de défis susceptibles d'accroître la complexité architecturale, la charge opérationnelle et les coûts. Les défis courants incluent la latence, les perturbations du réseau, la perte de données, la mise à l'échelle et le retard de réplication des données.
- Mettez en œuvre une gestion des erreurs et une journalisation robustes pour vous aider à résoudre les problèmes lorsque votre dépendance rencontre des problèmes.

Dépendance synchrone

Dans les communications synchrones, votre charge de travail envoie une demande à sa dépendance et bloque l'opération en attente de réponse. Lorsque sa dépendance reçoit la demande, elle essaie de la traiter le plus rapidement possible et renvoie une réponse à la charge de travail. L'un des principaux défis liés à la communication synchrone est qu'elle entraîne un couplage temporel, ce qui nécessite que la charge de travail et ses dépendances soient disponibles en même temps. Lorsque la charge de travail doit communiquer de manière synchrone avec ses dépendances, suivez les conseils ci-dessous :

- Votre charge de travail ne doit pas reposer sur plusieurs dépendances synchrones pour exécuter une seule fonction. Cette chaîne de dépendances augmente la fragilité globale, car toutes les dépendances sur le chemin doivent être disponibles pour que la demande soit traitée correctement.
- Lorsqu'une dépendance n'est pas saine ou n'est pas disponible, déterminez vos stratégies de gestion des erreurs et de nouvelles tentatives. Évitez d'utiliser un comportement bimodal. On parle de comportement bimodal lorsque la charge de travail présente un comportement différent en mode normal et en mode d'échec. Pour plus de détails sur le comportement bimodal, voir REL11-BP05 Utiliser la stabilité statique pour empêcher le comportement bimodal.
- N'oubliez pas qu'il vaut mieux échouer rapidement que faire attendre la charge de travail. Par exemple, le Guide du développeur AWS Lambda explique comment gérer les tentatives et les échecs lorsque vous invoquez des fonctions Lambda.
- Définissez des délais d'expiration lorsque la charge de travail appelle sa dépendance. Cette technique permet d'éviter d'attendre trop longtemps ou d'attendre indéfiniment une réponse. Vous trouverez une discussion utile sur ce sujet dans la section Réglage des paramètres de demande HTTP du SDK Java AWS pour les applications Amazon DynamoDB sensibles à la latence.
- Réduisez le nombre d'appels passés entre la charge de travail et sa dépendance pour répondre à une seule demande. Le fait d'avoir trop d'appels entre elles augmente le couplage et la latence.

Dépendance asynchrone

Pour pouvoir découpler temporellement la charge de travail de sa dépendance, elles doivent communiquer de manière asynchrone. En utilisant une approche asynchrone, la charge de travail peut poursuivre tout autre traitement sans avoir à attendre que sa dépendance, ou sa chaîne de dépendances, envoie une réponse.

Lorsque la charge de travail doit communiquer de manière asynchrone avec sa dépendance, suivez les conseils ci-dessous :

- Déterminez s'il convient d'utiliser la messagerie ou le streaming d'événements en fonction de votre cas d'utilisation et de vos exigences. La messagerie permet à votre charge de travail de communiquer avec ses dépendances en envoyant et en recevant des messages par le biais d'un agent de messages. Le streaming d'événements permet à votre charge de travail et à ses dépendances d'utiliser un service de streaming pour publier et s'abonner à des événements, diffusés sous forme de flux de données continus, qui doivent être traités dès que possible.

- La messagerie et le streaming d'événements traitent les messages différemment. Vous devez donc faire un choix en fonction des éléments suivants :
  - Priorité des messages : les agents de messages peuvent traiter les messages prioritaires avant les messages normaux. Dans le cadre du streaming d'événements, tous les messages ont la même priorité.
  - Consommation de messages : les agents de messages veillent à ce que les consommateurs reçoivent le message. Les consommateurs qui diffusent des événements doivent suivre le dernier message qu'ils ont lu.
  - Ordre des messages : avec la messagerie, la réception des messages dans l'ordre exact dans lequel ils sont envoyés n'est pas garantie, sauf si vous utilisez une approche « premier entré, premier sorti » (FIFO). Le streaming d'événements préserve toujours l'ordre dans lequel les données ont été produites.
  - Suppression du message : dans le cas de la messagerie, le consommateur doit supprimer le message après l'avoir traité. Le service de streaming d'événements ajoute le message à un flux et y reste jusqu'à l'expiration de la période de conservation du message. Cette politique de suppression rend le streaming d'événements adapté à la rediffusion de messages.
- Définissez comment la charge de travail comprend que sa dépendance a mené à bien sa tâche. Par exemple, lorsque votre charge de travail invoque une fonction Lambda de manière asynchrone, Lambda place l'événement dans une file d'attente, et renvoie une réponse de succès sans plus d'informations. Une fois le traitement terminé, la fonction Lambda peut envoyer le résultat à une destination configurable en fonction du succès ou de l'échec.
- Augmentez votre charge de travail pour gérer les messages dupliqués en tirant parti de l'idempotence. L'idempotence signifie que les résultats de la charge de travail ne changent pas même si elle est générée plusieurs fois pour le même message. Il est important de souligner que les services de messagerie ou de streaming redistribueront un message en cas de panne du réseau ou en l'absence de réception d'un accusé de réception.
- Si la charge de travail n'obtient pas de réponse de sa dépendance, elle doit soumettre à nouveau la demande. Envisagez de limiter le nombre de tentatives pour préserver le processeur, la

mémoire et les ressources réseau de votre charge de travail afin de gérer d'autres demandes. La documentation AWS Lambda montre comment gérer les erreurs lors d'une invocation asynchrone.

- Tirez parti des outils d'observabilité, de débogage et de suivi appropriés pour gérer et exploiter la communication asynchrone de la charge de travail avec ses dépendances. Vous pouvez utiliser Amazon CloudWatch pour surveiller les services de messagerie et de streaming d'événements. Vous pouvez également utiliser votre charge de travail avec AWS X-Ray pour obtenir rapidement des informations permettant de résoudre les problèmes.

Dépendance par lots

Les systèmes par lots prennent les données d'entrée, lancent une série de tâches pour les traiter et produisent certaines données de sortie, sans intervention manuelle. En fonction de la taille des données, les tâches peuvent s'exécuter pendant une durée allant de quelques minutes à, dans certains cas, plusieurs jours. Lorsque la charge de travail communique avec sa dépendance par lots, suivez les conseils ci-dessous :

- Définissez la fenêtre de temps pendant laquelle la charge de travail doit exécuter le traitement par lots. La charge de travail peut configurer un modèle de récurrence pour invoquer un système de traitement par lots (par exemple, toutes les heures ou à la fin de chaque mois).
- Déterminez l'emplacement de l'entrée des données et de la sortie des données traitées. Choisissez un service de stockage, tel qu'Amazon Simple Storage Services (Amazon S3), Amazon Elastic File System (Amazon EFS) et Amazon FSx pour Lustre, qui permet à votre charge de travail de lire et d'écrire des fichiers à l'échelle.
- Si votre charge de travail doit invoquer plusieurs tâches par lots, vous pouvez en tirer parti de AWS Step Functions pour simplifier l'orchestration des tâches par lots exécutées dans AWS ou sur site. Cet exemple de projet illustre l'orchestration de traitements par lots à l'aide de Step Functions, AWS Batch et Lambda.
- Surveillez les tâches par lots pour détecter d'éventuelles anomalies, telles qu'une tâche qui prend plus de temps que prévu. Vous pouvez utiliser des outils tels que CloudWatch Container Insights pour surveiller les environnements AWS Batch et les tâches. Dans ce cas, la charge de travail empêcherait le début de la tâche suivante et informerait le personnel concerné de l'exception.

Ressources

Documents connexes :

- Opérations : Surveillance et observabilité AWS Cloud

- L’Amazon Builders’ Library : défis liés aux systèmes distribués
- REL11-BP05 Utiliser la stabilité statique pour éviter les comportements bimodaux
- Guide du développeur AWS Lambda : Gestion des erreurs et tentatives automatiques dans AWS Lambda
- Réglage des paramètres de demande HTTP d’AWS SDK Java pour les applications Amazon DynamoDB sensibles à la latence
- Messagerie AWS
- Qu’est-ce que le streaming de données ?
- Guide du développeur AWS Lambda : Invocation asynchrone
- FAQ Amazon Simple Queue Service : Files d’attente FIFO
- Guide du développeur Amazon Kinesis Data Streams : Gestion des enregistrements en double
- Guide du développeur Amazon Simple Queue Service : Métriques CloudWatch disponibles pour Amazon SQS
- Guide du développeur Amazon Kinesis Data Streams : Surveillance du service Amazon Kinesis Data Streams avec Amazon CloudWatch
- Guide du développeur AWS X-Ray : Concepts AWS X-Ray
- Exemples AWS sur GitHub : Application AWS Step Functions Complex Orchestrator
- Guide de l’utilisateur AWS Batch : AWS Batch CloudWatch Container Insights

Vidéos connexes :

- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS (COP310)

Outils associés :

- Amazon CloudWatch
- Amazon CloudWatch Logs
- AWS X-Ray
- Amazon Simple Storage Service (Amazon S3)
- Amazon Elastic File System (Amazon EFS)
- Amazon FSx pour Lustre

- AWS Step Functions
- AWS Batch

REL04-BP02 Implémenter des dépendances faiblement couplées

Des dépendances telles que des systèmes de file d'attente, des systèmes de streaming, des flux de travail et des équilibreurs de charge sont couplées faiblement. Le couplage faible permet d'isoler le comportement d'un composant des autres composants qui en dépendent, ce qui augmente la résilience et l'agilité.

Le découplage des dépendances, telles que les systèmes de file d'attente, les systèmes de streaming et les flux de travail, permet de minimiser l'impact des modifications ou des défaillances sur un système. Cette séparation empêche le comportement d'un composant d'affecter les autres qui en dépendent, améliorant ainsi la résilience et l'agilité.

Dans les systèmes couplés fortement, la modification d'un composant peut nécessiter de modifier d'autres composants qui en dépendent, ce qui entraîne une dégradation des performances de tous les composants. Le couplage faible rompt cette dépendance de sorte que les composants dépendants n'ont besoin que de connaître l'interface publiée et sa version. La mise en œuvre d'un couplage faible entre les dépendances permet d'isoler une défaillance dans l'une afin de ne pas en impacter une autre.

Le couplage faible vous permet de modifier le code ou d'ajouter des fonctionnalités à un composant tout en minimisant les risques pour les autres composants qui en dépendent. Il offre également une résilience granulaire au niveau des composants, ce qui vous permet d'augmenter horizontalement voire de modifier la mise en œuvre sous-jacente de la dépendance.

Pour améliorer encore la résilience par un couplage faible, dans la mesure du possible, rendez asynchrones les interactions des composants. Ce modèle convient à toute interaction qui ne nécessite pas une réponse immédiate et pour laquelle une confirmation de l'enregistrement d'une requête suffira. Il implique un composant qui génère des événements et un autre qui les consomme. Les deux composants ne s'intègrent pas par point-to-point interaction directe, mais généralement par le biais d'une couche de stockage durable intermédiaire, telle qu'une SQS file d'attente Amazon, une plateforme de données de streaming telle qu'Amazon Kinesis ou. AWS Step Functions

Figure 4 : Les dépendances telles que des systèmes de file d'attente et des équilibreurs de charge sont couplées faiblement

Amazon fait la SQS queue et ce ne AWS Step Functions sont que deux moyens d'ajouter une couche intermédiaire pour un couplage souple. Des architectures axées sur les événements peuvent également être créées à l'aide d' AWS Cloud Amazon EventBridge, qui peut isoler les clients (producteurs d'événements) des services sur lesquels ils comptent (consommateurs d'événements). Amazon Simple Notification Service (AmazonSNS) est une solution efficace lorsque vous avez besoin d'une messagerie push à haut débit. many-to-many À l'aide d'Amazon SNS Topics, les systèmes de vos éditeurs peuvent diffuser les messages vers un grand nombre de points de terminaison d'abonnés pour un traitement parallèle.

Bien que les files d’attente offrent plusieurs avantages, dans la plupart des systèmes en temps réel stricts, les requêtes antérieures à un seuil (souvent en secondes) sont considérées comme obsolètes (le client a abandonné et n’attend plus de réponse). En conséquence, elles ne sont pas traitées. De cette façon, les requêtes plus récentes (et probablement toujours valides) peuvent être traitées à la place.

Résultat souhaité : la mise en œuvre de dépendances faiblement couplées vous permet de minimiser la surface de défaillance au niveau du composant, ce qui permet de diagnostiquer et de résoudre les problèmes. Elle simplifie également les cycles de développement en permettant aux équipes de mettre en œuvre des modifications à un niveau modulaire sans affecter les performances des autres composants qui en dépendent. Avec cette approche, il est possible d’augmenter horizontalement un composant en fonction des besoins en ressources et de l’utilisation de ce composant, ce qui contribue à améliorer la rentabilité.

Anti-modèles courants :

- Déploiement d’une charge de travail monolithique.
- Invocation directe APIs entre les niveaux de charge de travail sans possibilité de basculement ou de traitement asynchrone de la demande.
- Couplage fort à l’aide de données partagées. Les systèmes couplés faiblement évitent de partager des données par le biais de bases de données partagées ou d’autres formes de stockage de données couplées fortement, qui peuvent réintroduire un couplage fort et entraver la capacité de mise à l’échelle.
- Ignorer la contre-pression. Votre charge de travail doit être capable de ralentir ou d’arrêter les données entrantes lorsqu’un composant ne peut pas les traiter au même rythme.

Avantages du respect de cette bonne pratique : le couplage faible permet d'isoler le comportement d'un composant des autres composants qui en dépendent, ce qui augmente la résilience et l'agilité. La défaillance d'un composant est isolée des autres.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Implémentez des dépendances couplées faiblement. Différentes solutions permettent de créer des applications couplées faiblement. Il s'agit notamment de services permettant de mettre en œuvre des files d'attente entièrement gérées, des flux de travail automatisés, de réaction aux événements, etc., qui peuvent aider à isoler le comportement des composants par rapport aux autres composants, augmentant ainsi la résilience et l'agilité. APIs

- Créez des architectures pilotées par les événements : Amazon vous EventBridge aide à créer des architectures pilotées par les événements faiblement couplées et distribuées.
- Implémenter des files d'attente dans les systèmes distribués : vous pouvez utiliser Amazon Simple Queue Service (AmazonSQS) pour intégrer et découpler les systèmes distribués.
- Conteneuriser les composants sous forme de microservices : les microservices permettent aux équipes de créer des applications composées de petits composants indépendants qui communiquent via des canaux bien définis. APIs Amazon Elastic Container Service (AmazonECS) et Amazon Elastic Kubernetes Service (EKSAmazon) peuvent vous aider à démarrer plus rapidement avec les conteneurs.
- Gérez les flux de travail avec Step Functions : Step Functions vous aide à coordonner plusieurs AWS services dans des flux de travail flexibles.
- Tirez parti des architectures de messagerie publish-subscribe (pub/sub) : Amazon Simple Notification Service (AmazonSNS) assure la transmission des messages des éditeurs aux abonnés (également appelés producteurs et consommateurs).

Étapes d'implémentation

- Les composants d'une architecture basée sur les événements sont initiés par des événements. Les événements sont des actions qui se produisent dans un système (par exemple, un utilisateur ajoute un article à un panier). Lorsque l'action aboutit, un événement est généré et active le composant suivant du système.
  - Création d'applications basées sur les événements avec Amazon EventBridge

    - AWS re:Invent 2022 - Conception d'intégrations pilotées par des événements à l'aide d'Amazon EventBridge
- Les systèmes de messagerie distribuée comportent trois parties principales qui doivent être mises en œuvre pour une architecture basée sur des files d'attente. Ils incluent les composants du système distribué, la file d'attente utilisée pour le découplage (distribuée sur les SQS serveurs Amazon) et les messages de la file d'attente. Dans un système classique, les producteurs envoient le message dans la file d'attente et le consommateur reçoit le message de la file d'attente. La file d'attente stocke les messages sur plusieurs SQS serveurs Amazon à des fins de redondance.
    - SQSArchitecture Amazon de base
    - Envoyer des messages entre des applications distribuées avec Amazon Simple Queue Service
- Lorsqu'ils sont bien utilisés, les microservices améliorent la maintenabilité et la capacité de mise à l'échelle, car les composants couplés faiblement sont gérés par des équipes indépendantes. Ils permettent également d'isoler les comportements d'un composant en cas de changement.
    - Implémentation de microservices sur AWS
    - Let's Architect! Architecting microservices with containers
- AWS Step Functions Vous pouvez notamment créer des applications distribuées, automatiser des processus, orchestrer des microservices. L'orchestration de plusieurs composants dans un flux de travail automatisé vous permet de découpler des dépendances dans votre application.
    - Créez un flux de travail sans serveur avec et AWS Step FunctionsAWS Lambda
    - Commencer avec AWS Step Functions

Ressources

Documents connexes :

- Amazon EC2 : garantir l'impuissance
- L'Amazon Builders' Library : défis liés aux systèmes distribués
- L'Amazon Builders' Library : fiabilité, travail constant et une bonne tasse de café
- Qu'est-ce qu'Amazon EventBridge ?
- Qu'est-ce qu'Amazon Simple Queue Service ?
- Rompre avec votre monolithe
- Orchestrez des microservices basés sur des files d'attente avec Amazon et Amazon AWS Step Functions SQS

- SQSArchitecture Amazon de base
- Architecture basée sur des files d’attente

Vidéos connexes :

- AWS Sommet de New York 2019 : introduction aux architectures événementielles et à Amazon EventBridge (05) MAD2
- AWS re:Invent 2018 : Boucles serrées et ouverture d'esprit : comment prendre le contrôle des systèmes, grands et petits ARC337 (y compris le couplage lâche, le travail constant, la stabilité statique)
- AWS re:Invent 2019 : Passage à des architectures pilotées par les événements (08) SVS3
- AWS re:Invent 2019 : applications évolutives pilotées par des événements sans serveur utilisant Amazon et Lambda SQS
- AWS re:Invent 2022 - Conception d'intégrations pilotées par des événements à l'aide d'Amazon EventBridge
- AWS re:Invent 2017 : Elastic Load Balancing : analyse approfondie et meilleures pratiques

REL04-BP03 Faire un travail constant

Les systèmes peuvent échouer en cas de modifications importantes et rapides de la charge. Par exemple, si votre charge de travail effectue une surveillance de l’état de milliers de serveurs, elle doit envoyer chaque fois une charge utile de la même taille (un instantané complet de l’état actuel). Qu’aucun des serveurs ne présente de problème ou qu’ils en connaissent tous, le système de surveillance de l’état effectue un travail constant sans modifications importantes ni rapides.

Par exemple, si le système de surveillance de l’état surveille 100 000 serveurs, la charge sur celui-ci est nominale avec le taux de défaillance normalement faible du serveur. En revanche, si un événement majeur rendait la moitié de ces serveurs défectueux, le système de surveillance de l’état serait submergé en tentant de mettre à jour les systèmes de notification et de communiquer l’état à ses clients. Le système de surveillance de l’état devrait donc envoyer un instantané complet de l’état actuel à chaque fois. 100 000 états de santé du serveur, chacun représenté par un octet, ne représenteraient qu’une charge utile de 12,5 Ko. Qu’aucun des serveurs ne présente de problème ou qu’ils en connaissent tous, le système de surveillance de l’état effectue un travail constant, et les modifications importantes et rapides ne menacent pas la stabilité du système. C’est ainsi qu’Amazon Route 53 gère les surveillances de l’état des points de terminaison (tels que les adresses IP) pour déterminer comment les utilisateurs finaux sont acheminés vers eux.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d’implémentation

- Effectuez un travail constant : les systèmes peuvent échouer lorsque la charge connaît des changements rapides et importants.
- Implémentez des dépendances couplées faiblement. Des dépendances telles que des systèmes de file d’attente, des systèmes de streaming, des flux de travail et des équilibreurs de charge sont couplées faiblement. Le couplage faible permet d’isoler le comportement d’un composant des autres composants qui en dépendent, ce qui augmente la résilience et l’agilité.
  - L’Amazon Builders’ Library : fiabilité, travail constant et une bonne tasse de café
  - AWS re:Invent 2018 : Boucles serrées et ouverture d'esprit : comment prendre le contrôle des systèmes, grands et petits ARC337 (y compris un travail constant)
    - Pour l’exemple d’un système de surveillance de l’état surveillant 100 000 serveurs, concevez les charges de travail de manière à ce que les tailles de charge utile restent constantes, quel que soit le nombre de réussites ou d’échecs.

Ressources

Documents connexes :

- Amazon EC2 : garantir l'impuissance
- L’Amazon Builders’ Library : défis liés aux systèmes distribués
- L’Amazon Builders’ Library : fiabilité, travail constant et une bonne tasse de café

Vidéos connexes :

- AWS Sommet de New York 2019 : introduction aux architectures événementielles et à Amazon EventBridge (05) MAD2
- AWS re:Invent 2018 : Boucles serrées et ouverture d'esprit : comment prendre le contrôle des systèmes, grands et petits ARC337 (y compris un travail constant)
- AWS re:Invent 2018 : Boucles serrées et ouverture d'esprit : comment prendre le contrôle des systèmes, grands et petits ARC337 (y compris le couplage lâche, le travail constant, la stabilité statique)
- AWS re:Invent 2019 : Passage à des architectures pilotées par les événements (08) SVS3

REL04-BP04 Rendre les opérations de mutation idempotentes

Un service idempotent garantit que chaque demande est traitée une seule fois, de sorte que la soumission de plusieurs demandes identiques ait le même effet que la soumission d'une seule demande. Il est ainsi plus facile pour un client d'implémenter de nouvelles tentatives sans craindre qu'une demande soit traitée plusieurs fois par erreur. Pour ce faire, les clients peuvent émettre des demandes d'API avec un jeton d'idempotence, qui est utilisé chaque fois que la demande est répétée. Une API de service idempotente utilise le jeton pour renvoyer une réponse identique à la réponse qui a été renvoyée la première fois que la demande a été traitée, même si l'état sous-jacent du système a changé.

Dans un système distribué, il est relativement simple d'effectuer une action au plus une fois (le client soumet une seule demande) ou au moins une fois (le client continue à soumettre des demandes jusqu'à ce qu'il reçoive une confirmation de succès). Il est plus difficile de garantir qu'une action est exécutée exactement une fois, de sorte que la soumission de plusieurs demandes identiques a le même effet qu'une seule demande. En utilisant des jetons d'idempotence dans les API, les services peuvent recevoir une demande de mutation une ou plusieurs fois sans avoir besoin de créer des enregistrements en double ou des effets secondaires.

Résultat escompté : vous disposez d'une approche cohérente, bien documentée et largement adoptée pour garantir l'idempotence sur l'ensemble des composants et services.

Anti-modèles courants :

- Vous appliquez l'idempotence sans distinction, même lorsque cela n'est pas nécessaire.
- Vous introduisez une logique trop complexe pour implémenter l'idempotence.
- Vous utilisez les horodatages comme des clés pour l'idempotence. Cela peut entraîner des inexactitudes en raison d'un décalage d'horloge ou du fait que plusieurs clients utilisent les mêmes horodatages pour appliquer les modifications.
- Vous stockez des données utiles complètes à des fins d'idempotence. Dans cette approche, vous enregistrez des données utiles complètes pour chaque demande et vous les remplacez à chaque nouvelle demande. Cela peut dégrader les performances et affecter la capacité de mise à l'échelle.
- Vous générez des clés de manière incohérente entre les services. En l'absence de clés cohérentes, les services peuvent ne pas reconnaître les demandes en double, ce qui peut conduire à des résultats indésirables.

Avantages liés au respect de cette bonne pratique :

- Capacité de mise à l'échelle accrue : le système peut gérer les nouvelles tentatives et les demandes en double sans avoir à appliquer une logique supplémentaire ni à gérer des états complexes.
- Fiabilité améliorée : l'idempotence aide les services à traiter plusieurs demandes identiques de manière cohérente, ce qui réduit le risque d'effets secondaires indésirables ou de doublons d'enregistrements. Cela est particulièrement important dans les systèmes distribués, où les défaillances du réseau et les nouvelles tentatives sont communes.
- Cohérence des données améliorée : étant donné qu'une même demande produit la même réponse, l'idempotence permet de maintenir la cohérence des données dans les systèmes distribués. Cela est essentiel pour maintenir l'intégrité des transactions et des opérations.
- Gestion des erreurs : les jetons d'idempotence facilitent la gestion des erreurs. Si un client ne reçoit pas de réponse en raison d'un problème, il peut renvoyer la demande en toute sécurité avec le même jeton d'idempotence.
- Transparence opérationnelle : l'idempotence permet une meilleure surveillance et une meilleure journalisation. Les services peuvent consigner les demandes avec leurs jetons d'idempotence, ce qui facilite le suivi et le débogage des problèmes.
- Contrat d'API simplifié : il peut simplifier le contrat entre les systèmes côté client et serveur et réduire la crainte d'un traitement erroné des données.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Dans un système distribué, il est relativement simple d'effectuer une action au plus une fois (le client soumet une seule demande) ou au moins une fois (le client continue à soumettre des demandes jusqu'à ce que le succès soit confirmé). Toutefois, il est difficile de mettre en œuvre un comportement exactement une fois. Pour ce faire, vos clients doivent générer et fournir un jeton d'idempotence pour chaque demande.

En utilisant des jetons d'idempotence, un service peut faire la distinction entre de nouvelles demandes et des demandes répétées. Lorsqu'un service reçoit une demande contenant un jeton d'idempotence, il vérifie si le jeton a déjà été utilisé. Si le jeton a été utilisé, le service extrait et retourne la réponse stockée. Si le jeton est nouveau, le service traite la demande, stocke la réponse avec le jeton, puis retourne la réponse. Ce mécanisme rend toutes les réponses idempotentes, ce qui améliore la fiabilité et la cohérence du système distribué.

L'idempotence est également un comportement important des architectures axées sur les événements. Ces architectures s'appuient généralement sur une file d'attente de messages telle qu'Amazon SQS, Amazon MQ, Amazon Kinesis Streams ou Amazon Managed Streaming for Apache Kafka (MSK). Dans certaines circonstances, un message qui n'a été publié qu'une seule fois peut être envoyé accidentellement plusieurs fois. Lorsqu'un diffuseur de publication génère et inclut des jetons d'idempotence dans des messages, il demande que le traitement de tout message reçu en double ne donne pas lieu à une action répétée pour le même message. Les consommateurs doivent suivre chaque jeton reçu et ignorer les messages contenant des jetons dupliqués.

Les services et les consommateurs doivent également transmettre le jeton d'idempotence reçu à tous les services en aval qu'il appelle. Chaque service en aval de la chaîne de traitement est de la même manière responsable de la mise en œuvre de l'idempotence afin d'éviter l'effet secondaire consistant à traiter un message plusieurs fois.

Étapes d'implémentation

1. Identification des opérations idempotentes

   Déterminez quelles opérations nécessitent l'idempotence. Il s'agit généralement des méthodes HTTP POST, PUT et DELETE et des opérations d'insertion, de mise à jour ou de suppression de base de données. Les opérations qui n'entraînent pas de mutation d'état, telles que les requêtes en lecture seule, ne nécessitent généralement pas l'idempotence, sauf si elles ont des effets secondaires.
2. Utilisation d'identifiants uniques

   Incluez un jeton unique dans chaque demande d'opération idempotente envoyée par l'expéditeur, soit directement dans la demande, soit au sein de ses métadonnées (par exemple, un en-tête HTTP). Cela permet au destinataire de reconnaître et de traiter les demandes ou opérations dupliquées. Les identifiants couramment utilisés pour les jetons incluent les identifiants uniques universels (UUID) et les identifiants KSUID (K-Sortable Unique Identifiers).
3. Suivi et gestion de l'état

   Tenez à jour l'état de chaque opération ou demande dans votre charge de travail. Pour ce faire, vous pouvez stocker le jeton d'idempotence et l'état correspondant (en attente, terminé ou échec) dans une base de données, un cache ou un autre stockage permanent. Ces informations d'état permettent à la charge de travail d'identifier et de traiter les demandes ou opérations en double.

   Maintenez la cohérence et l'atomicité en utilisant des mécanismes de contrôle de simultanéité appropriés si nécessaire, tels que des verrous, des transactions ou des contrôles de simultanéité

optimiste. Cela inclut le processus d'enregistrement du jeton idempotent et d'exécution de toutes les opérations de mutation associées au traitement de la demande. Cela contribue à éviter la survenue de conditions de concurrence et vérifie que les opérations idempotentes se déroulent correctement.

Supprimez régulièrement les anciens jetons d'idempotence de l'entrepôt de données pour gérer le stockage et les performances. Si votre système de stockage les prend en charge, pensez à utiliser des horodatages d'expiration pour les données (souvent appelés « durée de vie » ou valeurs TTL). La probabilité de réutilisation des jetons d'idempotence diminue avec le temps.

Les options de stockage AWS courantes généralement utilisées pour le stockage des jetons d'idempotence et de l'état associé incluent :

- Amazon DynamoDB : DynamoDB est un service de base de données NoSQL qui fournit des performances à faible latence et une haute disponibilité, ce qui le rend parfaitement adapté au stockage de données liées à l'idempotence. Le modèle de données clé-valeur et document de DynamoDB permet de stocker et d'extraire efficacement les jetons d'idempotence et les informations d'état associées. DynamoDB peut également faire expirer automatiquement les jetons d'idempotence si votre application définit une valeur TTL à leur insertion.
- Amazon ElastiCache : ElastiCache peut stocker des jetons d'idempotence à haut débit, à faible latence et à faible coût. ElastiCache (Redis) et ElastiCache (Memcached) peuvent tous les deux également faire expirer automatiquement les jetons d'idempotence si votre application définit une valeur TTL à leur insertion.
- Amazon Relational Database Service (RDS) : vous pouvez utiliser Amazon RDS pour stocker les jetons d'idempotence et les informations d'état associées, en particulier si votre application utilise déjà une base de données relationnelle à d'autres fins.
- Amazon Simple Storage Service (S3) : Amazon S3 est un service de stockage d'objets hautement évolutif et durable qui peut être utilisé pour stocker les jetons d'idempotence et les métadonnées associées. Les capacités de gestion des versions de S3 peuvent être particulièrement utiles pour maintenir l'état des opérations idempotentes. Le choix du service de stockage dépend généralement de facteurs tels que le volume de données liées à l'idempotence, les caractéristiques de performances requises, le besoin de durabilité et de disponibilité, et la manière dont le mécanisme d'idempotence s'intègre dans l'architecture globale de la charge de travail.

4. Mise en œuvre des opérations idempotentes

Concevez vos composants d'API et de charge de travail de manière à ce qu'ils soient idempotents. Incorporez des vérifications d'idempotence dans vos composants de charge de travail. Avant de traiter une demande ou d'effectuer une opération, vérifiez si l'identifiant unique a déjà été traité. Si tel est le cas, renvoyez le résultat précédent au lieu de réexécuter l'opération. Par exemple, si un client envoie une demande de création d'utilisateur, vérifiez si un utilisateur possédant le même identifiant unique existe déjà. Si un tel utilisateur existe, ses informations doivent être renvoyées au lieu de créer un nouvel utilisateur. De même, si un consommateur de file d'attente reçoit un message contenant un jeton d'idempotence dupliqué, le consommateur doit ignorer le message.

Créez des suites de tests complètes qui valident l'idempotence des demandes. Elles doivent couvrir un large éventail de scénarios, tels que des demandes réussies, des demandes ayant échoué et des demandes dupliquées.

Si votre charge de travail tire parti des fonctions AWS Lambda, envisagez d'utiliser Powertools for AWS Lambda. Powertools for AWS Lambda est une boîte à outils pour développeurs permettant de mettre en œuvre les bonnes pratiques en matière de technologies sans serveur et d'augmenter la rapidité des développeurs dans le cadre de l'utilisation des fonctions AWS Lambda. En particulier, il fournit un utilitaire pour convertir vos fonctions Lambda en opérations idempotentes qu'il est possible de réessayer en toute sécurité.

5. Communication claire de l'idempotence

   Documentez vos composants d'API et de charge de travail afin de communiquer clairement la nature idempotente des opérations. Cela permet aux clients de comprendre le comportement attendu et de savoir comment interagir de manière fiable avec votre charge de travail.

6. Surveillance et audit

   Mettez en œuvre des mécanismes de surveillance et d'audit pour détecter tout problème lié à l'idempotence des réponses, tel que des variations de réponse inattendues ou un traitement excessif des demandes dupliquées. Cela peut vous aider à détecter et à étudier tout problème ou comportement inattendu lié à votre charge de travail.

Ressources

Bonnes pratiques associées :

- REL05-BP03 Contrôler et limiter les appels de nouvelle tentative
- REL06-BP01 Surveiller tous les composants de la charge de travail (génération)

- REL06-BP03 Envoyer des notifications (traitement et alarmes en temps réel)
- REL08-BP02 Intégrer les tests fonctionnels dans le cadre de votre déploiement

Documents connexes :

- Amazon Builders' Library : sécurisation des nouvelles tentatives avec des API idempotentes
- L'Amazon Builders' Library : défis liés aux systèmes distribués
- L'Amazon Builders' Library : fiabilité, travail constant et une bonne tasse de café
- Amazon Elastic Container Service : garantie de l'idempotence
- Comment puis-je rendre ma fonction Lambda idempotente ?
- Procédure pour garantir l'idempotence dans les demandes d'API Amazon EC2

Vidéos connexes :

- Création d'applications distribuées avec une architecture axée sur les événements – AWS Online Tech Talks
- AWS re:Invent 2023 - Building next-generation applications with event-driven architecture
- AWS re:Invent 2023 - Advanced integration patterns & trade-offs for loosely coupled systems
- AWS re:Invent 2023 - Advanced event-driven patterns with Amazon EventBridge
- AWS re:Invent 2018 - Close Loops and Opening Minds: How to Take Control of Systems, Big and Small ARC337 (inclut les concepts de couplage faible, travail constant et stabilité statique)
- AWS re:Invent 2019 - Moving to event-driven architectures (SVS308)

Outils associés :

- Idempotence avec Powertools AWS Lambda (Java)
- Idempotence avec Powertools AWS Lambda (Python)
- Page GitHub de Powertools AWS Lambda

FIA 5. Comment concevoir les interactions dans un système distribué afin d'atténuer les défaillances ou d'y résister ?

Les systèmes distribués s'appuient sur des réseaux de communication pour interconnecter des composants (tels que des serveurs ou des services). Votre charge de travail doit fonctionner de

manière fiable malgré la perte de données ou la latence sur ces réseaux. Les composants du système distribué doivent fonctionner de manière à ne pas avoir d'impact négatif sur les autres composants ou sur la charge de travail. Ces bonnes pratiques permettent aux charges de travail de résister aux contraintes ou aux défaillances, de s'en remettre plus rapidement et d'atténuer l'impact de ces altérations. Il en résulte une amélioration du temps moyen de récupération (MTTR).

Bonnes pratiques

- REL05-BP01 Implémenter une dégradation progressive pour transformer les dépendances matérielles applicables en dépendances souples
- REL05-BP02 Demandes d'accélérateur
- REL05-BP03 Contrôler et limiter les appels de nouvelle tentative
- REL05-BP04 Échouer rapidement et limiter les files d'attente
- REL05-BP05 Définir les délais d'expiration des clients
- REL05-BP06 Rendre les systèmes apatrides dans la mesure du possible
- REL05-BP07 Mettre en œuvre des leviers de secours

REL05-BP01 Implémenter une dégradation progressive pour transformer les dépendances matérielles applicables en dépendances souples

Les composants de l'application doivent continuer à exécuter leur fonction principale même si les dépendances deviennent indisponibles. Ils peuvent fournir des données légèrement obsolètes, des données alternatives ou même aucune donnée. Cela garantit que le fonctionnement global du système n'est que très peu entravé par des défaillances localisées tout en fournissant une valeur commerciale centrale.

Résultat souhaité : lorsque les dépendances d'un composant ne sont pas en bon état, le composant lui-même peut continuer de fonctionner, mais de manière dégradée. Les modes de défaillance des composants doivent être considérés comme un fonctionnement normal. Les flux de travail doivent être conçus de manière à ce que ces défaillances n'aboutissent pas à une défaillance complète ou qu'elles aboutissent au moins à des états prévisibles et récupérables.

Anti-modèles courants :

- Ne pas identifier les fonctionnalités métier essentielles nécessaires. Ne pas tester le fonctionnement des composants, même en cas de défaillance des dépendances.
- Aucune donnée n'est diffusée en cas d'erreur ou lorsqu'une seule dépendance parmi plusieurs n'est pas disponible et que des résultats partiels peuvent toujours être renvoyés.

- Création d’un état incohérent lorsqu’une transaction échoue partiellement.
- Ne pas disposer d’un autre moyen d’accéder à un magasin de paramètres central.
- Invalider ou vider l’état local à la suite d’un échec d’actualisation sans prendre en compte les conséquences d’une telle opération.

Avantages du respect de cette bonne pratique : une dégradation progressive améliore la disponibilité du système dans son ensemble et maintient la fonctionnalité des fonctions les plus importantes même en cas de panne.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

La mise en œuvre d’une dégradation progressive permet de minimiser l’impact des défaillances de dépendance sur le fonctionnement des composants. Idéalement, un composant détecte les défaillances liées aux dépendances et les contourne de manière à avoir un impact minimal sur les autres composants ou les clients.

L’architecture permettant une dégradation progressive implique de prendre en compte les modes de défaillance potentiels lors de la conception des dépendances. Pour chaque mode de défaillance, déterminez un moyen de fournir la plupart des fonctionnalités, ou les plus critiques d’entre elles, du composant aux appelants ou aux clients. Ces considérations peuvent devenir des exigences supplémentaires qui peuvent être testées et vérifiées. Idéalement, un composant est capable d’exécuter sa fonction principale de manière acceptable, même en cas de défaillance d’une ou de plusieurs dépendances.

Il s’agit tout autant d’une discussion commerciale que technique. Toutes les exigences commerciales sont importantes et doivent être satisfaites dans la mesure du possible. Cependant, il est tout de même logique de se demander ce qui doit se passer lorsque toutes les exigences ne peuvent pas être satisfaites. Un système peut être conçu pour être disponible et cohérent, mais lorsqu’une exigence doit être supprimée, laquelle est la plus importante ? Pour le traitement des paiements, il peut s’agir de la cohérence. Pour une application en temps réel, il peut s’agir de la disponibilité. Pour un site Web orienté client, la réponse peut dépendre des attentes du client.

Ce que cela signifie dépend des exigences du composant et de ce qui doit être considéré comme sa fonction principale. Par exemple :

- un site Web d’e-commerce peut afficher des données provenant de plusieurs systèmes différents, par exemple des recommandations personnalisées, les produits les mieux classés et l’état des

commandes des clients sur la page de destination. Lorsqu'un système en amont est défaillant, il est tout de même judicieux d'afficher tout le reste au lieu d'afficher une page d'erreur à un client.

- Un composant effectuant des écritures par lots peut toujours continuer à traiter un lot si l'une des opérations individuelles échoue. La mise en œuvre d'un mécanisme de nouvelle tentative doit être simple. Cela peut être fait en renvoyant à l'appelant des informations indiquant quelles opérations ont réussi, lesquelles ont échoué et pourquoi elles ont échoué, ou en plaçant les demandes ayant échoué dans une file d'attente de lettres mortes pour implémenter des tentatives asynchrones. Les informations relatives aux opérations ayant échoué doivent également être consignées.
- Un système qui traite les transactions doit vérifier que toutes les mises à jour individuelles sont exécutées ou qu'aucune d'entre elles ne l'est. Pour les transactions distribuées, le modèle Saga peut être utilisé pour annuler les opérations précédentes en cas d'échec d'une opération ultérieure de la même transaction. Ici, la fonction principale est de maintenir la cohérence.
- Les systèmes soumis à des contraintes de temps doivent être en mesure de gérer les dépendances qui ne répondent pas en temps voulu. Dans ces cas de figure, le modèle du disjoncteur peut être utilisé. Lorsque les réponses d'une dépendance commencent à expirer, le système peut passer à un état fermé où aucun appel supplémentaire n'est effectué.
- Une application peut lire des paramètres à partir d'un magasin de paramètres. Il peut être utile de créer des images de conteneur avec un ensemble de paramètres par défaut et de les utiliser si le magasin de paramètres n'est pas disponible.

Notez que les chemins empruntés en cas de défaillance d'un composant doivent être testés et doivent être nettement plus simples que le chemin principal. Généralement, les stratégies de repli doivent être évitées.

Étapes d'implémentation

Identifiez les dépendances externes et internes. Déterminez quels types de défaillances peuvent y survenir. Réfléchissez à des moyens de minimiser l'impact négatif sur les systèmes en amont et en aval, ainsi que sur les clients lors de ces défaillances.

Vous trouverez ci-dessous une liste des dépendances et la manière de les dégrader de façon appropriée en cas d'échec :

1. Défaillance partielle des dépendances : un composant peut adresser plusieurs demandes à des systèmes en aval, soit sous la forme de demandes multiples adressées à un système, soit sous celle d'une demande adressée à plusieurs systèmes. Selon le contexte métier, différentes

méthodes de gestion peuvent être appropriées (pour plus de détails, voir les exemples précédents dans le guide de mise en œuvre).

2. Un système en aval est incapable de traiter les demandes en raison d'une charge élevée : si les demandes adressées à un système en aval échouent régulièrement, il n'est pas logique de continuer à réessayer. Cela peut créer une charge supplémentaire sur un système déjà surchargé et rendre la récupération plus difficile. Le modèle du disjoncteur peut être utilisé ici afin de surveiller les appels en échec vers un système en aval. Si un grand nombre d'appels échouent, il cessera d'envoyer d'autres demandes au système en aval et n'autorisera les appels qu'occasionnellement pour vérifier si le système en aval est à nouveau disponible.
3. Aucun magasin de paramètres n'est disponible : pour transformer un magasin de paramètres, vous pouvez utiliser la mise en cache des dépendances souples ou des valeurs par défaut saines incluses dans les images de conteneur ou de machine. Notez que ces valeurs par défaut doivent être conservées up-to-date et incluses dans les suites de tests.
4. Aucun service de surveillance ou autre dépendance non fonctionnelle n'est disponible : si un composant ne peut pas envoyer par intermittence des journaux, des métriques ou des traces à un service de surveillance central, il est souvent préférable de continuer à exécuter les fonctions métier comme d'habitude. Il est souvent inacceptable de ne pas enregistrer ni de pousser des métriques pendant une longue période. En outre, certains cas d'utilisation peuvent nécessiter des entrées d'audit complètes pour répondre aux exigences de conformité.
5. Il est possible qu'une instance principale d'une base de données relationnelle ne soit pas disponible : Amazon Relational Database Service, comme presque toutes les bases de données relationnelles, ne peut avoir qu'une seule instance de rédacteur principal. Cela crée un point de défaillance unique pour les charges de travail d'écriture et complique la mise à l'échelle. Ce problème peut être partiellement atténué en utilisant une configuration multi-AZ pour une haute disponibilité ou Amazon Aurora sans serveur pour une meilleure mise à l'échelle. Pour des exigences de très haute disponibilité, il peut être judicieux de ne pas se fier du tout au rédacteur principal. Pour les requêtes qui se limitent à la lecture, des répliques de lecture peuvent être utilisées, ce qui assure la redondance et la possibilité d'une augmentation horizontale, et pas seulement d'une augmentation verticale. Les écritures peuvent être mises en mémoire tampon, par exemple dans une file d'attente Amazon Simple Queue Service, afin que les demandes d'écriture des clients puissent toujours être acceptées même si le serveur principal est temporairement indisponible.

Ressources

Documents connexes :

- Amazon API Gateway : limitez les API demandes pour un meilleur débit
- CircuitBreaker(résume le disjoncteur de « Release It ! » livre)
- Rétentatives d'erreur et retard exponentiel dans AWS
- Michael Nygard « Release It! Design and Deploy Production-Ready Software »
- L'Amazon Builders' Library : éviter le basculement dans les systèmes distribués
- L'Amazon Builders' Library : éviter les retards de file d'attente insurmontables
- L'Amazon Builders' Library : défis et stratégies de mise en cache
- Amazon Builders' Library : délais d'attente, nouvelles tentatives et backoff avec instabilité

Vidéos connexes :

- Réessayez, attendez et agitez : AWS re:Invent 2019 : Introducing The Amazon Builders' Library () DOP328

Exemples connexes :

- Atelier Well-Architected : niveau 300 : implémentation de la surveillance de l'état et gestion des dépendances pour améliorer la fiabilité

REL05-BP02 Demandes d'accélérateur

Limitez les demandes pour atténuer l'épuisement des ressources en cas d'augmentation imprévue de la demande. Les demandes inférieures aux taux de limitation sont traitées tandis que celles dépassant la limite définie sont rejetées avec un message de retour indiquant que la demande a été limitée.

Résultat souhaité : les pics de volume importants, qu'ils soient dus à une augmentation soudaine du trafic client, à des inondations ou à des tempêtes de nouvelles tentatives, sont atténués par la limitation des demandes, ce qui permet aux charges de travail de poursuivre le traitement normal du volume de demandes pris en charge.

Anti-modèles courants :

- APIles limites des points de terminaison ne sont pas implémentées ou sont laissées aux valeurs par défaut sans tenir compte des volumes attendus.

- APIles points de terminaison ne sont pas testés en charge ou les limites d'étranglement ne sont pas testées.
- Limiter les taux de demandes sans tenir compte de la taille ou de la complexité des demandes.
- Tester les taux de demande maximaux ou la taille maximale des demandes, mais pas les deux simultanément.
- Les ressources ne sont pas provisionnées selon les mêmes limites établies lors des tests.
- Les plans d'utilisation n'ont pas été configurés ou pris en compte pour les API consommateurs d'application à application (A2A).
- Les utilisateurs de files d’attente qui mettent à l’échelle horizontalement ne disposent pas de paramètres de simultanéité maximaux configurés.
- La limitation du débit par adresse IP n’a pas été mise en œuvre.

Avantages du respect de cette bonne pratique : les charges de travail qui fixent des limites peuvent fonctionner normalement et traiter correctement le chargement des demandes acceptées en cas de pics de volume inattendus. Les pics soudains ou soutenus de demandes et de files d'attente sont limités APIs et n'épuisent pas les ressources de traitement des demandes. Les limites de débit limitent les demandes individuelles, de sorte que les volumes élevés de trafic provenant d'une seule adresse IP ou d'un seul API consommateur n'épuisent pas les ressources d'autres consommateurs.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Les services doivent être conçus pour traiter une capacité connue de demandes ; cette capacité peut être établie par des tests de charge. Si les taux d’arrivée des demandes dépassent les limites, la réponse appropriée indique qu’une demande a été limitée. Cela permet au consommateur de gérer l’erreur et de réessayer ultérieurement.

Lorsque votre service nécessite une implémentation de limitation, pensez à implémenter l’algorithme du compartiment à jetons, dans lequel un jeton compte pour une demande. Les jetons sont rechargés à une vitesse limitée par seconde et vidés de manière asynchrone à raison d’un jeton par demande.

L’algorithme du compartiment à jetons.

Amazon API Gateway implémente l'algorithme du token bucket en fonction des limites du compte et de la région et peut être configuré par client avec des plans d'utilisation. En outre, Amazon Simple Queue Service (AmazonSQS) et Amazon Kinesis peuvent mettre en mémoire tampon les demandes afin de réduire le taux de demandes et de permettre des taux de limitation plus élevés pour les demandes pouvant être traitées. Enfin, vous pouvez implémenter une limitation de débit AWS WAFpour limiter API les consommateurs spécifiques qui génèrent une charge anormalement élevée.

Étapes d’implémentation

Vous pouvez configurer API Gateway avec des limites de limitation pour vos erreurs APIs et renvoyer des `429 Too Many Requests` erreurs lorsque les limites sont dépassées. Vous pouvez l'utiliser AWS WAF avec vos points de terminaison AWS AppSync et API Gateway pour activer la limitation du débit par adresse IP. En outre, lorsque votre système peut tolérer un traitement asynchrone, vous pouvez placer les messages dans une file d’attente ou un flux pour accélérer les réponses aux clients du service, ce qui vous permet d’atteindre des taux de limitation plus élevés.

Avec le traitement asynchrone, lorsque vous avez configuré Amazon SQS comme source d'événements pour AWS Lambda, vous pouvez configurer une simultanéité maximale afin d'éviter que des taux d'événements élevés ne consomment le quota d'exécution simultanée du compte disponible nécessaire pour les autres services de votre charge de travail ou de votre compte.

Bien que API Gateway fournisse une implémentation gérée du bucket de jetons, dans les cas où vous ne pouvez pas utiliser API Gateway, vous pouvez tirer parti des implémentations open source

spécifiques au langage (voir les exemples connexes dans Ressources) du bucket de jetons pour vos services.

- Comprenez et configurez les limites de limitation de API Gateway au niveau du compte, par région, par étape, et API par API clé par niveau de plan d'utilisation.
- Appliquez des règles AWS WAF de limitation de débit à API Gateway et aux AWS AppSync terminaux pour vous protéger contre les inondations et bloquer les programmes malveillantsIPs. Les règles de limitation du débit peuvent également être configurées sur AWS AppSync API les clés pour les consommateurs A2A.
- Déterminez si vous avez besoin de plus de contrôle de la régulation que de la limitation du débit et AWS AppSync APIs, dans l'affirmative, configurez une API passerelle devant votre AWS AppSync terminal.
- Lorsque les SQS files d'attente Amazon sont configurées comme déclencheurs pour les consommateurs de files d'attente Lambda, définissez la simultanéité maximale sur une valeur qui traite suffisamment pour atteindre vos objectifs de niveau de service mais qui ne respecte pas les limites de simultanéité ayant un impact sur les autres fonctions Lambda. Envisagez de définir une simultanéité réservée pour d'autres fonctions Lambda du même compte et de la même région lorsque vous utilisez des files d'attente avec Lambda.
- Utilisez API Gateway avec des intégrations de services natives à Amazon SQS ou Kinesis pour mettre en mémoire tampon les demandes.
- Si vous ne pouvez pas utiliser API Gateway, consultez les bibliothèques spécifiques au langage pour implémenter l'algorithme Token bucket adapté à votre charge de travail. Consultez la section des exemples et faites vos propres recherches pour trouver une bibliothèque appropriée.
- Testez les limites que vous envisagez de définir ou d'autoriser à augmenter, et documentez les limites testées.
- N'augmentez pas les limites au-delà de ce que vous avez établi lors des tests. Lorsque vous augmentez une limite, vérifiez que les ressources provisionnées sont déjà équivalentes ou supérieures à celles des scénarios de test avant d'appliquer l'augmentation.

Ressources

Bonnes pratiques associées :

- REL04-BP03 Faire un travail constant
- REL05-BP03 Contrôler et limiter les appels de nouvelle tentative

Documents connexes :

- Amazon API Gateway : limitez les API demandes pour un meilleur débit
- AWS WAF : instruction de règle fréquentielle
- Introduction d'une simultanéité maximale AWS Lambda lors de l'utilisation d'Amazon SQS comme source d'événements
- AWS Lambda : simultanéité maximale

Exemples connexes :

- Les trois principales règles AWS WAF basées sur les taux
- Bucket4j Java
- Jeton-compartiment Python
- Nœud jeton-compartiment
- . NETLimitation du taux de filetage du système

Vidéos connexes :

- Implémentation des meilleures pratiques de API sécurité GraphQL avec AWS AppSync

Outils associés :

- APIPasserelle Amazon
- AWS AppSync
- Amazon SQS
- Amazon Kinesis
- AWS WAF

REL05-BP03 Contrôler et limiter les appels de nouvelle tentative

Utilisez le backoff exponentiel pour relancer les demandes à des intervalles de plus en plus longs entre chaque nouvelle tentative. Introduisez un décalage entre les tentatives afin de randomiser les intervalles entre les tentatives. Limitez le nombre maximal de tentatives.

Résultat escompté : Les composants typiques d'un système logiciel distribué incluent les serveurs, les équilibreurs de charge, les bases de données et DNS les serveurs. Pendant le fonctionnement normal, ces composants peuvent répondre aux demandes par des erreurs temporaires ou limitées, ainsi que par des erreurs qui persisteraient indépendamment des nouvelles tentatives. Lorsque des clients adressent des demandes à des services, celles-ci consomment des ressources, notamment de la mémoire, des threads, des connexions, des ports ou toute autre ressource limitée. Le contrôle et la limitation des nouvelles tentatives constituent une stratégie visant à libérer et à minimiser la consommation de ressources afin que les composants du système soumis à des contraintes ne soient pas surchargés.

Lorsque le client demande une expiration du délai ou reçoit des réponses d'erreur, il doit décider de réessayer ou non. S'il recommence, il le fait avec un backoff exponentiel avec une instabilité et une valeur de nouvelle tentative maximale. Par conséquent, les services et processus back-end sont moins sollicités et le temps nécessaire pour s'autoréparer est réduit, ce qui se traduit par une récupération plus rapide et un traitement efficace des demandes.

Anti-modèles courants :

- Implémentation de nouvelles tentatives sans ajouter de valeurs de backoff exponentiel, d'instabilité et de nouvelles tentatives maximales. Le backoff et l'instabilité permettent d'éviter les pics de trafic artificiels dus à des tentatives involontaires coordonnées à intervalles réguliers.
- Implémentation de nouvelles tentatives sans tester leurs effets ou en supposant que les nouvelles tentatives sont déjà intégrées ou SDK sans test de scénarios de nouvelle tentative.
- Incapacité à comprendre les codes d'erreur publiés à partir des dépendances, ce qui entraîne une nouvelle tentative pour toutes les erreurs, y compris celles dont la cause claire indique un manque d'autorisation, une erreur de configuration ou toute autre condition qui, comme on pouvait s'y attendre, ne sera pas résolue sans intervention manuelle.
- Ne pas aborder les pratiques d'observabilité, notamment la surveillance et l'envoi d'alertes en cas de pannes de service répétées afin que les problèmes sous-jacents soient connus et puissent être résolus.
- Développement de mécanismes de nouvelle tentative personnalisés lorsque des fonctionnalités de nouvelle tentative intégrées ou tierces suffisent.
- Réessayer à plusieurs couches de votre pile d'applications d'une manière qui complique les nouvelles tentatives et augmente la consommation de ressources lors d'une tempête de nouvelles tentatives. Assurez-vous de comprendre comment ces erreurs affectent votre application et les dépendances sur lesquelles vous vous appuyez, puis implémentez les nouvelles tentatives à un seul niveau.

- Réessayer les appels de service qui ne sont pas idempotents, ce qui peut entraîner des effets secondaires inattendus tels que des résultats dupliqués.

Avantages du respect de cette bonne pratique : les nouvelles tentatives aident les clients à obtenir les résultats souhaités lorsque les requêtes échouent, mais elles font également perdre plus de temps au serveur pour obtenir les réponses souhaitées. Lorsque les défaillances sont rares ou transitoires, les nouvelles tentatives fonctionnent bien. Lorsque les défaillances sont causées par une surcharge de ressources, les nouvelles tentatives peuvent aggraver la situation. L'ajout d'un backoff exponentiel avec instabilité aux nouvelles tentatives des clients permet aux serveurs de se rétablir en cas de défaillance provoquée par une surcharge de ressources. L'instabilité permet d'éviter l'alignement des demandes en pics, tandis que le backoff réduit l'escalade de charge provoquée par l'ajout de nouvelles tentatives au chargement normal des demandes. Enfin, il est important de configurer un nombre maximal de nouvelles tentatives ou un temps écoulé afin d'éviter de créer des backlogs susceptibles d'entraîner des échecs métastables.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Contrôler et limiter les appels de nouvelle tentative. Utilisez le backoff exponentiel pour réessayer après des intervalles progressivement plus longs. Introduisez l'instabilité pour randomiser les intervalles de nouvelle tentative et limiter le nombre maximal de nouvelles tentatives.

Certains AWS SDKs implémentent les nouvelles tentatives et le recul exponentiel par défaut. Utilisez ces AWS implémentations intégrées, le cas échéant, dans votre charge de travail. Implémentez une logique similaire dans votre charge de travail lorsque vous appelez des services qui sont idempotents et où les nouvelles tentatives améliorent la disponibilité de vos clients. Déterminez quels sont les délais d'expiration et quand les nouvelles tentatives doivent s'arrêter en fonction de votre cas d'utilisation. Créez et mettez en pratique des scénarios de test pour ces cas d'utilisation impliquant de nouvelles tentatives.

Étapes d'implémentation

- Déterminez la couche optimale de votre pile d'applications pour implémenter de nouvelles tentatives pour les services sur lesquels repose votre application.
- Tenez compte des stratégies de relance éprouvées SDKs qui mettent en œuvre des stratégies de relance éprouvées avec un retard et une instabilité exponentiels dans la langue de votre choix, et privilégiez ces stratégies par rapport à l'écriture de vos propres implémentations de nouvelle tentative.

- Vérifiez que les services sont idempotents avant d'implémenter de nouvelles tentatives. Une fois les nouvelles tentatives mises en œuvre, assurez-vous qu'elles sont à la fois testées et régulièrement mises en œuvre en production.
- Lorsque vous appelez le AWS serviceAPIs, utilisez les options de configuration AWS SDKsAWS CLIet et comprenez la nouvelle tentative. Déterminez si les valeurs par défaut conviennent à votre cas d'utilisation, testez-les et ajustez-les si nécessaire.

Ressources

Bonnes pratiques associées :

- REL04-BP04 Rendre les opérations de mutation idempotentes
- REL05-BP02 Demandes d'accélérateur
- REL05-BP04 Échouer rapidement et limiter les files d'attente
- REL05-BP05 Définir les délais d'expiration des clients
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances

Documents connexes :

- Rétentatives d'erreur et retard exponentiel dans AWS
- Amazon Builders' Library : délais d'attente, nouvelles tentatives et backoff avec instabilité
- Instabilité et backoff exponentiel
- Sécuriser les nouvelles tentatives avec idempotent APIs

Exemples connexes :

- Nouvelle tentative Spring
- Nouvelle tentative Resilience4j

Vidéos connexes :

- Réessayez, attendez et agitez : AWS re:Invent 2019 : Introducing The Amazon Builders' Library () DOP328

Outils associés :

- AWS SDKset outils : comportement des nouvelles tentatives
- AWS Command Line Interface: nouvelles AWS CLI tentatives

REL05-BP04 Échouer rapidement et limiter les files d'attente

Lorsqu'un service n'est pas en mesure de répondre correctement à une demande, procédez à son interruption immédiate. Cela permet la libération des ressources associées à une demande et donne la possibilité au service de récupérer s'il lui manque des ressources. L'interruption immédiate est un modèle de conception logicielle bien établi qui peut être exploité pour créer des charges de travail hautement fiables dans le cloud. La mise en file d'attente est également un modèle d'intégration d'entreprise bien établi qui permet de faciliter le chargement et de permettre aux clients de libérer des ressources lorsque le traitement asynchrone peut être toléré. Lorsqu'un service est capable de répondre correctement dans des conditions normales, mais échoue lorsque le taux de demandes est trop élevé, utilisez une file d'attente pour mettre les demandes en mémoire tampon. Toutefois, ne permettez pas l'accumulation de longs backlogs de files d'attente susceptibles d'entraîner le traitement de demandes obsolètes auxquelles un client a déjà renoncé.

Résultat souhaité : lorsque les systèmes sont confrontés à des problèmes de ressources, à des dépassements de délai, à des exceptions ou à des pannes grises qui rendent les objectifs de niveau de service irréalisables, les stratégies d'interruption immédiate permettent d'accélérer la récupération du système. Les systèmes qui doivent absorber les pics de trafic et peuvent prendre en charge le traitement asynchrone peuvent améliorer la fiabilité en permettant aux clients de lancer rapidement des demandes grâce à l'utilisation des files d'attente pour mettre en mémoire tampon les demandes envoyées aux services back-end. Lors de la mise en mémoire tampon des demandes dans des files d'attente, des stratégies de gestion des files d'attente sont mises en œuvre pour éviter des backlogs insurmontables.

Anti-modèles courants :

- Implémentation de files de messages, mais pas de configuration de files d'attente de lettres mortes (DLQ) ou d'alarmes sur les DLQ volumes pour détecter les défaillances d'un système.
- Il ne s'agit pas de mesurer l'ancienneté des messages dans une file d'attente, mais de mesurer la latence pour comprendre quand les utilisateurs prennent du retard ou si des erreurs entraînent de nouvelles tentatives.
- Conservation des messages en attente dans une file d'attente, alors qu'il n'est plus utile de traiter ces messages si l'entreprise n'en a plus besoin.

- La configuration de files d'attente « premier entré, premier sorti » (FIFO) lorsque « dernier entré, premier sorti » (LIFO) répondrait mieux aux besoins des clients, par exemple lorsqu'il n'est pas nécessaire de passer des commandes strictes et que le traitement des arriérés retarde toutes les nouvelles demandes urgentes, ce qui se traduit par un dépassement des niveaux de service pour tous les clients.
- Exposer les files d'attente internes aux clients au lieu de les exposer à ceux APIs qui gèrent l'admission au travail et placent les demandes dans les files d'attente internes.
- La combinaison d'un trop grand nombre de types de demandes de travail dans une seule file d'attente peut aggraver les problèmes de backlog en répartissant la demande de ressources entre les types de demandes.
- Traitement de demandes complexes et simples dans la même file d'attente, malgré la nécessité d'une surveillance, de délais d'expiration et d'allocations de ressources différents.
- Absence de validation des entrées ou utilisation des assertions pour implémenter des mécanismes d'interruption immédiate dans les logiciels qui génèrent des exceptions vers des composants de niveau supérieur capables de gérer les erreurs de façon appropriée.
- Absence de suppression des ressources défectueuses du routage des requêtes, en particulier lorsque les défaillances sont grises, ce qui indique à la fois des réussites et des échecs en raison d'un plantage et d'un redémarrage, d'une panne de dépendance intermittente, d'une capacité réduite ou d'une perte de paquets réseau.

Avantages du respect de cette bonne pratique : les systèmes avec interruption immédiate sont plus faciles à déboguer et à corriger, et présentent souvent des problèmes de codage et de configuration avant la publication des versions en production. Les systèmes qui intègrent des stratégies de mise en file d'attente efficaces offrent une résilience et une fiabilité accrues face aux pics de trafic et aux pannes intermittentes du système.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les stratégies d'interruption immédiate peuvent être codées dans des solutions logicielles ou configurées dans l'infrastructure. En plus de leur capacité d'interruption immédiate, les files d'attente constituent une technique architecturale simple mais puissante qui permet de découpler les composants du système en douceur. Amazon CloudWatch fournit des fonctionnalités de surveillance et d'alarme en cas de panne. Une fois que l'on sait qu'un système est défaillant, des stratégies d'atténuation peuvent être invoquées, notamment en cas de défaillance de ressources

altérées. Lorsque les systèmes mettent en œuvre des files d'attente avec Amazon SQS et d'autres technologies de file d'attente pour faciliter le chargement, ils doivent réfléchir à la manière de gérer les arriérés de files d'attente, ainsi que les défaillances liées à la consommation de messages.

Étapes d’implémentation

- Implémentez des assertions programmatiques ou des métriques spécifiques dans votre logiciel et utilisez-les pour signaler explicitement les problèmes du système. Amazon vous CloudWatch aide à créer des métriques et des alarmes en fonction du modèle de journal des applications et de SDK l'instrumentation.
- Utilisez CloudWatch les métriques et les alarmes pour éviter les ressources altérées qui ajoutent de la latence au traitement ou qui échouent à traiter les demandes à plusieurs reprises.
- Utilisez le traitement asynchrone en concevant de manière APIs à accepter les demandes et à les ajouter aux files d'attente internes à l'aide d'SQSAmazon, puis à répondre au client émetteur du message par un message de réussite afin que le client puisse libérer des ressources et passer à autre chose pendant que les clients de la file d'attente principale traitent les demandes.
- Mesurez et surveillez le temps de latence du traitement des files d'attente en produisant une CloudWatch métrique chaque fois que vous retirez un message d'une file d'attente en le comparant à l'horodatage du message.
- Lorsque des défaillances empêchent le bon traitement des messages ou que des pics de trafic concernent des volumes qui ne peuvent pas être traités conformément aux contrats de niveau de service, mettez de côté le trafic ancien ou excédentaire vers une file d’attente de débordement. Cela permet de traiter en priorité les nouvelles tâches et les tâches plus anciennes lorsque la capacité est disponible. Cette technique est une approximation du LIFO traitement et permet un traitement normal du système pour tous les nouveaux travaux.
- Utilisez des lettres mortes ou réadaptez des files d’attente afin de déplacer les messages qui ne peuvent pas être traités hors du backlog vers un emplacement pouvant faire l’objet de recherches et de résolutions ultérieures.
- Réessayez ou, si cela est acceptable, supprimez les anciens messages en comparant l’heure actuelle à l’horodatage du message et en supprimant les messages qui ne sont plus pertinents pour le client demandeur.

Ressources

Bonnes pratiques associées :

- REL04-BP02 Implémenter des dépendances faiblement couplées

- REL05-BP02 Demandes d'accélérateur
- REL05-BP03 Contrôler et limiter les appels de nouvelle tentative
- REL06-BP02 Définir et calculer des métriques (agrégation)
- REL06-BP07 Surveillez le end-to-end suivi des demandes via votre système

Documents connexes :

- Éviter les backlogs insurmontables dans les files d’attente
- Interruption immédiate
- Comment puis-je éviter un arriéré croissant de messages dans ma SQS file d'attente Amazon ?
- Elastic Load Balancing : changement de zone
- Amazon Application Recovery Controller : contrôle du routage pour le basculement du trafic

Exemples associés :

- Modèles d’intégration d’entreprise : canal des lettres mortes

Vidéos connexes :

- AWS re:Invent 2022 - Utilisation d'applications multi-AZ à haute disponibilité

Outils associés :

- Amazon SQS
- Amazon MQ
- AWS IoT Core
- Amazon CloudWatch

REL05-BP05 Définir les délais d'expiration des clients

Définissez les délais d’expiration de manière appropriée pour les connexions et les demandes, vérifiez-les systématiquement et ne vous fiez pas aux valeurs par défaut, car elles ne tiennent pas compte des spécificités de la charge de travail.

Résultat souhaité : les délais d'expiration du client doivent prendre en compte le coût pour le client, le serveur et la charge de travail associés à l'attente des demandes dont le traitement prend un temps anormal. Dans la mesure où il est impossible de connaître la cause exacte d'un délai d'attente, les clients doivent utiliser leur connaissance des services pour établir des attentes relatives aux causes probables et aux délais d'expiration appropriés.

Le délai d'expiration des connexions client dépend des valeurs configurées. Après avoir dépassé le délai imparti, les clients décident de revenir en arrière et de réessayer ou d'ouvrir un coupe-circuit. Ces modèles évitent d'émettre des demandes susceptibles d'exacerber un problème d'erreur sous-jacent.

Anti-modèles courants :

- Ne pas connaître les délais d'expiration du système ou les délais d'expiration par défaut.
- Ne pas connaître le délai normal d'exécution des demandes.
- Ne pas connaître les raisons pour lesquelles les demandes peuvent prendre un temps anormalement long à traiter, ni les coûts pour le client, le service ou les performances de la charge de travail associés à l'attente de ces traitements.
- Ne pas connaître la probabilité qu'un réseau défaillant entraîne l'échec d'une requête uniquement lorsque le délai d'expiration est atteint, ainsi que les coûts pour les performances du client et de la charge de travail si l'on n'adopte pas un délai d'expiration plus court.
- Ne pas tester les scénarios de délai d'expiration à la fois pour les connexions et les demandes.
- Définir des délais d'expiration trop élevés, ce qui peut entraîner de longs temps d'attente et augmenter l'utilisation des ressources.
- Définir des délais d'attente trop bas, ce qui entraîne des défaillances artificielles.
- Oublier les modèles pour gérer les erreurs de temporisation des appels distants, par exemple les disjoncteurs et les nouvelles tentatives.
- Ne pas envisager de surveiller les taux d'erreur des appels de service, les objectifs de niveau de service en matière de latence et les valeurs aberrantes en matière de latence. Ces métriques peuvent fournir des informations sur les délais d'attente agressifs ou permissifs.

Avantages du respect de cette bonne pratique : les délais d'expiration des appels distants sont configurés et les systèmes sont conçus pour gérer les délais d'expiration de façon appropriée afin de préserver les ressources lorsque les appels distants répondent de manière anormalement lente et que les erreurs de délai d'expiration sont gérées de façon appropriée par les clients du service.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Définissez un délai d’expiration de la connexion et un délai d’expiration de la demande pour tout appel de dépendance de service et, plus généralement, pour tout appel entre processus. De nombreux cadres proposent des capacités de délai d’expiration intégrées, mais soyez prudent, car certains ont des valeurs par défaut infinies ou supérieures à ce qui est acceptable pour vos objectifs de service. Une valeur trop élevée réduit l’utilité du délai d’attente, car les ressources continuent d’être consommées pendant que le client attend l’expiration du délai. Une valeur trop faible peut générer un trafic accru sur le back-end et une latence accrue en raison du nombre excessif de demandes réessayées. Dans certains cas, cela peut entraîner des interruptions complètes, car toutes les demandes font l’objet d’une nouvelle tentative.

Tenez compte des points suivants lorsque vous déterminez des stratégies de délai d’expiration :

- Le traitement des demandes peut prendre plus de temps que d’habitude en raison de leur contenu, de défaillances d’un service cible ou d’une panne de partition réseau.
- Les demandes dont le contenu est anormalement coûteux peuvent consommer des ressources inutiles du serveur et du client. Dans ce cas, le fait d’avoir un délai d’expiration pour ces demandes et de ne pas réessayer peut préserver les ressources. Les services doivent également se protéger contre les contenus anormalement coûteux avec des limites et des délais d’expiration côté serveur.
- Les demandes qui prennent anormalement longtemps en raison d’une défaillance du service peuvent être interrompues et réessayées. Il convient de tenir compte des coûts de service liés à la demande et à la nouvelle tentative, mais si la cause est une déficience localisée, une nouvelle tentative ne sera probablement pas coûteuse et réduira la consommation de ressources du client. Le délai d’expiration peut également libérer des ressources du serveur en fonction de la nature de la déficience.
- Les demandes dont l’exécution prend beaucoup de temps parce que la demande ou la réponse n’a pas été envoyée par le réseau peuvent être interrompues et réessayées. La demande ou la réponse n’ayant pas été envoyée, il en aurait résulté un échec indépendamment de la durée du délai imparti. Dans ce cas, l’expiration du délai ne libérera pas les ressources du serveur, mais des ressources client et cela améliorera les performances de la charge de travail.

Tirez parti de modèles de conception bien établis, tels que les nouvelles tentatives et les disjoncteurs, pour gérer les délais d'attente avec élégance et prendre en charge les approches rapides. AWS SDKsset AWS CLIpermettent de configurer les délais d'expiration des connexions et des demandes

ainsi que les nouvelles tentatives avec un retard et une instabilité exponentiels. AWS Lambdales fonctions prennent en charge la configuration des délais d'attente, et avec AWS Step Functions, vous pouvez créer des disjoncteurs low code qui tirent parti des intégrations prédéfinies avec les services et. AWS SDKs AWS App Mesh Envoy fournit des fonctionnalités de délai d’expiration et de disjoncteur.

Étapes d’implémentation

- Configurez les délais d’expiration pour les appels de service à distance et profitez des fonctionnalités de délai d’expiration spécifique à la langue intégrées ou des bibliothèques de délai d’expiration open source.
- Lorsque votre charge de travail passe des appels avec un AWS SDK, consultez la documentation pour connaître la configuration du délai d'expiration spécifique à la langue.
  - Python
  - PHP
  - .NET
  - Ruby
  - Java
  - Go
  - Node.js
  - C++
- Lorsque vous utilisez des AWS CLI commandes AWS SDKs or dans votre charge de travail, configurez les valeurs de délai d'expiration par défaut en définissant les AWS valeurs par défaut pour `connectTimeoutInMillis` et. `tlsNegotiationTimeoutInMillis`
- Appliquez des options de ligne de commande `cli-connect-timeout` et `cli-read-timeout` contrôlez des AWS CLI commandes ponctuelles aux AWS services.
- Surveillez les appels de service à distance pour détecter les délais d’expiration et définissez des alarmes en cas d’erreurs persistantes afin de pouvoir gérer de manière proactive les scénarios d’erreur.
- Mettez en œuvre CloudWatch des mesures et CloudWatch une détection des anomalies sur les taux d'erreur des appels, les objectifs de niveau de service en matière de latence et les valeurs aberrantes de latence afin de mieux comprendre la gestion des délais d'attente trop agressifs ou trop permissifs.
- Configurez les délais d’expiration des fonctions Lambda.

- APILes clients Gateway doivent implémenter leurs propres tentatives lors de la gestion des délais d'expiration. APIGateway prend en charge un délai d'intégration de 50 millisecondes à 29 secondes pour les intégrations en aval et ne réessaie pas lorsque les demandes d'intégration expirent.
- Implémentez le modèle de disjoncteur pour éviter de passer des appels à distance lorsque le délai d'expiration est écoulé. Ouvrez le circuit pour éviter les échecs d'appels et fermez-le lorsque les appels répondent normalement.
- Pour les charges de travail basées sur des conteneurs, consultez les fonctionnalités d'App Mesh Envoy pour tirer parti des délais d'attente et des disjoncteurs intégrés.
- AWS Step Functions À utiliser pour créer des disjoncteurs à faible code pour les appels de service à distance, en particulier lorsque vous faites appel à des intégrations Step Functions AWS natives SDKs et compatibles afin de simplifier votre charge de travail.

Ressources

Bonnes pratiques associées :

- REL05-BP03 Contrôler et limiter les appels de nouvelle tentative
- REL05-BP04 Échouer rapidement et limiter les files d'attente
- REL06-BP07 Surveillez le end-to-end suivi des demandes via votre système

Documents connexes :

- AWS SDK: tentatives et délais d'expiration
- Amazon Builders' Library : délais d'attente, nouvelles tentatives et backoff avec instabilité
- Quotas Amazon API Gateway et remarques importantes
- AWS Command Line Interface : options de ligne de commande
- AWS SDK for Java 2.x: Configurer les API délais
- AWS Botocore utilisant l'objet de configuration et la référence de configuration
- AWS SDK for .NET : nouvelles tentatives et délais d'expiration
- AWS Lambda : configuration des options de fonction Lambda

Exemples connexes :

- Utilisation du schéma du disjoncteur avec AWS Step Functions Amazon DynamoDB
- Martin Fowler : CircuitBreaker

Outils associés :

- AWS SDKs
- AWS Lambda
- Amazon SQS
- AWS Step Functions
- AWS Command Line Interface

REL05-BP06 Rendre les systèmes apatrides dans la mesure du possible

Les systèmes ne doivent pas exiger d'état ou doivent décharger un état de telle sorte qu'entre les différentes demandes client, il n'y ait pas de dépendance vis-à-vis des données stockées localement sur disque et en mémoire. Cela permet le remplacement à volonté des serveurs sans impact sur la disponibilité.

Lorsque des utilisateurs ou des services interagissent avec une application, ils exécutent souvent une série d'interactions qui forment une session. Une session correspond aux données uniques des utilisateurs qui persistent entre les requêtes pendant l'utilisation de l'application. Une application sans état n'a pas besoin de connaître les interactions précédentes et ne stocke pas d'informations de session.

Une fois conçu pour être apatride, vous pouvez ensuite utiliser des services de calcul sans serveur, tels que AWS Lambda ou. AWS Fargate

Outre le remplacement des serveurs, les applications apatrides présentent un autre avantage : elles peuvent évoluer horizontalement, car toutes les ressources informatiques disponibles (telles que les EC2 instances et les AWS Lambda fonctions) peuvent répondre à toutes les demandes.

Avantages du respect de cette bonne pratique : les systèmes conçus pour être sans état sont plus adaptables à la mise à l'échelle horizontale, ce qui permet d'ajouter ou de supprimer des capacités en fonction des fluctuations du trafic et de la demande. Ils sont également intrinsèquement résilients aux défaillances et offrent flexibilité et agilité dans le cadre du développement d'applications.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Rendre vos applications sans état. Les applications sans état permettent une mise à l'échelle horizontale et tolèrent la défaillance d'un nœud individuel. Analysez et comprenez les composants de votre application qui conservent leur état au sein de l'architecture. Cela vous permet d'évaluer l'impact potentiel de la transition vers une conception sans état. Une architecture sans état découple les données utilisateur et décharge les données de session. Cela permet de mettre à l'échelle chaque composant indépendamment pour répondre à des demandes variables de charge de travail et optimiser l'utilisation des ressources.

Étapes d'implémentation

- Identifiez et comprenez les composants avec état dans votre application.
- Découplez les données en séparant et en gérant les données utilisateur de la logique principale de l'application.
  - Amazon Cognito peut découpler les données utilisateur du code de l'application en utilisant des fonctionnalités telles que les groupes d'identités, les groupes d'utilisateurs et Amazon Cognito Sync.
  - Vous pouvez utiliser AWS Secrets Manager pour découpler les données utilisateur en stockant les secrets dans un emplacement sécurisé et centralisé. Cela signifie que le code de l'application n'a pas besoin de stocker de secrets, ce qui le rend plus sécurisé.
  - Envisagez d'utiliser Amazon S3 pour stocker des données volumineuses non structurées, telles que des images et des documents. Votre application peut récupérer ces données en cas de besoin, évitant ainsi d'avoir à les stocker en mémoire.
  - Utilisez Amazon DynamoDB pour stocker des informations telles que les profils utilisateur. Votre application peut interroger ces données en temps quasi réel.
- Déchargez les données de session dans une base de données, un cache ou des fichiers externes.
  - Amazon ElastiCache, Amazon DynamoDB, Amazon Elastic File System (Amazon) et EFS Amazon MemoryDB sont des exemples de services que vous pouvez utiliser pour AWS décharger des données de session.
- Concevez une architecture sans état après avoir identifié les données d'état et d'utilisateur qui doivent être conservées avec la solution de stockage de votre choix.

Ressources

Bonnes pratiques associées :

- REL11-BP03 Automatiser la guérison sur toutes les couches

Documents connexes :

- L'Amazon Builders' Library : éviter le basculement dans les systèmes distribués
- L'Amazon Builders' Library : éviter les retards de file d'attente insurmontables
- L'Amazon Builders' Library : défis et stratégies de mise en cache
- Bonnes pratiques pour Stateless Web Tier sur AWS

REL05-BP07 Mettre en œuvre des leviers de secours

Les leviers d'urgence sont des processus rapides qui peuvent réduire l'impact sur la disponibilité de votre charge de travail.

Les leviers d'urgence fonctionnent en désactivant, en limitant ou en modifiant le comportement des composants ou des dépendances à l'aide de mécanismes connus et testés. Ils permettent d'atténuer les perturbations de la charge de travail causées par l'épuisement des ressources dû à une augmentation inattendue de la demande et de réduire l'impact des défaillances des composants non stratégiques de votre charge de travail.

Résultat souhaité : en mettant en œuvre des leviers d'urgence, vous pouvez établir des processus dont le fonctionnement a été vérifié pour maintenir la disponibilité des composants essentiels de votre charge de travail. La charge de travail devrait se dégrader de manière appropriée et continuer à remplir ses fonctions stratégiques durant l'activation d'un levier d'urgence. Pour plus de détails sur la dégradation progressive, voir REL05-BP01 Implémenter la dégradation progressive pour transformer les dépendances strictes applicables en dépendances souples.

Anti-modèles courants :

- La défaillance des dépendances non stratégiques a un impact sur la disponibilité de votre charge de travail principale.
- Le comportement des composants stratégiques n'est pas testé ou vérifié lors d'une défaillance d'un composant non stratégique.
- Aucun critère clair et déterministe n'a été défini pour l'activation ou la désactivation d'un levier d'urgence.

Avantages du respect de cette bonne pratique : la mise en œuvre de leviers d'urgence peut améliorer la disponibilité des composants critiques de votre charge de travail en fournissant à vos résolveurs des processus établis pour répondre aux pics de demande inattendus ou aux défaillances liées à des dépendances non critiques.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

- Identifier les composants stratégiques de votre charge de travail.
- Concevoir et construire les composants stratégiques de votre charge de travail de manière à ce qu'ils résistent aux défaillances des composants non stratégiques.
- Effectuer des tests pour valider le comportement de vos composants stratégiques en cas de défaillance des composants non stratégiques.
- Définir et surveiller des métriques ou des déclencheurs pertinents pour lancer des procédures de levier d'urgence.
- Définir les procédures (manuelles ou automatisées) qui comprennent le levier d'urgence.

Étapes d'implémentation

- Identifier les composants stratégiques de votre charge de travail.
  - Chaque composant technique de votre charge de travail doit être associé à la fonction commerciale correspondante et classé comme stratégique ou non stratégique. Pour des exemples de fonctionnalités critiques et non critiques d'Amazon, consultez Any Day Can Be Prime Day: How Amazon.com Search Uses Chaos Engineering to Handle Over 84K Requests Per Second.
  - Il s'agit d'une décision à la fois technique et commerciale, qui varie en fonction de l'organisation et de la charge de travail.
- Concevoir et construire les composants stratégiques de votre charge de travail de manière à ce qu'ils résistent aux défaillances des composants non stratégiques.
  - Lors de l'analyse des dépendances, tenez compte de tous les modes de défaillance potentiels et vérifiez que vos mécanismes de levier d'urgence fournissent les fonctionnalités stratégiques aux composants en aval.
- Effectuer des tests pour valider le comportement de vos composants stratégiques pendant l'activation de vos leviers d'urgence.

  - Éviter les comportements bimodaux. Pour plus de détails, voir REL11-BP05 Utiliser la stabilité statique pour empêcher le comportement bimodal.
- Définir et surveiller des métriques pertinentes pour lancer des procédures de levier d'urgence.
  - La recherche des bonnes métriques à surveiller dépend de votre charge de travail. Parmi les métriques, citons la latence ou le nombre de demandes infructueuses à une dépendance.
- Définir les procédures (manuelles ou automatisées) qui comprennent le levier d'urgence.
  - Cela peut inclure des mécanismes tels que le délestage, les demandes de limitation ou la mise en œuvre d'une dégradation appropriée.

Ressources

Bonnes pratiques associées :

- REL05-BP01 Implémenter une dégradation progressive pour transformer les dépendances matérielles applicables en dépendances souples
- REL05-BP02 Demandes d'accélérateur
- REL11-BP05 Utiliser la stabilité statique pour empêcher le comportement bimodal

Documents connexes :

- Automatiser les déploiements sécurisés et sans intervention
- Any Day Can Be Prime Day: How Amazon.com Search Uses Chaos Engineering to Handle Over 84K Requests Per Second

Vidéos connexes :

- AWS re:Invent 2020 : fiabilité, cohérence et confiance grâce à l'immuabilité

Gestion des modifications

Questions

- FIA 6. Comment surveiller les ressources de charge de travail ?
- FIA 7. Comment concevoir votre charge de travail pour qu'elle s'adapte à l'évolution de la demande ?

- FIA 8. Comment mettre en œuvre des modifications ?

FIA 6. Comment surveiller les ressources de charge de travail ?

Les journaux et les métriques sont des outils puissants qui permettent de mieux comprendre l'état de santé de votre charge de travail. Vous pouvez configurer votre charge de travail pour qu'elle surveille les journaux et les métriques et envoie des notifications lorsque des seuils sont franchis ou que des événements importants se produisent. La surveillance permet à votre charge de travail de reconnaître quand des seuils de faibles performances sont franchis ou quand des défaillances se produisent, afin d'y répondre par une récupération automatique.

Bonnes pratiques

- REL06-BP01 Surveiller tous les composants pour la charge de travail (génération)
- REL06-BP02 Définir et calculer des métriques (agrégation)
- REL06-BP03 Envoyer des notifications (traitement en temps réel et alarme)
- REL06-BP04 Automatiser les réponses (traitement en temps réel et alarmes)
- REL06-BP05 Analyser les journaux
- REL06-BP06 Passer régulièrement en revue la portée et les métriques de surveillance
- REL06-BP07 Surveillez le end-to-end suivi des demandes via votre système

REL06-BP01 Surveiller tous les composants pour la charge de travail (génération)

Surveillez les composants de la charge de travail à l'aide d'Amazon CloudWatch ou d'outils tiers. Surveillez AWS les services avec le AWS Health tableau de bord.

Tous les composants de votre charge de travail doivent être surveillés, y compris le côté utilisateur, la logique métier et les niveaux de stockage. Au besoin, définissez des métriques clés, décrivez leur procédure d'extraction des journaux, puis spécifiez des seuils d'invocation pour les événements d'alarme correspondants. Assurez-vous que les métriques correspondent aux indicateurs de performance clés (KPIs) de votre charge de travail, et utilisez les métriques et les journaux pour identifier les signes avant-coureurs d'une dégradation du service. Par exemple, un indicateur lié aux résultats commerciaux, tel que le nombre de commandes traitées avec succès par minute, peut indiquer les problèmes de charge de travail plus rapidement qu'un indicateur technique, tel que CPU l'utilisation. Utilisez le AWS Health tableau de bord pour obtenir une vue personnalisée des performances et de la disponibilité des AWS services sous-jacents à vos AWS ressources.

La surveillance dans le cloud offre de nouvelles opportunités. La plupart des fournisseurs de cloud ont développé des hooks personnalisables et peuvent fournir des informations pour vous aider à surveiller plusieurs niveaux de votre charge de travail. AWS des services tels qu'Amazon CloudWatch appliquent des algorithmes statistiques et d'apprentissage automatique pour analyser en permanence les métriques des systèmes et des applications, déterminer des bases de référence normales et détecter des anomalies avec une intervention minimale de l'utilisateur. Les algorithmes de détection des anomalies tiennent compte de la saisonnalité et des changements de tendance des métriques.

AWS met à disposition une multitude d'informations de surveillance et de journalisation destinées à la consommation, qui peuvent être utilisées pour définir des mesures et des change-in-demand processus spécifiques à la charge de travail et adopter des techniques d'apprentissage automatique, quelle que soit l'expertise en machine learning.

En outre, surveillez l’ensemble de vos points de terminaison externes afin de vous assurer qu’ils sont indépendants de votre implémentation de base. Cette surveillance active peut être effectuée avec des transactions synthétiques (parfois appelées Canary utilisateurs à ne pas confondre avec les déploiements Canary) qui exécutent régulièrement un certain nombre de tâches courantes effectuées par les clients de la charge de travail. Maintenez ces tâches de courte durée et veillez à ne pas surcharger votre charge de travail pendant les tests. Amazon CloudWatch Synthetics vous permet de créer des canaris synthétiques pour surveiller vos points de terminaison et. APIs Vous pouvez également combiner les nœuds de clients synthétiques Canary avec la console AWS X-Ray pour identifier les scripts Canary synthétiques qui rencontrent des erreurs, des pannes ou des taux de limitation au cours de la période sélectionnée.

Résultat souhaité :

Collectez et utilisez des métriques critiques de tous les composants de la charge de travail pour garantir la fiabilité de la charge de travail et une expérience utilisateur optimale. Détecter qu’une charge de travail n’atteint pas les résultats vous permet de déclarer rapidement un sinistre et de vous remettre d’un incident.

Anti-modèles courants :

- Surveillance limitée aux interfaces externes de votre charge de travail.
- Ne pas générer de métriques spécifiques à la charge de travail et se fier uniquement aux métriques qui vous sont fournies par les AWS services utilisés par votre charge de travail.
- N'utilisez que des indicateurs techniques dans votre charge de travail et ne surveillez aucune métrique liée aux paramètres non techniques auxquels KPIs la charge de travail contribue.

- S’appuyer sur le trafic de production et de simples surveillances de l’état pour surveiller et évaluer l’état de la charge de travail.

Avantages du respect de cette bonne pratique : la surveillance à tous les niveaux de votre charge de travail vous permet d’anticiper et de résoudre plus rapidement les problèmes dans les composants qui constituent la charge de travail.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

1. Activez la journalisation lorsqu’elle est disponible. Les données de surveillance doivent être obtenues à partir de tous les composants des charges de travail. Activez la journalisation supplémentaire, telle que les journaux d’accès S3, et autorisez votre charge de travail à consigner des données qui lui sont spécifiques. Collectez des métriques relatives aux CPU E/S réseau et aux moyennes des E/S sur disque auprès de services tels qu'Amazon, ECS Amazon, EKS AmazonEC2, Elastic Load Balancing AWS Auto Scaling et Amazon. EMR Consultez la section AWS Services qui publient CloudWatch des métriques pour obtenir la liste des AWS services sur lesquels vous publiez des métriques CloudWatch.
2. Passez en revue toutes les métriques par défaut et explorez toutes les lacunes de collecte de données. Chaque service génère des métriques par défaut. La collecte des métriques par défaut vous permet de mieux comprendre les dépendances entre les composants de charge de travail et sur la manière dont la fiabilité et les performances des composants affectent la charge de travail. Vous pouvez également créer et publier vos propres statistiques à CloudWatch l'aide du AWS CLI ou d'unAPI.
3. Évaluez tous les indicateurs afin de déterminer ceux sur lesquels vous souhaitez émettre une alerte pour chaque AWS service de votre charge de travail. Vous pouvez choisir de sélectionner un sous-ensemble de métriques qui ont un impact majeur sur la fiabilité de la charge de travail. En vous concentrant sur les métriques et les seuils critiques, vous pouvez affiner le nombre d’alertes et réduire le nombre de faux positifs.
4. Définissez des alertes et le processus de récupération de votre charge de travail après l’invocation de l’alerte. La définition d'alertes vous permet de notifier, d'escalader et de suivre rapidement les étapes nécessaires pour vous remettre d'un incident et atteindre l'objectif de temps de rétablissement prescrit (RTO). Vous pouvez utiliser Amazon CloudWatch Alarms pour appeler des flux de travail automatisés et lancer des procédures de restauration en fonction de seuils définis.
5. Explorez l’utilisation de transactions synthétiques pour collecter des données pertinentes sur l’état des charges de travail. La surveillance synthétique suit les mêmes routes et effectue les mêmes

actions qu'un client, ce qui vous permet de vérifier en permanence l'expérience client même lorsque vous n'avez aucun trafic client sur vos charges de travail. En utilisant les transactions synthétiques, vous pouvez découvrir les problèmes avant vos clients.

Ressources

Bonnes pratiques associées :

- REL11-BP03 Automatiser la guérison sur toutes les couches

Documents connexes :

- Commencer à utiliser votre AWS Health tableau de bord — État de santé de votre compte
- AWS Services qui publient CloudWatch des métriques
- Journaux d'accès de votre Network Load Balancer
- Journaux d'accès pour votre Application Load Balancer
- Accès à Amazon CloudWatch Logs pour AWS Lambda
- Journalisation des accès au serveur Amazon S3
- Activation des journaux d'accès pour votre Classic Load Balancer
- Exporter les données du journal vers Amazon S3
- Installation de l' CloudWatch agent sur une EC2 instance Amazon
- Publication des métriques personnalisées
- Utilisation des tableaux de CloudWatch bord Amazon
- Utilisation d'Amazon CloudWatch Metrics
- Utilisation des Canaries (Amazon CloudWatch Synthetics)
- Qu'est-ce qu'Amazon CloudWatch Logs ?

Guides de l'utilisateur :

- Création d'un journal de suivi
- Surveillance des métriques relatives à la mémoire et au disque pour les instances Amazon EC2 Linux
- Utilisation CloudWatch des journaux avec des instances de conteneur
- Journaux de flux VPC
- Qu'est-ce qu'Amazon DevOps Guru ?

- Qu'est-ce que c'est AWS X-Ray ?

Blogs connexes :

- Débogage avec Amazon CloudWatch Synthetics et AWS X-Ray

Exemples et ateliers connexes :

- Ateliers AWS Well-Architected : Excellence opérationnelle - Surveillance des dépendances
- Amazon Builders' Library : Instrumentation des systèmes distribués au profit de la visibilité opérationnelle
- Atelier sur l'observabilité

REL06-BP02 Définir et calculer des métriques (agrégation)

Collectez les métriques et les journaux des composants de votre charge de travail, et calculez des métriques agrégées pertinentes à partir de ces derniers. Ces métriques permettent une observabilité large et approfondie de votre charge de travail et peuvent améliorer de manière significative votre posture de résilience.

L'observabilité ne se limite pas à collecter des métriques à partir des composants de votre charge de travail et à être en mesure de les visualiser et d'émettre des alertes à leur sujet. Elle permet de bénéficier d'une compréhension globale du comportement de votre charge de travail. Ces informations comportementales proviennent de tous les composants de vos charges de travail, notamment des services cloud dont elles dépendent, des journaux bien conçus et des métriques. Ces données vous permettent de surveiller le comportement de votre charge de travail dans son ensemble et de comprendre de manière détaillée l'interaction de chaque composant avec chaque unité de travail.

Résultat escompté :

- Vous collectez les journaux des composants de votre charge de travail et des dépendances des services AWS, et vous les publiez dans un emplacement central où ils peuvent être facilement consultés et traités.
- Vos journaux contiennent des horodatages fidèles et précis.
- Vos journaux contiennent des informations pertinentes sur le contexte du traitement, telles qu'un identifiant de suivi, un identifiant d'utilisateur ou de compte et une adresse IP distante.

- Vous créez des métriques agrégées à partir de vos journaux qui représentent le comportement de votre charge de travail d'un point de vue général.
- Vous pouvez interroger vos journaux agrégés pour obtenir des informations approfondies et pertinentes sur votre charge de travail et identifier les problèmes réels et potentiels.

Anti-modèles courants :

- Vous ne collectez pas de métriques ou de journaux pertinents à partir des instances de calcul sur lesquelles vos charges de travail s'exécutent ou des services cloud qu'elles utilisent.
- Vous négligez la collecte des journaux et métriques liés à vos indicateurs de performances clés (KPI) métier.
- Vous analysez la télémétrie liée à la charge de travail de manière isolée, sans agrégation ni corrélation.
- Vous laissez les métriques et les journaux expirer trop rapidement, ce qui entrave l'analyse des tendances et l'identification des problèmes récurrents.

Avantages du respect de ces bonnes pratiques : vous pouvez détecter davantage d'anomalies et corréler les événements et les métriques entre les différents composants de votre charge de travail. Vous pouvez créer des informations exploitables à partir des composants de votre charge de travail en vous basant sur les informations contenues dans les journaux, qui ne sont souvent pas disponibles dans les métriques seules. Vous pouvez déterminer plus rapidement les causes des défaillances en interrogeant vos journaux à grande échelle.

Niveau d'exposition au risque si ces bonnes pratiques ne sont pas respectées : élevé

Directives d'implémentation

Identifiez les sources des données de télémétrie pertinentes pour vos charges de travail et leurs composants. Ces données proviennent non seulement des composants qui publient des métriques, tels que votre système d'exploitation (OS) et les environnements d'exécution d'applications tels que Java, mais également des journaux des applications et des services cloud. Par exemple, les serveurs Web enregistrent généralement chaque demande avec des informations détaillées telles que l'horodatage, la latence de traitement, l'ID utilisateur, l'adresse IP distante, le chemin et la chaîne de requête. Le niveau de détail de ces journaux vous permet d'effectuer des requêtes détaillées et de générer des métriques qui n'auraient peut-être pas été disponibles autrement.

Collectez les métriques et les journaux à l'aide d'outils et de processus appropriés. Les journaux générés par les applications exécutées sur une instance Amazon EC2 peuvent être collectés par un agent tel que l'agent Amazon CloudWatch et publiés dans un service de stockage central tel qu'Amazon CloudWatch Logs. Les services de calcul gérés par AWS tels qu'AWS Lambda et Amazon Elastic Container Service publient automatiquement les journaux dans CloudWatch Logs pour vous. Activez la collecte de journaux pour les services de stockage et de traitement AWS utilisés par vos charges de travail, tels qu'Amazon CloudFront, Amazon S3, Elastic Load Balancing et Amazon API Gateway.

Enrichissez vos données de télémétrie avec des dimensions qui peuvent vous aider à mieux identifier les modèles comportementaux et à isoler les problèmes corrélés à des groupes de composants associés. Une fois ces dimensions ajoutées, vous pouvez observer le comportement des composants de manière plus détaillée, détecter les défaillances corrélées et prendre les mesures correctives appropriées. La zone de disponibilité, l'ID d'instance EC2 et l'ID de pod ou de tâche de conteneur sont des exemples de dimensions utiles.

Une fois que vous avez collecté les métriques et les journaux, vous pouvez rédiger des requêtes et générer des métriques agrégées à partir de ces éléments pour fournir des informations exploitables utiles sur les comportements normaux et anormaux. Par exemple, vous pouvez utiliser Amazon CloudWatch Logs Insights pour dériver des métriques personnalisées des journaux de vos applications, Amazon CloudWatch Metrics Insights pour interroger vos métriques à grande échelle, Amazon CloudWatch Container Insights pour collecter, agréger et résumer les métriques et les journaux de vos applications et microservices conteneurisés, ou Amazon CloudWatch Lambda Insights si vous utilisez des fonctions AWS Lambda. Pour créer une métrique de taux d'erreur agrégé, vous pouvez incrémenter un compteur chaque fois qu'une réponse ou un message d'erreur est détecté dans les journaux de vos composants ou calculer la valeur agrégée d'une métrique de taux d'erreur existante. Vous pouvez utiliser ces données pour générer des histogrammes illustrant le comportement de queue, par exemple les demandes ou les processus les moins performants. Vous pouvez également analyser ces données en temps réel pour détecter des modèles anormaux à l'aide de solutions telles que la détection des anomalies de CloudWatch Logs. Il est possible de placer ces informations exploitables dans des tableaux de bord afin de les organiser en fonction de vos besoins et préférences.

L'interrogation des journaux peut vous aider à comprendre comment des demandes spécifiques ont été traitées par les composants de votre charge de travail et à révéler les modèles de demandes ou tout autre contexte ayant un impact sur la résilience de votre charge de travail. Il peut être utile d'étudier et de préparer des requêtes à l'avance, en fonction de votre connaissance des comportements de vos applications et des autres composants, afin de pouvoir les utiliser plus

facilement en cas de besoin. Par exemple, CloudWatch Logs Insights vous permet de rechercher et d'analyser de façon interactive les données de vos journaux dans CloudWatch Logs. Vous pouvez également utiliser Amazon Athena pour interroger les journaux provenant de plusieurs sources, y compris de nombreux services AWS, à l'échelle du pétaoctet.

Lorsque vous définissez une politique de conservation des journaux, tenez compte de la valeur des journaux d'historique. Les journaux d'historique peuvent aider à identifier les modèles d'utilisation et comportementaux à long terme, les régressions et les améliorations des performances de votre charge de travail. Les journaux définitivement supprimés ne peuvent pas être analysés ultérieurement. Toutefois, la valeur des journaux d'historique tend à diminuer sur de longues périodes. Choisissez une politique capable d'équilibrer vos besoins de manière appropriée et conforme à toutes les exigences légales ou contractuelles auxquelles vous pourriez être soumis.

Étapes d'implémentation

1. Choisissez des mécanismes de collecte, de stockage, d'analyse et d'affichage de vos données d'observabilité.
2. Installez et configurez des collecteurs de métriques et de journaux sur les composants appropriés de votre charge de travail (par exemple, sur les instances Amazon EC2 et dans les conteneurs sidecar). Configurez ces collecteurs pour qu'ils redémarrent automatiquement s'ils s'arrêtent de façon inattendue. Activez la mise en mémoire tampon du disque ou de la mémoire pour les collecteurs afin que les échecs de publication temporaires n'aient pas d'impact sur vos applications ou n'entraînent aucune perte de données.
3. Activez la journalisation sur les services AWS que vous utilisez dans le cadre de vos charges de travail et transférez ces journaux au service de stockage que vous avez sélectionné si nécessaire. Reportez-vous au guide d'utilisateur ou au manuel du développeur des services respectifs pour obtenir des instructions détaillées.
4. Définissez les métriques opérationnelles pour vos charges de travail en fonction de vos données de télémétrie. Elles peuvent être basées sur des métriques directes émises par les composants de votre charge de travail, qui peuvent inclure des métriques liées aux KPI métier, ou sur les résultats de calculs agrégés tels que des sommes, des taux, des centiles ou des histogrammes. Calculez ces métriques à l'aide de votre analyseur de journaux et placez-les dans des tableaux de bord, de façon appropriée.
5. Préparez des requêtes de journaux appropriées pour analyser les composants de la charge de travail, les demandes ou le comportement des transactions selon les besoins.

6. Définissez et activez une politique de conservation des journaux pour les journaux de vos composants. Supprimez régulièrement les journaux lorsqu'ils sont plus anciens que la politique ne l'autorise.

Ressources

Bonnes pratiques associées :

- REL06-BP01 Surveiller tous les composants de la charge de travail (génération)
- REL06-BP03 Envoyer des notifications (traitement et alarmes en temps réel)
- REL06-BP04 Automatiser les réponses (traitement et alarmes en temps réel)
- REL06-BP05 Analyser les journaux
- REL06-BP06 Passer régulièrement en revue la portée et les métriques de surveillance
- REL06-BP07 Surveiller le suivi de bout en bout des demandes via votre système

Documentation connexe :

- Fonctionnement d'Amazon CloudWatch
- Amazon Managed Prometheus
- Amazon Managed Grafana
- Analyse des données de journaux avec CloudWatch Logs Insights
- Amazon CloudWatch Lambda Insights
- Amazon CloudWatch Container Insights
- Interrogation de vos métriques avec CloudWatch Metrics Insights
- Distro pour Open Telemetry AWS
- Exemples de requêtes pour Amazon CloudWatch Logs Insights
- Débogage avec Amazon CloudWatch Synthetics et AWS X-Ray
- Recherche et filtrage des données de journaux
- Envoi des journaux directement à Amazon S3
- Amazon Builders' Library : Instrumentation des systèmes distribués au profit de la visibilité opérationnelle

Ateliers connexes :

- Un atelier sur l'observabilité

Outils associés :

- AWS Distro for OpenTelemetry (GitHub)

REL06-BP03 Envoyer des notifications (traitement en temps réel et alarme)

Lorsque les organisations détectent des problèmes potentiels, elles envoient des notifications et des alertes en temps réel au personnel et aux systèmes appropriés afin de résoudre rapidement et efficacement ces problèmes.

Résultat souhaité : il est possible d'obtenir des réponses rapides aux événements opérationnels en configurant des alarmes pertinentes basées sur les métriques de service et d'application. Lorsque les seuils d'alarme sont dépassés, le personnel et les systèmes appropriés sont avertis afin de résoudre les problèmes sous-jacents.

Anti-modèles courants :

- Vous configurez les alarmes avec un seuil trop élevé, ce qui fait échouer l'envoi des notifications vitales.
- Vous configurez les alarmes avec un seuil trop bas, ce qui empêche la prise en compte des alertes importantes à cause du bruit généré par un trop grand nombre de notifications.
- Vous ne mettez pas à jour les alarmes et leur seuil en cas de changement d'utilisation.
- Pour les alarmes qu'il est préférable de traiter par des actions automatisées, vous envoyez la notification au personnel au lieu de générer l'action automatisée, ce qui entraîne l'envoi d'un trop grand nombre de notifications.

Avantages du respect de cette bonne pratique : en envoyant des notifications et des alertes en temps réel au personnel et aux systèmes appropriés, vous pouvez détecter rapidement les problèmes et réagir rapidement face aux incidents opérationnels.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les charges de travail doivent être équipées d'un système de traitement et d'avertissement en temps réel afin d'améliorer la détectabilité des problèmes susceptibles d'affecter la disponibilité de

l'application et de déclencher une réponse automatique. Les organisations peuvent procéder au traitement et à l'avertissement en temps réel en créant des alertes avec des métriques définies afin de recevoir des notifications chaque fois que des événements importants se produisent ou qu'une métrique dépasse un seuil.

Amazon vous CloudWatch permet de créer des alarmes métriques et composites à l'aide d' CloudWatch alarmes basées sur un seuil statique, la détection d'anomalies et d'autres critères. Pour plus de détails sur les types d'alarmes que vous pouvez configurer à l'aide CloudWatch, consultez la section sur les alarmes de la CloudWatch documentation.

Vous pouvez créer des vues personnalisées des indicateurs et des alertes concernant vos AWS ressources pour vos équipes à l'aide de CloudWatch tableaux de bord. Les pages d'accueil personnalisables de la CloudWatch console vous permettent de surveiller vos ressources dans une vue unique dans plusieurs régions.

Les alarmes peuvent effectuer une ou plusieurs actions, telles que l'envoi d'une notification à un SNSsujet Amazon, l'exécution d'une EC2 action Amazon ou d'une action Amazon EC2 Auto Scaling, ou la création d'un incident OpsItem ou dans AWS Systems Manager.

Amazon CloudWatch utilise Amazon SNS pour envoyer des notifications lorsque l'alarme change d'état, assurant ainsi la transmission des messages des éditeurs (producteurs) aux abonnés (consommateurs). Pour plus de détails sur la configuration SNS des notifications Amazon, consultez Configuration d'Amazon SNS.

CloudWatch envoie EventBridgedes événements chaque fois qu'une CloudWatch alarme est créée, mise à jour, supprimée ou que son état change. Vous pouvez utiliser ces événements pour créer EventBridge des règles qui exécutent des actions, telles que vous avertir chaque fois que l'état d'une alarme change ou déclencher automatiquement des événements dans votre compte à l'aide de l'automatisation de Systems Manager.

Quand devriez-vous utiliser Amazon EventBridge ou Amazon SNS ?

Amazon EventBridge et les deux SNS peuvent être utilisés pour développer des applications pilotées par des événements, et votre choix dépendra de vos besoins spécifiques.

Amazon EventBridge est recommandé lorsque vous souhaitez créer une application qui réagit aux événements de vos propres applications, applications SaaS et AWS services. EventBridge est le seul service basé sur des événements qui s'intègre directement à des partenaires SaaS tiers. EventBridge intègre également automatiquement les événements de plus de 200 AWS services sans que les développeurs n'aient à créer de ressources dans leur compte.

EventBridge utilise une structure définie JSON pour les événements et vous aide à créer des règles qui sont appliquées à l'ensemble du corps de l'événement afin de sélectionner les événements à transférer à une cible. EventBridge prend actuellement en charge plus de 20 AWS services en tant que cibles AWS Lambda, notamment Amazon SQS, AmazonSNS, Amazon Kinesis Data Streams et Amazon Data Firehose.

Amazon SNS est recommandé pour les applications nécessitant une ventilation élevée (des milliers ou des millions de points de terminaison). Une tendance courante que nous observons est que les clients utilisent Amazon SNS comme cible pour leur règle afin de filtrer les événements dont ils ont besoin et de les diffuser sur plusieurs points de terminaison.

Les messages ne sont pas structurés et peuvent être de n'importe quel format. Amazon SNS prend en charge le transfert de messages vers six types de cibles différents, notamment LambdaSQS, Amazon, HTTP /S endpointsSMS, mobile push et e-mail. La latence SNS typique d'Amazon est inférieure à 30 millisecondes. Un large éventail de AWS services envoient SNS des messages Amazon en configurant le service à cette fin (plus de 30, dont AmazonEC2, Amazon S3 et Amazon RDS).

Étapes d'implémentation

1. Créez une alarme à l'aide des CloudWatch alarmes Amazon.
   a. Une alarme métrique surveille une seule CloudWatch métrique ou une expression en fonction CloudWatch des métriques. L'alarme déclenche une ou plusieurs actions en fonction de la valeur de la métrique ou de l'expression par rapport à un seuil sur un certain nombre d'intervalles de temps. L'action peut consister à envoyer une notification à un SNSsujet Amazon, à effectuer une EC2 action Amazon ou une action Amazon EC2 Auto Scaling, ou à créer un incident OpsItem ou dans AWS Systems Manager.
   b. Une alarme composite est une expression de règle qui prend en compte les conditions d'alarme des autres alarmes que vous avez créées. L'alarme composite ne passe en état d'alarme que si toutes les conditions de la règle sont satisfaites. Les alarmes spécifiées dans l'expression de règle d'une alarme composite peuvent inclure des alarmes de métrique et des alarmes composites supplémentaires. Les alarmes composites peuvent envoyer SNS des notifications à Amazon lorsque leur état change et peuvent créer des Systems Manager OpsItemsou des incidents lorsqu'elles entrent dans l'état d'alarme, mais elles ne peuvent pas effectuer d'actions Amazon EC2 ou Auto Scaling.
2. Configurez SNSles notifications Amazon. Lorsque vous créez une CloudWatch alarme, vous pouvez inclure une SNS rubrique Amazon pour envoyer une notification lorsque l'alarme change d'état.

3. Créez des règles EventBridge qui correspondent aux CloudWatch alarmes spécifiées. Chaque règle prend en charge plusieurs cibles, y compris des fonctions Lambda. Par exemple, vous pouvez définir une alarme qui se déclenche lorsque l'espace disque disponible est insuffisant, ce qui déclenche une fonction Lambda par le biais EventBridge d'une règle, afin de nettoyer l'espace. Pour plus de détails sur EventBridge les cibles, voir EventBridge cibles.

Ressources

Bonnes pratiques Well-Architected connexes :

- REL06-BP01 Surveiller tous les composants pour la charge de travail (génération)
- REL06-BP02 Définir et calculer des métriques (agrégation)
- REL12-BP01 Utiliser des playbooks pour enquêter sur les causes des défaillances

Documents connexes :

- Amazon CloudWatch
- CloudWatch Informations sur les journaux
- Utilisation des CloudWatch alarmes Amazon
- Utilisation des tableaux de CloudWatch bord Amazon
- Utilisation des CloudWatch métriques Amazon
- Configuration des SNS notifications Amazon
- CloudWatch détection d'anomalies
- CloudWatch Protection des données des journaux
- Amazon EventBridge
- Amazon Simple Notification Service

Vidéos connexes :

- reinvent 2022 observability videos
- AWS re:Invent 2022 - Meilleures pratiques en matière d'observabilité sur Amazon

Exemples associés :

- Un atelier sur l’observabilité

- Amazon va EventBridge utiliser le contrôle AWS Lambda des commentaires par Amazon CloudWatch Alarms

REL06-BP04 Automatiser les réponses (traitement en temps réel et alarmes)

Utilisez l'automatisation pour agir en cas de détection d'événement, par exemple, pour remplacer les composants défectueux.

Un traitement automatique en temps réel des alarmes est mis en œuvre afin que les systèmes puissent prendre rapidement des mesures correctives et tenter d'éviter les pannes ou une dégradation du service lorsque les alarmes se déclenchent. Les réponses automatisées aux alarmes peuvent inclure le remplacement des composants défaillants, l'ajustement de la capacité de calcul, la redirection du trafic vers des hôtes, des zones de disponibilité ou d'autres régions en bonne santé, et la notification des opérateurs.

Résultat souhaité : les alarmes en temps réel sont identifiées et le traitement automatique des alarmes est configuré pour déclencher les actions appropriées prises pour maintenir les objectifs de niveau de service et les accords de niveau de service ()SLAs. L'automatisation peut aller de l'autoréparation de composants individuels au basculement complet du site.

Anti-modèles courants :

- Pas d'inventaire ou de catalogue clair des principales alarmes en temps réel.
- Aucune réponse automatique aux alarmes critiques (par exemple, lorsque la capacité de calcul est presque épuisée, une mise à l'échelle automatique se produit).
- Réponses aux alarmes contradictoires.
- Aucune procédure opérationnelle standard (SOPs) à suivre par les opérateurs lorsqu'ils reçoivent des notifications d'alerte.
- Pas de surveillance des modifications de configuration, alors que des changements de configuration non détectés peuvent entraîner des temps d'arrêt pour les charges de travail.
- Pas de stratégie pour annuler les modifications de configuration involontaires.

Avantages du respect de cette bonne pratique : l'automatisation du traitement des alarmes peut améliorer la résilience du système. Le système prend automatiquement des mesures correctives, réduisant ainsi les activités manuelles qui nécessitent des interventions humaines sujettes aux erreurs. L'exécution de la charge de travail permet d'atteindre les objectifs de disponibilité et de réduire les interruptions de service.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Pour gérer efficacement les alertes et automatiser leur réponse, classez les alertes en fonction de leur criticité et de leur impact, documentez les procédures de réponse et planifiez les réponses avant de classer les tâches.

Identifiez les tâches nécessitant des actions spécifiques (souvent détaillées dans les runbooks) et examinez tous les runbooks et playbooks pour déterminer les tâches qui peuvent être automatisées. Si les actions peuvent être définies, alors elles sont souvent automatisables. Si les actions ne peuvent pas être automatisées, documentez les étapes manuelles SOP et formez les opérateurs à ces étapes. Remettez continuellement en question les processus manuels pour trouver des opportunités d’automatisation où vous pouvez établir et maintenir un plan d’automatisation des réponses aux alertes.

Étapes d’implémentation

1. Créez un inventaire des alarmes : pour obtenir une liste de toutes les alarmes, vous pouvez AWS CLIutiliser la CloudWatch commande Amazon`describe-alarms`. Selon le nombre d'alarmes que vous avez configurées, vous devrez peut-être utiliser la pagination pour récupérer un sous-ensemble d'alarmes pour chaque appel, ou vous pouvez utiliser le pour obtenir les alarmes AWS SDK à l'aide d'un API appel.
2. Documentez toutes les actions d’alarme : mettez à jour un runbook avec toutes les alarmes et leurs actions, qu’elles soient manuelles ou automatisées. AWS Systems Manager fournit des runbooks prédéfinis. Pour plus d’informations sur les runbooks, consultez Travailler avec des runbooks. Pour plus de détails sur la façon d’afficher le contenu du runbook, consultez Afficher le contenu du runbook.
3. Configurer et gérer les actions d'alarme : pour toutes les alarmes nécessitant une action, spécifiez l'action automatisée à l'aide du CloudWatch SDK. Par exemple, vous pouvez modifier automatiquement l'état de vos EC2 instances Amazon en fonction CloudWatch d'une alarme en créant et en activant des actions sur une alarme ou en désactivant des actions sur une alarme.

   Vous pouvez également utiliser Amazon EventBridge pour répondre automatiquement aux événements du système, tels que les problèmes de disponibilité des applications ou les modifications des ressources. Vous pouvez créer des règles pour indiquer quels événements vous intéressent et les actions à effectuer quand un événement correspond à une règle. Les actions qui peuvent être initiées automatiquement incluent l'appel d'une AWS Lambdafonction, l'appel

d'Amazon EC2Run Command, le transfert de l'événement à Amazon Kinesis Data Streams et l'utilisation d'Automate Amazon. EC2 EventBridge

4. Procédures opérationnelles standard (SOPs) : en fonction des composants de votre application, AWS Resilience Hubrecommande plusieurs SOPmodèles. Vous pouvez les utiliser SOPs pour documenter tous les processus qu'un opérateur doit suivre en cas d'alerte. Vous pouvez également créer une application SOP basée sur les recommandations du Resilience Hub, dans laquelle vous avez besoin d'une application Resilience Hub associée à une politique de résilience, ainsi que d'une évaluation historique de la résilience par rapport à cette application. Les recommandations qui vous SOP sont adressées sont issues de l'évaluation de la résilience.

   Resilience Hub travaille avec Systems Manager pour automatiser vos étapes SOPs en fournissant un certain nombre de SSMdocuments que vous pouvez utiliser comme base pour celles-ciSOPs. Par exemple, Resilience Hub peut recommander un SOP pour ajouter de l'espace disque sur la base d'un document SSM d'automatisation existant.
5. Réalisez des actions automatisées à l'aide d'Amazon DevOps Guru : vous pouvez utiliser Amazon DevOps Guru pour surveiller automatiquement les ressources de l'application afin de détecter tout comportement anormal et de fournir des recommandations ciblées afin d'accélérer l'identification des problèmes et les délais de résolution. Avec DevOps Guru, vous pouvez surveiller les flux de données opérationnelles en temps quasi réel à partir de plusieurs sources, notamment Amazon CloudWatch Metrics AWS Config, AWS CloudFormation, et AWS X-Ray. Vous pouvez également utiliser DevOps Guru pour créer OpsItems OpsCenter et envoyer automatiquement des événements à des EventBridge fins d'automatisation supplémentaire.

Ressources

Bonnes pratiques associées :

- REL06-BP01 Surveiller tous les composants pour la charge de travail (génération)
- REL06-BP02 Définir et calculer des métriques (agrégation)
- REL06-BP03 Envoyer des notifications (traitement en temps réel et alarme)
- REL08-BP01 Utiliser des runbooks pour les activités standard telles que le déploiement

Documents connexes :

- AWS Systems Manager Automation
- Création d'une EventBridge règle déclenchant un événement à partir d'une AWS ressource

- Un atelier sur l'observabilité
- Amazon Builders' Library : Instrumentation des systèmes distribués au profit de la visibilité opérationnelle
- Qu'est-ce qu'Amazon DevOps Guru ?
- Utilisation des documents d'automatisation (playbooks)

Vidéos connexes :

- AWS re:Invent 2022 - Meilleures pratiques en matière d'observabilité sur Amazon
- AWS re:Invent 2020 : automatisez tout avec AWS Systems Manager
- Présentation de AWS Resilience Hub
- Créez des systèmes de tickets personnalisés pour les notifications Amazon DevOps Guru
- Activez l'agrégation d'informations sur plusieurs comptes avec Amazon Guru DevOps

Exemples associés :

- Ateliers de fiabilité
- Atelier Amazon CloudWatch et Systems Manager

REL06-BP05 Analyser les journaux

Collectez les fichiers journaux et les historiques de métriques, puis analysez-les pour obtenir des informations plus générales sur les tendances et la charge de travail.

Amazon CloudWatch Logs Insights prend en charge un langage de requête simple mais puissant que vous pouvez utiliser pour analyser les données des journaux. Amazon CloudWatch Logs prend également en charge les abonnements qui permettent aux données de circuler facilement vers Amazon S3, où vous pouvez les utiliser, ou vers Amazon Athena pour les interroger. Il prend également en charge les requêtes dans une grande variétés de formats. Consultez la section Formats de données SerDes et formats pris en charge dans le guide de l'utilisateur d'Amazon Athena pour plus d'informations. Pour analyser d'énormes ensembles de fichiers journaux, vous pouvez exécuter un EMR cluster Amazon pour effectuer des analyses à l'échelle du pétaoctet.

Il existe un certain nombre d'outils fournis par des AWS partenaires et des tiers qui permettent l'agrégation, le traitement, le stockage et l'analyse. Ces outils incluent New Relic, Splunk, Loggly, Logstash et Nagios. CloudHealth Cependant, la génération en dehors du système et des journaux

d'applications est propre à chaque fournisseur de cloud, et généralement, spécifique à chaque service.

Un partie souvent négligée de la surveillance des processus concerne la gestion des données. Vous devez déterminer les exigences de rétention des données de surveillance, puis appliquer des stratégies de cycle de vie en conséquence. Amazon S3 prend en charge la gestion du cycle de vie au niveau du compartiment S3. Cette gestion du cycle de vie peut être appliquée différemment à d'autres chemins dans le compartiment. Vers la fin du cycle de vie, vous pouvez transférer des données dans Amazon S3 Glacier pour un stockage à long terme, puis les laisser expirer une fois la fin de la période de rétention terminée. La classe de stockage S3 Intelligent-Tiering est conçue pour optimiser les coûts en transférant automatiquement les données vers le niveau d'accès le plus économique, sans impact sur les performances ni surcharge opérationnelle.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

- CloudWatch Logs Insights vous permet de rechercher et d'analyser de manière interactive les données de vos CloudWatch journaux dans Amazon Logs.
  - Analyse des données de journal avec CloudWatch Logs Insights
  - Exemples de requêtes Amazon CloudWatch Logs Insights
- Utilisez Amazon CloudWatch Logs pour envoyer des journaux à Amazon S3 où vous pouvez les utiliser ou à Amazon Athena pour interroger les données.
  - Comment analyser les journaux d'accès au serveur Amazon S3 à l'aide d'Athena ?
    - Créez une stratégie de cycle de vie S3 pour votre compartiment de journaux d'accès au serveur. Configurez la stratégie de cycle de vie pour supprimer périodiquement les fichiers journaux. Cela permet de réduire la quantité de données analysées par Athena pour chaque requête.
      - Comment créer la stratégie de cycle de vie d'un compartiment S3 ?

Ressources

Documents connexes :

- Exemples de requêtes Amazon CloudWatch Logs Insights
- Analyse des données de journal avec CloudWatch Logs Insights
- Débogage avec Amazon CloudWatch Synthetics et AWS X-Ray

- Comment créer la stratégie de cycle de vie d'un compartiment S3 ?
- Comment analyser les journaux d'accès au serveur Amazon S3 à l'aide d'Athena ?
- Un atelier sur l'observabilité
- Amazon Builders' Library : Instrumentation des systèmes distribués au profit de la visibilité opérationnelle

REL06-BP06 Passer régulièrement en revue la portée et les métriques de surveillance

Passez fréquemment en revue la manière dont la surveillance de la charge de travail est mise en œuvre et mettez-la à jour au fur et à mesure que votre charge de travail et son architecture évoluent. Des audits réguliers de votre surveillance permettent de réduire le risque de négliger ou d'omettre des indicateurs de panne et contribuent à aider votre charge de travail à atteindre ses objectifs de disponibilité.

Un suivi efficace s'appuie sur des métriques métier clés, qui évoluent en fonction des priorités de votre entreprise. Votre processus d'examen de la surveillance doit mettre l'accent sur les indicateurs de niveau de service (SLI) et incorporer des informations exploitables provenant de votre infrastructure, de vos applications, de vos clients et de vos utilisateurs.

Résultat escompté : vous disposez d'une stratégie de surveillance efficace qui est régulièrement revue et mise à jour, ainsi qu'après tout événement ou changement important. Vous vérifiez que les indicateurs d'intégrité clés des applications restent pertinents au fur et à mesure de l'évolution de votre charge de travail et de vos exigences professionnelles.

Anti-modèles courants :

- Vous collectez uniquement les métriques par défaut.
- Vous configurez une stratégie de surveillance, mais vous ne la passez jamais en revue.
- Vous ne remettez pas en question la surveillance lorsque des modifications majeures sont déployées.
- Vous vous fiez à des métriques obsolètes pour déterminer l'état de la charge de travail.
- Vos équipes d'exploitation sont submergées d'alertes faussement positives en raison de métriques et de seuils obsolètes.
- Vous ne bénéficiez pas de l'observabilité des composants d'application qui ne sont pas surveillés.
- Vous vous concentrez uniquement sur des métriques techniques de bas niveau et excluez les métriques métier de votre surveillance.

Avantages liés au respect de cette bonne pratique : lorsque vous passez régulièrement en revue votre surveillance, vous pouvez anticiper les problèmes potentiels et vérifier que vous êtes capable de les détecter. Cela vous permet également de découvrir des zones d'ombre que vous auriez pu manquer lors d'examens antérieurs, ce qui améliore encore votre capacité à détecter les problèmes.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Passez en revue les métriques et la portée de la surveillance au cours de votre processus d'examen de l'état de préparation opérationnelle (ORR). Effectuez des examens périodiques de l'état de préparation opérationnelle selon un calendrier cohérent afin d'évaluer s'il existe des écarts entre votre charge de travail actuelle et la surveillance que vous avez configurée. Établissez une fréquence régulière d'examen des performances opérationnelles et de partage des connaissances afin d'améliorer votre capacité à obtenir de meilleures performances de la part de vos équipes d'exploitation. Confirmez ou non que les seuils d'alerte existants sont toujours adéquats et vérifiez les situations dans lesquelles les équipes d'exploitation reçoivent des alertes faussement positives ou ne surveillent pas les aspects de l'application qui devraient être surveillés.

Le cadre d'analyse de résilience fournit des conseils utiles qui peuvent vous aider à naviguer dans le processus. L'objectif de ce cadre est d'identifier les modes de défaillance potentiels et les contrôles préventifs et correctifs que vous pouvez utiliser pour atténuer leur impact. Ces connaissances peuvent vous aider à identifier les métriques et les événements appropriés à surveiller pour émettre des alertes.

Étapes d'implémentation

1. Planifiez et effectuez des vérifications régulières des tableaux de bord de charge de travail. Vous pouvez avoir des cadences différentes selon la profondeur à laquelle vous inspectez.
2. Inspectez les tendances dans les métriques. Comparez les valeurs des métriques aux valeurs historiques pour voir si des tendances peuvent indiquer que quelque chose doit faire l'objet d'une enquête. Cela peut être une augmentation de la latence, une diminution de la fonction principale de l'entreprise ou une augmentation des réponses aux échecs.
3. Recherchez des valeurs aberrantes et des anomalies dans vos métriques, qui peuvent être masquées par des moyennes ou des médianes. Observez les maximales et les minimales sur une période donnée et étudiez les causes des observations qui figurent loin des normales attendues. Au fur et à mesure que vous éliminez ces causes, vous pouvez resserrer les limites des métriques attendues en réponse à l'amélioration de la cohérence des performances de votre charge de travail.

4. Recherchez des changements importants de comportement. Un changement immédiat de quantité ou de direction d'une métrique peut indiquer une modification de l'application ou des facteurs externes, dont le suivi peut nécessiter l'ajout de métriques supplémentaires.
5. Vérifiez si la stratégie de surveillance actuelle reste pertinente pour l'application. Sur la base d'une analyse des incidents précédents (ou du cadre d'analyse de résilience), déterminez si d'autres aspects de l'application devraient être incorporés dans la portée de la surveillance.
6. Passez en revue vos métriques de surveillance des utilisateurs réels (RUM) pour déterminer s'il existe des lacunes dans la couverture des fonctionnalités de l'application.
7. Passez en revue votre processus de gestion des modifications. Mettez à jour vos procédures, si nécessaire, pour inclure une étape d'analyse de surveillance à effectuer avant d'approuver une modification.
8. Mettez en œuvre un examen de surveillance dans le cadre de votre examen de l'état de préparation opérationnelle et de vos processus de correction des erreurs.

Ressources

Bonnes pratiques associées

- REL06-BP01 Surveiller tous les composants de la charge de travail (génération)
- REL06-BP02 Définir et calculer des métriques (agrégation)
- REL06-BP07 Surveiller le suivi de bout en bout des demandes via votre système
- REL12-BP02 Effectuer une analyse post-incident
- REL12-BP06 Organiser régulièrement des tests de simulation de panne

Documents connexes :

- Pourquoi mettre en place la correction des erreurs (COE)
- Utilisation des tableaux de bord Amazon CloudWatch
- Création de tableaux de bord pour une visibilité opérationnelle
- Modèles de résilience multi-AZ avancés – Défaillances grises
- Exemples de requêtes pour Amazon CloudWatch Logs Insights
- Débogage avec Amazon CloudWatch Synthetics et AWS X-Ray
- Un atelier sur l'observabilité

- Amazon Builders' Library : Instrumentation des systèmes distribués au profit de la visibilité opérationnelle
- Utilisation des tableaux de bord Amazon CloudWatch
- Bonnes pratiques AWS en matière d'observabilité
- Cadre d'analyse de résilience
- Cadre d'analyse de résilience – Observabilité
- Examen de l'état de préparation opérationnelle – ORR

REL06-BP07 Surveillez le end-to-end suivi des demandes via votre système

Suivez les demandes au fur et à mesure qu'elles sont traitées dans les composants du service afin que les équipes produits puissent plus facilement analyser et résoudre les problèmes et améliorer les performances.

Résultat escompté : les charges de travail dotées d'un suivi complet de tous les composants sont faciles à déboguer, ce qui améliore le temps moyen de résolution (MTTR) des erreurs et la latence en simplifiant la découverte des causes premières. End-to-endle suivi réduit le temps nécessaire pour découvrir les composants concernés et analyser en détail les causes profondes des erreurs ou des temps de latence.

Anti-modèles courants :

- Le traçage est utilisé pour certains composants, mais pas pour tous. Par exemple, sans suivi, les AWS Lambdaéquipes risquent de ne pas comprendre clairement la latence causée par les démarrages à froid dans le cadre d'une charge de travail exigeante.
- Les canaris synthétiques ou la surveillance par utilisateur réel (RUM) ne sont pas configurés avec le traçage. Sans canarisRUM, la télémétrie des interactions avec le client est omise de l'analyse des traces, ce qui donne un profil de performance incomplet.
- Les charges de travail hybrides incluent à la fois des outils de suivi natifs du cloud et des outils tiers, mais aucune mesure n'a été prise pour intégrer pleinement une solution de traçage unique. Sur la base de la solution de suivi choisie, le suivi natif du cloud SDKs doit être utilisé pour instrumenter des composants qui ne sont pas natifs du cloud ou des outils tiers doivent être configurés pour ingérer la télémétrie de suivi native du cloud.

Avantages du respect de cette bonne pratique : lorsque les équipes de développement sont alertées de problèmes, elles peuvent obtenir une image complète des interactions entre les composants du

système, y compris la corrélation composant par composant avec la journalisation, les performances et les défaillances. Dans la mesure où le traçage permet d'identifier visuellement les causes profondes, vous passez moins de temps à les étudier. Les équipes qui comprennent en détail les interactions entre les composants prennent de meilleures décisions plus rapidement lors de la résolution des problèmes. L'analyse des traces des systèmes permet d'améliorer la prise de décisions, par exemple quand il convient d'invoquer le basculement de reprise après sinistre (DR) ou de choisir le meilleur endroit pour mettre en œuvre des stratégies d'autoréparation, ce qui permet d'améliorer la satisfaction des clients envers vos services.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les équipes qui exploitent des applications distribuées peuvent utiliser des outils de traçage pour établir un identifiant de corrélation, collecter des traces de demandes et créer des cartes de service pour les composants connectés. Tous les composants de l'application doivent être inclus dans les traces des demandes, notamment les clients de service, les passerelles d'intergiciels et les bus d'événements, les composants de calcul et le stockage, y compris les magasins de clés-valeurs et les bases de données. Intégrez des canaris synthétiques et une surveillance des utilisateurs réels dans votre configuration de end-to-end suivi afin de mesurer les interactions avec les clients distants et la latence afin d'évaluer avec précision les performances de vos systèmes par rapport à vos accords de niveau de service et à vos objectifs.

Vous pouvez utiliser AWS X-Rayles services d'instrumentation Amazon CloudWatch Application Monitoring pour fournir une vue complète des demandes au fur et à mesure qu'elles transitent par votre application. X-Ray collecte la télémétrie des applications et vous permet de la visualiser et de la filtrer en fonction des charges utiles, des fonctions, des traces, des servicesAPIs, et peut être activée pour les composants du système sans code ou à faible code. CloudWatch la surveillance des applications inclut ServiceLens l'intégration de vos traces aux métriques, aux journaux et aux alarmes. CloudWatch la surveillance des applications inclut également des produits synthétiques pour surveiller vos terminauxAPIs, ainsi que la surveillance des utilisateurs réels pour instrumenter vos clients d'applications Web.

Étapes d'implémentation

- AWS X-Ray À utiliser sur tous les services natifs pris en charge tels qu'Amazon S3 et Amazon API Gateway. AWS Lambda Ces AWS services permettent à X-Ray de changer de configuration en utilisant l'infrastructure sous forme de code AWS SDKs, ou le. AWS Management Console
- Applications instrumentales AWS Distro pour Open Telemetry et X-Ray ou agents de collecte tiers.

- Consultez le Guide du développeur AWS X-Ray pour une implémentation spécifique au langage de programmation. Ces sections de documentation expliquent comment instrumenter les HTTP demandes, SQL les requêtes et les autres processus spécifiques à votre langage de programmation d'applications.
- Utilisez le suivi X-Ray pour Amazon CloudWatch Synthetic Canaries et Amazon CloudWatch RUM afin d'analyser le chemin des demandes de votre client utilisateur final via votre AWS infrastructure en aval.
- Configurez CloudWatch les métriques et les alarmes en fonction de l'état des ressources et de la télémétrie Canary afin que les équipes soient rapidement alertées des problèmes, puis puissent étudier en profondeur les traces et les cartes des services avec. ServiceLens
- Activez l'intégration de X-Ray pour les outils de suivi tiers tels que Datadog, New Relic ou Dynatrace si vous utilisez des outils tiers pour votre solution de suivi principale.

Ressources

Bonnes pratiques associées :

- REL06-BP01 Surveiller tous les composants pour la charge de travail (génération)
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances

Documents connexes :

- Qu'est-ce que c'est AWS X-Ray ?
- Amazon CloudWatch : surveillance des applications
- Débogage avec Amazon CloudWatch Synthetics et AWS X-Ray
- Amazon Builders' Library : Instrumentation des systèmes distribués au profit de la visibilité opérationnelle
- Intégration AWS X-Ray à d'autres AWS services
- AWS Distro pour et OpenTelemetry AWS X-Ray
- Amazon CloudWatch : utilisation de la surveillance synthétique
- Amazon CloudWatch : Utilisation CloudWatch RUM
- Configurer Amazon CloudWatch Synthetics Canary et Amazon Alarm CloudWatch
- Disponibilité et au-delà : comprendre et améliorer la résilience des systèmes distribués sur AWS

Exemples connexes :

- Un atelier sur l’observabilité

Vidéos connexes :

- AWS re:Invent 2022 - Comment surveiller les applications sur plusieurs comptes
- Comment surveiller vos AWS applications

Outils associés :

- AWS X-Ray
- Amazon CloudWatch
- Amazon Route 53

FIA 7. Comment concevoir votre charge de travail pour qu’elle s’adapte à l’évolution de la demande ?

Une charge de travail évolutive permet d’ajouter ou de supprimer automatiquement des ressources de manière à ce qu’elles correspondent à la demande actuelle à un moment donné.

Bonnes pratiques

- REL07-BP01 Utiliser l’automatisation lors de l’obtention des ressources ou de leur mise à l’échelle
- REL07-BP02 Obtenir des ressources en cas de détection d'une altération de la charge de travail
- REL07-BP03 Obtenir des ressources après avoir réalisé qu’un plus grand nombre de ressources est nécessaire pour une charge de travail
- REL07-BP04 Testez votre charge de travail

REL07-BP01 Utiliser l’automatisation lors de l’obtention des ressources ou de leur mise à l’échelle

La définition programmatique, le provisionnement et la gestion de votre infrastructure et de vos ressources constituent la pierre angulaire de la fiabilité dans le cloud. L’automatisation vous aide à rationaliser le provisionnement des ressources, à faciliter des déploiements cohérents et sécurisés et à mettre à l’échelle les ressources sur l’ensemble de votre infrastructure.

Résultat escompté : vous gérez votre infrastructure en tant que code (IaC). Vous définissez et gérez votre code d'infrastructure dans des systèmes de contrôle de version (VCS). Vous déléguez le provisionnement des ressources AWS à des mécanismes automatisés et vous tirez parti de services gérés tels qu'Application Load Balancer (ALB), Network Load Balancer (NLB) et des groupes Auto Scaling. Vous provisionnez vos ressources à l'aide de pipelines d'intégration continue et livraison continue (CI/CD) afin que les modifications du code déclenchent automatiquement des mises à jour des ressources, y compris des mises à jour de vos configurations Auto Scaling.

Anti-modèles courants :

- Vous déployez des ressources manuellement via la ligne de commande ou dans la AWS Management Console (processus également appelé ClickOps).
- Vous associez étroitement vos ressources et composants d'application et vous créez ainsi des architectures rigides.
- Vous mettez en œuvre des politiques de mise à l'échelle rigides qui ne s'adaptent pas à l'évolution des exigences commerciales, des modèles de trafic ou des nouveaux types de ressources.
- Vous estimez manuellement la capacité pour répondre de façon anticipée à la demande.

Avantages liés au respect de cette bonne pratique : l'infrastructure en tant que code (IaC) permet de définir programmatiquement l'infrastructure. Vous pouvez ainsi gérer les modifications d'infrastructure en suivant le même cycle de développement logiciel que pour des modifications d'application, ce qui favorise la cohérence et la reproductibilité et réduit le risque de tâches manuelles susceptibles d'engendrer des erreurs. Vous pouvez rationaliser davantage le processus de provisionnement et de mise à jour des ressources en implémentant l'infrastructure en tant que code avec des pipelines de livraison automatisés. Vous pouvez déployer des mises à jour d'infrastructure de manière fiable et efficace sans nécessiter d'intervention manuelle. Cette agilité est particulièrement importante lorsque vous mettez à l'échelle les ressources pour répondre à des demandes fluctuantes.

Vous pouvez obtenir une mise à l'échelle dynamique et automatisée des ressources en conjonction avec l'infrastructure en tant que code et les pipelines de livraison. En surveillant les métriques clés et en appliquant des politiques de mise à l'échelle prédéfinies, Auto Scaling peut automatiquement provisionner ou déprovisionner les ressources selon les besoins, ce qui améliore les performances et la rentabilité. Cela réduit le risque d'erreurs manuelles ou de retards en réponse aux modifications des exigences en matière d'application ou de charge de travail.

La combinaison de l'infrastructure en tant que code, des pipelines de livraison automatisés et d'Auto Scaling aide les organisations à provisionner, mettre à jour et mettre à l'échelle leurs environnements

en toute confiance. Cette automatisation est essentielle pour maintenir une infrastructure cloud réactive, résiliente et gérée efficacement.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Pour configurer l'automatisation avec des pipelines CI/CD et une infrastructure en tant que code (IaC) pour votre architecture AWS, choisissez un système de contrôle de version tel que Git pour stocker vos modèles et votre configuration IaC. Ces modèles peuvent être écrits à l'aide d'outils tels qu'AWS CloudFormation. Pour commencer, définissez les composants de votre infrastructure (tels que les VPC AWS, les groupes Amazon EC2 Auto Scaling et les bases de données Amazon RDS) dans ces modèles.

Ensuite, intégrez ces modèles d'infrastructure en tant que code à un pipeline CI/CD pour automatiser le processus de déploiement. AWS CodePipeline fournit une solution AWS native transparente, ou vous pouvez utiliser d'autres solutions CI/CD tierces. Créez un pipeline qui s'active lorsque des modifications surviennent dans votre référentiel de contrôle de version. Configurez le pipeline de manière à inclure des étapes qui vérifient et valident vos modèles d'infrastructure en tant que code, déploient l'infrastructure dans un environnement intermédiaire, exécutent des tests automatisés et déploient finalement la solution en production. Incorporez des étapes d'approbation si nécessaire pour garder le contrôle sur les modifications. Ce pipeline automatisé accélère le déploiement, mais favorise également la cohérence et la fiabilité entre les environnements.

Configurez la mise à l'échelle automatique de ressources telles que les instances Amazon EC2, les tâches Amazon ECS et les réplicas de base de données dans votre infrastructure en tant que code afin d'assurer l'augmentation horizontale et la réduction horizontale automatiques selon les besoins. Cette approche améliore la disponibilité et les performances des applications et optimise les coûts en ajustant dynamiquement les ressources en fonction de la demande. Pour obtenir la liste des ressources prises en charge, consultez Amazon EC2 Auto Scaling et AWS Auto Scaling.

Étapes d'implémentation

1. Créez et utilisez un référentiel de code source pour stocker le code qui contrôle la configuration de votre infrastructure. Validez les modifications apportées à ce référentiel afin de refléter les modifications continues que vous souhaitez apporter.
2. Sélectionnez une solution d'infrastructure en tant que code, telle qu'AWS CloudFormation pour maintenir à jour votre infrastructure et détecter les incohérences (dérive) par rapport à l'état prévu.
3. Intégrez votre plateforme IaC à votre pipeline CI/CD pour automatiser les déploiements.

4. Déterminez et collectez les métriques appropriées pour la mise à l'échelle automatique des ressources.
5. Configurez la mise à l'échelle automatique des ressources à l'aide de politiques d'augmentation horizontale et de réduction horizontale pour vos composants de charge de travail. Envisagez d'utiliser une mise à l'échelle planifiée pour les modèles d'utilisation prévisibles.
6. Surveillez les déploiements pour détecter les défaillances et les régressions. Mettez en œuvre des mécanismes de restauration au sein de votre plateforme CI/CD pour annuler les modifications si nécessaire.

Ressources

Documents connexes :

- AWS Auto Scaling : Fonctionnement des plans de dimensionnement
- AWS Marketplace : produits utilisables avec autoscaling
- Gestion automatique de la capacité de débit avec l'autoscaling de DynamoDB
- Utiliser un équilibreur de charge avec un groupe Auto Scaling
- Qu'est-ce qu'AWS Global Accelerator ?
- Qu'est-ce qu'Amazon EC2 Auto Scaling ?
- Présentation de AWS Auto Scaling
- Qu'est-ce qu'Amazon CloudFront ?
- Qu'est-ce qu'Amazon Route 53 ?
- Qu'est-ce qu'Elastic Load Balancing ?
- Qu'est-ce qu'un équilibreur de charge Network Load Balancer ?
- Qu'est-ce qu'un équilibreur de charge Application Load Balancer ?
- Intégration de Jenkins avec AWS CodeBuild et AWS CodeDeploy
- Création d'un pipeline à quatre étapes avec AWS CodePipeline

Vidéos connexes :

- Back to Basics : Déployer votre code sur Amazon EC2
- AWS Supports You | Démarrer votre solution d'infrastructure en tant que code en utilisant les modèles AWS CloudFormation
- Rationaliser votre processus de publication de logiciel en utilisant AWS CodePipeline

- Surveiller les ressources AWS à l'aide des tableaux de bord Amazon CloudWatch
- Créer des tableaux de bord CloudWatch inter-comptes et inter-régions | Amazon Web Services

REL07-BP02 Obtenir des ressources en cas de détection d'une altération de la charge de travail

Si la disponibilité est affectée, mettez à l'échelle les ressources de manière réactive si nécessaire, afin de restaurer la disponibilité de la charge de travail.

Vous devez commencer par configurer les surveillances de l'état et les critères de ces vérifications pour indiquer quand la disponibilité est affectée par le manque de ressources. Informez ensuite le personnel approprié qu'il doit mettre à l'échelle manuellement la ressource ou lancer l'automatisation pour procéder à une mise à l'échelle automatique.

L'échelle peut être ajustée manuellement en fonction de votre charge de travail (par exemple, en modifiant le nombre d'EC2instances dans un groupe Auto Scaling ou en modifiant le débit d'une table DynamoDB via le ou). AWS Management Console AWS CLI Cependant, l'automatisation doit être utilisée dans la mesure du possible (voir Utiliser l'automatisation lors de l'obtention ou de la mise à l'échelle des ressources).

Résultat souhaité : les activités de mise à l'échelle (automatiquement ou manuellement) sont lancées pour rétablir la disponibilité en cas de détection d'une panne ou d'une dégradation de l'expérience client.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Mettez en œuvre l'observabilité et la surveillance de tous les composants de votre charge de travail, afin de surveiller l'expérience client et de détecter les défaillances. Définissez les procédures, manuelles ou automatisées, qui permettent de dimensionner les ressources requis. o Pour plus d'informations, voir REL11-BP01 Surveiller tous les composants de la charge de travail afin de détecter les défaillances.

Étapes d'implémentation

- Définissez les procédures, manuelles ou automatisées, de mise à l'échelle des ressources requises.
  - Les procédures de mise à l'échelle dépendent de la conception des différents composants de votre charge de travail.

- Les procédures de mise à l'échelle varient également en fonction de la technologie sous-jacente utilisée.
  - Les composants qui l'utilisent AWS Auto Scaling peuvent utiliser des plans de dimensionnement pour configurer un ensemble d'instructions permettant de dimensionner vos ressources. Si vous utilisez des AWS ressources AWS CloudFormation ou si vous y ajoutez des balises, vous pouvez configurer des plans de dimensionnement pour différents ensembles de ressources par application. Auto Scaling fournit des recommandations pour les stratégies de mise à l'échelle personnalisées pour chaque ressource. Une fois que vous avez créé votre plan de mise à l'échelle, Auto Scaling combine les méthodes de mise à l'échelle dynamique et prédictive pour prendre en charge votre stratégie de mise à l'échelle. Pour plus de détails, consultez Comment fonctionnent les plans de mise à l'échelle.
  - Amazon EC2 Auto Scaling vérifie que vous disposez du nombre correct d'EC2instances Amazon disponibles pour gérer la charge de votre application. Vous créez des collections d'EC2instances, appelées groupes Auto Scaling. Vous pouvez spécifier le nombre minimum et maximum d'instances dans chaque groupe Auto Scaling, et Amazon EC2 Auto Scaling garantit que votre groupe ne passe jamais en dessous ou au-dessus de ces limites. Pour plus de détails, consultez Qu'est-ce qu'Amazon EC2 Auto Scaling ?
  - L'autoscaling d'Amazon DynamoDB utilise le service d'autoscaling d'application pour ajuster de manière dynamique la capacité de débit approvisionné en votre nom, en réponse aux schémas de trafic réels. Cela permet à une table ou à un index secondaire global d'augmenter les capacités en lecture et écriture qui lui sont allouées afin de gérer les hausses soudaines de trafic sans limitation. Pour plus d'informations, voir Gestion automatique de la capacité de débit avec l'autoscaling de DynamoDB.

Ressources

Bonnes pratiques associées :

- REL07-BP01 Utiliser l'automatisation pour obtenir ou dimensionner des ressources
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances

Documents connexes :

- AWS Auto Scaling : Fonctionnement des plans de dimensionnement
- Gestion automatique de la capacité de débit avec l'autoscaling de DynamoDB
- Qu'est-ce qu'Amazon EC2 Auto Scaling ?

REL07-BP03 Obtenir des ressources après avoir réalisé qu'un plus grand nombre de ressources est nécessaire pour une charge de travail

L'une des fonctionnalités les plus précieuses du cloud computing est sa capacité à provisionner des ressources de manière dynamique.

Dans les environnements informatiques sur site traditionnels, vous devez identifier et provisionner une capacité suffisante à l'avance pour répondre à un pic de demande. Cela pose problème car cela coûte cher et présente des risques pour la disponibilité si vous sous-estimez la capacité maximale requise par la charge de travail.

Dans le cloud, vous n'avez pas à le faire. Au lieu de cela, vous pouvez provisionner comme il se doit des capacités de calcul, de base de données et d'autres ressources pour répondre à la demande actuelle et prévue. Des solutions automatisées telles qu'Amazon EC2 Auto Scaling et Application Auto Scaling peuvent mettre en ligne des ressources pour vous sur la base de métriques que vous spécifiez. Cela peut faciliter le processus de mise à l'échelle et le rendre prévisible, et cela peut rendre votre charge de travail nettement plus fiable en garantissant que vous disposez de suffisamment de ressources à tout moment.

Résultat escompté : vous configurez la mise à l'échelle automatique des ressources de calcul et autres pour répondre à la demande. Vous prévoyez une marge de manœuvre suffisante dans vos politiques de mise à l'échelle pour permettre de répondre à des pics de trafic pendant que des ressources supplémentaires sont mises en ligne.

Anti-modèles courants :

- Vous provisionnez un nombre fixe de ressources évolutives.
- Vous choisissez une métrique de mise à l'échelle qui ne correspond pas à la demande réelle.
- Vous ne parvenez pas à prévoir une marge de manœuvre suffisante dans vos plans de mise à l'échelle pour faire face aux pics de demande.
- Vos politiques de mise à l'échelle tardent trop à augmenter la capacité, ce qui entraîne l'épuisement de la capacité et une dégradation du service lors de la mise en ligne de ressources supplémentaires.
- Vous ne parvenez pas à configurer correctement le nombre minimal et le nombre maximal de ressources, ce qui entraîne des échecs de mise à l'échelle.

Avantages liés au respect de cette bonne pratique : il est essentiel de disposer de suffisamment de ressources pour répondre à la demande actuelle afin de garantir une haute disponibilité de votre charge de travail et de respecter les objectifs de niveau de service (SLO) définis. La mise à l'échelle

automatique vous permet de fournir la bonne quantité de ressources de calcul, de base de données et autres dont votre charge de travail a besoin afin de répondre à la demande actuelle et prévue. Vous n'avez pas besoin de déterminer la capacité maximale requise et d'allouer statiquement des ressources pour la fournir. Au lieu de cela, à mesure que la demande augmente, vous pouvez allouer davantage de ressources pour y répondre et, une fois que la demande est retombée, vous pouvez désactiver les ressources pour réduire les coûts.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Tout d'abord, déterminez si le composant de charge de travail est adapté à la mise à l'échelle automatique. Ces composants sont appelés évolutifs horizontalement car ils fournissent les mêmes ressources et se comportent de manière identique. Parmi les composants évolutifs horizontalement, citons les instances EC2 configurées de la même manière, les tâches Amazon Elastic Container Service (ECS) et les pods exécutés sur Amazon Elastic Kubernetes Service (EKS). Ces ressources de calcul sont généralement situées derrière un équilibreur de charge et sont appelées réplicas.

Les autres ressources répliquées peuvent inclure des réplicas en lecture de base de données, des tables Amazon DynamoDB et des clusters Amazon ElastiCache (Redis OSS). Pour obtenir la liste complète des ressources prises en charge, consultez Services AWS que vous pouvez utiliser avec Application Auto Scaling.

Pour les architectures basées sur des conteneurs, il peut être nécessaire de procéder à une mise à l'échelle de deux manières différentes. Tout d'abord, vous devrez peut-être mettre à l'échelle les conteneurs qui fournissent des services évolutifs horizontalement. Ensuite, vous devrez peut-être mettre à l'échelle les ressources de calcul pour libérer de l'espace pour de nouveaux conteneurs. Il existe différents mécanismes de mise à l'échelle automatique pour chaque couche. Pour mettre à l'échelle les tâches ECS, vous pouvez utiliser Application Auto Scaling. Pour mettre à l'échelle les pods Kubernetes, vous pouvez utiliser Horizontal Pod Autoscaler (HPA) ou Kubernetes Event-driven Autoscaler (KEDA). Pour mettre à l'échelle les ressources de calcul, vous pouvez utiliser les fournisseurs de capacité pour ECS, ou pour Kubernetes, vous pouvez utiliser Karpenter ou Cluster Autoscaler.

Ensuite, sélectionnez de quelle façon vous voulez effectuer la mise à l'échelle automatique. Il existe trois options principales : la mise à l'échelle basée sur des métriques, la mise à l'échelle planifiée et la mise à l'échelle prédictive.

Mise à l'échelle basée sur des métriques

La mise à l'échelle basée sur des métriques provisionne les ressources en fonction de la valeur d'une ou de plusieurs métriques de mise à l'échelle. Une métrique de mise à l'échelle est une métrique qui correspond à la demande de votre charge de travail. Un bon moyen de déterminer les métriques de mise à l'échelle appropriées consiste à effectuer des tests de charge dans un environnement hors production. Pendant vos tests de charge, maintenez fixe le nombre de ressources évolutives et augmentez lentement la demande (par exemple, débit, simultanéité ou utilisateurs simulés). Recherchez ensuite des métriques qui augmentent (ou diminuent) à mesure que la demande augmente, et inversement diminuent (ou augmentent) lorsque la demande diminue. Les métriques de mise à l'échelle typiques incluent l'utilisation du processeur, la profondeur de la file d'attente de travail (telle qu'une file d'attente Amazon SQS), le nombre d'utilisateurs actifs et le débit du réseau.

> **Note**
>
> AWS a observé qu'avec la plupart des applications, l'utilisation de la mémoire augmente à mesure que l'application monte en intensité, puis atteint une valeur stable. Lorsque la demande diminue, l'utilisation de la mémoire reste généralement élevée au lieu de diminuer en parallèle. Étant donné que l'utilisation de la mémoire ne correspond pas à la demande dans les deux cas de figure (à savoir en augmentant ni en diminuant avec la demande), réfléchissez bien avant de sélectionner cette métrique pour la mise à l'échelle automatique.

La mise à l'échelle basée sur des métriques est une opération latente. Plusieurs minutes peuvent être nécessaires pour que les métriques d'utilisation se propagent aux mécanismes de mise à l'échelle automatique, et ces mécanismes attendent généralement un signal clair d'augmentation de la demande avant de réagir. Ensuite, à mesure que l'outil de mise à l'échelle automatique crée de nouvelles ressources, la mise en service complète de ces dernières peut prendre plus de temps. Pour cette raison, il est important de ne pas définir vos cibles de métriques de mise à l'échelle trop proches de l'utilisation totale (par exemple, 90 % d'utilisation du processeur). Cela risque d'épuiser la capacité de ressources existante avant qu'une capacité supplémentaire puisse être mise en ligne. Les cibles typiques d'utilisation des ressources peuvent varier entre 50 et 70 % pour une disponibilité optimale, en fonction des modèles de demande et du temps nécessaire pour provisionner des ressources supplémentaires.

Mise à l'échelle planifiée

La mise à l'échelle planifiée provisionne ou supprime des ressources en fonction du calendrier ou de l'heure de la journée. Elle est fréquemment utilisée pour les charges de travail dont la demande est prévisible, telles que les pics d'utilisation pendant les heures ouvrables de semaine ou lors de

soldes. Amazon EC2 Auto Scaling et Application Auto Scaling prennent tous deux en charge la mise à l'échelle planifiée. La fonctionnalité cron scaler de KEDA prend en charge la mise à l'échelle planifiée des pods Kubernetes.

Mise à l'échelle prédictive

La mise à l'échelle prédictive utilise le machine learning pour mettre à l'échelle automatiquement les ressources en fonction de la demande anticipée. La mise à l'échelle prédictive analyse la valeur historique d'une métrique d'utilisation que vous fournissez et prédit en permanence sa valeur future. La valeur prédite est ensuite utilisée pour augmenter ou réduire verticalement la ressource. Amazon EC2 Auto Scaling peut effectuer une mise à l'échelle prédictive.

Étapes d'implémentation

1. Déterminez si le composant de charge de travail est adapté à la mise à l'échelle automatique.
2. Déterminez le type de mécanisme de mise à l'échelle le plus approprié pour la charge de travail : mise à l'échelle basée sur des métriques, mise à l'échelle planifiée ou mise à l'échelle prédictive.
3. Sélectionnez le mécanisme de mise à l'échelle automatique approprié pour le composant. Pour les instances Amazon EC2, utilisez Amazon EC2 Auto Scaling. Pour les autres services AWS, utilisez Application Auto Scaling. Pour les pods Kubernetes (tels que ceux exécutés dans un cluster Amazon EKS), pensez à Horizontal Pod Autoscaler (HPA) ou à Kubernetes Event-driven Autoscaling (KEDA). Pour les nœuds Kubernetes ou EKS, pensez à Karpenter et à Cluster Auto Scaler (CAS).
4. Pour une mise à l'échelle basée sur des métriques ou planifiée, effectuez des tests de charge afin de déterminer les métriques de mise à l'échelle et les valeurs cibles appropriées pour votre charge de travail. Pour une mise à l'échelle planifiée, déterminez le nombre de ressources nécessaires aux dates et heures que vous sélectionnez. Déterminez le nombre maximal de ressources nécessaires pour répondre aux pics de trafic attendus.
5. Configurez l'outil de mise à l'échelle en fonction des informations collectées ci-dessus. Pour plus d'informations, consultez la documentation du service de mise à l'échelle automatique. Vérifiez que les limites de mise à l'échelle maximale et minimale sont correctement configurées.
6. Vérifiez que la configuration de mise à l'échelle fonctionne comme prévu. Effectuez des tests de charge dans un environnement hors production, observez comment le système réagit et ajustez le cas échéant. Lorsque vous activez la mise à l'échelle automatique en production, configurez les alarmes appropriées pour être averti de tout comportement inattendu.

Ressources

Documents connexes :

- Qu’est-ce qu’Amazon EC2 Auto Scaling ?
- Conseils prescriptifs AWS : tests de charge des applications
- AWS Marketplace : produits utilisables avec autoscaling
- Gestion automatique de la capacité de débit avec l’autoscaling de DynamoDB
- Mise à l’échelle prédictive pour EC2 alimentée par le machine learning
- Mise à l’échelle planifiée pour Amazon EC2 Auto Scaling
- Telling Stories About Little’s Law

REL07-BP04 Testez votre charge de travail

Adoptez une méthodologie de test de charge pour déterminer si la mise à l’échelle répond aux exigences de la charge de travail.

Il est important d’exécuter régulièrement des tests de charge. Les tests de charge doivent découvrir le point de rupture et tester les performances de votre charge de travail. AWS facilite la mise en place d'environnements de test temporaires qui modélisent l'échelle de votre charge de travail de production. Dans le Cloud, vous pouvez créer un environnement d’essai à l’échelle de la production et à la demande, exécuter les tests, puis désactiver les ressources. Puisque vous ne payez l’environnement de test que lorsqu’il s’exécute, vous pouvez simuler votre environnement réel pour une fraction du coût d’un test sur site.

Les tests de charge en production doivent également être intégrés aux tests de simulation de pannes, lors desquels le système de production est mis sous tension pendant les périodes où le client est moins utilisé et tout le personnel est disponible pour interpréter les résultats et résoudre les problèmes qui surviennent.

Anti-modèles courants :

- Exécution de tests de charge sur des déploiements qui n’ont pas la même configuration que votre production.
- Exécution d’un test de charge uniquement sur des éléments individuels de votre charge de travail, et non sur l’ensemble de la charge de travail.

- Exécution de tests de charge avec un sous-ensemble de demandes et non un ensemble représentatif de demandes réelles.
- Exécution de tests de charge avec un faible facteur de sécurité au-dessus de la charge prévue.

Avantages du respect de cette bonne pratique : vous savez quels composants de votre architecture échouent sous charge et vous pouvez identifier les métriques à surveiller qui indiquent suffisamment à temps que vous approchez de cette charge pour que vous résolviez le problème et empêchiez ainsi l'impact de cette défaillance.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

- Exécutez des tests de charge pour identifier l'aspect de votre charge de travail qui indique que vous devez ajouter ou supprimer de la capacité. Les tests de charge doivent avoir un trafic représentatif similaire à ce que vous recevez en production. Augmentez la charge tout en surveillant les métriques que vous avez instrumentées pour déterminer quelle métrique indique quand vous devez ajouter ou supprimer des ressources.
  - Test de charge distribué sur AWS : simulez des milliers d'utilisateurs connectés
    - Identifiez le mélange de demandes. Comme vous pouvez avoir divers mélanges de demandes, vous devez examiner les différentes périodes lors de l'identification de la combinaison de trafic.
    - Implémentez un pilote de charge. Vous pouvez utiliser un code personnalisé, un logiciel open source ou un logiciel commercial pour implémenter un pilote de charge.
    - Effectuez un test de charge initial avec une faible capacité. Vous constatez des effets immédiats en entraînant une charge moindre, éventuellement aussi petite qu'une instance ou un conteneur.
    - Effectuez un test de charge par rapport à une capacité plus importante. Étant donné que les effets seront différents sur une charge distribuée, vous devez procéder à des essais dans un environnement aussi proche que possible de celui du produit.

Ressources

Documents connexes :

- Test de charge distribué sur AWS : simulez des milliers d'utilisateurs connectés

- Tests de charge des applications

Vidéos connexes :

- AWS Sommet ANZ 2023 : Accélérez en toute confiance grâce AWS aux tests de charge distribués

FIA 8. Comment mettre en œuvre des modifications ?

Des modifications contrôlées sont nécessaires pour déployer de nouvelles fonctionnalités et vérifier que les charges de travail et l'environnement d'exploitation fonctionnent avec des logiciels connus et peuvent être corrigés ou remplacés de manière prévisible. Si ces modifications ne sont pas maîtrisées, il devient difficile d'en prévoir les effets ou de résoudre les problèmes qui en découlent.

Bonnes pratiques

- REL08-BP01 Utiliser des runbooks pour les activités standard telles que le déploiement
- REL08-BP02 Intégrer les tests fonctionnels dans le cadre de votre déploiement
- REL08-BP03 Intégrez les tests de résilience dans le cadre de votre déploiement
- REL08-BP04 Effectuer le déploiement à l'aide d'une infrastructure immuable
- REL08-BP05 Déployer les modifications avec l'automatisation

REL08-BP01 Utiliser des runbooks pour les activités standard telles que le déploiement

Les runbooks sont les procédures prédéfinies destinées à parvenir à un résultat spécifique. Utilisez des runbooks pour effectuer des tâches manuelles ou automatiques standard. Il peut s'agir du déploiement d'une charge de travail, de l'application de correctifs à une charge de travail ou de la modification du DNS.

Par exemple, mettez en place des processus pour assurer la sécurité des restaurations pendant les déploiements. Pour garantir la fiabilité d'un service, il est essentiel de s'assurer que vous pouvez restaurer un déploiement sans interruption pour vos clients.

Concernant les procédures de runbook, commencez par un processus manuel efficace valide, mettez-le en œuvre dans le code et, le cas échéant, déclenchez son exécution automatique.

Même pour les charges de travail sophistiquées hautement automatisées, les runbooks restent utiles pour exécuter des tests de simulation de pannes ou répondre à des exigences rigoureuses en matière de rapports et d'audit.

Notez que les playbooks sont utilisés en réponse à des incidents spécifiques et que les runbooks le sont pour obtenir des résultats spécifiques. En règle générale, les runbooks sont destinés aux activités de routine, tandis que les playbooks sont utilisés pour répondre à des événements non réguliers.

Anti-modèles courants :

- Exécution de modifications imprévues de la configuration en production.
- Ignorer les étapes de votre plan afin d'accélérer le déploiement, ce qui entraîne un échec du déploiement.
- Effectuer des modifications sans tester l'annulation de la modification.

Avantages du respect de cette bonne pratique : une planification efficace des modifications augmente votre capacité à exécuter correctement la modification, car vous êtes conscient de tous les systèmes concernés. Vous gagnez en confiance si vous réussissez à valider des modifications que vous apportez aux environnements de test.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

- Obtenez des réponses cohérentes et rapides à des événements bien compris en documentant les procédures dans des runbooks.
  - Concepts AWS Well-Architected Framework : runbook
- Utilisez le principe de l'infrastructure en tant que code pour définir votre infrastructure. En ayant recours à AWS CloudFormation ou à un tiers de confiance pour définir votre infrastructure, vous pouvez utiliser le contrôle de version et suivre les modifications apportées à la version du logiciel.
  - Utilisez AWS CloudFormation ou un fournisseur tiers de confiance pour définir votre infrastructure.
    - Présentation de AWS CloudFormation
  - Créez des modèles qui sont singuliers et découplés, en utilisant de bons principes de conception de logiciels.
    - Déterminez les autorisations, les modèles et les responsables de l'implémentation.
      - Contrôle de l'accès avec AWS Identity and Access Management

- Utilisez un système de gestion de code source hébergé basé sur une technologie populaire telle que Git pour stocker votre code source et votre configuration d'infrastructure en tant que code (IaC).

Ressources

Documents connexes :

- Partenaire APN : partenaires pouvant vous aider à créer des solutions de déploiement automatisées
- AWS Marketplace : produits pouvant être utilisés pour automatiser vos déploiements
- Concepts AWS Well-Architected Framework : runbook
- Présentation de AWS CloudFormation

Exemples connexes :

- Automatisation des opérations avec les playbooks et les runbooks

REL08-BP02 Intégrer les tests fonctionnels dans le cadre de votre déploiement

Utilisez des techniques telles que les tests unitaires et les tests d'intégration qui valident les fonctionnalités requises.

Un test unitaire est un processus consistant à tester la plus petite unité fonctionnelle du code afin de valider son comportement. Un test d'intégration vise à valider que chaque fonctionnalité de l'application respecte les exigences du logiciel. Alors que les tests unitaires visent à tester une partie d'une application de manière isolée, les tests d'intégration prennent en compte les effets secondaires (par exemple, l'effet de la modification des données lors d'une opération de mutation). Dans les deux cas, les tests doivent être intégrés dans un pipeline de déploiement et si les critères de réussite ne sont pas respectés, le pipeline est arrêté ou annulé. Ces tests sont exécutés dans un environnement de préproduction, qui est mis en place avant la production dans le pipeline.

Vous obtenez les meilleurs résultats lorsque ces tests sont exécutés automatiquement dans le cadre d'actions de génération et de déploiement. Par exemple, avec AWS CodePipeline, les développeurs valident les modifications apportées à un référentiel source dans lequel CodePipeline détecte automatiquement les modifications. L'application est générée et des tests unitaires sont exécutés. Une fois les tests unitaires réussis, le code ainsi généré est déployé sur les serveurs intermédiaires,

à des fins de test. Depuis le serveur intermédiaire, CodePipeline exécute d'autres tests, tels que des tests d'intégration ou de chargement. Une fois ces tests terminés avec succès, CodePipeline déploie le code testé et approuvé sur les instances de production.

Résultat escompté : vous utilisez l'automatisation pour effectuer des tests unitaires et d'intégration afin de valider que votre code se comporte comme prévu. Ces tests sont intégrés au processus de déploiement, et un échec entraîne l'abandon du déploiement.

Anti-modèles courants :

- Vous ignorez ou contournez les échecs de test et les plans pendant le processus de déploiement afin de raccourcir le délai de déploiement.
- Vous effectuez manuellement des tests en dehors du pipeline de déploiement.
- Vous sautez des étapes de test dans l'automatisation par le biais de flux de travail d'urgence manuels.
- Vous exécutez des tests automatisés dans un environnement qui ne ressemble vraiment pas à l'environnement de production.
- Vous créez une suite de tests trop peu flexible et difficile à maintenir, à mettre à jour ou à mettre à l'échelle à mesure de l'évolution de l'application.

Avantages liés au respect de cette bonne pratique : les tests automatisés effectués au cours du processus de déploiement détectent les problèmes à un stade précoce, ce qui réduit le risque d'une mise en production avec des bogues ou des comportements inattendus. Les tests unitaires valident le fait que le code se comporte comme souhaité et que les contrats d'API sont respectés. Les tests d'intégration valident le fait que le système fonctionne conformément aux exigences spécifiées. Ces types de tests vérifient l'état de fonctionnement escompté de composants tels que les interfaces utilisateur, les API, les bases de données et le code source.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Adoptez une approche de développement piloté par les tests (TDD) pour écrire des logiciels, dans laquelle vous développez des cas de test pour spécifier et valider votre code. Pour commencer, créez des cas de test pour chaque fonction. Si le test échoue, vous devez écrire un nouveau code pour réussir le test. Cette approche vous permet de valider le résultat escompté de chaque fonction. Exécutez des tests unitaires et validez leur réussite avant de valider le code dans un référentiel de code source.

Mettez en œuvre des tests unitaires et d'intégration dans le cadre des étapes de création, de test et de déploiement du pipeline CI/CD. Automatisez les tests et lancez automatiquement les tests chaque fois qu'une nouvelle version de l'application est prête à être déployée. Si les critères de réussite ne sont pas respectés, le pipeline est arrêté ou annulé.

Dans le cas d'une application Web ou mobile, effectuez des tests d'intégration automatisés sur plusieurs navigateurs de bureau ou sur des appareils réels. Cette approche est particulièrement utile pour valider la compatibilité et les fonctionnalités des applications mobiles sur un large éventail d'appareils.

Étapes d'implémentation

1. Rédigez des tests unitaires avant d'écrire du code fonctionnel (développement piloté par les tests ou TDD). Établissez des lignes directrices en matière de code afin que l'écriture et l'exécution de tests unitaires soient une exigence non fonctionnelle de codage.
2. Créez une suite de tests d'intégration automatisés qui couvrent les fonctionnalités testables identifiées. Ces tests doivent simuler les interactions avec les utilisateurs et valider les résultats attendus.
3. Créez l'environnement de test nécessaire pour exécuter les tests d'intégration. Cela peut inclure des environnements de préparation ou de pré-production qui imitent étroitement l'environnement de production.
4. Configurez vos étapes source, de construction, de test et de déploiement à l'aide de la console AWS CodePipeline ou de l'AWS Command Line Interface (CLI).
5. Déployez l'application une fois que le code a été créé et testé. AWS CodeDeploy peut la déployer dans vos environnements intermédiaire (tests) et de production. Ces environnements peuvent inclure des instances Amazon EC2, des fonctions AWS Lambda et des serveurs sur site. Le même mécanisme de déploiement doit être utilisé pour déployer l'application dans tous les environnements.
6. Surveillez la progression de votre pipeline et le statut de chaque étape. Utilisez des contrôles de qualité pour bloquer le pipeline en fonction du statut des tests. Vous pouvez aussi recevoir des notifications en cas d'échec ou d'achèvement d'une étape du pipeline.
7. Surveillez en permanence les résultats des tests et recherchez des modèles, des régressions ou des domaines nécessitant une plus grande attention. Utilisez ces informations pour améliorer la suite de tests, identifier les domaines de l'application nécessitant des tests plus approfondis et optimiser le processus de déploiement.

Ressources

Bonnes pratiques associées :

- REL07-BP04 Effectuer un test de charge de votre charge de travail
- REL08-BP03 Intégrer les tests de résilience dans le cadre de votre déploiement
- REL12-BP04 Tester la résilience à l'aide de l'ingénierie du chaos

Documents connexes :

- Conseils prescriptifs AWS : automatisation des tests
- Intégration et livraison continues
- Indicateurs pour les tests fonctionnels
- Surveillance des pipelines
- Utilisation d'AWS CodePipeline avec AWS CodeBuild pour tester le code et exécuter des générations
- AWS Device Farm

REL08-BP03 Intégrez les tests de résilience dans le cadre de votre déploiement

Intégrez des tests de résilience en introduisant consciemment des défaillances dans le système afin de mesurer sa fonctionnalité en cas de scénarios perturbateurs. Les tests de résilience sont différents des tests unitaires et fonctionnels qui sont généralement intégrés dans les cycles de déploiement, car ils se concentrent sur l'identification des défaillances imprévues de votre système. Bien qu'il soit prudent de commencer par l'intégration des tests de résilience en préproduction, fixez-vous comme objectif d'implémenter ces tests en production dans le cadre de vos journées de simulation.

Résultat escompté : les tests de résilience contribuent à renforcer la confiance dans la capacité du système à résister à la dégradation en cours de production. Les expériences identifient les points faibles susceptibles d'entraîner une défaillance, ce qui vous permet d'améliorer le système afin d'atténuer automatiquement et efficacement les défaillances et la dégradation.

Anti-modèles courants :

- Manque d'observabilité et de surveillance dans les processus de déploiement
- Dépendance à l'humain pour résoudre les défaillances du système
- Mécanismes d'analyse de mauvaise qualité

- Accent mis sur les problèmes connus d'un système et manque d'expérimentation pour identifier les problèmes inconnus
- Identification des défaillances, mais pas de solution
- Aucune documentation sur les résultats ni aucun runbook

Avantages de l'établissement de bonnes pratiques : les tests de résilience intégrés à vos déploiements permettent d'identifier les problèmes inconnus du système qui, autrement, passeraient inaperçus, ce qui peut entraîner des interruptions de production. L'identification de ces problèmes système inconnus vous permet de documenter les résultats, d'intégrer les tests dans votre processus CI/CD et de créer des runbooks, ce qui simplifie leur atténuation grâce à des mécanismes efficaces et reproductibles.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les formes de test de résilience les plus courantes pouvant être intégrées dans les déploiements de votre système sont la reprise après sinistre et l'ingénierie du chaos.

- Incluez des mises à jour de vos plans de reprise après sinistre et de vos procédures opérationnelles standard (SOPs) lors de tout déploiement important.
- Intégrez des tests de fiabilité à vos pipelines de déploiement automatisés. Des services tels que AWS Resilience Hub peuvent être intégrés à votre pipeline CI/CD pour établir des évaluations continues de la résilience qui sont automatiquement évaluées dans le cadre de chaque déploiement.
- Définissez vos applications dans AWS Resilience Hub. Les évaluations de résilience génèrent des extraits de code qui vous aident à créer des procédures de restauration sous forme de documents AWS Systems Manager pour vos applications et fournissent une liste de CloudWatch moniteurs et d'alarmes Amazon recommandés.
- Une fois que vos plans de reprise après sinistre SOPs sont planifiés et mis à jour, effectuez des tests de reprise après sinistre pour vérifier leur efficacité. Les tests de reprise après sinistre contribuent à déterminer si vous pouvez restaurer votre système après un événement et revenir à un fonctionnement normal. Vous pouvez simuler différentes stratégies de reprise après sinistre et déterminer si votre planification est suffisante pour répondre à vos exigences de disponibilité. Les stratégies courantes de reprise après sinistre incluent la sauvegarde et la restauration, l'environnement en veille, la veille à froid, la veille à chaud, la veille permanente et la veille active/active. Elles diffèrent toutes en matière de coût et de complexité. Avant les tests de reprise après

sinistre, nous vous recommandons de définir votre objectif de temps de reprise (RTO) et votre objectif de point de reprise (RPO) afin de simplifier le choix de la stratégie à simuler. AWS propose des outils de reprise après sinistre destinés AWS Elastic Disaster Recoveryà vous aider à démarrer votre planification et vos tests.

- Les expériences d'ingénierie du chaos introduisent des perturbations dans le système, telles que des pannes de réseau et des pannes de service. En simulant le système avec des pannes contrôlées, vous pouvez en découvrir les vulnérabilités tout en limitant les impacts des pannes injectées. Comme pour les autres stratégies, exécutez des simulations de défaillances contrôlées dans des environnements non liés à la production en utilisant des services tels que AWS Fault Injection Service pour gagner en confiance avant de les déployer en production.

Ressources

Documents connexes :

- Tester les défaillances à l'aide de tests de résilience pour renforcer la préparation à la reprise
- Évaluation continue de la résilience des applications avec AWS Resilience Hub et AWS CodePipeline
- Architecture de reprise après sinistre (DR) sur AWS, partie 1 : Stratégies de reprise dans le cloud
- Vérifier la résilience de vos charges de travail à l'aide de Chaos Engineering
- Principes de l'ingénierie du chaos
- Atelier d'ingénierie du chaos

Vidéos connexes :

- AWS re:Invent 2020: Testing Resilience using Chaos Engineering
- Améliorez la résilience des applications grâce au service d'injection de AWS défauts
- Préparez et protégez vos applications contre les perturbations avec AWS Resilience Hub

REL08-BP04 Effectuer le déploiement à l'aide d'une infrastructure immuable

Une infrastructure immuable est un modèle qui exige qu'aucune mise à jour, aucune application de correctifs de sécurité ni aucun changement de configuration ne se produise sur place sur les charges de travail de production. Lorsqu'un changement est nécessaire, l'architecture est intégrée à la nouvelle infrastructure et déployée en production.

Suivez une stratégie de déploiement d'infrastructure immuable pour améliorer la fiabilité, la cohérence et la reproductibilité de vos déploiements de charges de travail.

Résultat escompté : avec une infrastructure immuable, aucune modification sur place n'est autorisée pour exécuter les ressources de l'infrastructure dans le cadre d'une charge de travail. Lorsqu'une modification est nécessaire, un nouvel ensemble de ressources d'infrastructure contenant toutes les modifications nécessaires est déployé parallèlement à vos ressources existantes. Ce déploiement est validé automatiquement et, en cas de succès, le trafic est progressivement transféré vers ce nouvel ensemble de ressources.

Cette stratégie de déploiement s'applique notamment aux mises à jour logicielles, aux correctifs de sécurité, aux modifications de l'infrastructure, ainsi qu'aux mises à jour de la configuration et des applications.

Anti-modèles courants :

- Modifications sur place des ressources d'infrastructure en cours d'exécution.

Avantages liés au respect de cette bonne pratique :

- Cohérence accrue entre les environnements : comme il n'existe aucune différence dans les ressources d'infrastructure entre les environnements, la cohérence est améliorée et les tests simplifiés.
- Réduction des dérives de configuration : en remplaçant fréquemment les ressources d'infrastructure à partir d'une configuration de base connue et contrôlée par les versions, l'infrastructure est réinitialisée à un état connu, testé et fiable, ce qui évite les dérives de configuration.
- Déploiements atomiques fiables : les déploiements se terminent avec succès ou rien ne change, ce qui améliore la cohérence et la fiabilité du processus de déploiement.
- Déploiements simplifiés : les déploiements sont simplifiés, car ils n'ont pas besoin de prendre en charge les mises à niveau. Les mises à niveau sont simplement de nouveaux déploiements.
- Déploiements plus sûrs avec des processus de restauration et de récupération rapides : les déploiements sont plus sûrs, car la version de travail précédente n'est pas modifiée. Vous pouvez la restaurer si des erreurs sont détectées.
- Position de sécurité améliorée : en interdisant les modifications de l'infrastructure, les mécanismes d'accès à distance (tels que SSH) peuvent être désactivés. Vous pouvez ainsi réduire les vecteurs d'attaque tout en renforçant la sécurité de votre organisation.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Automation

Lors de la définition d'une stratégie de déploiement d'infrastructure immuable, il est recommandé de recourir autant que possible à l'automatisation afin d'accroître la reproductibilité et de minimiser le risque d'erreur humaine. Pour plus de détails, voir REL08-BP05 Déployer les modifications grâce à l'automatisation et Automatisation de déploiements sûrs et sans intervention directe.

Avec l'infrastructure en tant que code (IaC), les étapes de provisionnement, d'orchestration et de déploiement de l'infrastructure sont définies de manière programmatique, descriptive et déclarative et stockées dans un système de contrôle des sources. L'utilisation de l'infrastructure en tant que code simplifie l'automatisation du déploiement de l'infrastructure et contribue à garantir l'immuabilité de cette dernière.

Modèles de déploiement

Lorsqu'une modification de la charge de travail est requise, la stratégie de déploiement d'infrastructure immuable impose le déploiement d'un nouvel ensemble de ressources d'infrastructure comprenant toutes les modifications nécessaires. Il est important que ce nouvel ensemble de ressources suive un schéma de déploiement qui minimise l'impact sur les utilisateurs. Il existe deux stratégies principales pour ce type de déploiement :

Déploiement Canary : consiste à diriger un petit nombre de vos clients vers la nouvelle version, généralement exécutée sur une seule instance de service (la version Canary). Examinez ensuite en profondeur les modifications de comportement ou les erreurs générées. Vous pouvez supprimer le trafic du Canary si vous rencontrez des problèmes critiques et faire basculer les utilisateurs vers la version précédente. Si le déploiement est réussi, vous pouvez le continuer à la vitesse souhaitée, tout en surveillant les modifications (pour éviter les erreurs), jusqu'à ce qu'il soit terminé. AWS CodeDeploy peut être configuré avec une configuration de déploiement qui permettra un déploiement canary.

Déploiement bleu/vert : semblable au déploiement Canary si ce n'est qu'un parc complet de l'application est déployé en parallèle. Vos déploiements alternent entre deux piles (bleu et vert). Une fois encore, vous pouvez faire basculer le trafic vers la nouvelle version et revenir à l'ancienne si vous rencontrez des problèmes lors du déploiement. Généralement, tout le trafic est commuté en même temps. Vous pouvez toutefois également utiliser des fractions de votre trafic vers chaque version pour modifier l'adoption de la nouvelle version à l'aide des capacités de routage DNS pondéré

d'Amazon Route 53. AWS CodeDeploy et AWS Elastic Beanstalk peuvent être configurés avec une configuration de déploiement qui permet un déploiement bleu/vert.

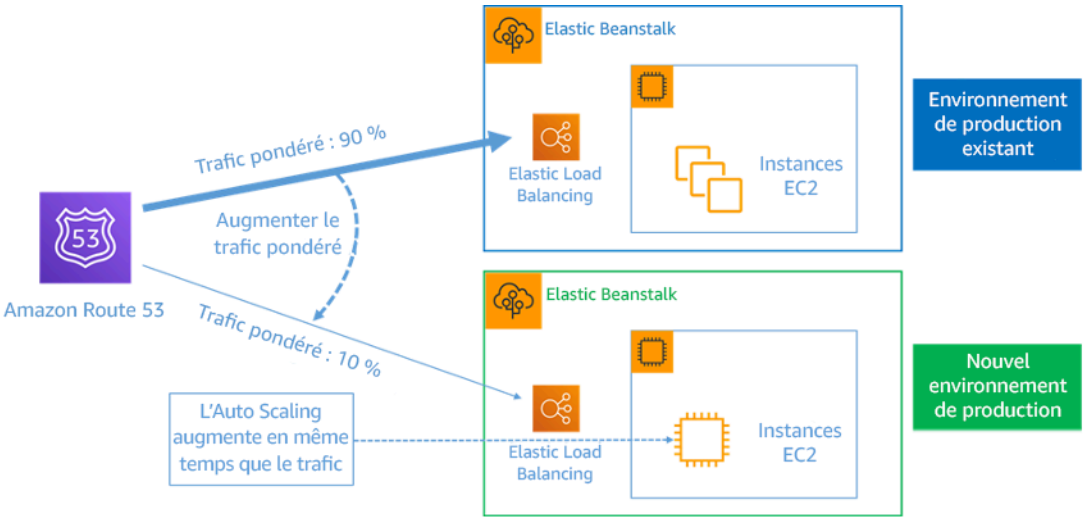


Figure 8 : Déploiement bleu/vert avec AWS Elastic Beanstalk et Amazon Route 53

Détection des écarts

L'écart est défini comme tout changement qui fait qu'une ressource d'infrastructure présente un état ou une configuration différent de ce qui est attendu. Toute modification de configuration non gérée va à l'encontre de la notion d'infrastructure immuable et doit être détectée et corrigée afin de garantir la mise en œuvre d'une infrastructure immuable.

Étapes d'implémentation

- Interdisez la modification sur place des ressources d'infrastructure en cours d'exécution.
  - Vous pouvez utiliser AWS Identity and Access Management (IAM) pour spécifier qui ou quoi peut accéder aux services et aux ressources dans AWS, gérer de manière centralisée les autorisations détaillées et analyser les accès pour affiner les autorisations dans AWS.
- Automatisez le déploiement des ressources d'infrastructure pour améliorer la reproductibilité et minimiser le risque d'erreur humaine.
  - Comme décrit dans le livre blanc Introduction au DevOps sur AWS, l'automatisation est la pierre angulaire des services AWS et est prise en charge en interne dans tous les services, fonctionnalités et offres.
  - La préintégration de votre Amazon Machine Image (AMI) peut accélérer son lancement. EC2 Image Builder est un service AWS entièrement géré qui vous permet d'automatiser la création, la maintenance, la validation, le partage et le déploiement d'une AMI Linux ou Windows personnalisée, fiable et à jour.

- Les services qui prennent en charge l'automatisation incluent :
  - AWS Elastic Beanstalk est un service permettant de déployer et de mettre à l'échelle rapidement des applications et des services web développés avec Java, .NET, PHP, Node.js, Python, Ruby, Go et Docker sur des serveurs courants, tels qu'Apache, NGINX, Passenger et IIS.
  - AWS Proton aide les équipes de plateforme à connecter et à coordonner les différents outils dont vos équipes de développement ont besoin pour le provisionnement de l'infrastructure, les déploiements de code, la surveillance et les mises à jour. AWS Proton permet une infrastructure automatisée sous forme de provisionnement de code et de déploiement d'applications sans serveur et basées sur des conteneurs.
- L'utilisation d'une infrastructure en tant que code facilite l'automatisation du déploiement de l'infrastructure et contribue à garantir l'immuabilité de l'infrastructure. AWS fournit des services de création, de déploiement et de maintenance programmatique, descriptive et déclarative de l'infrastructure.
  - AWS CloudFormation aide les développeurs à créer des ressources AWS de manière ordonnée et prévisible. Les ressources sont écrites dans des fichiers texte au format JSON ou YAML. Les modèles nécessitent une syntaxe et une structure spécifiques, qui dépendent des types de ressources créées et gérées. Vous créez vos ressources au format JSON ou YAML avec n'importe quel éditeur de code, vous les archivez dans un système de contrôle de version, puis AWS CloudFormation crée les services spécifiés d'une manière sûre et reproductible.
  - AWS Serverless Application Model (AWS SAM) est un cadre open source que vous pouvez utiliser pour construire des applications sans serveur sur AWS. AWS SAM s'intègre à d'autres services AWS et constitue une extension de AWS CloudFormation.
  - AWS Cloud Development Kit (AWS CDK) est un cadre de développement logiciel open source que vous pouvez utiliser pour modéliser et allouer vos ressources d'applications cloud à l'aide de langages de programmation familiers. Vous pouvez utiliser AWS CDK pour modéliser l'infrastructure d'applications avec TypeScript, Python, Java et .NET. AWS CDK utilise AWS CloudFormation en arrière-plan pour provisionner les ressources de manière sécurisée et reproductible.
  - AWS Cloud Control API introduit un ensemble commun d'API CRUDL (Create, Read, Update, Delete, and List) pour aider les développeurs à gérer leur infrastructure cloud de manière simple et cohérente. Les API courantes de Cloud Control API permettent aux développeurs de gérer de manière uniforme le cycle de vie des services AWS et tiers.
- Mettez en œuvre des modèles de déploiement qui minimisent l'impact sur les utilisateurs.

  - Déploiements canary :
    - Configuration d'un déploiement de la version canary API Gateway
    - Créer un pipeline avec des déploiements Canary pour Amazon ECS à l'aide de AWS App Mesh
  - Déploiements bleu/vert : le livre blanc sur les déploiements bleu/vert sur AWS décrit des exemples de techniques pour mettre en œuvre des stratégies de déploiement bleu/vert.
- Détectez les écarts de configuration ou d'état. Pour plus de détails, consultez Détection de modifications non gérées de la configuration des piles et des ressources.

Ressources

Bonnes pratiques associées :

- REL08-BP05 Déployer les modifications avec l'automatisation

Documents connexes :

- Automatiser les déploiements sécurisés et sans intervention
- Tirer parti de AWS CloudFormation pour créer une infrastructure immuable chez Nubank
- Infrastructure en tant que code
- Implémentation d'une alarme pour détecter automatiquement l'écart dans les piles AWS CloudFormation

Vidéos connexes :

- AWS re:Invent 2020: Reliability, consistency, and confidence through immutability

REL08-BP05 Déployer les modifications avec l'automatisation

Les déploiements et l'application de correctifs sont automatisés pour éliminer l'impact négatif.

Les modifications apportées aux systèmes de production sont l'un des secteurs de risque les plus importants pour de nombreuses organisations. Nous considérons les déploiements comme un problème de premier ordre à résoudre, tout comme les problèmes opérationnels que le logiciel rencontre. Aujourd'hui, il convient d'appliquer l'automatisation dès que les opérations le permettent,

y compris lors des tests et du déploiement de modifications, lors de l'ajout ou de la suppression de capacités et lors de la migration des données.

Résultat escompté : vous intégrez la sécurité des déploiements automatisés dans le processus de publication grâce à des tests de pré-production approfondis, à des annulations automatiques et à des déploiements de production échelonnés. Cette automatisation minimise l'impact potentiel de l'échec des déploiements sur la production, et les développeurs n'ont plus besoin de surveiller activement les déploiements jusqu'à la production.

Anti-modèles courants :

- Vous effectuez des modifications manuelles.
- Vous sautez des étapes de l'automatisation grâce à des flux de travail d'urgence manuels.
- Vous ne suivez pas les plans et processus établis au profit de délais accélérés.
- Vous effectuez des déploiements de suivi rapides sans prévoir de durée d'intégration.

Avantages du respect de cette bonne pratique : lorsque vous utilisez l'automatisation pour déployer toutes les modifications, vous éliminez le risque d'erreur humaine et vous offrez la possibilité de tester avant de modifier la production. L'exécution de ce processus avant la phase de production permet de vérifier que vos plans sont complets. En outre, la restauration automatique de votre processus de publication peut identifier les problèmes de production et ramener votre charge de travail à son état de fonctionnement antérieur.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Automatisez votre pipeline de déploiement. Le déploiement des pipelines vous permet d'une part d'invoquer des tests automatisés et la détection des anomalies et, d'autre part, d'arrêter le pipeline à une certaine étape avant le déploiement en production ou de restaurer automatiquement l'environnement d'avant la modification. L'adoption de la culture de l'intégration continue et de la livraison/du déploiement continus (CI/CD) en fait partie intégrante. Lors de celle-ci, un commit ou une modification de code passe par différentes étapes automatisées, des étapes de construction et de test au déploiement dans les environnements de production.

Bien que les principes traditionnels suggèrent d'impliquer l'intervention humaine pour les procédures opérationnelles les plus complexes, nous vous conseillons d'automatiser ces mêmes procédures pour cette même raison.

Étapes d'implémentation

Vous pouvez automatiser les déploiements pour supprimer les opérations manuelles en procédant comme suit :

- Configurez un référentiel de code pour stocker votre code en toute sécurité : utilisez un système de gestion de code source hébergé basé sur une technologie populaire telle que Git pour stocker votre code source et votre configuration d'infrastructure en tant que code (IaC).
- Configurez un service d'intégration continue pour compiler votre code source, exécuter des tests et créer des artefacts de déploiement : pour configurer un projet de génération à cette fin, voir Commencer avec l'utilisation de la console par AWS CodeBuild.
- Configurez un service de déploiement qui automatise les déploiements d'applications et gère la complexité des mises à jour des applications sans recourir à des déploiements manuels sujets aux erreurs : AWS CodeDeploy automatise les déploiements de logiciels vers divers services informatiques, tels qu'Amazon EC2 AWS Fargate, AWS Lambda et vos serveurs sur site. Pour configurer ces étapes, consultez Premiers pas avec CodeDeploy.
- Configurez un service de livraison continue qui automatise vos pipelines de publication pour des mises à jour plus rapides et plus fiables des applications et de l'infrastructure : envisagez d'utiliser AWS CodePipeline pour vous aider à automatiser vos pipelines de publication. Pour plus de détails, consultez les didacticiels CodePipeline.

Ressources

Bonnes pratiques associées :

- OPS05-BP04 Utiliser des systèmes de gestion du développement et du déploiement
- OPS05-BP10 Automatiser complètement l'intégration et le déploiement
- OPS06-BP02 Déploiements de tests
- OPS06-BP04 Automatiser les tests et les restaurations

Documents connexes :

- Livraison continue de piles AWS CloudFormation imbriquées à l'aide de AWS CodePipeline
- Partenaire APN : partenaires pouvant vous aider à créer des solutions de déploiement automatisées
- AWS Marketplace : produits pouvant être utilisés pour automatiser vos déploiements

- Automatisez les messages de chat avec les webhooks.
- L'Amazon Builders' Library : Garantir la sécurité des restaurations pendant les déploiements
- L'Amazon Builders' Library : Aller plus vite avec la distribution continue
- Présentation de AWS CodePipeline
- Qu'est-ce que CodeDeploy ?
- Gestionnaire de correctifs d'AWS Systems Manager
- Qu'est-ce qu'Amazon SES ?
- Qu'est-ce qu'Amazon Simple Notification Service ?

Vidéos connexes :

- AWS Summit 2019: CI/CD on AWS

Gestion des défaillances

Questions

- FIA 9. Comment sauvegarder les données ?
- FIA 10. Comment utiliser l'isolation des pannes pour protéger votre charge de travail ?
- FIA 11. Comment concevoir votre charge de travail pour la rendre résistante aux défaillances de composants ?
- FIA 12. Comment tester la fiabilité ?
- FIA 13. Comment planifier la reprise après sinistre (DR) ?

FIA 9. Comment sauvegarder les données ?

Sauvegardez les données, les applications et la configuration pour répondre à vos exigences en matière d'objectifs de délai de reprise (RTO) et de points de reprise (RPO).

Bonnes pratiques

- REL09-BP01 Identifiez et sauvegardez toutes les données qui doivent être sauvegardées, ou reproduisez les données à partir des sources
- REL09-BP02 Sauvegardes sécurisées et cryptées

- REL09-BP03 Effectuer une sauvegarde automatique des données
- REL09-BP04 Effectuer une restauration périodique des données pour vérifier l'intégrité et les processus de sauvegarde

REL09-BP01 Identifiez et sauvegardez toutes les données qui doivent être sauvegardées, ou reproduisez les données à partir des sources

Identifiez et utilisez les fonctionnalités de sauvegarde des services et ressources de données utilisés par votre charge de travail. La plupart des services offrent des fonctionnalités permettant de sauvegarder vos données de charge de travail.

Résultat escompté : les sources de données ont été identifiées et classées en fonction de leur ordre d'importance. Ensuite, établissez une stratégie de récupération de données basée surRPO. Cette stratégie implique soit de sauvegarder ces sources de données, soit d'avoir la capacité de reproduire des données provenant d'autres sources. En cas de perte de données, la stratégie mise en œuvre permet la récupération ou la reproduction des données dans les limites définies RPO etRTO.

Phase de maturité du cloud : fondamentale

Anti-modèles courants :

- Ne pas connaître toutes les sources de données pour la charge de travail ni leur ordre d'importance.
- Ne pas effectuer de sauvegardes des sources de données critiques.
- Sauvegarder uniquement certaines sources de données sans utiliser leur ordre d'importance comme critère.
- Aucune fréquence définieRPO, ou la fréquence de sauvegarde ne peut pas être atteinteRPO.
- Ne pas évaluer si une sauvegarde est nécessaire ou si les données peuvent être reproduites à partir d'autres sources.

Avantages liés au respect de cette bonne pratique : identifier les emplacements où les sauvegardes sont nécessaires et mettre en place un mécanisme pour créer des sauvegardes, ou être capable de reproduire les données à partir d'une source externe améliore la capacité de restauration et de récupération des données lors d'une panne.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d’implémentation

Tous les magasins AWS de données offrent des fonctionnalités de sauvegarde. Des services tels qu'Amazon RDS et Amazon DynamoDB prennent également en charge la sauvegarde automatique qui point-in-time permet de récupérer PITR (), ce qui vous permet de restaurer une sauvegarde à tout moment jusqu'à cinq minutes ou moins avant l'heure actuelle. De nombreux AWS services offrent la possibilité de copier des sauvegardes vers un autre Région AWS. AWS Backup est un outil qui vous permet de centraliser et d'automatiser la protection des données dans l'ensemble des AWS services. AWS Elastic Disaster Recoveryvous permet de copier les charges de travail complètes du serveur et de maintenir une protection continue des données sur site, entre zones azimuts ou entre régions, avec un objectif de point de restauration (RPO) mesuré en secondes.

Amazon S3 peut être utilisé comme destination de sauvegarde pour les sources de données autogérées et AWS gérées. AWS des services tels qu'Amazon EBSRDS, Amazon et Amazon DynamoDB ont intégré des fonctionnalités permettant de créer des sauvegardes. Vous pouvez aussi utiliser des logiciels de sauvegarde tiers.

Les données sur site peuvent être sauvegardées à l' AWS Cloud aide de AWS Storage Gatewayou AWS DataSync. Les compartiments Amazon S3 peuvent être utilisés pour stocker ces données sur AWS. Amazon S3 propose plusieurs niveaux de stockage tels qu’Amazon S3 Glacier ou S3 Glacier Deep Archive pour réduire les coûts du stockage de données.

Il se peut que vous puissiez répondre aux besoins de récupération de données en reproduisant les données à partir d’autres sources. Par exemple, les nœuds de réplication Amazon ou les ElastiCache répliques de RDS lecture Amazon peuvent être utilisés pour reproduire des données en cas de perte du nœud principal. Dans les cas où de telles sources peuvent être utilisées pour atteindre votre objectif de point de restauration (RPO) et votre objectif de temps de restauration (RTO), il se peut que vous n'ayez pas besoin de sauvegarde. Autre exemple : si vous travaillez avec AmazonEMR, il n'est peut-être pas nécessaire de sauvegarder votre HDFS banque de données, tant que vous pouvez reproduire les données dans Amazon EMR à partir d'Amazon S3.

Lors de la sélection d’une stratégie de sauvegarde, tenez compte du temps nécessaire pour récupérer les données. Le temps nécessaire pour récupérer les données dépend du type de sauvegarde (dans le cas d’une stratégie de sauvegarde) ou de la complexité du mécanisme de reproduction des données. Ce temps doit être conforme à la RTO charge de travail.

Étapes d’implémentation

1. Identifiez toutes les sources de données pour la charge de travail. Les données peuvent être stockées sur un certain nombre de ressources telles que les bases de données, les volumes, les

systèmes de fichiers, les systèmes de journalisation et le stockage d'objets. Reportez-vous à la section Ressources pour trouver des documents connexes sur les différents AWS services où les données sont stockées et sur la capacité de sauvegarde que ces services fournissent.

2. Classez les sources de données en fonction de leur ordre d'importance. Différents jeux de données ont différents niveaux d'importance pour une charge de travail, et donc différentes exigences en matière de résilience. Par exemple, certaines données peuvent être critiques et nécessiter une valeur RPO proche de zéro, tandis que d'autres données peuvent être moins critiques et peuvent tolérer une perte de données plus élevée RPO et une certaine perte de données. De même, différents ensembles de données peuvent également avoir RTO des exigences différentes.
3. Utilisez AWS des services tiers pour créer des sauvegardes des données. AWS Backupest un service géré qui permet de créer des sauvegardes de différentes sources de données sur AWS. AWS Elastic Disaster Recoverygère la réplication automatique des données en moins d'une seconde vers un Région AWS. La plupart AWS des services disposent également de fonctionnalités natives permettant de créer des sauvegardes. AWS Marketplace Il propose également de nombreuses solutions qui offrent ces fonctionnalités. Consultez Ressources ci-dessous pour découvrir comment créer des sauvegardes de données à partir de divers services AWS .
4. Pour les données non sauvegardées, définissez un mécanisme de reproduction des données. Vous pouvez choisir de ne pas sauvegarder les données qui peuvent être reproduites à partir d'autres sources pour diverses raisons. Il peut arriver qu'il soit moins coûteux de reproduire des données à partir de sources en cas de besoin plutôt que de créer une sauvegarde, car le stockage des sauvegardes peut impliquer un coût. Autre exemple : la restauration à partir d'une sauvegarde prend plus de temps que la reproduction des données à partir des sources, ce qui entraîne une violation. RTO Dans de telles situations, envisagez les avantages et inconvénients de chaque approche et définissez un processus clair sur la façon dont les données peuvent être reproduites à partir de ces sources lorsque la récupération des données est nécessaire. Par exemple, si vous avez chargé des données depuis Amazon S3 dans un entrepôt de données (comme Amazon Redshift) ou dans un MapReduce cluster (comme AmazonEMR) pour analyser ces données, il peut s'agir d'un exemple de données pouvant être reproduites à partir d'autres sources. Tant que les résultats de ces analyses sont stockés quelque part ou sont reproductibles, vous ne subirez aucune perte de données en cas de défaillance de l'entrepôt de données ou du MapReduce cluster. Parmi les autres exemples qui peuvent être reproduits à partir de sources, citons les caches (comme Amazon ElastiCache) ou les répliques de RDS lecture.
5. Spécifiez un rythme de sauvegarde des données. La création de sauvegardes des sources de données est un processus périodique dont la fréquence doit dépendre duRPO.

Niveau d’effort du plan d’implémentation : modéré

Ressources

Bonnes pratiques associées :

REL13-BP01 Définir les objectifs de reprise en termes de durée d’indisponibilité et de perte de données

REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise

Documents connexes :

- Qu'est-ce que c'est AWS Backup ?
- Qu'est-ce que c'est AWS DataSync ?
- Qu’est-ce que la sauvegarde en volumes ?
- APNPartenaire : partenaires qui peuvent vous aider en matière de sauvegarde
- AWS Marketplace : produits pouvant être utilisés pour la sauvegarde
- EBSInstantanés Amazon
- Sauvegarde d'Amazon EFS
- Sauvegarde du serveur de fichiers Amazon FSx pour Windows
- Backup and Restore ElastiCache pour Redis
- Création d’un instantané de cluster de bases de données dans Neptune
- Création d’un instantané de base de données
- Création d'une EventBridge règle qui se déclenche selon un calendrier
- Réplication entre Régions avec Amazon S3
- EFS-à- EFS AWS Backup
- Exporter les données du journal vers Amazon S3
- Gestion du cycle de vie des objets
- Sauvegarde et restauration à la demande pour DynamoDB
- oint-in-timeRestauration IP pour DynamoDB
- Utilisation des instantanés d'Amazon OpenSearch Service Index
- Qu'est-ce que c'est AWS Elastic Disaster Recovery ?

Vidéos connexes :

- AWS re:Invent 2021 - Backup, reprise après sinistre et protection contre les ransomwares avec AWS
- AWS Backup Démo : Sauvegarde entre comptes et entre régions
- AWS re:Invent 2019 : Plongez en profondeur AWS Backup, ft. Rackspace () STG341

Exemples connexes :

- Well-Architected Lab - Implémentation de la réplication bidirectionnelle entre régions () pour Amazon S3 CRR
- Atelier Well-Architected : test de la sauvegarde et de la restauration de données
- Atelier Well-Architected : sauvegarde et restauration avec failback pour la charge de travail d'analyse
- Atelier Well-Architected : reprise après sinistre - sauvegarde et restauration

REL09-BP02 Sauvegardes sécurisées et cryptées

Contrôlez et détectez l'accès aux sauvegardes à l'aide de l'authentification et de l'autorisation. Assurez la prévention et détectez si l'intégrité des données des sauvegardes est compromise à l'aide du chiffrement.

Anti-modèles courants :

- Avoir le même accès aux sauvegardes et à l'automatisation de la restauration que vous le faites pour les données.
- Absence de chiffrement de vos sauvegardes.

Avantages du respect de cette bonne pratique : la sécurisation de vos sauvegardes empêche la falsification des données. De même, le chiffrement des données empêche l'accès à ces données si elles sont accidentellement exposées.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Contrôlez et détectez l'accès aux sauvegardes à l'aide d'une authentification et d'une autorisation, telles que AWS Identity and Access Management (IAM). Assurez la prévention et détectez si l'intégrité des données des sauvegardes est compromise à l'aide du chiffrement.

Amazon S3 prend en charge plusieurs méthodes de chiffrement de vos données au repos. Grâce au chiffrement côté serveur, Amazon S3 accepte vos objets sous forme de données non chiffrées, puis les chiffre lors de leur stockage. Avec le chiffrement côté client, votre application de charge de travail s'occupe du chiffrement des données avant leur transmission à Amazon S3. Les deux méthodes vous permettent d'utiliser AWS Key Management Service (AWS KMS) pour créer et stocker la clé de données, ou vous pouvez fournir votre propre clé, dont vous êtes alors responsable. À l'aide de AWS KMS, vous pouvez définir des politiques indiquant IAM qui peut ou ne peut pas accéder à vos clés de données et à vos données déchiffrées.

Pour AmazonRDS, si vous avez choisi de chiffrer vos bases de données, vos sauvegardes sont également cryptées. Les sauvegardes DynamoDB sont toujours chiffrées. Lors de l'utilisation AWS Elastic Disaster Recovery, toutes les données en transit et au repos sont cryptées. Avec Elastic Disaster Recovery, les données au repos peuvent être chiffrées à l'aide de la clé de EBS chiffrement Amazon Encryption Volume Encryption par défaut ou d'une clé personnalisée gérée par le client.

Étapes d'implémentation

1. Utilisez le chiffrement sur chacun de vos magasins de données. La sauvegarde est également chiffrée si vos données sources le sont.
   - Utilisez le chiffrement sur AmazonRDS. . Vous pouvez configurer le chiffrement au repos AWS Key Management Service lors de la création d'une RDS instance.
   - Utilisez le chiffrement sur les EBS volumes Amazon. . Vous pouvez configurer le chiffrement par défaut ou spécifier une clé unique lors de la création du volume.
   - Utilisez le chiffrement Amazon DynamoDB requis. DynamoDB chiffre toutes les données au repos. Vous pouvez utiliser une AWS KMS clé AWS détenue ou une KMS clé AWS gérée, en spécifiant une clé stockée dans votre compte.
   - Chiffrez vos données stockées sur Amazon EFS. Configurez le chiffrement lorsque vous créez votre système de fichiers.
   - Configurez le chiffrement dans les régions source et de destination. Vous pouvez configurer le chiffrement au repos dans Amazon S3 à l'aide des clés stockées dansKMS, mais celles-ci sont spécifiques à chaque région. Vous pouvez spécifier les clés de destination lorsque vous configurez la réplication.
   - Choisissez d'utiliser le EBSchiffrement Amazon par défaut ou personnalisé pour Elastic Disaster Recovery. Cette option permet de chiffrer les données au repos répliquées sur les disques du sous-réseau de la zone de transit et sur les disques répliqués.

2. Mettez en œuvre les autorisations de moindre privilège pour accéder à vos sauvegardes. Suivez les bonnes pratiques pour limiter l'accès aux sauvegardes, instantanés et réplicas conformément aux bonnes pratiques de sécurité.

Ressources

Documents connexes :

- AWS Marketplace : produits pouvant être utilisés pour la sauvegarde
- Amazon EBS Encryption
- Amazon S3 : protection des données à l'aide du chiffrement
- CRRConfiguration supplémentaire : réplication d'objets créés avec le chiffrement côté serveur (SSE) à l'aide de clés de chiffrement stockées dans AWS KMS
- Chiffrement de DynamoDB au repos
- Chiffrer les ressources Amazon RDS
- Chiffrement des données et des métadonnées sur Amazon EFS
- Chiffrement des sauvegardes dans AWS
- Gestion des tables chiffrées
- Pilier de sécurité - AWS Well-Architected Framework
- Qu'est-ce que c'est AWS Elastic Disaster Recovery ?

Exemples connexes :

- Well-Architected Lab - Implémentation de la réplication bidirectionnelle entre régions () pour Amazon S3 CRR

REL09-BP03 Effectuer une sauvegarde automatique des données

Configurez les sauvegardes à effectuer automatiquement en fonction d'un calendrier périodique basé sur l'objectif du point de restauration (RPO) ou sur les modifications apportées au jeu de données. Les jeux de données critiques dont le seuil de tolérance pour la perte de données est faible doivent être sauvegardés automatiquement et fréquemment, tandis que les données moins critiques où certaines données peuvent être perdues peuvent être sauvegardées moins fréquemment.

Résultat souhaité : un processus automatisé qui crée des sauvegardes de sources de données à une cadence établie.

Anti-modèles courants :

- Exécution manuelle des sauvegardes.
- Utilisation de ressources qui ont une capacité de sauvegarde, mais sans inclure la sauvegarde dans votre automatisation.

Avantages de la mise en place de cette bonne pratique : l'automatisation des sauvegardes permet de vérifier qu'elles sont effectuées régulièrement en fonction de vos besoins RPO et de vous avertir si elles ne le sont pas.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

AWS Backup peut être utilisé pour créer des sauvegardes de données automatisées de différentes sources de AWS données. Les RDS instances Amazon peuvent être sauvegardées presque en continu toutes les cinq minutes et les objets Amazon S3 peuvent être sauvegardés presque en continu toutes les quinze minutes, ce qui permet une point-in-time restauration (PITR) à un moment précis dans l'historique des sauvegardes. AWS Les autres sources de données, telles que les EBS volumes Amazon, les tables Amazon DynamoDB ou les systèmes de fichiers FSx Amazon AWS Backup , peuvent exécuter des sauvegardes automatiques toutes les heures. Ces services offrent également des fonctionnalités de sauvegarde natives. AWS les services proposant une sauvegarde automatique avec point-in-time restauration incluent Amazon DynamoDB, RDS Amazon et Amazon Keyspaces (pour Apache Cassandra), qui peuvent être restaurés à un moment précis dans l'historique des sauvegardes. La plupart des autres services de stockage de données AWS offrent la possibilité de planifier des sauvegardes périodiques, à une fréquence de sauvegarde pouvant atteindre toutes les heures.

Amazon RDS et Amazon DynamoDB proposent une sauvegarde continue avec restauration. point-in-time Une fois activée, la gestion des versions Amazon S3 est automatique. Amazon Data Lifecycle Manager peut être utilisé pour automatiser la création, la copie et la suppression de EBS snapshots Amazon. Il peut également automatiser la création, la copie, la dépréciation et le désenregistrement d'Amazon Machine Images AMIs () EBS soutenues par Amazon et de leurs instantanés Amazon sous-jacents. EBS

AWS Elastic Disaster Recovery assure une réplication continue au niveau des blocs depuis l'environnement source (sur site ou AWS) vers la région de restauration cible. Point-in-timeLes EBS instantanés Amazon sont automatiquement créés et gérés par le service.

Pour une vue centralisée de l'automatisation et de l'historique de vos sauvegardes, AWS Backup fournit une solution de sauvegarde entièrement gérée et basée sur des règles. Cette solution centralise et automatise la sauvegarde des données sur plusieurs services AWS dans le cloud et sur site à l'aide d' AWS Storage Gateway.

Outre la gestion des versions, Amazon S3 intègre la réplication. L'intégralité du compartiment S3 peut être automatiquement répliquée vers un autre compartiment de la même ou d'une autre Région AWS.

Étapes d'implémentation

1. Identifiez les sources de données qui sont actuellement sauvegardées manuellement. Pour en savoir plus, veuillez consulter REL09-BP01 Identifiez et sauvegardez toutes les données qui doivent être sauvegardées, ou reproduisez les données à partir des sources.
2. Déterminez le correspondant RPO à la charge de travail. Pour en savoir plus, veuillez consulter REL13-BP01 Définir les objectifs de reprise en termes de durée d'indisponibilité et de perte de données.
3. Utilisez une solution de sauvegarde automatisée ou un service géré. AWS Backup est un service entièrement géré qui facilite la centralisation et l'automatisation de la protection des données dans l'ensemble AWS des services, dans le cloud et sur site. En utilisant les plans de sauvegarde dans AWS Backup, créez des règles qui définissent les ressources à sauvegarder, et la fréquence à laquelle ces sauvegardes doivent être créées. Cette fréquence doit être déterminée par la fréquence RPO établie à l'étape 2. Pour obtenir des conseils pratiques sur la façon de créer des sauvegardes automatisées à l'aide de AWS Backup cette méthode, consultez la section Testing Backup and Restore of Data. Des fonctionnalités de sauvegarde natives sont proposées par la plupart des AWS services qui stockent des données. Par exemple, RDS peut être utilisé pour des sauvegardes automatisées avec point-in-time recovery (PITR).
4. Pour les sources de données non prises en charge par une solution de sauvegarde automatisée ou un service géré tel que des sources de données sur site ou des files d'attente de messages, envisagez d'utiliser une solution tierce de confiance pour créer des sauvegardes automatisées. Vous pouvez également créer une automatisation pour ce faire à l'aide du AWS CLI ouSDKs. Vous pouvez utiliser AWS Lambda Functions ou AWS Step Functions définir la logique impliquée dans la création d'une sauvegarde de données, et utiliser Amazon EventBridge pour l'invoquer à une fréquence adaptée à votreRPO.

Niveau d'effort du plan d'implémentation : faible

Ressources

Documents connexes :

- APNPartenaire : partenaires qui peuvent vous aider en matière de sauvegarde
- AWS Marketplace : produits pouvant être utilisés pour la sauvegarde
- Création d'une EventBridge règle qui se déclenche selon un calendrier
- Qu'est-ce que c'est AWS Backup ?
- Qu'est-ce que c'est AWS Step Functions ?
- Qu'est-ce que c'est AWS Elastic Disaster Recovery ?

Vidéos connexes :

- AWS re:Invent 2019 : Plongez en profondeur AWS Backup, ft. Rackspace () STG341

Exemples connexes :

- Atelier Well-Architected : test de la sauvegarde et de la restauration de données

REL09-BP04 Effectuer une restauration périodique des données pour vérifier l'intégrité et les processus de sauvegarde

Vérifiez que la mise en œuvre de votre processus de sauvegarde répond à vos objectifs de temps de restauration (RTO) et de point de restauration (RPO) en effectuant un test de restauration.

Résultat souhaité : Les données issues des sauvegardes sont périodiquement restaurées à l'aide de mécanismes bien définis afin de vérifier que la restauration est possible dans le délai de restauration fixé (RTO) pour la charge de travail. Vérifiez que la restauration à partir d'une sauvegarde aboutit à une ressource contenant les données d'origine sans qu'aucune de celles-ci ne soit corrompue ou inaccessible, et qu'elle entraîne une perte de données conforme à l'objectif du point de restauration (RPO).

Anti-modèles courants :

- Restauration d’une sauvegarde, mais sans interroger ou récupérer des données pour vérifier l’utilisation de la restauration.
- Supposer qu’une sauvegarde existe.

- Supposer que la sauvegarde d’un système est pleinement opérationnelle et que les données peuvent être récupérées à partir de celle-ci.
- En supposant que le délai de restauration ou de restauration des données à partir d'une sauvegarde correspond RTO à la charge de travail.
- En supposant que les données contenues dans la sauvegarde correspondent à la RPO charge de travail
- Effectuez une restauration si nécessaire, sans utiliser de runbook ou en dehors d’une procédure automatisée établie.

Avantages de cette bonne pratique : le test de restauration des sauvegardes permet de vérifier que les données peuvent être restaurées en cas de besoin sans craindre qu'elles soient manquantes ou endommagées, que la restauration et la restauration sont possibles dans le cadre de la RTO charge de travail et que toute perte de données est conforme à la charge RPO de travail.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Tester la fonctionnalité de sauvegarde et de restauration permet de garantir que ces actions peuvent être effectuées pendant une panne. Restaurez périodiquement les sauvegardes vers un nouvel emplacement et exécutez des tests pour vérifier l’intégrité des données. Certains tests courants à effectuer consistent à vérifier si toutes les données sont disponibles, ne sont pas corrompues, sont accessibles et si toute perte de données est conforme à la RPO charge de travail. De tels tests peuvent également aider à déterminer si les mécanismes de restauration sont suffisamment rapides pour s'adapter à la charge de travailRTO.

Vous pouvez ainsi mettre en AWS place un environnement de test et restaurer vos sauvegardes pour évaluer les RPO fonctionnalités, RTO et exécuter des tests sur le contenu et l'intégrité des données.

En outre, Amazon RDS et Amazon DynamoDB point-in-time autorisent la restauration (). PITR Grâce à la sauvegarde continue, vous pouvez restaurer votre jeu de données à l’état dans lequel il était à une date et une heure spécifiées.

Si toutes les données sont disponibles, ne sont pas endommagées, sont accessibles et que toute perte de données est imputable à la RPO charge de travail. De tels tests peuvent également aider à déterminer si les mécanismes de restauration sont suffisamment rapides pour s'adapter à la charge de travailRTO.

AWS Elastic Disaster Recovery propose des instantanés point-in-time de restauration continue des volumes AmazonEBS. Au fur et à mesure que les serveurs source sont répliqués, les point-in-time états sont chroniqués dans le temps en fonction de la politique configurée. Elastic Disaster Recovery vous aide à vérifier l'intégrité de ces instantanés en lançant des instances à des fins de test et d'analyse sans rediriger le trafic.

Étapes d'implémentation

1. Identifiez les sources de données qui sont actuellement sauvegardées et où ces sauvegardes sont stockées. Pour obtenir des conseils de mise en œuvre, consultez REL09-BP01 Identifiez et sauvegardez toutes les données qui doivent être sauvegardées, ou reproduisez les données à partir des sources.
2. Établissez des critères de validation des données pour chaque source de données. Différents types de données ont des propriétés différentes qui pourraient nécessiter des mécanismes de validation distincts. Réfléchissez à la manière dont ces données pourraient être validées avant de vous assurer que vous pouvez les utiliser en production. Certaines méthodes courantes de validation des données consistent à utiliser des propriétés de données et de sauvegarde telles que le type de données, le format, la somme de contrôle, la taille ou une combinaison de ces propriétés avec une logique de validation personnalisée. Par exemple, il peut s'agir d'une comparaison des valeurs de somme de contrôle entre la ressource restaurée et la source de données au moment de la création de la sauvegarde.
3. Établissez RTO et RPO restaurez les données en fonction de leur criticité. Pour obtenir des conseils de mise en œuvre, consultez REL13-BP01 Définir les objectifs de reprise en termes de durée d'indisponibilité et de perte de données.
4. Évaluez votre capacité de récupération. Passez en revue votre stratégie de sauvegarde et de restauration pour déterminer si elle répond à vos RPO attentes, RTO et ajustez-la si nécessaire. À l'aide de AWS Resilience Hub, vous pouvez exécuter une évaluation de votre charge de travail. L'évaluation évalue la configuration de votre application par rapport à la politique de résilience et indique si vos RPO objectifs RTO et vos objectifs peuvent être atteints.
5. Effectuez un test de restauration avec les processus établis utilisés en production pour la restauration des données. Ces processus dépendent de la façon dont la source de données d'origine a été sauvegardée, du format et de l'emplacement de stockage de la sauvegarde elle-même, ou ils varient selon que les données sont reproduites à partir d'autres sources. Par exemple, si vous utilisez un service géré tel que AWS Backup, cela peut être aussi simple que de restaurer la sauvegarde dans une nouvelle ressource. Si vous avez utilisé AWS Elastic Disaster Recovery vous pouvez lancer une simulation de récupération.

6. Validez la récupération des données à partir de la ressource restaurée en fonction des critères que vous avez définis précédemment pour la validation des données. Les données restaurées et récupérées contiennent-elles l'enregistrement/l'élément le plus récent au moment de la sauvegarde ? Ces données correspondent-elles à la RPO charge de travail ?
7. Mesurez le temps nécessaire à la restauration et à la restauration et comparez-le à celui requisRTO. Ce processus s'inscrit-il dans le cadre RTO de la charge de travail ? Par exemple, comparez les horodatages du début du processus de restauration et de la fin de la validation de la récupération pour calculer la durée de ce processus. Tous les AWS API appels sont horodatés et ces informations sont disponibles dans. AWS CloudTrail Bien que ces informations puissent fournir des détails sur le début du processus de restauration, l'horodatage indiquant la fin de la validation doit être enregistré par votre logique de validation. Si vous utilisez un processus automatisé, des services tels qu'Amazon DynamoDB peuvent être utilisés pour stocker ces informations. En outre, de nombreux AWS services fournissent un historique des événements qui fournit des informations horodatées lorsque certaines actions se sont produites. Dans AWS Backup ce cadre, les actions de sauvegarde et de restauration sont appelées tâches, et ces tâches contiennent des informations d'horodatage dans le cadre de leurs métadonnées, qui peuvent être utilisées pour mesurer le temps nécessaire à la restauration et à la restauration.
8. Informez les parties prenantes si la validation des données échoue ou si le temps nécessaire à la restauration et à la restauration dépasse le délai fixé RTO pour la charge de travail. Lors de la mise en œuvre de l'automatisation à cette fin, comme dans cet atelier, des services tels qu'Amazon Simple Notification Service (AmazonSNS) peuvent être utilisés pour envoyer des notifications push telles que des e-mails ou SMS aux parties prenantes. Ces messages peuvent également être publiés sur des applications de messagerie telles qu'Amazon Chime, Slack ou Microsoft Teams ou utilisés pour créer des tâches, comme dans le cas de Systems OpsItems Manager AWS. OpsCenter
9. Automatisez ce processus pour qu'il s'exécute périodiquement. Par exemple, des services tels AWS Lambda qu'un State Machine in AWS Step Functions peuvent être utilisés pour automatiser les processus de restauration et de restauration, et Amazon EventBridge peut être utilisé pour invoquer régulièrement ce flux de travail d'automatisation, comme indiqué dans le schéma d'architecture ci-dessous. Découvrez comment automatiser la validation de la récupération de données avec AWS Backup. De plus, cet atelier Well-Architected apporte une expérience pratique sur une façon d'automatiser plusieurs des étapes indiquées ici.

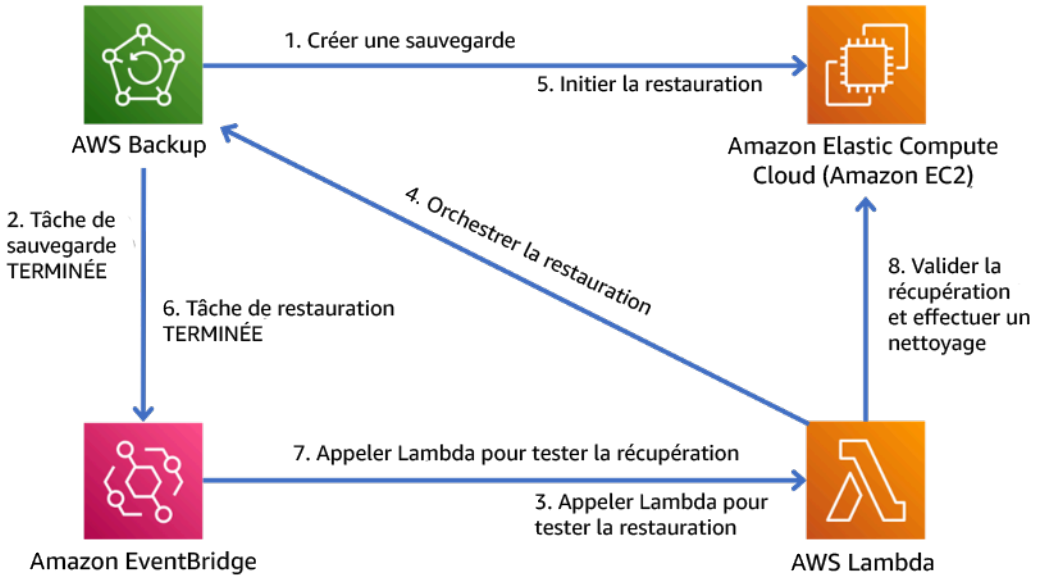


Figure 9. Un processus de sauvegarde et de restauration automatisé

Niveau d'effort pour le plan de mise en œuvre : modéré à élevé selon la complexité des critères de validation.

Ressources

Documents connexes :

- Automatisez la validation de la récupération des données avec AWS Backup
- APNPartenaire : partenaires qui peuvent vous aider en matière de sauvegarde
- AWS Marketplace : produits pouvant être utilisés pour la sauvegarde
- Création d'une EventBridge règle qui se déclenche selon un calendrier
- Sauvegarde et restauration à la demande pour DynamoDB
- Qu'est-ce que c'est AWS Backup ?
- Qu'est-ce que c'est AWS Step Functions ?
- Qu'est-ce que AWS Elastic Disaster Recovery
- AWS Elastic Disaster Recovery

Exemples connexes :

- Atelier Well-Architected : test de la sauvegarde et de la restauration de données

FIA 10. Comment utiliser l’isolation des pannes pour protéger votre charge de travail ?

L’isolation des défaillances limite l’impact de la défaillance d’un composant ou d’un système à une limite définie. Si l’isolation est correcte, les composants situés en dehors de cette limite ne sont pas affectés par la défaillance. L’exécution de votre charge de travail au-delà de plusieurs limites d’isolation des défaillances peut la rendre plus résistante aux défaillances.

Bonnes pratiques

- REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements
- REL10-BP02 Automatiser la récupération des composants limités à un seul emplacement
- REL10-BP03 Utiliser des architectures cloisonnées pour limiter la portée de l’impact

REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements

Distribuez les données et les ressources de charge de travail sur plusieurs zones de disponibilité ou, si nécessaire, entre Régions AWS.

L’un des principes fondamentaux de la conception de services dans AWS est d’éviter les points de défaillance uniques, y compris l’infrastructure physique sous-jacente. AWS fournit des ressources et des services de cloud computing à l’échelle mondiale, sur plusieurs sites géographiques appelés régions. Chaque région est physiquement et logiquement indépendante et se compose de trois zones de disponibilité (AZ) ou plus. Les zones de disponibilité sont géographiquement proches les unes des autres, mais sont physiquement séparées et isolées. La répartition de vos charges de travail entre les zones de disponibilité et les régions vous permet de réduire le risque de menaces telles que les incendies, les inondations, les catastrophes météorologiques, les tremblements de terre et les erreurs humaines.

Créez une stratégie de localisation pour assurer une haute disponibilité adaptée à vos charges de travail.

Résultat escompté : les charges de travail de production sont réparties entre plusieurs zones de disponibilité (AZ) ou régions afin de garantir la tolérance aux pannes et la haute disponibilité.

Anti-modèles courants :

- Votre charge de travail de production n’existe que dans une seule zone de disponibilité.
- Vous mettez en œuvre une architecture multirégionale alors qu’une architecture multi-AZ répondrait aux exigences.

- Vos déploiements ou vos données sont désynchronisés, ce qui entraîne une dérive de la configuration ou une sous-réplication des données.
- Vous ne tenez pas compte des dépendances entre les composants de l'application si les exigences en matière de résilience et de multi-localisation diffèrent entre ces composants.

Avantages liés au respect de cette bonne pratique :

- Votre charge de travail est plus résiliente face aux incidents, tels que les pannes d'alimentation, les défaillances de contrôle environnemental, les catastrophes naturelles, les pannes de service en amont ou les problèmes réseau affectant une zone de disponibilité ou une région entière.
- Vous pouvez accéder à un inventaire plus large d'instances Amazon EC2 et réduire le risque d'exceptions InsufficientCapacityExceptions (ICE) lors du lancement de types d'instances EC2 spécifiques.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Déployez et gérez toutes les charges de travail de production dans au moins deux zones de disponibilité (AZ) d'une région.

Utilisation de plusieurs zones de disponibilité

Les zones de disponibilité sont des lieux d'hébergement de ressources qui sont physiquement séparés les uns des autres afin d'éviter les défaillances corrélées dues à des risques tels que les incendies, les inondations et les tornades. Chaque zone de disponibilité possède une infrastructure physique indépendante comprenant des connexions électriques, des sources d'alimentation de secours, des services mécaniques et une connectivité réseau. Cette disposition limite les défaillances d'un de ces composants à la seule zone de disponibilité affectée. Par exemple, si un incident à l'échelle de la zone de disponibilité rend les instances EC2 indisponibles dans la zone de disponibilité affectée, vos instances situées dans une autre zone de disponibilité restent disponibles.

Bien que physiquement séparées, les zones de disponibilité situées dans la même Région AWS sont suffisamment proches pour fournir une mise en réseau à haut débit et à faible latence (moins de dix millisecondes). Vous pouvez répliquer les données de manière synchrone entre les zones de disponibilité pour la plupart des charges de travail sans affecter de manière significative l'expérience utilisateur. Cela signifie que vous pouvez utiliser les zones de disponibilité d'une région dans une configuration active/active ou active/veille.

Tous les calculs associés à votre charge de travail doivent être répartis entre les différentes zones de disponibilité. Cela inclut les instances Amazon EC2, les tâches AWS Fargate et les fonctions AWS Lambda associées au VPC. Les services de calcul AWS, notamment EC2 Auto Scaling, Amazon Elastic Container Service (ECS) et Amazon Elastic Kubernetes Service (EKS), vous permettent de lancer et de gérer les calculs sur l'ensemble des zones de disponibilité. Configurez-les pour remplacer automatiquement les calculs selon les besoins dans une autre zone de disponibilité afin de maintenir la disponibilité. Pour diriger le trafic vers les zones de disponibilité disponibles, placez un équilibreur de charge devant vos ressources de calcul, tel qu'un Application Load Balancer ou un Network Load Balancer. Les équilibreurs de charge AWS peuvent rediriger le trafic vers les instances disponibles en cas d'altération de la zone de disponibilité.

Vous devez également répliquer les données pour votre charge de travail et les rendre disponibles dans plusieurs zones de disponibilité. Certains services de données AWS gérés, tels qu'Amazon S3, Amazon Elastic File Service (EFS), Amazon Aurora, Amazon DynamoDB, Amazon Simple Queue Service (SQS) et Amazon Kinesis Data Streams répliquent les données dans plusieurs zones de disponibilité par défaut et résistent à l'altération de la zone de disponibilité. Avec d'autres services de données AWS gérés, tels qu'Amazon Relational Database Service (RDS), Amazon Redshift et Amazon ElastiCache, vous devez activer la réplication multi-AZ. Une fois l'option activée, ces services détectent automatiquement une altération de la zone de disponibilité, redirigent les demandes vers une zone de disponibilité disponible et répliquent à nouveau les données selon les besoins après reprise, sans intervention du client. Familiarisez-vous avec le guide de l'utilisateur de chaque service de données AWS géré que vous utilisez pour comprendre ses fonctionnalités, ses comportements et son fonctionnement multi-AZ.

Si vous utilisez un stockage autogéré, tel que les volumes Amazon Elastic Block Store (EBS) ou le stockage d'instances Amazon EC2, vous devez gérer vous-même la réplication multi-AZ.

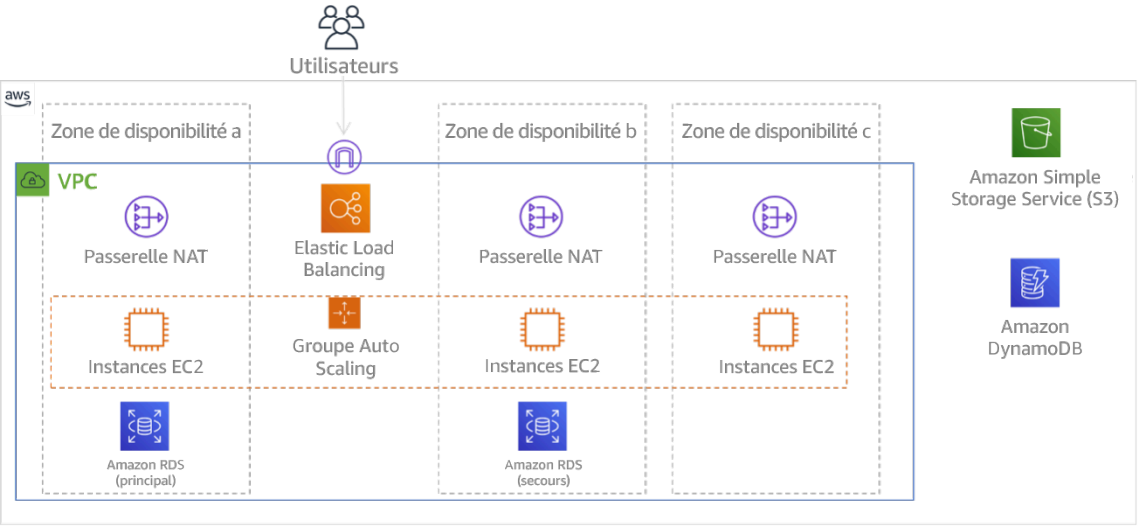


Figure 9 : Architecture multiniveau déployée sur trois zones de disponibilité. Notez qu'Amazon S3 et Amazon DynamoDB comportent toujours automatiquement plusieurs zones de disponibilités. L'ELB est également déployé dans les trois zones.

Utilisation de plusieurs Régions AWS

Si vos charges de travail nécessitent une résilience extrême (telles qu'une infrastructure critique, des applications liées à la santé ou des services répondant à des exigences strictes de disponibilité imposées par le client ou par le législateur), vous pouvez avoir besoin d'une disponibilité supérieure à ce qu'une Région AWS unique peut fournir. Dans ce cas, vous devez déployer et exploiter votre charge de travail sur au moins deux Régions AWS (en supposant que vos exigences en matière de résidence des données le permettent).

Les Régions AWS sont situées dans différentes régions géographiques du monde et sur plusieurs continents. Les Régions AWSprésentent une séparation physique et une isolation encore plus importantes que les zones de disponibilité. À quelques exceptions près, les services AWS profitent de cette conception pour fonctionner de manière totalement indépendante entre les différentes régions (on parle alors de services régionaux). Un service Région AWSal est conçu de manière à ce qu'une défaillance n'ait pas d'impact sur le service dans une autre région.

Lorsque vous gérez votre charge de travail dans plusieurs régions, vous devez prendre en compte des exigences supplémentaires. Les ressources des différentes régions étant séparées et indépendantes les unes des autres, vous devez dupliquer les composants de votre charge de travail dans chaque région. Cela inclut l'infrastructure de base, telle que les VPC, en plus des services de calcul et de données.

REMARQUE : Lorsque vous envisagez une conception multirégionale, vérifiez que votre charge de travail est capable de s'exécuter dans une région unique. Si vous créez des dépendances entre des régions de manière à ce qu'un composant d'une région repose sur des services ou des composants d'une autre région, vous pouvez augmenter le risque de défaillance et affaiblir considérablement votre position en matière de fiabilité.

Pour faciliter les déploiements multirégionaux et maintenir la cohérence, AWS CloudFormation StackSets peut répliquer l'ensemble de votre infrastructure AWS entre plusieurs régions. AWS CloudFormation peut également détecter une dérive de configuration et vous informer lorsque vos ressources AWS d'une région ne sont pas synchronisées. De nombreux services AWS proposent la réplication multirégionale des ressources de charge de travail importantes. Par exemple, EC2 Image Builder peut publier vos images de machine EC2 (AMI) après chaque génération dans chaque région

que vous utilisez. Amazon Elastic Container Registry (ECR) peut répliquer vos images de conteneur dans les régions que vous avez sélectionnées.

Vous devez également répliquer vos données dans chacune des régions que vous avez choisies. De nombreux services de données AWS gérés offrent une capacité de réplication interrégionale, notamment Amazon S3, Amazon DynamoDB, Amazon RDS, Amazon Aurora, Amazon Redshift, Amazon Elasticache, et Amazon EFS. Les tables globales Amazon DynamoDB acceptent les écritures dans toute région prise en charge et répliqueront les données entre toutes vos autres régions configurées. Pour les autres services, vous devez désigner une région principale pour les écritures, car les autres régions contiennent des réplicas en lecture seule. Consultez le guide de l'utilisateur et le manuel du développeur de chaque service de données AWS géré utilisé par votre charge de travail pour comprendre ses capacités et ses limites multirégionales. Accordez une attention particulière à l'endroit où les écritures doivent être dirigées, aux capacités et aux limites transactionnelles, à la manière dont la réplication est effectuée et à la manière de surveiller la synchronisation entre les régions.

AWS permet également d'acheminer de façon très flexible le trafic de demandes vers vos déploiements régionaux. Par exemple, vous pouvez configurer vos enregistrements DNS à l'aide d'Amazon Route 53 pour diriger le trafic vers la région disponible la plus proche de l'utilisateur. Vous pouvez également configurer vos enregistrements DNS dans une configuration active/en veille, dans laquelle vous désignez une région comme principale et ne vous rabattez sur un réplica régional que si la région principale devient défectueuse. Vous pouvez configurer la surveillance de l'état Route 53 pour détecter les points de terminaison défectueux et effectuer un basculement automatique. Vous pouvez également utiliser Amazon Application Recovery Controller (ARC) pour fournir un contrôle de routage hautement disponible permettant de réacheminer manuellement le trafic selon les besoins.

Même si vous choisissez de ne pas opérer dans plusieurs régions pour des raisons de haute disponibilité, considérez plusieurs régions dans le cadre de votre stratégie de reprise après sinistre (DR). Si possible, répliquez les composants et les données de l'infrastructure de votre charge de travail dans une configuration de secours à chaud ou d'environnement en veille dans une région secondaire. Dans cette conception, vous répliquez l'infrastructure de base de la région principale, telle que les VPC, les groupes Auto Scaling, les orchestrateurs de conteneurs et d'autres composants, mais vous configurez les composants de taille variable dans la région de secours (tels que le nombre d'instances EC2 et de réplicas de base de données) de manière à ce qu'ils aient une taille minimale exploitable. Vous organisez également une réplication continue des données de la région principale vers la région de secours. En cas d'incident, vous pouvez augmenter horizontalement ou accroître les ressources de la région de secours, puis la promouvoir en région principale.

Étapes d’implémentation

1. Travaillez avec les parties prenantes de l’entreprise et les experts en résidence des données pour déterminer les Régions AWS qui peuvent être utilisées pour héberger vos ressources et vos données.
2. Travaillez avec les parties prenantes techniques et commerciales pour évaluer votre charge de travail et déterminer si ses besoins de résilience peuvent être satisfaits par une approche multi-AZ (une seule Région AWS) ou s’ils nécessitent une approche multirégionale (si plusieurs régions sont autorisées). L’utilisation de plusieurs régions permet de bénéficier d’une plus grande disponibilité, mais peut entraîner une complexité et des coûts supplémentaires. Tenez compte des facteurs suivants dans votre évaluation :
    a. Objectifs commerciaux et exigences des clients : quelle est la durée d’indisponibilité autorisée en cas d’incident affectant la charge de travail dans une zone de disponibilité ou une région ? Évaluez vos objectifs de point de récupération tels qu’ils sont présentés dans REL13-BP01 Définir les objectifs de reprise en termes de durée d’indisponibilité et de perte de données.
    b. Exigences relatives à la reprise après sinistre (DR) : contre quel type de sinistre potentiel souhaitez-vous vous assurer ? Envisagez la possibilité d’une perte de données ou d’une indisponibilité à long terme à différents niveaux d’impact, d’une simple zone de disponibilité à une région entière. Si vous répliquez des données et des ressources entre des zones de disponibilité et qu’une seule zone de disponibilité connaît une défaillance prolongée, vous pouvez récupérer le service dans une autre zone de disponibilité. Si vous répliquez des données et des ressources entre plusieurs régions, vous pouvez récupérer le service dans une autre région.
3. Déployez vos ressources de calcul dans plusieurs zones de disponibilité.
    a. Dans votre VPC, créez plusieurs sous-réseaux dans des zones de disponibilité différentes. Configurez chacune d’elles de manière à ce qu’elle soit suffisamment grande pour accueillir les ressources nécessaires pour répondre à la charge de travail, même en cas d’incident. Pour plus d’informations, consultez REL02-BP03 S’assurer que l’allocation des sous-réseaux IP tient compte de l’expansion et de la disponibilité.
    b. Si vous utilisez des instances Amazon EC2, utilisez EC2 Auto Scaling pour gérer vos instances. Spécifiez les sous-réseaux que vous avez choisis à l’étape précédente lorsque vous créez vos groupes Auto Scaling.
    c. Si vous utilisez le calcul AWS Fargate pour Amazon ECS ou Amazon EKS, sélectionnez les sous-réseaux que vous avez choisis à la première étape lors de la création d’un service ECS, lancez une tâche ECS ou créez un profil Fargate pour EKS.

   d. Si vous utilisez des fonctions AWS Lambda qui doivent être exécutées dans votre VPC, sélectionnez les sous-réseaux que vous avez choisis à la première étape lors de la création de la fonction Lambda. Pour toutes les fonctions qui n'ont pas de configuration VPC, AWS Lambda gère automatiquement la disponibilité pour vous.

   e. Placez des redirecteurs de trafic tels que des équilibreurs de charge devant vos ressources de calcul. Si l'équilibrage de charge entre zones est activé, les équilibreurs AWS Application Load Balancers et Network Load Balancers détectent quand des cibles telles que des instances et des conteneurs EC2 sont inaccessibles en raison d'une altération de la zone de disponibilité et redirigent le trafic vers des cibles situées dans des zones de disponibilité saines. Si vous désactivez l'équilibrage de charge entre zones, utilisez Amazon Application Recovery Controller (ARC) pour fournir une fonctionnalité de changement de zone. Si vous utilisez un équilibreur de charge tiers ou si vous avez implémenté vos propres équilibreurs de charge, configurez-les avec plusieurs front ends répartis dans différentes zones de disponibilité.

4. Répliquez les données de votre charge de travail sur plusieurs zones de disponibilité.

   a. Si vous utilisez un service de données AWS géré tel qu'Amazon RDS, Amazon ElastiCache ou Amazon FSx, étudiez son guide de l'utilisateur pour comprendre ses capacités de réplication de données et de résilience. Activez la réplication et le basculement entre zones de disponibilité si nécessaire.

   b. Si vous utilisez des services de stockage AWS gérés tels qu'Amazon S3, Amazon EFS et Amazon FSx, évitez d'utiliser des configurations mono-AZ ou à zone unique pour des données qui requièrent une durabilité élevée. Utilisez une configuration multi-AZ pour ces services. Consultez le guide de l'utilisateur du service correspondant pour déterminer si la réplication multi-AZ est activée par défaut ou si vous devez l'activer.

   c. Si vous exécutez une base de données, une file d'attente ou un autre service de stockage autogéré, organisez la réplication multi-AZ conformément aux instructions ou aux bonnes pratiques de l'application. Familiarisez-vous avec les procédures de basculement de votre application.

5. Configurez votre service DNS pour détecter une altération de la zone de disponibilité et rediriger le trafic vers une zone de disponibilité saine. Amazon Route 53, lorsqu'il est utilisé en combinaison avec des Elastic Load Balancers, peut le faire automatiquement. Route 53 peut également être configuré avec des enregistrements de basculement qui utilisent la surveillance de l'état pour répondre aux requêtes avec uniquement des adresses IP saines. Pour tous les enregistrements DNS utilisés pour le basculement, spécifiez une faible valeur de durée de vie (TTL) (par exemple, 60 secondes ou moins) afin d'éviter que la mise en cache des enregistrements n'entrave la reprise (les enregistrements d'alias Route 53 fournissent des durées de vie (TTL) appropriées pour vous).

Étapes supplémentaires lors de l'utilisation de plusieurs Régions AWS

1. Répliquez l'ensemble du code d'application et de système d'exploitation (OS) utilisé par votre charge de travail dans les régions que vous avez sélectionnées. Répliquez les images Amazon Machine Image (AMI) utilisées par vos instances EC2, si nécessaire, à l'aide de solutions telles qu'Amazon EC2 Image Builder. Répliquez les images de conteneur stockées dans des registres à l'aide de solutions telles que la réplication entre régions Amazon ECR. Activez la réplication régionale pour tous les compartiments Amazon S3 utilisés pour stocker les ressources d'application.
2. Déployez vos ressources de calcul et vos métadonnées de configuration (telles que les paramètres stockés dans AWS Systems Manager Parameter Store) dans plusieurs régions. Utilisez les mêmes procédures que celles décrites dans les étapes précédentes, mais répliquez la configuration pour chaque région que vous utilisez pour votre charge de travail. Utilisez des solutions d'infrastructure en tant que code, telles qu'AWS CloudFormation pour reproduire uniformément les configurations entre les régions. Si vous utilisez une région secondaire dans une configuration d'environnement en veille pour la reprise après sinistre, vous pouvez réduire le nombre de vos ressources de calcul à une valeur minimale afin de réduire les coûts, avec une augmentation correspondante du temps de reprise.
3. Répliquez vos données de votre région principale vers vos régions secondaires.
    a. Les tables globales Amazon DynamoDB fournissent des réplicas globaux de vos données sur lesquels vous pouvez écrire depuis n'importe quelle région prise en charge. Avec d'autres services de données AWS gérés, tels qu'Amazon RDS, Amazon Aurora et Amazon Elasticache, vous désignez une région principale (lecture/écriture) et des régions de réplica (lecture seule). Consultez les guides de l'utilisateur et les manuels du développeur des services respectifs pour plus de détails sur la réplication régionale.
    b. Si vous exécutez une base de données autogérée, organisez la réplication multirégionale conformément aux instructions ou aux bonnes pratiques de l'application. Familiarisez-vous avec les procédures de basculement de votre application.
    c. Si votre charge de travail utilise AWS EventBridge, vous devrez peut-être transférer certains événements de votre région principale vers vos régions secondaires. Pour ce faire, spécifiez les bus d'événements dans vos régions secondaires comme cibles pour les événements correspondants dans votre région principale.
4. Déterminez si et dans quelle mesure vous souhaitez utiliser des clés de chiffrement identiques entre les régions. Une approche standard conciliant sécurité et facilité d'utilisation consiste à utiliser des clés régionales pour les données et l'authentification locales d'une région, et à utiliser des clés globales pour le chiffrement des données répliquées entre différentes régions. AWS

Key Management Service (KMS) prend en charge les clés multirégionales pour répartir en toute sécurité et protéger les clés partagées entre les régions.

5. Envisagez d'utiliser AWS Global Accelerator pour améliorer la disponibilité de votre application en dirigeant le trafic vers les régions qui contiennent des points de terminaison sains.

Ressources

Bonnes pratiques associées :

- REL02-BP03 S'assurer que l'allocation des sous-réseaux IP tient compte de l'expansion et de la disponibilité
- REL11-BP05 Utiliser la stabilité statique pour éviter les comportements bimodaux
- REL13-BP01 Définir les objectifs de reprise en termes de durée d'indisponibilité et de perte de données

Documents connexes :

- Infrastructure mondiale AWS
- Livre blanc : Limites d'isolation des défaillances des services AWS
- Résilience dans Amazon EC2 Auto Scaling
- Amazon EC2 Auto Scaling : exemple : répartition des instances entre les zones de disponibilité
- Fonctionnement d'EC2 Image Builder
- Comment Amazon ECS place les tâches sur les instances de conteneur (y compris Fargate)
- Résilience dans AWS Lambda
- Amazon S3 : présentation de la réplication d'objets
- Réplication d'images privées sur Amazon ECR
- Tables globales : réplication multirégion avec DynamoDB
- Amazon Elasticache for Redis OSS : réplication entre Régions AWS à l'aide d'entrepôts de données globaux
- Résilience dans Amazon RDS
- Utilisation de bases de données globales Amazon Aurora
- Manuel du développeur AWS Global Accelerator
- Clés multi-régions dans AWS KMS

- Amazon Route 53 : configuration du basculement DNS
- Manuel du développeur Amazon Application Recovery Controller (ARC)
- Envoi et réception d'événements Amazon EventBridge entre régions Régions AWS
- Série de blog sur la création d'une application multirégion avec les services AWS
- Architecture de reprise après sinistre (DR) sur AWS, partie I : stratégies de reprise dans le cloud
- Architecture de reprise après sinistre sur AWS, partie III : Environnement en veille et secours à chaud

Vidéos connexes :

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications
- AWS re:Invent 2019: Innovation and operation of the AWS global network infrastructure

REL10-BP02 Automatiser la récupération des composants limités à un seul emplacement

Si les composants de la charge de travail ne peuvent s'exécuter que dans une seule zone de disponibilité ou un centre de données sur site, implémentez la capacité permettant d'effectuer une reconstruction complète de la charge de travail dans le cadre de vos objectifs de reprise définis.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Si la bonne pratique de déploiement de la charge de travail sur plusieurs emplacements n'est pas possible en raison de contraintes technologiques, vous devez implémenter une autre solution de résilience. Vous devez automatiser la possibilité de recréer l'infrastructure nécessaire, de redéployer les applications et de recréer les données nécessaires pour ces situations.

Par exemple, Amazon EMR lance tous les nœuds d'un cluster donné dans la même zone de disponibilité, car l'exécution d'un cluster dans la même zone améliore les performances des flux de travail en fournissant un taux d'accès aux données plus élevé. Si ce composant est requis pour la résilience de la charge de travail, vous devez pouvoir redéployer le cluster et ses données. De même, pour Amazon EMR, vous devez assurer la redondance autrement qu'en utilisant plusieurs zones de disponibilité. Vous pouvez passer par plusieurs nœuds. Avec le système de fichiers EMR (EMRFS), les données EMR peuvent être conservées dans Amazon S3, et ainsi être répliquées sur plusieurs zones de disponibilité ou Régions AWS.

De même, pour Amazon Redshift, il met en service, par défaut, votre cluster dans une zone de disponibilité sélectionnée de façon aléatoire au sein de la Région AWS que vous sélectionnez. Tous les nœuds de cluster sont provisionnés dans la même zone.

Pour les charges de travail basées sur des serveurs avec état déployés dans un centre de données sur site, vous pouvez utiliser AWS Elastic Disaster Recovery pour protéger vos charges de travail dans AWS. Si vous êtes déjà hébergé dans AWS, vous pouvez utiliser Elastic Disaster Recovery pour protéger votre charge de travail dans une autre zone de disponibilité ou région. Elastic Disaster Recovery utilise une réplication continue au niveau des blocs vers une zone de stockage légère afin de fournir une récupération rapide et fiable des applications sur site et dans le cloud.

Étapes d'implémentation

1. Implémentation de l'autorégénération. Dans la mesure du possible, déployez vos instances ou vos conteneurs en utilisant la mise à l'échelle automatique. Si vous ne pouvez pas utiliser la mise à l'échelle automatique, utilisez la récupération automatique pour les instances EC2 ou mettez en place un mécanisme d'autoréparation basé sur Amazon EC2 ou des événements de cycle de vie de conteneur ECS.
   - Utilisez les groupes Amazon EC2 Auto Scaling pour les instances et les charges de travail de conteneur qui n'ont aucune exigence en matière d'adresse IP d'instance, d'adresse IP privée, d'adresse IP élastique et de métadonnées d'instance.
     - Les données utilisateur du modèle de lancement peuvent être utilisées pour mettre en place un mécanisme permettant la récupération automatique de la plupart des charges de travail.
   - Utilisez la récupération automatique des instances Amazon EC2 pour les charges de travail nécessitant une seule adresse d'ID d'instance, une adresse IP privée, une adresse IP élastique et les métadonnées d'instance.
     - La récupération automatique envoie des alertes de statut de récupération à une rubrique SNS lorsque la défaillance de l'instance est détectée.
   - Utilisez les événements du cycle de vie de l'instance Amazon EC2 ou les événements Amazon ECS pour automatiser l'autoréparation lorsque la mise à l'échelle automatique ou la récupération de votre instance EC2 ne peuvent pas être utilisées.
     - Utilisez les événements pour invoquer le mécanisme vous permettant de réparer votre composant selon la logique de processus dont vous avez besoin.
   - Protégez les charges de travail avec état limitées à un seul emplacement à l'aide de AWS Elastic Disaster Recovery.

Ressources

Documents connexes :

- Événements Amazon ECS
- Hooks de cycle de vie Amazon EC2 Auto Scaling
- Récupération de votre instance.
- Mise à l'échelle automatique des services
- Qu'est-ce qu'Amazon EC2 Auto Scaling ?
- AWS Elastic Disaster Recovery

REL10-BP03 Utiliser des architectures cloisonnées pour limiter la portée de l'impact

Mettez en œuvre des architectures de cloisonnement (également connues sous le nom d'architectures cellulaires) pour restreindre l'effet d'une panne au sein d'une charge de travail à un nombre limité de composants.

Résultat escompté : une architecture cellulaire utilise plusieurs instances isolées d'une charge de travail, chaque instance étant appelée cellule. Chaque cellule est indépendante, ne partage pas d'état avec les autres cellules et traite un sous-ensemble des demandes de la charge de travail globale. Cela réduit l'impact potentiel d'une panne, telle qu'une mauvaise mise à jour logicielle, sur une cellule individuelle et les demandes qu'elle traite. Si une charge de travail utilise 10 cellules pour traiter 100 demandes, lorsqu'une panne survient, 90 % des demandes globales ne sont pas affectées par la panne.

Anti-modèles courants :

- Permettre aux cellules de se développer sans limites.
- Appliquer des mises à jour ou des déploiements de code à toutes les cellules en même temps.
- Partage de l'état ou des composants entre les cellules (à l'exception de la couche routeur).
- Ajout d'une logique métier ou de routage complexe à la couche routeur.
- Ne pas minimiser les interactions entre les cellules.

Avantages du respect de cette bonne pratique : avec les architectures cellulaires, de nombreux types de défaillances courants sont contenus dans la cellule elle-même, ce qui permet une isolation

supplémentaire des pannes. Ces limites de défaillances peuvent apporter de la résilience face à des types de défaillances difficiles à contenir, tels que des déploiements de code infructueux ou des demandes corrompues ou invoquant un mode de défaillance spécifique (également appelées demandes de pilules empoisonnées).

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Sur un navire, les cloisons permettent de contenir une brèche dans la coque dans une seule section de la coque. Dans les systèmes complexes, ce modèle est souvent répliqué pour permettre d’isoler des pannes. Les limites isolées pour les défaillances restreignent l’effet d’une panne au sein d’une charge de travail à un nombre limité de composants. Les composants situés en dehors du périmètre ne sont pas affectés par la défaillance. En utilisant plusieurs périmètres d’isolation des pannes, vous pouvez limiter l’impact sur votre charge de travail. Sur AWS, les clients peuvent utiliser plusieurs zones de disponibilité et régions pour isoler des pannes, mais le concept d’isolement des pannes peut également être étendu à l’architecture de votre charge de travail.

La charge de travail globale est divisée en cellules par une clé de partition. Cette clé doit s’aligner sur la base de granularité du service, ou sur la manière naturelle dont la charge de travail d’un service peut être subdivisée avec un minimum d’interactions entre les cellules. Des exemples de clés de partition sont l’ID du client, l’ID de la ressource ou tout autre paramètre facilement accessible dans la plupart des appels d’API. Une couche de routage des cellules distribue les requêtes aux cellules individuelles en fonction de la clé de partition et présente un point de terminaison unique aux clients.

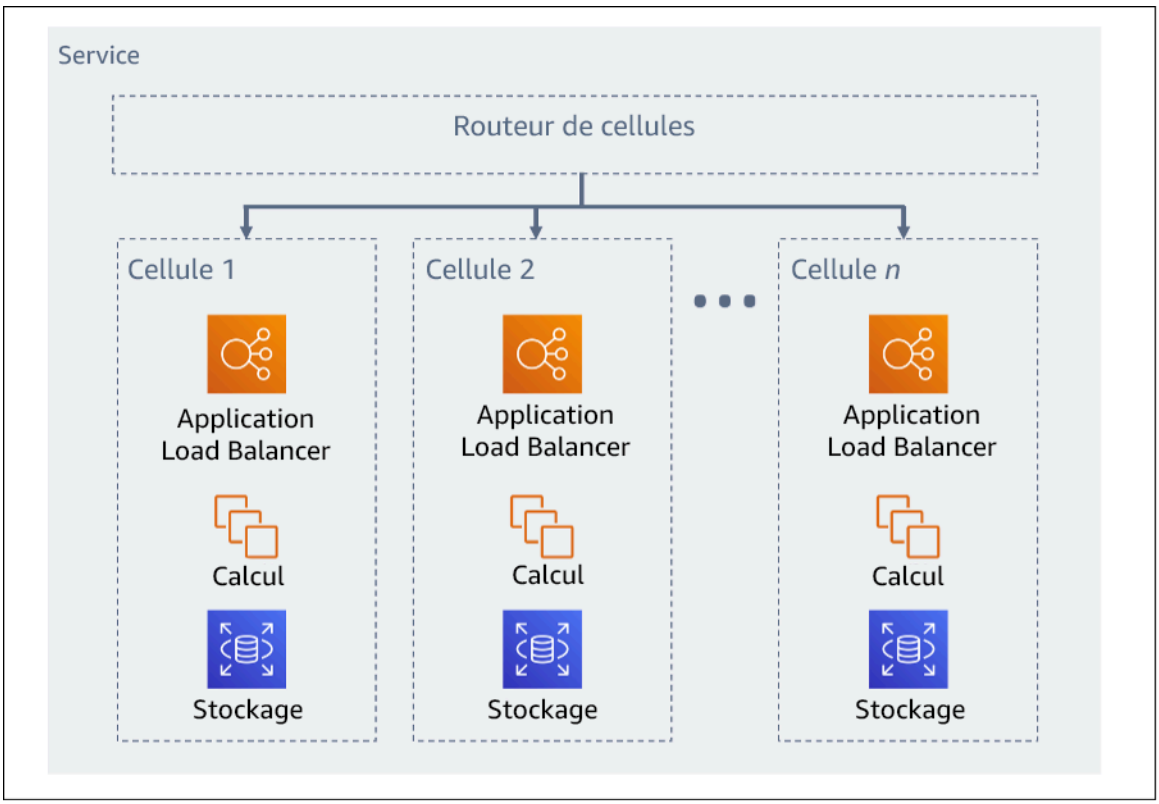


Figure 11 : Architecture cellulaire

Étapes d'implémentation

Lors de la conception d'une architecture cellulaire, vous devez tenir compte de plusieurs éléments :

1. Clé de partition : une attention particulière doit être prise lors du choix de la clé de partition.
   - Celle-ci doit s'aligner sur la base de granularité du service, ou sur la manière naturelle dont la charge de travail d'un service peut être subdivisée avec un minimum d'interactions entre les cellules. Exemples : `customer ID` ou `resource ID`.
   - La clé de partition doit être disponible dans toutes les requêtes, soit directement, soit d'une manière qui pourrait être facilement déduite de façon déterministe par d'autres paramètres.
2. Mappage cellulaire persistant : les services en amont ne doivent interagir qu'avec une seule cellule pendant le cycle de vie de leurs ressources.
   - En fonction de la charge de travail, vous devrez peut-être concevoir une stratégie de migration de cellules pour faire migrer les données d'une cellule à l'autre. La migration d'une cellule peut s'avérer nécessaire si un utilisateur ou une ressource particulière de votre charge de travail devient trop importante et nécessite une cellule dédiée.
   - Les cellules ne doivent pas partager d'état ou de composants entre elles.

- Par conséquent, les interactions entre cellules doivent être évitées ou réduites au minimum, car elles créent des dépendances entre les cellules et diminuent donc les bénéfices de l'isolement des défaillances.

3. Couche routeur : la couche routeur est un composant partagé entre les cellules et ne peut donc pas suivre la même stratégie de compartimentage qu'avec les cellules.
   - Nous recommandons de paramétrer la couche routeur pour distribuer les requêtes à des cellules individuelles à l'aide d'un algorithme de mappage de partition d'une manière efficace sur le plan des calculs. Par exemple, en combinant des fonctions de hachage cryptographiques et de l'arithmétique modulaire pour mapper les clés de partition aux cellules.
   - Pour éviter les impacts sur plusieurs cellules, la couche de routage doit rester aussi simple et évolutive horizontalement que possible, ce qui nécessite d'éviter toute logique métier complexe au sein de cette couche. Cela présente l'avantage supplémentaire de faciliter la compréhension de son comportement attendu à tout moment, favorisant ainsi une testabilité approfondie. Comme l'explique Colm MacCárthaigh dans Fiabilité, travail constant et une bonne tasse de café, des conceptions simples et des modèles de travail constants produisent des systèmes fiables et réduisent la fragilité.
4. Taille des cellules : les cellules doivent avoir une taille maximale et ne doivent pas être autorisées à croître au-delà de cette taille.
   - La taille maximale doit être identifiée en effectuant des tests approfondis, jusqu'à ce que les points de rupture soient atteints et que des marges de fonctionnement sûres soient établies. Pour obtenir plus de détails sur la mise en œuvre des pratiques de test, consultez REL07-BP04 Testez votre charge de travail
   - La charge de travail globale doit se développer en ajoutant des cellules supplémentaires, ce qui lui permet de s'adapter à l'augmentation de la demande.
5. Stratégies multi-AZ ou multi-régions : plusieurs niveaux de résilience doivent être exploités pour se protéger contre différents domaines de défaillance.
   - Pour la résilience, vous devez adopter une approche qui repose sur des couches de défense. Une couche protège contre les perturbations de petite envergure et courantes en créant une architecture hautement disponible à l'aide de plusieurs AZ. Une autre couche de défense est destinée à protéger contre les événements rares tels que les catastrophes naturelles généralisées et les perturbations au niveau régional. Cette deuxième couche implique de concevoir l'architecture de votre application pour qu'elle s'étende sur plusieurs Régions AWS. La mise en œuvre d'une stratégie multirégion pour votre charge de travail permet de la protéger contre les catastrophes naturelles généralisées qui affectent une grande région géographique d'un pays, ou les défaillances techniques à l'échelle régionale. Sachez que la mise en œuvre

d'une architecture multirégion peut être très complexe et n'est généralement pas requise pour la plupart des charges de travail. Pour en savoir plus, veuillez consulter REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements.

6. Déploiement du code : une stratégie de déploiement de code échelonnée doit être préférée au déploiement de modifications de code dans toutes les cellules en même temps.
   - Cela permet de minimiser les risques de panne de plusieurs cellules en raison d'un mauvais déploiement ou d'une erreur humaine. Pour plus de détails, consulter Automatiser un déploiement sûr et sans intervention.

Ressources

Bonnes pratiques associées :

- REL07-BP04 Testez votre charge de travail
- REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements

Documents connexes :

- Fiabilité, travail constant et une bonne tasse de café
- AWS et compartimentage
- Isolation de la charge de travail à l'aide du partitionnement aléatoire
- Automatiser un déploiement sûr et sans intervention

Vidéos connexes :

- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small
- AWS re:Invent 2018: How AWS Minimizes the Blast Radius of Failures (ARC338)
- Shuffle-sharding: AWS re:Invent 2019: Introducing The Amazon Builders' Library (DOP328)
- AWS Summit ANZ 2021 - Everything fails, all the time: Designing for resilience

Exemples connexes :

- Atelier Well-Architected : Isolement des pannes avec le partitionnement aléatoire

FIA 11. Comment concevoir votre charge de travail pour la rendre résistante aux défaillances de composants ?

Les charges de travail exigeant une haute disponibilité et un faible temps moyen de récupération (MTTR) doivent être conçues pour être résilientes.

Bonnes pratiques

- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP02 Basculez vers des ressources saines
- REL11-BP03 Automatiser la guérison sur toutes les couches
- REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle lors de la restauration
- REL11-BP05 Utiliser la stabilité statique pour empêcher le comportement bimodal
- REL11-BP06 Envoyer des notifications lorsque des événements ont un impact sur la disponibilité
- REL11-BP07 Concevoir votre produit pour atteindre les objectifs de disponibilité et les contrats de niveau de service (SLA)

REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances

Surveillez en continu l'état de votre charge de travail afin que vous et vos systèmes automatisés ayez connaissance des dégradations ou des défaillances dès qu'elles se produisent. Surveillez les indicateurs de performance clés (KPIs) en fonction de la valeur commerciale.

Tous les mécanismes de récupération et de réparation doivent commencer par la capacité à détecter rapidement les problèmes. Les défaillances techniques doivent être détectées au préalable pour être résolues. Cependant, la disponibilité dépend de la capacité de votre charge de travail à générer de la valeur commerciale. Les indicateurs de performance clés (KPIs) qui mesurent cette valeur doivent donc faire partie de votre stratégie de détection et de correction.

Résultat escompté : les composants essentiels d'une charge de travail sont surveillés de manière indépendante afin de détecter les défaillances et de les signaler au moment et à l'emplacement où elles se produisent.

Anti-modèles courants :

- Aucune alarme n'a été configurée. Les pannes se produisent donc sans notification.

- Des alarmes existent, mais les seuils ne laissent pas assez de temps pour réagir.
- Les métriques ne sont pas collectées assez souvent pour atteindre l'objectif de temps de rétablissement (RTO).
- Seules les interfaces de la charge de travail axées directement sur le client sont activement surveillées.
- Collecte uniquement des métriques techniques et non des métriques de fonction commerciale.
- Aucune métrique ne mesure l’expérience utilisateur de la charge de travail.
- Trop de contrôleurs sont créés.

Avantages liés au respect de cette bonne pratique : la surveillance appropriée à tous les niveaux vous permet de raccourcir le délai de reprise en réduisant le temps de détection.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élévé

Directives d’implémentation

Identifiez toutes les charges de travail qui seront examinées à des fins de surveillance. Une fois que vous avez identifié tous les composants de la charge de travail à surveiller, déterminez l’intervalle de surveillance. Cet intervalle a un impact direct sur la rapidité avec laquelle la restauration peut être initiée en fonction du temps nécessaire pour détecter une panne. Le délai moyen de détection (MTTD) est le délai entre la survenue d'une panne et le début des opérations de réparation. La liste des services doit être longue et complète.

La surveillance doit couvrir toutes les couches de la pile d’applications, y compris l’application, la plateforme, l’infrastructure et le réseau.

Votre stratégie de surveillance doit tenir compte de l’impact des défaillances grises. Pour en savoir plus sur les défaillances grises, consultez la section Défaillances grises dans le livre blanc Modèles de résilience Multi-AZ avancée.

Étapes d’implémentation

- Votre intervalle de surveillance dépend de la vitesse à laquelle vous devez effectuer la récupération. Votre temps de rétablissement est déterminé par le temps nécessaire pour récupérer. Vous devez donc déterminer la fréquence de collecte en tenant compte de ce temps et de votre objectif de temps de rétablissement (RTO).
- Configurez la surveillance détaillée des composants et des services gérés.

  - Déterminez si une surveillance détaillée des EC2 instances et d'Auto Scaling est nécessaire. La surveillance détaillée fournit des métriques à intervalle d'une minute, et la surveillance par défaut fournit des métriques à intervalle de cinq minutes.
  - Déterminez s'il RDS est nécessaire d'améliorer la surveillance pour. La surveillance améliorée utilise un agent sur RDS les instances pour obtenir des informations utiles sur différents processus ou threads.
  - Déterminez les exigences de surveillance des composants sans serveur critiques pour Lambda API, Gateway, AmazonEKS, ECSAmazon et tous les types d'équilibreurs de charge.
  - Déterminez les exigences de surveillance des composants de stockage pour Amazon S3FSx, AmazonEFS, Amazon et Amazon EBS.
- Créez des métriques personnalisées pour mesurer les indicateurs de performance clés de l'entreprise (KPIs). Les charges de travail mettent en œuvre des fonctions commerciales clés, qui doivent être utilisées pour aider à identifier lorsqu'un problème indirect survient. KPIs
- Surveillez l'expérience utilisateur pour détecter les défaillances à l'aide de tests canary utilisateur. Les tests de transaction synthétiques (également appelés « tests canary », à ne pas confondre avec les déploiements canary) qui peuvent exécuter et simuler le comportement des clients font partie des processus de test les plus importants. Exécutez ces tests en permanence sur vos points de terminaison de charge de travail à partir de divers emplacements distants.
- Créez des métriques personnalisées qui suivent l'expérience utilisateur. Si vous pouvez analyser l'expérience du client, vous pouvez savoir à quel moment l'expérience du consommateur se dégrade.
- Définissez des alarmes pour détecter quand une partie de votre charge de travail ne fonctionne pas correctement et pour indiquer quand mettre à l'échelle automatiquement les ressources. Les alarmes peuvent être affichées visuellement sur les tableaux de bord, envoyer des alertes via Amazon SNS ou par e-mail, et fonctionner avec Auto Scaling pour augmenter ou diminuer les ressources de charge de travail.
- Créez des tableaux de bord pour la visualisation de vos métriques. Les tableaux de bord peuvent être utilisés pour afficher visuellement des tendances, des valeurs aberrantes et d'autres indicateurs de problèmes potentiels ou pour fournir une indication des problèmes que vous pourriez vouloir examiner.
- Créez un système de suivi distribué pour vos services. La surveillance distribuée vous permet d'analyser les performances de votre application et de ses services sous-jacents, afin d'identifier et de dépanner la cause première des problèmes et des erreurs de performances.

- Créez des tableaux de bord de systèmes de surveillance (à l'aide CloudWatchde X-Ray) et collectez des données dans une région et un compte distincts.
- Créez une intégration pour la surveillance Amazon Health Aware afin de permettre de surveiller la visibilité AWS des ressources susceptibles de présenter des dégradations. Pour les charges de travail essentielles à l'entreprise, cette solution donne accès à des alertes proactives et en temps réel pour les AWS services.

Ressources

Bonnes pratiques associées :

- Définition de la disponibilité
- REL11-BP06 Envoyer des notifications lorsque des événements ont un impact sur la disponibilité

Documents connexes :

- Amazon CloudWatch Synthetics vous permet de créer des canaries d'utilisateurs
- Activer ou désactiver la surveillance détaillée pour votre instance
- Surveillance améliorée
- Surveillance de vos groupes et instances Auto Scaling à l'aide d'Amazon CloudWatch
- Publication des métriques personnalisées
- Utilisation d'Amazon CloudWatch Alarms
- Utilisation de CloudWatch tableaux de bord
- Utilisation de tableaux de CloudWatch bord multicomptes interrégionaux
- Utilisation du suivi X-Ray interrégional entre comptes
- Compréhension de la disponibilité
- Implémentation d'Amazon Health Aware (AHA)

Vidéos connexes :

- Mitigating gray failures

Exemples connexes :

- Atelier Well-Architected : niveau 300 : implémentation de la surveillance de l'état et gestion des dépendances pour améliorer la fiabilité
- Un atelier sur l'observabilité : explorer X-Ray

Outils associés :

- CloudWatch
- CloudWatch X-Ray

REL11-BP02 Basculez vers des ressources saines

En cas de défaillance des ressources, les ressources saines doivent continuer à répondre aux requêtes. En cas de problèmes de localisation (tels que la zone de disponibilité ou Région AWS), assurez-vous que vous disposez de systèmes permettant de basculer vers des ressources saines situées dans des lieux intacts.

Lorsque vous concevez un service, répartissez la charge entre les ressources, les zones de disponibilité ou les régions. Ainsi, la défaillance d'une ressource individuelle ou l'altération peut être atténuée en déplaçant le trafic vers les ressources saines restantes. Réfléchissez à la manière dont les services sont découverts et acheminés en cas de défaillance.

Concevez vos services en tenant compte de la restauration après panne. Chez AWS, nous concevons des services de manière à minimiser le temps de restauration en cas de panne et d'impact sur les données. Nos services utilisent principalement des magasins de données qui valident les requêtes uniquement lorsque les données sont stockées durablement sur plusieurs réplicas au sein d'une région. Ils sont élaborés de manière à utiliser l'isolation basée sur les cellules et à faire appel à l'isolement des pannes fourni par des zones de disponibilité. Nous utilisons largement l'automatisation dans nos procédures opérationnelles. Nous optimisons également nos replace-and-restart fonctionnalités afin de nous remettre rapidement en cas d'interruption.

Les modèles et les conceptions qui permettent le basculement varient pour chaque service de plateforme AWS . De nombreux services gérés AWS nativement sont des zones de disponibilité multiples (comme Lambda API ou Gateway). D'autres AWS services (tels que EC2 etEKS) nécessitent des conceptions de bonnes pratiques spécifiques pour prendre en charge le basculement des ressources ou le stockage des données entre AZs eux.

La surveillance doit être configurée pour vérifier que la ressource de basculement est saine, suivre la progression du basculement des ressources et surveiller le rétablissement des processus métier.

Résultat souhaité : les systèmes sont capables d'utiliser automatiquement ou manuellement de nouvelles ressources pour se remettre d'une dégradation.

Anti-modèles courants :

- La planification des défaillances ne fait pas partie de la phase de planification et de conception.
- RTOet ne RPO sont pas établis.
- Surveillance insuffisante pour détecter les ressources défaillantes.
- Isolement approprié des domaines de défaillance.
- Le basculement multirégional n'est pas pris en compte.
- La détection des défaillances est trop sensible ou trop agressive lors de la décision de basculer.
- Pas de test ni de validation de la conception du basculement.
- Automatisation de la réparation automatique sans notification indiquant que la réparation était nécessaire.
- Pas de période d'attente pour éviter un failback trop précoce.

Avantages du respect de cette bonne pratique : vous pouvez créer des systèmes plus résilients qui préservent leur fiabilité en cas de défaillance en se dégradant progressivement et en se rétablissant rapidement.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

AWS des services tels que Elastic Load Balancing et Amazon EC2 Auto Scaling aident à répartir la charge entre les ressources et les zones de disponibilité. Par conséquent, la défaillance d'une ressource individuelle (telle qu'une EC2 instance) ou la détérioration d'une zone de disponibilité peuvent être atténuées en transférant le trafic vers des ressources saines.

Pour les charges de travail multirégionales, les conceptions sont plus complexes. Par exemple, les répliques de lecture entre régions vous permettent de déployer vos données sur plusieurs régions. Régions AWS Cependant, le basculement est toujours nécessaire pour faire passer le réplica en lecture au niveau principal, puis pour rediriger le trafic vers le nouveau point de terminaison. Amazon Route 53, Amazon Application Recovery Controller (ARC) CloudFront, Amazon et AWS Global Accelerator peuvent aider à acheminer le trafic Régions AWS.

AWS les services, tels qu'Amazon S3, Lambda, API Gateway, Amazon, SQS Amazon, SNS AmazonSES, Amazon Pinpoint, Amazon, ou ECR AWS Certificate Manager Amazon DynamoDB

EventBridge, sont automatiquement déployés dans plusieurs zones de disponibilité par. AWS En cas de panne, ces AWS services acheminent automatiquement le trafic vers des emplacements sains. Les données sont stockées de manière redondante dans plusieurs zones de disponibilité et restent disponibles.

Pour AmazonRDS, Amazon Aurora, Amazon Redshift, Amazon ou Amazon EKSECS, le multi-AZ est une option de configuration. AWS peut diriger le trafic vers l'instance saine si le basculement est initié. Cette action de basculement peut être prise par le AWS client ou selon ses exigences.

Pour les EC2 instances Amazon, Amazon Redshift, Amazon ECS Tasks ou Amazon EKS Pods, vous choisissez les zones de disponibilité dans lesquelles vous souhaitez effectuer le déploiement. Pour certaines conceptions, Elastic Load Balancing fournit la solution pour détecter les instances dans les zones défectueuses et acheminer le trafic vers les zones saines. Elastic Load Balancing peut même acheminer le trafic vers les composants de votre centre de données sur site.

Pour le basculement du trafic multirégional, le réacheminement peut tirer parti d'Amazon Route 53, d'Amazon Application Recovery Controller, de Route 53 Private DNS for AWS Global Accelerator VPCs, ou CloudFront pour fournir un moyen de définir des domaines Internet et d'attribuer des politiques de routage, y compris des bilans de santé, afin d'acheminer le trafic vers des régions saines. AWS Global Accelerator fournit des adresses IP statiques qui agissent comme un point d'entrée fixe vers votre application, puis routent vers les points de terminaison Régions AWS de votre choix, en utilisant le réseau AWS mondial plutôt qu'Internet pour de meilleures performances et une meilleure fiabilité.

Étapes d’implémentation

- Créez des modèles de basculement pour toutes les applications et tous les services appropriés. Isolez chaque composant de l'architecture et créez des modèles de basculement RTO correspondant RPO à chaque composant.
- Configurez les environnements de bas niveau (tels que le développement ou les tests) avec tous les services requis pour disposer d’un plan de basculement. Déployez les solutions en utilisant l’infrastructure en tant que code (IaC) pour garantir la reproductibilité.
- Configurez un site de reprise tel qu’une deuxième région pour implémenter et tester les modèles de basculement. Si nécessaire, les ressources pour les tests peuvent être configurées temporairement afin de limiter les coûts supplémentaires.
- Déterminez les plans de basculement automatisés AWS, ceux qui peuvent être automatisés par un DevOps processus et ceux qui peuvent être manuels. Documentez et mesurez chaque service RTO etRPO.

- Créez un manuel de basculement et incluez toutes les étapes nécessaires au basculement de chaque ressource, application et service.
- Créez un manuel de failback et incluez toutes les étapes nécessaires (avec calendrier) pour chaque ressource, application et service.
- Créez un plan pour lancer et répéter le manuel. Utilisez des simulations et des tests de chaos pour tester les étapes du manuel et l'automatisation.
- En cas de détérioration de l'emplacement (par exemple, zone de disponibilité ou Région AWS), assurez-vous de disposer de systèmes permettant de basculer vers des ressources saines situées dans des lieux intacts. Vérifiez le quota, les niveaux de mise à l'échelle automatique et les ressources en cours d'exécution avant le test de basculement.

Ressources

Bonnes pratiques Well-Architected connexes :

- REL13- Planifiez la DR
- REL10 - Utilisez l'isolation des pannes pour protéger votre charge de travail

Documents connexes :

- Cadre RTO et RPO objectifs
- Basculement à l'aide du routage pondéré Route 53
- Reprise après sinistre avec Amazon Application Recovery Controller
- EC2avec mise à l'échelle automatique
- EC2Déploiements - Multi-AZ
- ECSDéploiements - Multi-AZ
- Changez de trafic à l'aide d'Amazon Application Recovery Controller
- Lambda avec Application Load Balancer et basculement
- ACMRéplication et basculement
- Réplication et basculement du magasin de paramètres
- ECRréplication entre régions et basculement
- Configuration de la réplication entre régions du gestionnaire de secrets
- Activer la réplication entre régions pour EFS et le basculement

- EFSRéplication entre régions et basculement
- Basculement du réseau
- Basculement du terminal S3 à l'aide de MRAP
- Création d’une réplication entre régions pour S3
- Conseils pour le basculement interrégional et le retour en arrière progressif sur AWS
- Basculement à l’aide d’un accélérateur mondial multirégional
- Failover avec DRS
- Création de mécanismes de reprise après sinistre à l’aide d’Amazon Route 53

Exemples connexes :

- Reprise après sinistre activée AWS
- Elastic Disaster Recovery activé AWS

REL11-BP03 Automatiser la guérison sur toutes les couches

Utilisez des capacités automatisées pour effectuer des actions correctives en cas de détection d’une défaillance. Les dégradations peuvent être automatiquement corrigées par le biais de mécanismes de service internes ou peuvent nécessiter le redémarrage ou la suppression des ressources par le biais d’actions correctives.

Pour les applications autogérées et la réparation interrégionale, les modèles de restauration et les processus de réparation automatisés peuvent être extraits des bonnes pratiques existantes.

Pouvoir redémarrer ou supprimer une ressource est important pour remédier aux défaillances. Une bonne pratique consiste à rendre les services sans état dans la mesure du possible. Cela évite toute perte de données ou de disponibilité au redémarrage des ressources. Dans le cloud, vous pouvez (et devriez généralement) remplacer la totalité de la ressource (par exemple, une instance de calcul ou une fonction sans serveur) dans le cadre du redémarrage. Le redémarrage proprement dit est un moyen simple et fiable de récupération après une défaillance. De nombreux types de défaillances différents se produisent dans les charges de travail. Les défaillances peuvent se produire au niveau du matériel, des logiciels, des communications et des opérations.

Le redémarrage ou la nouvelle tentative s’appliquent également aux requêtes réseau. Appliquez la même approche de récupération à la fois pour un délai d’expiration réseau et une défaillance de

la dépendance, si la dépendance renvoie une erreur. Comme ces deux événements ont un effet semblable sur le système, plutôt que d'essayer de traiter l'un ou l'autre comme un « cas particulier », appliquez une stratégie semblable de nouvelle tentative limitée avec un backoff exponentiel et une instabilité. La possibilité d'exécuter un redémarrage est un mécanisme de récupération présenté dans l'informatique orientée récupération et dans les architectures de cluster haute disponibilité.

Résultat souhaité : des actions automatisées sont effectuées pour remédier à la détection d'une panne.

Anti-modèles courants :

- Allocation des ressources sans mise à l'échelle automatique.
- Déploiement d'applications une par une dans des instances ou des conteneurs.
- Déploiement d'applications qui ne peuvent pas être déployées dans plusieurs emplacements sans utiliser la récupération automatique.
- Réparation manuelle des applications impossible à réparer par la mise à l'échelle et la récupération automatiques.
- Aucune automatisation pour le basculement des bases de données.
- Absence de méthodes automatisées pour rediriger le trafic vers de nouveaux points de terminaison.
- Aucune réplication du stockage.

Avantages du respect de cette bonne pratique : la réparation automatisée contribue à réduire le temps moyen de récupération et à améliorer votre disponibilité.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les conceptions d'Amazon EKS ou d'autres services Kubernetes doivent inclure à la fois un minimum et un maximum de répliques ou d'ensembles dynamiques, ainsi que le dimensionnement minimal des clusters et des groupes de nœuds. Ces mécanismes mettent à disposition un minimum de ressources de traitement en permanence tout en corrigeant automatiquement les défaillances à l'aide du plan de contrôle Kubernetes.

Les modèles de conception accessibles via un équilibreur de charge utilisant des clusters de calcul doivent tirer parti des groupes Auto Scaling. Elastic Load Balancing (ELB) distribue automatiquement

le trafic applicatif entrant sur plusieurs cibles et appliances virtuelles dans une ou plusieurs zones de disponibilité (AZs).

La taille des conceptions basées sur le calcul en cluster qui n'utilisent pas l'équilibrage de charge doit être conçue pour la perte d'au moins un nœud. Cela permet au service de continuer à fonctionner avec une capacité potentiellement réduite pendant la restauration d'un nouveau nœud. Mongo, DynamoDB Accelerator, Amazon Redshift, Amazon, Cassandra, MSK Kafka, EC2 -, Couchbase et EMR Amazon Service sont des exemples de services. ELK OpenSearch Bon nombre de ces services peuvent être conçus avec des fonctionnalités supplémentaires de réparation automatique. Certaines technologies de cluster doivent générer une alerte en cas de perte d'un nœud, déclenchant un flux de travail automatique ou manuel pour recréer un nœud. Ce flux de travail peut être automatisé AWS Systems Manager pour résoudre rapidement les problèmes.

Amazon EventBridge peut être utilisé pour surveiller et filtrer des événements tels que des CloudWatch alarmes ou des changements d'état dans d'autres AWS services. Sur la base des informations relatives aux événements, il peut ensuite invoquer AWS Lambda Systems Manager Automation ou d'autres cibles pour exécuter une logique de correction personnalisée sur votre charge de travail. Amazon EC2 Auto Scaling peut être configuré pour vérifier l'état de l'EC2instance. Si l'instance est dans un état autre qu'en cours d'exécution, ou si l'état du système est altéré, Amazon EC2 Auto Scaling considère que l'instance n'est pas saine et lance une instance de remplacement. Pour les remplacements à grande échelle (comme la perte d'une zone de disponibilité complète), il est préférable d'opter pour la stabilité statique pour une haute disponibilité.

Étapes d'implémentation

- Utilisez des groupes Auto Scaling pour déployer des niveaux dans une charge de travail. L'autoscaling peut effectuer une autoréparation sur les applications sans état et ajouter ou supprimer de la capacité.
- Pour les instances de calcul mentionnées précédemment, utilisez l'équilibrage de charge et choisissez le type d'équilibreur de charge approprié.
- Envisagez de guérir pour AmazonRDS. Avec les instances de secours, configurez le basculement automatique vers l'instance de secours. Pour Amazon RDS Read Replica, un flux de travail automatisé est nécessaire pour créer une réplique de lecture principale.
- Mettez en œuvre la restauration automatique sur les EC2 instances dont les applications ne peuvent pas être déployées sur plusieurs sites et qui peuvent tolérer le redémarrage en cas de panne. La récupération automatique peut être utilisée pour remplacer du matériel défaillant et redémarrer l'instance lorsque l'application ne peut pas être déployée sur plusieurs emplacements.

Les métadonnées de l'instance et les adresses IP associées sont conservées, ainsi que les EBSvolumes et les points de montage vers Amazon Elastic File System ou File Systems for Lustre et Windows. À l'aide de AWS OpsWorks, vous pouvez configurer la guérison automatique des EC2 instances au niveau de la couche.

- Implémentez la récupération automatique à l'aide d'AWS Step Functions et d'AWS Lambda lorsque vous ne pouvez pas utiliser la mise à l'échelle automatique ou la récupération automatique, ou lorsque la récupération automatique échoue. Lorsque vous ne pouvez pas utiliser le dimensionnement automatique, que vous ne pouvez pas utiliser la restauration automatique ou que la restauration automatique échoue, vous pouvez automatiser la guérison à l'aide de AWS Step Functions et AWS Lambda.
- Amazon EventBridge peut être utilisé pour surveiller et filtrer des événements tels que des CloudWatchalarmes ou des changements d'état dans d'autres AWS services. En fonction des informations d'événement, il peut ensuite invoquer AWS Lambda (ou d'autres cibles) pour exécuter une logique de correction personnalisée sur votre charge de travail.

Ressources

Bonnes pratiques associées :

- Définition de la disponibilité
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances

Documents connexes :

- Fonctionnement d' AWS Auto Scaling
- Amazon EC2 Automatic Recovery
- Boutique Amazon Elastic Block (AmazonEBS)
- Amazon Elastic File System (AmazonEFS)
- Qu'est-ce qu'Amazon FSx for Lustre ?
- Qu'est-ce qu'Amazon FSx pour Windows File Server ?
- AWS OpsWorks : Utilisation de la réparation automatique pour remplacer les instances en échec
- Qu'est-ce que c'est AWS Step Functions ?
- Qu'est-ce que c'est AWS Lambda ?
- Qu'est-ce qu'Amazon EventBridge ?

- Utilisation d'Amazon CloudWatch Alarms
- Amazon RDS Failover
- SSM- Systems Manager Automation
- Bonnes pratiques en matière d’architecture résiliente

Vidéos connexes :

- Provisionner et dimensionner automatiquement OpenSearch le service
- Amazon RDS Failover automatique

Exemples connexes :

- Atelier sur Auto Scaling
- Atelier Amazon RDS Failover

Outils associés :

- CloudWatch
- CloudWatch X-Ray

REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle lors de la restauration

Les plans de contrôle fournissent les outils administratifs APIs nécessaires pour créer, lire et décrire, mettre à jour, supprimer et répertorier (CRUDL) les ressources, tandis que les plans de données gèrent le trafic de day-to-day service. Lorsque vous mettez en œuvre des réponses de restauration ou d’atténuation en cas d’événements susceptibles d’avoir un impact sur la résilience, concentrez-vous sur l’utilisation d’un nombre minimal d’opérations du plan de contrôle pour récupérer, redimensionner, restaurer, réparer ou basculer le service. L’action du plan de données doit remplacer toute activité lors de ces événements de dégradation.

Par exemple, les actions suivantes font toutes partie du plan de contrôle : lancement d’une nouvelle instance de calcul, création d’un stockage par blocs et description des services de file d’attente. Lorsque vous lancez des instances de calcul, le plan de contrôle doit effectuer plusieurs tâches, telles que la recherche d’un hôte physique avec la capacité suffisante, l’allocation d’interfaces réseau, la préparation de volumes locaux de stockage par blocs, la génération d’informations d’identification et l’ajout de règles de sécurité. Les plans de contrôle relèvent souvent d’une orchestration complexe.

Résultat souhaité : lorsqu'une ressource passe à un état altéré, le système peut être rétabli automatiquement ou manuellement en transférant le trafic des ressources altérées vers des ressources saines.

Anti-modèles courants :

- Dépendance à l'égard de la modification DNS des enregistrements pour réacheminer le trafic.
- Nécessité de réaliser des opérations de mise à l'échelle du plan de contrôle pour remplacer les composants endommagés en raison de ressources sous-provisionnées.
- S'appuyer sur des actions étendues, multiservices et API multicommandes pour remédier à toute catégorie de déficience.

Avantages du respect de cette bonne pratique : l'augmentation du taux de réussite de la correction automatisée contribue à réduire le temps moyen de récupération et à améliorer la disponibilité de la charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen : pour certains types de dégradations de service, les plans de contrôle sont affectés. Les dépendances liées à une utilisation intensive du plan de contrôle pour la correction peuvent augmenter le temps de restauration (RTO) et le temps moyen de restauration (MTTR).

Directives d'implémentation

Pour limiter les actions du plan de données, évaluez chaque service pour déterminer les actions nécessaires afin de restaurer le service.

Tirez parti d'Amazon Application Recovery Controller pour déplacer le DNS trafic. Ces fonctionnalités surveillent en permanence la capacité de votre application à se rétablir en cas de défaillance et vous permettent de contrôler la restauration de votre application dans plusieurs Régions AWS zones de disponibilité et sur site.

Les politiques de routage Route 53 utilisent le plan de contrôle. Ne vous fiez donc pas à celui-ci pour la récupération. Les plans de données Route 53 répondent aux DNS requêtes et effectuent et évaluent les bilans de santé. Ils sont distribués dans le monde entier et conçus pour un accord de niveau de service de disponibilité à 100 % (SLA).

La gestion de Route 53 APIs et les consoles dans lesquelles vous créez, mettez à jour et supprimez les ressources Route 53 s'exécutent sur des plans de contrôle conçus pour donner la priorité à la cohérence et à la durabilité dont vous avez besoin lors de la gestionDNS. Pour ce faire, les plans

de contrôle sont situés dans une seule région : USA Est (Virginie du Nord). Bien que les deux systèmes soient conçus pour être très fiables, les plans de contrôle ne sont pas inclus dans leSLA. Dans de rares cas, la conception résiliente du plan de données permet de maintenir la disponibilité alors que les plans de contrôle ne le font pas. Pour les mécanismes de reprise après sinistre et de basculement, utilisez les fonctions du plan de données pour assurer la meilleure fiabilité possible.

Concevez votre infrastructure informatique de manière à ce qu'elle soit statiquement stable afin d'éviter d'utiliser le plan de contrôle lors d'un incident. Par exemple, si vous utilisez des EC2 instances Amazon, évitez de provisionner de nouvelles instances manuellement ou de demander à Auto Scaling Groups d'ajouter des instances en réponse. Pour obtenir les niveaux de résilience les plus élevés, allouez une capacité suffisante dans le cluster utilisé pour le basculement. Si ce seuil de capacité doit être limité, réglez l'ensemble du end-to-end système afin de limiter en toute sécurité le trafic total atteignant l'ensemble limité de ressources.

Pour les services tels qu'Amazon DynamoDB, API Amazon Gateway, les équilibreurs AWS Lambda de charge et les services sans serveur, l'utilisation de ces services permet de tirer parti du plan de données. Cependant, la création de nouvelles fonctions, d'équilibreurs de charge, de API passerelles ou de tables DynamoDB est une action du plan de contrôle qui doit être terminée avant la dégradation afin de préparer un événement et de répéter les actions de basculement. Pour AmazonRDS, les actions du plan de données permettent d'accéder aux données.

Pour plus d'informations sur les plans de données, les plans de contrôle et sur la manière dont les services AWS sont conçus pour atteindre les objectifs de haute disponibilité, consultez la section Stabilité statique à l'aide des zones de disponibilité.

Comprendre quelles opérations relèvent du plan de données et quelles opérations relèvent du plan de contrôle.

Étapes d’implémentation

Pour chaque charge de travail qui doit être restaurée après un événement de dégradation, évaluez le runbook de basculement, la conception de la haute disponibilité, la conception de la réparation automatique ou le plan de restauration des ressources haute disponibilité. Identifiez chaque action qui pourrait être considérée comme une action du plan de contrôle.

Envisagez de remplacer l’action du plan de contrôle par une action de plan de données :

- Auto Scaling (plan de contrôle) vers des EC2 ressources Amazon pré-dimensionnées (plan de données)

- Mise à l'échelle d'une EC2 instance Amazon (plan de contrôle) vers une AWS Lambda mise à l'échelle (plan de données)
- Évaluez toutes les conceptions utilisant Kubernetes, ainsi que la nature des actions du plan de contrôle. L'ajout de pods est une action du plan de données dans Kubernetes. Les actions doivent se limiter à l'ajout de pods et non à l'ajout de nœuds. L'utilisation de nœuds surapprovisionnés est la méthode préférée pour limiter les actions du plan de contrôle

Envisagez d'autres approches qui permettent aux actions du plan de données d'affecter les mêmes mesures correctives.

- Route 53 Modification d'enregistrement (plan de contrôle) ou Amazon Application Recovery Controller (plan de données)
- Surveillance de l'état Route 53 pour des mises à jour plus automatisées

Envisagez certains services dans une région secondaire, s'ils sont critiques, afin de permettre davantage d'actions du plan de contrôle et du plan de données dans une région non affectée.

- Amazon EC2 Auto Scaling ou Amazon EKS dans une région principale par rapport à Amazon EC2 Auto Scaling ou Amazon EKS dans une région secondaire et acheminement du trafic vers la région secondaire (action du plan de contrôle)
- Réalisez un réplica en lecture dans la région secondaire ou tentez la même action dans la région principale (action du plan de contrôle).

Ressources

Bonnes pratiques associées :

- Définition de la disponibilité
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances

Documents connexes :

- APNPartenaire : partenaires qui peuvent vous aider à automatiser votre tolérance aux pannes
- AWS Marketplace : produits pouvant être utilisés pour la tolérance aux pannes
- L'Amazon Builders' Library : éviter la surcharge des systèmes distribués en plaçant sous contrôle le plus petit service

- Amazon API DynamoDB (plan de contrôle et plan de données)
- AWS Lambda Exécutions (réparties entre le plan de contrôle et le plan de données)
- AWS Elemental MediaStore Plan de données
- Création d'applications hautement résilientes à l'aide d'Amazon Application Recovery Controller, partie 1 : pile à région unique
- Création d'applications hautement résilientes à l'aide d'Amazon Application Recovery Controller, partie 2 : pile multirégionale
- Création de mécanismes de reprise après sinistre à l’aide d’Amazon Route 53
- Qu'est-ce qu'Amazon Application Recovery Controller
- Plan de contrôle et plan de données Kubernetes

Vidéos connexes :

- Back to Basics - Using Static Stability
- Création de charges de travail multisites résilientes à l'aide de services mondiaux AWS

Exemples connexes :

- Présentation d'Amazon Application Recovery Controller
- L’Amazon Builders’ Library : éviter la surcharge des systèmes distribués en plaçant sous contrôle le plus petit service
- Création d'applications hautement résilientes à l'aide d'Amazon Application Recovery Controller, partie 1 : pile à région unique
- Création d'applications hautement résilientes à l'aide d'Amazon Application Recovery Controller, partie 2 : pile multirégionale
- Stabilité statique avec les zones de disponibilité

Outils associés :

- Amazon CloudWatch
- AWS X-Ray

REL11-BP05 Utiliser la stabilité statique pour empêcher le comportement bimodal

Les charges de travail doivent être statiquement stables et ne fonctionner que dans un seul mode normal. On parle de comportement bimodal lorsque la charge de travail présente un comportement différent en mode normal et en mode d'échec.

Par exemple, vous pouvez essayer de récupérer une défaillance de la zone de disponibilité en lançant de nouvelles instances dans une zone de disponibilité différente. Il peut en résulter une réponse bimodale lors d'un mode de défaillance. Pour éviter ce type de comportement, vous devez créer des charges de travail stables statiquement et qui fonctionnent dans un seul mode. Dans cet exemple, ces instances auraient dû être provisionnées dans la deuxième zone de disponibilité avant la panne. Ce modèle de stabilité statique permet de vérifier que la charge de travail ne fonctionne que dans un seul mode.

Résultat souhaité : les charges de travail ne présentent pas de comportement bimodal en mode normal et en mode d'échec.

Anti-modèles courants :

- Supposer que les ressources peuvent toujours être provisionnées quelle que soit l'étendue de la défaillance.
- Essayer d'acquérir dynamiquement des ressources lors d'une panne.
- Ne pas provisionner les ressources adéquates dans les zones ou les régions jusqu'à ce qu'une panne se produise.
- Envisager des modèles statiques et stables pour les ressources informatiques uniquement.

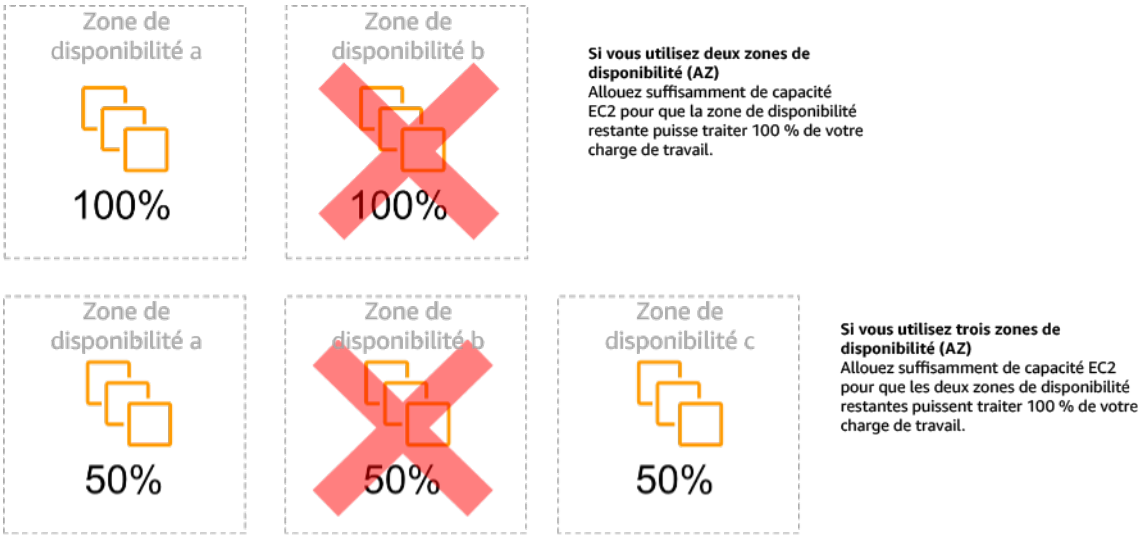
Avantages du respect de cette bonne pratique : les charges de travail exécutées avec des modèles statiquement stables sont capables d'avoir des résultats prévisibles lors d'événements normaux et de défaillances.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Un comportement bimodal survient lorsque votre charge de travail adopte un comportement différent en mode normal et en mode de défaillance (par exemple, en s'appuyant sur le lancement de nouvelles instances en cas de défaillance d'une zone de disponibilité). Un exemple de comportement bimodal est celui où les EC2 conceptions stables d'Amazon fournissent suffisamment d'instances dans chaque zone de disponibilité pour gérer la charge de travail si une AZ était supprimée. Elastic

Load Balancing ou Amazon Route 53 vérifieraient l'état pour éloigner une charge de l'instance défaillante. Une fois le trafic transféré, remplacez-le AWS Auto Scaling de manière asynchrone depuis la zone défaillante et lancez-les dans les zones saines. La stabilité statique pour les déploiements informatiques (tels que EC2 les instances ou les conteneurs) se traduit par une fiabilité maximale.



Stabilité statique des EC2 instances dans les zones de disponibilité

Cela doit être comparé au coût de ce modèle et à la valeur commerciale du maintien de la charge de travail dans tous les cas de résilience. Il est moins coûteux de provisionner moins de capacité de calcul et de compter sur le lancement de nouvelles instances en cas de panne. Cependant, pour les pannes à grande échelle (comme une zone de disponibilité ou une panne régionale), cette approche se révèle moins efficace, car elle repose à la fois sur un plan opérationnel et sur la disponibilité de ressources suffisantes dans les zones ou les régions non affectées.

Votre solution doit également tenir compte de la fiabilité par rapport aux coûts nécessaires pour votre charge de travail. Les architectures de stabilité statique s'appliquent à diverses architectures, notamment les instances de calcul réparties entre les zones de disponibilité, les conceptions de répliques de lecture de bases de données, les conceptions de clusters Kubernetes (AmazonEKS) et les architectures de basculement multirégionales.

Il est également possible de mettre en œuvre un modèle plus stable sur le plan statique en utilisant davantage de ressources dans chaque zone. En ajoutant davantage de zones, vous réduisez la quantité de calcul supplémentaire nécessaire à la stabilité statique.

Autre exemple de comportement bimodal : un délai d'expiration du réseau peut amener un système à tenter d'actualiser l'état de configuration de l'ensemble du système. Cela ajouterait une

charge inattendue à un autre composant et pourrait provoquer sa défaillance, entraînant d'autres conséquences inattendues. Cette boucle de rétroaction négative a un impact sur la disponibilité de votre charge de travail. Vous pourriez donc créer des systèmes stables statiquement et fonctionnant dans un seul mode. Un modèle statiquement stable consisterait à effectuer un travail constant et à toujours actualiser l'état de la configuration selon une cadence fixe. Lorsqu'un appel échoue, la charge de travail utilise la valeur précédemment mise en cache et déclenche une alarme.

Un autre exemple de comportement bimodal consiste à autoriser les clients à contourner votre cache de charge de travail lorsque des défaillances se produisent. Cette solution peut sembler répondre aux besoins des clients, mais elle peut modifier considérablement les exigences de votre charge de travail et risque d'entraîner des échecs.

Évaluez les charges de travail critiques afin de déterminer celles qui nécessitent ce type de modèle de résilience. Pour celles qui sont jugées critiques, chaque composant de l'application doit être examiné. Voici quelques exemples de services nécessitant une évaluation de la stabilité statique :

- Calcul : AmazonEC2, EKS -EC2, ECS -EC2, EMR - EC2
- Bases de données : Amazon Redshift, AmazonRDS, Amazon Aurora
- Stockage : Amazon S3 (zone unique), Amazon EFS (supports), Amazon FSx (supports)
- Équilibreurs de charge : selon certains modèles

Étapes d'implémentation

- Créez des systèmes stables statiquement et qui fonctionnent dans un seul mode. Dans ce cas, provisionnez suffisamment d'instances dans chaque zone de disponibilité ou région pour gérer la capacité de la charge de travail si une zone de disponibilité ou une région était supprimée. Plusieurs services peuvent être utilisés pour l'acheminement vers des ressources saines, par exemple :
  - DNSRoutage entre régions
  - MRAP MultiRegion Routage Amazon S3
  - AWS Global Accelerator
  - Contrôleur Amazon Application Recovery
- Configurez des réplicas en lecture de base de données pour tenir compte de la perte d'une instance primaire unique ou d'un réplica en lecture. Si le trafic est desservi par des réplicas en lecture, la quantité dans chaque zone de disponibilité et chaque région doit correspondre au besoin global en cas de défaillance de la zone ou de la région.

- Configurez les données critiques dans un stockage Amazon S3 conçu pour être statiquement stable pour les données stockées en cas de défaillance d'une zone de disponibilité. Si la classe de stockage Amazon S3 unizone – Accès peu fréquent est utilisée, elle ne doit pas être considérée comme statiquement stable, car la perte de cette zone minimise l'accès aux données stockées.
- Des équilibreurs de charge sont parfois configurés de manière incorrecte ou sciemment pour desservir une zone de disponibilité spécifique. Dans ce cas, la conception statiquement stable peut consister à répartir une charge de travail sur plusieurs AZs dans le cadre d'une conception plus complexe. Le modèle original peut être utilisé pour réduire le trafic interzone pour des raisons de sécurité, de latence ou de coût.

Ressources

Bonnes pratiques Well-Architected connexes :

- Définition de la disponibilité
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle lors de la restauration

Documents connexes :

- Minimiser les dépendances dans un plan de reprise après sinistre
- L'Amazon Builders' Library : stabilité statique avec les zones de disponibilité
- Fault Isolation Boundaries
- Stabilité statique avec les zones de disponibilité
- Multizone RDS
- Minimiser les dépendances dans un plan de reprise après sinistre
- DNSRoutage entre régions
- MRAP MultiRegion Routage Amazon S3
- AWS Global Accelerator
- Contrôleur Amazon Application Recovery
- Amazon S3 à zone unique
- Équilibrage de charge entre zones

Vidéos connexes :

- Stabilité statique dans AWS : AWS re:Invent 2019 : Introducing The Amazon Builders' Library () DOP328

REL11-BP06 Envoyer des notifications lorsque des événements ont un impact sur la disponibilité

Des notifications sont envoyées en cas de détection de dépassement de seuils, même si l'événement à l'origine du problème a été automatiquement résolu.

La réparation automatisée assure la fiabilité de votre charge de travail. Cependant, elle peut également masquer les problèmes sous-jacents à résoudre. Implémentez une surveillance et des événements appropriés afin de pouvoir détecter les schémas de problèmes, y compris ceux résolus par la réparation automatique, afin de pouvoir résoudre les problèmes de cause racine.

Les systèmes résilients sont conçus de manière à ce que les événements de dégradation soient immédiatement communiqués aux équipes concernées. Ces notifications doivent être envoyées par un ou plusieurs canaux de communication.

Résultat souhaité : Des alertes sont immédiatement envoyées aux équipes opérationnelles lorsque des seuils sont dépassés, tels que les taux d'erreur, la latence ou d'autres indicateurs de performance clés (KPI) critiques, afin que ces problèmes soient résolus dès que possible et que l'impact sur les utilisateurs soit évité ou minimisé.

Anti-modèles courants :

- Envoyer un trop grand nombre d'alarmes.
- Envoyer des alarmes non exploitables.
- Régler les seuils d'alarme à un niveau trop élevé (sensibilité excessive) ou trop faible (sensibilité insuffisante).
- Ne pas envoyer d'alarmes pour les dépendances externes.
- Ne pas prendre en compte les défaillances grises lors de la conception de la surveillance et des alarmes.
- Effectuer des réparations automatisées, mais ne pas notifier l'équipe appropriée que des réparations étaient nécessaires.

Avantages de l'établissement de cette meilleure pratique : les notifications de reprise informent les équipes opérationnelles et commerciales des dégradations de service afin qu'elles puissent réagir

immédiatement afin de minimiser à la fois le temps moyen de détection (MTTD) et le temps moyen de réparation (MTTR). Les notifications d'événements de reprise vous permettent également de ne pas ignorer les problèmes qui se produisent peu fréquemment.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen L'absence de mise en œuvre de mécanismes appropriés de surveillance et de notification des événements peut entraîner l'incapacité à détecter des schémas de problèmes, y compris ceux traités par la réparation automatisée. Une équipe ne sera informée de la dégradation du système que lorsque les utilisateurs contacteront le service clientèle ou par hasard.

Directives d'implémentation

Lors de la définition d'une stratégie de surveillance, le déclenchement d'une alarme est un événement courant. Cet événement contiendra probablement un identifiant pour l'alarme, l'état de l'alarme (comme `IN ALARM` ou `OK`) et les détails de ce qui l'a déclenchée. Dans de nombreux cas, un événement d'alarme doit être détecté et une notification par e-mail doit être envoyée. Voici un exemple d'action sur une alarme. La notification d'alarme est essentielle pour l'observabilité, car elle permet d'informer les bonnes personnes de l'existence d'un problème. Cependant, lorsque l'action sur les événements arrive à maturité dans votre solution d'observabilité, elle peut automatiquement remédier au problème sans nécessiter d'intervention humaine.

Une fois les alarmes de KPI surveillance établies, les alertes doivent être envoyées aux équipes appropriées lorsque les seuils sont dépassés. Ces alertes peuvent également être utilisées pour déclencher des processus automatisés qui tenteront de remédier à la dégradation.

Pour une surveillance plus complexe des seuils, des alarmes composites doivent être envisagées. Les alarmes composites utilisent un certain nombre d'alarmes de KPI surveillance pour créer une alerte basée sur la logique métier opérationnelle. CloudWatchLes alarmes peuvent être configurées pour envoyer des e-mails ou pour enregistrer des incidents dans des systèmes de suivi des incidents tiers à l'aide de l'SNSintégration d'Amazon ou d'Amazon EventBridge.

Étapes d'implémentation

Créez différents types d'alarmes en fonction des charges de travail surveillées, par exemple :

- Les alarmes d'application permettent de détecter si une partie de votre charge de travail ne fonctionne pas correctement.
- Les alarmes relatives à l'infrastructure indiquent à quel moment il faut mettre les ressources à l'échelle. Les alarmes peuvent être affichées visuellement sur les tableaux de bord, envoyer

des alertes via Amazon SNS ou par e-mail, et fonctionner avec Auto Scaling pour augmenter ou diminuer les ressources de charge de travail.

- Des alarmes statiques simples peuvent être créées pour surveiller le dépassement d'un seuil statique par une métrique pendant un nombre spécifié de périodes d'évaluation.
- Des alarmes composites peuvent prendre en compte des alarmes complexes provenant de sources multiples.
- Une fois l'alarme créée, créez les événements de notification appropriés. Vous pouvez directement invoquer un Amazon SNS API pour envoyer des notifications et associer toute automatisation à des fins de correction ou de communication.
- Intégrez la surveillance Amazon Health Aware pour permettre de surveiller la visibilité AWS des ressources susceptibles de présenter des dégradations. Pour les charges de travail essentielles à l'entreprise, cette solution donne accès à des alertes proactives et en temps réel pour les AWS services.

Ressources

Bonnes pratiques Well-Architected connexes :

- Définition de la disponibilité

Documents connexes :

- Création d'une CloudWatch alarme basée sur un seuil statique
- Qu'est-ce qu'Amazon EventBridge ?
- Qu'est-ce qu'Amazon Simple Notification Service ?
- Publication des métriques personnalisées
- Utilisation d'Amazon CloudWatch Alarms
- Amazon Health Aware (AHA)
- Configuration d'alarmes CloudWatch composites
- Nouveautés en matière d' AWS observabilité à re:Invent 2022

Outils associés :

- CloudWatch

- CloudWatch X-Ray

REL11-BP07 Concevoir votre produit pour atteindre les objectifs de disponibilité et les contrats de niveau de service (SLA)

Concevez votre produit de manière à atteindre les objectifs de disponibilité et les contrats de niveau de service (SLA). Si vous publiez ou convenez en privé d'objectifs de disponibilité ou d'accords de niveau de service, vérifiez que votre architecture et vos processus opérationnels sont conçus pour les prendre en charge.

Résultat souhaité : chaque application dispose d'un objectif défini en matière de disponibilité et d'un contrat de niveau de service pour les indicateurs de performance, qui peuvent être surveillés et maintenus afin d'atteindre les résultats commerciaux.

Anti-modèles courants :

- Concevoir et déployer des charges de travail sans fixer de contrats de niveau de service.
- Les métriques des SLA sont fixées à un niveau trop élevé sans justification ni exigences commerciales.
- Fixer des contrats de niveau de service sans tenir compte des dépendances et des contrats de niveau de service sous-jacents.
- Les conceptions d'applications sont créées sans tenir compte du modèle de responsabilité partagée pour la résilience.

Avantages du respect de cette bonne pratique : la conception d'applications reposant sur des objectifs de résilience clés vous aide à atteindre les objectifs commerciaux et les attentes des clients. Ces objectifs contribuent à orienter le processus de conception de l'application qui évalue les différentes technologies et envisage divers compromis.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

La conception des applications doit tenir compte d'un ensemble diversifié d'exigences découlant d'objectifs commerciaux, opérationnels et financiers. Dans le cadre des exigences opérationnelles, les charges de travail doivent avoir des objectifs spécifiques en matière de métriques de résilience afin qu'elles puissent être correctement surveillées et prises en charge. Les métriques de résilience

ne doivent pas être définies ou déduites après le déploiement de la charge de travail. Elles doivent être définies pendant la phase de conception et aider à guider les diverses décisions et compromis.

- Chaque charge de travail doit disposer de son propre ensemble de métriques de résilience. Ces métriques peuvent être différentes de celles d'autres applications commerciales.
- La réduction des dépendances peut avoir un impact positif sur la disponibilité. Chaque charge de travail doit tenir compte de ses dépendances et de leurs contrats de niveau de service. En général, sélectionnez les dépendances dont les objectifs de disponibilité sont égaux ou supérieurs à ceux de votre charge de travail.
- Envisagez des conceptions faiblement couplées afin que votre charge de travail puisse fonctionner correctement malgré l'altération des dépendances, lorsque cela est possible.
- Réduisez les dépendances du plan de contrôle, notamment lors de la reprise ou d'une dégradation. Évaluez les conceptions statiques stables pour les charges de travail critiques. Utilisez le partage des ressources pour augmenter la disponibilité de ces dépendances dans une charge de travail.
- L'observabilité et l'instrumentation sont essentielles pour respecter les contrats de niveau de service en réduisant le temps moyen de détection (MTTD) et le temps moyen de réparation (MTTR).
- Des défaillances moins fréquentes (MTBF plus long), des temps de détection des défaillances plus courts (MTTD plus court) et des temps de réparation plus courts (MTTR plus court) sont les trois facteurs utilisés pour améliorer la disponibilité des systèmes distribués.
- L'établissement et le respect des métriques de résilience pour une charge de travail sont à la base de toute conception efficace. Ces conceptions doivent tenir compte des compromis entre la complexité de la conception, les dépendances des services, les performances, la mise à l'échelle et les coûts.

Étapes d'implémentation

- Examinez et documentez la conception de la charge de travail en tenant compte des questions suivantes :
  - Où les plans de contrôle sont-ils utilisés dans la charge de travail ?
  - Comment la charge de travail met-elle en œuvre la tolérance aux pannes ?
  - Quels sont les modèles de conception pour la mise à l'échelle, la mise à l'échelle automatique, la redondance et les composants hautement disponibles ?
  - Quelles sont les exigences en matière de cohérence et de disponibilité des données ?

  - Y a-t-il des considérations relatives à l'économie des ressources ou à la stabilité statique des ressources ?
  - Quelles sont les dépendances des services ?
- Définissez les métriques SLA en fonction de l'architecture de la charge de travail tout en travaillant avec les parties prenantes. Tenez compte des SLA de toutes les dépendances utilisées par la charge de travail.
- Une fois l'objectif du SLA fixé, optimisez l'architecture pour le respecter.
- Une fois que la conception a été définie de manière à respecter le contrat de niveau de service, il faut mettre en œuvre les changements opérationnels, l'automatisation des processus et les runbooks qui visent également à réduire les délais d'attente et les temps de réponse.
- Une fois le déploiement effectué, surveillez et rendez compte du contrat de niveau de service.

Ressources

Bonnes pratiques associées :

- REL03-BP01 Choisissez comment segmenter votre charge de travail
- REL10-BP01 Déploiement de la charge de travail sur plusieurs emplacements
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL11-BP03 Automatiser la guérison sur toutes les couches
- REL12-BP04 Tester la résilience à l'aide de l'ingénierie du chaos
- REL13-BP01 Définir les objectifs de reprise en termes de durée d'indisponibilité et de perte de données
- Comprendre l'état de la charge de travail

Documents connexes :

- Disponibilité avec redondance
- Pilier de fiabilité - Disponibilité
- Mesurer la disponibilité
- AWS Fault Isolation Boundaries
- Modèle de responsabilité partagée pour la résilience
- Stabilité statique avec les zones de disponibilité

- Accords de niveau de service (SLA) AWS
- Conseils pour l'architecture cellulaire sur AWS
- Infrastructure AWS
- Livre blanc sur les modèles de résilience multi-AZ avancés

Services connexes :

- Amazon CloudWatch
- AWS Config
- AWS Trusted Advisor

FIA 12. Comment tester la fiabilité ?

Une fois que vous avez conçu votre charge de travail pour qu'elle soit résiliente aux sollicitations de la production, les tests sont le seul moyen de s'assurer qu'elle fonctionne comme prévu et d'obtenir la résilience voulue.

Bonnes pratiques

- REL12-BP01 Utiliser des playbooks pour enquêter sur les causes des défaillances
- REL12-BP02 Effectuer une analyse post-incident
- REL12-BP03 Tester les exigences de capacité de mise à l'échelle et de performances
- REL12-BP04 Tester la résilience à l'aide de l'ingénierie du chaos
- REL12-BP05 Organiser régulièrement des tests de simulation de panne

REL12-BP01 Utiliser des playbooks pour enquêter sur les causes des défaillances

Consignez le processus d'enquête dans des playbooks afin de faciliter l'application de réponses cohérentes et rapides face aux scénarios de défaillance qui ne sont pas bien compris. Les playbooks sont les étapes prédéfinies suivies pour identifier les facteurs adjuvants à un scénario de défaillance. Les résultats des étapes du processus sont utilisés pour déterminer les prochaines mesures à prendre jusqu'à ce que la question soit identifiée ou remontée.

Le playbook est une planification proactive que vous devez appliquer afin de pouvoir prendre efficacement des mesures réactives. Lorsque des scénarios de défaillance ne figurant pas dans le

playbook sont rencontrés en production, commencez par résoudre le problème (éteindre l'incendie). Procédez ensuite à une rétrospective en examinant les étapes suivies pour résoudre le problème et utilisez-les pour ajouter une nouvelle entrée dans le playbook.

Notez que les playbooks sont utilisés en réponse à des incidents spécifiques, tandis que les runbooks le sont pour obtenir des résultats spécifiques. En règle générale, les runbooks sont employés pour les activités de routine et les playbooks pour répondre à des événements non réguliers.

Anti-modèles courants :

- Planification du déploiement d'une charge de travail sans connaître les processus permettant de diagnostiquer les problèmes ou de réagir aux incidents.
- Décisions imprévues sur les systèmes à partir desquels peut se faire la collecte des journaux et métriques lors de l'examen d'un événement.
- Non-conservation des métriques et événements pendant suffisamment longtemps pour pouvoir récupérer les données.

Avantages du respect de cette bonne pratique : la capture de playbooks garantit le respect constant des processus. La codification de vos playbooks limite l'introduction d'erreurs à partir de l'activité manuelle. L'automatisation des playbooks accélère le temps de réponse à un événement en évitant aux membres de l'équipe d'intervenir ou en leur fournissant des informations supplémentaires lorsque leur intervention commence.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

- Utilisez des playbooks pour identifier les problèmes. Les playbooks sont des processus documentés pour enquêter sur les problèmes. Mettez en œuvre des réponses cohérentes et rapides aux échecs en documentant les processus dans des playbooks. Les playbooks doivent contenir les informations et les instructions nécessaires pour permettre à une personne compétente de recueillir les informations pertinentes, identifier les causes potentielles de défaillance, isoler les pannes et déterminer les facteurs adjuvants (c'est-à-dire effectuer une analyse post-incident).
  - Mettez en œuvre les playbooks en tant que code. Effectuez vos opérations en tant que code en scriptant vos playbooks afin d'en assurer la cohérence et de limiter les erreurs causées par les processus manuels. Les playbooks peuvent être composés de plusieurs scripts représentant les différentes étapes qui pourraient être nécessaires pour identifier les facteurs contribuant à un problème. Les activités de runbook peuvent être invoquées ou effectuées dans le cadre

d’activités de playbook, ou peuvent demander l’exécution d’un playbook en réponse à des événements identifiés.

- Automatisez vos playbooks opérationnels avec AWS Systems Manager
- Run Command d’AWS Systems Manager
- AWS Systems Manager Automation
- Présentation de AWS Lambda
- Qu’est-ce qu’Amazon EventBridge ?
- Utilisation d’alarmes Amazon CloudWatch

Ressources

Documents connexes :

- AWS Systems Manager Automation
- Run Command d’AWS Systems Manager
- Automatisez vos playbooks opérationnels avec AWS Systems Manager
- Utilisation d’alarmes Amazon CloudWatch
- Utilisation de scripts Canary (Amazon CloudWatch Synthetics)
- Qu’est-ce qu’Amazon EventBridge ?
- Présentation de AWS Lambda

Exemples connexes :

- Automatisation des opérations avec les playbooks et les runbooks

REL12-BP02 Effectuer une analyse post-incident

Passez en revue les événements ayant un impact sur le client et identifiez les facteurs adjuvants et les mesures préventives. Utilisez ces informations pour développer des mesures d’atténuation afin de limiter ou d’empêcher la récurrence. Développez des procédures pour fournir des réponses rapides et efficaces. Publiez, le cas échéant, les facteurs adjuvants et les mesures correctives adaptées au public ciblé. Vous devez disposer d’une méthode pour communiquer ces causes à d’autres si nécessaire.

Évaluez pourquoi les tests existants n'ont pas permis de résoudre le problème. Ajoutez des tests pour ce cas si aucun test correspondant n'existe.

Résultat escompté : vos équipes ont adopté une approche cohérente et convenue pour gérer l'analyse post-incident. L'un des mécanismes est le processus de correction d'erreur (COE). Celui-ci aide vos équipes à identifier, comprendre et traiter les causes profondes des incidents, tout en mettant en place des mécanismes et des barrières de protection pour limiter la probabilité qu'un incident se reproduise.

Anti-modèles courants :

- Trouver des facteurs adjuvants sans pour autant continuer à chercher plus en profondeur d'autres problèmes et approches potentiels pour atténuer les risques.
- Identification limitée aux causes d'erreur humaine et sans formation ou automatisation pouvant empêcher les erreurs humaines.
- Se concentrer sur les reproches plutôt que sur la compréhension des causes profondes, ce qui crée une culture de la peur et empêche de communiquer ouvertement
- Absence de partage d'informations, qui entrave la circulation des résultats de l'analyse de l'incident et empêche les autres de bénéficier des enseignements tirés
- Absence de mécanisme permettant de capturer les connaissances institutionnelles, ce qui engendre une perte d'informations précieuses en ne conservant pas les enseignements tirés sous la forme de bonnes pratiques actualisées, et entraîne la répétition d'incidents ayant des causes profondes identiques ou similaires

Avantages du respect de cette bonne pratique : une analyse post-incident et le partage des résultats permettent à d'autres charges de travail d'atténuer les risques si elles ont mis en œuvre les mêmes facteurs adjuvants. Elle permet aussi de mettre en œuvre l'atténuation des risques ou la récupération automatisée avant qu'un incident ne se produise.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Une bonne analyse post-incident permet de proposer des solutions courantes pour les problèmes avec des modèles d'architecture utilisés dans d'autres compartiments de vos systèmes.

La documentation et la résolution des problèmes sont l'une des pierres angulaires du processus COE. Il est recommandé de définir une méthode normalisée pour documenter les causes profondes

et de veiller à ce qu'elles soient examinées et traitées. Attribuez clairement la responsabilité du processus d'analyse post-incident. Désignez une équipe ou une personne chargée de superviser les enquêtes et le suivi de l'incident.

Encouragez une culture axée sur l'apprentissage et l'amélioration plutôt que sur les reproches. Insistez sur le fait que l'objectif est de prévenir de futurs incidents, et non de pénaliser des individus.

Élaborez des procédures bien définies pour mener les analyses post-incident. Ces procédures doivent décrire les étapes à suivre, les informations à collecter et les principales questions à aborder lors de l'analyse. Enquêtez en profondeur sur les incidents, en allant au-delà des causes immédiates afin d'identifier les causes profondes et les facteurs contributifs. Utilisez des techniques telles que les « cinq pourquoi » pour approfondir les problèmes sous-jacents.

Tenez un répertoire des enseignements tirés des analyses des incidents. Ces connaissances institutionnelles peuvent servir de référence pour les incidents futurs et les efforts de prévention. Partagez les résultats et les réflexions tirées des analyses post-incident, et envisagez d'organiser des réunions de synthèse post-incident ouvertes à tous pour discuter des enseignements tirés.

Étapes d'implémentation

- Veillez à ce que l'analyse post-incident soit exempte de tout reproche. Cela permet aux personnes impliquées dans l'incident de faire preuve d'objectivité quant aux actions correctives proposées, et de promouvoir une auto-évaluation et une collaboration honnêtes entre les équipes.
- Définissez une méthode standardisée pour documenter les problèmes critiques. Voici un exemple de structure :
  - Que s'est-il passé ?
  - Quel a été l'impact sur vos clients et votre activité ?
  - Quelle était la cause profonde ?
  - Quelles sont les données à votre disposition pour étayer votre raisonnement ?
    - Par exemple, des métriques et des graphiques.
  - Quelles ont été les principales répercussions, notamment en termes de sécurité ?
    - Lors de la conception des charges de travail, vous faites un compromis entre les piliers en fonction de votre activité. Ces décisions professionnelles peuvent orienter vos priorités en matière d'ingénierie. Vous pouvez opter pour l'optimisation afin de réduire les coûts au détriment de la fiabilité dans les environnements de développement ou, pour les solutions stratégiques, vous pouvez optimiser la fiabilité pour des coûts plus importants. La sécurité est toujours une priorité, car vous devez protéger vos clients.

    - Quelles leçons avez-vous apprises ?
    - Quelles mesures correctives allez-vous prendre ?
        - Éléments d’action
        - Éléments connexes
- Élaborez des procédures opérationnelles standard bien définies pour mener les analyses post-incident.
- Mettez en place un processus standardisé de signalement des incidents. Documentez tous les incidents de manière exhaustive, y compris le rapport d’incident initial, les journaux, les communications et les mesures prises pendant l’incident.
- N’oubliez pas qu’un incident n’est pas forcément une panne. Il peut s’agir d’un accident évité de justesse ou d’un système qui fonctionne de manière inattendue tout en remplissant sa fonction.
- Améliorez sans cesse votre processus d’analyse post-incident en fonction des retours et des enseignements tirés.
- Capturez les principales conclusions dans un système de gestion des connaissances et examinez les modèles qui devraient être ajoutés aux guides du développeur ou aux listes de contrôle de prédéploiement.

Ressources

Documents connexes :

- Pourquoi mettre en place la correction des erreurs (COE)

Vidéos connexes :

- Amazon’s approach to failing successfully
- AWS re:Invent 2021 - Amazon Builders’ Library: Operational Excellence at Amazon

REL12-BP03 Tester les exigences de capacité de mise à l’échelle et de performances

Utilisez des techniques telles que les tests de charge pour valider que la charge de travail répond aux exigences de mise à l’échelle et de performances.

Dans le cloud, vous pouvez créer un environnement de test à l’échelle de la production pour votre charge de travail à la demande. Au lieu de vous fier à un environnement de test réduit, qui pourrait

entraîner des prévisions inexactes des comportements de production, vous pouvez utiliser le cloud pour provisionner un environnement de test qui reflète étroitement votre environnement de production attendu. Cet environnement vous permet de réaliser des tests dans le cadre d'une simulation plus précise des conditions réelles auxquelles votre application est confrontée.

Parallèlement aux tests de performance, il est essentiel de vérifier que vos ressources de base, vos paramètres de mise à l'échelle, vos quotas de service et votre conception de la résilience fonctionnent comme prévu sous charge. Cette approche globale garantit que votre application peut être mise à l'échelle de manière fiable et fonctionner selon les besoins, même dans les conditions les plus exigeantes.

Résultat escompté : votre charge de travail conserve son comportement attendu même lorsqu'elle est soumise à des pics de charge. Vous abordez de manière proactive tous les problèmes liés aux performances susceptibles de survenir au fur et à mesure que l'application grandit et évolue.

Anti-modèles courants :

- Vous utilisez des environnements de test qui ne correspondent pas étroitement à l'environnement de production.
- Vous considérez les tests de charge comme une activité ponctuelle distincte plutôt que comme une partie intégrante du pipeline d'intégration continue (CI) du déploiement.
- Vous ne définissez pas d'exigences de performances claires et mesurables, telles que des cibles de temps de réponse, de débit et de capacité de mise à l'échelle.
- Vous effectuez des tests avec des scénarios de charge non réalistes ou insuffisants, et vous ne parvenez pas à tester les pics de charge, les pics soudains ou une charge élevée prolongée.
- Vous n'effectuez pas de test de stress de la charge de travail en dépassant les limites de charge attendues.
- Vous utilisez des outils de test de charge ou de profilage des performances inadaptés ou inappropriés.
- Vous ne disposez pas de systèmes complets de surveillance et d'alerte pour effectuer le suivi des métriques de performances et détecter les anomalies.

Avantages liés au respect de cette bonne pratique :

- Les tests de charge vous aident à identifier les goulots d'étranglement potentiels de performances de votre système avant sa mise en production. Lorsque vous simulez le trafic et les charges

de travail au niveau de la production, vous pouvez identifier les domaines dans lesquels votre système peut avoir du mal à gérer la charge, tels que de longs délais de réponse, des contraintes de ressources ou des défaillances du système.

- En testant votre système dans différentes conditions de charge, vous pouvez mieux comprendre les exigences en matière de ressources pour prendre en charge votre charge de travail. Ces informations peuvent vous aider à prendre des décisions éclairées concernant l'allocation des ressources et à éviter un surprovisionnement ou un sous-provisionnement des ressources.
- Pour identifier les points de défaillance potentiels, vous pouvez observer les performances de votre charge de travail dans des conditions de charge élevée. Ces informations vous aident à améliorer la fiabilité et la résilience de votre charge de travail en mettant en œuvre des mécanismes de tolérance aux pannes, des stratégies de basculement ou des mesures de redondance, selon le cas.
- Vous identifiez et traitez les problèmes de performances à un stade précoce, ce qui vous permet d'éviter les conséquences coûteuses de pannes du système, de longs délais de réponse et d'utilisateurs mécontents.
- Les données de performances et les informations de profilage détaillées collectées lors des tests peuvent vous aider à résoudre les problèmes liés aux performances susceptibles de survenir en production. Cela peut conduire à une réponse et une résolution plus rapides des incidents, ce qui réduit l'impact sur les utilisateurs et les opérations de votre organisation.
- Dans certains secteurs, les tests de performances proactifs peuvent aider votre charge de travail à respecter les normes de conformité, réduisant ainsi le risque de pénalités ou de problèmes juridiques.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

La première étape consiste à définir une stratégie de tests complète qui couvre tous les aspects des exigences de mise à l'échelle et de performances. Pour commencer, définissez clairement les objectifs de niveau de service (SLO) de votre charge de travail en fonction des besoins de votre entreprise, tels que le débit, l'histogramme de latence et le taux d'erreur. Concevez ensuite une suite de tests capables de simuler différents scénarios de charge allant d'une utilisation moyenne à des pics soudains et des pics de charge prolongés, et vérifiez que le comportement de la charge de travail respecte vos objectifs de niveau de service. Ces tests doivent être automatisés et intégrés dans votre pipeline d'intégration et de déploiement continus afin de détecter les régressions de performances de façon précoce dans le processus de développement.

Pour tester efficacement la mise à l'échelle et les performances, investissez dans les outils et l'infrastructure appropriés. Cela inclut des outils de test de charge capables de générer un trafic utilisateur réaliste, des outils de profilage des performances pour identifier les goulots d'étranglement et des solutions de surveillance pour suivre les métriques clés. Il est important de vérifier que vos environnements de test correspondent étroitement à l'environnement de production en termes d'infrastructure et de conditions d'environnement afin que les résultats de vos tests soient aussi précis que possible. Pour faciliter la réplication et la mise à l'échelle fiables de configurations de type production, utilisez une infrastructure en tant que code et des applications basées sur des conteneurs.

Les tests de mise à l'échelle et de performances sont un processus continu et non une activité ponctuelle. Mettez en œuvre une surveillance et des alertes complètes pour suivre les performances de l'application en production, et utilisez ces données pour affiner en permanence vos stratégies de test et vos efforts d'optimisation. Analysez régulièrement les données de performances pour identifier les problèmes émergents, tester les nouvelles stratégies de mise à l'échelle et mettre en œuvre des optimisations afin d'améliorer l'efficacité et la fiabilité de l'application. Lorsque vous adoptez une approche itérative et que vous tirez constamment des enseignements des données de production, vous pouvez vérifier que votre application peut s'adapter aux demandes variables des utilisateurs et maintenir une résilience et des performances optimales au fil du temps.

Étapes d'implémentation

1. Établissez des exigences de performances claires et mesurables, telles que des cibles de temps de réponse, de débit et de capacité de mise à l'échelle. Ces exigences doivent être basées sur les modèles d'utilisation de votre charge de travail, les attentes des utilisateurs et les besoins de votre entreprise.
2. Sélectionnez et configurez un outil de test de charge capable d'imiter avec précision les modèles de charge et le comportement des utilisateurs dans votre environnement de production.
3. Configurez un environnement de test correspondant étroitement à l'environnement de production, y compris aux conditions d'infrastructure et d'environnement, afin d'améliorer la précision des résultats de vos tests.
4. Créez une suite de tests couvrant un large éventail de scénarios, allant de modèles d'utilisation moyenne à des pics de charge, à des pics rapides et à des charges élevées prolongées. Intégrez ces tests dans vos processus d'intégration et de déploiement continus afin de détecter les régressions de performances de façon précoce dans le processus de développement.
5. Effectuez des tests de charge pour simuler le trafic utilisateur réel et comprendre le comportement de votre application dans différentes conditions de charge. Pour effectuer un test de stress

de votre application, dépassez la charge attendue et observez son comportement, tel qu'une dégradation du temps de réponse, l'épuisement des ressources ou des défaillances du système, afin d'identifier le point de rupture de votre application et d'élaborer des stratégies de mise à l'échelle. Évaluez la capacité de mise à l'échelle de votre charge de travail en augmentant progressivement la charge, et mesurez l'impact sur les performances pour identifier les limites de mise à l'échelle et planifier les besoins futurs de capacité.

6. Mettez en œuvre une surveillance et des alertes complètes pour suivre les métriques de performances, détecter les anomalies et lancer des actions de mise à l'échelle ou des notifications lorsque les seuils sont dépassés.
7. Surveillez et analysez en permanence les données de performances pour identifier les domaines à améliorer. Itérez sur vos stratégies de test et vos efforts d'optimisation.

Ressources

Bonnes pratiques associées :

- REL01-BP04 Surveiller et gérer les quotas
- REL06-BP01 Surveiller tous les composants de la charge de travail (génération)
- REL06-BP03 Envoyer des notifications (traitement et alarmes en temps réel)

Documents connexes :

- Tests de charge des applications
- Test de charge distribuée sur AWS
- Surveillance des performances d'application
- Politique de test d'Amazon EC2

Exemples connexes :

- Test de charge distribuée sur AWS (GitHub)

Outils associés :

- Amazon CodeGuru Profiler
- Amazon CloudWatch RUM

- Apache JMeter
- K6
- Vegeta
- Hey
- ab
- wrk

REL12-BP04 Tester la résilience à l'aide de l'ingénierie du chaos

Exécutez régulièrement des expériences de chaos dans des environnements dont les conditions se rapprochent autant que possible de la production pour comprendre comment nos systèmes réagissent à des conditions défavorables.

Résultat escompté :

la résilience de la charge de travail est régulièrement vérifiée en appliquant l'ingénierie du chaos sous la forme d'expériences d'injection de pannes ou de charge inattendue, en plus des tests de résilience qui confirment le comportement attendu connu de votre charge de travail lors d'un événement. Associez l'ingénierie du chaos aux tests de résilience pour avoir l'assurance que votre charge de travail peut résister en cas de défaillance des composants et récupérer suite à des perturbations inattendues avec peu ou pas d'impact.

Anti-modèles courants :

- Conception à des fins de résilience, mais pas de vérification du fonctionnement de la charge de travail dans son ensemble en cas de défaillances.
- Pas d'expériences dans des conditions concrètes et pour la charge prévue.
- Pas de traitement de vos expériences en tant que code ou de maintenance de vos expériences tout au long du cycle de développement.
- Pas d'exécution d'expériences de chaos dans le cadre de votre pipeline CI/CD, ainsi qu'en dehors des déploiements.
- Pas d'utilisation des analyses passées post-incident pour déterminer les défaillances à tester.

Avantages du respect de cette bonne pratique : l'injection de défaillances pour vérifier la résilience de votre charge de travail vous permet d'avoir l'assurance que les procédures de récupération de votre conception résiliente fonctionneront en cas de défaillances réelles.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

L'ingénierie du chaos offre la possibilité à vos équipes d'injecter en continu des perturbations concrètes (simulations) de manière contrôlée au niveau du fournisseur de services, de l'infrastructure, de la charge de travail et des composants, avec peu ou pas d'impact pour vos clients. Ainsi, vos équipes tirent les leçons de ces défaillances et observent, mesurent et améliorent la résilience de vos charges de travail, tout en confirmant que les alertes se déclenchent et que les équipes sont informées en cas d'événement.

Une pratique de l'ingénierie du chaos en continu peut mettre en évidence des défaillances dans vos charges de travail qui, si elles ne sont pas résolues, peuvent impacter de manière négative la disponibilité et le fonctionnement.

> **Note**
>
> L'ingénierie du chaos est la discipline d'expérimentation d'un système. Elle permet de s'assurer de la capacité du système à résister à des conditions de production difficiles. – Principes de l'ingénierie du chaos

Si un système est capable de résister à ces perturbations, l'expérience de chaos doit être maintenue en tant que test de régression automatisé. De cette façon, les expériences de chaos doivent être réalisées dans le cadre de votre cycle de développement des systèmes et de votre pipeline CI/CD.

Pour veiller à ce que votre charge de travail résiste en cas de défaillance des composants, injectez des événements concrets dans le cadre de vos expériences. Par exemple, expérimentez une perte des instances Amazon EC2 ou un basculement de l'instance de base de données Amazon RDS principale, puis vérifiez que votre charge de travail n'est pas impactée (ou très peu). Utilisez plusieurs défaillances des composants pour simuler des événements capables de causer une perturbation dans une zone de disponibilité.

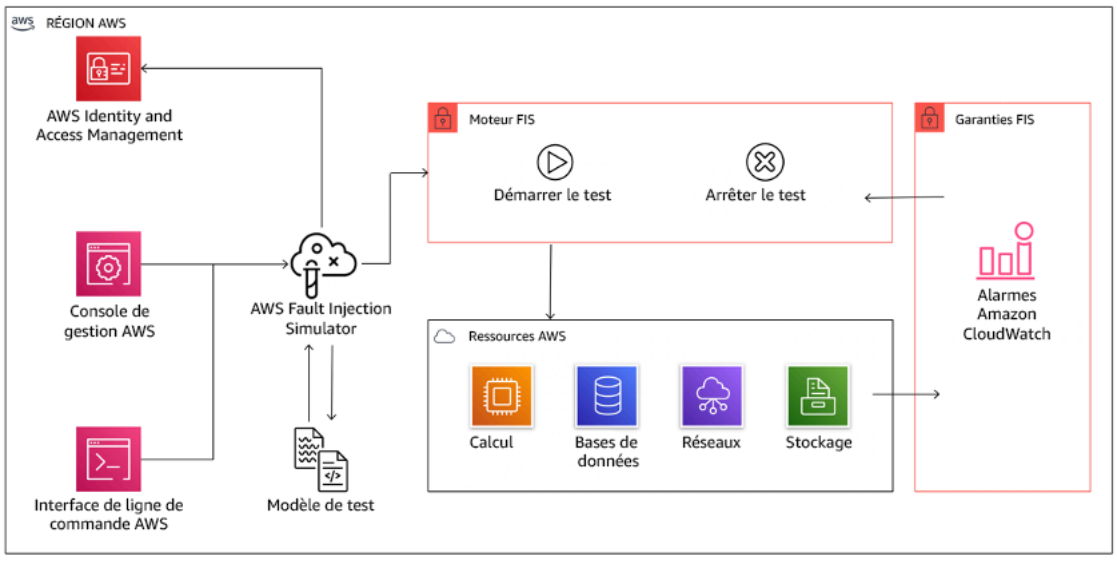
Pour les défaillances de niveau application (telles que les plantages), commencez par des tests de stress comme l'épuisement de la mémoire et du processeur.

Afin de valider des mécanismes de remplacement ou de basculement pour les dépendances externes dues aux pannes réseau intermittentes, vos composants doivent simuler un tel événement en bloquant l'accès aux fournisseurs tiers pendant une durée spécifiée pouvant aller de quelques secondes à plusieurs heures.

D’autres modes de dégradation peuvent entraîner des fonctionnalités limitées et ralentir les réponses, ce qui se traduit par une perturbation de vos services. Généralement, cette dégradation résulte d’une latence accrue sur les services critiques et d’une communication réseau peu fiable (perte de paquets). Les expériences avec ces défaillances, dont les effets de mise en réseau tels que la latence, les messages supprimés et les défaillances DNS, peuvent inclure l’incapacité de résoudre un nom, d’atteindre un service DNS ou de se connecter aux services dépendants.

Outils de l’ingénierie du chaos :

AWS Fault Injection Service (AWS FIS) est un service entièrement géré permettant l’exécution d’expériences d’injection de pannes qui peuvent être utilisées dans le cadre de votre pipeline CD, ou en dehors du pipeline. AWS FIS s’impose donc comme un choix judicieux lors des tests de simulation de pannes. Il prend en charge l’introduction simultanée de défaillances sur différents types de ressources, notamment Amazon EC2, Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) et Amazon RDS. Ces défaillances incluent l’arrêt des ressources, les basculements forcés, le stress du processeur ou de la mémoire, la limitation, la latence et la perte de paquets. Comme il est intégré aux alarmes Amazon CloudWatch, vous pouvez définir des conditions d’arrêt comme barrières de protection pour annuler une expérience si elle provoque un impact inattendu.



AWS Fault Injection Service s’intègre aux ressources AWS pour vous permettre d’exécuter des expériences d’injection de pannes pour vos charges de travail.

Il existe également plusieurs options tierces pour les expériences d’injection de pannes. Il s’agit notamment d’outils open source tels que Chaos Toolkit, Chaos Mesh et Litmus Chaos, ainsi que d’options commerciales telles que Gremlin. Pour élargir le champ des pannes pouvant être injectées

sur AWS, AWS FIS s'intègre à Chaos Mesh et Litmus Chaos, ce qui vous permet de coordonner les flux de travail d'injection de pannes entre plusieurs outils. Par exemple, vous pouvez exécuter un test de stress sur un processeur de pod à l'aide des défaillances Chaos Mesh ou Litmus, tout en arrêtant un pourcentage de nœuds de cluster sélectionné de façon aléatoire grâce aux actions des défaillances AWS FIS.

Étapes d'implémentation

1. Déterminez les défaillances à utiliser pour les expériences.

   Évaluez la conception de votre charge de travail à des fins de résilience. Ces conceptions (créées selon les bonnes pratiques du cadre Well-Architected) tiennent compte des risques basés sur les dépendances critiques, les événements passés, les problèmes connus et les exigences de conformité. Répertoriez chaque élément de la conception destiné à maintenir la résilience et les défaillances qu'il entend réduire. Pour plus d'informations sur la création de telles listes, consultez le livre blanc consacré à l'examen de la préparation opérationnelle, qui explique comment créer un processus visant à empêcher que de tels incidents ne se reproduisent. Le processus de Failure Modes and Effects Analysis (FMEA) ou d'analyse des modes de défaillance et de leurs effets vous propose un framework pour réaliser une analyse de niveau composant des défaillances et de leur impact sur votre charge de travail. Le FMEA est décrit plus en détail par Adrian Cockcroft dans Failure Modes and Continuous Resilience.
2. Attribuer une priorité à chaque défaillance.

   Commencez par définir une classification grossière telle que élevée, moyenne et basse. Pour évaluer les priorités, tenez compte de la fréquence de la défaillance et de son impact sur la charge de travail dans son ensemble.

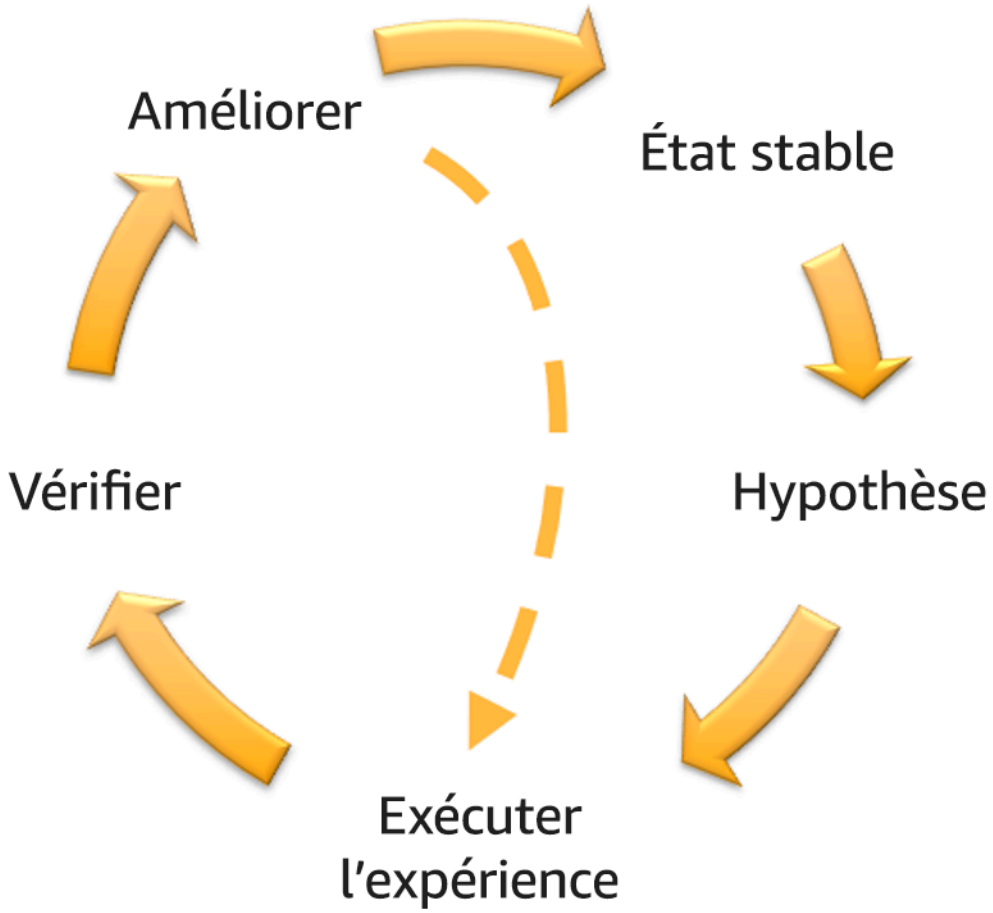
   Lors de la prise en compte de la fréquence d'une défaillance donnée, analysez les données passées de cette charge de travail, le cas échéant. Si aucune donnée passée n'est disponible, utilisez les données des autres charges de travail s'exécutant dans un environnement semblable.

   Lors de la prise en compte de l'impact d'une défaillance donnée, souvenez-vous qu'en général plus le champ de la défaillance est large, plus grand est l'impact. Tenez compte également de la conception de la charge de travail et de son objectif. Par exemple, la capacité à accéder aux magasins de données sources est essentielle pour une charge de travail effectuant des transformations et de l'analyse de données. Dans ce cas, vous donnerez la priorité aux expériences liées aux défaillances d'accès ainsi qu'aux accès limités et à l'insertion de la latence.

Les analyses post-incident constituent une excellente source de données pour comprendre à la fois la fréquence et l'impact des modes de défaillance.

Utilisez la priorité attribuée pour déterminer les défaillances à expérimenter en premier lieu, puis l'ordre dans lequel développer de nouvelles expériences d'injection de pannes.

3. Suivre le volant d'inertie de l'ingénierie du chaos et de la résilience continue figurant sur la figure suivante pour chaque expérience réalisée.



Volant d'inertie de l'ingénierie du chaos et de la résilience continue réalisé grâce à la méthode scientifique d'Adrian Hornsby.

a. Définir l'état stable comme le résultat mesurable d'une charge de travail qui indique un comportement normal.

Votre charge de travail présente un état stable si elle fonctionne de manière fiable et comme prévu. Par conséquent, confirmez que votre charge de travail est saine avant de définir un état

stable. L'état stable ne signifie pas forcément sans impact pour la charge de travail en cas de défaillance, car un certain pourcentage des défaillances n'excède pas des limites supportables. L'état stable constitue le repère que vous observerez pendant l'expérience, qui mettra en évidence des anomalies si votre hypothèse formulée dans l'étape suivante ne donne pas les résultats escomptés.

Par exemple, un état stable d'un système de paiements peut être défini comme le traitement de 300 TPS avec un taux de réussite de 99 % et un temps de transmission aller-retour de 500 ms.

b. Formuler une hypothèse sur la façon dont la charge de travail réagira à la défaillance.

Une bonne hypothèse repose sur la façon dont la charge de travail est destinée à réduire la défaillance pour maintenir l'état stable. L'hypothèse indique que vu qu'il s'agit d'une défaillance d'un type particulier, le système ou la charge de travail maintiendra un état stable, car la charge de travail a été conçue avec une atténuation des risques spécifique. Le type particulier de défaillance et d'atténuation des risques doit être spécifié dans l'hypothèse.

Le modèle suivant peut être utilisé pour l'hypothèse (mais une autre formulation est aussi acceptable) :

> **Note**
>
> En cas de *`panne spécifique`*, le *`nom de la charge de travail décrira les contrôles d'atténuation`* visant à maintenir l'*`impact des métriques commerciales ou techniques`*.

Par exemple :

- Si 20 % des nœuds du node-group Amazon EKS sont supprimés, l'API Transaction Create continue de répondre au 99e centile des demandes en moins de 100 ms (état stable). Les nœuds Amazon EKS seront opérationnels dans les cinq minutes, et les pods seront planifiés et traiteront le trafic huit minutes après le début de l'expérience. Les alertes se déclencheront sous trois minutes.
- En cas de défaillance d'une seule instance Amazon EC2, la surveillance de l'état Elastic Load Balancing du système de commandes permet à Elastic Load Balancing d'envoyer uniquement des demandes aux instances saines restantes, tandis qu'Amazon EC2 Auto Scaling remplace l'instance en échec, tout en maintenant une augmentation des erreurs (5xx) côté serveur (état stable) inférieure à 0,01 %.

- Si l'instance de base de données Amazon RDS principale échoue, la charge de travail de collecte des données Chaîne d'approvisionnement basculera et se connectera à l'instance de base de données Amazon RDS de secours pour maintenir les erreurs de lecture ou d'écriture de base de données (état stable) inférieures à 1 minute.

c. Exécuter l'expérience en injectant la défaillance.

Une expérience doit par défaut être sécurisée et tolérée par la charge de travail. Si vous savez que la charge de travail va échouer, n'exécutez pas l'expérience. L'ingénierie du chaos doit être utilisée pour rechercher les risques connus ou inconnus. Les risques connus sont des choses dont vous êtes conscient mais que vous ne comprenez pas complètement, et les risques inconnus sont des choses dont vous n'êtes pas conscient et que vous ne comprenez pas complètement. Exécuter une expérience sur une charge de travail que vous savez défaillante ne vous apportera rien de plus. Votre expérience doit être soigneusement préparée, disposer d'un champ d'impact défini et fournir un mécanisme de protection pouvant être appliqué en cas de perturbations inattendues. Si votre vérification préalable indique que votre charge de travail doit résister à l'expérience, exécutez cette dernière. Il existe plusieurs moyens d'injecter les défaillances. Pour les charges de travail sur AWS, AWS FIS fournit de nombreuses simulations de pannes prédéfinies appelées actions. Vous pouvez également définir des actions personnalisées qui s'exécutent dans AWS FIS à l'aide des documents AWS Systems Manager.

Nous déconseillons l'utilisation de scripts personnalisés pour les expériences de chaos, sauf si ces derniers sont capables de comprendre l'état actuel de la charge de travail, d'émettre des journaux, de fournir des mécanismes de protection pour annuler une expérience et des conditions d'arrêt dans la mesure du possible.

Un framework ou des outils efficaces capables de prendre en charge l'ingénierie du chaos doivent suivre l'état actuel d'une expérience, émettre des journaux et fournir des mécanismes de protection pour prendre en charge l'exécution contrôlée d'une expérience. Commencez par un service établi comme AWS FIS qui vous permet d'exécuter des expériences avec un champ clairement défini et des mécanismes de sécurité capables de protéger l'expérience en cas de perturbations inattendues. Pour en savoir plus sur une plus grande variété d'expériences utilisant AWS FIS, consultez également l'atelier Applications résilientes et Well-Architected avec l'ingénierie du chaos. AWS Resilience Hub analysera votre charge de travail et créera des expériences que vous pourrez choisir d'implémenter et d'exécuter dans AWS FIS.

> **Note**
>
> Pour chaque expérience, vous devez bien comprendre le champ et son impact. Nous recommandons que les défaillances soient d'abord simulées sur un environnement hors production avant d'être exécutées en production.

Les expériences doivent être menées en production sous une charge réelle à l'aide de déploiements Canary qui permettent de déployer à la fois un système de contrôle et un déploiement de système expérimental, dans la mesure du possible. L'exécution d'expériences pendant les heures creuses est une bonne pratique pour réduire l'impact potentiel de la première expérience en production. De plus, si l'utilisation du trafic client réel s'avère trop risquée, vous pouvez exécuter des expériences à l'aide du trafic synthétique sur l'infrastructure de production pour des déploiements de système de contrôles et d'expériences. Lorsqu'une exécution en production n'est pas possible, exécutez les expériences dans des environnements de pré-production aussi proches que possible de la production.

Vous devez définir des garde-fous pour veiller à ce que l'expérience n'impacte pas le trafic de la production ou d'autres systèmes au-delà des limites acceptables. Définissez des conditions d'arrêt pour stopper une expérience si elle atteint le seuil d'une métrique de barrière de protection défini par vos soins. Ces conditions doivent inclure les métriques de l'état stable de la charge de travail, ainsi que celles sur les composants dans lesquels vous injectez la défaillance. Un moniteur synthétique (également appelée un utilisateur canary) est une métrique que vous devez généralement inclure en tant que proxy utilisateur. Les conditions d'arrêt de AWS FIS sont prises en charge dans le cadre d'un modèle de test, autorisant jusqu'à cinq conditions d'arrêt par modèle.

L'un des principes de l'ingénierie du chaos est de minimiser le champ de l'expérience et son impact :

Bien qu'un impact négatif à court terme soit autorisé, l'ingénieur du chaos a la responsabilité et l'obligation de minimiser et de maîtriser les conséquences des expériences.

Pour vérifier le champ et l'impact potentiel, vous pouvez dans un premier temps exécuter l'expérience dans un environnement hors production, en vérifiant que les seuils des conditions d'arrêt s'activent comme prévu pendant l'expérience et que l'observabilité est en place pour détecter une exception, plutôt que d'exécuter l'expérience directement en production.

Lorsque vous exécutez des expériences d'injection de pannes, vérifiez que toutes les parties responsables sont bien informées. Communiquez avec les équipes appropriées, telles que les équipes en charge des opérations, les équipes chargées de la fiabilité du service et le service client pour leur indiquer quand les expériences seront exécutées et à quoi ils doivent s'attendre. Donnez à ces équipes les outils de communication nécessaires pour informer les personnes en charge de l'exécution de l'expérience si elles constatent des effets négatifs.

Vous devez restaurer la charge de travail et ses systèmes sous-jacents dans leur état fonctionnel et connu d'origine. En général, la conception résiliente de la charge de travail lui permet de s'auto-réparer. Cependant, certaines conceptions défaillantes ou échecs d'expériences peuvent laisser votre charge de travail dans un état d'échec inattendu. À la fin de l'expérience, vous devez en être conscient et restaurer la charge de travail et les systèmes. Avec AWS FIS, vous pouvez définir une configuration de barrière de protection (également appelée post action) dans les paramètres d'action. Une post action restaure la cible dans l'état dans lequel elle se trouvait avant l'exécution de l'action. Qu'elles soient automatisées (comme lorsque vous utilisez AWS FIS) ou manuelles, ces post actions doivent faire partie d'un playbook décrivant la façon de détecter et de gérer les échecs.

d. Vérifier l'hypothèse.

Principes de l'ingénierie du chaos donne des conseils sur la façon de vérifier l'état stable de votre charge de travail :

Concentrez-vous sur le résultat mesurable d'un système, plutôt que sur les attributs internes du système. Les mesures de ce résultat sur une courte période de temps constituent un proxy pour l'état stable du système. Le débit général du système, les taux d'erreur et les centiles de latence peuvent tous être des métriques d'intérêt représentant un comportement d'état stable. En se focalisant sur les modèles de comportement systémique pendant les expériences, l'ingénierie du chaos vérifie que le système fonctionne, au lieu d'essayer de confirmer qu'il fonctionne.

Dans nos deux exemples précédents, nous incluons la métrique de l'état stable inférieure à 0,01 % d'augmentation des erreurs (5xx) côté serveur et la métrique inférieure à 1 minute d'erreurs de lecture ou d'écriture de base de données.

Les erreurs 5xx constituent une bonne métrique, car elles sont une conséquence du mode de défaillance dont le client de la charge de travail fera l'expérience directement. La mesure des erreurs de base de données est correcte en tant que conséquence directe de la défaillance, mais doit être également complétée par une mesure d'impact, telle que les échecs de

demandes client ou les erreurs remontées. Par ailleurs, incluez une surveillance synthétique (également appelée utilisateur canary) sur n'importe quelle API ou URI directement accessible par le client de votre charge de travail.

e. Améliorer la conception de la charge de travail à des fins de résilience.

   Si l'état stable n'a pas été maintenu, enquêtez sur les moyens d'améliorer la conception de la charge de travail afin de réduire la défaillance, tout en appliquant les bonnes pratiques du pilier AWS Well-Architected Reliability. Vous trouverez des conseils et des ressources supplémentaires dans la AWS Builder's Library, qui contient des articles sur la manière d'améliorer vos surveillances de l'état ou d'utiliser des nouvelles tentatives avec retard dans le code de votre application, entre autres.

   Une fois ces changements implémentés, exécutez de nouveau l'expérience (illustrée par la ligne pointillée dans le volant d'inertie de l'ingénierie du chaos) pour déterminer son efficacité. Si l'étape de vérification indique que l'hypothèse est vraie, alors la charge de travail sera en état stable et le cycle continuera.

4. Exécuter régulièrement des expériences.

   Une expérience de chaos est un cycle, et les expériences doivent être exécutées régulièrement dans le cadre de l'ingénierie du chaos. Lorsqu'une charge de travail correspond à l'hypothèse d'une expérience, cette dernière doit être automatisée pour s'exécuter en continu en tant que test de régression de votre pipeline CI/CD. Pour savoir comment procéder, consultez ce blog sur la façon de mener des expériences AWS FIS à l'aide d'AWS CodePipeline. Ce laboratoire sur les expériences AWS FIS récurrentes dans un pipeline CI/CD vous permet de travailler sur le terrain.

   Les expériences d'injection de pannes font également partie des tests de simulation de pannes (consultez REL12-BP05 Organiser régulièrement des tests de simulation de panne). Les tests de simulation de pannes simulent une défaillance ou un événement pour vérifier les systèmes, les processus et la réponse de l'équipe. L'objectif est d'effectuer les actions que l'équipe effectuerait si un événement exceptionnel se produisait.

5. Enregistrer et stocker les résultats des expériences.

   Les résultats des expériences d'injection de pannes doivent être enregistrés et conservés. Incluez toutes les données nécessaires (telles que l'heure, la charge de travail et les conditions) afin de pouvoir analyser ultérieurement les résultats de l'expérience et les tendances. Les exemples de résultats peuvent inclure des captures d'écran des tableaux de bord, des fichiers CSV de la base de données de votre/vos métriques ou un registre manuscrit des événements et observations

pendant l'expérience. La journalisation des expériences avec AWS FIS peut faire partie de cette capture de données.

Ressources

Bonnes pratiques associées :

- REL08-BP03 Intégrez les tests de résilience dans le cadre de votre déploiement
- REL13-BP03 Tester la mise en œuvre de la reprise après sinistre pour valider la mise en œuvre

Documents connexes :

- Présentation de AWS Fault Injection Service
- Présentation de AWS Resilience Hub
- Principes de l'ingénierie du chaos
- Ingénierie du chaos : préparation de votre première expérience
- Ingénierie de résilience : apprendre à intégrer les pannes
- Témoignages d'utilisation de l'ingénierie du chaos
- Éviter les solutions de secours dans les systèmes distribués
- Déploiement canary pour des expériences de chaos

Vidéos connexes :

- AWS re:Invent 2020: Testing resiliency using chaos engineering (ARC316)
- AWS re:Invent 2019: Improving resiliency with chaos engineering (DOP309-R1)
- AWS re:Invent 2019: Performing chaos engineering in a serverless world (CMY301)

Exemples connexes :

- Atelier Well-Architected : niveau 300 : test de la résilience d'Amazon EC2, Amazon RDS et Amazon S3
- Atelier L'ingénierie du chaos dans AWS
- Atelier Applications résilientes et Well-Architected avec l'ingénierie du chaos
- Atelier Chaos sans serveur

- Atelier Mesurer et améliorer la résilience de votre application avec AWS Resilience Hub

Outils associés :

- AWS Fault Injection Service
- AWS Marketplace : Gremlin Chaos Engineering Platform
- Chaos Toolkit
- Chaos Mesh
- Litmus

REL12-BP05 Organiser régulièrement des tests de simulation de panne

Organisez des tests de simulation de panne pour tester régulièrement vos procédures de réponse aux événements et aux déficiences ayant un impact sur la charge de travail. Impliquez les mêmes équipes qui seraient chargées de traiter les scénarios de production. Ces exercices permettent de mettre en œuvre des mesures visant à prévenir l'impact des événements de production sur les utilisateurs. Lorsque vous mettez en pratique vos procédures de réponse dans des conditions réalistes, vous pouvez identifier et corriger toute lacune ou faiblesse avant l'avènement d'un événement réel.

Les tests de simulation de panne simulent des événements dans des environnements de type production pour tester les systèmes, les processus et la réponse de votre équipe. L'objectif est d'effectuer les mêmes actions que l'équipe effectuerait si l'événement se produisait réellement. Ces exercices vous aident à comprendre où apporter des améliorations et comment développer une expérience de gestion des événements et des déficiences au sein de votre organisation. Ces exercices doivent être effectués régulièrement afin que votre équipe développe des automatismes pour mieux réagir.

Les tests de simulation de panne préparent les équipes à gérer les événements de production en toute confiance. Les équipes expérimentées sont plus à même de détecter différents scénarios et d'y réagir rapidement. Cela se traduit par une amélioration significative de l'état de préparation et de la posture de résilience.

Résultat escompté : vous planifiez et effectuez régulièrement des tests de simulation sur la résilience. Ces tests de simulation de panne sont considérés comme un élément normal et attendu de l'activité de l'entreprise. Votre organisation a développé une culture de préparation et lorsque des problèmes

de production surviennent, vos équipes sont bien préparées pour réagir promptement, résoudre efficacement les problèmes et atténuer leur impact sur les clients.

Anti-modèles courants :

- Vous documentez vos procédures sans jamais vous exercer à les appliquer.
- Vous excluez les décideurs d'entreprise des exercices de test.
- Vous organisez des tests de simulation de panne, mais vous n'informez pas toutes les parties prenantes concernées.
- Vous vous concentrez uniquement sur les défaillances techniques, mais vous n'impliquez pas les parties prenantes de l'entreprise.
- Vous n'incorporez pas les leçons apprises lors des tests de simulation de panne dans vos processus de reprise.
- Vous blâmez les équipes pour les échecs et les bogues.

Avantages liés au respect de cette bonne pratique :

- Amélioration des compétences en matière de réponse : lors des tests de simulation de panne, les équipes s'exercent à réaliser leurs tâches et testent leurs mécanismes de communication, ce qui leur permet de réagir de façon plus coordonnée et efficace dans le cadre de situations de production.
- Identification et traitement des dépendances : les environnements complexes impliquent souvent des dépendances complexes entre différents systèmes, services et composants. Les tests de simulation de panne peuvent vous aider à identifier et à traiter ces dépendances, ainsi qu'à vérifier que vos systèmes et services critiques sont correctement couverts par vos procédures de dossier d¦exploitation et peuvent être augmentés verticalement ou récupérés en temps opportun.
- Promotion d'une culture de résilience : les tests de simulation de panne peuvent aider à développer un état d'esprit de résilience au sein d'une organisation. Lorsqu'ils impliquent des parties prenantes et des équipes interfonctionnelles, ces exercices favorisent la prise de conscience, la collaboration et une compréhension commune de l'importance de la résilience dans l'ensemble de l'organisation.
- Amélioration et adaptation continues : des tests de simulation de panne réguliers vous aident à évaluer en permanence vos stratégies de résilience et à les adapter, afin de les maintenir pertinentes et efficaces face à des circonstances changeantes.

- Renforcement de la confiance dans le système : des tests de simulation de panne réussis peuvent vous aider à renforcer la confiance dans la capacité du système à résister aux perturbations et à s'en remettre.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Une fois que vous avez conçu et mis en œuvre les mesures de résilience nécessaires, organisez des tests de simulation de panne pour confirmer que tout fonctionne comme prévu en production. Les tests de simulation de panne, en particulier la première fois, doivent impliquer tous les membres de l'équipe, et l'ensemble des parties prenantes et des participants doivent être informés à l'avance de la date, de l'heure et des scénarios simulés.

Pendant les tests de simulation de panne, les équipes impliquées simulent divers événements et scénarios potentiels conformément aux procédures prescrites. Les participants surveillent de près et évaluent l'impact de ces événements simulés. Si le système fonctionne comme prévu, les mécanismes automatisés de détection, de mise à l'échelle et de réparation automatique devraient s'activer et n'entraîner que peu ou pas d'impact sur les utilisateurs. Si l'équipe constate un impact négatif, elle doit annuler le test et résoudre les problèmes identifiés, soit par des moyens automatisés, soit par une intervention manuelle documentée dans les dossiers d¦exploitation applicables.

Pour améliorer continuellement la résilience, il est essentiel de documenter et d'incorporer les leçons apprises. Ce processus constitue une boucle de rétroaction qui capture systématiquement les informations exploitables recueillies pendant les tests de simulation de panne et les utilise pour améliorer les systèmes, les processus et les compétences des équipes.

Pour vous aider à reproduire des scénarios réels dans lesquels des services ou des composants du système peuvent tomber en panne de façon inattendue, injectez des défauts simulés dans un exercice de test de simulation. Les équipes peuvent tester la résilience et la tolérance aux pannes de leurs systèmes et simuler leurs processus de réponse aux incidents et de reprise dans un environnement contrôlé.

Dans AWS, vos tests de simulation de panne peuvent être réalisés avec des réplicas de votre environnement de production en utilisant une infrastructure en tant que code. Ce processus vous permert d'effectuer des tests dans un environnement sûr qui ressemble étroitement à votre environnement de production. Envisagez d'utiliser AWS Fault Injection Service pour créer différents scénarios de panne. Utilisez des services tels qu'Amazon CloudWatch et AWS X-Ray pour surveiller

le comportement du système pendant les tests de simulation de panne. Utilisez AWS Systems Manager pour gérer et exécuter les playbooks, et AWS Step Functions pour orchestrer les flux de travail récurrents des tests de simulation de panne.

Étapes d'implémentation

- Établissez un programme de tests de simulation de panne : élaborez un programme structuré qui définit la fréquence, la portée et les objectifs des tests de simulation de panne. Impliquez les principales parties prenantes et les experts du domaine concerné dans la planification et l'exécution de ces exercices.
- Préparez les tests de simulation de panne :
  1. Identifiez les principaux services essentiels à l'entreprise qui seront au centre des tests de simulation de panne. Cataloguez et cartographiez les personnes, les processus et les technologies qui prennent en charge ces services.
  2. Définissez l'ordre du jour des tests de simulation de panne et préparez les équipes impliquées à participer à l'événement. Préparez vos services d'automatisation pour simuler les scénarios planifiés et exécuter les processus de récupération appropriés. Les services AWS tels que AWS Fault Injection Service, AWS Step Functions et AWS Systems Manager peuvent vous aider à automatiser divers aspects des tests de simulation de panne, tels que l'injection des défauts et le lancement des actions de récupération.
- Exécutez votre simulation : dans le cadre des tests de simulation de panne, exécutez le scénario planifié. Observez et documentez la façon dont les personnes, les processus et les technologies réagissent à l'événement simulé.
- Réalisez le bilan de l'exercice : après les tests de simulation de panne, organisez une séance rétrospective pour passer en revue les enseignements tirés. Identifiez les domaines d'amélioration et les actions nécessaires pour améliorer la résilience opérationnelle. Consignez vos résultats et effectuez le suivi des modifications nécessaires pour améliorer vos stratégies de résilience et votre préparation aux travaux à entreprendre.

Ressources

Bonnes pratiques associées :

- REL12-BP01 Utiliser des playbooks pour enquêter sur les causes des défaillances
- REL12-BP04 Tester la résilience à l'aide de l'ingénierie du chaos
- OPS04-BP01 Identification des indicateurs clés de performance

- OPS07-BP03 Utilisation de runbooks pour effectuer des procédures
- OPS10-BP01 Utilisation d'un processus pour la gestion des événements, des incidents et des problèmes

Documents connexes :

- Qu'est-ce qu'AWS GameDay ?
- Concepts AWS Well-Architected – Tests de simulation de panne

Vidéos connexes :

- AWS re:Invent 2023 - Practice like you play: How Amazon scales resilience to new heights

Exemples connexes :

- Atelier AWS – Surmonter la tempête : déclenchement d'un chaos contrôlé pour systèmes résilients
- Élaboration de tests de simulation de panne en vue de soutenir la résilience opérationnelle

FIA 13. Comment planifier la reprise après sinistre (DR) ?

La mise en place de sauvegardes et de composants de charge de travail redondants constitue le début de votre stratégie de DR. L'objectif de délai de reprise (RTO) et l'objectif de point de reprise (RPO) sont vos objectifs pour la restauration de votre charge de travail. Définissez-les en fonction des besoins de l'entreprise. Mettez en œuvre une stratégie pour atteindre ces objectifs, en particulier en tenant compte de l'emplacement et de la fonction des données et des ressources de charge de travail. La probabilité d'une perturbation et le coût de la reprise sont également des facteurs clés qui permettent de déterminer la valeur opérationnelle de la reprise après sinistre d'une charge de travail.

Bonnes pratiques

- REL13-BP01 Définir les objectifs de reprise en termes de durée d'indisponibilité et de perte de données
- REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise
- REL13-BP03 Tester la mise en œuvre de la reprise après sinistre pour valider la mise en œuvre
- REL13-BP04 Gérer la dérive de configuration au niveau du site ou de la région de reprise après sinistre

- REL13-BP05 Automatiser la reprise

REL13-BP01 Définir les objectifs de reprise en termes de durée d'indisponibilité et de perte de données

Les défaillances peuvent avoir un impact sur votre activité de plusieurs manières. Tout d'abord, les défaillances peuvent entraîner une interruption de service (durée d'indisponibilité). Ensuite, les défaillances peuvent entraîner la perte, l'incohérence ou l'obsolescence des données. Pour déterminer la manière de réagir et de récupérer après une défaillance, définissez un objectif de délai de reprise (RTO) et un objectif de point de reprise (RPO) pour chaque charge de travail. L'objectif de délai de reprise (RTO) est le délai maximal acceptable entre l'interruption du service et son rétablissement. L'objectif de point de reprise (RPO) est le temps maximal acceptable après le dernier point de récupération des données.

Résultat escompté : chaque charge de travail dispose d'un RTO et d'un RPO basés sur des considérations techniques et l'impact sur l'activité.

Anti-modèles courants :

- Vous n'avez pas défini d'objectifs de récupération.
- Vous sélectionnez des objectifs de récupération arbitraires.
- Vous sélectionnez des objectifs de récupération trop souples qui ne répondent pas aux objectifs de l'entreprise.
- Vous n'avez pas évalué l'impact de la durée d'indisponibilité et de la perte de données.
- Vous sélectionnez des objectifs de récupération non réalistes pour la configuration de votre charge de travail, tels qu'un délai de récupération nul ou une perte de données nulle, qui ne sont pas réalisables.
- Vous sélectionnez des objectifs de récupération plus rigoureux que les objectifs métier réels. Cela entraîne une mise en œuvre de la récupération plus coûteuse et plus compliquée que ce dont la charge de travail a besoin.
- Vous sélectionnez des objectifs de récupération incompatibles avec ceux d'une charge de travail dépendante.
- Vous ne tenez pas compte des exigences réglementaires et de conformité.

Avantages liés au respect de cette bonne pratique : lorsque vous définissez des RTO et des RPO pour vos charges de travail, vous définissez des objectifs de récupération clairs et mesurables en

fonction des besoins de votre entreprise. Une fois ces objectifs définis, vous pouvez créer des plans de reprise après sinistre (DR) spécialement conçus pour les atteindre.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Créez une matrice ou une fiche qui vous aidera à planifier la reprise après sinistre. Dans cette matrice, créez différentes catégories ou niveaux de charge de travail en fonction de leur impact sur l'activité (critique, élevé, moyen ou faible) et des RTO et RPO associés à cibler pour chacun d'entre eux. La matrice suivante fournit un exemple (notez que vos valeurs RTO et RPO peuvent différer) que vous pouvez suivre :

Matrice de reprise après sinistre						
		Objectif de point de reprise				
		Moins de 1 minute	Moins de 1 heure	Moins de 6 heures	Moins de 1 jour	+ de 1 jour
Durée maximale d'interruption	Moins de 10 minutes	Critique	Critique	Débit	Moyenne entreprise	Moyenne entreprise
	Moins de 2 heures	Critique	Débit	Moyenne entreprise	Moyenne entreprise	Faible
	Moins de 8 heures	Débit	Moyenne entreprise	Moyenne entreprise	Faible	Faible
	Moins de 24 heures	Moyenne entreprise	Moyenne entreprise	Faible	Faible	Faible
	+ de 24 heures	Moyenne entreprise	Faible	Faible	Faible	Faible

Exemple de matrice de reprise après sinistre

Pour chaque charge de travail, vous devez étudier et comprendre l'impact de la durée d'indisponibilité et de la perte de données sur votre activité. Cet impact augmente généralement avec la durée d'indisponibilité et la perte de données, mais sa forme peut varier en fonction du type de charge de travail. Par exemple, une durée d'indisponibilité d'une heure peut avoir un impact faible, mais après cela, l'impact peut croître rapidement. L'impact peut prendre de nombreuses formes, y compris la forme d'un impact financier (tel qu'une perte de chiffre d'affaires), d'un impact sur la réputation (y compris la perte de confiance des clients), d'un impact opérationnel (comme une paie manquée ou une baisse de productivité) et d'un risque réglementaire. Une fois terminé, attribuez la charge de travail au niveau approprié.

Lorsque vous analysez l'impact d'une panne, posez-vous les questions suivantes :

1. Quelle est la durée maximale d'indisponibilité de la charge de travail avant qu'elle n'ait un impact inacceptable sur l'activité ?
2. Quelle est l'ampleur et la nature de l'impact d'une interruption de la charge de travail sur l'entreprise ? Tenez compte de tous les types d'impact, notamment financier, de réputation, opérationnel et réglementaire.
3. Quelle est la quantité maximale de données pouvant être perdues ou irrécupérables avant que l'activité ne subisse un impact inacceptable ?
4. Les données perdues peuvent-elles être recréées à partir d'autres sources (également appelées données dérivées) ? Si tel est le cas, considérez également les RPO de toutes les données sources utilisées pour recréer les données de la charge de travail.
5. Quels sont les objectifs de récupération et les attentes de disponibilité des charges de travail dont celle-ci dépend (en aval) ? Vos objectifs en matière de charge de travail doivent être réalisables compte tenu des capacités de récupération de ses dépendances en aval. Envisagez des solutions de contournement ou d'atténuation des dépendances en aval susceptibles d'améliorer la capacité de récupération de cette charge de travail.
6. Quels sont les objectifs de récupération et les attentes de disponibilité des charges de travail qui dépendent de celle-ci (en amont) ? Les objectifs de charge de travail en amont peuvent exiger que cette charge de travail soit dotée de capacités de récupération plus strictes qu'il n'y paraît initialement.
7. Les objectifs de récupération sont-ils différents en fonction du type d'incident ? Par exemple, vous pouvez avoir des RTO et des RPO différents selon que l'incident a un impact sur une zone de disponibilité ou sur une région entière.
8. Vos objectifs de récupération changent-ils au cours de certains événements ou à certaines périodes de l'année ? Par exemple, vous pouvez avoir des RTO et des RPO différents pour les fêtes de fin d'année, lors d'événements sportifs, en périodes de soldes et lors du lancement de nouveaux produits.
9. Comment les objectifs de récupération s'alignent-ils sur la stratégie de reprise après sinistre de votre secteur d'activité et de votre organisation ?
10. Y a-t-il des ramifications juridiques ou contractuelles à prendre en compte ? Par exemple, avez-vous l'obligation contractuelle de fournir un service avec un RTO ou un RPO donné ? Quelles pénalités pourriez-vous encourir en cas de non-respect de cette obligation ?
11. Devez-vous maintenir l'intégrité des données pour répondre aux exigences réglementaires ou de conformité ?

La fiche suivante peut vous aider à évaluer chaque charge de travail. Vous pouvez modifier cette fiche en fonction de vos besoins spécifiques, par exemple en ajoutant des questions supplémentaires.

Étape 2 : Questions principales	S'applique à la charge de travail ?	RPO de la charge de travail	RPO de la charge de travail	Ajustement de RTO.	Ajustement de RPO.	Instructions
[1] durée maximale pendant laquelle la charge de travail peut être inactive						mesuré en temps depuis le début de la panne jusqu'à la récupération
[2] quantité maximale de données pouvant être perdues						mesuré dans le temps depuis le dernier jeu de données restaurable
[3a] dépendances en amont						saisissez les objectifs de récupération en amont les plus stricts
[3b] dépendances en aval						saisissez les objectifs de récupération en aval les moins stricts
[3a] dépendances en amont rapprochées						Si la valeur en amont est inférieure aux valeurs actuelles et la valeur en aval supérieure,
[3b] dépendances en aval rapprochées						manipulez les dépendances pour les rapprocher et entrez les valeurs rapprochées ici
[3] dépendances						réduisez les valeurs pour répondre aux dépendances en amont ou augmentez-les selon les fonctionnalités de dépendance en aval
Étape 2 : Questions supplémentaires						Indiquez si la question s'applique. Dans le cas contraire, ignorez-la
RTO/RPO de base						Transférez les valeurs RTO et RPO d'en haut jusqu'ici
[4] type de panne	[ ]O / [ ]N					Indiquez les objectifs de récupération pour les événements avec les exigences les plus strictes
[5] objectifs temporels spécifiques	[ ]O / [ ]N					Indiquez les objectifs de récupération pour les durées avec les exigences les plus strictes
[6] clients perturbés	[ ]O / [ ]N					Représentez graphiquement les clients impactés en fonction du temps d'arrêt ou de la perte de données. Utilisez ces informations pour saisir le RTO et le RPO maximum autorisés en fonction de l'impact sur le client
[7] impact sur la réputation	[ ]O / [ ]N					Déterminez avec l'entreprise le RTO et le RPO maximum en fonction de l'impact sur la réputation
[8] impact opérationnel	[ ]O / [ ]N					Indiquez un RTO et un RPO maximum en fonction de l'impact opérationnel
[9] alignement organisationnel	[ ]O / [ ]N					Indiquez le RTO et le RPO maximum pour les charges de travail de ce type selon les exigences LOB et organisationnelles
[10] obligations contractuelles	[ ]O / [ ]N					Indiquez un RTO et un RPO maximum en fonction des obligations contractuelles
[11] conformité réglementaire	[ ]O / [ ]N					Indiquez le RTO et le RPO maximum en fonction de la conformité réglementaire applicable
cible basée sur des questions supplémentaires						Prenez la valeur minimale (valeur plus stricte) des Q 11-4 et entrez-la ici
cible ajustée						Si les objectifs de la ligne ci-dessus ne peuvent pas être atteints, collaborez avec les parties prenantes pour assouplir les contraintes et entrez un nouveau minimum ici
RTO/RPO ajusté						Indiquez les valeurs RPO/RTO de base, ou la cible ajustée, selon la valeur la plus basse
Étape 3						
Mapper vers une catégorie ou un niveau prédéfini						Ajustez les deux valeurs vers le bas (méthode plus stricte) pour vous aligner sur le niveau défini le plus proche

Fiche

Étapes d'implémentation

1. Identifiez les parties prenantes et les équipes techniques responsables de chaque charge de travail et interagissez avec elles.
2. Créez des catégories ou des niveaux de criticité pour déterminer l'impact de la charge de travail dans votre organisation. Exemples de catégories : critique, élevé, moyen et faible. Pour chaque catégorie, choisissez un RTO et un RPO qui reflètent les objectifs et les exigences de votre activité.
3. Attribuez l'une des catégories d'impact que vous avez créées à l'étape précédente à chaque charge de travail. Pour déterminer la correspondance d'une charge de travail à une catégorie, tenez compte de l'importance de la charge de travail pour l'activité et de l'impact de son interruption ou d'une perte de données, et utilisez les questions ci-dessus pour vous guider. Cela se traduit par un RTO et un RPO pour chaque charge de travail.

4. Pour chaque charge de travail, tenez compte du RTO et du RPO déterminés à l'étape précédente. Impliquez les équipes stratégiques et techniques chargées de la charge de travail afin de déterminer si les objectifs doivent être ajustés. Par exemple, les parties prenantes de l'entreprise peuvent déterminer que des objectifs plus stricts sont requis. Par ailleurs, les équipes techniques peuvent décider de la nécessité de modifier les cibles pour les rendre réalisables dans le cadre des contraintes technologiques et des ressources disponibles.

Ressources

Bonnes pratiques associées :

- REL09-BP04 Effectuer une récupération périodique des données pour vérifier l'intégrité et les processus de sauvegarde
- REL12-BP01 Utiliser des playbooks pour enquêter sur les causes des défaillances
- REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise
- REL13-BP03 Effectuer un test de validation de la mise en œuvre de la reprise après sinistre

Documents connexes :

- AWS Architecture Blog: Disaster Recovery Series
- Disaster Recovery of Workloads on AWS: Recovery in the Cloud (AWS Whitepaper)
- Gestion des politiques de résilience avec AWS Resilience Hub
- Partenaire APN : partenaires pouvant faciliter la reprise après sinistre
- AWS Marketplace : produits pouvant être utilisés pour la reprise après sinistre

Related videos:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications
- Disaster Recovery of Workloads on AWS

REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise

Définissez une stratégie de reprise après sinistre qui répond aux objectifs de reprise de votre charge de travail. Choisissez une stratégie telle que : sauvegarde et restauration, mode secours (actif/passif) ou actif/actif.

Résultat escompté : pour chaque charge de travail, il existe une stratégie de reprise après sinistre définie et implémentée qui permet à cette charge de travail d'atteindre les objectifs de reprise. Les stratégies de reprise après sinistre entre les charges de travail utilisent des modèles réutilisables (comme les stratégies décrites précédemment).

Anti-modèles courants :

- Mettre en œuvre des procédures de récupération incohérentes pour les charges de travail avec des objectifs de reprise après sinistre similaires.
- Conserver l'implémentation ad hoc de la stratégie de reprise après sinistre lorsqu'un sinistre se produit.
- Ne pas avoir de plan de reprise après sinistre.
- Être dépendant des opérations du plan de contrôle pendant la récupération.

Avantages liés au respect de cette bonne pratique :

- L'utilisation de stratégies de reprise définies vous permet d'utiliser des outils et des procédures de test courantes.
- L'utilisation de stratégies de reprise définies améliore le partage des connaissances entre les équipes et la mise en œuvre de la reprise après sinistre sur les charges de travail qu'elles possèdent.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé Sans une stratégie de reprise après sinistre planifiée, mise en œuvre et testée, il est peu probable que vous atteigniez vos objectifs de reprise en cas de sinistre.

Directives d'implémentation

Une stratégie de reprise après sinistre repose sur la capacité à rétablir votre charge de travail sur un site de reprise si votre emplacement principal ne parvient plus à exécuter cette charge de travail. Les objectifs de récupération les plus courants sont le RTO et le RPO, comme indiqué dans REL13-BP01 Définir les objectifs de reprise en termes de durée d'indisponibilité et de perte de données.

Une stratégie de reprise après sinistre sur plusieurs zones de disponibilité (AZ) au sein d'une seule Région AWS peut vous prémunir contre les événements catastrophiques tels que les incendies, les inondations et les pannes de courant majeures. S'il est nécessaire de mettre en œuvre une protection contre un événement improbable qui empêcherait votre charge de travail de s'exécuter dans une Région AWS donnée, optez pour une stratégie de reprise après sinistre qui utilise plusieurs régions.

Lors de la conception d'une stratégie de reprise après sinistre dans plusieurs régions, vous devez choisir l'une des approches suivantes. Ils sont classés par ordre croissant de coût et de complexité, et par ordre décroissant de RTO et RPO. La région de restauration fait référence à une région Région AWS autre que la région principale utilisée pour votre charge de travail.

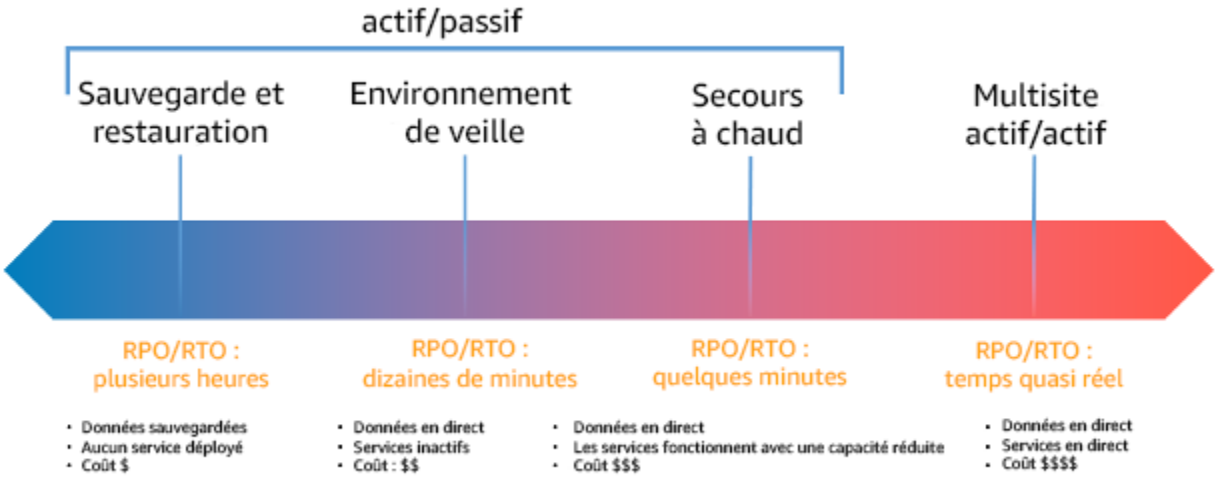


Figure 17 : stratégies de reprise après sinistre

- Sauvegarde et restauration (RPO en heures, RTO de 24 heures maximum) : sauvegardez vos données et applications dans la région de reprise après sinistre. L'utilisation de sauvegardes automatisées ou continues permet une reprise ponctuelle (PITR), ce qui peut réduire le RPO à seulement 5 minutes dans certains cas. En cas de sinistre, vous déployez votre infrastructure (en utilisant l'infrastructure en tant que code pour réduire le RTO), déployez votre code et restaurez les données sauvegardées pour vous remettre d'un sinistre dans la région de reprise.
- Veilleuse (RPO de quelques minutes, RTO de dizaines de minutes) : allouez une copie de votre infrastructure de charge de travail principale dans la région de reprise. Répliquez vos données dans la région de reprise et créez-y des sauvegardes. Les ressources requises pour prendre en charge la réplication et la sauvegarde des données, telles que les bases de données et le stockage d'objets, sont toujours actives. D'autres éléments tels que les serveurs d'applications ou le calcul sans serveur ne sont pas déployés, mais peuvent être créés si nécessaire avec la configuration et le code d'application requis.
- Secours semi-automatique (RPO de quelques secondes, RTO de quelques minutes) : maintenez une version réduite verticalement d'une charge de travail entièrement fonctionnelle qui s'exécute toujours dans la région de reprise. Les systèmes stratégiques sont entièrement dupliqués et sont

toujours opérationnels, mais avec une flotte réduite. Les données sont répliquées dans la région de reprise et y sont hébergées. Lorsque vient le moment de la reprise, le système est rapidement mis à l'échelle pour gérer la charge de production. Plus l'échelle du secours à chaud est élevée, plus la dépendance au RTO et au plan de contrôle est faible. Lorsqu'elle est entièrement mise à l'échelle, on parle de veille permanente.

- Multi-région (multi-site) active-active (RPO proche de zéro, RTO potentiellement nul) : votre charge de travail est déployée et dessert activement le trafic à partir de plusieurs Régions AWS. Cette stratégie vous oblige à synchroniser les données entre les régions. Il est important d'éviter ou de gérer les éventuels conflits causés par des écritures sur le même enregistrement dans deux réplicas régionaux différents, ce qui peut être complexe. La réplication des données est utile pour la synchronisation des données et vous protège contre certains types de sinistres. Toutefois, elle ne vous protège pas contre la corruption ou la destruction des données à moins que votre solution n'inclue également des options de récupération ponctuelle.

> **Note**
>
> La différence entre l'environnement en veille et le secours à chaud est parfois difficile à cerner. Ces deux stratégies incluent un environnement dans votre région de reprise avec des copies des ressources de votre région principale. L'environnement en veille diffère en ce qu'il ne peut pas traiter les demandes sans qu'une action supplémentaire soit entreprise au préalable, tandis que le secours à chaud peut gérer le trafic (à des niveaux de capacité réduits) immédiatement. L'environnement en veille vous oblige à allumer des serveurs, à déployer éventuellement une infrastructure supplémentaire (non essentielle) et à augmenter verticalement, tandis que le secours à chaud nécessite uniquement une augmentation verticale (tout est déjà déployé et en cours d'exécution). Choisissez entre ces options en fonction de vos besoins en matière de RTO et de RPO.
>
> Si le coût est un problème et que vous souhaitez atteindre des objectifs de RPO et RTO similaires à ceux définis dans la stratégie de secours à chaud, vous pouvez envisager des solutions natives du cloud, comme AWS Elastic Disaster Recovery, qui adoptent l'approche de l'environnement de veille et offrent des objectifs de RPO et RTO améliorés.

Étapes d'implémentation

1. Déterminez une stratégie de reprise après sinistre qui répond aux exigences de récupération pour cette charge de travail.

Le choix d'une stratégie de reprise après sinistre vise à trouver un juste milieu entre la réduction des temps d'arrêt et de la perte de données (RTO et RPO) et le coût et la complexité liées à la mise en œuvre de cette stratégie. Évitez de mettre en œuvre une stratégie plus stricte que nécessaire, car cela entraînerait des coûts inutiles.

Par exemple, dans le diagramme suivant, l'entreprise a déterminé son RTO maximal autorisé ainsi que la limite de dépenses possible pour sa stratégie de restauration de service. Compte tenu des objectifs de l'entreprise, les stratégies de reprise après sinistre en veille et secours à chaud satisfont à la fois aux critères de RTO et de coût.

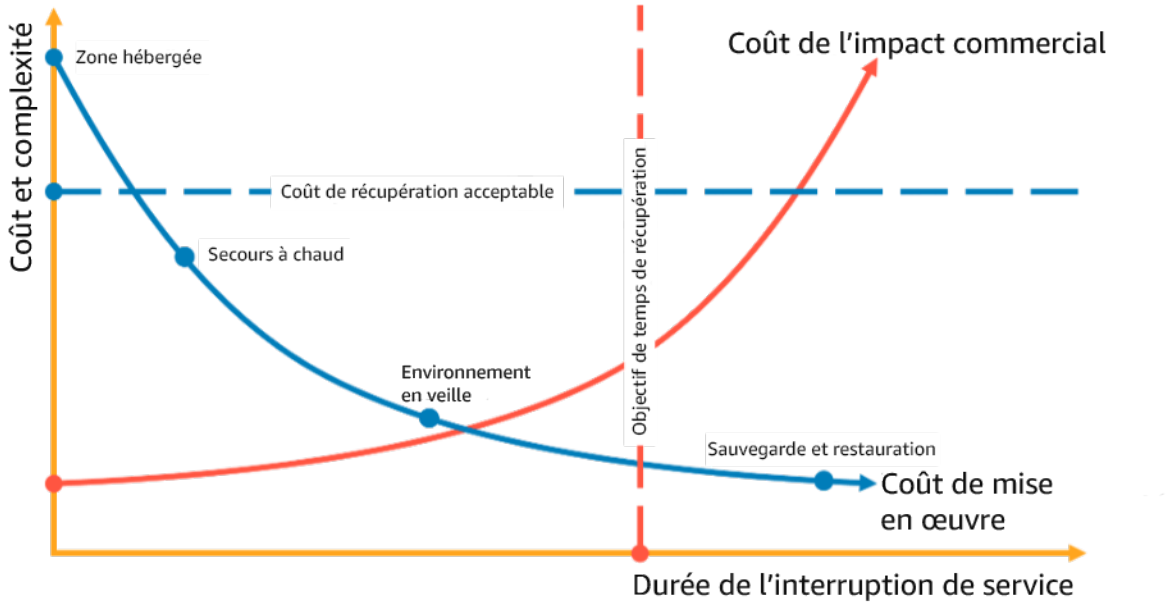


Figure 18 : choix d'une stratégie de reprise après sinistre basée sur le RTO et le coût

Pour en savoir plus, consultez Plan de continuité d'activité (BCP).

2. Passez en revue les modèles de mise en œuvre de la stratégie de reprise après sinistre sélectionnée.

   Cette étape consiste à comprendre comment mettre en œuvre la stratégie sélectionnée. Les stratégies reposent sur l'utilisation de Régions AWS comme site principal et site de reprise. Cependant, vous pouvez également choisir d'utiliser des zones de disponibilité dans une seule région comme stratégie de reprise après sinistre, ce qui permet d'exploiter des éléments de plusieurs de ces stratégies.

   Dans les étapes suivantes, vous pouvez appliquer la stratégie à votre charge de travail spécifique.

Sauvegarde et restauration

Sauvegarde et restauration est la stratégie la moins complexe à mettre en œuvre, mais nécessite plus de temps et d'efforts pour la restauration de la charge de travail, ce qui entraîne un RTO et un RPO plus élevés. Il est conseillé de toujours faire des sauvegardes de vos données et de les copier sur un autre site (comme une autre Région AWS).

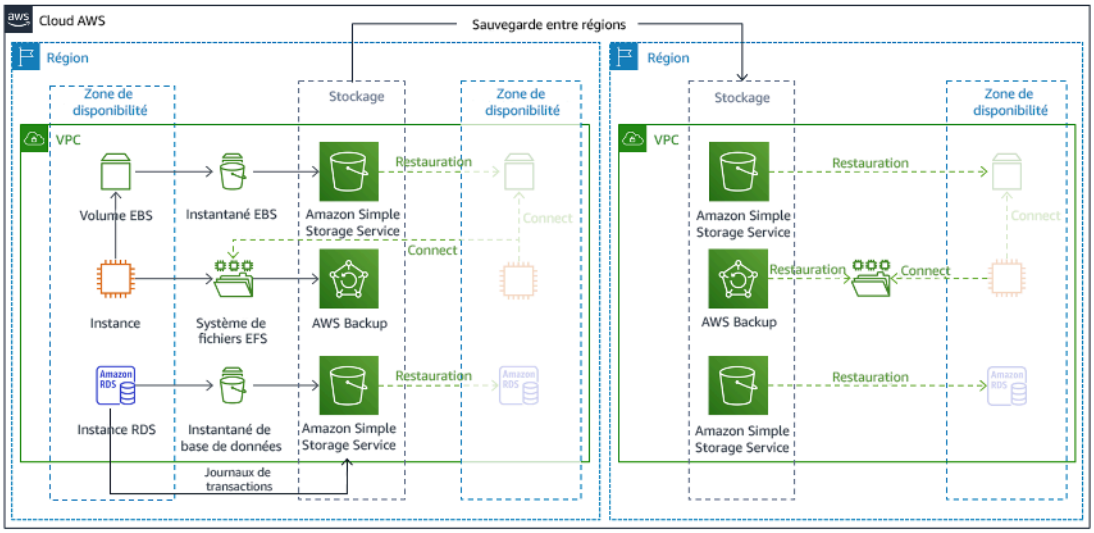


Figure 19 : architecture de sauvegarde et de restauration

Pour en savoir plus sur cette stratégie, consultezArchitecture de reprise après sinistre (DR) sur AWS, partie 2 : sauvegarde et restauration avec récupération rapide.

Veilleuse

L'approche de veilleuse, vous permet de répliquer vos données depuis la région principale vers la région de reprise. Les ressources principales utilisées pour l'infrastructure de charge de travail sont déployées dans la région de reprise, mais des ressources supplémentaires et toutes les dépendances sont toujours nécessaires pour en faire une pile fonctionnelle. Par exemple, dans la figure 20, aucune instance de calcul n'est déployée.

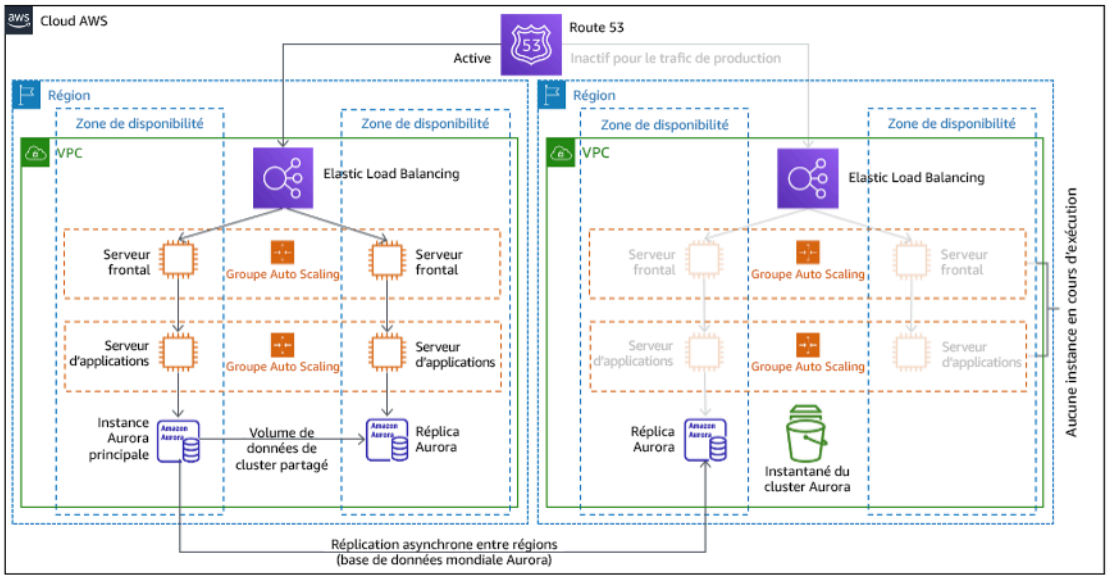


Figure 20 : architecture avec environnement en veille

Pour en savoir plus sur cette stratégie, consultez Architecture de reprise après sinistre sur AWS, partie 3 : environnement en veille et secours à chaud.

Secours semi-automatique

L'approche du secours semi-automatique consiste à s'assurer qu'il existe une copie réduite verticalement, mais entièrement fonctionnelle, de votre environnement de production dans une autre région. Cette approche étend le concept d'environnement en veille et réduit le temps de récupération, car votre charge de travail reste active dans une autre région. Si la région de reprise est déployée à pleine capacité, on parle de veille permanente.

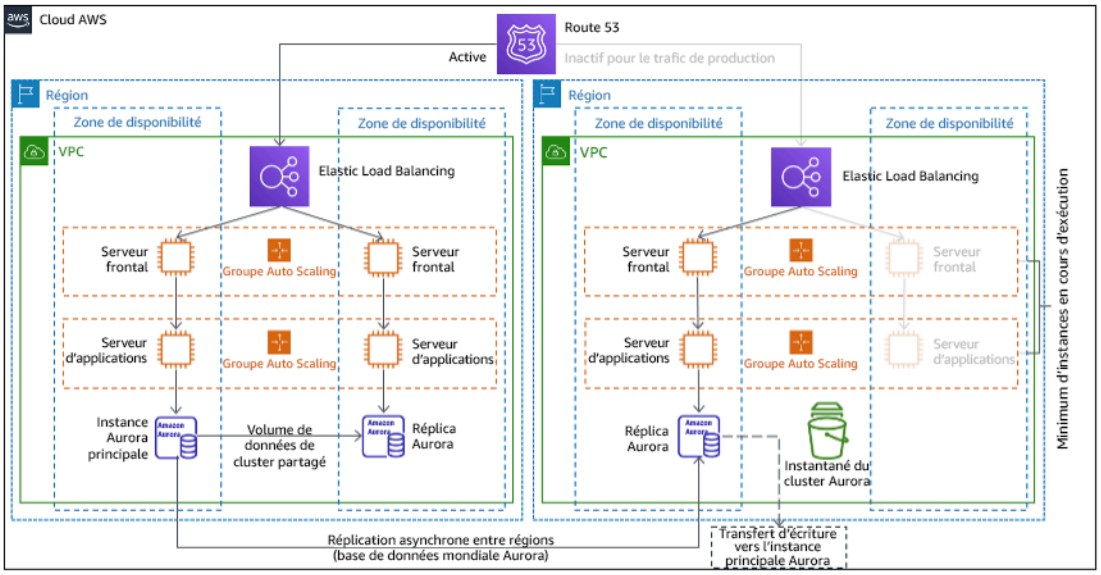


Figure 21 : Architecture de secours à chaud

L'utilisation du secours à chaud ou de l'environnement en veille nécessite une augmentation verticale des ressources dans la région de reprise. Pour vérifier que la capacité est disponible en cas de besoin, envisagez de l'utiliser pour les réservations de capacité pour les instances EC2. Si vous utilisez AWS Lambda, la simultanéité provisionnée peut fournir des environnements d'exécution afin qu'ils soient prêts à répondre immédiatement aux invocations de votre fonction.

Pour en savoir plus sur cette stratégie, consultez Architecture de reprise après sinistre sur AWS, partie 3 : environnement en veille et secours à chaud.

Multisite actif/actif

Vous pouvez exécuter votre charge de travail simultanément dans plusieurs régions dans le cadre d'une stratégie multisite active/active. Une stratégie multisite actif/actif dessert le trafic de toutes les régions dans lesquelles il est déployé. Les clients peuvent sélectionner cette stratégie pour des raisons autres que la reprise après sinistre. Elle peut être utilisée pour augmenter la disponibilité ou lors du déploiement d'une charge de travail auprès d'une audience mondiale (pour rapprocher le point de terminaison des utilisateurs et/ou déployer des piles localisées pour l'audience de cette région). En tant que stratégie de reprise après sinistre, si la charge de travail ne peut pas être prise en charge dans l'une des Régions AWS vers lesquelles elle est déployée, cette région est évacuée, et les régions restantes sont utilisées pour assurer la disponibilité. La stratégie de reprise après sinistre multisite actif/actif est la plus complexe sur le plan opérationnel et ne doit être sélectionnée que lorsque les besoins de l'entreprise l'exigent.

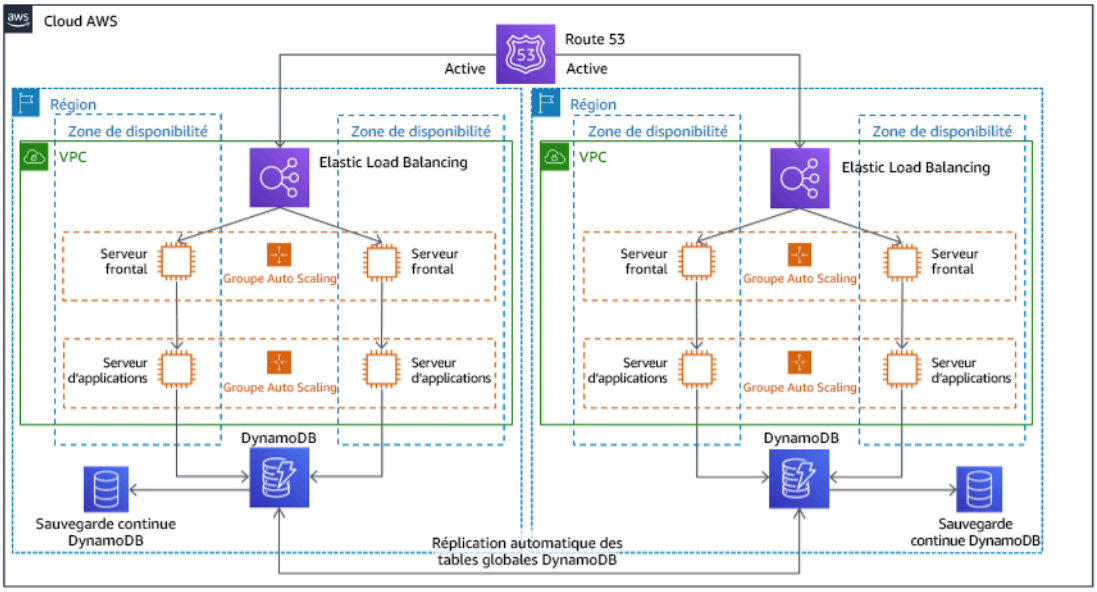


Figure 22 : architecture multisite de type actif/actif

Pour en savoir plus sur cette stratégie, consultez Architecture de reprise après sinistre sur AWS, partie 4 : multisite actif/actif.

AWS Elastic Disaster Recovery

Si vous envisagez d'adopter une stratégie de veilleuse ou de secours à chaud pour la reprise après sinistre, AWS Elastic Disaster Recovery peut proposer une autre approche offrant de meilleurs avantages. Elastic Disaster Recovery peut offrir un objectif de RPO et de RTO similaire à celui du mode de secours à chaud, tout en conservant l'approche peu coûteuse de la veilleuse. Elastic Disaster Recovery réplique vos données de votre région principale vers votre région de reprise, en utilisant une protection continue des données pour atteindre un RPO mesuré en secondes et un RTO mesurable en minutes. Seules les ressources nécessaires à la réplication des données sont déployées dans la région de reprise, ce qui permet de limiter les coûts, à l'instar de la stratégie de l'environnement de veille. En cas d'utilisation de Elastic Disaster Recovery, le service coordonne et orchestre la récupération des ressources informatiques lorsqu'elle est initiée dans le cadre d'un basculement ou d'une opération.

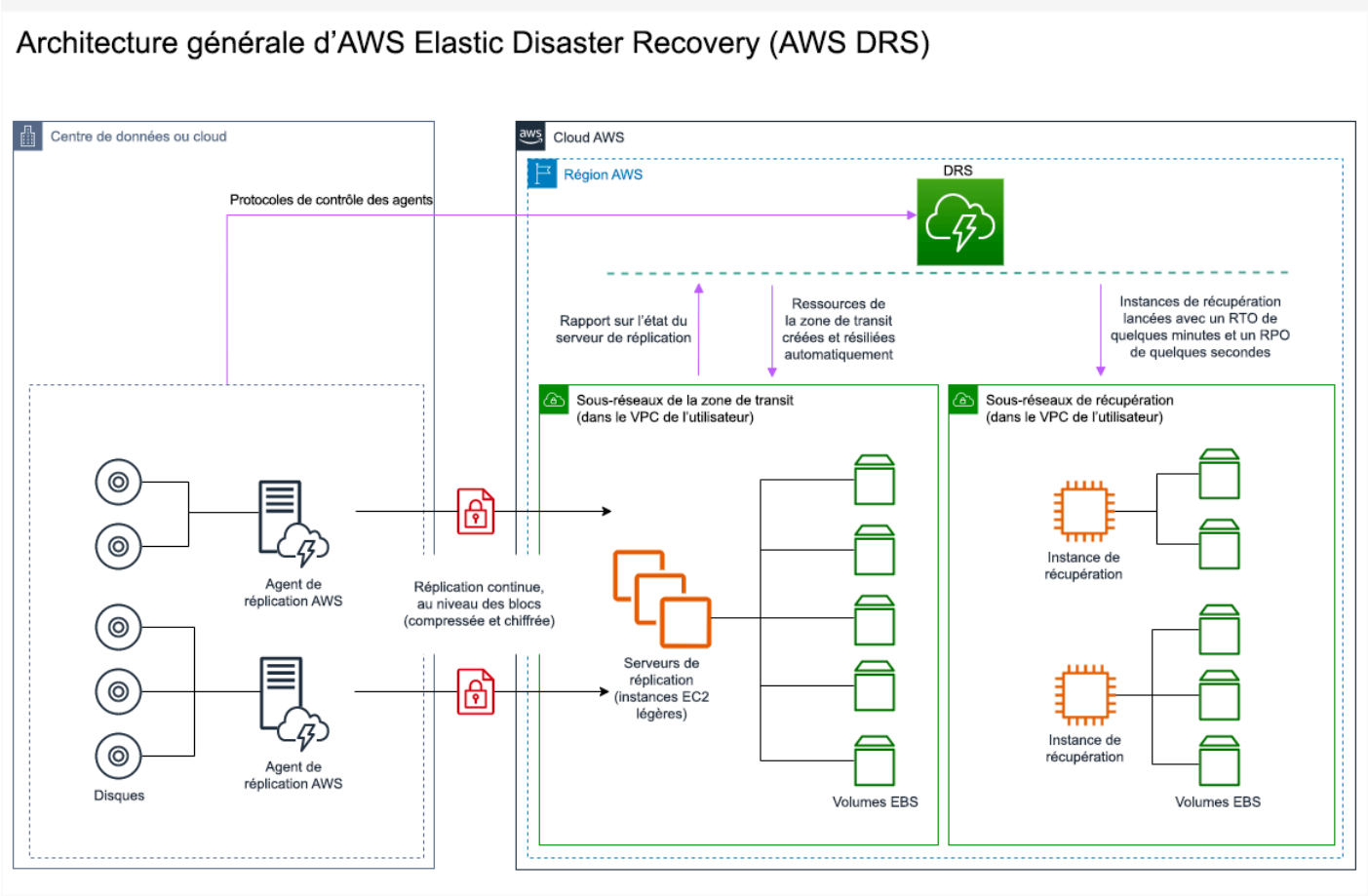


Figure 23 : architecture AWS Elastic Disaster Recovery

Pratiques supplémentaires de protection des données

Avec toutes les stratégies, vous devez également vous prémunir contre les catastrophes liées aux données La réplication continue des données vous protège contre certains types de sinistres, mais ne vous protège pas toujours contre la corruption ou la destruction des données, à moins que votre stratégie n'inclue également la gestion des versions des données stockées ou des options de récupération ponctuelle. Vous devez également sauvegarder les données répliquées sur le site de reprise pour créer des sauvegardes ponctuelles en plus des réplicas.

Utilisation de plusieurs zones de disponibilité (AZ) dans une seule Région AWS

Lorsque vous utilisez plusieurs AZ dans une même région, l'implémentation de la reprise après sinistre exploite plusieurs éléments des stratégies ci-dessus. Vous devez d'abord créer une architecture haute disponibilité (HA), en utilisant plusieurs AZ, comme illustré à la figure 23. Cette architecture utilise une approche multisite actif/actif, car les instances Amazon EC2 et Elastic

Load Balancer disposent de ressources déployées dans plusieurs zones de disponibilité, ce qui permet de traiter activement les demandes. L'architecture fait également appel au mode de veille permanente : en cas de défaillance de l'instance Amazon RDS principale (ou de l'AZ elle-même), l'instance de secours est promue en instance principale.

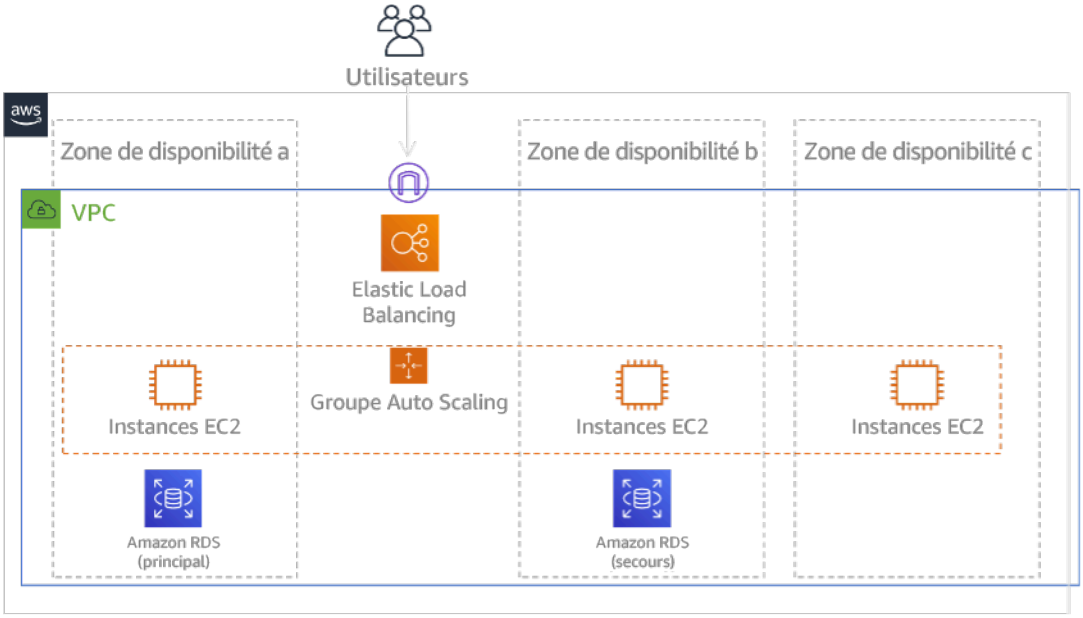


Figure 24 : architecture de multi-AZ

En plus de cette architecture haute disponibilité, vous devez ajouter des sauvegardes de toutes les données requises pour exécuter votre charge de travail. Une telle mesure est particulièrement importante pour les données limitées à une seule zone, telles que les volumes Amazon EBS ou les clusters Amazon Redshift. Si une zone de disponibilité tombe en panne, vous devrez restaurer ces données dans une autre zone de disponibilité. Dans la mesure du possible, vous devez également copier les sauvegardes de données dans une autre Région AWS comme couche de protection supplémentaire.

Une approche alternative moins courante à la reprise après sinistre multi-AZ à une seule région est illustrée dans le billet de blog intitulé Création d'applications hautement résilientes à l'aide d'Amazon Application Recovery Controller, partie 1 : pile dans une seule région. Dans ce cas, la stratégie consiste à maintenir autant que possible l'isolement entre les zones de disponibilité, à l'instar du fonctionnement des régions. Avec cette stratégie alternative, vous pouvez choisir une approche active/active ou active/passive.

> **Note**
>
> Certaines charges de travail sont soumises à des exigences réglementaires en matière de résidence des données. Si cela s'applique à votre charge de travail dans une localité qui n'a actuellement qu'une seule Région AWS, plusieurs régions ne répondront pas aux besoins de votre entreprise. Les stratégies multi-AZ assurent une bonne protection contre la plupart des catastrophes.

3. Évaluez les ressources de votre charge de travail et déterminez quelle sera leur configuration dans la région de reprise avant le basculement (pendant le fonctionnement normal).

   Pour l'infrastructure et les ressources AWS, utilisez l'infrastructure sous forme de code tel que AWS CloudFormation ou des outils tiers tels que Hashicorp Terraform. Pour un déploiement sur plusieurs comptes et régions en une seule opération, vous pouvez utiliser AWS CloudFormation StackSets. Pour les stratégies « Multisite actif/actif » et « Veille permanente », l'infrastructure déployée dans la région de reprise dispose des mêmes ressources que la région principale. Pour les stratégies « Environnement en veille » et « Secours à chaud », l'infrastructure déployée nécessitera des actions supplémentaires pour être prête pour la production. À l'aide des paramètres et de la logique conditionnelle de CloudFormation, vous pouvez contrôler si une pile déployée est active ou en veille avec un seul modèle. En utilisant Elastic Disaster Recovery, le service répliquera et orchestrera la restauration des configurations d'applications et des ressources informatiques.

   Toutes les stratégies de reprise après sinistre nécessitent que les sources de données soient sauvegardées dans la Région AWS, puis que ces sauvegardes soient copiées dans la région de restauration. AWS Backup fournit une vue centralisée dans laquelle vous pouvez configurer, planifier et surveiller les sauvegardes de ces ressources. Pour les stratégies « Environnement en veille », « Secours à chaud » et « Multisite actif/actif », vous devez également répliquer les données de la région principale vers les ressources de données de la région de reprise, telles que des instances de base de données Amazon Relational Database Service (Amazon RDS) ou les tables Amazon DynamoDB. Ces ressources de données sont donc actives et prêtes à répondre aux demandes dans la région de reprise.

   Pour en savoir plus sur le fonctionnement des services AWS dans les différentes régions, consultez cette série de blogs sur la création d'une application multi-régionale avec des services AWS.

4. Déterminez et mettez en œuvre la manière dont vous préparerez votre région de reprise pour le basculement en cas de besoin (lors d'un sinistre).

   Pour la stratégie multisite actif/actif, le basculement consiste à évacuer une région et à s'appuyer sur les régions actives restantes. En général, ces régions sont prêtes à accepter du trafic. Pour les stratégies Environnement en veille et Secours à chaud, vos actions de reprise devront déployer les ressources manquantes, telles que les instances EC2 de la figure 20, ainsi que toute autre ressource manquante.

   Pour toutes les stratégies ci-dessus, vous devrez peut-être promouvoir les instances en lecture seule des bases de données au rang d'instances principales en lecture/écriture.

   Pour la sauvegarde et la restauration, la restauration des données à partir de la sauvegarde crée des ressources pour ces données, telles que des volumes EBS, des instances de base de données RDS et des tables DynamoDB. Vous devez également restaurer l'infrastructure et déployer le code. Vous pouvez utiliser AWS Backup pour restaurer les données dans la région de reprise. Pour plus d'informations, consultez REL09-BP01 Identifiez et sauvegardez toutes les données qui doivent être sauvegardées, ou reproduisez les données à partir des sources. La reconstruction de l'infrastructure inclut la création de ressources telles que des instances EC2 en plus du Virtual Private Cloud (VPC) Amazon, des sous-réseaux et des groupes de sécurité nécessaires. Vous pouvez automatiser une grande partie du processus de restauration. Pour savoir comment procéder, consultez ce billet de blog.

5. Déterminez et mettez en œuvre la manière dont vous redirigerez le trafic vers le basculement en cas de besoin (lors d'un sinistre).

   Cette opération de basculement peut être lancée automatiquement ou manuellement. Le basculement lancé automatiquement sur la base de la surveillance de l'état ou d'alarmes doit être utilisé avec prudence, car un basculement inutile (fausse alerte) entraînerait des coûts tels que l'indisponibilité et la perte de données. Le basculement manuel est donc souvent utilisé. Dans ce cas, nous vous conseillons tout de même d'automatiser les étapes de basculement, de sorte que vous n'ayez à appuyer que sur un bouton pour lancer le basculement.

   Il existe plusieurs options de gestion du trafic à prendre en compte lors de l'utilisation des services AWS. L'une des options consiste à utiliser Amazon Route 53. Amazon Route 53 vous permet d'associer plusieurs points de terminaison IP dans une ou plusieurs Régions AWS avec un nom de domaine Route 53. Pour mettre en œuvre le basculement initié manuellement, vous pouvez utiliser Amazon Application Recovery Controller, qui fournit une API de plan de données hautement disponible pour rediriger le trafic vers la région de récupération. Lors de la mise en œuvre du

basculement, utilisez les opérations du plan de données et évitez celles du plan de contrôle, comme décrit dans REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle lors de la restauration.

Pour en savoir plus à ce sujet et sur d'autres options, consultez cette section du livre blanc sur la reprise après sinistre.

6. Élaborez un plan pour déterminer la façon dont votre charge de travail se rétablira.

   Failback consiste à renvoyer l'exploitation de la charge de travail à la région principale, après qu'un événement de sinistre s'est atténué. La mise en service de l'infrastructure et du code dans la région principale suit généralement les mêmes étapes que celles utilisées initialement. Elle s'appuie notamment sur l'infrastructure en tant que code et les pipelines de déploiement de code. Le défi posé par failback consiste à restaurer les magasins de données et à garantir leur cohérence avec la région de reprise en cours d'exécution.

   Lors de l'état de basculement, les bases de données de la région de reprise sont actives et disposent des données à jour. L'objectif est alors de resynchroniser les données de la région de reprise vers la région principale, en s'assurant qu'elle est à jour.

   Certains services AWS effectuent cette opération automatiquement. Si vous utilisez des tables globales Amazon DynamoDB, même si la table de la région principale devenait indisponible, DynamoDB reprendrait la propagation de toutes les écritures en attente lorsqu'elle se reconnecterait. Si vous utilisez Amazon Aurora Global Database et que vous utilisez un basculement planifié géré, la topologie de réplication existante de la base de données globale Aurora est conservée. Par conséquent, l'ancienne instance en lecture/écriture de la région principale deviendra un réplica et recevra les mises à jour de la région de reprise.

   Dans les cas où cela n'est pas automatique, vous devrez rétablir la base de données dans la région principale en tant que réplica de la base de données dans la région de reprise. Dans de nombreux cas, cela implique la suppression de l'ancienne base de données principale et la création de nouveaux réplicas.

   Après un basculement, si vous pouvez poursuivre l'exécution dans la région de reprise, envisagez d'en faire la nouvelle région principale. Vous devriez alors suivre toutes les étapes ci-dessus pour convertir l'ancienne région principale en région de reprise. Certaines organisations effectuent une rotation planifiée, en échangeant périodiquement leurs régions principale et de reprise (par exemple tous les trois mois).

Toutes les étapes nécessaires au basculement et au rétablissement doivent être conservées dans un playbook accessible à tous les membres de l'équipe et révisé périodiquement.

En utilisant Elastic Disaster Recovery, le service permettra d'orchestrer et d'automatiser le processus de failback. Pour en savoir plus, consultez la section Réalisation d'un failback.

Niveau d'effort du plan d'implémentation : élevé

Ressources

Bonnes pratiques associées :

- the section called "REL09-BP01 Identifiez et sauvegardez toutes les données qui doivent être sauvegardées, ou reproduisez les données à partir des sources"
- the section called "REL11-BP04 S'appuyer sur le plan de données et non sur le plan de contrôle lors de la restauration"
- the section called "REL13-BP01 Définir les objectifs de reprise en termes de durée d'indisponibilité et de perte de données"

Documents connexes :

- AWS Architecture Blog: Disaster Recovery Series
- Disaster Recovery of Workloads on AWS: Recovery in the Cloud (AWS Whitepaper)
- Options de reprise après sinistre dans le cloud
- Créer une solution dorsale active-active sans serveur sur plusieurs régions en une heure
- Solution dorsale sans serveur sur plusieurs régions – rechargé
- RDS : réplication d'un réplica en lecture entre les régions
- Route 53 : configuration du basculement DNS
- S3 : réplication entre régions
- Présentation de AWS Backup
- Qu'est-ce qu'Amazon Application Recovery Controller ?
- Reprise après sinistre Elastic AWS
- HashiCorp Terraform : Démarrage – AWS
- Partenaire APN : partenaires pouvant faciliter la reprise après sinistre

- AWS Marketplace : produits pouvant être utilisés pour la reprise après sinistre

Related videos:

- Disaster Recovery of Workloads on AWS
- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
- Get Started with AWS Elastic Disaster Recovery | Amazon Web Services

Exemples connexes :

- Atelier Well-Architected – Reprise après sinistre – Série d'ateliers illustrant les stratégies de reprise après sinistre

REL13-BP03 Tester la mise en œuvre de la reprise après sinistre pour valider la mise en œuvre

Testez régulièrement le basculement sur votre site de restauration pour vérifier qu'il fonctionne correctement RTO et RPO qu'il est respecté.

Anti-modèles courants :

- Ne jamais exécuter de basculements en production.

Avantages du respect de cette bonne pratique : en testant régulièrement votre plan de reprise après sinistre, vous vous assurez qu'il fonctionnera quand il le faudra et que votre équipe sait comment exécuter la stratégie.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

S'il y a bien un modèle à éviter, c'est celui qui consiste à développer des chemins de récupération rarement testés. Par exemple, vous pouvez avoir un magasin de données secondaire qui est utilisé pour les requêtes en lecture seule. Lorsque vous écrivez dans un magasin de données et que l'instance principale connaît une défaillance, vous pouvez basculer vers le magasin de données secondaire. Si vous ne testez pas fréquemment ce basculement, vous constaterez peut-être que vos hypothèses sur les capacités du magasin de données secondaire sont incorrectes. La capacité du magasin de données secondaire, qui peut avoir été suffisante lors de votre dernier test, peut ne plus

être en mesure de tolérer la charge dans le cadre de ce scénario. Notre expérience a montré que le seul chemin de récupération après erreur qui fonctionne est celui que vous testez fréquemment. C'est pourquoi l'idéal est de n'avoir qu'un petit nombre de chemins de récupération. Vous pouvez établir des modèles de reprise et tester ceux-ci régulièrement. Si vous avez un chemin de récupération complexe ou critique, vous devez toujours exécuter régulièrement cette panne en production pour vous assurer du bon fonctionnement de ce chemin de récupération. Dans l'exemple que nous venons de présenter, vous devez procéder régulièrement au basculement vers l'instance de secours, quel que soit le besoin.

Étapes d'implémentation

1. Préparez vos charges de travail pour la reprise. Testez régulièrement vos chemins de récupération. L'informatique orientée récupération identifie les caractéristiques des systèmes qui améliorent la récupération : isolement et redondance, capacité de l'ensemble du système à réduire les modifications, capacité à surveiller et déterminer l'état de santé, capacité à fournir des diagnostics, reprise automatique, conception modulaire et capacité à redémarrer. Entraînez votre chemin de reprise pour vérifier qu'il peut s'effectuer au moment et à l'état spécifiés. Utilisez vos runbooks au cours de cette reprise pour documenter les problèmes et trouver des solutions pour les résoudre avant le prochain test.
2. Pour les charges de travail EC2 basées sur Amazon, utilisez-le AWS Elastic Disaster Recoverypour implémenter et lancer des instances de forage dans le cadre de votre stratégie de reprise après sinistre. AWS Elastic Disaster Recovery permet d'exécuter des exercices de manière efficace, ce qui vous aide à vous préparer en cas de basculement. Vous pouvez également lancer fréquemment vos instances en utilisant Elastic Disaster Recovery à des fins de test et d'opération sans rediriger le trafic.

Ressources

Documents connexes :

- APNPartenaire : partenaires qui peuvent aider à la reprise après sinistre
- Blog d'architecture AWS : série sur la reprise après sinistre
- AWS Marketplace : produits pouvant être utilisés pour la reprise après sinistre
- AWS Elastic Disaster Recovery
- Reprise après sinistre des charges de travail sur AWS : restauration dans le cloud (AWS livre blanc)
- AWS Elastic Disaster Recovery Préparation au basculement

- Projet informatique orientée reprise Berkeley/Stanford
- Qu'est-ce que le simulateur d'injection de AWS défauts ?

Vidéos connexes :

- AWS re:Invent 2018 : Modèles d'architecture pour les applications active-active multirégionales
- AWS re:Invent 2019 : Backup-and-restore et des solutions de reprise après sinistre avec AWS

Exemples connexes :

- Ateliers Well-Architected : tester la résilience

REL13-BP04 Gérer la dérive de configuration au niveau du site ou de la région de reprise après sinistre

Pour mener à bien une procédure de reprise après sinistre (DR), votre charge de travail doit être en mesure de reprendre son fonctionnement normal en temps opportun, sans aucune perte de fonctionnalité ni de données une fois que l'environnement de reprise après sinistre a été mis en ligne. Pour atteindre cet objectif, il est essentiel de maintenir une infrastructure, des données et des configurations cohérentes entre votre environnement de reprise après sinistre et l'environnement principal.

Résultat escompté : la configuration et les données de votre site de reprise après sinistre sont identiques à celles du site principal, ce qui permet une récupération rapide et complète au moment requis.

Anti-modèles courants :

- Vous ne mettez pas à jour les emplacements de récupération lorsque des modifications sont apportées aux emplacements principaux, ce qui entraîne une obsolescence des configurations, susceptible d'entraver les efforts de récupération.
- Vous ne tenez pas compte des limitations potentielles telles que les différences de service entre les emplacements principaux et de récupération, ce qui peut entraîner des échecs inattendus lors du basculement.
- Vous vous appuyez sur des processus manuels pour mettre à jour et synchroniser l'environnement de reprise après sinistre, ce qui augmente le risque d'erreur humaine et d'incohérence.

- Vous ne détectez pas une dérive de configuration et avez une fausse impression de l'état de préparation du site de reprise après sinistre avant un incident.

Avantages liés au respect de cette bonne pratique : la cohérence entre l'environnement de reprise après sinistre et l'environnement principal améliore considérablement les chances de réussite de la récupération après un incident et réduit le risque d'échec de la procédure de récupération.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Une approche globale de la gestion de la configuration et de la préparation au basculement peut vous aider à vérifier que le site de reprise après sinistre est régulièrement mis à jour et prêt à prendre le relais en cas de défaillance du site principal.

Pour garantir la cohérence entre votre environnement principal et votre environnement de reprise après sinistre (DR), assurez-vous que vos pipelines de distribution répartissent les applications à la fois sur votre site principal et sur votre site de reprise après sinistre. Déployez les modifications sur les sites de reprise après une période d'évaluation appropriée (on parle alors de déploiements échelonnés) afin de détecter les problèmes sur le site principal et d'arrêter le déploiement avant qu'ils ne se propagent. Mettez en œuvre une surveillance pour détecter les dérives de configuration et suivre les modifications et la conformité dans l'ensemble de vos environnements. Procédez à des corrections automatisées sur le site de reprise après sinistre pour qu'il reste totalement cohérent et prêt à prendre le relais en cas d'incident.

Étapes d'implémentation

1. Vérifiez que la région de reprise après sinistre contient les services et fonctionnalités AWS nécessaires à la bonne exécution de votre plan de reprise après sinistre.
2. Utilisez une infrastructure en tant que code (IaC). Maintenez l'exactitude de vos modèles de configuration de l'infrastructure de production et des applications, et appliquez-les régulièrement à votre environnement de reprise après sinistre. AWS CloudFormation peut détecter des dérives entre ce que vos modèles CloudFormation spécifient et ce qui est réellement déployé.
3. Configurez des pipelines CI/CD pour déployer des applications et des mises à jour d'infrastructure dans tous les environnements, y compris les sites principaux et de reprise après sinistre. Les solutions CI/CD telles qu'AWS CodePipeline peuvent automatiser le processus de déploiement, ce qui réduit le risque de dérive de la configuration.

4. Échelonnez les déploiements entre l'environnement principal et l'environnement de reprise après sinistre. Cette approche permet de déployer et de tester initialement les mises à jour dans l'environnement principal, ce qui permet d'isoler les problèmes sur le site principal avant qu'ils ne soient propagés au site de reprise après sinistre. Cette approche empêche la transmission simultanée des défauts en production et au site de reprise après sinistre et préserve l'intégrité de l'environnement de reprise après sinistre.
5. Surveillez en permanence les configurations des ressources dans l'environnement principal et l'environnement de reprise après sinistre. Des solutions telles qu'AWS Config peuvent aider à renforcer la conformité des configurations et à détecter les dérives, ce qui permet de maintenir la cohérence des configurations dans tous les environnements.
6. Mettez en œuvre des mécanismes d'alerte pour suivre et signaler toute dérive de configuration, ainsi que toute interruption ou tout retard de réplication des données.
7. Automatisez la correction des dérives de configuration détectées.
8. Planifiez des audits et des contrôles de conformité réguliers pour vérifier l'alignement continu entre les configurations principale et de reprise après sinistre. Les examens périodiques vous aident à maintenir la conformité aux règles définies et à identifier les éventuelles anomalies à corriger.
9. Recherchez des disparités au niveau de la capacité provisionnée, des quotas de service, des limitations et des différences de configuration et de version AWS.

Ressources

Bonnes pratiques associées :

- REL01-BP01 Connaître les quotas de service et les contraintes
- REL01-BP02 Gérer les quotas de service entre les comptes et les régions
- REL01-BP04 Surveiller et gérer les quotas
- REL13-BP03 Effectuer un test de validation de la mise en œuvre de la reprise après sinistre

Documents connexes :

- Correction des ressources AWS non conformes par AWS Config Rules
- AWS Systems Manager Automation
- AWS CloudFormation : Détection de modifications non gérées de la configuration des piles et des ressources

- AWS CloudFormation : détection de tout écart à l'échelle d'une pile CloudFormation
- AWS Systems Manager Automation
- Reprise après sinistre des charges de travail sur AWS : reprise dans le cloud (livre blanc AWS)
- Comment mettre en œuvre une solution de gestion de configuration d'infrastructure sur AWS ?
- Correction des ressources AWS non conformes par AWS Config Rules

Vidéos connexes :

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)

Exemples connexes :

- Registre AWS CloudFormation
- Quota Monitor for AWS
- Mise en œuvre de la correction automatique des dérives pour AWS CloudFormation à l'aide d'Amazon CloudWatch et d'AWS Lambda
- Blog d'architecture AWS : série sur la reprise après sinistre
- AWS Marketplace : produits pouvant être utilisés pour la reprise après sinistre
- Automating safe, hands-off deployments

REL13-BP05 Automatiser la reprise

Mettez en œuvre des mécanismes de reprise testés et automatisés, à la fois fiables, observables et reproductibles afin de réduire le risque et l'impact sur l'activité d'une panne.

Résultat escompté : vous avez mis en œuvre un flux de travail d'automatisation bien documenté, standardisé et entièrement testé pour les processus de récupération. L'automatisation de la récupération corrige automatiquement les problèmes mineurs qui présentent un faible risque d'indisponibilité ou de perte de données. Vous êtes en mesure d'invoquer rapidement des processus de récupération pour des incidents graves, d'observer le comportement de correction pendant leur fonctionnement et de mettre fin aux processus si vous observez des situations dangereuses ou des défaillances.

Anti-modèles courants :

- Dans le cadre de votre plan de reprise, vous dépendez de composants ou de mécanismes défaillants ou dégradés.
- Vos processus de récupération nécessitent une intervention manuelle, telle que l'accès à la console (également appelé ClickOps).
- Vous lancez automatiquement les procédures de récupération dans les situations présentant un risque élevé d'indisponibilité ou de perte de données.
- Vous omettez d'inclure un mécanisme permettant d'annuler une procédure de récupération (comme un système Andon ou un bouton d'arrêt d'urgence) qui ne fonctionne pas ou qui présente des risques supplémentaires.

Avantages liés au respect de cette bonne pratique :

- Fiabilité, prévisibilité et cohérence accrues des opérations de récupération.
- Capacité à atteindre des objectifs de reprise plus stricts, notamment l'objectif de délai de reprise (RTO) et l'objectif de point de reprise (RPO).
- Diminution du risque d'échec de la récupération lors d'un incident.
- Réduction du risque d'échec associé aux processus de récupération manuels susceptibles de provoquer des erreurs humaines.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Pour mettre en œuvre la restauration automatique, vous avez besoin d'une approche globale qui utilise les services et les bonnes pratiques AWS. Pour commencer, identifiez les composants critiques et les points de défaillance potentiels de votre charge de travail. Développez des processus automatisés capables de récupérer vos charges de travail et vos données en cas de panne sans intervention humaine.

Développez l'automatisation de la récupération en utilisant les principes de l'infrastructure en tant que code (IaC). Cela rend votre environnement de récupération cohérent avec l'environnement source et permet de contrôler la version de vos processus de récupération. Pour orchestrer des flux de travail de récupération complexes, envisagez des solutions telles que AWS Systems Manager Automations ou AWS Step Functions.

L'automatisation des processus de récupération présente des avantages considérables et peut vous aider à atteindre plus facilement votre objectif de délai de reprise (RTO) et votre objectif de point

de reprise (RPO). Toutefois, vous pouvez rencontrer des situations inattendues susceptibles de provoquer un échec ou de créer de nouveaux risques, tels qu'une durée d'indisponibilité et une perte de données supplémentaires. Pour atténuer ce risque, offrez la possibilité d'arrêter rapidement une automatisation de récupération en cours. Une fois celle-ci arrêtée, vous pouvez enquêter et prendre des mesures correctives.

Pour les charges de travail prises en charge, envisagez des solutions telles qu'AWS Elastic Disaster Recovery (AWS DRS) pour fournir un basculement automatisé. AWS DRS réplique en continu vos machines (notamment le système d'exploitation, la configuration d'état du système, les bases de données, les applications et les fichiers) dans une zone intermédiaire de votre Compte AWS cible et de votre région préférée. En cas d'incident, AWS DRS automatise la conversion de vos serveurs répliqués en charges de travail entièrement provisionnées dans votre région de récupération sur AWS.

La maintenance et l'amélioration de la récupération automatisée sont un processus continu. Testez et affinez continuellement vos procédures de récupération sur la base des enseignements acquis, et tenez-vous au fait des nouveaux services et fonctionnalités AWS susceptibles d'améliorer vos capacités de récupération.

Étapes d'implémentation

1. Planifier une récupération automatisée
    a. Réalisez un examen approfondi de l'architecture, des composants et des dépendances de votre charge de travail afin d'identifier et de planifier des mécanismes de récupération automatisés. Classez les dépendances de votre charge de travail en dépendances strictes et souples. Les dépendances strictes sont celles sans lesquelles la charge de travail ne peut pas fonctionner et que rien ne peut substituer. Les dépendances souples sont celles que la charge de travail utilise habituellement, mais qui peuvent être remplacées par des systèmes ou des processus de substitution temporaires ou qui peuvent être traitées par une dégradation appropriée.
    b. Établissez des processus pour identifier et récupérer les données manquantes ou corrompues.
    c. Définissez les étapes permettant de confirmer le rétablissement d'un état stable après l'exécution des actions de récupération.
    d. Envisagez toutes les actions nécessaires pour préparer le système récupéré à être pleinement opérationnel, telles que la préparation et le remplissage des caches.
    e. Tenez compte des problèmes susceptibles d'être rencontrés au cours du processus de récupération et de la manière de les détecter et de les corriger.

   f. Envisagez des scénarios dans lesquels le site principal et son plan de contrôle sont inaccessibles. Vérifiez que les actions de récupération peuvent être effectuées indépendamment, sans avoir recours au site principal. Envisagez des solutions telles qu'Amazon Application Recovery Controller (ARC) pour rediriger le trafic sans qu'il soit nécessaire de muter manuellement les enregistrements DNS.

2. Développer un processus de récupération automatisé
   a. Mettez en œuvre des mécanismes automatisés de détection des pannes et de basculement pour une récupération sans intervention manuelle. Créez des tableaux de bord avec des outils tels qu'Amazon CloudWatch pour rendre compte de la progression et de l'état des procédures de récupération automatisées. Incluez des procédures pour valider la réussite de la récupération. Fournissez un mécanisme permettant d'annuler une récupération en cours.
   b. Créez des playbooks comme processus de secours pour les pannes qui ne permettent pas une récupération automatique, et tenez compte de votre plan de reprise après sinistre.
   c. Testez les processus de récupération comme indiqué dans le document REL13-BP03.
3. Préparer la récupération
   a. Évaluez l'état de votre site de reprise et déployez-y les composants stratégiques à l'avance. Pour plus de détails, consultez REL13-BP04.
   b. Définissez des rôles, des responsabilités et des processus décisionnels clairs pour les opérations de récupération, en impliquant les parties prenantes et les équipes sur l'ensemble de l'organisation.
   c. Définissez les conditions pour lancer vos processus de récupération.
   d. Créez un plan pour annuler le processus de récupération et revenir à votre site principal si nécessaire ou une fois que celui-ci est considéré comme sûr.

Ressources

Bonnes pratiques associées :

- REL07-BP01 Utiliser l'automatisation lors de l'obtention des ressources ou de leur mise à l'échelle
- REL11-BP01 Surveiller tous les composants de la charge de travail pour détecter les défaillances
- REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise
- REL13-BP03 Effectuer un test de validation de la mise en œuvre de la reprise après sinistre
- REL13-BP04 Gérer la dérive de configuration au niveau du site ou de la région de reprise après sinistre

Documents connexes :

- AWS Architecture Blog: Disaster Recovery Series
- Disaster Recovery of Workloads on AWS: Recovery in the Cloud (AWS Whitepaper)
- Orchestration de l'automatisation de la reprise après sinistre à l'aide d'Amazon Route 53 ARC et d'AWS Step Functions
- Création de dossiers d¦exploitation AWS Systems Manager Automation à l'aide d'AWS CDK
- AWS Marketplace : produits pouvant être utilisés pour la reprise après sinistre
- AWS Systems Manager Automation
- Reprise après sinistre Elastic AWS
- Utilisation d'Elastic Disaster Recovery pour le basculement et le failback
- Ressources de reprise après sinistre Elastic AWS
- Partenaire APN : partenaires pouvant faciliter la reprise après sinistre

Vidéos connexes :

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)
- AWS re:Invent 2022: AWS On Air ft. AWS Failback for AWS Elastic Disaster Recovery

Efficacité des performances

Le pilier Efficacité des performances englobe la capacité à utiliser efficacement les ressources du cloud pour satisfaire aux exigences système et à maintenir cette efficacité au fur et à mesure que la demande change et que les technologies évoluent. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Efficacité des performances.

Domaines de bonnes pratiques

- Sélection d'architecture
- Informatique et matériel
- Gestion des données
- Réseau et diffusion de contenu
- Processus et culture

Sélection d’architecture

Questions

- PERF 1. Comment sélectionner les ressources et l’architecture cloud adaptées à votre charge de travail ?

PERF 1. Comment sélectionner les ressources et l’architecture cloud adaptées à votre charge de travail ?

La solution optimale pour une charge de travail peut varier, et les solutions combinent souvent plusieurs approches. Les charges de travail Well-Architected utilisent plusieurs solutions et permettent d’exploiter différentes fonctionnalités pour améliorer les performances.

Bonnes pratiques

- PERF01-BP01 Découvrez et comprenez les services et fonctionnalités cloud disponibles
- PERF01-BP02 Suivez les conseils de votre fournisseur de cloud ou d'un partenaire approprié pour en savoir plus sur les modèles d'architecture et les meilleures pratiques
- PERF01-BP03 Intégrer les coûts dans les décisions architecturales
- PERF01-BP04 Évaluer l'impact des compromis sur les clients et l'efficacité de l'architecture
- PERF01-BP05 Politiques d'utilisation et architectures de référence
- PERF01-BP06 Utilisation du benchmarking pour éclairer vos décisions architecturales
- PERF01-BP07 Utiliser une approche axée sur les données pour les choix architecturaux

PERF01-BP01 Découvrez et comprenez les services et fonctionnalités cloud disponibles

Découvrez en continu les services et configurations disponibles qui vous aident à prendre de meilleures décisions architecturales et à améliorer l’efficacité des performances de votre architecture de charge de travail.

Anti-modèles courants :

- Vous utilisez le cloud comme centre de données hébergé.
- Vous ne modernisez pas votre application après la migration vers le cloud.
- Vous n’utilisez qu’un seul type de stockage pour tout ce que vous devez conserver.

- Vous utilisez les types d’instances qui correspondent le plus à vos standards actuels. Elles peuvent être de plus grande taille au besoin.
- Vous déployez et gérez les technologies disponibles en tant que services gérés.

Avantages liés au respect de cette bonne pratique : en envisageant de nouveaux services et de nouvelles configurations, vous pourriez être en mesure d’améliorer considérablement vos performances, de réduire les coûts et d’optimiser les efforts requis pour maintenir votre charge de travail. Cela peut également vous aider à accélérer le développement time-to-value des produits compatibles avec le cloud.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

AWS publie en permanence de nouveaux services et fonctionnalités susceptibles d'améliorer les performances et de réduire le coût des charges de travail dans le cloud. Il est essentiel up-to-date de rester fidèle à ces nouveaux services et fonctionnalités pour maintenir l'efficacité des performances dans le cloud. La modernisation de votre architecture de charge de travail vous permet également d’accélérer la productivité, de stimuler l’innovation et de générer de nouvelles opportunités de croissance.

Étapes d’implémentation

- Faites l’inventaire de vos charges de travail logicielles et de l’architecture des services connexes. Déterminez la catégorie de produits sur laquelle vous souhaitez en savoir plus.
- Explorez les AWS offres pour identifier et découvrir les services et options de configuration pertinents qui peuvent vous aider à améliorer les performances et à réduire les coûts et la complexité opérationnelle.
  - Amazon Web Services Cloud
  - AWS Académie
  - Quoi de neuf avec AWS ?
  - AWS Blog
  - AWS Générateur de compétences
  - AWS Événements et webinaires
  - AWS Training et certifications
  - AWS Chaîne Youtube

  - AWS Ateliers
  - Communautés AWS
- Utilisez Amazon Q pour obtenir des informations pertinentes et des conseils sur les services.
- Utilisez des environnements de test (sandbox) (hors production) pour découvrir et tester de nouveaux services sans frais supplémentaires.
- Découvrez en permanence les nouveaux services et fonctionnalités du cloud.

Ressources

Documents connexes :

- Présentation d'Amazon Web Services
- EC2Fonctionnalités d'Amazon
- Apprenez step-by-step avec le plan de formation d'un AWS partenaire
- AWS Formation et certification
- Mon parcours d'apprentissage pour devenir architecte de AWS solutions
- AWS Centre d'architecture
- AWS Partner Network
- AWS Bibliothèque de solutions
- AWS Centre de connaissances
- Créez des applications modernes sur AWS

Vidéos connexes :

- AWS re:Invent 2023 - Nouveautés d'Amazon EC2
- AWS re:Invent 2022 - Réduisez vos coûts d'exploitation et d'infrastructure avec Amazon ECS
- AWS re:Invent 2023 - Développez avec l'efficacité, l'agilité et l'innovation du cloud avec AWS
- AWS re:Invent 2022 - Déployez des modèles de machine learning pour l'inférence à des performances élevées et à moindre coût
- This is my Architecture

Exemples connexes :

- AWS Exemples
- AWS SDKExemples

PERF01-BP02 Suivez les conseils de votre fournisseur de cloud ou d'un partenaire approprié pour en savoir plus sur les modèles d'architecture et les meilleures pratiques

Utilisez les ressources cloud de l’entreprise, telles que la documentation, les architectes de solutions, les services professionnels ou les partenaires appropriés pour éclairer vos décisions architecturales. Ces ressources vous aident à vérifier et à améliorer votre architecture pour obtenir des performances optimales.

Anti-modèles courants :

- Vous l'utilisez AWS en tant que fournisseur de cloud commun.
- Vous utilisez AWS les services d'une manière pour laquelle ils n'ont pas été conçus.
- Vous suivez toutes les recommandations sans tenir compte du contexte de votre entreprise.

Avantage de l’établissement de cette bonne pratique : en suivant les recommandations d’un fournisseur de cloud ou d’un partenaire approprié, vous pouvez faire les bons choix architecturaux pour votre charge de travail et vous avez confiance dans vos décisions.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

AWS propose un large éventail de conseils, de documentation et de ressources qui peuvent vous aider à créer et à gérer des charges de travail cloud efficaces. AWS la documentation fournit des exemples de code, des didacticiels et des explications détaillées sur les services. Outre la documentation, AWS propose des programmes de formation et de certification, des architectes de solutions et des services professionnels qui peuvent aider les clients à explorer différents aspects des services cloud et à mettre en œuvre une architecture cloud efficace sur AWS.

Tirez parti de ces ressources pour obtenir des informations précieuses et des bonnes pratiques, gagner du temps et obtenir de meilleurs résultats dans le AWS Cloud.

Étapes d’implémentation

- Consultez AWS la documentation et les directives et suivez les meilleures pratiques. Ces ressources peuvent vous aider à choisir et à configurer efficacement les services, ainsi qu’à améliorer les performances.
  - AWS documentation (comme les guides d'utilisation et les livres blancs)
  - AWS Blog
  - AWS Training et certifications
  - AWS Chaîne Youtube
- Participez à des événements organisés par des AWS partenaires (tels que AWS Global Summits, AWS re:Invent, groupes d'utilisateurs et ateliers) pour découvrir les meilleures pratiques d'utilisation des services auprès d' AWS experts. AWS
  - Apprenez step-by-step avec un plan de formation pour AWS partenaires
  - AWS Événements et webinaires
  - AWS Ateliers
  - AWS Communautés
- Demandez de l' AWS aide si vous avez besoin de conseils supplémentaires ou d'informations sur les produits. AWS Les architectes de solutions et les services AWS professionnels fournissent des conseils pour la mise en œuvre des solutions. AWS Les partenaires fournissent AWS leur expertise pour vous aider à optimiser l'agilité et l'innovation au sein de votre entreprise.
- Utilisez Support si vous avez besoin d’une assistance technique pour utiliser un service de manière efficace. Nos plans de Support sont conçus pour vous fournir la bonne combinaison d'outils et l'accès à l'expertise afin que vous puissiez réussir AWS tout en optimisant les performances, en gérant les risques et en maîtrisant les coûts.

Ressources

Documents connexes :

- Centre d’architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- AWS Entreprise Support

Vidéos connexes :

- This is my Architecture
- AWS re:Invent 2023 - Modèles avancés basés sur les événements avec Amazon EventBridge
- AWS re:Invent 2023 - Implémentation de modèles de conception distribués sur AWS
- AWS re:Invent 2023 - Architecture d'application sous forme de code

Exemples connexes :

- Exemples AWS
- AWS SDKExemples
- AWS Architecture de référence analytique

PERF01-BP03 Intégrer les coûts dans les décisions architecturales

Tenez compte des coûts dans vos décisions architecturales afin d'améliorer l'utilisation des ressources et l'efficacité des performances de votre charge de travail cloud. Lorsque vous êtes conscient des implications financières de votre charge de travail cloud, vous êtes plus susceptible de tirer parti de ressources efficaces et de réduire les pratiques inutiles.

Anti-modèles courants :

- Vous n'utilisez qu'une seule famille d'instances.
- Vous n'évaluez pas les solutions sous licence par rapport aux solutions open source.
- Vous ne définissez pas de stratégies de cycle de vie pour le stockage.
- Vous ne passez pas en revue les nouveaux services et fonctionnalités du AWS Cloud.
- Vous utilisez uniquement le stockage par blocs.

Avantages liés au respect de cette bonne pratique : en tenant compte des coûts dans vos prises de décision, vous pouvez utiliser des ressources plus efficaces et explorer d'autres investissements.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

L'optimisation des charges de travail en matière de coûts peut améliorer l'utilisation des ressources et éviter le gaspillage dans une charge de travail cloud. La prise en compte des coûts dans les

décisions architecturales implique généralement de dimensionner correctement les composants de la charge de travail et de renforcer l'élasticité, ce qui se traduit par une amélioration de l'efficacité des performances de la charge de travail cloud.

Étapes d'implémentation

- Fixez des objectifs de coûts tels que des limites budgétaires pour votre charge de travail cloud.
- Identifiez les composants clés (tels que les instances et le stockage) qui augmentent le coût de votre charge de travail. AWS Pricing Calculator et AWS Cost Explorer vous permettent d'identifier les principaux facteurs de coûts dans votre charge de travail.
- Comprendre les modèles de tarification dans le cloud, tels que la demande, les instances réservées, les Savings Plans et les instances ponctuelles.
- Utilisez les bonnes pratiques d'optimisation des coûts de Well-Architected pour optimiser ces composants clés en matière de coûts.
- Surveillez et analysez en permanence les coûts afin d'identifier les opportunités d'optimisation des coûts dans votre charge de travail.
  - Utilisez les budgets AWS pour recevoir des alertes en cas de coûts inacceptables.
  - Utilisez AWS Compute Optimizer ou AWS Trusted Advisor pour obtenir des recommandations en matière d'optimisation des coûts.
  - Utilisez la détection des anomalies de coûts AWS pour obtenir une détection automatisée des anomalies de coûts et une analyse des causes profondes.

Ressources

Documents connexes :

- Qu'est-ce que AWS Billing and Cost Management ?
- Optimisation des coûts avec AWS
- Choix d'une stratégie de gestion des AWS coûts
- Guide de gestion des AWS coûts pour débutants
- Présentation détaillée du tableau de bord Cost Intelligence Dashboard
- Centre d'architecture AWS
- Bibliothèque de solutions AWS
- Centre de connaissances AWS

Vidéos connexes :

- This is my Architecture
- AWS re:Invent 2023 - Nouveautés en matière d'optimisation des coûts AWS
- AWS re:Invent 2023 - Optimisez les coûts et les performances et suivez les progrès en matière d'atténuation
- AWS re:Invent 2023 - meilleures pratiques en matière d'optimisation des coûts AWS de stockage
- AWS re:Invent 2023 - Optimisez les coûts dans vos environnements multi-comptes

Exemples connexes :

- AWS Compute Optimizer Code de démonstration
- Atelier d’optimisation des coûts
- Playbooks de mise en œuvre technique de la gestion financière dans le cloud
- Optimisation du démarrage : ajustement des performances des applications pour une efficacité maximale
- Atelier d’optimisation sans serveur (performances et coûts)
- Mise à l’échelle d’architectures rentables

PERF01-BP04 Évaluer l'impact des compromis sur les clients et l'efficacité de l'architecture

Lors de l’évaluation des améliorations liées à la performance, identifiez les choix qui affectent vos clients et l’efficacité de la charge de travail. Par exemple, si l’utilisation d’un magasin de données clé-valeur augmente les performances du système, il est important d’évaluer l’impact de la nature constante de cette modification à terme sur les clients.

Anti-modèles courants :

- Vous supposez que tous les gains de performances doivent être mis en œuvre, même s’il existe des compromis en termes d’implémentation.
- Vous n’évaluez les modifications apportées aux charges de travail que lorsqu’un problème de performances a atteint un point critique.

Avantages liés au respect de cette bonne pratique : lorsque vous évaluez les améliorations potentielles liées aux performances, vous devez décider si les compromis concernant les

modifications sont compatibles avec les exigences de charge de travail. Dans certains cas, vous devrez peut-être mettre en place des contrôles supplémentaires pour compenser les compromis.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Identifiez les domaines critiques de votre architecture en termes de performances et d'impact sur les clients. Déterminez la façon dont vous pouvez apporter des améliorations ainsi que les compromis que ces améliorations entraînent et la façon dont ils affectent le système et l'expérience de l'utilisateur. Par exemple, la mise en œuvre de la mise en cache des données permet d'améliorer de manière significative les performances, mais nécessite une stratégie précise concernant la manière et le moment où mettre à jour ou invalider les données mises en cache pour empêcher un comportement incorrect du système.

Étapes d'implémentation

- Comprenez vos exigences en matière de charge de travail etSLAs.
- Définissez clairement les facteurs d'évaluation. Les facteurs peuvent être liés au coût, à la fiabilité, à la sécurité et aux performances de votre charge de travail.
- Sélectionnez l'architecture et les services qui répondent à vos besoins.
- Mener des expériences et des validations de concepts (POCs) pour évaluer les facteurs de compromis et leur impact sur les clients et l'efficacité de l'architecture. En général, les charges de travail hautement disponibles, performantes et sécurisées consomment davantage de ressources cloud tout en offrant une meilleure expérience client. Comprenez les compromis entre la complexité, les performances et les coûts de votre charge de travail. Généralement, la priorisation de deux des facteurs se fait au détriment du troisième.

Ressources

Documents connexes :

- Bibliothèque Amazon Builders' Library
- Amazon QuickSight KPIs
- Amazon CloudWatch RUM
- Documentation X-Ray
- Comprenez les modèles de résilience et les compromis pour concevoir une architecture efficace dans le cloud

Vidéos connexes :

- Optimisez les applications via Amazon CloudWatch RUM
- AWS re:Invent 2023 - Capacité, disponibilité, rentabilité : choisissez trois
- AWS re:Invent 2023 - Modèles d'intégration avancés et compromis pour les systèmes faiblement couplés

Exemples connexes :

- Mesurez le temps de chargement des pages avec Amazon CloudWatch Synthetics
- Client CloudWatch RUM Web Amazon

PERF01-BP05 Politiques d'utilisation et architectures de référence

Utilisez les stratégies internes et les architectures de référence existantes lors de la sélection des services et des configurations en vue d'augmenter votre efficacité lorsque vous concevez et mettez en œuvre votre charge de travail.

Anti-modèles courants :

- Vous autorisez un large éventail de technologies qui peuvent avoir un impact sur les frais généraux de gestion de votre entreprise.

Avantages liés au respect de cette bonne pratique : l'établissement d'une stratégie pour les choix d'architecture, de technologie et de fournisseur permet de prendre des décisions rapidement.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Le fait de disposer de stratégies internes en matière de sélection des ressources et de l'architecture fournit des normes et des directives à suivre lors des choix architecturaux. Ces directives simplifient le processus de prise de décision lors du choix du bon service cloud et peuvent contribuer à améliorer l'efficacité des performances. Déployez votre charge de travail à l'aide de stratégies ou d'architectures de référence. Intégrez les services à votre déploiement dans le cloud. Utilisez ensuite vos tests de performance pour vérifier que vous pouvez continuer à répondre à vos exigences de performance.

Étapes d’implémentation

- Comprenez clairement les exigences de votre charge de travail cloud.
- Passez en revue les stratégies internes et externes pour identifier les plus pertinentes.
- Utilisez les architectures de référence appropriées fournies par AWS ou les bonnes pratiques de votre secteur.
- Créez un continuum composé de stratégies, de normes, d’architectures de référence et de directives normatives pour les situations courantes. Vos équipes pourront ainsi agir plus rapidement. Adaptez les ressources à votre secteur d’activité, le cas échéant.
- Validez ces stratégies et architectures de référence pour votre charge de travail dans les environnements de test (sandbox).
- up-to-dateRespectez les normes et les AWS mises à jour du secteur pour vous assurer que vos politiques et architectures de référence contribuent à optimiser votre charge de travail dans le cloud.

Ressources

Documents connexes :

- Centre d’architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- AWS Blogue d'architecture

Vidéos connexes :

- This is my Architecture
- AWS re:Invent 2022 - Accélérez la création de valeur pour votre entreprise grâce à une architecture de SAP référence AWS

Exemples connexes :

- Exemples AWS
- AWS SDKExemples

PERF01-BP06 Utilisation du benchmarking pour éclairer vos décisions architecturales

Définissez des points de référence pour les performances d'une charge de travail existante afin de comprendre ses performances sur le cloud et prendre des décisions architecturales sur la base de ces données.

Anti-modèles courants :

- Vous comptez sur des points de référence courants qui ne reflètent pas les caractéristiques de votre charge de travail.
- Vous utilisez les commentaires et la perception des clients comme seule référence.

Avantages de l'établissement de cette bonne pratique : le benchmarking de votre implémentation actuelle vous permet de mesurer les améliorations de performance.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Utilisez la définition de points de référence avec des tests synthétiques pour évaluer les performances des composants de votre charge de travail. Le benchmarking est généralement plus rapide à configurer que les tests de charge. Il est utilisé pour évaluer la technologie pour un composant en particulier. Le benchmarking est souvent utilisé au début d'un nouveau projet, lorsque vous n'avez pas de solution complète pour le test de charge.

Vous pouvez créer vos propres tests de performances, ou bien utiliser un test conforme aux normes du secteur, comme le TPC-DS pour comparer vos charges de travail. Les points de référence du secteur sont utiles lorsque vous comparez différents environnements. Les points de référence personnalisés sont utiles pour cibler certains types d'opérations que vous souhaitez effectuer dans votre architecture.

Avec le benchmarking, il est important de préparer votre environnement de test pour obtenir des résultats valides. Exécutez plusieurs fois le même point de référence pour vous assurer d'avoir capturé toute variabilité au fil du temps.

Étant donné que les points de référence sont généralement plus rapides à exécuter que les tests de charge, ils peuvent être utilisés plus tôt dans le pipeline de déploiement et fournir un retour rapide sur les écarts de performances. Lorsque vous évaluez un changement important dans un composant ou un service, un point de référence peut être un moyen rapide pour voir si la modification a un intérêt.

L'utilisation du benchmarking avec un test de charge est essentielle, car un test de charge vous indique comment votre charge de travail se comporte dans un environnement de production.

Étapes d'implémentation

- Planification et définition :
  - Définissez les objectifs, la base de référence, les scénarios de test, les métriques (telles que l'utilisation du processeur, la latence ou le débit) et les indicateurs de rendement clés de votre test de performances.
  - Concentrez-vous sur les exigences des utilisateurs en matière d'expérience utilisateur et sur des facteurs tels que le temps de réponse et l'accessibilité.
  - Identifiez un outil de benchmarking adapté à votre charge de travail. Vous pouvez utiliser des services AWS tels qu'Amazon CloudWatch ou un outil tiers compatible avec votre charge de travail.
- Configuration et instrumentation :
  - Configurez votre environnement et vos ressources.
  - Mettez en œuvre la surveillance et la journalisation pour capturer les résultats des tests.
- Comparaison et surveillance :
  - Effectuez vos tests de performances et surveillez les métriques pendant le test.
- Analyse et documentation :
  - Documentez votre processus de benchmarking et vos résultats.
  - Analysez les résultats pour identifier les goulots d'étranglement, les tendances et les domaines d'amélioration.
  - Utilisez les résultats des tests pour prendre des décisions architecturales et ajuster votre charge de travail. Cet ajustement peut impliquer la modification des services ou l'adoption de nouvelles fonctionnalités.
- Optimisation et répétition :
  - Ajustez les configurations et les allocations des ressources en fonction de vos critères de référence.
  - Testez à nouveau votre charge de travail après ajustement pour valider vos améliorations.
  - Documentez vos conclusions et répétez le processus pour identifier d'autres domaines d'amélioration.

Ressources

Documents connexes :

- Centre d’architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Flux de travail génomiques, partie 5 : benchmarking automatisé
- Évaluation et optimisation du déploiement des points de terminaison dans Amazon SageMaker AI JumpStart

Vidéos connexes :

- AWS re:Invent 2023 - Benchmarking AWS Lambda cold starts
- Benchmarking stateful services in the cloud
- This is my Architecture
- Optimize applications through via Amazon CloudWatch RUM
- Présentation d’Amazon CloudWatch Synthetics

Exemples connexes :

- Exemples AWS
- Exemples de kit SDK AWS
- Tests de charge distribuée
- Mesure du temps de chargement des pages avec Amazon CloudWatch Synthetics
- Client Web Amazon CloudWatch RUM

PERF01-BP07 Utiliser une approche axée sur les données pour les choix architecturaux

Définissez une approche orientée données claire pour les choix architecturaux afin de vérifier que les services et configurations cloud appropriés sont utilisés pour répondre aux besoins spécifiques de votre entreprise.

Anti-modèles courants :

- Vous supposez que votre architecture actuelle est statique et ne devrait pas être mise à jour au fil du temps.
- Vos choix architecturaux sont basés sur des suppositions et des hypothèses.
- Vous introduisez des modifications d'architecture au fil du temps sans justification.

Avantages liés au respect de cette bonne pratique : en adoptant une approche bien définie pour les choix architecturaux, vous utilisez les données pour influencer la conception de votre charge de travail et prendre des décisions éclairées au fil du temps.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Mobilisez l'expérience et l'expertise des ressources cloud internes ou faites appel à des ressources externes, comme des cas d'utilisation publiés ou des livres blancs pour choisir les ressources et services dans votre architecture. Vous devriez disposer d'un processus bien défini qui encourage l'expérimentation et le benchmarking avec les services qui pourraient être utilisés dans votre charge de travail.

Les backlogs relatifs aux charges de travail critiques doivent non seulement comprendre des témoignages d'utilisateurs proposant des fonctionnalités pertinentes pour les entreprises et les utilisateurs, mais également des récits techniques qui constituent une piste architecturale pour la charge de travail. Cette piste s'inspire des nouvelles avancées technologiques et des nouveaux services et les adopte sur la base de données et de justifications appropriées. Cela permet de vérifier que l'architecture reste pérenne et ne stagne pas.

Étapes d'implémentation

- Collaborez avec les principales parties prenantes pour définir les exigences en matière de charge de travail, y compris les considérations relatives aux performances, à la disponibilité et aux coûts. Tenez compte de facteurs tels que le nombre d'utilisateurs et le modèle d'utilisation de votre charge de travail.

- Créez une piste architecturale ou un backlog technologique qui est axé en priorité sur le backlog fonctionnel.
- Évaluez les différents services cloud (pour en savoir plus, consultez PERF01-BP01 Découvrez et comprenez les services et fonctionnalités cloud disponibles).
- Explorez les différents modèles architecturaux, tels que les microservices ou le modèle sans serveur, qui répondent à vos exigences en termes de performances (pour en savoir plus, consultez PERF01-BP02 Suivez les conseils de votre fournisseur de cloud ou d'un partenaire approprié pour en savoir plus sur les modèles d'architecture et les meilleures pratiques).

- Consultez d'autres équipes, des diagrammes d'architecture et des ressources, telles que AWS Solutions Architects, AWS Architecture Center, etc. AWS Partner Network, pour vous aider à choisir l'architecture adaptée à votre charge de travail.

- Définissez des métriques de performances telles que le débit et le temps de réponse qui peuvent vous aider à évaluer les performances de votre charge de travail.
- Testez et utilisez des métriques définies pour valider les performances de l’architecture sélectionnée.
- Surveillez en continu les performances et effectuez les ajustements nécessaires pour maintenir un niveau optimal de performance pour votre architecture.
- Documentez l’architecture que vous avez sélectionnée et les décisions que vous avez prises comme référence pour les futures mises à jour et les futurs apprentissages.
- Vérifiez en permanence l’approche de sélection de l’architecture et mettez-la à jour en fonction des apprentissages, des nouvelles technologies et des métriques indiquant un changement nécessaire ou un problème dans l’approche actuelle.

Ressources

Documents connexes :

- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- Modèles architecturaux sur lesquels créer End-to-End des applications basées sur les données AWS

Vidéos connexes :

- This is my Architecture
- AWS re:Invent 2021 - L'entreprise axée sur les données : passer de la vision à la valeur
- AWS re:Invent 2022 - Fournir des architectures durables et performantes
- AWS re:Invent 2023 - Optimisez les coûts et les performances et suivez les progrès en matière d'atténuation
- AWS re:Invent 2022 - AWS optimisation : étapes réalisables pour des résultats immédiats

Exemples connexes :

- Exemples AWS
- AWS SDKExemples

Informatique et matériel

Questions

- PERF 2. Comment sélectionner et utiliser les ressources de calcul dans votre charge de travail ?

PERF 2. Comment sélectionner et utiliser les ressources de calcul dans votre charge de travail ?

Le choix d'une solution de calcul optimale pour une charge de travail particulière peut varier selon la conception de l'application, les modèles d'utilisation et les paramètres de configuration. Les architectures peuvent utiliser différentes solutions de calcul pour divers composants et permettent différentes fonctionnalités pour améliorer les performances. Le choix d'une solution de calcul inadaptée à une architecture peut nuire à ses performances.

Bonnes pratiques

- PERF02-BP01 Sélectionnez les meilleures options de calcul pour votre charge de travail
- PERF02-BP02 Comprendre la configuration et les fonctionnalités de calcul disponibles
- PERF02-BP03 Collecter des métriques liées au calcul
- PERF02-BP04 Configurer et dimensionner correctement les ressources de calcul
- PERF02-BP05 Adaptez dynamiquement vos ressources informatiques

- PERF02-BP06 Utiliser des accélérateurs de calcul matériels optimisés

PERF02-BP01 Sélectionnez les meilleures options de calcul pour votre charge de travail

La sélection de l'option de calcul la mieux adaptée à votre charge de travail vous permet d'améliorer les performances, de réduire les coûts d'infrastructure inutiles et de diminuer les efforts opérationnels nécessaires pour maintenir votre charge de travail.

Anti-modèles courants :

- Vous utilisez la même option de calcul que celle utilisée sur site.
- Vous manquez de connaissances sur les options, les fonctionnalités et les solutions de cloud computing et sur la manière dont elles pourraient améliorer vos performances de calcul.
- Vous surprovisionnez une option de calcul existante pour répondre aux exigences de mise à l'échelle ou de performances, alors qu'une autre option de calcul s'alignerait plus précisément sur les caractéristiques de votre charge de travail.

Avantages liés au respect de cette bonne pratique : en identifiant les exigences de calcul et en les comparant aux options disponibles, vous pouvez optimiser votre charge de travail en termes de ressources.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Pour optimiser vos charges de travail dans le cloud en termes d'efficacité des performances, il est important de sélectionner les options de calcul les plus adaptées à votre cas d'utilisation et à vos exigences de performance. AWS fournit une variété d'options de calcul adaptées aux différentes charges de travail dans le cloud. Par exemple, vous pouvez utiliser Amazon EC2 pour lancer et gérer des serveurs virtuels, AWS Lambdapour exécuter du code sans avoir à approvisionner ou à gérer des serveurs, Amazon ECS ou Amazon EKS pour exécuter et gérer des conteneurs, ou AWS Batchpour traiter de gros volumes de données en parallèle. En fonction de vos besoins en termes de mise à l'échelle et de calcul, vous devez choisir et configurer la solution de calcul optimale pour votre situation. Vous pouvez également envisager d'utiliser plusieurs types de solutions de calcul dans une seule charge de travail, car chacune présente ses avantages et ses inconvénients.

Les étapes suivantes vous guident dans la sélection des options de calcul adaptées aux caractéristiques de votre charge de travail et à vos exigences de performances.

Étapes d'implémentation

- Comprenez les exigences de calcul de votre charge de travail. Les exigences clés à prendre en compte incluent les besoins de traitement, les modèles de trafic, les modèles d'accès aux données, les besoins de mise à l'échelle et les exigences de latence.
- Découvrez les différents services AWS informatiques adaptés à votre charge de travail. Pour de plus amples informations, veuillez consulter PERF01-BP01 Découvrez et comprenez les services et fonctionnalités cloud disponibles. Voici quelques options de calcul AWS clés, leurs caractéristiques et leurs cas d'utilisation courants :

AWS service	Principales caractéristiques	Cas d'utilisation courants
Amazon Elastic Compute Cloud (AmazonEC2)	Possède une option dédiée pour le matériel, les exigences de licence, une large sélection de différent es familles d'instances, les types de processeurs et les accélérateurs de calcul	Migration « lift-and-shift », application monolithique, environnements hybrides, applications d'entreprise
Amazon Elastic Container Service (AmazonECS), Amazon Elastic Kubernetes Service (Amazon) EKS	Déploiement facile, environne ments cohérents, évolutivité	Microservices, environne ments hybrides
AWS Lambda	Service de calcul sans serveur qui exécute du code en réponse à des événement s et gère automatiquement les ressources de calcul sous-jacentes.	Microservices, applications basées sur les événements
AWS Batch	Provisionne et fait évoluer de manière efficace et dynamique Amazon Elastic Container Service (AmazonECS), Amazon	HPC, train les modèles ML

AWS service	Principales caractéristiques	Cas d’utilisation courants
	Elastic Kubernetes Service (EKSAmazon) et les ressources de calcul AWS Fargate, avec la possibilité d'utiliser des instances à la demande ou ponctuelles en fonction des exigences de votre poste	
Amazon Lightsail	Application Linux et Windows préconfigurée pour exécuter de petites charges de travail	Applications web simples, site web personnalisé

- Évaluez les coûts (tels que le tarif horaire ou le transfert de données) et les frais de gestion (tels que l’application de correctifs et la mise à l’échelle) associés à chaque option de calcul.
- Réalisez des tests et procédez au benchmarking dans un environnement hors production afin d’identifier l’option de calcul la mieux adaptée à vos exigences en termes de charge de travail.
- Après avoir testé et identifié votre nouvelle solution de calcul, planifiez votre migration et validez vos métriques de performance.
- Utilisez AWS des outils de surveillance tels qu'Amazon CloudWatch et des services d'optimisation AWS Compute Optimizerpour optimiser en permanence vos ressources informatiques en fonction de modèles d'utilisation réels.

Ressources

Documents connexes :

- Cloud Compute with AWS
- Types d'EC2instances Amazon
- EKSConteneurs Amazon : Amazon EKS Worker Nodes
- Amazon ECS Containers : instances de ECS conteneurs Amazon
- Fonctions : configuration des fonctions Lambda
- Recommandations pour les conteneurs

- Recommandations pour les modèles sans serveur

Vidéos connexes :

- AWS re:Invent 2023 - AWS Graviton : le meilleur rapport qualité/prix pour vos charges de travail AWS
- AWS re:Invent 2023 - Nouvelles fonctionnalités d'IA générative d'Amazon Elastic Compute Cloud dans AMS
- AWS re:Invent 2023 - What’s new with Amazon Elastic Compute Cloud
- AWS re:Invent 2023 - Smart savings: Amazon Elastic Compute Cloud cost-optimization strategies
- AWS re:Invent 2021 - Powering next-gen Amazon Elastic Compute Cloud: Deep dive on the Nitro System
- AWS re:Invent 2019 - Optimisez les performances et les coûts de votre calcul AWS
- AWS re:Invent 2019 - Amazon Elastic Compute Cloud foundations
- AWS re:Invent 2022 - Déployez des modèles ML pour l'inférence à des performances élevées et à moindre coût
- AWS re:Invent 2019 - Optimisez les performances et les coûts de votre calcul AWS
- EC2Fondations Amazon
- Deploy ML models for inference at high performance and low cost

Exemples connexes :

- Migration de l’application Web vers des conteneurs
- Exécution d’un modèle Hello World sans serveur
- EKSAtelier Amazon
- EC2Atelier Amazon
- Charges de travail efficaces et résilientes avec l’autoscaling Amazon EC2 Auto Scaling
- Migrer vers AWS Graviton avec Container Services

PERF02-BP02 Comprendre la configuration et les fonctionnalités de calcul disponibles

Découvrez les options et les fonctionnalités de configuration disponibles pour votre service de calcul qui vous aideront à allouer la quantité de ressources appropriée et à améliorer l’efficacité des performances.

Anti-modèles courants :

- Vous ne comparez pas les options de calcul ni les familles d'instances disponibles avec les caractéristiques de la charge de travail.
- Vous surprovisionnez les ressources de calcul pour répondre aux pics de demande.

Avantages de l'établissement de cette meilleure pratique : familiarisez-vous avec les fonctionnalités et les configurations de AWS calcul afin de pouvoir utiliser une solution informatique optimisée pour répondre aux caractéristiques et aux besoins de votre charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Chaque solution de calcul dispose de configurations et de fonctionnalités uniques pour prendre en charge différentes caractéristiques et exigences de charge de travail. Découvrez comment ces options soutiennent votre charge de travail et déterminez celles qui sont optimales pour votre système. Ces options incluent par exemple la famille d'instances, les tailles, les fonctionnalités (E/S)GPU, le bursting, les délais d'expiration, la taille des fonctions, les instances de conteneur et la simultanéité. Si votre charge de travail utilise la même option de calcul depuis plus de quatre semaines et que vous pensez que les caractéristiques resteront les mêmes à l'avenir, vous pouvez vérifier si votre option de calcul actuelle est adaptée aux charges de travail CPU et du point de vue de la mémoire. AWS Compute Optimizer

Étapes d'implémentation

- Comprenez les exigences en matière de charge de travail (comme les CPU besoins, la mémoire et le temps de latence).
- Consultez AWS la documentation et les meilleures pratiques pour découvrir les options de configuration recommandées qui peuvent contribuer à améliorer les performances de calcul. Voici quelques options de configuration clés à prendre en compte :

Option de configuration	Exemples
Type d'instance	• Les instances optimisées pour le calcul sont idéales pour les charges de travail qui nécessitent un rapport v/mémoire CPU élevé.

Option de configuration	Exemples
	• Les instances à mémoire optimisée offrent de grandes quantités de mémoire pour soutenir les charges de travail gourmandes en mémoire. • Les instances optimisées pour le stockage sont conçues pour les charges de travail qui nécessitent un accès séquentiel élevé en lecture et en écriture (IOPS) au stockage local.
Modèle de tarification	• Les instances à la demande vous permettent d'utiliser la capacité de calcul à l'heure ou à la seconde, sans engagement à long terme. Ces instances sont idéales pour dépasser les besoins de base en matière de performances. • Les Savings Plans permettent de réaliser des économies importantes par rapport aux instances à la demande, en échange d'un engagement à utiliser une quantité spécifique de puissance de calcul pour une période d'un ou de trois ans. • Les instances Spot vous permettent de tirer parti de la capacité d'instance inutilisée à un prix réduit pour vos charges de travail sans état et tolérantes aux pannes.
Auto Scaling	Utilisez la configuration Auto Scaling pour faire correspondre les ressources de calcul aux modèles de trafic.

Option de configuration	Exemples
Dimensionnement	• Utilisez Compute Optimizer pour obtenir des recommandations basées sur le machine learning sur la configuration de calcul qui correspond le mieux à vos caractéristiques de calcul. • Utilisez AWS Lambda Power Tuning pour sélectionner la meilleure configuration pour votre fonction Lambda.
Accélérateurs de calcul matériels	• Les instances de calcul accéléré exécutent des fonctions telles que le traitement graphique ou la mise en correspondance de modèles de données de manière plus efficace que les alternatives CPU basées sur des données. • Pour les charges de travail liées à l'apprentissage automatique, profitez d'un matériel spécialement conçu pour votre charge de travail, tel que AWS Trainium, Inferentia et Amazon AWS EC2 DL1

Ressources

Documents connexes :

- Cloud Compute with AWS
- Types d'EC2instances Amazon
- Contrôle de l'état du processeur pour votre EC2 instance Amazon
- EKSConteneurs Amazon : Amazon EKS Worker Nodes
- Amazon ECS Containers : instances de ECS conteneurs Amazon
- Fonctions : configuration des fonctions Lambda

Vidéos connexes :

- AWS re:Invent 2023 — AWS Graviton : le meilleur rapport qualité/prix pour vos charges de travail AWS
- AWS re:Invent 2023 — Nouvelles fonctionnalités d'IA EC2 générative d'Amazon dans AWS Management Console
- AWS re:Invent 2023 — Nouveautés d'Amazon EC2
- AWS re:Invent 2023 — Économies intelligentes : stratégies d'optimisation des coûts d'Amazon EC2
- AWS re:Invent 2021 — Au service d'EC2Amazon de nouvelle génération : étude approfondie du système Nitro
- AWS re:Invent 2019 — Amazon Foundations EC2
- AWS re:Invent 2022 — Optimisation des performances et EKS des coûts d'Amazon AWS

Exemples connexes :

- Code de démonstration de Compute Optimizer
- Atelier sur les instances Amazon EC2 Spot
- Charges de travail efficaces et résilientes avec Amazon EC2 AWS Auto Scaling
- Atelier pour développeurs Graviton
- AWS journée d'immersion pour les charges de travail Microsoft
- AWS journée d'immersion pour les charges de travail Linux
- AWS Compute Optimizer Code de démonstration
- EKSAtelier Amazon

PERF02-BP03 Collecter des métriques liées au calcul

Enregistrez et suivez les métriques liées au calcul pour mieux comprendre comment fonctionnent vos ressources de calcul et améliorer leurs performances et leur utilisation.

Anti-modèles courants :

- Vous utilisez uniquement la recherche manuelle des fichiers journaux pour les métriques.
- Vous n’utilisez que les métriques par défaut enregistrées par votre logiciel de surveillance.
- Vous n’examinez les métriques qu’en cas de problème.

Avantages liés au respect de cette bonne pratique : en collectant des métriques liées aux performances, vous pouvez aligner les performances des applications sur les exigences de l'entreprise afin de garantir que vous répondez à vos besoins en matière de charge de travail. Cela peut également vous aider à améliorer en continu les performances et l'utilisation des ressources de votre charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les charges de travail cloud peuvent générer de gros volumes de données telles que des métriques, des journaux et des événements. Dans ce contexte AWS Cloud, la collecte de métriques est une étape cruciale pour améliorer la sécurité, la rentabilité, les performances et la durabilité. AWS fournit un large éventail de mesures liées aux performances à l'aide de services de surveillance tels qu'Amazon CloudWatch pour vous fournir des informations précieuses. Des indicateurs tels que CPU l'utilisation, l'utilisation de la mémoire, les E/S du disque et les entrées et sorties du réseau peuvent fournir des informations sur les niveaux d'utilisation ou les goulots d'étranglement des performances. Utilisez ces métriques dans le cadre d'une approche fondée sur les données pour ajuster activement et optimiser les ressources de votre charge de travail. Dans un scénario idéal, vous devriez collecter toutes les métriques relatives à vos ressources de calcul sur une plateforme unique, avec des stratégies de conservation mises en œuvre pour atteindre les objectifs financiers et opérationnels.

Étapes d'implémentation

- Identifiez les métriques liées aux performances qui sont pertinentes pour votre charge de travail. Vous devriez collecter des métriques relatives à l'utilisation des ressources et au fonctionnement de votre charge de travail cloud (comme le temps de réponse et le débit).
  - Métriques EC2 par défaut d'Amazon
  - Métriques ECS par défaut d'Amazon
  - Métriques EKS par défaut d'Amazon
  - Métriques par défaut de Lambda
  - Métriques relatives à EC2 la mémoire et au disque Amazon
- Choisissez et configurez la solution de journalisation et de surveillance adaptée à votre charge de travail.
  - Observabilité native AWS
  - AWS Distro pour OpenTelemetry
  - Amazon Managed Service for Prometheus

- Définissez le filtre et l’agrégation requis pour les métriques en fonction de vos exigences en matière de charge de travail.
  - Quantifiez les métriques personnalisées des applications avec Amazon CloudWatch Logs et les filtres métriques
  - Collectez des statistiques personnalisées grâce au balisage CloudWatch stratégique d'Amazon
- Configurez des stratégies de conservation des données pour vos métriques afin qu’elles correspondent à vos objectifs sécuritaires et opérationnels.
  - Conservation des données par défaut pour les CloudWatch métriques
  - Conservation des données par défaut pour les CloudWatch journaux
- Si nécessaire, créez des alarmes et des notifications pour vos métriques afin de vous aider à résoudre de manière proactive les problèmes liés aux performances.
  - Créez des alarmes pour des métriques personnalisées à l'aide de la détection des CloudWatch anomalies Amazon
  - Créez des métriques et des alarmes pour des pages Web spécifiques avec Amazon CloudWatch RUM
- Utilisez l’automatisation pour déployer vos agents d’agrégation de métriques et de journaux.
  - AWS Systems Manager automatisation
  - OpenTelemetryCollectionneur

Ressources

Documents connexes :

- Surveillance et observabilité
- Bonnes pratiques : mise en œuvre de l'observabilité avec AWS
- CloudWatch Documentation Amazon
- Collectez des métriques et des journaux à partir d'EC2instances Amazon et de serveurs sur site avec l'agent CloudWatch
- Accès à Amazon CloudWatch Logs pour AWS Lambda
- Utilisation CloudWatch des journaux avec des instances de conteneur
- Publier des métriques personnalisées
- AWS Réponse : journalisation centralisée
- AWS Services qui publient CloudWatch des métriques

- Surveillance d'Amazon EKS sur AWS Fargate

Vidéos connexes :

- AWS re:Invent 2023 — [LAUNCH] Surveillance des applications pour les charges de travail modernes
- AWS re:Invent 2023 — Mise en œuvre de l'observabilité des applications
- AWS re:Invent 2023 — Élaborer une stratégie d'observabilité efficace
- AWS re:Invent 2023 — Une observabilité sans faille avec Distro pour AWS OpenTelemetry
- Gestion des performances des applications sur AWS

Exemples connexes :

- AWS Journée d'immersion pour les charges de travail Linux - Amazon CloudWatch
- Surveillance des ECS clusters et des conteneurs Amazon
- Surveillance à l'aide des tableaux de CloudWatch bord Amazon
- EKSAtelier Amazon

PERF02-BP04 Configurer et dimensionner correctement les ressources de calcul

Configurez et dimensionnez correctement les ressources de calcul en fonction des exigences de performance de votre charge de travail et évitez de sous-utiliser ou de surexploiter les ressources.

Anti-modèles courants :

- Vous ignorez les exigences de performance de votre charge de travail, ce qui entraîne un surprovisionnement ou un sous-provisionnement des ressources de calcul.
- Vous ne choisissez que la plus grande ou la plus petite instance disponible pour toutes les charges de travail.
- Vous n’utilisez qu’une seule famille d’instances pour faciliter la gestion.
- Vous ignorez les recommandations de AWS Cost Explorer Compute Optimizer concernant le dimensionnement correct.
- Vous ne réévaluez pas la charge de travail pour voir si de nouveaux types d’instances pourraient convenir.
- Vous ne certifiez qu’un petit nombre de configurations d’instance pour votre organisation.

Avantages liés au respect de cette bonne pratique : dimensionner correctement les ressources de calcul garantit le fonctionnement optimal dans le cloud en évitant le surprovisionnement et le sous-provisionnement des ressources. Le dimensionnement correct des ressources de calcul se traduit généralement par des performances renforcées, une meilleure expérience client et une baisse des coûts.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Le dimensionnement correct permet aux organisations d'exploiter leur infrastructure cloud de manière efficace et rentable tout en répondant aux besoins de l'entreprise. Le surprovisionnement des ressources cloud peut entraîner des coûts supplémentaires, tandis que le sous-provisionnement peut entraîner des performances médiocres et une expérience client négative. AWS fournit des outils tels que AWS Compute Optimizeret AWS Trusted Advisorqui utilisent des données historiques pour fournir des recommandations afin de dimensionner correctement vos ressources informatiques.

Étapes d'implémentation

- Choisissez le type d'instance qui correspond le mieux à vos besoins :
  - Comment choisir le type d'EC2instance Amazon adapté à ma charge de travail ?
  - Sélection du type d'instance basée sur les attributs pour Amazon Fleet EC2
  - Créer un groupe Auto Scaling en utilisant la sélection du type d'instance basée sur des attributs
  - Optimisation de vos coûts de calcul Kubernetes avec la consolidation Karpenter
- Analysez les différentes caractéristiques de performance de votre charge de travail et le lien entre ces caractéristiques et la mémoire, le réseau et CPU l'utilisation. Utilisez ces données pour choisir les ressources qui correspondent le mieux aux objectifs de votre charge de travail en matière de profil et de performance.
- Surveillez l'utilisation de vos ressources à l'aide d'outils de AWS surveillance tels qu'Amazon CloudWatch.
- Sélectionnez la configuration adaptée à vos ressources de calcul.
  - Pour les charges de travail éphémères, évaluez les CloudWatch indicateurs Amazon de l'instance, `CPUUtilization` afin de déterminer si l'instance est sous-utilisée ou surutilisée.
  - Pour des charges de travail stables, vérifiez les AWS outils de redimensionnement tels que AWS Compute Optimizer et AWS Trusted Advisor à intervalles réguliers pour identifier les opportunités d'optimisation et de dimensionnement correct de la ressource de calcul.

- Testez les changements de configuration dans un environnement hors production avant de les implémenter dans un environnement réel.
- Réévaluez en permanence les nouvelles offres de calcul et comparez-les aux besoins de votre charge de travail.

Ressources

Documents connexes :

- Cloud Compute avec AWS
- Types d'EC2instances Amazon
- Amazon ECS Containers : instances de ECS conteneurs Amazon
- EKSConteneurs Amazon : Amazon EKS Worker Nodes
- Fonctions : configuration des fonctions Lambda
- Contrôle de l'état du processeur pour votre EC2 instance Amazon

Vidéos connexes :

- EC2Fondations Amazon
- AWS re:Invent 2023 — AWS Graviton : le meilleur rapport qualité/prix pour vos charges de travail AWS
- AWS re:Invent 2023 — Nouvelles fonctionnalités d'IA EC2 générative d'Amazon dans AWS Management Console
- AWS re:Invent 2023 — Nouveautés d'Amazon EC2
- AWS re:Invent 2023 — Économies intelligentes : stratégies d'optimisation des coûts d'Amazon EC2
- AWS re:Invent 2021 — Au service d'EC2Amazon de nouvelle génération : étude approfondie du système Nitro
- AWS re:Invent 2019 — Amazon Foundations EC2

Exemples connexes :

- AWS Compute Optimizer Code de démonstration
- EKSAtelier Amazon
- Recommandations en matière de redimensionnement

PERF02-BP05 Adaptez dynamiquement vos ressources informatiques

Utilisez l’élasticité du cloud pour mettre à l’échelle vos ressources de calcul de manière dynamique afin de répondre à vos besoins et d’éviter de surprovisionner ou de sous-provisionner la capacité de votre charge de travail.

Anti-modèles courants :

- Vous réagissez aux alertes en augmentant manuellement la capacité.
- Vous utilisez les mêmes recommandations de dimensionnement (généralement, infrastructure statique) que sur site.
- Vous conservez une capacité accrue après un événement de mise à l’échelle au lieu de la réduire.

Avantages liés au respect de cette bonne pratique : en configurant et en testant l’élasticité des ressources de calcul, vous pouvez économiser de l’argent, maintenir les points de référence des performances et améliorer la fiabilité en fonction de l’évolution du trafic.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

AWS offre la flexibilité nécessaire pour augmenter ou diminuer vos ressources de manière dynamique grâce à divers mécanismes de mise à l'échelle afin de répondre à l'évolution de la demande. Combinée aux métriques liées au calcul, la mise à l’échelle dynamique permet aux charges de travail de réagir automatiquement aux changements et d’utiliser l’ensemble optimal de ressources de calcul pour atteindre son objectif.

Vous pouvez utiliser plusieurs approches pour rapprocher l’offre de ressources de la demande.

- Approche de suivi des objectifs : surveillez votre métrique de capacité de mise à l’échelle et augmentez ou réduisez automatiquement votre capacité selon vos besoins.
- Mise à l’échelle prédictive : mettez à l’échelle en prévision des tendances quotidiennes et hebdomadaires.
- Approche basée sur le calendrier : définissez votre propre calendrier de mise à l’échelle en fonction de changements de charge prévisibles.
- Mise à l’échelle des services : choisissez des services (sans serveur, par exemple) conçus pour se mettre à l’échelle automatiquement.

Vous devez vous assurer que les déploiements de charge de travail peuvent gérer les événements de mise à l'échelle ascendante et descendante.

Étapes d'implémentation

- Les instances de calcul, les conteneurs et les fonctions fournissent des mécanismes d'élasticité, soit en combinaison avec l'autoscaling, soit en tant que fonctionnalité du service. Voici des exemples de mécanismes d'autoscaling :

Mécanisme d'autoscaling	Où utiliser
Amazon EC2 Auto Scaling	Pour vous assurer que vous disposez du nombre correct d'EC2instances Amazon disponibles pour gérer la charge utilisateur de votre application.
Application Autoscaling	Pour dimensionner automatiquement les ressources pour des AWS services individue ls autres qu'Amazon, EC2 tels que AWS Lambdales fonctions ou les services Amazon Elastic Container Service (AmazonECS).
Outil Cluster Autoscaler/Karpenter de Kubernetes	Pour mettre à l'échelle automatiquement les clusters Kubernetes.

- La mise à l'échelle est souvent évoquée en lien avec les services de calcul tels que EC2 les instances ou AWS Lambda les fonctions Amazon. Assurez-vous également de prendre en compte la configuration des services non liés au calcul tels que AWS Glue pour répondre à la demande.
- Vérifiez que les métriques de mise à l'échelle correspondent aux caractéristiques de la charge de travail en cours de déploiement. Si vous déployez une application de transcodage vidéo, un taux d'CPUutilisation de 100 % est attendu et ne doit pas être votre indicateur principal. Utilisez plutôt la profondeur de la file d'attente des tâches de transcodage. Le cas échéant, vous pouvez utiliser une métrique personnalisée pour votre politique de dimensionnement. Pour choisir les bons indicateurs, prenez en compte les conseils suivants destinés à Amazon EC2 :
  - La métrique doit être une métrique d'utilisation valide et décrire à quel point l'instance est occupée.
  - La valeur de métrique doit augmenter ou diminuer en proportion du nombre d'instances présentes dans le groupe Auto Scaling.

- Assurez-vous d'utiliser une mise à l'échelle dynamique plutôt qu'une mise à l'échelle manuelle pour votre groupe Auto Scaling. Nous vous recommandons également d'utiliser des politiques de dimensionnement pour le suivi des cibles dans votre dimensionnement dynamique.
- Vérifiez que les déploiements de charges de travail peuvent gérer les deux événements de mise à l'échelle (augmentation et diminution des charges de travail). Par exemple, vous pouvez utiliser l'historique des activités pour vérifier une activité de mise à l'échelle dans un groupe Auto Scaling.
- Évaluez votre charge de travail pour les modèles prédictifs et mettez-la à l'échelle de manière proactive pour anticiper les changements prévisibles et prévus de la demande. Avec la mise à l'échelle prédictive, vous pouvez supprimer le besoin de surprovisionner de la capacité. Pour plus de détails, consultez Predictive Scaling with Amazon EC2 Auto Scaling.

Ressources

Documents connexes :

- Cloud Compute avec AWS
- Types d'EC2instances Amazon
- Amazon ECS Containers : instances de ECS conteneurs Amazon
- EKSConteneurs Amazon : Amazon EKS Worker Nodes
- Fonctions : configuration des fonctions Lambda
- Contrôle de l'état du processeur pour votre EC2 instance Amazon
- Présentation approfondie d'Amazon ECS Cluster Auto Scaling
- Présentation de Karpenter, un Cluster Autoscaler de Kubernetes hautement performant et open source

Vidéos connexes :

- AWS re:Invent 2023 — AWS Graviton : le meilleur rapport qualité/prix pour vos charges de travail AWS
- AWS re:Invent 2023 — Nouvelles fonctionnalités d'IA EC2 générative d'Amazon dans Management Console AWS
- AWS re:Invent 2023 — Nouveautés d'Amazon EC2
- AWS re:Invent 2023 — Économies intelligentes : stratégies d'optimisation des coûts d'Amazon EC2

- AWS re:Invent 2021 — Au service d'EC2Amazon de nouvelle génération : étude approfondie du système Nitro
- AWS re:Invent 2019 — Amazon Foundations EC2

Exemples connexes :

- Exemples de groupes Amazon EC2 Auto Scaling
- EKSAtelier Amazon
- Faites évoluer vos EKS charges de travail Amazon en exécutant sur IPv6

PERF02-BP06 Utiliser des accélérateurs de calcul matériels optimisés

Utilisez des accélérateurs matériels pour exécuter certaines fonctions de manière plus efficace que les alternatives basées sur l'UC.

Anti-modèles courants :

- En ce qui concerne votre charge de travail, vous n'avez pas comparé une instance à usage général à une instance dédiée qui est capable de fournir de meilleures performances à moindre coût.
- Vous utilisez des accélérateurs de calcul matériels pour les tâches qui peuvent être plus efficaces en utilisant des alternatives basées sur l'UC.
- Vous ne surveillez pas l'utilisation du GPU.

Avantages liés au respect de cette bonne pratique : en utilisant des accélérateurs matériels, tels que des unités de traitement graphique (GPU) et une matrice de portes programmables sur site (FPGA), vous pouvez exécuter certaines fonctions de traitement de manière plus efficace.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les instances de calcul accéléré donnent accès à des accélérateurs de calcul matériels tels que les GPU et les FPGA. Ces accélérateurs matériels exécutent certaines fonctions comme le traitement graphique ou la correspondance de modèles de données plus efficacement que les alternatives basées sur l'UC. De nombreuses charges de travail accélérées, telles que le rendu, le transcodage et le machine learning, sont très variables en matière d'utilisation des ressources. Exécutez ce matériel

uniquement pendant le temps nécessaire et mettez-le hors service grâce à l'automatisation lorsque vous n'en avez plus besoin afin d'améliorer l'efficacité globale des performances.

Étapes d'implémentation

- Identifiez les instances de calcul accéléré qui peuvent répondre à vos besoins.
- Pour les charges de travail de machine learning, tirez parti d'un matériel conçu spécialement pour votre charge de travail, par exemple AWS, Trainum, AWS Inferentia et Amazon EC2 DL1. AWS Les instances Inferentia telles que les instances Inf2 offrent des performances/watt jusqu'à 50 % supérieures à celles des instances Amazon EC2 comparables.
- Collectez des métriques d'utilisation pour vos instances de calcul accéléré. Par exemple, vous pouvez utiliser l'agent CloudWatch pour collecter des métriques telles que `utilization_gpu` et `utilization_memory` pour vos GPU, comme indiqué dans Collecter les métriques des GPU NVIDIA avec Amazon CloudWatch.
- Optimisez le code, le fonctionnement du réseau et les paramètres des accélérateurs matériels pour veiller à ce que le matériel sous-jacent soit pleinement utilisé.
  - Optimiser les paramètres GPU
  - Surveillance et optimisation des GPU dans l'AMI Deep Learning
  - Optimisation des E/S pour le réglage des performances de GPU pour l'entraînement du deep learning dans l'IA Amazon SageMaker
- Utilisez les dernières bibliothèques performantes et les pilotes GPU.
- Utilisez l'automatisation pour libérer les instances GPU lorsqu'elles ne sont pas utilisées.

Ressources

Documents connexes :

- Fonctionnement d'Amazon Elastic Container Service
- Instances GPU
- Instances avec AWS Trainium
- Instances avec AWS Inferentia
- Passons à l'architecture ! Architecture avec des puces personnalisées et des accélérateurs

- Calcul accéléré
- Instances Amazon EC2 VT1

- Comment choisir le type d'instance EC2 approprié pour ma charge de travail ?
- Choix du meilleur accélérateur d'IA et de la meilleure compilation de modèles pour l'inférence de vision par ordinateur avec l'IA Amazon SageMaker

Vidéos connexes :

- AWSre:Invent 2021 - Comment sélectionner les instances Amazon Elastic Compute Cloud GPU pour le deep learning
- AWSre:INVENT 2022 - [NOUVEAU LANCEMENT !] Présentation des instances AWS Amazon EC2 Inf2 basées sur Inferentia2
- AWSre:Invent 2022 - Accélérez le deep learning et innovez plus rapidement avec Trainium AWS
- AWS re:Invent 2022 - Deep learning on AWS with NVIDIA: From training to deployment

Exemples connexes :

- IA Amazon SageMaker et NVIDIA GPU Cloud (NGC)
- Utilisation de l'IA SageMaker avec Trainium et Inferentia pour optimiser les charges de travail d'inférence et d'entraînement du deep learning
- Optimisation des modèles NLP avec les instances Amazon Elastic Compute Cloud Inf1 dans l'IA Amazon SageMaker

Gestion des données

Questions

- PERF 3. Comment stocker les données de votre charge de travail, comment les gérer et comment y accéder ?

PERF 3. Comment stocker les données de votre charge de travail, comment les gérer et comment y accéder ?

La solution optimale de gestion des données pour un système particulier varie en fonction du type de données (bloc, fichier ou objet), des modèles d'accès (aléatoire ou séquentiel), du débit requis, de la fréquence d'accès (en ligne, hors ligne, archivage), de la fréquence de mise à jour (WORM, dynamique), ainsi que des contraintes de disponibilité et de durabilité. Les charges de travail Well-

Architected utilisent des magasins de données sur mesure qui intègrent différentes fonctionnalités pour améliorer les performances.

Bonnes pratiques

- PERF03-BP01 Utilisez un magasin de données spécialement conçu pour répondre au mieux à vos besoins en matière d'accès aux données et de stockage
- PERF03-BP02 Évaluer les options de configuration disponibles pour le magasin de données
- PERF03-BP03 Collecter et enregistrer les indicateurs de performance du magasin de données
- PERF03-BP04 Mise en œuvre de stratégies pour améliorer les performances des requêtes dans un magasin de données
- PERF03-BP05 Implémenter des modèles d'accès aux données qui utilisent la mise en cache

PERF03-BP01 Utilisez un magasin de données spécialement conçu pour répondre au mieux à vos besoins en matière d'accès aux données et de stockage

Comprenez les caractéristiques des données (telles que la possibilité de partage, la taille, la taille du cache, les modèles d’accès, la latence, le débit et la persistance des données) afin de sélectionner les magasins de données dédiés (stockage ou base de données) adaptés à votre charge de travail.

Anti-modèles courants :

- Vous vous en tenez à un magasin de données, car l’équipe interne sait comment tirer parti de ce type de solution en particulier.
- Vous partez du principe que toutes les charges de travail ont des exigences similaires en termes de stockage de données et d’accès aux données.
- Vous n’avez pas implémenté de catalogue de données pour inventorier vos ressources de données.

Avantages liés au respect de cette bonne pratique : en comprenant l’importance des caractéristiques et des exigences des données, vous pouvez déterminer la technologie de stockage la plus efficace et la plus performante adaptée à vos besoins en matière de charge de travail.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Lors de la sélection et de la mise en œuvre du stockage des données, assurez-vous que les caractéristiques de requête, de dimensionnement et de stockage répondent aux exigences relatives aux données de charge de travail. AWS fournit de nombreuses technologies de stockage de données et de base de données, notamment le stockage par blocs, le stockage d'objets, le stockage en continu, les systèmes de fichiers, les bases de données relationnelles, les bases de données à valeur clé, les bases de données documentaires, en mémoire, les graphiques, les séries chronologiques et les bases de données de registre. Chaque solution de gestion de données propose des options et des configurations pour prendre en charge vos cas d’utilisation et vos modèles de données. En comprenant les caractéristiques et les exigences des données, vous pouvez vous affranchir de la technologie de stockage monolithique et des one-size-fits-all approches restrictives pour vous concentrer sur la gestion appropriée des données.

Étapes d’implémentation

- Procédez à l’inventaire des différents types de données qui existent dans votre charge de travail.
- Comprenez et documentez les caractéristiques et les exigences des données, notamment :
  - Type de données (non structurées, semi-structurées, relationnelles)
  - Volume et croissance des données
  - Durabilité des données : persistantes, éphémères, temporaires
  - ACIDexigences (atomicité, consistance, isolation, durabilité)
  - Modèles d’accès aux données (à lecture intensive ou à écriture intensive)
  - Latence
  - Débit
  - IOPS(opérations d'entrée/sortie par seconde)
  - Période de conservation des données
- Découvrez les différents magasins de données (services de stockage et de base de données) disponibles pour votre charge de travail AWS qui peuvent répondre aux caractéristiques de vos données, comme indiqué dansPERF01-BP01 Découvrez et comprenez les services et fonctionnalités cloud disponibles. Voici quelques exemples de technologies de stockage AWS et leurs principales caractéristiques :

Type	AWS Services	Principales caractéristiques
Stockage d'objets	Amazon S3	Capacité de mise à l'échelle illimitée, haute disponibilité et plusieurs options d'accessi bilité. Le transfert et l'accès à des objets à l'intérieur et à l'extérieur d'Amazon S3 peuvent utiliser un service, tel que Transfer Accelerat ion ou Access Points (points d'accès), pour répondre à votre localisation, à vos besoins en matière de sécurité et à vos modèles d'accès.
Archivage et stockage	Amazon S3 Glacier	Conçu pour l'archivage des données.
Stockage en streaming	Amazon Kinesis Amazon Managed Streaming pour Apache Kafka (Amazon) MSK	Ingestion et stockage efficaces des données de streaming.
Système de fichiers partagé	Amazon Elastic File System (AmazonEFS)	Système de fichiers montable auquel plusieurs types de solutions informatiques peuvent accéder.

Type	AWS Services	Principales caractéristiques
Système de fichiers partagé	Amazon FSx	Construit sur les dernières solutions AWS informati ques pour prendre en charge quatre systèmes de fichiers couramment utilisés : Open NetApp ONTAPZFS, Windows File Server et Lustre. FSxLa latence, le débit et le débit d'Amazon IOPS varient selon le système de fichiers et doivent être pris en compte lors de la sélection du système de fichiers adapté à vos besoins en matière de charge de travail.
Stockage en mode bloc	Boutique Amazon Elastic Block (AmazonEBS)	Service de stockage par blocs évolutif et performant conçu pour Amazon Elastic Compute Cloud (AmazonEC 2). Amazon EBS inclut le stockage SSD sauvegardé pour les charges de travail transactionnelles intensives et le stockage HDD sauvegard é pour les charges de travail IOPS gourmandes en débit.

Type	AWS Services	Principales caractéristiques
Base de données relationnelle	Amazon Aurora, Amazon RDS, Amazon Redshift.	Conçu pour prendre en charge les transactions ACID (atomicité, cohérence, isolation, durabilité) et pour maintenir l'intégrité référentielle et la forte cohérence des données. De nombreuses applications traditionnelles, de planification des ressources d'entreprise (ERP), de gestion de la relation client (CRM) et de commerce électronique utilisent des bases de données relationnelles pour stocker leurs données.
Base de données clé-valeur	Amazon DynamoDB	Optimisées pour les modèles d’accès courants, généralement pour stocker et récupérer de gros volumes de données. Les applications Web à trafic élevé, les systèmes d’e-commerce et les applications de jeu sont des cas d’utilisation typiques pour les bases de données de valeurs-clés.

Type	AWS Services	Principales caractéristiques
Base de données documentaire	Amazon DocumentDB	Conçu pour stocker des données semi-structurées sous JSON forme de documents similaires. Ces bases de données aident les développeurs à créer et mettre à jour rapidement des applications telles que la gestion de contenu, les catalogues et les profils utilisateur.
Base de données en mémoire	Amazon ElastiCache, Amazon MemoryDB pour Redis	Utilisées pour les applications qui nécessitent un accès en temps réel aux données, la latence la plus faible et le débit le plus élevé. Vous pouvez utiliser des bases de données en mémoire pour la mise en cache des applications, la gestion des sessions, les classements des jeux, le magasin de fonctionnalités ML à faible latence, le système de messagerie à microservices et un mécanisme de streaming à haut débit

Type	AWS Services	Principales caractéristiques
Base de données orientée graphe	Amazon Neptune	Destinées aux applications qui doivent parcourir et interroger des millions de relations entre des jeux de données graphiques hautement connectés avec une latence de millisecondes à grande échelle. De nombreuses entreprises utilisent des bases de données de graphiques pour la détection des fraudes, les réseaux sociaux et les moteurs de recommandation.
Base de données de séries temporelles	Amazon Timestream	Utilisées pour collecter, synthétiser et extraire efficacement des informations à partir de données qui changent au fil du temps. Les applications IoT et la DevOps télémétrie industrielle peuvent utiliser des bases de données de séries chronologiques.

Type	AWS Services	Principales caractéristiques
Larges colonnes	Amazon Keyspaces (pour Apache Cassandra)	Utilise des tables, des lignes et des colonnes, mais contrairement à une base de données relationnelle, les noms et le format des colonnes peuvent varier d'une ligne à l'autre dans la même table. Généralement, vous voyez un magasin de colonnes larges dans les applications industrielles à grande échelle pour la maintenance des équipements, la gestion des parcs et l'optimisation des itinéraires.
Registre	Base de données Amazon Quantum Ledger (AmazonQLDB)	Fournit une autorité centralisée et fiable pour conserver un enregistrement évolutif, immuable et vérifiable grâce au chiffrement des transactions pour chaque application. Il n'est pas rare de voir des bases de données de registre utilisées pour les systèmes d'enregistrement, la chaîne d'approvisionnement, les inscriptions et même les transactions bancaires.

- Si vous créez une plate-forme de données, tirez parti de l'architecture de données moderne AWS pour intégrer votre lac de données, votre entrepôt de données et vos magasins de données spécialement conçus.
- Les principales questions que vous devez vous poser lors du choix d'un magasin de données pour votre charge de travail sont les suivantes :

Question	Éléments à prendre en compte
Comment sont structurées les données ?	• Si les données ne sont pas structurées, envisagez un magasin d'objets tel qu'Amazon S3 ou une base de données sans SQL données sans base de données telle qu'Amazon DocumentDB • Pour les données clé-valeur, pensez à DynamoDB, Amazon (ElastiCache Redis) ou Amazon MemoryDB OSS
Quel niveau d’intégrité référentielle est requis ?	• En ce qui concerne les contraintes liées aux clés étrangères, les bases de données relationnelles telles qu'Amazon RDS et Aurora peuvent fournir ce niveau d'intégrité. • Généralement, dans un SQL modèle sans données, vous dénormaliseriez vos données en un seul document ou en un ensemble de documents à récupérer en une seule demande plutôt que de joindre plusieurs documents ou tableaux.
La conformité ACID (atomicité, consistance, isolation, durabilité) est-elle requise ?	• Si les ACID propriétés associées aux bases de données relationnelles sont requises, envisagez une base de données relationnelle telle qu'Amazon et RDS Aurora. • Si une forte cohérence est requise pour Aucune SQL base de données, vous pouvez utiliser des lectures fortement cohérentes avec DynamoDB.

Question	Éléments à prendre en compte
Comment les exigences de stockage vont-elles évoluer au fil du temps ? Comment cela affectera-t-il la capacité de mise à l'échelle ?	• Les bases de données sans serveur telles que DynamoDB et Amazon Quantum Ledger Database (QLDBAmazon) évolueront de manière dynamique. • Les bases de données relationnelles ont des limites supérieures sur le stockage alloué et doivent souvent être partitionnées horizontalement à l'aide de mécanismes tels que le partitionnement une fois qu'elles atteignent ces limites.
Quelle est la proportion de requêtes en lecture par rapport aux requêtes en écriture ? La mise en cache pourrait-elle améliorer les performances ?	• Les charges de travail gourmandes en lecture peuvent bénéficier d'une couche de mise en cache, comme ElastiCacheou DAXsi la base de données est DynamoDB. • Les lectures peuvent également être déchargées pour lire des répliques avec des bases de données relationnelles telles qu'Amazon. RDS
Le stockage et la modification (OLTP- Traitement des transactions en ligne) ou la récupération et le reporting (OLAP- Traitement analytique en ligne) ont-ils une priorité plus élevée ?	• Pour un traitement transactionnel en lecture telle quelle à haut débit, envisagez une base de données sans base de données SQL telle que DynamoDB. • Pour des modèles de lecture complexes et à haut débit (comme la jointure) cohérents, utilisez Amazon. RDS • Pour les requêtes analytiques, envisagez d'utiliser une base de données en colonnes telle qu'Amazon Redshift ou d'exporter les données vers Amazon S3 et d'effectuer des analyses à l'aide d'Athena ou d'Amazon. QuickSight

Question	Éléments à prendre en compte
Quel est le niveau de durabilité requis pour les données ?	• Aurora réplique automatiquement vos données sur trois zones de disponibilité au sein d'une région. Autrement dit, vos données sont très durables avec moins de risque de perte de données. • DynamoDB est automatiquement répliqué sur plusieurs zones de disponibilité, assurant ainsi la haute disponibilité et la durabilité des données. • Amazon S3 offre une durabilité de 99,999999999 %. De nombreux services de base de données, tels qu'Amazon RDS et DynamoDB, prennent en charge l'exporta tion de données vers Amazon S3 pour une conservation et un archivage à long terme.
Souhaitez-vous vous éloigner des moteurs de base de données commerciaux ou des coûts de licence ?	• Pensez aux moteurs open source tels que Postgre SQL et My on SQL Amazon ou RDS Aurora. • Tirez parti de AWS Database Migration Service et AWS Schema Conversion Tool pour passer des moteurs de bases de données commerciaux vers des moteurs open source.
Qu'attendez-vous de la base de données du point de vue opérationnel ? Le passage aux services gérés est-il une préoccupation majeure ?	• Tirer parti d'Amazon RDS au lieu d'AmazonEC2, et de DynamoDB ou d'Amazon DocumentDB au lieu d'héberge r vous-même une SQL base de données « No » peut réduire les frais d'exploitation.

Question	Éléments à prendre en compte
Comment accédez-vous actuellement à la base de données ? S'agit-il uniquement d'un accès aux applications ou existe-t-il des utilisateurs de Business Intelligence (BI) et d'autres off-the-shelf applications connectées ?	• Si vous dépendez d’outils externes, vous devrez peut-être maintenir la compatibilité avec les bases de données qu’ils prennent en charge. Amazon RDS est entièrement compatible avec les différentes versions de moteurs qu'il prend en charge, notamment Microsoft SQL Server, OracleSQL, My et PostgreSQL.

- Réalisez des tests et procédez au benchmarking dans un environnement hors production afin d’identifier le magasin de données qui peut répondre à vos exigences en termes de charge de travail.

Ressources

Documents connexes :

- Types de EBS volumes Amazon
- EC2Stockage Amazon
- Amazon EFS : Amazon EFS Performance
- Amazon FSx pour Lustre Performance
- Performances du serveur de fichiers Amazon FSx pour Windows
- Amazon Glacier S3 : documentation Amazon Glacier S3
- Amazon S3 : directives en matière de débit de demandes et de performances
- Stockage dans le cloud avec AWS
- Caractéristiques d'Amazon EBS I/O
- Bases de données cloud avec AWS
- AWS Mise en cache de bases de données
- DynamoDB Accelerator
- Bonnes pratiques Amazon Aurora
- Performances d’Amazon Redshift
- Amazon Athena top 10 de conseils en matière de performance

- Bonnes pratiques Amazon Redshift Spectrum
- Bonnes pratiques Amazon DynamoDB
- Choisissez entre Amazon EC2 et Amazon RDS
- Bonnes pratiques pour la mise en œuvre d'Amazon ElastiCache

Vidéos connexes :

- AWS re:Invent 2023 : Améliorez l'efficacité d'Amazon Elastic Block Store et soyez plus rentable
- AWS re:Invent 2023 : Optimisation du prix et des performances du stockage avec Amazon Simple Storage Service
- AWS re:Invent 2023 : Création et optimisation d'un lac de données sur Amazon Simple Storage Service
- AWS re:Invent 2022 : Création d'architectures de données modernes sur AWS
- AWS re:Invent 2022 : Création d'architectures de maillage de données sur AWS
- AWS re:Invent 2023 : présentation approfondie d'Amazon Aurora et de ses innovations
- AWS re:Invent 2023 : Modélisation avancée des données avec Amazon DynamoDB
- AWS re:Invent 2022 : Modernisez les applications avec des bases de données spécialement conçues
- Amazon DynamoDB deep dive: Advanced design patterns

Exemples connexes :

- AWS Atelier sur les bases de données spécialement conçues
- Bases de données pour développeurs
- AWS Journée d'immersion dans l'architecture de données moderne
- Créez un maillage de données sur AWS
- Exemples Amazon S3
- Optimisation du modèle de données à l’aide du partage de données Amazon Redshift
- Migrations des bases de données
- MS SQL Server - AWS Database Migration Service (AWS DMS) Démo de réplication
- Atelier pratique sur la modernisation des bases de données
- Échantillons Amazon Neptune

PERF03-BP02 Évaluer les options de configuration disponibles pour le magasin de données

Comprenez et évaluez les différentes fonctionnalités et options de configuration disponibles pour vos magasins de données afin d'optimiser l'espace de stockage et les performances de votre charge de travail.

Anti-modèles courants :

- Vous n'utilisez qu'un seul type de stockage, tel qu'AmazonEBS, pour toutes les charges de travail.
- Vous utilisez le provisionné IOPS pour toutes les charges de travail sans effectuer de tests réels sur tous les niveaux de stockage.
- Vous ne connaissez pas les options de configuration de la solution de gestion de données que vous avez choisie.
- Vous vous concentrez uniquement sur l'augmentation de la taille de l'instance sans examiner les autres options de configuration disponibles.
- Vous ne testez pas les caractéristiques de mise à l'échelle de votre magasin de données.

Avantages liés au respect de cette bonne pratique : en explorant et en expérimentant les configurations de magasin de données, vous pourriez réduire le coût de l'infrastructure, améliorer les performances et réduire l'effort requis pour maintenir vos charges de travail.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Une charge de travail peut comporter un ou plusieurs magasins de données utilisés en fonction des exigences de stockage des données et d'accès aux données. Pour optimiser l'efficacité et le coût de vos performances, vous devez évaluer les modèles d'accès aux données afin de déterminer les configurations de magasin de données appropriées. Pendant que vous explorez les options de magasin de données, tenez compte de divers aspects tels que les options de stockage, la mémoire, le calcul, le réplica en lecture, les exigences de cohérence, le regroupement de connexions et les options de mise en cache. Testez ces différentes options de configuration pour améliorer les métriques d'efficacité des performances.

Étapes d'implémentation

- Comprenez les configurations actuelles (comme le type d'instance, la taille de stockage ou la version du moteur de base de données) de votre magasin de données.

- Consultez AWS la documentation et les meilleures pratiques pour découvrir les options de configuration recommandées qui peuvent contribuer à améliorer les performances de votre magasin de données. Les principales options de magasin de données à prendre en compte sont les suivantes :

Option de configuration	Exemples
Déchargement des lectures (comme les réplicas en lecture et la mise en cache)	• Pour les tables DynamoDB, vous pouvez décharger les lectures à l'aide de la mise en cache. DAX • Vous pouvez créer un cluster Amazon ElastiCache (RedisOSS) et configurer votre application pour qu'elle lise d'abord dans le cache, puis revenir à la base de données si l'élément demandé n'est pas présent. • Bases de données relationnelles telles qu'Amazon RDS et Aurora, et mises en service Aucune SQL base de données telle que Neptune et Amazon DocumentDB ne prend toutes en charge l'ajout de répliques de lecture pour décharger les parties de lecture de la charge de travail. • Les bases de données sans serveur comme DynamoDB se mettent à l’échelle automatiq uement. Assurez-vous de disposer de suffisamment d'unités de capacité de lecture (RCU) pour gérer la charge de travail.

Option de configuration	Exemples
Mise à l’échelle des écritures (comme le partitionnement des clés de partition ou l’introduction d’une file d’attente)	• Pour les bases de données relationnelles, vous pouvez augmenter la taille de l'instanc e pour faire face à une charge de travail accrue ou augmenter le provisionnement IOPs pour augmenter le débit du stockage sous-jacent. • Vous pouvez également ajouter une file d’attente devant votre base de données plutôt que d’écrire directement dans la base de données. Ce modèle vous permet de dissocier l’ingestion de la base de données et de contrôler le débit afin que la base de données ne soit pas submergée. • Regrouper vos demandes d’écriture plutôt que de créer de nombreuses transactions de courte durée contribue à améliorer le débit dans les bases de données relationn elles à volume d’écriture élevé. • Les bases de données sans serveur telles que DynamoDB peuvent augmenter le débit d'écriture automatiquement ou en ajustant les unités de capacité d'écriture allouées WCU () en fonction du mode de capacité. • Vous pouvez toujours rencontrer des problèmes avec les partitions à chaud lorsque vous atteignez les limites de débit pour une clé de partition donnée. Pour pallier ce problème, choisissez une clé de partition distribuée plus uniformément ou partitionnez en écriture la clé de partition.

Option de configuration	Exemples
Politiques de gestion du cycle de vie de vos jeux de données	• Vous pouvez utiliser Amazon S3 Lifecycle afin de gérer vos objets au cours de leur cycle de vie. Si vos schémas d'accès sont inconnus, changeants ou imprévisibles, vous pouvez utiliser Amazon S3 Intelligent-Tiering, qui surveille les schémas d'accès et déplace automatiquement les objets qui n'ont pas été accédés vers des niveaux d'accès moins coûteux. Vous pouvez tirer parti des métriques Amazon S3 Storage Lens pour identifier les opportunités d'optimisation et les lacunes dans la gestion du cycle de vie. • Amazon EFS Lifecycle Management gère automatiquement le stockage de fichiers pour vos systèmes de fichiers.
Gestion et regroupement des connexions	• Amazon RDS Proxy peut être utilisé avec Amazon RDS et Aurora pour gérer les connexions à la base de données. • Les bases de données sans serveur comme DynamoDB n'ont pas de connexions associées, mais tenez compte de la capacité allouée et des stratégies de mise à l'échelle automatique pour faire face aux pics de charge.

- Réalisez des tests et procédez au benchmarking dans un environnement hors production afin d'identifier l'option de configuration qui répond à vos exigences en termes de charge de travail.
- Après avoir réalisé vos tests, planifiez votre migration et validez vos métriques de performance.
- Utilisez AWS des outils de surveillance (comme Amazon CloudWatch) et d'optimisation (comme Amazon S3 Storage Lens) pour optimiser en permanence votre magasin de données en utilisant des modèles d'utilisation réels.

Ressources

Documents connexes :

- Stockage cloud avec AWS
- Types de EBS volumes Amazon
- EC2Stockage Amazon
- Amazon EFS : Amazon EFS Performance
- Amazon FSx pour Lustre Performance
- Performances du serveur de fichiers Amazon FSx pour Windows
- Amazon Glacier S3 : documentation Amazon Glacier S3
- Amazon S3 : directives en matière de débit de demandes et de performances
- Caractéristiques d'Amazon EBS I/O
- Bases de données cloud avec AWS
- AWS Mise en cache de bases de données
- DynamoDB Accelerator
- Bonnes pratiques Amazon Aurora
- Performances d’Amazon Redshift
- Amazon Athena top 10 de conseils en matière de performance
- Bonnes pratiques Amazon Redshift Spectrum
- Bonnes pratiques Amazon DynamoDB

Vidéos connexes :

- AWS re:Invent 2023: Improve Amazon Elastic Block Store efficiency and be more cost-efficient
- AWS re:Invent 2023: Optimize storage price and performance with Amazon Simple Storage Service
- AWS re:Invent 2023: Building and optimizing a data lake on Amazon Simple Storage Service
- AWS re:Invent 2023 : Nouveautés en matière de stockage de fichiers AWS
- AWS re:Invent 2023: Dive deep into Amazon DynamoDB

Exemples connexes :

- AWS Atelier sur les bases de données spécialement conçues
- Bases de données pour développeurs
- AWS Journée d'immersion dans l'architecture de données moderne
- Amazon EBS Autoscale
- Exemples Amazon S3
- Exemple Amazon DynamoDB
- AWS Exemples de migration de base de données
- Atelier sur la modernisation des bases de données
- Utilisation des paramètres de votre base de données Amazon RDS pour Postgress

PERF03-BP03 Collecter et enregistrer les indicateurs de performance du magasin de données

Suivez et archivez les métriques de performance pertinentes pour votre magasin de données afin de comprendre comment fonctionnent vos solutions de gestion des données. Ces métriques peuvent vous aider à optimiser votre magasin de données, à vérifier que les exigences de votre charge de travail sont satisfaites et à fournir une vue d'ensemble claire sur le fonctionnement de la charge de travail.

Anti-modèles courants :

- Vous utilisez uniquement la recherche manuelle des fichiers journaux pour les métriques.
- Vous publiez uniquement des métriques sur les outils internes utilisés par votre équipe et vous n'avez pas une visibilité complète de votre charge de travail.
- Vous n'utilisez que les métriques par défaut enregistrées par le logiciel de surveillance que vous avez sélectionné.
- Vous n'examinez les métriques qu'en cas de problème.
- Vous ne surveillez que les métriques au niveau du système et vous ne capturez pas les métriques d'accès aux données ou d'utilisation des données.

Avantages liés au respect de cette bonne pratique : la définition de points de référence pour les performances vous permet de mieux comprendre le comportement normal et les exigences des charges de travail. Les modèles anormaux peuvent être identifiés et débogués plus rapidement, ce qui améliore les performances et la fiabilité du magasin de données.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d’implémentation

L’enregistrement de plusieurs métriques de performance sur une période donnée est nécessaire pour la surveillance des performances de vos magasins de données. Cette surveillance vous permet non seulement de détecter les anomalies, mais aussi d’évaluer les performances par rapport aux métriques métier afin de vérifier que vous répondez aux besoins de votre charge de travail.

Ces métriques doivent inclure à la fois le système sous-jacent qui prend en charge le magasin de données et les métriques de la base de données. Les indicateurs système sous-jacents peuvent inclure CPU l'utilisation, la mémoire, le stockage sur disque disponible, les E/S sur disque, le taux d'accès au cache et les mesures entrantes et sortantes du réseau, tandis que les indicateurs du magasin de données peuvent inclure les transactions par seconde, les requêtes les plus fréquentes, les taux de requêtes moyens, les temps de réponse, l'utilisation de l'index, les blocages de table, les délais d'attente des requêtes et le nombre de connexions ouvertes. Ces données sont essentielles pour comprendre comment fonctionne la charge de travail et comment la solution de gestion des données est utilisée. Utilisez ces métriques dans le cadre d’une approche fondée sur les données pour ajuster et optimiser les ressources de votre charge de travail.

Utilisez des outils, des bibliothèques et des systèmes qui enregistrent des mesures de performances liées aux performances de la base de données.

Étapes d’implémentation

- Identifiez les métriques de performances clés que votre magasin de données doit suivre.
  - Métriques et dimensions d’Amazon S3
  - Mesures de surveillance pour une RDS instance Amazon
  - Surveillance de la charge de base de données avec Performance Insights sur Amazon RDS
  - Vue d’ensemble de la surveillance améliorée
  - Métriques et dimensions DynamoDB
  - Surveillance de l’accélérateur DynamoDB
  - Surveillance d'Amazon MemoryDB avec Amazon CloudWatch
  - Quelles métriques dois-je surveiller ?
  - Surveillance des performances de cluster Amazon Redshift
  - Métriques et dimensions Timestream
  - CloudWatch Métriques Amazon pour Amazon Aurora
  - Journalisation et surveillance dans Amazon Keyspaces (pour Apache Cassandra)

  - Surveillance des ressources Amazon Neptune
- Utilisez une solution de journalisation et de surveillance approuvée pour collecter ces métriques. Amazon CloudWatch peut collecter des métriques sur l'ensemble des ressources de votre architecture. Vous pouvez également récupérer et publier des métriques personnalisées pour faire apparaître des métriques d'entreprise ou des métriques dérivées. Utilisez CloudWatch ou utilisez des solutions tierces pour définir des alarmes indiquant lorsque les seuils sont dépassés.
- Vérifiez si la surveillance du magasin de données peut bénéficier d'une solution de machine learning qui détecte les anomalies de performance.
  - Amazon DevOps Guru for Amazon RDS fournit de la visibilité sur les problèmes de performance et recommande des mesures correctives.
- Configurez la conservation des données dans votre solution de surveillance et de journalisation en fonction de vos objectifs sécuritaires et opérationnels.
  - Conservation des données par défaut pour les CloudWatch métriques
  - Conservation des données par défaut pour les CloudWatch journaux

Ressources

Documents connexes :

- Mise en cache de bases de données AWS
- Amazon Athena top 10 de conseils en matière de performance
- Bonnes pratiques Amazon Aurora
- DynamoDB Accelerator
- Bonnes pratiques Amazon DynamoDB
- Bonnes pratiques Amazon Redshift Spectrum
- Performances d'Amazon Redshift
- Bases de données cloud avec AWS
- Amazon RDS Performance Insights

Vidéos connexes :

- AWS re:Invent 2022 - Surveillance des performances avec Amazon et RDS Aurora, avec Autodesk
- Surveillance et optimisation des performances des bases de données avec Amazon DevOps Guru pour Amazon RDS

- AWS re:Invent 2023 - Nouveautés en matière de stockage de fichiers AWS
- AWS re:Invent 2023 - Découvrez en détail Amazon DynamoDB
- AWS re:Invent 2023 - Création et optimisation d'un lac de données sur Amazon S3
- AWS re:Invent 2023 - Nouveautés en matière de stockage de fichiers AWS
- AWS re:Invent 2023 - Découvrez en détail Amazon DynamoDB
- Meilleures pratiques pour surveiller les charges de travail Redis sur Amazon ElastiCache

Exemples connexes :

- Cadre de collecte de métriques pour l’ingestion des jeux de données AWS
- Atelier RDS de surveillance Amazon
- AWS Atelier sur les bases de données spécialement conçues

PERF03-BP04 Mise en œuvre de stratégies pour améliorer les performances des requêtes dans un magasin de données

Mettez en œuvre des stratégies pour optimiser les données et améliorer les requêtes sur les données afin de renforcer la capacité de mise à l’échelle et l’efficacité des performances pour votre charge de travail.

Anti-modèles courants :

- Vous ne partitionnez pas les données dans votre magasin de données.
- Vous ne stockez les données que dans un seul format de fichier dans votre magasin de données.
- Vous n’utilisez pas d’index dans votre magasin de données.

Avantages liés au respect de cette bonne pratique : en optimisant les performances des données et des requêtes, vous augmentez leur efficacité, vous réduisez les coûts et vous améliorez l’expérience utilisateur.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

L’optimisation des données et des requêtes sont des aspects essentiels de l’efficacité des performances d’un magasin de données, car ils ont un impact sur les performances et la réactivité de l’ensemble de la charge de travail dans le cloud. Les données non optimisées peuvent augmenter

l'utilisation des ressources et les goulots d'étranglement, ce qui réduit l'efficacité globale d'un magasin de données.

L'optimisation des données inclut plusieurs techniques pour garantir un stockage de données et un accès aux données efficaces. Cela permet également d'améliorer les performances des requêtes dans un magasin de données. Les principales stratégies incluent le partitionnement des données, la compression des données et la dénormalisation des données, qui permettent d'optimiser les données à la fois pour le stockage et l'accès.

Étapes d'implémentation

- Comprenez et analysez les requêtes essentielles sur les données effectuées dans votre magasin de données.
- Identifiez les requêtes lentes dans votre magasin de données et utilisez des plans de requêtes pour comprendre leur état actuel.
  - Analyse du plan de requêtes dans Amazon Redshift
  - Utilisation d'EXPLAIN et EXPLAIN ANALYZE sur Athena
- Mettez en œuvre des stratégies pour améliorer les performances des requêtes. Les stratégies clés incluent :
  - L'utilisation d'un format de fichier en colonnes (comme Parquet ou ORC).
  - La compression des données dans le magasin de données pour réduire l'espace de stockage et les opérations d'E/S.
  - Le partitionnement des données pour diviser les données en parties plus petites et réduire le temps d'analyse des données.
    - Partitionnement de données dans Athena
    - Partitions et distribution des données
  - L'indexation des données sur les colonnes communes de la requête.
  - Utilisez des vues matérialisées pour les requêtes fréquentes.
    - Compréhension des vues matérialisées
    - Création de vues matérialisées dans Amazon Redshift
  - Choisissez l'opération de jointure appropriée pour la requête. Lorsque vous joignez deux tables, spécifiez la table la plus grande sur le côté gauche de la jointure et la plus petite sur le côté droit de la jointure.
  - La solution de mise en cache distribué pour améliorer la latence et réduire le nombre d'opérations d'E/S dans la base de données.

  - Maintenance régulière, telle que l'aspiration, la réindexation et les statistiques d'exécution.
- Expérimentez et testez les stratégies dans un environnement hors production.

Ressources

Documents connexes :

- Bonnes pratiques Amazon Aurora
- Performances d'Amazon Redshift
- Amazon Athena top 10 de conseils en matière de performance
- Mise en cache de bases de données AWS
- Bonnes pratiques de mise en œuvre d'Amazon ElastiCache
- Partitionnement de données dans Athena

Vidéos connexes :

- AWS re:Invent 2023 - AWS storage cost-optimization best practices
- AWS re:Invent 2022 - Performance monitoring with Amazon RDS and Aurora, featuring Autodesk
- Optimize Amazon Athena Queries with New Query Analysis Tools

Exemples connexes :

- AWS Atelier sur les bases de données sur mesure

PERF03-BP05 Implémenter des modèles d'accès aux données qui utilisent la mise en cache

Mettez en œuvre des modèles d'accès qui peuvent tirer parti de la mise en cache des données pour une récupération rapide des données fréquemment consultées.

Anti-modèles courants :

- Vous mettez en cache des données qui changent fréquemment.
- Vous utilisez les données mises en cache comme si elles étaient stockées de manière durable et toujours disponibles.
- Vous ne tenez pas compte de la cohérence de vos données mises en cache.

- Vous ne surveillez pas l'efficacité de la mise en cache.

Avantages liés au respect de cette bonne pratique : le stockage des données dans un cache contribue à améliorer la latence et le débit de lecture, l'expérience utilisateur et l'efficacité globale, tout en réduisant les coûts.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Un cache est un composant logiciel ou matériel destiné à stocker des données afin que les requêtes futures portant sur les mêmes données puissent être traitées plus rapidement ou plus efficacement. Les données stockées dans un cache peuvent être reconstruites en cas de perte en répétant un calcul antérieur ou en les récupérant dans un autre magasin de données.

La mise en cache des données peut être l'une des stratégies les plus efficaces pour améliorer les performances globales de votre application et réduire la charge qui pèse sur vos sources de données principales sous-jacentes. Les données peuvent être mises en cache à plusieurs niveaux de l'application, par exemple au sein de l'application en effectuant des appels à distance ou mise en cache côté client ou en utilisant un service secondaire rapide pour stocker les données mise en cache à distance.

Mise en cache côté client

Grâce à la mise en cache côté client, chaque client (une application ou un service qui interroge l'entrepôt de données dorsales) peut stocker les résultats de ses requêtes uniques localement pendant une durée spécifiée. Cela permet de réduire le nombre de requêtes adressées à un entrepôt de données sur le réseau en vérifiant d'abord le cache du client local. En l'absence de résultats, l'application peut alors interroger l'entrepôt de données et stocker ces résultats localement. Ce modèle permet à chaque client de stocker les données dans l'emplacement le plus proche possible (le client lui-même), ce qui se traduit par la latence la plus faible possible. Les clients peuvent également continuer à répondre à certaines requêtes lorsque l'entrepôt de données dorsales n'est pas disponible, ce qui augmente la disponibilité de l'ensemble du système.

L'un des inconvénients de cette approche est que lorsque plusieurs clients sont impliqués, ils peuvent stocker les mêmes données mises en cache localement. Cela entraîne à la fois une double utilisation du stockage et une incohérence des données entre ces clients. Un client peut mettre en cache les résultats d'une requête et, une minute plus tard, un autre client peut exécuter la même requête et obtenir un résultat différent.

Mise en cache à distance

Pour résoudre le problème de duplication de données entre clients, un service externe rapide ou un cache distant, peut être utilisé pour stocker les données demandées. Au lieu de vérifier un magasin de données local, chaque client vérifie le cache distant avant d'interroger l'entrepôt de données dorsales. Cette stratégie permet d'obtenir des réponses plus cohérentes entre les clients, d'améliorer l'efficacité des données stockées et d'augmenter le volume de données mises en cache, car l'espace de stockage évolue indépendamment des clients.

L'inconvénient d'un cache distant est que l'ensemble du système peut connaître une latence plus élevée, car un saut de réseau à réseau supplémentaire est nécessaire pour vérifier le cache distant. La mise en cache côté client peut être utilisée parallèlement à la mise en cache à distance pour une mise en cache à plusieurs niveaux afin d'améliorer la latence.

Étapes d'implémentation

- Identifiez les bases de données APIs et les services réseau susceptibles de bénéficier de la mise en cache. Les services dont la charge de travail de lecture est importante, dont le read-to-write ratio est élevé ou dont la mise à l'échelle est coûteuse sont candidats à la mise en cache.
  - Mise en cache de bases de données
  - Activation de la API mise en cache pour améliorer la réactivité
- Identifiez le type de stratégie de mise en cache le mieux adapté à votre modèle d'accès.
  - Stratégies de mise en cache
  - Solutions de mise en cache AWS
- Suivez les bonnes pratiques de mise en cache pour votre banque de données.
- Configurez une stratégie d'invalidation du cache, telle que a time-to-live (TTL), pour toutes les données afin d'équilibrer la fraîcheur des données et de réduire la pression sur la banque de données principale.
- Activez des fonctionnalités telles que les nouvelles tentatives de connexion automatiques, le backoff exponentiel, les délais d'attente côté client et le regroupement des connexions dans le client, le cas échéant, car elles peuvent améliorer les performances et la fiabilité.
  - Bonnes pratiques : clients Redis et Amazon ElastiCache (RedisOSS)
- Surveillez le taux d'accès au cache en visant un objectif de 80 % ou plus. Des valeurs inférieures peuvent indiquer une taille de cache insuffisante ou un modèle d'accès qui ne bénéficie pas de la mise en cache.
  - Quelles métriques dois-je surveiller ?

  - Bonnes pratiques pour surveiller les charges de travail Redis sur Amazon ElastiCache
  - Surveillance des meilleures pratiques avec Amazon ElastiCache (RedisOSS) à l'aide d'Amazon CloudWatch
- Mettre en œuvre la réplication des données pour transférer les lectures vers plusieurs instances et améliorer les performances et la disponibilité de lecture des données.

Ressources

Documents connexes :

- Utilisation de l'objectif Amazon ElastiCache Well-Architected
- Surveillance des meilleures pratiques avec Amazon ElastiCache (RedisOSS) à l'aide d'Amazon CloudWatch
- Quelles métriques dois-je surveiller ?
- ElastiCache Livre blanc sur les performances à grande échelle avec Amazon
- Défis et stratégies en matière de mise en cache

Vidéos connexes :

- Parcours de ElastiCache formation Amazon
- Concevez pour réussir grâce aux ElastiCache meilleures pratiques d'Amazon
- AWS re:Invent 2020 - Concevez pour réussir grâce aux meilleures pratiques d'Amazon ElastiCache
- AWS re:Invent 2023 - [LAUNCH] Présentation d'Amazon Serverless ElastiCache
- AWS re:Invent 2022 - 5 excellentes façons de réinventer votre couche de données avec Redis
- AWS re:Invent 2021 - Présentation approfondie d'Amazon ElastiCache (Redis) OSS

Exemples connexes :

- Améliorer les performances SQL de ma base de données avec Amazon ElastiCache (RedisOSS)

Réseau et diffusion de contenu

Questions

PERF 4. Comment sélectionner et configurer les ressources de mise en réseau de votre charge de travail ?

La solution de mise en réseau optimale pour une charge de travail varie en fonction de la latence, des exigences de débit, de l'instabilité et de la bande passante. Le choix des options d'emplacement est tributaire des contraintes physiques telles que les ressources pour utilisateur ou sur site. Ces contraintes peuvent être compensées avec les emplacements périphériques ou le placement des ressources.

Bonnes pratiques

- PERF04-BP01 Comprendre l'impact du réseau sur les performances
- PERF04-BP02 Évaluer les fonctionnalités réseau disponibles
- PERF04-BP03 Choisissez une connectivité dédiée adaptée à votre charge VPN de travail
- PERF04-BP04 Utiliser l'équilibrage de charge pour répartir le trafic entre plusieurs ressources
- PERF04-BP05 Choisissez les protocoles réseau pour améliorer les performances
- PERF04-BP06 Choisissez l'emplacement de votre charge de travail en fonction des exigences du réseau
- PERF04-BP07 Optimiser la configuration du réseau en fonction des métriques

PERF04-BP01 Comprendre l'impact du réseau sur les performances

Analysez et comprenez l'impact des décisions liées au réseau sur votre charge de travail afin de fournir des performances efficaces et une meilleure expérience utilisateur.

Anti-modèles courants :

- Tout le trafic passe par vos centres de données existants.
- Vous acheminez l'ensemble du trafic via des pare-feux centralisés au lieu d'utiliser des outils de sécurité réseau natifs cloud.
- Vous configurez AWS Direct Connect des connexions sans connaître les exigences d'utilisation réelles.
- Vous ne tenez pas compte des caractéristiques de la charge de travail et de la surcharge de chiffrement lors de la définition de vos solutions de mise en réseau.

- Vous utilisez des concepts et des stratégies sur site pour les solutions de mise en réseau dans le cloud.

Avantages liés au respect de cette bonne pratique : comprendre comment la mise en réseau affecte les performances de la charge de travail vous aide à identifier les goulots d'étranglement potentiels, à améliorer l'expérience utilisateur, à accroître la fiabilité et à réduire la maintenance opérationnelle à mesure que la charge de travail évolue.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Le réseau est responsable de la connectivité entre les composants d'application, les services cloud, les réseaux périphériques et les données sur site et, par conséquent, il peut avoir un impact majeur sur les performances de la charge de travail. Outre les performances de la charge de travail, l'expérience utilisateur peut également être affectée par la latence du réseau, la bande passante, les protocoles, l'emplacement, la congestion du réseau, l'instabilité, le débit et les règles de routage.

Veillez à avoir une liste documentée des exigences de mise en réseau de la charge de travail, y compris la latence, la taille des paquets, les règles de routage, les protocoles et les modèles de trafic pris en charge. Passez en revue les solutions de mise en réseau disponibles et identifiez le service qui répond aux caractéristiques de mise en réseau de votre charge de travail. Les réseaux basés sur le cloud peuvent être rapidement recréés. L'évolution de votre architecture réseau au fil du temps est donc nécessaire pour améliorer l'efficacité des performances.

Étapes d'implémentation :

- Définissez et documentez les exigences de performance réseau, y compris les métriques telles que la latence du réseau, la bande passante, les protocoles, les emplacements, les modèles de trafic (pics et fréquence), le débit, le chiffrement, l'inspection et les règles de routage.
- Découvrez les principaux services AWS réseau tels que VPCsElastic Load Balancing (ELB) et Amazon Route 53. AWS Direct Connect
- Capturez les principales caractéristiques réseau suivantes :

Caractéristiques	Outils et métriques
Caractéristiques de mise en réseau fondamentales	• VPCJournaux de flux • AWS Transit Gateway Journaux de flux

Caractéristiques	Outils et métriques
	• AWS Transit Gateway métriques • AWS PrivateLink métriques
Caractéristiques de mise en réseau des applications	• Elastic Fabric Adapter (EFA) • AWS App Mesh métriques • Métriques Amazon API Gateway
Caractéristiques de mise en réseau à la périphérie	• CloudFront Métriques Amazon • Métriques Amazon Route 53 • AWS Global Accelerator métriques
Caractéristiques de mise en réseau hybride	• AWS Direct Connect métriques • AWS Site-to-Site VPN métriques • AWS Client VPN métriques • AWS Cloud WANmétriques
Caractéristiques de mise en réseau de la sécurité	• AWS ShieldAWS WAF, et AWS Network Firewall métriques
Caractéristiques de traçage	• AWS X-Ray • VPCAnalyseur de Reachability • Analyseur d’accès réseau • Amazon Inspector • Amazon CloudWatch RUM

- Définition de points de référence et test des performances du réseau :
  - Comparez le débit du réseau, car certains facteurs peuvent affecter les performances EC2 du réseau Amazon lorsque les instances se trouvent dans les mêmes VPC instances. Mesurez la bande passante réseau entre les instances Amazon EC2 Linux d'une même instanceVPC.
  - Effectuez des tests de charge pour expérimenter des solutions et des options de mise en réseau.

Ressources

Documents connexes :

- Application Load Balancer
- EC2Mise en réseau améliorée sous Linux
- EC2Mise en réseau améliorée sous Windows
- EC2Groupes de placement
- Activation de la mise en réseau améliorée avec l'adaptateur réseau Elastic (ENA) sur les instances Linux
- Network Load Balancer
- Produits de mise en réseau avec AWS
- Passerelle de transit
- Transition vers le routage basé sur la latence dans Amazon Route 53
- VPCPoints de terminaison

Vidéos connexes :

- AWS re:Invent 2023 - AWS mise en réseau des fondations
- AWS re:Invent 2023 - Que peut apporter le réseau à votre application ?
- AWS re:Invent 2023 - VPC Designs avancés et nouvelles fonctionnalités
- AWS re:Invent 2023 - Guide du développeur sur les réseaux cloud
- AWS re:Invent 2019 - Connectivité AWS et architectures réseau hybrides AWS
- AWS re:Invent 2019 - Optimisation des performances réseau pour les instances Amazon EC2
- AWS Summit Online - Améliorez les performances du réseau mondial pour les applications
- AWS re:Invent 2020 - Meilleures pratiques et astuces de mise en réseau avec le Well-Architected Framework
- AWS re:Invent 2020 : meilleures pratiques de AWS mise en réseau pour les migrations à grande échelle

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- AWS Ateliers de réseautage
- Atelier pratique sur le pare-feu réseau
- Observation et diagnostic de votre réseau sur AWS

- Recherche et résolution des erreurs de configuration réseau sur AWS

PERF04-BP02 Évaluer les fonctionnalités réseau disponibles

Évaluez les fonctions de mise en réseau dans le cloud qui peuvent améliorer les performances. Mesurez l'impact de ces fonctions au moyen de tests, de métriques et de l'analyse. Par exemple, tirez parti des fonctionnalités au niveau du réseau qui sont disponibles pour réduire la latence, la distance réseau ou l'instabilité.

Anti-modèles courants :

- Vous restez au sein d'une même région, car c'est là que votre siège social se trouve physiquement.
- Vous utilisez des pare-feux plutôt que des groupes de sécurité pour filtrer le trafic.
- Vous faites une pause TLS pour inspecter le trafic plutôt que de vous fier aux groupes de sécurité, aux politiques relatives aux terminaux et à d'autres fonctionnalités natives du cloud.
- Vous utilisez uniquement la segmentation basée sur un sous-réseau au lieu des groupes de sécurité.

Avantages liés au respect de cette bonne pratique : l'évaluation de toutes les options et fonctionnalités de service peut augmenter les performances de vos charges de travail, baisser le coût d'infrastructure, réduire les efforts nécessaires à la maintenance de vos charges de travail et améliorer votre posture générale en matière de sécurité. Vous pouvez utiliser le AWS backbone mondial pour offrir une expérience réseau optimale à vos clients.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

AWS propose des services tels AWS Global Accelerator CloudFrontqu'Amazon qui peuvent contribuer à améliorer les performances du réseau, tandis que la plupart AWS des services proposent des fonctionnalités (telles que la fonctionnalité Amazon S3 Transfer Acceleration) permettant d'optimiser le trafic réseau.

Examinez les options de configuration liées au réseau disponibles et leur impact potentiel sur votre charge de travail. L'optimisation des performances dépend de la compréhension de la manière dont ces options interagissent avec votre architecture et de l'impact qu'elles auront à la fois sur les performances mesurées et sur l'expérience utilisateur.

Étapes d’implémentation

- Créer une liste des composants de la charge de travail.
  - Envisagez AWS Cloud WANde l'utiliser pour créer, gérer et surveiller le réseau de votre organisation lors de la création d'un réseau mondial unifié.
  - Surveillez vos réseaux mondiaux et principaux à l'aide des métriques Amazon CloudWatch Logs. Tirez parti d'Amazon CloudWatch RUM, qui fournit des informations permettant d'identifier, de comprendre et d'améliorer l'expérience numérique des utilisateurs.
  - Visualisez la latence réseau globale entre les zones de disponibilité Régions AWS et au sein de chaque zone de disponibilité, AWS Network Managerafin de mieux comprendre le lien entre les performances de votre application et les performances du AWS réseau sous-jacent.
  - Utilisez un outil ou un service de base de données de gestion de configuration (CMDB) existant, par exemple AWS Configpour créer un inventaire de votre charge de travail et de sa configuration.
- Identifier et documenter le test comparatif pour vos métriques de performances s’il s’agit d’une charge de travail existante, en vous concentrant sur les goulots d’étranglement et les zones à améliorer. Les métriques de mise en réseau liées aux performances diffèrent par charge de travail en fonction des exigences métier et des caractéristiques de charge de travail. Pour commencer, il pourrait être important d’examiner ces métriques pour votre charge de travail : bande passante, latence, perte de paquets, instabilité et retransmissions.
- S’il s’agit d’une nouvelle charge de travail, effectuez des tests de charge pour identifier les goulots d’étranglement liés aux performances.
- Concernant l’identification des goulots d’étranglement au niveau des performances, examiner les options de configuration pour les solutions afin d’identifier les opportunités d’amélioration des performances. Découvrez les principales options et fonctionnalités de mise en réseau suivantes :

Opportunité d’amélioration	Solution
Chemin ou itinéraires réseau	Utilisez l’analyseur d’accès réseau pour identifier les chemins ou les itinéraires.
Protocoles réseau	Consultez PERF04-BP05 Choisissez les protocoles réseau pour améliorer les performances.

Opportunité d’amélioration	Solution
Topologie du réseau	Évaluez vos compromis opérationnels et de performance entre le VPCpeering et AWS Transit Gatewaylors de la connexion de plusieurs comptes. AWS Transit Gateway simplifie la façon dont vous interconnectez tous vos VPCs réseaux, qui peuvent s'étendre sur Comptes AWS des milliers de réseaux locaux. Partagez votre compte AWS Transit Gateway entre plusieurs comptes en utilisant AWS Resource Access Manager. Consultez PERF04-BP03 Choisissez une connectivité dédiée adaptée à votre charge VPN de travail.

Opportunité d’amélioration	Solution
Services de réseau	AWS Global Acceleratorest un service réseau qui améliore les performances du trafic de vos utilisateurs jusqu'à 60 % en utilisant l'infrast ructure réseau AWS mondiale. Amazon CloudFront peut améliorer les performances de votre charge de travail, de diffusion de contenu et de latence à l'échelle mondiale. Utilisez Lambda @edge pour exécuter des fonctions qui personnalisent le contenu au plus près CloudFront des utilisateurs, réduisent le temps de latence et améliorent les performances. Amazon Route 53 propose des options de routage basées sur la latence, de routage de géolocalisation, de routage de géolocali sation et de routage basé sur IP pour vous aider à améliorer les performances de votre charge de travail auprès d’un public mondial. Identifiez l’option de routage qui optimiserait les performances de votre charge de travail en examinant le trafic de votre charge de travail et la localisation des utilisateurs lorsque votre charge de travail est distribuée dans le monde entier.

Opportunité d’amélioration	Solution
Fonctionnalités des ressources de stockage	Amazon S3 Transfer Acceleration est une fonctionnalité qui permet aux utilisateurs externes de bénéficier des optimisations du réseau CloudFront pour télécharger des données vers Amazon S3. Cela améliore le transfert d’importants volumes de données à partir d’emplacements distants qui n’ont pas de connectivité dédiée au AWS Cloud. Les points d’accès multi-régions Amazon S3 répliquent le contenu vers plusieurs régions et simplifient la charge de travail en fournissa nt un point d’accès. Lorsqu’un point d’accès multi-région est utilisé, vous pouvez demander ou écrire des données à Amazon S3 tandis que le service identifie le compartiment à la latence la plus faible.

Opportunité d’amélioration	Solution
Fonctionnalités des ressources informatiques	Les interfaces réseau élastiques (ENA) utilisées par EC2 les instances Amazon, les conteneurs et les fonctions Lambda sont limitées par flux. Passez en revue vos groupes de placement pour optimiser le débit EC2 de votre réseau. Pour éviter un goulot d’étrangl ement par flux, créez votre application pour qu’elle utilise plusieurs flux. Pour surveille r et obtenir une visibilité sur vos métriques réseau liées au calcul, utilisez CloudWatch Metrics et ethtool. La `ethtool` commande est incluse dans le ENA pilote et expose des métriques supplémentaires liées au réseau qui peuvent être publiées sous forme de métrique personnalisée sur. CloudWatch Les Amazon Elastic Network Adapters (ENA) fournissent une optimisation supplémen taire en fournissant un meilleur débit à vos instances au sein d'un groupe de placement de clusters. Elastic Fabric Adapter (EFA) est une interface réseau pour les EC2 instances Amazon qui vous permet d'exécuter des charges de travail nécessitant des niveaux élevés de communica tions entre nœuds à grande échelle. AWS Les instances EBS optimisées pour Amazon utilisent une pile de configuration optimisée et fournissent une capacité dédiée supplémen taire pour augmenter les EBS E/S Amazon.

Ressources

Documents connexes :

- Application Load Balancer
- EC2Mise en réseau améliorée sous Linux
- EC2Mise en réseau améliorée sous Windows
- EC2Groupes de placement
- Activation de la mise en réseau améliorée avec l'adaptateur réseau Elastic (ENA) sur les instances Linux
- Network Load Balancer
- Produits de mise en réseau avec AWS
- Transition vers le routage basé sur la latence dans Amazon Route 53
- VPCPoints de terminaison
- Journaux de flux VPC

Vidéos connexes :

- AWS re:Invent 2023 — Prêts pour la suite ? Designing networks for growth and flexibility
- AWS re:Invent 2023 — VPC Designs avancés et nouvelles fonctionnalités
- AWS re:Invent 2023 — Guide du développeur sur les réseaux cloud
- AWS re:Invent 2022 — Approfondissez l'infrastructure réseau AWS
- AWS re:Invent 2019 — Connectivité AWS et architectures réseau hybrides AWS
- AWS re:Invent 2018 — Optimisation des performances réseau pour les instances Amazon EC2
- AWS Global Accelerator

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- AWS Ateliers de réseautage
- Observation et diagnostic de votre réseau
- Détecter et corriger les erreurs de configuration du réseau sur AWS

PERF04-BP03 Choisissez une connectivité dédiée adaptée à votre charge VPN de travail

Lorsque la connectivité hybride est requise pour connecter des ressources sur site et dans le cloud, allouez une bande passante adéquate pour répondre à vos exigences de performance. Estimez les exigences en matière de bande passante et de latence pour votre charge de travail hybride. Ces chiffres détermineront vos exigences en matière de dimensionnement.

Anti-modèles courants :

- Vous évaluez uniquement les VPN solutions en fonction des exigences de chiffrement de votre réseau.
- Vous n'évaluez pas les options de sauvegarde ni de connectivité redondante.
- Vous n'identifiez pas toutes les exigences de la charge de travail (chiffrement, protocole, bande passante et trafic requis).

Avantages liés au respect de cette bonne pratique : la sélection et la configuration de solutions de connectivité appropriées renforcent la fiabilité de votre charge de travail et optimisent les performances. En identifiant les exigences en matière de charge de travail, en planifiant à l'avance et en évaluant les solutions hybrides, vous pouvez minimiser les modifications coûteuses du réseau physique et les frais d'exploitation tout en augmentant votre time-to-value

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Développez une architecture réseau hybride en fonction de vos besoins en bande passante. AWS Direct Connect vous permet de connecter votre réseau sur site en privé à AWS. Cette solution convient lorsque vous avez besoin d'une bande passante élevée et d'une faible latence tout en conservant des performances constantes. Une VPN connexion établit une connexion sécurisée via Internet. Elle sert uniquement lorsque seule une connexion temporaire est requise, lorsque le coût est un facteur, ou en cas d'urgence en attendant qu'une connectivité réseau physique résiliente soit établie lors de l'utilisation d' AWS Direct Connect.

Si vos besoins en bande passante sont élevés, vous pouvez envisager plusieurs VPN services AWS Direct Connect ou services. Le trafic peut être équilibré entre les services, mais nous ne recommandons pas l'équilibrage de charge entre AWS Direct Connect et en VPN raison des différences de latence et de bande passante.

Étapes d’implémentation

- Évaluez les besoins en bande passante et en latence de vos applications existantes.
  - Pour les charges de travail existantes qui sont transférées AWS, exploitez les données de vos systèmes de surveillance réseau internes.
  - Pour les nouvelles charges de travail ou pour les charges de travail existantes pour lesquelles vous ne disposez pas de données de suivi, contactez les propriétaires du produit pour obtenir des métriques de performance adéquates et offrir une bonne expérience utilisateur.
- Sélectionnez une connexion dédiée ou VPN comme option de connectivité. En fonction de toutes les exigences en matière de charge de travail (chiffrement, bande passante et besoins en trafic), vous pouvez choisir AWS Direct Connect ou AWS VPN(ou les deux). Le schéma suivant peut vous aider à choisir le type de connexion approprié.
  - AWS Direct Connect fournit une connectivité dédiée à l’environnement AWS , de 50 Mbit/s à 100 Gbit/s, en utilisant des connexions dédiées ou des connexions hébergées. Cela vous permet de gérer et de contrôler la latence et de profiter d’une bande passante provisionnée. Ainsi, vos charges de travail peuvent se connecter efficacement à d’autres environnements. En faisant appel à des AWS Direct Connect partenaires, vous pouvez bénéficier d'une end-to-end connectivité à partir de plusieurs environnements, fournissant ainsi un réseau étendu aux performances constantes. AWS permet de dimensionner la bande passante de connexion directe en utilisant 100 Gbit/s natifs, un groupe d'agrégation de liens (LAG) ou un multipath à BGP coût égal (). ECMP
  - AWS Site-to-Site VPNIl fournit un VPN service géré prenant en charge la sécurité du protocole Internet (IPsec). Lorsqu'une VPN connexion est créée, chaque VPN connexion inclut deux tunnels pour une haute disponibilité.
- Suivez AWS la documentation pour choisir l'option de connectivité appropriée :
  - Si vous décidez de l'utiliser AWS Direct Connect, sélectionnez la bande passante adaptée à votre connectivité.
  - Si vous utilisez un réseau AWS Site-to-Site VPN sur plusieurs sites pour vous connecter à un Région AWS, utilisez une Site-to-SiteVPNconnexion accélérée afin d'améliorer les performances du réseau.
  - Si la conception de votre réseau consiste en IPSec VPN une connexion via une connexion AWS Direct Connect, pensez à utiliser une adresse IP privée VPN pour améliorer la sécurité et réaliser une segmentation. AWS Site-to-Site L'adresse IP privée VPN est déployée au-dessus de l'interface virtuelle de transit (VIF).

  - AWS Direct Connect SiteLinkpermet de créer des connexions redondantes et à faible latence entre vos centres de données du monde entier en envoyant les données sur le chemin le plus rapide entre les AWS Direct Connect sites, en les contournant. Régions AWS
- Validez votre configuration de connectivité avant le déploiement en production. Effectuez des tests de sécurité et de performance pour vous assurer qu'elle répond à vos exigences en matière de bande passante, de fiabilité, de latence et de conformité.
- Surveillez régulièrement les performances et l'utilisation de votre connectivité et optimisez-les si nécessaire.

Organigramme des performances déterministes

Ressources

Documents connexes :

- Produits de mise en réseau avec AWS
- AWS Transit Gateway
- VPCPoints de terminaison
- Création d'une infrastructure VPC AWS multiréseau évolutive et sécurisée
- Cliente VPN

Vidéos connexes :

- AWS re:Invent 2023 — Création d'une connectivité réseau hybride avec AWS
- AWS re:Invent 2023 — Connectivité à distance sécurisée pour AWS
- AWS re:Invent 2022 — Optimisation des performances avec Amazon CloudFront
- AWS re:Invent 2019 — Connectivité AWS et architectures réseau hybrides AWS
- AWS re:Invent 2020 — Connect AWS Transit Gateway

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- AWS Ateliers de réseautage

PERF04-BP04 Utiliser l'équilibrage de charge pour répartir le trafic entre plusieurs ressources

Répartissez le trafic sur plusieurs ressources ou services pour permettre à votre charge de travail de tirer parti de l’élasticité fournie par le cloud. Vous pouvez également utiliser l’équilibrage de charge afin de décharger la terminaison du chiffrement en vue d’améliorer les performances, d’assurer la fiabilité et de gérer et acheminer efficacement le trafic.

Anti-modèles courants :

- Vous ne tenez pas compte des exigences de votre charge de travail lorsque vous choisissez le type d’équilibreur de charge.
- Vous ne tirez pas parti des fonctionnalités de l’équilibreur de charge pour optimiser les performances.
- La charge de travail est exposée directement à Internet sans équilibreur de charge.
- Vous acheminez tout le trafic Internet via des équilibreurs de charge existants.
- Vous utilisez un équilibrage de TCP charge générique et vous faites en sorte que chaque nœud de calcul gère SSL le chiffrement.

Avantages liés au respect de cette bonne pratique : un équilibreur de charge gère la charge variable du trafic de votre application dans une seule zone de disponibilité ou entre plusieurs zones de disponibilité et permet une haute disponibilité, une mise à l’échelle automatique et une meilleure utilisation de votre charge de travail.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Les équilibreurs de charge constituent le point d’entrée de votre charge de travail, à partir duquel ils distribuent le trafic vers vos cibles principales, telles que les instances de calcul ou les conteneurs, afin d’améliorer l’utilisation.

Le choix du bon type d’équilibreur de charge est la première étape de l’optimisation de votre architecture. Commencez par répertorier les caractéristiques de votre charge de travail, telles que le protocole (comme TCPHTTP,TLS, ou WebSockets), le type de cible (comme les instances, les conteneurs ou les applications sans serveur), les exigences de l'application (telles que les connexions de longue durée, l'authentification des utilisateurs ou la rigidité) et le placement (par exemple, région, zone locale, avant-poste ou isolation zonale).

AWS fournit plusieurs modèles permettant à vos applications d'utiliser l'équilibrage de charge. Application Load Balancer convient parfaitement à l'équilibrage de la charge HTTP et du HTTPS trafic et fournit un routage avancé des demandes destiné à la fourniture d'architectures d'applications modernes, notamment des microservices et des conteneurs.

Network Load Balancer convient parfaitement à l'équilibrage de charge du TCP trafic lorsque des performances extrêmes sont requises. Il est capable de traiter des millions de requêtes par seconde tout en maintenant de très faibles latences. Il est optimisé pour gérer les tendances soudaines et instables du trafic.

Elastic Load Balancing intègre la gestion et le SSL TLS déchiffrement des certificats, ce qui vous permet de gérer de manière centralisée les SSL paramètres de l'équilibreur de charge et de décharger les tâches CPU intensives de votre charge de travail.

Après avoir choisi le bon équilibreur de charge, vous pouvez commencer à tirer parti de ses fonctionnalités pour réduire les efforts que votre système dorsal doit fournir pour servir le trafic.

Par exemple, en utilisant à la fois Application Load Balancer (ALB) et Network Load Balancer NLB (), vous pouvez SSL effectuer un déchargement par TLS chiffrement, ce qui vous permet d'éviter que vos cibles ne se lancent CPU dans une poignée de main TLS intensive et d'améliorer la gestion des certificats.

Lorsque vous configurez SSL ou TLS déchargez dans votre équilibreur de charge, celui-ci devient responsable du chiffrement du trafic en provenance et à destination des clients, tout en distribuant le trafic non chiffré à vos backends, en libérant les ressources de votre backend et en améliorant le temps de réponse des clients.

Application Load Balancer peut également desservir HTTP /2 trafic sans avoir à le prendre en charge sur vos cibles. Cette simple décision peut améliorer le temps de réponse de votre application, car HTTP /2 utilise TCP les connexions de manière plus efficace.

Les exigences de latence de votre charge de travail doivent être prises en compte lors de la définition de l’architecture. Par exemple, si vous avez une application sensible à la latence, vous pouvez décider d’utiliser Network Load Balancer, qui offre des latences extrêmement faibles. Vous pouvez également décider de rapprocher votre charge de travail de vos clients en tirant parti d’Application Load Balancer dans AWS Local Zones or même AWS Outposts.

L’équilibrage de charge entre zones est un autre élément à prendre en compte pour les charges de travail sensibles à la latence. Avec l’équilibrage de charge inter-zone, chaque nœud d’équilibreur de charge distribue le trafic sur les cibles enregistrées dans toutes les zones de disponibilité activées.

Intégrez l'autoscaling à votre équilibreur de charge. L'un des aspects essentiels d'un système performant est le dimensionnement adéquat de vos ressources dorsales. Pour ce faire, vous pouvez tirer parti des intégrations d'équilibreurs de charge pour les ressources cibles du système dorsal. Grâce à l'intégration de l'équilibreur de charge avec les groupes Auto Scaling, les cibles seront ajoutées à l'équilibreur de charge ou retirées de l'équilibreur de charge selon les besoins en fonction du trafic entrant. Les équilibreurs de charge peuvent également s'intégrer à Amazon ECS et Amazon EKS pour les charges de travail conteneurisées.

- Amazon ECS - Équilibrage de charge des services
- Équilibrage de charge des applications sur Amazon EKS
- Équilibrage de charge réseau sur Amazon EKS

Étapes d'implémentation

- Définissez vos exigences en matière d'équilibrage de charge, notamment en termes de volume de trafic, de disponibilité et de capacité de mise à l'échelle des applications.
- Choisissez le type d'équilibreur de charge adapté à votre application.
  - Utilisez Application Load Balancer pour les charges de travailHTTP/HTTPS.
  - Utilisez Network Load Balancer pour les charges autres que les HTTP charges de travail qui s'exécutent sur ou. TCP UDP
  - Utilisez une combinaison des deux (ALBcomme cible NLB) si vous souhaitez tirer parti des fonctionnalités des deux produits. Par exemple, vous pouvez le faire si vous souhaitez utiliser la statique IPs de NLB avec le routage basé sur les HTTP en-têtes depuisALB, ou si vous souhaitez exposer votre HTTP charge de travail à un AWS PrivateLink.
  - Pour une comparaison complète des équilibreurs de charge, consultez la comparaison des ELB produits.
- UtilisezSSL/TLSoffloading si possible.
  - ConfigurezHTTPS/TLSlisteners avec Application Load Balancer et Network Load Balancer intégrés à. AWS Certificate Manager
  - Notez que certaines charges de travail peuvent nécessiter un end-to-end chiffrement pour des raisons de conformité. Dans ce cas, il est nécessaire de permettre le chiffrement au niveau des cibles.
  - Pour connaître les meilleures pratiques en matière de sécurité, voir SEC09-BP02 Appliquer le chiffrement en transit.

- Sélectionnez le bon algorithme de routage (uniquementALB).
  - L'algorithme de routage peut faire une réelle différence dans la manière d'utiliser vos cibles dorsales et donc dans leur impact sur les performances. Par exemple, ALB propose deux options pour les algorithmes de routage :
  - Demandes en suspens les moins nombreuses : cette option permet d'obtenir une meilleure répartition de la charge sur vos cibles dorsales dans les cas où les requêtes de votre application varient en complexité ou vos cibles varient en capacité de traitement.
  - Tour de rôle : utilisez cette méthode lorsque les requêtes et les cibles sont similaires, ou si vous devez distribuer les requêtes de manière égale entre les cibles.
- Envisagez l'option inter-zone ou l'isolement par zone.
  - Désactivez l'option désactivée (utilisez l'isolement par zone) pour améliorer la latence et les domaines de panne par zone. Il est désactivé par défaut dans NLB et dans, ALBvous pouvez le désactiver par groupe cible.
  - Activez l'option inter-zone pour une disponibilité et une flexibilité accrues. Par défaut, l'interzone est activée pour ALB et NLBvous pouvez l'activer pour chaque groupe cible.
- Activez HTTP Keep-Alives pour vos HTTP charges de travail (uniquement). ALB Grâce à cette fonctionnalité, l'équilibreur de charge peut réutiliser les connexions du backend jusqu'à l'expiration du délai de conservation, améliorant ainsi votre temps de HTTP demande et de réponse et réduisant également l'utilisation des ressources sur vos cibles de backend. Pour plus de détails sur la façon de procéder pour Apache et Nginx, consultez Quels sont les paramètres optimaux pour utiliser Apache ou en NGINX tant que serveur principal pour ? ELB
- Activez la surveillance pour votre équilibreur de charge.
  - Activez les journaux d'accès pour votre Application Load Balancer et Network Load Balancer.
  - Les principaux domaines à prendre en compte ALB sont `request_processing_timerequest_processing_time`, et`response_processing_time`.
  - Les principaux domaines à prendre en compte NLB sont `connection_time` et`tls_handshake_time`.
  - Soyez prêt à interroger les journaux lorsque vous en aurez besoin. Vous pouvez utiliser Amazon Athena pour interroger à la fois les ALBjournaux et NLB les journaux.
  - Créez des alarmes pour les indicateurs liés aux performances, tels que `TargetResponseTime`pour ALB.

Ressources

Documents connexes :

- ELBcomparaison de produits
- AWS Infrastructure mondiale
- Amélioration des performances et réduction des coûts grâce à l'affinité des zones de disponibilité
- Procédure détaillée d'analyse des journaux avec Amazon Athena
- Interrogation des journaux de l'Application Load Balancer
- Surveillance de vos Application Load Balancers
- Surveillance de votre Network Load Balancer
- Utiliser Elastic Load Balancing pour répartir le trafic sur les instances dans votre groupe Auto Scaling

Vidéos connexes :

- AWS re:INVENT 2023 : Qu'est-ce que le réseau peut apporter à votre application ?
- AWS Re:inForce 20 : Comment utiliser Elastic Load Balancing pour améliorer votre niveau de sécurité à grande échelle
- AWS re:Invent 2018 : Elastic Load Balancing : analyse approfondie et meilleures pratiques
- AWS re:Invent 2021 - Comment choisir le bon équilibreur de charge pour vos charges de travail AWS
- AWS re:Invent 2019 : Tirez le meilleur parti d'Elastic Load Balancing pour différentes charges de travail

Exemples connexes :

- Équilibreur de charge de passerelle
- CDKet AWS CloudFormation des exemples pour l'analyse des journaux avec Amazon Athena

PERF04-BP05 Choisissez les protocoles réseau pour améliorer les performances

Prenez des décisions concernant les protocoles de communication entre les systèmes et les réseaux en fonction de l'impact sur les performances de la charge de travail.

Il existe une relation entre la latence et la bande passante pour atteindre le débit. Si votre transfert de fichiers utilise le protocole de contrôle de transmission (TCP), des latences plus élevées réduiront probablement le débit global. Il existe des approches pour résoudre ce problème en TCP ajustant et en optimisant les protocoles de transfert, mais l'une des solutions consiste à utiliser le protocole User Datagram (UDP).

Anti-modèles courants :

- Vous l'utilisez TCP pour toutes les charges de travail, quelles que soient les exigences de performance.

Avantages liés au respect de cette bonne pratique : vérifiez que vous utilisez un protocole approprié pour la communication entre les utilisateurs et les composants de la charge de travail, afin d'améliorer l'expérience globale des utilisateurs de vos applications. Par exemple, le mode sans connexion UDP permet une vitesse élevée, mais il n'offre pas de retransmission ni une fiabilité élevée. TCPest un protocole complet, mais il nécessite une charge plus importante pour le traitement des paquets.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Si vous avez la possibilité de choisir différents protocoles pour votre application et que vous possédez l'expertise nécessaire dans ce domaine, optimisez votre application et l'expérience de l'utilisateur final en utilisant un autre protocole. Notez que cette approche présente des difficultés importantes et ne doit être tentée que si vous avez d'abord optimisé votre application à d'autres égards.

Pour améliorer les performances de votre charge de travail, il est essentiel de comprendre les exigences en matière de latence et de débit, puis de choisir des protocoles réseau qui optimisent les performances.

Quand envisager d'utiliser TCP

TCPfournit des données fiables et peut être utilisé pour la communication entre les composants de la charge de travail lorsque la fiabilité et la garantie de livraison des données sont importantes. De nombreuses applications Web s'appuient sur des protocoles TCP basés, tels que HTTP etHTTPS, pour ouvrir des TCP sockets permettant la communication entre les composants de l'application. Le transfert de données de courrier électronique et de fichiers est une application courante qui est également utiliséeTCP, car il s'agit d'un mécanisme de transfert simple et fiable entre les composants

de l'application. L'utilisation de TLS with TCP peut alourdir la communication, ce qui peut entraîner une augmentation de la latence et une réduction du débit, mais elle présente l'avantage de la sécurité. La surcharge provient principalement de la charge supplémentaire du processus de liaison, qui peut prendre plusieurs allers-retours pour se terminer. Une fois la liaison établie, la charge de chiffrement et de déchiffrement des données devient relativement faible.

Quand envisager d'utiliser UDP

UDPest un connection-less-oriented protocole et convient donc aux applications nécessitant une transmission rapide et efficace, telles que les données de journalisation, de surveillance et de VoIP. Pensez également à les utiliser UDP si vous disposez de composants de charge de travail qui répondent à de petites requêtes provenant d'un grand nombre de clients afin de garantir des performances optimales de la charge de travail. Datagram Transport Layer Security (DTLS) est l'UDPéquivalent de Transport Layer Security (TLS). Lors de l'utilisation DTLS avecUDP, la surcharge provient du chiffrement et du déchiffrement des données, car le processus de prise de contact est simplifié. DTLSajoute également une petite surcharge aux UDP paquets, car il inclut des champs supplémentaires pour indiquer les paramètres de sécurité et détecter les altérations.

Quand envisager d'utiliser SRD

Le datagramme fiable évolutif (SRD) est un protocole de transport réseau optimisé pour les charges de travail à haut débit en raison de sa capacité à équilibrer la charge du trafic sur plusieurs chemins et à récupérer rapidement en cas de perte de paquets ou de défaillance de liaison. SRDest donc mieux utilisé pour les charges de travail de calcul haute performance (HPC) qui nécessitent une communication à haut débit et à faible latence entre les nœuds de calcul. Il peut s’agir de tâches de traitement parallèle telles que la simulation, la modélisation et l’analyse de données qui impliquent un transfert important de données entre les nœuds.

Étapes d’implémentation

- Utilisez les services AWS Global Accelerator et AWS Transfer Family pour améliorer le débit de vos applications de transfert de fichiers en ligne. Le AWS Global Accelerator service vous aide à réduire le temps de latence entre vos appareils clients et votre charge de travail AWS. Vous pouvez utiliser des AWS Transfer Family protocoles TCP basés tels que Secure Shell File Transfer Protocol (SFTP) et File Transfer Protocol over SSL (FTPS) pour dimensionner et gérer en toute sécurité vos transferts de fichiers vers les services AWS de stockage.
- Utilisez la latence du réseau pour déterminer si la communication entre les composants de la charge de travail TCP est appropriée. Si la latence du réseau entre votre application cliente et votre serveur est élevée, la prise de TCP contact à trois peut prendre un certain temps, ce qui a un

impact sur la réactivité de votre application. Des mesures telles que le délai jusqu'au premier octet (TTFB) et le temps d'aller-retour (RTT) peuvent être utilisées pour mesurer la latence du réseau. Si votre charge de travail fournit du contenu dynamique aux utilisateurs, pensez à utiliser Amazon CloudFront, qui établit une connexion permanente à chaque origine pour le contenu dynamique afin de supprimer le temps de configuration de la connexion qui ralentirait autrement chaque demande du client.

- L'utilisation TLS avec TCP ou UDP peut entraîner une augmentation de la latence et une réduction du débit pour votre charge de travail en raison de l'impact du chiffrement et du déchiffrement. Pour de telles charges de travail, pensez à SSL TLS /offloading sur Elastic Load Balancing afin d'améliorer les performances de la charge de travail en permettant à l'équilibreur de charge de gérer SSL les processus de TLS chiffrement et de déchiffrement au lieu de laisser les instances principales s'en charger. Cela peut contribuer à réduire l'CPUutilisation des instances principales, ce qui peut améliorer les performances et augmenter la capacité.
- Utilisez le Network Load Balancer (NLB) pour déployer des services qui s'appuient sur le UDP protocole, tels que l'authentification et l'autorisation, la journalisation, l'DNSIoT et le streaming multimédia, afin d'améliorer les performances et la fiabilité de votre charge de travail. Le UDP trafic NLB entrant est réparti sur plusieurs cibles, ce qui vous permet d'adapter votre charge de travail horizontalement, d'augmenter la capacité et de réduire les frais généraux d'une seule cible.
- Pour vos charges de travail informatiques à hautes performances (HPC), pensez à utiliser la fonctionnalité Elastic Network Adapter (ENA) Express qui utilise le SRD protocole pour améliorer les performances du réseau en fournissant une bande passante à flux unique plus élevée (25 Gbit/s) et une latence de queue plus faible (99,9 centile) pour le trafic réseau entre les instances. EC2
- Utilisez l'Application Load Balancer (ALB) pour acheminer et équilibrer la charge de votre trafic g RPC (Remote Procedure Calls) entre les composants de la charge de travail ou entre les RPC clients et les services g. g RPC utilise le protocole de transport TCP basé sur HTTP /2 et offre des avantages en termes de performances tels qu'un encombrement réseau réduit, une compression, une sérialisation binaire efficace, la prise en charge de nombreuses langues et un streaming bidirectionnel.

Ressources

Documents connexes :

- Comment acheminer le UDP trafic vers Kubernetes
- Application Load Balancer
- EC2Mise en réseau améliorée sous Linux

- EC2Mise en réseau améliorée sous Windows
- EC2Groupes de placement
- Activation de la mise en réseau améliorée avec l'adaptateur réseau Elastic (ENA) sur les instances Linux
- Network Load Balancer
- Produits de mise en réseau avec AWS
- Transition vers le routage basé sur la latence dans Amazon Route 53
- VPCPoints de terminaison

Vidéos connexes :

- AWS re:Invent 2022 — Augmenter les performances du réseau sur les instances Amazon Elastic Compute Cloud de nouvelle génération
- AWS re:Invent 2022 — Les bases de la mise en réseau des applications

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- Ateliers sur la mise en réseau AWS

PERF04-BP06 Choisissez l'emplacement de votre charge de travail en fonction des exigences du réseau

Évaluez les options de placement des ressources afin de réduire la latence du réseau et d’améliorer le débit, offrant ainsi une expérience utilisateur optimale en réduisant les temps de chargement des pages et de transfert des données.

Anti-modèles courants :

- Vous regroupez toutes les ressources de charge de travail dans un seul emplacement géographique.
- Vous avez choisi la région la plus proche de votre emplacement, pas celle de l’utilisateur final de la charge de travail.

Avantages liés au respect de cette bonne pratique : l'expérience utilisateur est fortement affectée par le temps de latence entre l'utilisateur et votre application. En utilisant un réseau mondial AWS privé Régions AWS et approprié, vous pouvez réduire le temps de latence et offrir une meilleure expérience aux utilisateurs distants.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les ressources, telles que EC2 les instances Amazon, sont placées dans des zones de disponibilité Régions AWSinternes AWS Outposts, des zones AWS locales ou AWS Wavelengthdes zones. Le choix de cet emplacement influence la latence et le débit du réseau à partir d'un emplacement donné de l'utilisateur. Les services périphériques tels qu'Amazon CloudFront AWS Global Acceleratorpeuvent également être utilisés pour améliorer les performances du réseau soit en mettant en cache le contenu sur des sites périphériques, soit en fournissant aux utilisateurs un chemin optimal vers la charge de travail via le réseau AWS mondial.

Amazon EC2 propose des groupes de placement pour la mise en réseau. Un groupe de placement est un regroupement logique d'instances permettant de réduire la latence. L'utilisation de groupes de placement dotés de types d'instances compatibles et d'un adaptateur réseau élastique (ENA) permet aux charges de travail de participer à un réseau de 25 Gbit/s à faible latence et à instabilité réduite. Les groupes de placement sont recommandés pour les charges de travail nécessitant une latence réseau faible, un débit réseau élevé ou les deux.

Les services sensibles à la latence sont fournis sur des sites périphériques via un réseau AWS mondial, tel qu'Amazon. CloudFront Ces emplacements périphériques fournissent généralement des services tels que le réseau de diffusion de contenu (CDN) et le système de noms de domaine (DNS). En disposant de ces services à la périphérie, les charges de travail peuvent répondre avec une faible latence aux demandes de contenu ou de DNS résolution. Ces services fournissent également des services géographiques tels que le ciblage géographique du contenu (qui fournit des contenus différents en fonction de l'emplacement des utilisateurs finaux) ou le routage en fonction de la latence pour diriger les utilisateurs finaux vers la région plus proche (latence minimum).

Utilisez des services en périphérie pour réduire la latence et permettre la mise en cache de contenu. Configurez correctement le contrôle du cache pour les deux DNS etHTTP/HTTPSafin de tirer le meilleur parti de ces approches.

Étapes d'implémentation

- Capturez des informations sur le trafic IP entrant et sortant des interfaces réseau.

  - Enregistrement du trafic IP à l'aide de VPC Flow Logs
  - Comment l'adresse IP du client est-elle préservée dans AWS Global Accelerator
- Analysez les modèles d’accès au réseau dans votre charge de travail afin d’identifier comment les utilisateurs utilisent votre application.
  - Utilisez des outils de surveillance, tels qu'Amazon CloudWatch AWS CloudTrail, pour recueillir des données sur les activités du réseau.
  - Analysez les données pour identifier le modèle d’accès au réseau.
- Choisissez les régions pour le déploiement de votre charge de travail en fonction des éléments clés suivants :
  - Lieu de stockage de vos données : pour les applications utilisant de grandes quantités de données (telles que le big data et le machine learning). Le code de l’application doit s’exécuter aussi près que possible des données.
  - Lieu de stockage de vos données : pour les applications orientées utilisateur, choisissez une région (ou des régions) proche des utilisateurs de votre charge de travail.
  - Autres contraintes : tenez compte des contraintes telles que le coût et la conformité, comme expliqué dans la section Éléments à prendre en compte lors de la sélection d’une région pour vos charges de travail.
- Utilisez des zones locales AWS pour exécuter des charges de travail telles que le rendu vidéo. Les zones locales vous permettent de profiter des avantages liés à la présence de ressources de calcul et de stockage plus proches des utilisateurs finaux.
- Utilisez AWS Outposts pour les charges de travail qui doivent rester sur site et dont vous souhaitez qu’elles fonctionnent de manière transparente avec le reste de vos charges de travail dans AWS.
- Les applications telles que le streaming vidéo en direct haute résolution, le son haute fidélité et la réalité augmentée ou virtuelle (AR/VR) nécessitent ultra-low-latency des appareils 5G. Pour de telles applications, considérez AWS Wavelength. AWS Wavelength intègre des services de AWS calcul et de stockage dans les réseaux 5G, fournissant une infrastructure informatique de pointe mobile pour le développement, le déploiement et la mise à l'échelle d' ultra-low-latencyapplications.
- Utilisez des solutions de mise en cache locale ou proposées par AWS pour les ressuorces fréquemment utilisées afin d’améliorer les performances, de réduire les déplacements de données et de diminuer l’impact environnemental.

Service	Utilisation
Amazon CloudFront	Utilisez-le pour mettre en cache du contenu statique tel que des images, des scripts et des vidéos, ainsi que du contenu dynamique tel que API des réponses ou des applications Web.
Amazon ElastiCache	Permet de mettre en cache du contenu pour les applications Web.
DynamoDB Accelerator	Permet d'ajouter une accélération en mémoire à vos tables DynamoDB.

- Utilisez des services capables de vous aider à exécuter le code plus près des utilisateurs de votre charge de travail, tels que les suivants :

Service	Utilisation
Lambda@Edge	Destiné aux opérations exigeantes en puissance de calcul qui sont lancées lorsque des objets ne sont pas dans le cache.
CloudFront Fonctions Amazon	À utiliser pour des cas d'utilisation simples tels que HTTP des requêtes ou des manipulations de réponses qui peuvent être initiées par des fonctions de courte durée.
AWS IoT Greengrass	Permet d'exécuter du calcul local, une messagerie et une mise en cache de données pour les appareils connectés.

- Certaines applications nécessitent des points d'entrée fixes ou des performances plus élevées en réduisant la latence et l'instabilité du premier octet et en augmentant le débit. Ces applications peuvent bénéficier de services réseau qui fournissent des adresses IP anycast statiques et des TCP terminaisons aux emplacements périphériques. AWS Global Acceleratorpeut améliorer les performances de vos applications jusqu'à 60 % et permettre un basculement rapide pour

les architectures multirégionales. AWS Global Accelerator vous fournit des adresses IP anycast statiques qui servent de point d'entrée fixe pour vos applications hébergées dans une ou plusieurs d' Régions AWS entre elles. Ces adresses IP permettent au trafic de pénétrer sur le réseau AWS mondial aussi près que possible de vos utilisateurs. AWS Global Accelerator réduit le temps de configuration de la connexion initiale en établissant une TCP connexion entre le client et l'emplacement AWS périphérique le plus proche du client. Passez en revue l'utilisation de AWS Global Accelerator pour améliorer les performances de vosTCP/UDPworkloads et permettre un basculement rapide pour les architectures multirégionales.

Ressources

Bonnes pratiques associées :

- COST07-BP02 Mettre en œuvre les régions en fonction des coûts
- COST08-BP03 Mettre en œuvre des services pour réduire les coûts de transfert de données
- REL10-BP01 Déployer la charge de travail sur plusieurs sites
- REL10-BP02 Sélectionnez les emplacements appropriés pour votre déploiement multisite
- SUS01-BP01 Choisissez la région en fonction des exigences commerciales et des objectifs de durabilité
- SUS02-BP04 Optimiser le placement géographique des charges de travail en fonction de leurs exigences en matière de réseau
- SUS04-BP07 Minimiser le mouvement des données sur les réseaux

Documents connexes :

- AWS Infrastructure mondiale
- AWS Zones locales et AWS Outposts choix de la technologie adaptée à votre charge de travail périphérique
- Groupes de placement
- AWS Zones Locales
- AWS Outposts
- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator

- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Vidéos connexes :

- AWS Vidéo explicative sur les Zones Locales
- AWS Outposts: Overview and How it Works
- AWS re:Invent 2023 - Une stratégie de migration pour les charges de travail en périphérie et sur site
- AWS re:INVENT 2021 - AWS Outposts : Apporter l' AWS expérience sur site
- AWS re:Invent 2020 : AWS Wavelength : Exécutez des applications avec une latence extrêmement faible à la périphérie de la 5G
- AWS re:Invent 2022 - Zones AWS locales : création d'applications pour une périphérie distribuée
- AWS re:Invent 2021 - Création de sites Web à faible latence avec Amazon CloudFront
- AWS re:Invent 2022 - Améliorez les performances et la disponibilité avec AWS Global Accelerator
- AWS re:Invent 2022 - Construisez votre réseau étendu mondial en utilisant AWS
- AWS re:Invent 2020 : gestion du trafic mondial avec Amazon Route 53

Exemples connexes :

- AWS Global Accelerator Atelier de routage personnalisé
- Gestion des réécritures et des redirections à l’aide des fonctions de périphérie

PERF04-BP07 Optimiser la configuration du réseau en fonction des métriques

Utilisez les données collectées et analysées pour prendre des décisions avisées concernant l’optimisation de votre configuration réseau.

Anti-modèles courants :

- Vous supposez que tous les problèmes liés aux performances sont liés à l’application.
- Vous testez uniquement les performances de votre réseau à partir d’un emplacement proche de l’endroit où vous avez déployé la charge de travail.

- Vous utilisez des configurations par défaut pour tous les services du réseau.
- Vous surdimensionnez la ressource réseau afin de fournir une capacité suffisante.

Avantages liés au respect de cette bonne pratique : la collecte des métriques nécessaires de votre réseau AWS et la mise en œuvre d'outils de surveillance du réseau vous permettent de comprendre les performances du réseau et d'optimiser les configurations du réseau.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

La surveillance du trafic en provenance VPCs et à destination des sous-réseaux ou des interfaces réseau est essentielle pour comprendre comment utiliser les ressources AWS réseau et optimiser les configurations réseau. À l'aide des outils AWS réseau suivants, vous pouvez examiner plus en détail les informations relatives à l'utilisation du trafic, à l'accès au réseau et aux journaux.

Étapes d'implémentation

- Identifiez les indicateurs de performance clés tels que la latence ou la perte de paquets à collecter. AWS fournit plusieurs outils qui peuvent vous aider à collecter ces statistiques. Les outils suivants vous permettent d'obtenir des informations supplémentaires sur l'utilisation du trafic, l'accès au réseau et les journaux.

AWS outil	Où utiliser
Gestionnaire d'adresses VPC IP Amazon.	IPAMUtilisez-le pour planifier, suivre et surveiller les adresses IP pour vos charges de travail AWS et celles sur site. Il s'agit d'une bonne pratique pour optimiser l'utilisation et l'allocation des adresses IP.
VPCJournaux de flux	Utilisez les journaux de VPC flux pour capturer des informations détaillées sur le trafic à destination et en provenance des interfaces réseau de votreVPCs. Avec VPC Flow Logs, vous pouvez diagnostiquer les règles de groupe de sécurité trop restrictives ou trop permissives et déterminer la direction du trafic

AWS outil	Où utiliser
	à destination et en provenance des interfaces réseau.
AWS Transit Gateway Journaux de flux	Utilisez les journaux de AWS Transit Gateway flux pour capturer des informations sur le trafic IP à destination et en provenance de vos passerelles de transit.
DNSjournalisation des requêtes	Enregistrez les informations relatives aux DNS requêtes publiques ou privées reçues par Route 53. DNSLes journaux vous permetten t d'optimiser les DNS configurations en comprenant le domaine ou le sous-domaine qui a été demandé ou les EDGE emplaceme nts Route 53 qui ont répondu aux DNS requêtes.
Reachability Analyzer	Reachability Analyzer vous aide à analyser et à déboguer l’accessibilité du réseau. Reachability Analyzer est un outil d'analyse de configuration qui vous permet d'effectuer des tests de connectivité entre une ressource source et une ressource de destination dans votre. VPCs Cet outil vous aide à vérifier que votre configuration réseau correspond à la connectivité souhaitée.

AWS outil	Où utiliser
Analyseur d’accès réseau	Vous pouvez utiliser l’Analyseur d’accès réseau pour comprendre l’accès réseau à vos ressources. Vous pouvez utiliser l’analyseur d’accès réseau pour spécifier vos exigences en matière d’accès au réseau et identifier les chemins d’accès potentiels qui ne répondent pas à vos exigences spécifiées. En optimisan t la configuration de votre réseau correspon dant, vous pouvez comprendre et vérifier l’état de votre réseau et démontrer si votre réseau sur AWS répond à vos exigences de conformit é.
Amazon CloudWatch	Utilisez Amazon CloudWatch et activez les métriques appropriées pour les options de réseau. Veillez à choisir la métrique de réseau adaptée à votre charge de travail. Par exemple, vous pouvez activer les métriques pour l'utilisation des adresses VPC réseau, la VPC NAT passerelle AWS Transit Gateway, le VPN tunnel AWS Network Firewall, Elastic Load Balancing et AWS Direct Connect. La surveillance continue des métriques est une bonne pratique pour observer et comprendre l’état et l’utilisation de votre réseau. Elle vous aide à optimiser la configuration du réseau en fonction de vos observations.

AWS outil	Où utiliser
AWS Network Manager	Vous pouvez ainsi surveiller les performances historiques et en temps réel du réseau AWS mondial à des fins opérationnelles et de planification. AWS Network Manager Network Manager fournit une latence réseau globale entre les zones de disponibilité Régions AWS et au sein de chaque zone de disponibilité, ce qui vous permet de mieux comprendre le lien entre les performances de votre application et les performances du AWS réseau sous-jacent.
Amazon CloudWatch RUM	Utilisez Amazon CloudWatch RUM pour collecter les statistiques qui vous fournissent les informations qui vous aideront à identifier, à comprendre et à améliorer l'expérience utilisateur.

- Identifiez les principaux intervenants et les modèles de trafic des applications à l'aide VPC des journaux de AWS Transit Gateway flux.
- Évaluez et optimisez votre architecture réseau actuelleVPCs, y compris les sous-réseaux et le routage. Par exemple, vous pouvez évaluer dans quelle mesure le VPC peering est différent ou vous AWS Transit Gateway aider à améliorer la mise en réseau dans votre architecture.
- Évaluez les chemins de routage de votre réseau pour vérifier que le chemin le plus court entre les destinations est toujours utilisé. L'Analyseur d'accès réseau vous aide à le faire.

Ressources

Documents connexes :

- Journalisation des DNS requêtes publiques
- Qu'est-ce que c'est IPAM ?
- Définir Reachability Analyzer
- Définir l'Analyseur d'accès réseau
- CloudWatchindicateurs pour votre VPCs

- Optimisez les performances et réduisez les coûts d'analyse du réseau avec VPC Flow Logs au format Apache Parquet
- Surveillance de vos réseaux mondiaux et principaux à l'aide des CloudWatch métriques Amazon
- Surveiller en permanence le trafic et les ressources du réseau

Vidéos connexes :

- AWS re:Invent 2023 — Guide du développeur sur les réseaux cloud
- AWS re:Invent 2023 — Prêts pour la suite ? Designing networks for growth and flexibility
- AWS re:Invent 2023 — VPC Designs avancés et nouvelles fonctionnalités
- AWS re:Invent 2022 — Approfondissez l'infrastructure réseau AWS
- AWS re:Invent 2020 — Meilleures pratiques et astuces de mise en réseau avec le cadre Well-Architected AWS
- AWS re:Invent 2020 — Surveillance et résolution des problèmes du trafic réseau

Exemples connexes :

- Ateliers sur la mise en réseau AWS
- Surveillance réseau AWS
- Observation et diagnostic de votre réseau sur AWS
- Détecter et corriger les erreurs de configuration du réseau sur AWS

Processus et culture

Questions

- PERF 5. Comment vos pratiques et votre culture organisationnelles contribuent-elles à l’efficacité des performances de votre charge de travail ?

PERF 5. Comment vos pratiques et votre culture organisationnelles contribuent-elles à l’efficacité des performances de votre charge de travail ?

Lors de la création de l’architecture des charges de travail, vous pouvez adopter certains principes et certaines pratiques pour optimiser l’exécution de charges de travail cloud efficaces et performantes.

Pour adopter une culture qui favorise l’efficacité des performances des charges de travail dans le cloud, tenez compte des principes et pratiques clés suivants :

Bonnes pratiques

- PERF05-BP01 Établir des indicateurs de performance clés (KPIs) pour mesurer la santé et le rendement de la charge de travail
- PERF05-BP02 Utiliser des solutions de surveillance pour comprendre les domaines dans lesquels les performances sont les plus critiques
- PERF05-BP03 Définir un processus pour améliorer les performances de la charge de travail
- PERF05-BP04 Testez votre charge de travail
- PERF05-BP05 Utiliser l'automatisation pour résoudre de manière proactive les problèmes liés aux performances
- PERF05-BP06 Maintenez votre charge de travail et vos services up-to-date
- PERF05-BP07 Passez en revue les métriques à intervalles réguliers

PERF05-BP01 Établir des indicateurs de performance clés (KPIs) pour mesurer la santé et le rendement de la charge de travail

Identifiez ceux KPIs qui mesurent quantitativement et qualitativement les performances de la charge de travail. KPIsvous aider à mesurer l'état et les performances d'une charge de travail liée à un objectif commercial.

Anti-modèles courants :

- Vous surveillez uniquement les métriques au niveau du système pour avoir un aperçu de votre charge de travail et ne comprenez pas les impacts commerciaux possibles.
- Vous supposez que vos données KPIs sont déjà publiées et partagées sous forme de données métriques standard.
- Vous ne définissez pas une donnée quantitative ou mesurableKPI.
- Vous n'êtes pas en accord KPIs avec les objectifs ou les stratégies de l'entreprise.

Avantages de l'établissement de cette meilleure pratique : l'identification des éléments spécifiques KPIs représentatifs de la santé et des performances de la charge de travail permet d'aligner les équipes sur leurs priorités et de définir des résultats commerciaux réussis. Le partage de ces

métriques avec tous les départements offre une visibilité et un alignement sur les seuils, les attentes et l'impact commercial.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

KPIspermettre aux équipes commerciales et d'ingénierie de s'aligner sur la mesure des objectifs et des stratégies et sur la manière dont ces facteurs se combinent pour produire des résultats commerciaux. Par exemple, une charge de travail de site Web peut utiliser le temps de chargement de la page comme indication des performances globales. Cette métrique serait l'un des éléments de données pris en compte qui mesure l'expérience d'un utilisateur. En plus d'identifier les temps limites de chargement des pages, vous devez documenter le résultat attendu ou le risque commercial si les performances idéales ne sont pas atteintes. Un temps de chargement long des pages affecte directement vos utilisateurs finaux, nuit à leur expérience utilisateur et peut entraîner une perte de clients. Lorsque vous définissez vos KPI seuils, combinez à la fois les points de référence du secteur et les attentes de vos utilisateurs finaux. Par exemple, si la référence actuelle du secteur est une page Web qui se charge dans un délai de deux secondes, mais que vos utilisateurs finaux s'attendent à ce qu'une page Web se charge dans un délai d'une seconde, vous devez prendre en compte ces deux points de données lors de l'établissement du. KPI

Votre équipe doit évaluer votre charge de travail à KPIs l'aide de données granulaires en temps réel et de données historiques à titre de référence et créer des tableaux de bord qui effectuent des calculs métriques sur vos KPI données afin d'en tirer des informations opérationnelles et d'utilisation. KPIsdoit être documenté et inclure des seuils qui soutiennent les objectifs et les stratégies de l'entreprise, et doit être mappé aux indicateurs surveillés. KPIsdoit être revu lorsque les objectifs commerciaux, les stratégies ou les exigences des utilisateurs finaux changent.

Étapes d'implémentation

- Identifier les parties prenantes : identifier et documenter les principales parties prenantes de l'entreprise, y compris les équipes de développement et d'exploitation.
- Fixer des objectifs : collaborez avec ces parties prenantes pour définir et documenter les objectifs de votre charge de travail. Tenez compte des aspects critiques des performances de vos charges de travail, tels que le débit, le temps de réponse et le coût, ainsi que des objectifs métier, tels que la satisfaction des utilisateurs.
- Passez en revue les meilleures pratiques du secteur : passez en revue les meilleures pratiques du secteur pour identifier celles qui KPIs correspondent aux objectifs de votre charge de travail.

- Identifiez les indicateurs : identifiez les indicateurs qui correspondent aux objectifs de votre charge de travail et qui peuvent vous aider à mesurer les performances et les objectifs commerciaux. Établissez KPIs sur la base de ces indicateurs. Les mesures telles que le temps de réponse moyen ou le nombre d'utilisateurs simultanés sont des exemples de métriques.
- Définition et documentation KPIs : utilisez les meilleures pratiques du secteur et vos objectifs de charge de travail pour définir des cibles pour votre charge de travailKPI. Utilisez ces informations pour définir KPI des seuils de gravité ou de niveau d'alarme. Identifiez et documentez le risque et l'impact d'KPIune non-satisfaction.
- Mettre en œuvre la surveillance : utilisez des outils de surveillance tels qu'Amazon CloudWatch ou AWS Configpour collecter des métriques et mesurerKPIs.
- Communiquez visuellement KPIs : utilisez des outils de tableau de bord tels QuickSight qu'Amazon pour visualiser les parties prenantes et communiquer KPIs avec elles.
- Analyser et optimiser : passez régulièrement en revue et analysez KPIs pour identifier les domaines de votre charge de travail qui doivent être améliorés. Collaborez avec les parties prenantes pour mettre en œuvre ces améliorations.
- Revoir et affiner : passez régulièrement en revue les indicateurs et KPIs évaluez leur efficacité, en particulier lorsque les objectifs commerciaux ou les performances de la charge de travail changent.

Ressources

Documents connexes :

- CloudWatchdocumentation
- Surveillance, journalisation et performances AWS Partner s
- AWS outils d'observabilité
- L'importance des indicateurs de performance clés (KPIs) pour les migrations cloud à grande échelle
- Comment suivre l'optimisation de vos coûts à l'KPIsaide du KPI tableau de bord
- Documentation X-Ray
- Utilisation des tableaux de CloudWatch bord Amazon
- Amazon QuickSight KPIs

Vidéos connexes :

- AWS re:Invent 2023 - Optimisez les coûts et les performances et suivez les progrès en matière d'atténuation
- AWS re:Invent 2023 - Gérez les événements du cycle de vie des ressources à grande échelle avec AWS Health
- AWS re:Invent 2023 - Performances et efficacité sur Pinterest : optimisation des dernières instances
- AWS re:Invent 2022 - AWS optimisation : étapes réalisables pour des résultats immédiats
- AWS re:Invent 2023 - Élaborer une stratégie d'observabilité efficace
- AWS Summit SF 2022 - Observabilité complète et surveillance des applications avec AWS
- AWS re:Invent 2023 - Mise à l'échelle AWS des 10 premiers millions d'utilisateurs
- AWS re:Invent 2022 - Comment Amazon utilise de meilleurs indicateurs pour améliorer les performances de son site Web
- Création d'une stratégie de mesures efficace pour votre entreprise | AWS Événements

Exemples connexes :

- Création d'un tableau de bord avec Amazon QuickSight

PERF05-BP02 Utiliser des solutions de surveillance pour comprendre les domaines dans lesquels les performances sont les plus critiques

Comprenez et identifiez les domaines où l’augmentation des performances de votre charge de travail aura un impact positif sur l’efficacité ou l’expérience client. Par exemple, un site Web qui comporte un grand nombre d’interactions clients pourrait gagner à utiliser des services de périphérie pour rapprocher la diffusion de contenus des clients.

Anti-modèles courants :

- Vous supposez que les mesures de calcul standard telles que CPU l'utilisation ou la pression de la mémoire sont suffisantes pour détecter les problèmes de performances.
- Vous n’utilisez que les métriques par défaut enregistrées par le logiciel de surveillance que vous avez sélectionné.
- Vous n’examinez les métriques qu’en cas de problème.

Avantages de l'établissement de cette meilleure pratique : la compréhension des domaines de performance critiques aide les responsables de la charge de travail à surveiller KPIs et à hiérarchiser les améliorations à fort impact.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Configurez le end-to-end suivi pour identifier les modèles de trafic, la latence et les domaines de performance critiques. Surveillez vos modèles d'accès aux données afin d'identifier les requêtes lentes ou les données mal fragmentées et partitionnées. Identifiez les zones de charge de travail limitées à l'aide de tests ou de surveillance des charges.

améliorer l'efficacité des performances en comprenant votre architecture, vos modèles de trafic et d'accès aux données, et identifier vos temps de latence et de traitement. Identifier les goulots d'étranglement potentiels qui pourraient avoir une incidence sur l'expérience client à mesure que la charge de travail augmente. Après avoir enquêté sur ces domaines, déterminez quelle solution vous pouvez déployer afin de surmonter ces problèmes de performances.

Étapes d'implémentation

- Configurez end-to-end la surveillance pour capturer tous les composants et mesures de la charge de travail. Voici des exemples de solutions de surveillance sur AWS.

Service	Où utiliser
Surveillance CloudWatch des utilisateurs réels d'Amazon () RUM	Pour capturer les métriques de performances des applications à partir de sessions réelles côté client et front-end.
AWS X-Ray	Pour tracer le trafic à travers les couches applicatives et identifier la latence entre les composants et les dépendances. Utilisez les cartographies de services X-Ray afin de voir les relations et la latence entre les composant s de la charge de travail.

Service	Où utiliser
Informations sur les performances d'Amazon Relational Database Service	Pour consulter les métriques de performances de la base de données et identifier les améliorations des performances.
Surveillance RDS améliorée d'Amazon	Pour consulter les métriques de performances du système d'exploitation de la base de données.
Amazon DevOps Guru	Pour détecter les modèles de fonctionnement anormaux afin que vous puissiez identifier les problèmes opérationnels avant qu'ils n'affectent vos clients.

- Effectuez des tests afin de générer des métriques, d'identifier les tendances de trafic, les goulots d'étranglement et les domaines de performance critiques. Voici quelques exemples de méthodes de test :
  - Configurez CloudWatchSynthetic Canaries pour imiter les activités des utilisateurs basées sur le navigateur de manière programmatique à l'aide de tâches cron Linux ou d'expressions de taux afin de générer des métriques cohérentes au fil du temps.
  - Utiliser le test de charge distribuéAWS afin de générer un trafic de pointe ou de tester la charge de travail au taux de croissance attendu.
- Évaluez les métriques et la télémétrie pour identifier vos domaines de performances critiques. Examinez ces domaines avec votre équipe afin de discuter de la surveillance et des solutions pour éviter les goulots d'étranglement.
- Expérimentez des améliorations des performances et mesurez ces changements avec des données. Par exemple, vous pouvez utiliser CloudWatchEvidently pour tester les nouvelles améliorations et les impacts sur les performances de votre charge de travail.

Ressources

Documents connexes :

- Quoi de neuf en matière d' AWS observabilité à re:Invent 2023
- Bibliothèque Amazon Builders' Library
- Documentation X-Ray

- Amazon CloudWatch RUM
- Amazon DevOps Guru

Vidéos connexes :

- AWS re:Invent 2023 - [LAUNCH] Surveillance des applications pour les charges de travail modernes
- AWS re:Invent 2023 - Mise en œuvre de l'observabilité des applications
- AWS re:Invent 2023 - Élaboration d'une stratégie d'observabilité efficace
- AWS Summit SF 2022 - Observabilité complète et surveillance des applications avec AWS
- AWS re:Invent 2022 - AWS optimisation : étapes réalisables pour des résultats immédiats
- AWS re:Invent 2022 - La bibliothèque Amazon Builders' Library : 25 ans d'excellence opérationnelle d'Amazon
- AWS re:Invent 2022 - Comment Amazon utilise de meilleurs indicateurs pour améliorer les performances de son site Web
- Surveillance visuelle des applications avec Amazon CloudWatch Synthetics

Exemples connexes :

- Mesurez le temps de chargement des pages avec Amazon CloudWatch Synthetics
- Client CloudWatch RUM Web Amazon
- X-Ray SDK pour Python
- Test de charge distribué sur AWS

PERF05-BP03 Définir un processus pour améliorer les performances de la charge de travail

Définissez un processus d’évaluation de nouveaux services, modèles de conception, types de ressources et configurations au fur et à mesure qu’ils deviennent disponibles. Par exemple, exécutez des tests de performances existants sur de nouvelles offres d’instances afin de déterminer leur potentiel d’amélioration de votre charge de travail.

Anti-modèles courants :

- Vous supposez que votre architecture actuelle est statique et ne sera pas mise à jour au fil du temps.

- Vous introduisez des modifications d'architecture au fil du temps sans justification basée sur les métriques.

Avantages liés au respect de cette bonne pratique : un processus défini pour les modifications d'architecture rend possible l'utilisation des données collectées pour influencer la conception de votre charge de travail au fil du temps.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les performances de votre charge de travail présentent quelques contraintes clés. Documentez-les pour connaître les types d'innovations qui pourraient améliorer les performances de votre charge de travail. Utilisez ces informations lors de l'apprentissage de nouveaux services ou la technologie au fur et à mesure de leur disponibilité afin d'identifier les moyens d'atténuer des contraintes ou des goulets d'étranglement.

Identifiez les principales contraintes de performance pour votre charge de travail. Documentez les contraintes environnementales de votre charge de travail pour connaître les types d'innovations qui pourraient améliorer les performances de celle-ci.

Étapes d'implémentation

- Identifier KPIs : Identifiez les performances de votre charge de travail KPIs comme indiqué dans la section PERF05-BP01 Établir des indicateurs de performance clés (KPIs) pour mesurer la santé et le rendement de la charge de travail pour établir une base de référence de votre charge de travail.
- Mettre en œuvre le suivi : utilisez des outils AWS d'observabilité pour collecter des indicateurs de performance et les mesurerKPIs.
- Réalisation d'une analyse : effectuez une analyse approfondie pour identifier les domaines (tels que la configuration et le code d'application) de votre charge de travail qui ne sont pas performants, comme indiqué dans PERF05-BP02 Utiliser des solutions de surveillance pour comprendre les domaines dans lesquels les performances sont les plus critiques. Utilisez vos outils d'analyse et de performance pour identifier les stratégies d'amélioration des performances.
- Validation des améliorations : utilisez des environnements de test (sandbox) ou en préproduction pour valider l'efficacité des stratégies d'amélioration.
- Mise en œuvre des modifications : mettez en œuvre les modifications en production et surveillez en permanence les performances de la charge de travail. Documentez les améliorations et communiquez-les aux parties prenantes.

- Révision et affinage : passez régulièrement en revue votre processus d'amélioration des performances afin d'identifier les domaines à améliorer.

Ressources

Documents connexes :

- Blog AWS
- Quoi de neuf avec AWS
- AWS Générateur de compétences

Vidéos connexes :

- AWS re:Invent 2022 - Fournir des architectures durables et performantes
- AWS re:Invent 2023 - Optimisez les coûts et les performances et suivez les progrès en matière d'atténuation
- AWS re:Invent 2022 - AWS optimisation : étapes réalisables pour des résultats immédiats
- AWS re:Invent 2022 - Optimisez vos AWS charges de travail grâce à des conseils sur les meilleures pratiques

Exemples connexes :

- AWS Github

PERF05-BP04 Testez votre charge de travail

Effectuez un test de charge de votre charge de travail pour vérifier qu'elle peut supporter la charge de production et identifier les éventuels goulots d'étranglement en termes de performances.

Anti-modèles courants :

- Vous testez les différentes parties et non la totalité de votre charge de travail.
- Vous testez la charge sur une infrastructure qui n'est pas la même que votre environnement de production.
- Vous n'effectuez le test de charge que pour la charge prévue sans aller au-delà, avec pour but de prévoir où vous pourriez rencontrer des problèmes à l'avenir.

- Vous effectuez des tests de charge sans consulter la politique de EC2 test d'Amazon et sans soumettre de formulaire de soumission d'événements simulés. Cela entraîne l'échec de votre test, car il ressemble à un denial-of-service événement.

Avantages liés au respect de cette bonne pratique : la mesure de vos performances dans le cadre d'un test de charge vous indiquera où vous serez affecté au fil de l'augmentation de la charge. Cela peut vous permettre d'anticiper les changements nécessaires avant qu'ils n'affectent votre charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Les tests de charge dans le cloud sont un processus visant à mesurer les performances de la charge de travail cloud dans des conditions réalistes avec la charge utilisateur attendue. Ce processus implique la mise en service d'un environnement cloud de type production, l'utilisation d'outils de test de charge pour générer la charge et l'analyse de métriques pour évaluer la capacité de votre charge de travail à gérer une charge réaliste. Pour effectuer un test de charge, vous devez exécuter des versions de données de production factices ou légèrement altérées (supprimez les données sensibles ou les informations d'identification). Effectuez automatiquement des tests de charge dans le cadre de votre pipeline de livraison et comparez les résultats par rapport à des seuils KPIs et à des seuils prédéfinis. Ce processus vous permet de continuer à atteindre les performances requises.

Étapes d'implémentation

- Définition de vos objectifs de test : identifiez les aspects de performance de votre charge de travail que vous souhaitez évaluer, tels que le débit et le temps de réponse.
- Sélection d'un outil de test : choisissez et configurez l'outil de test de charge adapté à votre charge de travail.
- Configuration de votre environnement : configurez l'environnement de test en fonction de votre environnement de production. Vous pouvez utiliser AWS les services pour exécuter des environnements de production afin de tester votre architecture.
- Mettez en œuvre la surveillance : utilisez des outils de surveillance tels qu'Amazon CloudWatch pour collecter des métriques sur les ressources de votre architecture. Vous pouvez également collecter et publier des métriques personnalisées.
- Définition de des scénarios : définissez les scénarios et les paramètres de test de charge (tels que la durée du test et le nombre d'utilisateurs).

- Tests de charge : réalisez des scénarios de test à grande échelle. Profitez-en AWS Cloud pour tester votre charge de travail afin de découvrir où elle ne parvient pas à évoluer ou si elle évolue de manière non linéaire. Par exemple, utilisez les instances Spot pour générer des charges à faible coût et découvrir les goulots d'étranglement avant de les rencontrer en production.
- Analyse des résultats des tests : analysez les résultats pour identifier les goulots d'étranglement en matière de performances et les domaines à améliorer.
- Documentation et partage des résultats : documentez et rendez compte des résultats et des recommandations. Partagez ces informations avec les parties prenantes pour les aider à prendre des décisions éclairées concernant les stratégies d'optimisation des performances.
- Itération continuelle : les tests de charge doivent être effectués à une cadence régulière, en particulier après un changement ou une mise à jour du système.

Ressources

Documents connexes :

- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Test de charge distribué sur AWS

Vidéos connexes :

- AWS Sommet ANZ 2023 : Accélérez en toute confiance grâce AWS aux tests de charge distribués
- AWS re:Invent 2022 - Tirez parti AWS de vos 10 premiers millions d'utilisateurs
- Résoudre avec AWS des solutions : tests de charge distribués

- AWS re:Invent 2021 - Optimisez les applications grâce aux informations des utilisateurs finaux avec Amazon CloudWatch RUM
- Démo d'Amazon CloudWatch Synthetics

Exemples connexes :

- Test de charge distribué sur AWS

PERF05-BP05 Utiliser l'automatisation pour résoudre de manière proactive les problèmes liés aux performances

Utilisez des indicateurs de performance clés (KPIs), combinés à des systèmes de surveillance et d'alerte, pour résoudre de manière proactive les problèmes liés aux performances.

Anti-modèles courants :

- Vous autorisez uniquement le personnel des opérations à apporter des modifications opérationnelles à la charge de travail.
- Vous confiez toutes les activités de filtre des alarmes à l’équipe des opérations sans correction proactive.

Avantages liés au respect de cette bonne pratique : la correction proactive des actions d’alarme permet au personnel d’assistance de se concentrer sur les éléments qui ne sont pas exploitables automatiquement. Cela permet au personnel des opérations de gérer toutes les alarmes sans être submergé et de se concentrer uniquement sur les alarmes critiques.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d’implémentation

Utilisez des alarmes pour déclencher des actions automatisées afin de corriger les problèmes dans la mesure du possible. Faites remonter l’alarme aux personnes qui peuvent répondre si une réponse automatique n’est pas possible. Par exemple, vous pouvez disposer d'un système capable de prédire les valeurs attendues des indicateurs de performance clés (KPI) et de déclencher une alarme lorsqu'ils dépassent certains seuils, ou d'un outil capable d'arrêter ou d'annuler automatiquement les déploiements s'ils KPIs sont en dehors des valeurs attendues.

Mettez en place des processus qui rendent visibles les performances pendant que votre charge de travail est en cours d’exécution. Créez des tableaux de bord de surveillance et établissez des normes de référence pour les attentes en matière de performances pour déterminer si les performances de la charge de travail sont optimales.

Étapes d’implémentation

- Identification du processus de remédiation : identifiez et comprenez le problème lié aux performances qui peut être résolu automatiquement. Utilisez des solutions de AWS surveillance telles qu'Amazon CloudWatch ou AWS X-Ray pour vous aider à mieux comprendre la cause première du problème.

- Définissez le processus d'automatisation : créez un processus step-by-step de correction qui peut être utilisé pour résoudre automatiquement le problème.
- Configuration de l'événement d'initiation : configurez l'événement pour lancer automatiquement le processus de correction. Par exemple, vous pouvez définir un déclencheur pour redémarrer automatiquement une instance lorsqu'elle atteint un certain seuil d'CPUutilisation.
- Automatisez la correction : utilisez les AWS services et les technologies pour automatiser le processus de correction. Par exemple, AWS Systems Manager Automation fournit une solution sécurisée et évolutive d'automatisation du processus de résolution. Veillez à utiliser une logique d'auto-réparation pour annuler les modifications si elles ne permettent pas de résoudre le problème.
- Test du flux de travail : testez le processus de résolution automatisé dans un environnement de pré-production.
- Mise en œuvre du flux de travail : implémentez la correction automatique dans l'environnement de production.
- Élaboration d'un manuel : élaborez et documentez un manuel qui décrit les étapes du plan de remédiation, y compris les événements initiateurs, la logique de remédiation et les mesures prises. Veillez à former les parties prenantes pour les aider à répondre efficacement aux événements de résolution automatisée.
- Révision et affinage : évaluez régulièrement l'efficacité du flux de travail de correction automatisé. Ajustez les événements de lancement et la logique de résolution, si nécessaire.

Ressources

Documents connexes :

- CloudWatchDocumentation
- AWS Partner Network Partenaires de surveillance, de journalisation et de performance
- Documentation X-Ray
- Utilisation des alarmes et des actions d'alarme dans CloudWatch
- Élaborez une pratique d'automatisation du cloud pour l'excellence opérationnelle : les meilleures pratiques de AWS Managed Services
- Automatisez le réglage des performances de votre Amazon Redshift grâce à l'optimisation automatique des tables

Vidéos connexes :

- AWS re:Invent 2023 - Stratégies de mise à l'échelle automatisée, de correction et d'autoréparation intelligente
- AWS re:Invent 2023 - [LAUNCH] Surveillance des applications pour les charges de travail modernes
- AWS re:Invent 2023 - Mise en œuvre de l'observabilité des applications
- AWS re:Invent 2021 - Automatisation intelligente des opérations dans le cloud
- AWS re:Invent 2022 - Configuration de contrôles à grande échelle dans votre environnement AWS
- AWS re:Invent 2022 - Automatisation de la gestion des correctifs et de la conformité à l'aide de AWS
- AWS re:Invent 2022 - Comment Amazon utilise de meilleurs indicateurs pour améliorer les performances de son site Web
- AWS re:Invent 2023 - Prenez le dessus : diagnostiquez et résolvez les problèmes de performance avec Amazon RDS
- AWS re:Invent 2021 - {New Launch} Détectez et résolvez automatiquement les problèmes avec Amazon Guru DevOps
- AWS re:Invent 2023 - Centralisez vos opérations

Exemples connexes :

- CloudWatch Journaux, personnalisation des alarmes

PERF05-BP06 Maintenez votre charge de travail et vos services up-to-date

Restez up-to-date sur les nouveaux services et fonctionnalités du cloud pour adopter des fonctionnalités efficaces, résoudre les problèmes et améliorer l'efficacité globale des performances de votre charge de travail.

Anti-modèles courants :

- Vous supposez que votre architecture actuelle est statique et ne sera pas mise à jour au fil du temps.
- Vous ne disposez pas de systèmes ou de rythme régulier pour évaluer la compatibilité des packages et des logiciels mis à jour avec votre charge de travail.

Avantages de la mise en place de cette meilleure pratique : en établissant un processus pour rester à up-to-date jour avec les nouveaux services et offres, vous pouvez adopter de nouvelles fonctionnalités, résoudre les problèmes et améliorer les performances de la charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Évaluez les méthodes d'amélioration des performances au fur et à mesure que de nouveaux services, modèles de conception et fonctionnalités de produits entrent en scène. Identifiez celles de ces méthodes qui sont susceptibles d'améliorer les performances ou d'accroître l'efficacité de la charge de travail via l'évaluation, la discussion interne ou l'analyse externe. Mettez en place un processus permettant d'évaluer les mises à jour, les nouvelles fonctions et les services pertinents pour votre charge de travail. Par exemple, créez une démonstration de faisabilité qui utilise les nouvelles technologies ou consultez un groupe interne. Lorsque vous essayez de nouvelles idées ou services, exécutez des tests de performances pour mesurer leur impact sur les performances de la charge de travail.

Étapes d'implémentation

- Inventaire de votre charge de travail : établissez l'inventaire de votre logiciel de charge de travail et de l'architecture, et identifiez les composants pouvant être mis à jour.
- Identification des sources de mise à jour : identifiez les actualités et mettez à jour les sources liées aux composants de votre charge de travail. Par exemple, vous pouvez vous abonner au AWS blog What's New at pour découvrir les produits correspondant à votre composante de charge de travail. Vous pouvez vous abonner au RSS flux ou gérer vos abonnements par e-mail.
- Définition d'un calendrier de mise à jour : définissez un calendrier pour évaluer les nouveaux services et les nouvelles fonctionnalités adaptés à votre charge de travail.
  - Vous pouvez utiliser AWS Systems Manager Inventory pour collecter les métadonnées du système d'exploitation (OS), des applications et des instances à partir de vos EC2 instances Amazon et comprendre rapidement quelles instances exécutent le logiciel et les configurations requises par votre politique logicielle et quelles instances doivent être mises à jour.
- Évaluation de la nouvelle mise à jour : comprenez comment mettre à jour les composants de votre charge de travail. Profitez de l'agilité du cloud pour tester rapidement la façon dont les nouvelles fonctionnalités peuvent améliorer votre charge de travail afin de gagner en efficacité.
- Utiliser l'automatisation : utilisez l'automatisation pour le processus de mise à jour afin de réduire le niveau d'effort nécessaire au déploiement des nouvelles fonctionnalités et de limiter les erreurs causées par les processus manuels.

  - Vous pouvez utiliser CI/CD pour mettre à jour AMIs automatiquement des images de conteneur et d'autres artefacts liés à votre application cloud.
  - Vous pouvez utiliser des outils tels que AWS Systems Manager Patch Manager pour automatiser le processus de mise à jour du système et planifier l'activité à l'aide de AWS Systems Manager Maintenance Windows.
- Documentation du processus : documentez votre processus d'évaluation des mises à jour et des nouveaux services. Donnez aux propriétaires le temps et l'espace nécessaires pour rechercher, tester, expérimenter et valider les mises à jour et les nouveaux services. Reportez-vous aux exigences commerciales documentées et aidez KPIs à hiérarchiser les mises à jour qui auront un impact commercial positif.

Ressources

Documents connexes :

- Blog AWS
- Quoi de neuf avec AWS
- up-to-dateImplémentation d'images avec des pipelines EC2 Image Builder automatisés

Vidéos connexes :

- AWS Re:inForce 2022 - Automatisation de la gestion des correctifs et de la conformité à l'aide de AWS
- All Things Patch : AWS Systems Manager | AWS Événements

Exemples connexes :

- Gestion de l'inventaire et des correctifs
- Un atelier sur l'observabilité

PERF05-BP07 Passez en revue les métriques à intervalles réguliers

Vérifiez les métriques qui sont collectées dans le cadre de la maintenance de routine ou en réponse à des événements ou des incidents. Utilisez ces vérifications pour identifier d'une part les métriques qui ont été essentielles pour traiter les problèmes, et d'autre part les métriques supplémentaires, si elles ont été suivies, qui pourraient aider à identifier, traiter ou empêcher les problèmes.

Anti-modèles courants :

- Vous autorisez les métriques à rester dans un état d'alarme pendant longtemps.
- Vous créez des alarmes qui ne sont pas exploitables par un système d'automatisation.

Avantages liés au respect de cette bonne pratique : passez en revue en permanence les métriques qui sont collectées pour vérifier qu'elles identifient, résolvent ou préviennent correctement les problèmes. Les métriques peuvent également devenir caduques si vous les laissez dans un état d'alarme pendant longtemps.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Améliorez constamment la surveillance et la collecte des métriques. Lorsque vous répondez aux incidents ou aux événements, évaluez les métriques qui ont été utiles dans la gestion du problème et les métriques qui auraient pu aider mais ne sont pas suivies actuellement. Utilisez cette méthode pour améliorer la qualité des métriques que vous collectez afin de pouvoir prévenir ou résoudre plus rapidement les incidents futurs.

Lorsque vous répondez aux incidents ou aux événements, évaluez les métriques qui ont été utiles dans la gestion du problème et les métriques qui auraient pu aider mais ne sont pas suivies actuellement. Utilisez ce processus pour améliorer la qualité des métriques que vous collectez afin de pouvoir prévenir ou résoudre plus rapidement les incidents futurs.

Étapes d'implémentation

- Définition de métriques : définissez des métriques de performance critiques à surveiller qui correspondent à votre objectif de charge de travail, notamment des métriques telles que le temps de réponse et l'utilisation des ressources.
- Établissement de bases de référence : définissez une base de référence et une valeur souhaitable pour chaque métrique. La base de référence doit fournir des points de référence pour identifier les écarts ou les anomalies.
- Établissement d'une cadence : définissez une cadence (hebdomadaire ou mensuelle, par exemple) pour examiner les métriques critiques.
- Identification des problèmes de performance : au cours de chaque examen, évaluez les tendances et les écarts par rapport aux valeurs de référence. Recherchez les goulots d'étranglement ou les

anomalies au niveau des performances. Pour les problèmes identifiés, effectuez une analyse détaillée des causes profondes afin de comprendre la raison principale du problème.

- Identification des actions correctives : utilisez votre analyse pour identifier les actions correctives. Cela peut inclure le réglage des paramètres, la correction de bogues et la mise à l'échelle des ressources.
- Documentation des résultats : documentez vos conclusions, y compris les problèmes identifiés, les causes profondes et les mesures correctives.
- Répétition et amélioration : évaluez et améliorez en permanence le processus de révision des métriques. Utilisez les enseignements tirés de la révision précédente pour améliorer le processus au fil du temps.

Ressources

Documents connexes :

- CloudWatchDocumentation
- Collectez des métriques et des journaux à partir d'EC2instances Amazon et de serveurs sur site avec l'agent CloudWatch
- Interrogez vos indicateurs avec CloudWatch Metrics Insights
- AWS Partner Network Partenaires de surveillance, de journalisation et de performance
- Documentation X-Ray

Vidéos connexes :

- AWS re:Invent 2022 - Configuration de contrôles à grande échelle dans votre environnement AWS
- AWS re:Invent 2022 - Comment Amazon utilise de meilleurs indicateurs pour améliorer les performances de son site Web
- AWS re:Invent 2023 - Élaboration d'une stratégie d'observabilité efficace
- AWS Summit SF 2022 - Observabilité complète et surveillance des applications avec AWS
- AWS re:Invent 2023 - Prenez le dessus : diagnostiquez et résolvez les problèmes de performance avec Amazon RDS

Exemples connexes :

- Création d'un tableau de bord avec Amazon QuickSight

- CloudWatch Tableaux de bord

Optimisation des coûts

Le pilier Optimisation des coûts comprend la possibilité d'exécuter des systèmes pour offrir une valeur métier au prix le plus bas. Vous trouverez des conseils prescriptifs sur la mise en œuvre dans le livre blanc Pilier Optimisation des coûts.

Domaines de bonnes pratiques

- Pratiques en matière de gestion financière du cloud
- Sensibilisation aux dépenses et à l'utilisation
- Ressources rentables
- Gestion de la demande et offre de ressources
- Optimisation au fil du temps

Pratiques en matière de gestion financière du cloud

Question

- COÛT 1. Comment mettre en œuvre la gestion financière du cloud ?

COÛT 1. Comment mettre en œuvre la gestion financière du cloud ?

La mise en œuvre de la gestion financière du cloud (CFM) permet aux organisations de générer de la valeur ajoutée et d'être financièrement performantes en optimisant leurs coûts et l'utilisation, et en se mettant à l'échelle sur AWS.

Bonnes pratiques

- COST01-BP01 Assumer la responsabilité de l'optimisation des coûts
- COST01-BP02 Établir un partenariat entre la finance et la technologie
- COST01-BP03 Établissement de budgets et de prévisions cloud
- COST01-BP04 Intégrez la prise en compte des coûts dans vos processus organisationnels
- COST01-BP05 Rapport et notification sur l'optimisation des coûts
- COST01-BP06 Surveiller les coûts de manière proactive
- COST01-BP07 Tenez-vous au courant up-to-date des nouvelles versions de service

- COST01-BP08 Créer une culture axée sur les coûts
- COST01-BP09 Quantifier la valeur commerciale grâce à l'optimisation des coûts

COST01-BP01 Assumer la responsabilité de l'optimisation des coûts

Créez une équipe (bureau commercial cloud, centre d'excellence cloud ou FinOps équipe) chargée d'établir et de maintenir la connaissance des coûts au sein de votre organisation. Le propriétaire de l’optimisation des coûts peut être un individu ou une équipe (nécessite des personnes des équipes financières, technologiques et commerciales) qui comprend l’ensemble de l’organisation et la partie finance du cloud.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Il s'agit de l'introduction d'une fonction ou d'une équipe du Cloud Business Office (CBO) ou du Cloud Center of Excellence (CCOE) chargée d'établir et de maintenir une culture de prise en compte des coûts dans le domaine du cloud computing. Cette fonction peut être un individu existant, une équipe au sein de votre organisation ou une nouvelle équipe composée des principales parties prenantes de la finance, de la technologie et de l’organisation issues de toute l’entreprise.

La fonction (individu ou équipe) établit des priorités et consacre le pourcentage de temps requis aux activités de gestion et d’optimisation des coûts. Pour une petite organisation, la fonction peut consacrer un pourcentage de temps plus faible qu’une fonction à temps plein pour une grande entreprise.

La fonction exige une approche pluridisciplinaire, avec des capacités en gestion de projet, en science des données, en analyse financière et en développement de logiciels ou d’infrastructures. Elle peut améliorer l’efficacité de la charge de travail en procédant à des optimisations de coûts au sein de trois propriétés différentes :

- Centralisé : grâce à des équipes désignées telles que FinOps l'équipe, l'équipe de gestion financière du cloud (CFM), le bureau commercial du cloud (CBO) ou le centre d'excellence du cloud (CCoE), les clients peuvent concevoir et mettre en œuvre des mécanismes de gouvernance et promouvoir les meilleures pratiques à l'échelle de l'entreprise.
- Décentralisée : influence sur les équipes technologiques pour qu’elles optimisent les coûts.
- Hybride : une combinaison des équipes centralisée et décentralisée peut collaborer pour exécuter les optimisations de coûts.

La fonction peut être mesurée par rapport à leur capacité à exécuter et à atteindre les objectifs d'optimisation des coûts (par exemple, les métriques d'efficacité de la charge de travail).

Vous devez obtenir un parrainage de la direction pour cette fonction, ce qui est un facteur de réussite clé. Le parrain est considéré comme un défenseur d'une consommation efficace du cloud et apporte son soutien dans le cadre de la remontée pour l'équipe afin de garantir que les activités d'optimisation des coûts sont traitées avec le niveau de priorité défini par l'organisation. Sinon, les conseils peuvent être ignorés et les opportunités d'économies ne seront pas prioritaires. Ensemble, le sponsor et l'équipe aident votre organisation à utiliser le cloud de manière efficace et apportent une valeur ajoutée.

Si vous disposez du plan de support Business Enterprise-On-Ramp ou Enterprise et que vous avez besoin d'aide pour créer cette équipe ou cette fonction, contactez vos experts en gestion financière du cloud (CFM) par l'intermédiaire de votre équipe chargée de votre compte.

Étapes d'implémentation

- Définir des membres clés : toutes les parties concernées de votre organisation doivent contribuer à la gestion des coûts et s'y intéresser. Les équipes communes au sein des organisations incluent généralement : les responsables des finances, des applications ou des produits, la direction et les équipes techniques (DevOps). Certaines sont impliquées à temps plein (finance ou technique), tandis que d'autres le sont périodiquement, en fonction des besoins. Les personnes ou les équipes performantes CFM ont besoin des compétences suivantes :
  - Développement logiciel : dans le cas où des scripts et une automatisation sont créés.
  - Ingénierie d'infrastructure : pour déployer des scripts, automatiser des processus et comprendre comment les services et les ressources sont provisionnés.
  - Perspicacité opérationnelle : CFM il s'agit d'opérer efficacement sur le cloud en mesurant, en surveillant, en modifiant, en planifiant et en développant l'utilisation efficace du cloud.
- Définir des objectifs et des métriques : la fonction doit apporter de la valeur à l'organisation de différentes manières. Ces objectifs sont définis et évoluent continuellement au rythme de l'organisation. Les activités courantes incluent la création et l'exécution de programmes de formation sur l'optimisation des coûts au sein de l'organisation, le développement de normes à l'échelle de l'organisation, telles que la surveillance et la création de rapports pour l'optimisation des coûts, et la définition d'objectifs de charge de travail pour l'optimisation. Cette fonction doit également rendre compte régulièrement à l'organisation de sa capacité à optimiser les coûts.

  Vous pouvez définir des indicateurs de performance clés basés sur la valeur ou les coûts (KPIs). Lorsque vous définissez leKPIs, vous pouvez calculer le coût attendu en termes d'efficacité et de

résultats commerciaux attendus. Les indicateurs basés sur la valeur KPIs relient les indicateurs de coût et d'utilisation aux facteurs de valeur commerciale et aident à rationaliser l'évolution des dépenses. AWS La première étape de la dérivation basée sur la valeur KPIs consiste à travailler ensemble, au niveau interorganisationnel, pour sélectionner et convenir d'un ensemble standard de. KPIs

- Établir une cadence régulière : le groupe (équipes financières, technologiques et commerciales) doit se réunir régulièrement pour examiner ses objectifs et métriques. Une cadence type implique d'examiner l'état de l'organisation, de passer en revue les programmes en cours, puis de vérifier les métriques financières et d'optimisation globales. Par la suite, les principales charges de travail font l'objet d'un rapport plus détaillé.

  Pendant ces examens réguliers, vous pouvez examiner l'efficacité (le coût) de la charge de travail et les résultats métier. Par exemple, une hausse de 20 % du coût d'une charge de travail peut correspondre avec une utilisation client accrue. Dans ce cas, cette hausse de 20 % du coût peut être interprétée comme un investissement. Ces appels réguliers peuvent aider les équipes à identifier les valeurs KPIs qui donnent du sens à l'ensemble de l'organisation.

Ressources

Documents connexes :

- Blog AWS CCOE
- Création d'un bureau d'affaires du cloud
- CCOE- Centre d'excellence du cloud

Vidéos connexes :

- L'histoire du succès de Vanguard CCOE

Exemples connexes :

- Utiliser un centre d'excellence dans le cloud (CCOE) pour transformer l'ensemble de l'entreprise
- Construire un CCOE pour transformer l'ensemble de l'entreprise
- 7 pièges à éviter lors de la construction CCOE

COST01-BP02 Établir un partenariat entre la finance et la technologie

Impliquez les équipes financières et technologiques dans les discussions sur les coûts et l'utilisation à toutes les étapes de votre transition vers le cloud. Les équipes se réunissent régulièrement et discutent de sujets tels que les objectifs et les cibles organisationnels, l'état actuel des coûts et l'utilisation et les pratiques financières et comptables.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Dans le cloud, les équipes technologiques innovent plus rapidement grâce à la réduction de la durée des cycles d'approbation, d'achat et de déploiement des infrastructures. Il peut s'agir d'un ajustement pour les organisations financières auparavant habituées à exécuter des processus longs et gourmands en ressources pour l'acquisition et le déploiement de capitaux dans les centres de données et les environnements sur site, et la répartition des coûts uniquement lors de l'approbation du projet.

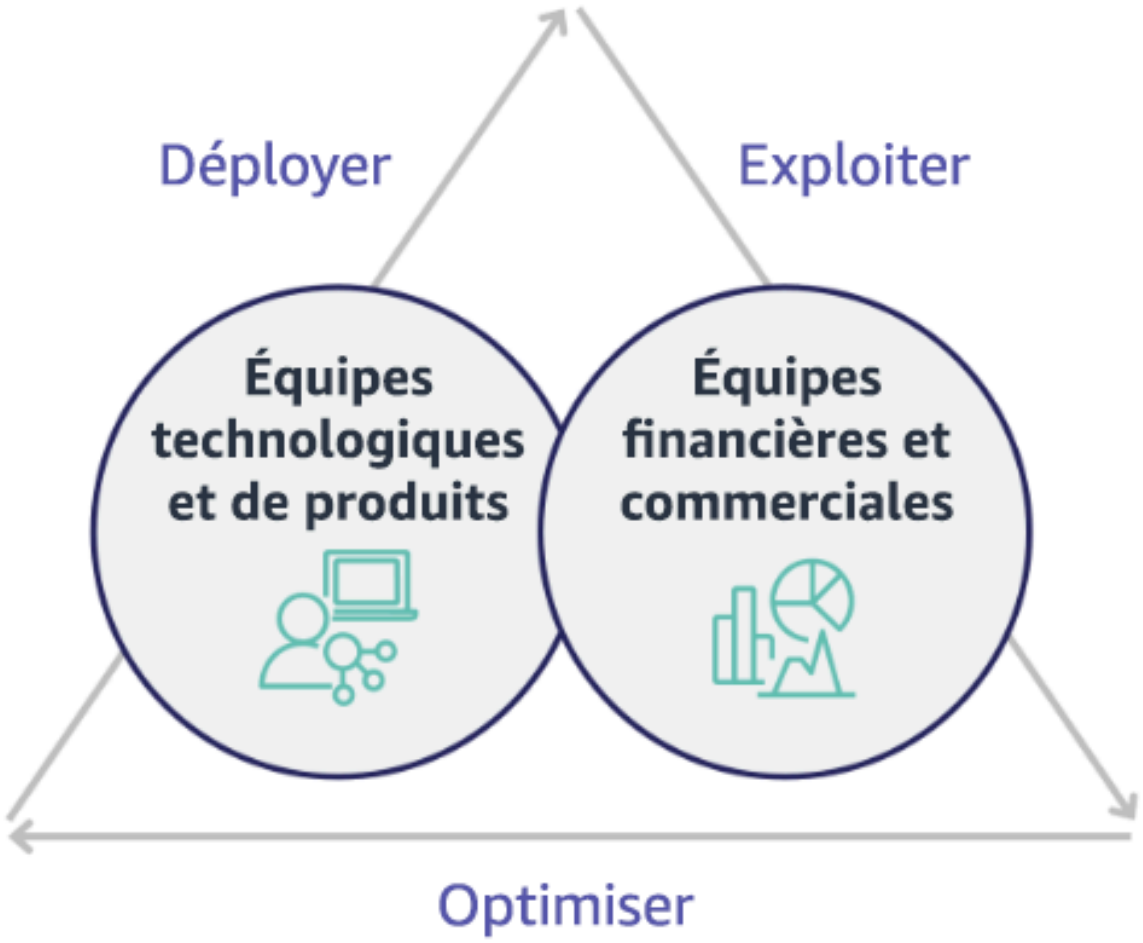
Du point de vue d'un organisme financier et d'acquisition, le processus de budgétisation des capitaux, de demandes de capitaux, d'approbations, d'acquisitions et d'installation d'une infrastructure physique a été appris et standardisé durant des décennies :

- Les équipes ingénierie ou informatiques sont généralement les demandeurs.
- Plusieurs équipes financières agissent en tant qu'approbateurs et acheteurs.
- Les équipes opérationnelles mettent en place, empilent et transfèrent ready-to-use l'infrastructure

Avec l'adoption du cloud, la consommation et l'acquisition d'infrastructure n'obéissent plus à une chaîne de dépendances. Dans le modèle cloud, les équipes technologiques et de produits ne se contentent plus de créer, mais sont les opérateurs et les propriétaires de leurs produits, responsables de la plupart des activités historiquement associées aux équipes financières et d'opérations, y compris l'acquisition et le déploiement.

Pour acquérir des ressources cloud, il suffit d'un compte et des bonnes autorisations. C'est également ce qui réduit les risques informatiques et financiers, ce qui signifie que les équipes sont toujours à quelques clics ou API appels pour mettre fin à des ressources cloud inactives ou inutiles. C'est également ce qui permet aux équipes technologiques d'innover plus rapidement : l'agilité et la capacité à mettre en place et à supprimer des expériences. Bien que la nature variable

de la consommation cloud puisse impacter la prévisibilité du point de vue de la prévision et de la budgétisation du capital, le cloud offre aux entreprises la possibilité de réduire les coûts de sur-approvisionnement, tout en diminuant les coûts d'opportunités associés au sous-approvisionnement conservateur.



Établissez un partenariat entre les principaux acteurs financiers et technologiques afin de créer une compréhension commune des objectifs organisationnels et de développer des mécanismes pour réussir financièrement dans le modèle de dépenses variables du cloud computing. Les équipes concernées au sein de votre organisation doivent être impliquées dans les discussions sur les coûts et l'utilisation à toutes les étapes de votre transition vers le cloud, y compris :

- Prospects financiers : les contrôleurs financiersCFOs, les planificateurs financiers, les analystes commerciaux, les responsables des achats, de l'approvisionnement et des comptes fournisseurs doivent comprendre le modèle cloud de consommation, les options d'achat et le processus de facturation mensuelle. Les services financiers doivent s'associer aux équipes technologiques

pour créer et socialiser une histoire de la valeur des TI et, ainsi, aider les équipes commerciales à comprendre le lien entre les dépenses en technologie et les résultats commerciaux. Prises sous cet angle, les dépenses technologiques ne sont pas considérées comme des coûts, mais plutôt comme des investissements. En raison des différences fondamentales entre le cloud (telles que le taux de changement d'utilisation, la tarification à l'usage, la tarification progressive, les modèles de tarification et les informations détaillées sur la facturation et l'utilisation) par rapport à l'exploitation sur site, il est essentiel que l'organisme financier comprenne comment l'utilisation du cloud peut influer sur les aspects commerciaux, notamment les processus d'acquisition, le suivi des incitations, la répartition des coûts et les états financiers.

- Responsables de technologiques : les responsables des technologies (y compris les propriétaires de produits et d'applications) doivent être conscients des exigences financières (par exemple, les contraintes budgétaires), ainsi que des exigences commerciales (par exemple, les contrats de niveau de service). Cela permet de mettre en œuvre la charge de travail pour atteindre les objectifs souhaités de l'organisation.

Le partenariat entre la finance et la technologie offre les avantages suivants :

- Les équipes financières et technologiques bénéficient d'une visibilité quasiment en temps réel sur les coûts et l'utilisation.
- Les équipes financières et technologiques établissent une procédure d'exploitation standard pour gérer les variations des dépenses liées au cloud.
- Les acteurs financiers agissent en tant que conseillers stratégiques en ce qui concerne la manière dont le capital est utilisé pour acheter des remises d'engagement (par exemple, les instances réservées ou les AWS Savings Plans) et la manière dont le cloud est utilisé pour développer l'organisation.
- Les processus existants de comptes fournisseurs et d'acquisition sont utilisés avec le cloud.
- Les équipes financières et technologiques collaborent pour prévoir les AWS coûts et l'utilisation futurs afin d'aligner et d'établir les budgets organisationnels.
- Une meilleure communication sur toute l'organisation grâce à un langage partagé et une compréhension commune des concepts financiers.

Les autres parties prenantes au sein de votre organisation qui doivent être impliquées dans les discussions sur les coûts et l'utilisation sont notamment :

- Propriétaires d’unités commerciales : les propriétaires d’unités commerciales doivent comprendre le modèle commercial du cloud afin de pouvoir orienter les unités commerciales et l’entreprise dans son ensemble. Cette connaissance du cloud est essentielle lorsqu’il est nécessaire de prévoir la croissance et l’utilisation de la charge de travail, et d’évaluer les options d’achat à plus long terme, telles que les instances réservées ou les Savings Plans.
- Équipe d'ingénierie : l'établissement d'un partenariat entre les équipes financières et technologiques est essentiel pour créer une culture axée sur les coûts qui encourage les ingénieurs à prendre des mesures en matière de gestion financière dans le cloud (CFM). L'un des problèmes courants des professionnels CFM des opérations financières et des équipes financières consiste à amener les ingénieurs à comprendre l'ensemble de l'activité sur le cloud, à suivre les meilleures pratiques et à prendre les mesures recommandées.
- Tierces parties : si votre organisation fait appel à des tiers (par exemple, des consultants ou des outils), assurez-vous qu'ils sont alignés sur vos objectifs financiers et qu'ils peuvent démontrer à la fois cet alignement grâce à leurs modèles d'engagement et leur retour sur investissement (ROI). En règle générale, les tiers contribueront à l’établissement de rapports et à l’analyse de toute charge de travail qu’ils gèrent, et ils fourniront une analyse des coûts de toute charge de travail qu’ils conçoivent.

La mise en œuvre CFM et la réussite nécessitent une collaboration entre les équipes financières, technologiques et commerciales, ainsi qu'un changement dans la façon dont les dépenses liées au cloud sont communiquées et évaluées au sein de l'organisation. Incluez les équipes d’ingénierie afin qu’elles participent aux discussions sur le coût et l’utilisation à chaque étape, et les encourager à suivre les bonnes pratiques ainsi qu’à prendre les mesures convenues en conséquence.

Étapes d’implémentation

- Définir des membres clés : veillez à ce que tous les membres concernés de vos équipes financières et technologiques s’impliquent dans le partenariat. Les membres concernés dans l’équipe financière sont ceux qui interagissent avec le projet de loi sur le cloud. Il s'agira généralement de contrôleurs financiersCFOs, de planificateurs financiers, d'analystes commerciaux, d'achats et d'approvisionnement. Les membres technologiques sont généralement les propriétaires de produits et d’applications, les responsables techniques et les représentants de toutes les équipes qui s’appuient sur le cloud. Les autres membres peuvent inclure les propriétaires d’unités commerciales, tels que le marketing qui influencera l’utilisation des produits. Il y a également des tiers, tels que des consultants afin d’assurer l’adéquation avec vos objectifs et vos mécanismes ainsi qu’une assistance pour les rapports d’activité.

- Définir de sujets de discussion : définissez les sujets communs aux équipes ou qui nécessitent une compréhension commune. Suivez le coût à partir de sa création jusqu'au paiement de la facture. Notez tous les membres impliqués, ainsi que les processus organisationnels qui doivent être appliqués. Ayez une compréhension de chacune de ses étapes ou de chacun de ses processus et des informations associées, telles que les modèles de tarification disponibles, la tarification progressive, les modèles de réduction, la budgétisation et les exigences financières.
- Établir une cadence régulière : pour créer un partenariat financier et technologique, mettez en place une cadence de communication régulière pour créer et maintenir un alignement. Le groupe doit se réunir régulièrement par rapport à ses objectifs et métriques. Une cadence type implique d'examiner l'état de l'organisation, de passer en revue les programmes en cours, puis de vérifier les métriques financières et d'optimisation globales. Les principales charges de travail font l'objet d'un rapport plus détaillé.

Ressources

Documents connexes :

- AWS Blog d'actualités

COST01-BP03 Établissement de budgets et de prévisions cloud

Ajustez les processus existants de budgétisation et de prévision d'organisation afin qu'ils soient compatibles avec la nature hautement variable des coûts et de l'utilisation du cloud. Les processus doivent être dynamiques en utilisant des algorithmes basés sur les tendances ou les facteurs d'activité, ou une combinaison des deux.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Dans les configurations informatiques traditionnelles sur site, les clients rencontrent souvent des difficultés pour planifier les coûts fixes qui ne changent que de temps en temps, généralement lors de l'achat de nouveau matériel informatique et de nouveaux services pour répondre à une forte hausse de la demande. En revanche, AWS Cloud adopte une approche différente : les clients paient les ressources qu'ils utilisent en fonction de leurs besoins informatiques et commerciaux réels. Dans l'environnement cloud, la demande peut fluctuer sur une base mensuelle, quotidienne ou même horaire.

L'utilisation du cloud apporte efficacité, rapidité et agilité, ce qui se traduit par des coûts et des modèles d'utilisation très variables. Les coûts peuvent diminuer ou parfois augmenter en réponse à une meilleure efficacité de la charge de travail ou au déploiement de nouvelles charges de travail et fonctionnalités. Alors que les charges de travail se mettent à l'échelle pour répondre aux besoins d'une clientèle croissante, l'utilisation du cloud et les coûts augmentent en conséquence en raison de l'accessibilité accrue des ressources. Cette flexibilité des services cloud s'étend aux coûts et aux prévisions, ce qui crée une certaine élasticité.

Il est essentiel de s'aligner étroitement sur ces besoins commerciaux et ces moteurs de demande changeants, et de viser une planification aussi précise que possible. Les processus budgétaires organisationnels traditionnels doivent être adaptés pour tenir compte de cette variabilité.

Envisagez de modéliser les coûts lorsque vous prévoyez le coût des nouvelles charges de travail. La modélisation des coûts crée une compréhension de base des coûts attendus du cloud. Cela vous permet d'effectuer des analyses du coût total de possession (TCO), du retour sur investissement (ROI) et d'autres analyses financières, de définir des objectifs et des attentes avec les parties prenantes et d'identifier les opportunités d'optimisation des coûts.

Votre organisation doit comprendre les définitions des coûts et les regroupements acceptés. Le niveau de détail auquel vous faites vos prévisions peut varier en fonction de la structure de votre organisation et des flux de travail internes. Sélectionnez une granularité adaptée à vos besoins spécifiques et à votre configuration organisationnelle. Il est important de comprendre à quel niveau la prévision est réalisée :

- Compte de gestion ou niveau AWS Organizations : le compte de gestion est le compte que vous utilisez pour créer AWS Organizations. Les organisations ont un seul compte de gestion par défaut.
- Compte lié ou membre : un compte dans Organisations est un Compte AWS standard qui contient vos ressources AWS et les identités qui peuvent accéder à ces ressources.
- Environnement : un environnement est une collection de ressources AWS qui exécute une version de l'application. Un environnement peut être créé avec plusieurs comptes liés ou comptes membres.
- Projet : un projet est une combinaison d'objectifs ou de tâches définis à accomplir au cours d'une période déterminée. Il est important de prendre en compte le cycle de vie du projet lors de votre prévision.
- Services AWS : groupes ou catégories tels que les services de calcul ou de stockage dans lesquels vous pouvez regrouper des services AWS pour votre prévision.

- Regroupement personnalisé : vous pouvez créer des groupes personnalisés en fonction des besoins de votre organisation, tels que des unités commerciales, des centres de coûts, des équipes, des étiquettes de répartition des coûts, des catégories de coûts, des comptes liés ou une combinaison de ces éléments.

Identifiez les facteurs commerciaux susceptibles d'avoir un impact sur votre coût d'utilisation et établissez des prévisions pour chacun d'entre eux séparément afin de calculer l'utilisation prévue à l'avance. Certains de ces facteurs peuvent être liés aux équipes informatiques et aux équipes produit au sein de l'organisation. D'autres facteurs commerciaux, tels que les événements commerciaux, les promotions, les expansions géographiques, les fusions et les acquisitions, sont connus de vos responsables des ventes, de vos responsables marketing et des responsables de l'entreprise. Il est donc important de collaborer et de tenir compte de tous ces moteurs de la demande également.

Vous pouvez utiliser AWS Cost Explorer pour effectuer des prévisions basées sur les tendances dans une plage temporelle future définie en fonction de vos dépenses passées. Le moteur de prévision de AWS Cost Explorer segmente vos données historiques en fonction des types de frais (par exemple, les instances réservées) et utilise une combinaison de modèles de machine learning et de modèles basés sur des règles pour prédire les dépenses sur tous les types de frais individuellement.

Une fois que vous avez établi votre processus de prévision et créé des modèles, vous pouvez utiliser AWS Budgets pour définir des budgets personnalisés à un niveau granulaire en spécifiant la période, la récurrence ou le montant (fixe ou variable) et en ajoutant des filtres tels que le service, Région AWS et des balises. Le budget est généralement préparé pour une seule année et reste fixe, exigeant un respect strict de la part de toutes les personnes concernées. En revanche, les prévisions sont plus souples. Elles permettent des réajustements tout au long de l'année et fournissent des projections dynamiques sur une période d'un, deux ou trois ans. Les budgets et les prévisions jouent un rôle crucial dans l'établissement des attentes financières parmi les différents acteurs technologiques et commerciaux. Des prévisions et une mise en œuvre précises permettent également de responsabiliser les parties prenantes qui sont directement en charge des coûts de provisionnement. Cela permet aussi de les sensibiliser aux coûts en général.

Pour suivre les performances de vos budgets existants, vous pouvez créer des rapports AWS Budgets et programmer leur envoi par e-mail à vous-même ainsi qu'à vos parties prenantes à un rythme régulier. Vous pouvez également créer des alertes AWS Budgets basées sur les coûts réels (qui sont réactives par essence) ou sur les coûts prévus, ce qui vous donne le temps de mettre en place des mesures d'atténuation contre les dépassements de coûts potentiels. Vous pouvez être alerté lorsque votre coût ou votre utilisation dépasse un certain niveau ou si les prévisions indiquent qu'ils vont dépasser le montant que vous avez fixé dans votre budget.

Ajustez les processus de budgétisation et de prévision existants pour qu'ils soient plus dynamiques à l'aide d'algorithmes basés sur les tendances (avec les coûts historiques comme entrées) et d'algorithmes basés sur des facteurs (par exemple, le lancement de nouveaux produits, l'expansion régionale ou de nouveaux environnements pour les charges de travail). Ces algorithmes sont idéaux pour un environnement de dépenses dynamique et variable. Une fois que vous avez déterminé votre prévision basée sur les tendances à l'aide de Cost Explorer ou de tout autre outil, utilisez AWS Pricing Calculator pour estimer votre cas d'utilisation AWS et les coûts futurs en fonction de l'utilisation prévue (trafic, demandes par seconde ou instances Amazon EC2 requises).

Surveillez l'exactitude de ces prévisions, car les budgets doivent être établis sur la base de ces calculs et estimations prévisionnels. Contrôlez la précision et l'efficacité des prévisions de coûts intégrées du cloud. Passez régulièrement en revue les dépenses réelles par rapport à vos prévisions et ajustez-les si nécessaire pour améliorer la précision des prévisions. Suivez l'écart des prévisions et effectuez une analyse des causes profondes de l'écart signalé pour agir et ajuster les prévisions.

Comme indiqué dans la rubrique COST01-BP02 Établir un partenariat entre la finance et la technologie, il est important d'encourager un partenariat et une cadence entre les services informatiques, les secteurs financiers et d'autres parties prenantes afin de vérifier qu'ils utilisent tous les mêmes outils ou processus dans un souci de cohérence. Au cas où les budgets devraient être modifiés, augmentez le nombre de points de contact chargés de la cadence afin de réagir plus rapidement à ces changements.

Étapes d'implémentation

- Définissez le langage des coûts au sein de l'organisation : créez un langage de coût AWS commun au sein de l'organisation avec plusieurs dimensions et groupements. Assurez-vous que les parties prenantes comprennent la granularité des prévisions, les modèles de tarification et le niveau de vos prévisions de coûts.
- Analysez les prévisions basées sur les tendances : utilisez des outils de prévision basés sur les tendances tels que AWS Cost Explorer et Amazon Forecast. Analysez votre coût d'utilisation en fonction de plusieurs dimensions comme le service, le compte, les balises et les catégories de coûts.
- Analysez les précisions basées sur des facteurs : identifiez l'impact des facteurs commerciaux sur votre utilisation du cloud et établissez des prévisions pour chacun d'entre eux séparément afin de calculer à l'avance le coût d'utilisation prévu. Travaillez en étroite collaboration avec les propriétaires d'unités commerciales et les parties prenantes pour comprendre l'impact sur les nouveaux facteurs et calculer les changements de coûts attendus afin de définir des budgets précis.

- Mettez à jour les processus existants de prévisions et de budget : définissez vos processus de prévisions et de budget en vous basant sur les méthodes de prévision adoptées, telles que les méthodes basées sur les tendances, sur les facteurs commerciaux ou une combinaison de ces deux méthodes. Les budgets doivent être calculés, réalistes et basés sur vos prévisions.
- Configuration des alertes et des notifications : utilisez les alertes AWS Budgets et la détection des anomalies de coûts pour recevoir des alertes et des notifications.
- Effectuez des révisions régulières avec des parties prenantes clés : par exemple, alignez-vous sur les changements de direction de l'entreprise et d'utilisation avec les parties prenantes des secteurs informatiques et des secteurs financiers, les équipes de plateforme et d'autres secteurs de l'entreprise.

Ressources

Documents connexes :

- AWS Cost Explorer
- AWS Cost and Usage Report
- Prévisions avec Cost Explorer
- Prévisions Amazon QuickSight
- AWS Budgets

Vidéos connexes :

- Comment utiliser AWS Budgets pour suivre mes dépenses et mon utilisation
- Série sur l'optimisation des coûts AWS : AWS Budgets

Exemples connexes :

- Compréhension et établissement de prévisions basées sur des facteurs
- Comment établir et promouvoir une culture de prévision
- Comment améliorer vos prévisions des coûts du cloud
- Utilisation des bons outils pour prévoir les coûts du cloud

COST01-BP04 Intégrez la prise en compte des coûts dans vos processus organisationnels

Mettez en œuvre la sensibilisation aux coûts, créez une transparence et intégrez une sensibilisation à l'égard des coûts dans les processus nouveaux ou existants qui ont une incidence sur l'utilisation, et tirez parti des processus existants pour la sensibilisation aux coûts. Intégrez la sensibilisation aux coûts dans la formation des employés.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

La sensibilisation aux coûts doit être mise en œuvre dans les processus organisationnels nouveaux et existants. Il s'agit de l'une des capacités prérequises fondamentales pour les autres bonnes pratiques. Il est recommandé de réutiliser et de modifier les processus existants dans la mesure du possible, ce qui réduit l'impact sur l'agilité et la vitesse. Signalez les coûts du cloud aux équipes technologiques et aux décideurs des équipes commerciales et financières afin de les sensibiliser aux coûts et d'établir des indicateurs de performance clés d'efficacité (KPIs) pour les acteurs financiers et commerciaux. Les recommandations suivantes vous aideront à mettre en œuvre la sensibilisation aux coûts dans votre charge de travail :

- Vérifiez que la gestion des modifications comprend une mesure des coûts pour quantifier l'impact financier des modifications. Cela permet de répondre de manière proactive aux préoccupations liées aux coûts et de mettre en évidence les économies réalisées.
- Vérifiez que l'optimisation des coûts est une composante essentielle de vos capacités d'exploitation. Par exemple, vous pouvez tirer parti des processus de gestion des incidents existants pour investiguer et identifier les causes racines des anomalies de coût et d'utilisation ou surcoûts.
- Accélérez la réduction des coûts et la génération de valeur métier avec l'automatisation ou l'utilisation d'outils. Lorsque vous réfléchissez au coût de la mise en œuvre, encadrez la conversation de manière à inclure un élément de retour sur investissement (ROI) afin de justifier l'investissement en temps ou en argent.
- Allouez les coûts de cloud en implémentant des relevés des services reçus ou des facturations internes pour les dépenses de cloud, y compris les options d'achat basées sur l'engagement, les services partagés et les achats marketplace afin de stimuler la plupart de la consommation de cloud sensible aux coûts.
- Étendez les programmes de formation et de développement existants afin d'y inclure une formation de sensibilisation aux coûts dans toute votre entreprise. Il est recommandé d'inclure une

formation et une certification continues. Cela permettra de créer une organisation capable de gérer automatiquement les coûts et l'utilisation.

- Profitez d'outils AWS natifs gratuits tels que AWS Cost Anomaly DetectionAWS Budgets, et AWS Budgets Reports.

Lorsque les organisations adoptent systématiquement des pratiques de gestion financière dans le cloud (CFM), ces comportements s'enracinent dans leur manière de travailler et de prendre des décisions. Il en résulte une culture plus attentive aux coûts, que ce soit pour les développeurs qui conçoivent l'architecture d'une nouvelle born-in-the-cloud application ou pour les responsables financiers qui analysent ces nouveaux investissements dans ROI le cloud.

Étapes d'implémentation

- Identifiez les processus organisationnels pertinents : chaque unité organisationnelle passe en revue ses processus et identifie les processus qui ont un impact sur les coûts et l'utilisation. Tous les processus qui entraînent la création ou l'arrêt d'une ressource doivent être inclus dans la vérification. Recherchez des processus qui peuvent soutenir la sensibilisation aux coûts dans votre entreprise, tels que la gestion des incidents et la formation.
- Instaurez une culture autonome consciente des coûts : assurez-vous que toutes les parties prenantes concernées s'alignent sur les coûts cause-of-change et qu'elles ont un impact en termes de coûts afin qu'elles comprennent le coût du cloud. Cela permettra à votre entreprise de mettre en place une culture de l'innovation consciente des coûts et autonome.
- Mettez à jour les processus avec la sensibilisation aux coûts : chaque processus est modifié pour tenir compte des coûts. Le processus peut nécessiter des contrôles préalables supplémentaires, tels que l'évaluation de l'impact du coût ou des contrôles a posteriori confirmant que les changements attendus en matière de coût et d'utilisation se sont produits. Les processus de soutien, tels que la formation et la gestion des incidents, peuvent être étendus pour inclure des éléments relatifs au coût et à l'utilisation.

Pour obtenir de l'aide, contactez CFM des experts par l'intermédiaire de l'équipe chargée de votre compte ou consultez les ressources et les documents connexes ci-dessous.

Ressources

Documents connexes :

- AWS Gestion financière dans le cloud

Exemples connexes :

- Stratégie pour une gestion des coûts de cloud efficace
- Série de blogs sur le contrôle des coûts n° 3 : comment gérer les augmentations de coûts
- Un guide pour débutants AWS Cost Management

COST01-BP05 Rapport et notification sur l'optimisation des coûts

Mettez en place des budgets pour le cloud et configurez des mécanismes pour détecter les anomalies d'utilisation. Configurez les outils connexes pour les alertes de coût et d'utilisation par rapport à des objectifs prédéfinis et recevez des notifications lorsqu'une utilisation dépasse ces objectifs. Organisez des réunions régulières pour analyser la rentabilité de vos charges de travail et promouvoir la sensibilisation aux coûts.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Vous devez régulièrement signaler l'optimisation des coûts et de l'utilisation au sein de votre entreprise. Vous pouvez mettre en œuvre des sessions dédiées pour discuter des performances en matière de coûts, ou inclure l'optimisation des coûts dans vos cycles de rapports opérationnels réguliers pour vos charges de travail. Utilisez des services et des outils pour contrôler régulièrement vos performances en matière de coûts et mettre en œuvre des possibilités d'économies.

Affichez votre coût et votre utilisation avec plusieurs filtres et niveaux de précision avec AWS Cost Explorer, qui fournit des tableaux de bord et des rapports tels que les coûts par service ou par compte, les coûts quotidiens ou les coûts du marketplace. Suivez l'évolution de vos coûts et de votre utilisation par rapport aux budgets configurés avec les rapports AWS Budgets.

AWS BudgetsÀ utiliser pour définir des budgets personnalisés afin de suivre vos coûts et votre utilisation et de répondre rapidement aux alertes reçues par e-mail ou aux notifications Amazon Simple Notification Service (AmazonSNS) si vous dépassez votre seuil. Définissez votre période budgétaire préférée sur une période quotidienne, mensuelle, trimestrielle ou annuelle, et établissez des limites budgétaires spécifiques pour rester informé de l'évolution des coûts et de l'utilisation réels ou prévus par rapport à votre seuil budgétaire. Vous pouvez également configurer des alertes et des actions automatiques par rapport à ces alertes ou via un processus d'approbation en cas de dépassement d'une cible budgétaire.

Mettez en œuvre des notifications sur les coûts et l'utilisation pour garantir que les modifications des coûts et de l'utilisation puissent être prises en compte rapidement en cas d'imprévu. AWS Cost Anomaly Detectionvous permet de réduire les imprévus en matière de coûts et d'améliorer le contrôle sans ralentir l'innovation. AWS Cost Anomaly Detection identifie les dépenses anormales et leurs causes profondes, ce qui contribue à réduire le risque de surprises liées à la facturation. En trois étapes simples, vous pouvez créer votre propre surveillance contextualisée et recevoir des alertes en cas de dépense irrégulière détectée.

Vous pouvez également utiliser Amazon QuickSight with AWS Cost and Usage Report (CUR) data, pour fournir des rapports hautement personnalisés avec des données plus détaillées. Amazon vous QuickSight permet de planifier des rapports et de recevoir des e-mails de rapports de coûts périodiques pour connaître l'historique des coûts et de l'utilisation ou des opportunités de réduction des coûts. Consultez notre solution Cost Intelligence Dashboard (CID) développée sur Amazon QuickSight, qui vous offre une visibilité avancée.

Utilisation AWS Trusted Advisor, qui fournit des conseils pour vérifier si les ressources allouées sont conformes aux AWS meilleures pratiques en matière d'optimisation des coûts.

Vérifiez vos recommandations en matière de Savings Plans à l’aide de graphiques visuels en fonction de vos coûts et de votre utilisation. Des graphiques horaires présentent les dépenses à la demande en regard de l’engagement recommandé des Savings Plans, fournissant un aperçu des économies estimées, de la couverture et de l’utilisation des Savings Plans. Cela permet aux organisations de comprendre comment s’appliquent leurs Savings Plans à chaque heure de dépenses sans avoir à investir du temps et des ressources dans l’élaboration de modèles pour analyser leurs dépenses.

Créez régulièrement des rapports contenant un aperçu des plans Savings Plans, des instances réservées et des recommandations de EC2 redimensionnement d'Amazon AWS Cost Explorer pour commencer à réduire les coûts associés aux charges de travail stables, aux ressources inactives et sous-utilisées. Identifiez et récupérez les dépenses inutiles liées au cloud pour les ressources déployées. Les dépenses inutiles liées au cloud se produisent lorsque des ressources de taille inappropriée sont créées, ou des modèles d’utilisation différents sont observés au lieu de ce qui est prévu. Suivez les AWS meilleures pratiques pour réduire le gaspillage ou demandez à l'équipe chargée de votre compte et à votre partenaire de vous aider à optimiser et à réduire vos coûts liés au cloud.

Générez des rapports réguliers pour profiter de meilleures options d’achat pour vos ressources afin de réduire les coûts unitaires de vos charges de travail. Les options d'achat telles que Savings Plans, Reserved Instances ou Amazon EC2 Spot Instances offrent les économies les plus importantes

pour les charges de travail tolérantes aux pannes et permettent aux parties prenantes (propriétaires d'entreprises, équipes financières et techniques) de participer à ces discussions d'engagement.

Partagez les rapports contenant des opportunités ou des annonces de nouvelles versions susceptibles de vous aider à réduire le coût total de possession (TCO) du cloud. Adoptez de nouveaux services, régions, fonctionnalités, solutions ou moyens de réduire davantage les coûts.

Étapes d’implémentation

- Configuration AWS Budgets : configurez AWS Budgets sur tous les comptes en fonction de votre charge de travail. Définissez un budget pour les dépenses globales des comptes et un budget pour la charge de travail à l’aide de balises.
  - Ateliers Well-Architected : utilisation des coûts et de la gouvernance
- Création de rapports sur l’optimisation des coûts : définissez un cycle régulier pour discuter de l’efficacité de la charge de travail et pour l’analyser. À l’aide des métriques définies, rendez compte des métriques atteintes et du coût associé. Identifiez et corrigez les tendances négatives, tout en ciblant les tendances positives que vous pouvez promouvoir dans votre organisation. Les rapports doivent impliquer des représentants des finances, des équipes d’application et des propriétaires, ainsi que des décideurs clés en ce qui concerne les dépenses liées au cloud.

Ressources

Documents connexes :

- AWS Cost Explorer
- AWS Trusted Advisor
- AWS Budgets
- AWS Cost and Usage Report
- AWS Budgets Bonnes pratiques
- Analytique Amazon S3

Exemples connexes :

- Ateliers Well-Architected : utilisation des coûts et de la gouvernance
- Principaux moyens de commencer à optimiser vos coûts liés au AWS cloud

COST01-BP06 Surveiller les coûts de manière proactive

Mettez en œuvre des outils et des tableaux de bord pour surveiller de manière proactive les coûts de la charge de travail. Vérifiez régulièrement les coûts grâce aux outils configurés ou prêts à l'emploi. Ne vous contentez pas d'examiner les coûts et les catégories lorsque vous recevez des notifications. La surveillance et l'analyse proactives des coûts permettent d'identifier les tendances positives et de les promouvoir dans toute votre organisation.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Il est recommandé de surveiller les coûts et l'utilisation de manière proactive au sein de votre entreprise et pas seulement lorsque des exceptions ou des anomalies se présentent. Des tableaux de bord très visibles dans votre bureau ou votre environnement de travail garantissent que les personnes clés ont accès aux informations dont elles ont besoin et indiquent que l'organisation se concentre sur l'optimisation des coûts. Des tableaux de bord visibles vous permettent de promouvoir activement les résultats positifs et de les mettre en œuvre dans toute votre organisation.

Créez une routine quotidienne ou fréquente à utiliser AWS Cost Explorerou tout autre tableau de bord tel qu'Amazon QuickSight pour voir les coûts et les analyser de manière proactive. Analysez l'utilisation et les coûts des AWS services au AWS niveau du compte, de la charge de travail ou au niveau du AWS service spécifique à l'aide du regroupement et du filtrage, et validez s'ils sont attendus ou non. Utilisez les balises ainsi que la granularité horaire et au niveau des ressources pour filtrer et identifier les coûts facturés pour les ressources principales. Vous pouvez également créer vos propres rapports avec le tableau de bord Cost Intelligence, une QuickSight solution Amazon conçue par AWS Solutions Architects, et comparer vos budgets aux coûts et à l'utilisation réels.

Étapes d'implémentation

- Création de rapports sur l'optimisation des coûts : définissez un cycle régulier pour discuter de l'efficacité de la charge de travail et pour l'analyser. À l'aide des métriques définies, rendez compte des métriques atteintes et du coût associé. Identifiez et corrigez les tendances négatives, et identifiez les tendances positives à promouvoir dans votre organisation. Les rapports doivent impliquer des représentants des équipes et des propriétaires d'application, de la finance et de la gestion.
- Créez et activez une granularité AWS Budgetsquotidienne des coûts et de l'utilisation afin de prendre des mesures rapides pour éviter tout dépassement de coûts potentiel : AWS Budgets vous pouvez configurer des notifications d'alerte, afin de rester informé si l'un de vos types de budget

dépasse les seuils préconfigurés. La meilleure façon d'en tirer parti AWS Budgets est de définir vos limites en termes de coûts et d'utilisation prévus, de sorte que tout montant supérieur à vos budgets puisse être considéré comme une dépense excessive.

- Créez AWS Cost Anomaly Detection pour surveiller les coûts : AWS Cost Anomaly Detectionutilise une technologie avancée de Machine Learning pour identifier les dépenses anormales et les causes profondes, afin que vous puissiez agir rapidement. Cela vous permet de configurer des surveillances de coûts qui définissent les segments de dépenses que vous souhaitez évaluer (par exemple, services AWS individuels, comptes membres, balises de répartition des coûts et catégories de coûts), mais aussi de définir quand, où et comment vous recevez vos notifications d'alerte. Pour chaque surveillance, attachez plusieurs abonnements à des alertes pour les propriétaires d'entreprise et les équipes technologiques, notamment un nom, un seuil d'impact du coût et une fréquence d'alerte (alertes individuelles, résumé quotidien, résumé hebdomadaire) pour chaque abonnement.
- Utilisez AWS Cost Explorer ou intégrez vos AWS Cost and Usage Report (CUR) données aux QuickSight tableaux de bord Amazon pour visualiser les coûts de votre organisation : AWS Cost Explorer possède une easy-to-use interface qui vous permet de visualiser, de comprendre et de gérer vos AWS coûts et votre utilisation au fil du temps. Le tableau de bord Cost Intelligence Dashboard est un tableau de bord personnalisable et accessible pour vous aider à créer la base de votre propre outil de gestion et d'optimisation des coûts.

Ressources

Documents connexes :

- AWS Budgets
- AWS Cost Explorer
- Budgets d'utilisation et de coûts au quotidien
- AWS Cost Anomaly Detection

Exemples connexes :

- Ateliers Well-Architected : visualisation
- Ateliers Well-Architected : visualisation avancée
- Ateliers Well-Architected : Cloud Intelligence Dashboards
- Ateliers Well-Architected : visualisation des coûts

- AWS Cost Anomaly Detection Alerte avec Slack

COST01-BP07 Tenez-vous au courant up-to-date des nouvelles versions de service

Consultez régulièrement des experts ou des AWS partenaires pour déterminer quels services et fonctionnalités sont les moins coûteux. Passez en revue AWS les blogs et autres sources d'information.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

AWS ajoute constamment de nouvelles fonctionnalités afin que vous puissiez tirer parti des dernières technologies pour expérimenter et innover plus rapidement. Vous pourriez être en mesure de mettre en œuvre de nouveaux AWS services et fonctionnalités afin d'améliorer la rentabilité de votre charge de travail. Consultez régulièrement Gestion des coûts AWS, le blog Actualités AWS, le blog Gestion des coûts AWS et Nouveautés d’ AWS pour plus d’informations sur les nouvelles versions de service et de fonctionnalités. Les articles « Nouveautés » fournissent un bref aperçu de toutes les annonces d'extension des AWS services, des fonctionnalités et des régions au fur et à mesure de leur publication.

Étapes d’implémentation

- Abonnez-vous aux blogs : Accédez aux pages AWS des blogs et abonnez-vous au blog What's New et à d'autres blogs pertinents. Vous pouvez vous inscrire sur la page de préférence de communication avec votre adresse e-mail.
- Abonnez-vous aux AWS actualités : consultez régulièrement le blog d'AWS actualités et les nouveautés AWS pour obtenir des informations sur les nouveaux services et fonctionnalités. Abonnez-vous au RSS fil d'actualité ou utilisez votre e-mail pour suivre les annonces et les publications.
- Suivez les AWS baisses de prix : Les baisses de prix régulières sur tous nos services constituent un moyen standard de AWS répercuter sur nos clients les gains d'efficacité économique réalisés grâce à notre envergure. Au 20 septembre 2023, AWS elle a réduit ses prix 134 fois depuis 2006. Si vous avez des décisions métier en attente en raison d’inquiétudes concernant les prix, vous pouvez les examiner de nouveau après les réductions de prix et l’intégration de nouveaux services. Vous pouvez en savoir plus sur les précédents efforts de réduction de prix, notamment sur les instances Amazon Elastic Compute Cloud (AmazonEC2), dans la catégorie des réductions de prix du AWS News Blog.

- AWS événements et rencontres : participez à votre AWS sommet local et à toute réunion locale avec d'autres organisations de votre région. Si vous ne pouvez pas y assister en personne, essayez d'assister à des événements virtuels pour en savoir plus sur AWS les experts et les études de cas d'autres clients.
- Rencontre avec l’équipe chargée de votre compte : planifiez un rythme régulier avec l’équipe chargée de votre compte, réunissez-vous et discutez des tendances du secteur et des services AWS . Parlez à votre gestionnaire de compte, à votre architecte de solutions et à votre équipe de support.

Ressources

Documents connexes :

- AWS Gestion des coûts
- Quoi de neuf avec AWS
- AWS Blog d'actualités

Exemples connexes :

- Amazon EC2 — 15 ans d'optimisation et de réduction de vos coûts informatiques
- AWS Blog d'actualités - Baisse de prix

COST01-BP08 Créer une culture axée sur les coûts

Mettez en œuvre des modifications ou des programmes dans toute votre entreprise afin de créer une culture de sensibilisation aux coûts. Il est recommandé de commencer petit, puis, au fur et à mesure que vos capacités augmentent et que votre organisation utilise le cloud, de mettre en œuvre des programmes de grande envergure.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d’implémentation

Une culture de sensibilisation aux coûts vous permet de mettre à l’échelle l’optimisation des coûts et la gestion financière dans le cloud (opérations financières, centre d’excellence cloud, équipes des opérations cloud, et bien plus encore) grâce aux bonnes pratiques qui sont appliquées de manière organique et décentralisée dans toute votre entreprise. La sensibilisation aux coûts permet de créer

de hauts niveaux de capacité dans toute votre organisation avec un minimum d'efforts, par rapport à une approche centralisée et descendante stricte.

La création d'une sensibilisation aux coûts dans le cloud computing, notamment pour les principaux facteurs de coût, permet aux équipes de comprendre les résultats attendus de n'importe quel changement en matière de coût. Les équipes qui accèdent aux environnements de cloud doivent connaître les modèles de tarification et la différence entre les centres de données sur site traditionnels et le cloud computing.

Le principal avantage d'une culture de sensibilisation aux coûts est que les équipes technologiques optimisent les coûts de manière proactive et continue (par exemple, ces éléments sont considérés comme une exigence non fonctionnelle lors de la création de l'architecture des nouvelles charges de travail ou de la modification de charges de travail existantes) au lieu de procéder à des optimisations de coûts réactives si nécessaire.

De petits changements de culture peuvent avoir de grandes répercussions sur l'efficacité de votre charge de travail actuelle et future. En voici quelques exemples :

- Donnez de la visibilité et créez de la sensibilisation dans les équipes ingénierie pour comprendre ce qu'elles font et leur impact en termes de coûts.
- Ludification des coûts et de l'utilisation dans votre entreprise. Cela peut se faire par le biais d'un tableau de bord visible publiquement ou d'un rapport qui compare les coûts et l'utilisation normalisés entre les équipes (par exemple, cost-per-workload et cost-per-transaction).
- Reconnaissance de la rentabilité. Récompensez les réalisations volontaires ou non sollicitées en matière d'optimisation des coûts, publiquement ou en privé, et tirez les leçons des erreurs pour éviter de les répéter à l'avenir.
- Créez des exigences organisationnelles hiérarchisées pour que les charges de travail soient exécutées selon des budgets prédéfinis.
- Questionnez les exigences métier en matière de changements, et l'impact du coût des changements demandés apportés à l'infrastructure de l'architecture ou la configuration de charge de travail, pour veiller à payer uniquement ce dont vous avez besoin.
- Veillez à ce que le planificateur de changements soit informé des changements attendus ayant un impact sur le coût, et qu'ils soient confirmés par les parties prenantes pour fournir des résultats métier de manière rentable.

Étapes d'implémentation

- Signalez les coûts du cloud aux équipes technologiques : pour mieux faire connaître les coûts et renforcer l'efficacité KPIs des acteurs financiers et commerciaux.
- Informez les parties prenantes ou les membres des équipes des changements planifiés : créez un point à l'ordre du jour pour discuter des changements planifiés et de l'impact coût-avantage sur la charge de travail lors des réunions hebdomadaires sur les changements.
- Rencontrez l'équipe de gestion de votre compte : établissez une cadence régulière de réunions avec l'équipe chargée de votre compte et discutez des tendances du secteur et des services AWS . Parlez à votre gestionnaire de compte, à votre architecte et à votre équipe de support.
- Partagez des histoires de réussite : partagez des histoires de réussite en matière de réduction des coûts pour toute charge de travail ou organisation afin de créer une attitude positive et d'encourager l'optimisation des coûts. Compte AWS
- Formation : Assurez-vous que les équipes techniques ou les membres de l'équipe sont formés pour être conscients des coûts liés aux ressources AWS Cloud.
- AWS événements et rencontres : participez à des AWS sommets locaux et à toute réunion locale avec d'autres organisations de votre région.
- Abonnez-vous aux blogs : Accédez aux pages des AWS blogs et abonnez-vous au blog What's New et à d'autres blogs pertinents pour suivre les nouvelles versions, les implémentations, les exemples et les modifications partagés par AWS.

Ressources

Documents connexes :

- AWS Blog
- AWS Gestion des coûts
- AWS Blog d'actualités

Exemples connexes :

- AWS Gestion financière dans le cloud
- AWS Well-Architected Labs : gestion financière dans le cloud

COST01-BP09 Quantifier la valeur commerciale grâce à l'optimisation des coûts

La quantification de la valeur métier générée par l’optimisation des coûts permet de comprendre l’ensemble des avantages pour votre entreprise. Parce que l’optimisation des coûts est un investissement nécessaire, la quantification de la valeur ajoutée vous permet d’expliquer le retour sur investissement aux parties prenantes. La quantification de la valeur ajoutée peut vous aider à obtenir une meilleure adhésion des parties prenantes aux investissements futurs en matière d’optimisation des coûts, et fournit un cadre pour mesurer les résultats des activités d’optimisation des coûts de votre organisation.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Quantifier la valeur métier signifie mesurer le bénéfice que les entreprises retirent des actions et des décisions qu’elles prennent. La valeur métier peut être tangible (comme la réduction des dépenses ou l’augmentation des profits) ou intangible (comme l’amélioration de la réputation de la marque ou l’augmentation de la satisfaction client).

Quantifier la valeur métier résultant de l’optimisation des coûts signifie déterminer la valeur ou le bénéfice que vous retirez de vos efforts pour dépenser plus efficacement. Par exemple, si une entreprise dépense 100 000 dollars pour déployer une charge de travail AWS et l'optimise ultérieurement, le nouveau coût passe à seulement 80 000 dollars sans pour autant sacrifier la qualité ou le rendement. Dans ce scénario, la valeur métier quantifiée résultant de l’optimisation des coûts représenterait une économie de 20 000 USD. Mais au-delà des simples économies, l’entreprise peut également quantifier la valeur en termes de rapidité de livraison, d’amélioration de la satisfaction client ou d’autres indicateurs résultant des efforts d’optimisation des coûts. Les parties prenantes doivent prendre des décisions concernant la valeur potentielle de l’optimisation des coûts, le coût de l’optimisation de la charge de travail et la valeur de retour.

En plus de faire état des économies réalisées grâce à l’optimisation des coûts, il est recommandé de quantifier la valeur supplémentaire générée. Les avantages de l’optimisation des coûts sont généralement quantifiés en termes de réduction des coûts par résultat commercial. Par exemple, vous pouvez quantifier les économies Amazon Elastic Compute Cloud(AmazonEC2) lorsque vous achetez des Savings Plans, qui réduisent les coûts et maintiennent les niveaux de production de la charge de travail. Vous pouvez quantifier les réductions de coûts AWS liées à la suppression des EC2 instances Amazon inactives ou à la suppression de volumes Amazon Elastic Block Store EBS (Amazon) indépendants.

Les avantages de l’optimisation des coûts vont toutefois au-delà de la réduction ou de l’évitement des coûts. Envisagez de capturer des données supplémentaires pour mesurer les améliorations de l’efficacité et la valeur ajoutée.

Étapes d’implémentation

- Évaluer les avantages commerciaux : Il s'agit du processus d'analyse et d'ajustement des AWS Cloud coûts de manière à maximiser les avantages tirés de chaque dollar dépensé. Au lieu de vous concentrer sur la réduction des coûts sans valeur métier, tenez compte des avantages commerciaux et du retour sur investissement de l’optimisation des coûts, ce qui peut rentabiliser davantage l’argent que vous dépensez. Il s’agit de dépenser judicieusement et de réaliser des investissements et des dépenses dans les secteurs qui génèrent le meilleur retour.
- Analysez les AWS coûts prévisionnels : les prévisions aident les parties prenantes du secteur financier à définir leurs attentes avec les autres parties prenantes internes et externes de l'organisation, et peuvent améliorer la prévisibilité financière de votre organisation. AWS Cost Explorerpeut être utilisé pour effectuer des prévisions de vos coûts et de votre utilisation.

Ressources

Documents connexes :

- AWS Cloud Économie
- AWS Blog
- AWS Gestion des coûts
- AWS Blog d'actualités
- Livre blanc du pilier Fiabilité de Well-Architected
- AWS Cost Explorer

Vidéos connexes :

- Tirez parti de la valeur commerciale avec Windows activé AWS

Exemples connexes :

- Mesure et maximisation de la valeur commerciale du client 360
- Valeur commerciale de l’adoption des bases de données gérées par Amazon Web Services

- Valeur commerciale d'Amazon Web Services pour les éditeurs de logiciels indépendants
- Valeur commerciale de la modernisation du cloud
- Valeur commerciale de la migration vers Amazon Web Services

Sensibilisation aux dépenses et à l'utilisation

Questions

- COÛT 2. Comment gérer l'utilisation ?
- COÛT 3. Comment surveillez-vous vos coûts et votre utilisation ?
- COÛT 4. Comment mettez-vous les ressources hors service ?

COÛT 2. Comment gérer l'utilisation ?

Définissez des stratégies et des mécanismes pour vous assurer que les coûts appropriés sont facturés tout en atteignant les objectifs. En adoptant une approche d'équilibre des pouvoirs, vous pouvez innover sans dépense excessive.

Bonnes pratiques

- COST02-BP01 Élaborez des politiques basées sur les exigences de votre organisation
- COST02-BP02 Mettre en œuvre les objectifs et les cibles
- COST02-BP03 Implémenter une structure de compte
- COST02-BP04 Implémenter des groupes et des rôles
- COST02-BP05 Mettre en œuvre le contrôle des coûts
- COST02-BP06 Suivez le cycle de vie du projet

COST02-BP01 Élaborez des politiques basées sur les exigences de votre organisation

Développez des stratégies qui définissent la manière dont les ressources sont gérées par votre organisation et inspectez-les régulièrement. Les stratégies doivent couvrir les aspects de coût des ressources et des charges de travail, y compris la création, la modification et la mise hors service pendant la durée de vie des ressources.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d’implémentation

Il est essentiel de comprendre les coûts et les facteurs de coûts de votre entreprise pour les gérer efficacement et identifier les possibilités de réduction. Les organisations exécutent généralement plusieurs charges de travail exécutées par plusieurs équipes. Ces équipes peuvent appartenir à différentes unités commerciales, chacune ayant ses propres sources de revenus. La possibilité d’attribuer le coût des ressources aux charges de travail, à l’organisation individuelle ou aux propriétaires de produits permet d’adopter un comportement d’utilisation efficace et contribue à réduire les pertes. La surveillance précise des coûts et de l’utilisation vous aide à comprendre dans quelle mesure une charge de travail est optimisée, ainsi que le degré de rentabilité des unités organisationnelles et des produits. Ces connaissances permettent de prendre des décisions plus éclairées quant à l’affectation des ressources au sein de votre organisation. La sensibilisation à l’utilisation à tous les niveaux de l’organisation est la clé du changement, car les changements d’utilisation entraînent des changements dans les coûts. Envisagez d’adopter une approche multidimensionnelle pour prendre conscience de votre utilisation et de vos dépenses.

La première étape de la gouvernance consiste à utiliser les exigences de votre entreprise pour élaborer des stratégies d’utilisation du cloud. Ces stratégies définissent la façon dont votre organisation utilise le cloud et dont les ressources sont gérées. Les stratégies doivent couvrir tous les aspects des ressources et des charges de travail qui ont trait au coût ou à l’utilisation, y compris la création, la modification et la mise hors service pendant la durée de vie d’une ressource. Vérifiez que les stratégies et les procédures sont suivies et mises en œuvre en cas de changement dans un environnement cloud. Lors de vos réunions sur la gestion des changements informatiques, posez des questions afin de connaître l’impact du coût des changements prévus, que ce soit une augmentation ou une baisse, la justification opérationnelle et le résultat attendu.

Les stratégies doivent être simples afin qu’elles soient aisément compréhensibles et puissent être mises en œuvre efficacement dans toute l’entreprise. Les stratégies doivent également être faciles à suivre et à interpréter (afin qu’elles soient utilisées) et être spécifiques (aucune mauvaise interprétation entre les équipes). En outre, elles doivent être inspectées périodiquement (comme nos mécanismes) et mises à jour à mesure que les conditions commerciales ou les priorités des clients évoluent, ce qui rendrait la stratégie obsolète.

Commencez par des stratégies générales de haut niveau, telles que la région géographique à utiliser ou les moments de la journée où les ressources doivent fonctionner. Affinez progressivement les stratégies des différentes unités organisationnelles et des charges de travail. Les stratégies communes comprennent les services et les fonctionnalités qui peuvent être utilisés (par exemple, un stockage moins performant dans les environnements de test et de développement), les types

de ressources qui peuvent être utilisés par différents groupes (par exemple, la plus grande taille de ressource dans un compte de développement est moyenne) et la durée d'utilisation de ces ressources (qu'elle soit temporaire, courte ou spécifique).

Exemple de stratégie

Vous trouverez ci-dessous un exemple de stratégie que vous pouvez consulter pour créer vos propres stratégies de gouvernance du cloud, axées sur l'optimisation des coûts. Assurez-vous d'ajuster la politique en fonction des exigences de votre organisation et des demandes de vos parties prenantes.

- Nom de la stratégie : définissez un nom de stratégie clair, par exemple stratégie d'optimisation des ressources et de réduction des coûts.
- Objectif : expliquez pourquoi cette stratégie doit être utilisée et quel est le résultat attendu. L'objectif de cette stratégie est de vérifier qu'il existe un coût minimum requis pour déployer et exécuter la charge de travail souhaitée afin de répondre aux exigences de l'organisation.
- Champ d'application : définissez clairement qui doit utiliser cette politique et quand elle doit être utilisée, par exemple DevOps X Team pour utiliser cette politique chez les clients de l'est des États-Unis pour l'environnement X (production ou non-production).

Déclaration de stratégie

1. Sélectionnez la région 1 de l'est des États-Unis ou plusieurs régions de l'est des États-Unis, en fonction de l'environnement de votre charge de travail et des exigences métier (développement, tests d'acceptation par les utilisateurs, préproduction ou production).
2. Planifiez l'exécution des RDS instances Amazon EC2 et Amazon entre six heures du matin et huit heures du soir (heure normale de l'Est (EST)).
3. Arrêtez toutes les EC2 instances Amazon inutilisées au bout de huit heures et les RDS instances Amazon non utilisées après 24 heures d'inactivité.
4. Mettez fin à toutes les EC2 instances Amazon non utilisées après 24 heures d'inactivité dans des environnements hors production. Rappelez au propriétaire de l'EC2instance Amazon (sur la base des balises) de vérifier ses EC2 instances Amazon arrêtées en production et de l'informer que ses EC2 instances Amazon seront résiliées dans les 72 heures si elles ne sont pas utilisées.
5. Utilisez une famille et une taille d'instance génériques telles que m5.large, puis redimensionnez l'instance en fonction de l'utilisation de la mémoire à l'CPUaide. AWS Compute Optimizer

6. Priorisez l’utilisation de l’autoscaling pour ajuster dynamiquement le nombre d’instances en cours d’exécution en fonction du trafic.
7. Utilisez des instances Spot pour les charges de travail non critiques.
8. Passez en revue les exigences en matière de capacité pour valider des Savings Plans ou des instances réservées pour des charges de travail prévisibles et informez l’équipe de gestion financière du cloud.
9. Utilisez des stratégies de cycle de vie Amazon S3 pour déplacer les données rarement consultées vers des niveaux de stockage moins coûteux. Si aucune stratégie de rétention n’est définie, utilisez Amazon S3 Intelligent-Tiering pour déplacer automatiquement les objets vers le niveau archivé.
10. Surveillez l'utilisation des ressources et définissez des alarmes pour déclencher des événements de dimensionnement à l'aide d'Amazon CloudWatch.
11. Pour chacun Compte AWS, utilisez AWS Budgets pour définir les budgets de coûts et d'utilisation de votre compte en fonction du centre de coûts et des unités commerciales.
12. Le AWS Budgets fait de définir les budgets de coûts et d'utilisation de votre compte peut vous aider à maîtriser vos dépenses et à éviter les factures imprévues, ce qui vous permet de mieux contrôler vos coûts.

Procédure : fournissez des procédures détaillées pour la mise en œuvre de cette stratégie ou consultez d’autres documents qui décrivent comment mettre en œuvre chaque déclaration de stratégie. Cette section doit fournir des step-by-step instructions pour la mise en œuvre des exigences de la politique.

Pour mettre en œuvre cette politique, vous pouvez utiliser divers outils ou AWS Config règles tiers pour vérifier la conformité avec la déclaration de politique et déclencher des actions correctives automatisées à l'aide de AWS Lambda fonctions. Vous pouvez également l'utiliser AWS Organizations pour appliquer la politique. En outre, vous devez régulièrement revoir votre utilisation des ressources et ajuster la stratégie si nécessaire pour vérifier qu’elle continue de répondre aux besoins de votre organisation.

Étapes d’implémentation

- Rencontre avec les parties prenantes : pour élaborer des stratégies, demandez aux parties prenantes (bureau commercial du cloud, ingénieurs ou décideurs fonctionnels chargés de l’application des stratégies) au sein de votre organisation de spécifier leurs exigences et de les documenter. Adoptez une approche itérative en commençant par les grandes lignes et en affinant continuellement jusqu’aux plus petites unités à chaque étape. Les membres de l’équipe incluent

ceux qui sont directement impliqués dans la charge de travail, tels que les unités d'organisation ou les propriétaires d'application, ainsi que les groupes de soutien, tels que les équipes de sécurité et les équipes financières.

- Obtention d'une confirmation : assurez-vous que les équipes s'accordent sur les stratégies décrivant qui peut accéder au AWS Cloud et y faire des déploiements. Vérifiez qu'elles suivent les stratégies de votre organisation et confirmez que leurs créations de ressources s'alignent sur les stratégies et les procédures convenues.
- Création de sessions de formation d'intégration : demandez aux nouveaux membres de l'organisation de suivre des cours de formation d'intégration afin de les sensibiliser aux coûts et aux exigences de l'organisation. Ils peuvent supposer des stratégies différentes issues de leur expérience passée ou ne pas y penser du tout.
- Définition d'emplacement pour votre charge de travail : définissez l'emplacement d'exécution de votre charge de travail, y compris le pays et la zone du pays. Ces informations sont utilisées pour le mappage Régions AWS et les zones de disponibilité.
- Définition et regroupement de services et de ressources : définissez les services requis par les charges de travail. Pour chaque service, spécifiez les types, la taille et le nombre de ressources requis. Définissez des groupes pour les ressources par fonction, tels que les serveurs d'applications ou le stockage de base de données. Les ressources peuvent appartenir à plusieurs groupes.
- Définition et regroupement des utilisateurs par fonction : définissez les utilisateurs qui interagissent avec la charge de travail, en vous concentrant sur ce qu'ils font et sur la façon dont ils l'utilisent, et non sur leur identité ou sur leur poste au sein de l'organisation. Regroupez les utilisateurs ou fonctions similaires. Vous pouvez utiliser les politiques AWS gérées comme guide.
- Définition des actions : en utilisant les emplacements, les ressources et les utilisateurs identifiés précédemment, définissez les actions requises par chacun pour atteindre les résultats de la charge de travail pendant sa durée de vie (développement, exploitation et mise hors service). Identifiez les actions en fonction des groupes, et non pas des éléments individuels des groupes, dans chaque emplacement. Commencez globalement avec la lecture ou l'écriture, puis affinez vers des actions spécifiques pour chaque service.
- Définition de la période de révision : les charges de travail et les exigences organisationnelles peuvent changer au fil du temps. Définissez le calendrier de révision de la charge de travail pour qu'il reste conforme aux priorités de l'organisation.
- Documentation des stratégies : assurez-vous que les stratégies définies sont accessibles en fonction des besoins de votre organisation. Ces stratégies sont utilisées pour mettre en œuvre, gérer et auditer l'accès à vos environnements.

Ressources

Documents connexes :

- Gestion des changements dans le cloud
- AWS Politiques gérées pour les fonctions professionnelles
- AWS stratégie de facturation pour comptes multiples
- Actions, ressources et clés de condition pour les AWS services
- AWS Gestion et gouvernance
- Contrôler l'accès aux IAM politiques Régions AWS d'utilisation
- Infrastructures mondiales, régions et AZs

Vidéos connexes :

- AWS Gestion et gouvernance à grande échelle

Exemples connexes :

- VMware- Que sont les politiques relatives au cloud ?

COST02-BP02 Mettre en œuvre les objectifs et les cibles

Mettez en œuvre des objectifs et des cibles de coût et d'utilisation de votre charge de travail. Les objectifs fournissent une orientation à votre organisation sur les résultats attendus, et les cibles fournissent des résultats mesurables spécifiques à atteindre pour vos charges de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Développez des objectifs et des cibles de coût et d'utilisation pour votre entreprise. En tant qu'entreprise en pleine croissance AWS, il est important de définir et de suivre des objectifs d'optimisation des coûts. Ces objectifs ou indicateurs de performance clés (KPIs) peuvent inclure des éléments tels que le pourcentage des dépenses à la demande ou l'adoption de certains services optimisés tels que les instances AWS Graviton ou les types de EBS volumes gp3. Fixez des objectifs mesurables et réalisables pour pouvoir plus facilement mesurer les améliorations de l'efficacité, ce qui est important pour les opérations métier. Les objectifs fournissent des conseils et des directives à votre organisation en ce qui concerne les résultats attendus.

Les cibles fournissent des résultats mesurables spécifiques à atteindre. En bref, un objectif est la direction que vous souhaitez prendre, et une cible est le chemin parcouru dans cette direction et le moment où cet objectif doit être atteint (utilisez des conseils spécifiques, mesurables, assignables, réalistes et opportuns, ouSMART). Voici un exemple d'objectif : l'utilisation de la plateforme doit augmenter de manière significative, avec seulement une augmentation mineure (non linéaire) des coûts. Voici un exemple de cible : une augmentation de 20 % de l'utilisation de la plateforme, avec une augmentation des coûts inférieure à 5 %. Voici un autre exemple d'objectif commun : les charges de travail doivent être plus efficaces tous les six mois. La cible qui correspondrait à cet objectif serait de faire en sorte que les indicateurs du coût par entreprise diminuent de 5 % tous les six mois. Utilisez les bons indicateurs et définissez des paramètres calculés KPIs pour votre organisation. Vous pouvez commencer par le basique KPIs et évoluer ultérieurement en fonction des besoins de l'entreprise.

L'un des objectifs de l'optimisation des coûts est d'accroître l'efficacité de la charge de travail, ce qui doit se traduire par une réduction du coût par résultat métier correspondant à cette charge de travail au fil du temps. Mettez en œuvre cet objectif pour toutes les charges de travail et fixez une cible telle qu'une augmentation de 5 % de l'efficacité tous les six mois à un an. Dans le cloud, vous pouvez y parvenir en mettant en place des capacités d'optimisation des coûts, ainsi qu'en lançant de nouveaux services et fonctionnalités.

Les cibles sont les points de référence quantifiables que vous souhaitez atteindre pour réaliser vos objectifs, et les points de référence comparent vos résultats réels par rapport à une cible. Établissez des KPIs points de référence en fonction du coût par unité des services informatiques (tels que l'adoption de Spot, l'adoption de Graviton, les derniers types d'instances et la couverture à la demande), des services de stockage (tels que EBS GP3 l'adoption, les EBS instantanés obsolètes et le stockage standard Amazon S3) ou de l'utilisation des services de base de données (tels que les moteurs RDS open source, l'adoption de Graviton et la couverture à la demande). Ces points de référence KPIs peuvent vous aider à vérifier que vous utilisez les AWS services de la manière la plus rentable possible.

Le tableau suivant fournit une liste de AWS mesures standard à titre de référence. Chaque organisation peut avoir des valeurs cibles différentes pour ceux-ciKPIs.

Catégorie	KPI (%)	Description
Calcul	EC2couverture d'utilisation	EC2instances (en coût ou en heures) utilisant SP+RI+Spot par rapport au total (en coût

Catégorie	KPI (%)	Description
		ou en heures) des instances EC2
Calcul	Calcul de l’utilisation de SP/RI	Heures SP ou RI utilisées par rapport au nombre total d’heures SP ou RI disponibles
Calcul	EC2/Coût horaire	EC2coût divisé par le nombre d'EC2instances exécutées au cours de cette heure
Calcul	v CPU coût	Coût par v CPU pour toutes les instances
Calcul	Dernière génération d’instances	Pourcentage d’instances sur Graviton (ou sur d’autres types d’instances de génération moderne)
Base de données	RDScouverture	RDSinstances (en coût ou en heures) utilisant RI par rapport au total (en coût ou en heures) des RDS instances
Base de données	RDSutilisation	Heures de RI utilisées par rapport au nombre total d’heures de RI disponibles
Base de données	RDSdisponibilité	RDScoût divisé par le nombre d'RDSinstances exécutées au cours de cette heure
Base de données	Dernière génération d’instances	Pourcentage d’instances sur Graviton (ou sur d’autres types d’instances modernes)

Catégorie	KPI (%)	Description
Stockage	Utilisation du stockage	Coût de stockage optimisé (par exemple Glacier, archivage approfondi ou accès peu fréquent) divisé par le coût de stockage total
Identification	Ressources non balisées	Cost Explorer 1. Filtrez les crédits, les remises, les taxes, les remboursements, la place de marché et copiez le dernier coût mensuel. 2. Sélectionnez Afficher uniquement les ressource s non balisées dans Cost Explorer. 3. Divisez le montant en ressources non balisées par votre coût mensuel.

À l'aide de ce tableau, incluez des valeurs cibles ou de référence, qui doivent être calculées en fonction des objectifs de votre organisation. Vous devez mesurer certains indicateurs pour votre entreprise et comprendre les résultats commerciaux liés à cette charge de travail afin de définir des indicateurs précis et réalistesKPIs. Lorsque vous évaluez les métriques de performance au sein d'une organisation, distinguez les différents types de métriques qui répondent à des objectifs distincts. Ces métriques mesurent principalement les performances et l'efficacité de l'infrastructure technique plutôt que l'impact commercial global. Par exemple, elles peuvent suivre les temps de réponse des serveurs, la latence du réseau ou la disponibilité du système. Ces métriques sont essentielles pour évaluer dans quelle mesure l'infrastructure soutient les opérations techniques de l'organisation. Cependant, elles ne fournissent pas d'informations directes sur les objectifs commerciaux plus généraux tels que la satisfaction des clients, la croissance des revenus ou la part de marché. Pour obtenir une compréhension complète des performances de l'entreprise, complétez ces métriques

d’efficacité par des métriques commerciales stratégiques directement corrélées aux résultats commerciaux.

Établissez une visibilité en temps quasi réel de vos KPIs opportunités d'épargne et des opportunités associées et suivez vos progrès au fil du temps. Pour commencer à définir et suivre les KPI objectifs, nous vous recommandons le KPI tableau de bord de Cloud Intelligence Dashboards (CID). Sur la base des données du rapport sur les coûts et l'utilisation (CUR), le KPI tableau de bord fournit une série de recommandations d'optimisation des coûtsKPIs, avec la possibilité de définir des objectifs personnalisés et de suivre les progrès au fil du temps.

Si vous disposez d'autres solutions pour définir et suivre les KPI objectifs, assurez-vous que ces méthodes sont adoptées par tous les acteurs de la gestion financière du cloud au sein de votre organisation.

Étapes d’implémentation

- Définition des niveaux d’utilisation attendus : pour commencer, concentrez-vous sur les niveaux d’utilisation. Collaborez avec les propriétaires d’application, les équipes marketing et les équipes stratégiques concernées afin de comprendre quels seront les niveaux d’utilisation attendus pour la charge de travail. Comment la demande des clients est-elle susceptible d’évoluer dans le temps et existe-t-il des changements potentiels dus à des augmentations saisonnières ou à des campagnes marketing ?
- Définition des ressources et des coûts de la charge de travail : une fois les niveaux d’utilisation définis, quantifiez les modifications des ressources de charge de travail nécessaires pour atteindre ces niveaux d’utilisation. Il sera peut-être nécessaire d’augmenter la taille des ressources d’un composant de la charge de travail ou leur nombre, d’accroître le transfert de données ou de remplacer les composants de la charge de travail par un service différent à un niveau spécifique. Spécifiez les coûts à chacun de ces points principaux et anticipez l’évolution des coûts en cas de modification de l’utilisation.
- Définition d’objectifs commerciaux : combinez les résultats des modifications attendues en termes d’utilisation et de coût aux modifications technologiques attendues ou aux programmes que vous exécutez, et développez des objectifs pour la charge de travail. Les objectifs doivent tenir compte de l’utilisation et des coûts, ainsi que de la relation entre les deux. Les objectifs doivent être simples, généraux et aider les personnes à comprendre les attentes de l’entreprise en termes de résultats (par exemple, s’assurer que les ressources inutilisées restent en dessous d’un certain niveau de coût). Vous n’avez pas besoin de définir d’objectifs pour chaque type de ressource inutilisé ni de définir de coûts qui peuvent entraîner des pertes pour les objectifs et les cibles. Assurez-vous qu’il existe des programmes organisationnels (par exemple la création de capacités

avec la formation et l'éducation) si des variations de coûts sont attendues sans changement dans l'utilisation.

- Définition de cibles : pour chacun des objectifs définis, spécifiez une cible mesurable. Si l'objectif est d'augmenter l'efficacité de la charge de travail, la cible doit quantifier le degré d'amélioration (généralement en matière de résultats commerciaux par dollar dépensé) et le moment où cette amélioration doit avoir lieu. Par exemple, vous pouvez vous fixer comme objectif de minimiser le gaspillage dû à un surprovisionnement. Avec cet objectif, votre cible peut être que le gaspillage dû au surprovisionnement des ressources de calcul pour le premier niveau des charges de travail de production ne dépasse pas 10 % du coût de calcul du niveau. En outre, une deuxième cible pourrait être que le gaspillage dû à un surprovisionnement des ressources de calcul dans le deuxième niveau des charges de travail de production ne dépasse pas 5 % du coût de calcul du niveau.

Ressources

Documents connexes :

- Stratégies gérées par AWS pour les activités professionnelles
- Stratégie de facturation multicompte AWS
- Contrôler l'accès aux IAM politiques Régions AWS d'utilisation
- Objectifs S.M.A.R.T.
- Comment suivre l'optimisation de vos coûts KPIs avec le CID KPI tableau de bord

Vidéos connexes :

- Ateliers Well-Architected : objectifs et cibles (niveau 100)

Exemples connexes :

- Qu'est-ce qu'une métrique unitaire ?
- Sélection d'une métrique unitaire pour soutenir votre entreprise
- Métriques unitaires en pratique : leçons apprises
- Comment les métriques unitaires aident à créer un alignement entre les fonctions commerciales
- Ateliers Well-Architected : mise hors service des ressources (objectifs et cibles)
- Ateliers Well-Architected : type de ressource, taille et nombre (objectifs et cibles)

COST02-BP03 Implémenter une structure de compte

Implémentez une structure de compte mappée sur votre organisation. Cela vous aide à répartir et à gérer les coûts dans toute votre organisation.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

AWS Organizations vous permet d'en créer plusieurs Comptes AWS , ce qui peut vous aider à gérer de manière centralisée votre environnement à mesure que vous augmentez vos charges de travail. AWS Vous pouvez modéliser la hiérarchie de votre organisation en la regroupant Comptes AWS dans une structure d'unité organisationnelle (UO) et en créant plusieurs unités Comptes AWS sous chaque UO. Pour créer une structure de compte, vous devez d'abord décider lequel de vos Comptes AWS sera le compte de gestion. Ensuite, vous pouvez créer de nouveaux comptes Comptes AWS ou sélectionner des comptes existants en tant que comptes membres en fonction de la structure de compte que vous avez conçue en suivant les meilleures pratiques en matière de comptes de gestion et les meilleures pratiques en matière de comptes de membre.

Il est recommandé de toujours lier au moins un compte membre au compte de gestion, quelle que soit la taille de votre entreprise ou l'utilisation prévue. Toutes les ressources liées aux charges de travail doivent se trouver uniquement dans les comptes membres et aucune ressource ne doit être créée dans le compte de gestion. Il n'y a pas de réponse universelle au nombre Comptes AWS que vous devriez avoir. Évaluez vos modèles opérationnels et de coûts actuels et futurs pour vous assurer que la structure de votre entreprise Comptes AWS reflète les objectifs de votre organisation. Certaines entreprises en créent plusieurs Comptes AWS pour des raisons commerciales, par exemple :

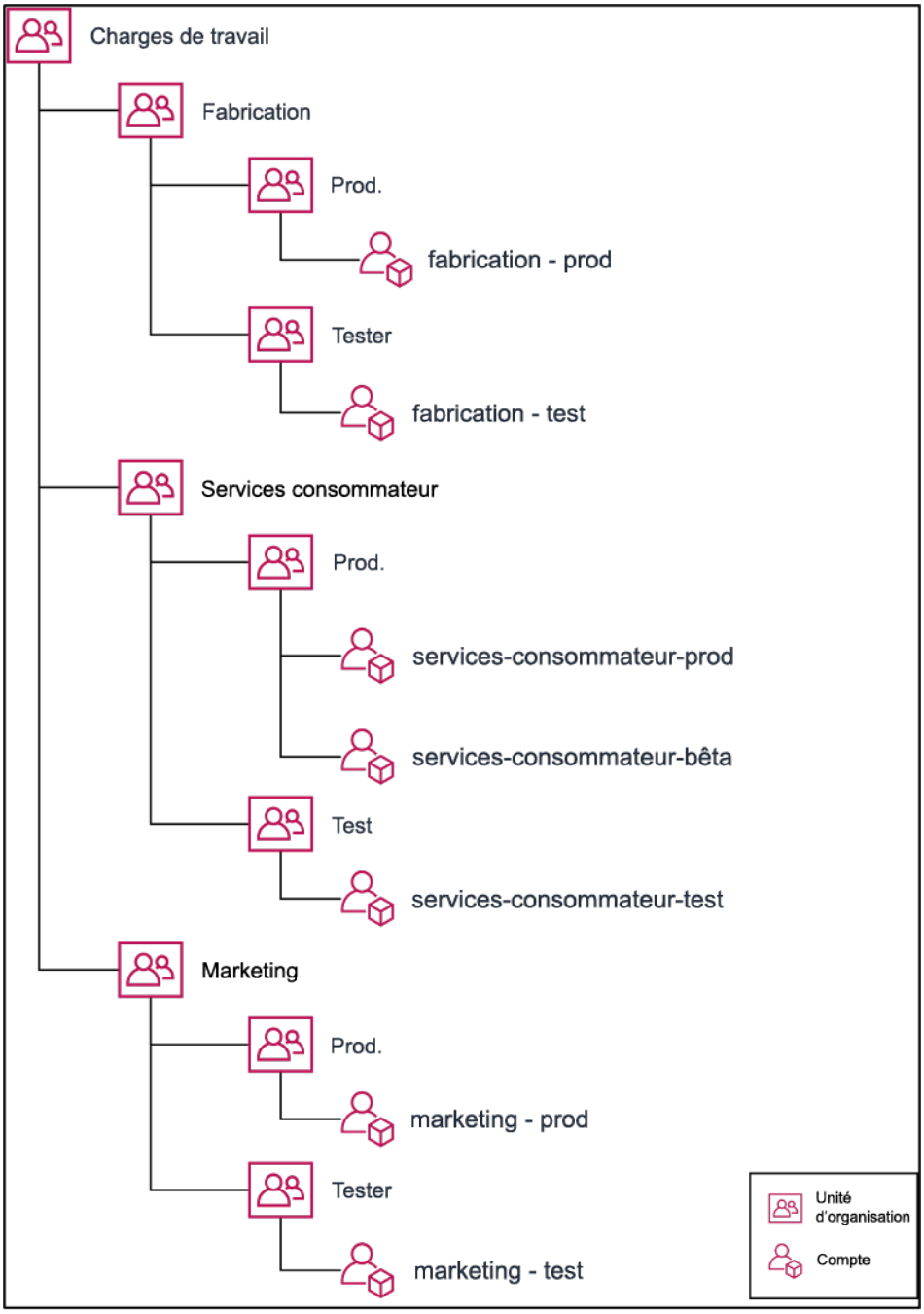
- Une isolation administrative, fiscale ou en matière de facturation est nécessaire entre les unités d'organisation, les centres de coûts ou les charges de travail spécifiques.
- AWS les limites de service sont définies pour être spécifiques à des charges de travail particulières.
- Il existe une exigence d'isolation et de séparation entre les charges de travail et les ressources.

Dans AWS Organizations, la facturation consolidée crée le lien entre un ou plusieurs comptes membres et le compte de gestion. Les comptes membres vous permettent d'isoler et de distinguer votre coût et votre utilisation par groupes. Une pratique courante consiste à avoir des comptes membres séparés pour chaque unité d'organisation (comme les finances, le marketing et les

ventes), ou pour chaque cycle de vie de l’environnement (comme le développement, les tests et la production), ou pour chaque charge de travail (charge de travail a, b et c), puis à regrouper ces comptes liés en utilisant la facturation consolidée.

La facturation consolidée vous permet de regrouper les paiements de plusieurs membres Comptes AWS sous un seul compte de gestion, tout en assurant la visibilité de l’activité de chaque compte lié. Comme les coûts et l’utilisation sont regroupés dans le compte de gestion, cela vous permet de maximiser vos réductions sur le volume de services et l’utilisation de vos remises sur engagement (Savings Plans et instances réservées) pour obtenir les remises les plus élevées.

Le schéma suivant montre comment utiliser AWS Organizations les unités organisationnelles (UO) pour regrouper plusieurs comptes et en placer plusieurs Comptes AWS sous chaque UO. Il est recommandé de l'utiliser OUs pour divers cas d'utilisation et charges de travail, ce qui fournit des modèles d'organisation des comptes.



Exemple de regroupement de Comptes AWS plusieurs unités organisationnelles.

AWS Control Towerpeut rapidement mettre en place et configurer plusieurs AWS comptes, en veillant à ce que la gouvernance soit conforme aux exigences de votre organisation.

Étapes d'implémentation

- Définition d'exigences de séparation : les exigences de séparation combinent plusieurs facteurs, notamment la sécurité, la fiabilité et les structures financières. Examinez chaque facteur dans

l'ordre et précisez si la charge de travail ou son environnement doivent être séparés des autres charges de travail. La sécurité favorise le respect des exigences en matière d'accès et de données. La fiabilité gère les limites afin que les environnements et les charges de travail n'affectent pas les autres. Examinez périodiquement les piliers de sécurité et de fiabilité du cadre Well-Architected et suivez les bonnes pratiques fournies. Les structures financières créent une séparation financière stricte (pour les multiples centres de coûts, et les différentes responsabilités et propriétés liées aux charges de travail). Parmi les exemples courants de séparation, citons : les charges de travail de production et de test exécutées dans des comptes distincts ou l'utilisation d'un compte distinct afin que les données de facture et de facturation soient fournies aux unités commerciales, aux services individuels au sein de l'organisation ou à la partie prenante qui détient le compte.

- Définition d'exigences de regroupement : les exigences de regroupement ne remplacent pas les exigences de séparation, mais sont utilisées pour faciliter la gestion. Regroupez les environnements ou les charges de travail similaires qui ne nécessitent pas de séparation. Par exemple, regroupez plusieurs environnements de test ou de développement d'une ou de plusieurs charges de travail.
- Définition d'une structure de compte : à l'aide de ces séparations et regroupements, spécifiez un compte pour chaque groupe et maintenez les exigences de séparation. Ces comptes sont vos comptes membres ou liés. En regroupant ces comptes membres au sein d'un seul compte de gestion ou compte payeur, vous rassemblez les données d'utilisation, ce qui vous permet d'obtenir des remises plus importantes sur le volume. Cela génère une seule facture pour tous les comptes. Il est possible de séparer les données de facturation et de fournir à chaque compte membre une vue individuelle de ses données de facturation. Si les données d'utilisation ou de facturation d'un compte membre ne doivent être visibles par aucun autre compte, ou si un formulaire de facture distinct AWS est requis, définissez plusieurs comptes de gestion ou de paiement. Dans ce cas, chaque compte membre possède son propre compte de gestion ou compte payeur. Les ressources doivent toujours être placées dans des comptes membres ou comptes liés. Les comptes de gestion ou comptes payeurs doivent être uniquement utilisés pour la gestion.

Ressources

Documents connexes :

- Utilisation des balises de répartition des coûts
- Stratégies gérées par AWS pour les fonctions de tâches
- Stratégie de facturation multicompte AWS
- Contrôler l'accès aux IAM politiques Régions AWS d'utilisation

- AWS Control Tower
- AWS Organizations
- Bonnes pratiques pour les comptes de gestion et les comptes membres
- Organisation de votre AWS environnement à l'aide de plusieurs comptes
- Activation des remises sur les Savings Plans et sur les instances réservées partagées
- Facturation consolidée
- Facturation consolidée

Exemples connexes :

- Séparer CUR et partager l'accès

Vidéos connexes :

- Présentant AWS Organizations
- Configurez un AWS environnement multi-comptes qui utilise les meilleures pratiques pour AWS Organizations

Exemples connexes :

- Well-Architected Labs : créer AWS une organisation (niveau 100)
- Séparer AWS Cost and Usage Report et partager l'accès
- Définition d'une stratégie AWS multi-comptes pour les entreprises de télécommunications
- Bonnes pratiques d'optimisation Comptes AWS
- Bonnes pratiques pour les unités organisationnelles avec AWS Organizations

COST02-BP04 Implémenter des groupes et des rôles

Mettez en œuvre des groupes et des rôles conformes à vos stratégies et qui indiquent qui crée, modifie ou met hors service des instances et des ressources dans chaque groupe. Par exemple, mettez en place des groupes de développement, de test et de production. Cela s'applique aux AWS services et aux solutions tierces.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d’implémentation

Les rôles et les groupes d’utilisateurs sont des éléments fondamentaux de la conception et de l’implémentation de systèmes sécurisés et efficaces. Les rôles et les groupes aident les organisations à trouver un équilibre entre le besoin de contrôle et le besoin de flexibilité et de productivité, et répondent ainsi aux objectifs de l’organisation et aux besoins des utilisateurs. Comme recommandé dans la section Gestion des identités et des accès de AWS Well-Architected Framework Security Pillar, vous avez besoin d'une gestion des identités robuste et d'autorisations en place pour fournir un accès aux bonnes ressources aux bonnes personnes dans les bonnes conditions. Les utilisateurs disposent uniquement de l’accès dont ils ont besoin pour effectuer leurs tâches. Le risque d’accès non autorisé ou d’utilisation abusive s’en trouve ainsi réduit.

Après avoir élaboré des stratégies, vous pouvez créer des groupes logiques et des rôles d’utilisateurs au sein de votre organisation. Vous pouvez alors attribuer des autorisations, contrôler les utilisations et mettre en œuvre des mécanismes robustes de contrôle des accès pour empêcher tout accès non autorisé à des informations sensibles. Commencez par des groupes de personnes de haut niveau. Ils correspondent généralement à des unités organisationnelles et à des fonctions (par exemple, administrateur système au sein du service informatique, contrôleur financier ou analyste commercial). Les groupes classent par catégories les personnes qui effectuent des tâches similaires et ont besoin d’un accès similaire. Les rôles définissent ce qu’un groupe doit faire. Il est plus facile de gérer les autorisations pour les groupes et les rôles que pour les utilisateurs individuels. Les rôles et les groupes attribuent des autorisations de manière cohérente et systématique à tous les utilisateurs, ce qui permet d’éviter les erreurs et les incohérences.

Lorsqu’un utilisateur voit son rôle changer, les administrateurs peuvent modifier son accès au niveau du rôle ou du groupe au lieu de reconfigurer des comptes d’utilisateur individuels. Par exemple, un administrateur de système du service informatique a besoin d’un accès pour créer toutes les ressources, mais un membre de l’équipe d’analytique n’a besoin que de créer des ressources d’analytique.

Étapes d’implémentation

- Mise en œuvre de groupes : en utilisant les groupes d’utilisateurs définis dans vos politiques organisationnelles, mettez en œuvre les groupes correspondants, si nécessaire. Pour connaître les meilleures pratiques relatives aux utilisateurs, aux groupes et à l'authentification, consultez le pilier de sécurité du AWS Well-Architected Framework.
- Mise en œuvre de rôles et de stratégies : à l’aide des actions définies dans vos politiques organisationnelles, créez les rôles et stratégies d’accès requis. Pour connaître les meilleures

pratiques relatives aux rôles et aux politiques, consultez le pilier de sécurité du AWS Well-Architected Framework.

Ressources

Documents connexes :

- Stratégies gérées par AWS pour les activités professionnelles
- Stratégie de facturation multicompte AWS
- AWS Pilier de sécurité du framework Well-Architected
- AWS Identity and Access Management (IAM)
- AWS Identity and Access Management politiques

Vidéos connexes :

- Pourquoi utiliser la gestion de l’identité et des accès

Exemples connexes :

- Identité et accès de base de l’atelier Well-Architected
- Contrôler l'accès aux IAM politiques Régions AWS d'utilisation
- Démarrage de la transition vers la gestion financière dans le cloud : opérations liées aux coûts du cloud

COST02-BP05 Mettre en œuvre le contrôle des coûts

Mettez en œuvre des contrôles reposant sur des politiques organisationnelles et les groupes et rôles définis. Il s’agit de s’assurer que les coûts encourus sont toujours conformes aux exigences de l’organisation, notamment en termes de contrôle d’accès aux régions ou aux types de ressources.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

En matière de contrôle des coûts, la première étape consiste souvent à configurer l’envoi de notifications lorsque des événements liés aux coûts ou à l’utilisation sortant du cadre des stratégies

en vigueur se produisent. Vous pouvez agir rapidement et vérifier si une action corrective est nécessaire, sans restreindre ni affecter négativement les charges de travail ou la nouvelle activité. Une fois que vous connaissez les limites de charge de travail et d'environnement, vous pouvez appliquer la gouvernance. AWS Budgetsvous permet de définir des notifications et de définir des budgets mensuels pour vos AWS coûts, votre utilisation et vos remises d'engagement (Savings Plans et instances réservées). Vous pouvez créer des budgets à un niveau de coût global (intégralité des coûts, par exemple) ou plus précis, si vous n'incluez que les dimensions spécifiques pertinentes, telles que les comptes liés, les services, les balises ou les zones de disponibilité.

Une fois que vous avez défini vos limites budgétaires AWS Budgets, utilisez-les AWS Cost Anomaly Detectionpour réduire vos coûts imprévus. AWS Cost Anomaly Detection est un service de gestion des coûts qui utilise l'apprentissage automatique pour surveiller en permanence vos coûts et votre utilisation afin de détecter les dépenses inhabituelles. Il vous permet d'identifier les dépenses anormales et les causes profondes, afin que vous puissiez prendre des mesures rapidement. Créez d'abord un outil de surveillance des coûts dans AWS Cost Anomaly Detection, puis choisissez votre préférence en matière d'alerte en définissant un seuil en dollars (par exemple, une alerte en cas d'anomalie ayant un impact supérieur à 1 000$). Une fois les alertes reçues, vous pouvez analyser la cause profonde de l'anomalie et son impact sur vos coûts. Vous pouvez également surveiller et analyser les anomalies dans AWS Cost Explorer.

Appliquez les politiques de gouvernance dans les politiques de contrôle AWS Identity and Access Managementdes AWS traversées et des AWS Organizations services (SCP). IAMvous permet de gérer en toute sécurité l'accès aux AWS services et aux ressources. Vous pouvez ainsi contrôler qui peut créer ou gérer les AWS ressources, le type de ressources qui peuvent être créées et l'endroit où elles peuvent être créées. IAM Cela réduit au minimum les risques que des ressources soient créées en dehors du cadre de la politique définie. Utilisez les rôles et les groupes créés précédemment et attribuez IAMdes politiques pour garantir une utilisation correcte. SCPpermet de contrôler de manière centralisée les autorisations maximales disponibles pour tous les comptes de votre organisation, en veillant à ce que vos comptes respectent vos directives de contrôle d'accès. SCPsne sont disponibles que dans une organisation dont toutes les fonctionnalités sont activées, et vous pouvez les configurer SCPs pour refuser ou autoriser les actions pour les comptes membres par défaut. Pour en savoir plus sur la mise en œuvre de la gestion des accès, reportez-vous au livre blanc sur le pilier de sécurité Well-Architected.

La gouvernance peut également être mise en œuvre grâce à la gestion des quotas de services AWS. En vous assurant que les quotas de service sont fixés avec un coût minimum et gérés avec précision, vous pouvez minimiser la création de ressources en dehors du cadre des exigences de votre organisation. Pour ce faire, vous devez comprendre à quel point vos exigences peuvent

rapidement changer, appréhender les projets en cours (tant la création que la mise hors service des ressources) et tenir compte de l'accélération des délais de mise en œuvre de ces quotas. Les Service Quotas peuvent être utilisés pour augmenter vos quotas, si nécessaire.

Étapes d'implémentation

- Implémentation de notifications sur les dépenses : à l'aide des stratégies définies par votre organisation, créez un système AWS Budgets qui vous avertira lorsque les dépenses ne seront pas conformes à vos stratégies. Configurez plusieurs budgets de coûts, un pour chaque compte, afin d'être averti des dépenses globales du compte. Configurez des budgets de coûts supplémentaires dans chaque compte pour les plus petites unités du compte. Ces unités varient en fonction de la structure de votre compte. Parmi les exemples courants Régions AWS, citons les charges de travail (à l'aide de balises) ou les AWS services. Configurez une liste de distribution comme destinataire des notifications au lieu d'utiliser le compte de messagerie d'un individu. Vous pouvez définir un budget réel en cas de dépassement d'un montant ou utiliser un budget prévisionnel pour notifier l'utilisation prévue. Vous pouvez également préconfigurer des actions AWS budgétaires qui peuvent appliquer des SCP politiques IAM ou des politiques spécifiques, ou arrêter des RDS instances Amazon EC2 ou Amazon cibles. Les actions Budget peuvent être lancées automatiquement ou nécessiter l'approbation du flux de travail.
- Implémentation de notifications en cas de dépenses anormales : utilisez AWS Cost Anomaly Detection pour réduire les coûts imprévus dans votre organisation et analyser la cause première des dépenses anormales potentielles. Une fois que vous avez créé un moniteur des coûts pour identifier les dépenses inhabituelles à la granularité spécifiée et que vous avez configuré les notifications AWS Cost Anomaly Detection, il vous envoie une alerte lorsque des dépenses inhabituelles sont détectées. Cela vous permettra d'analyser la cause première de l'anomalie et de comprendre l'impact sur votre coût. Utilisez AWS Cost Categories lors de la configuration AWS Cost Anomaly Detection pour identifier l'équipe de projet ou l'équipe de l'unité commerciale qui peut analyser la cause première des coûts imprévus et prendre les mesures nécessaires en temps opportun.
- Mettez en œuvre des contrôles d'utilisation : à l'aide des politiques d'organisation que vous avez définies, implémentez des IAM politiques et des rôles pour spécifier les actions que les utilisateurs peuvent effectuer et celles qu'ils ne peuvent pas effectuer. Plusieurs politiques organisationnelles peuvent être incluses dans une AWS politique. De la même manière que vous avez défini les stratégies, commencez de manière générale et appliquez ensuite des contrôles plus fins à chaque étape. Les limites de service constituent également un contrôle efficace de l'utilisation. Mettez en œuvre les limites de service correctes sur tous vos comptes.

Ressources

Documents connexes :

- Stratégies gérées par AWS pour les activités professionnelles
- Stratégie de facturation multicompte AWS
- Contrôler l'accès aux IAM politiques Régions AWS d'utilisation
- AWS Budgets
- AWS Cost Anomaly Detection
- Maîtrisez vos AWS coûts

Vidéos connexes :

- Comment puis-je AWS Budgets suivre mes dépenses et mon utilisation

Exemples connexes :

- Exemples de politiques de gestion des IAM accès
- Exemples de stratégies de contrôle des services
- AWS Mesures relatives aux budgets
- Créez une IAM politique pour contrôler l'accès aux EC2 ressources Amazon à l'aide de balises
- Restreindre l'accès d'IAMIdentity à des EC2 ressources Amazon spécifiques
- Créez une IAM politique pour restreindre l'EC2utilisation d'Amazon par famille
- Ateliers Well-Architected : gouvernance de l’utilisation des coûts et de l’utilisation (niveau 100)
- Ateliers Well-Architected : gouvernance de l’utilisation des coûts et de l’utilisation (niveau 200)
- Intégrations Slack pour la détection des anomalies de coûts à l'aide de AWS Chatbot

COST02-BP06 Suivez le cycle de vie du projet

Suivez, mesurez et auditez le cycle de vie des projets, des équipes et des environnements pour éviter d’utiliser et de payer des ressources superflues.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d’implémentation

En suivant efficacement le cycle de vie des projets, les organisations peuvent mieux contrôler les coûts grâce à une amélioration de la planification, de la gestion et de l’optimisation des ressources. Les informations obtenues dans le cadre du suivi sont précieuses pour prendre des décisions éclairées qui contribuent à la rentabilité et à la réussite globale du projet.

Le suivi du cycle de vie complet de la charge de travail vous aide à comprendre quand les charges de travail ou leurs composants ne sont plus nécessaires. Les charges de travail et les composants existants peuvent sembler utilisés, mais lors de la sortie de AWS nouveaux services ou fonctionnalités, ils peuvent être mis hors service ou adoptés. Consultez les précédentes étapes des charges de travail. Une fois qu’une charge de travail est en production, les environnements précédents peuvent être mis hors service ou leur capacité fortement réduite jusqu’à ce qu’ils soient de nouveau requis.

Vous pouvez associer des ressources à un calendrier ou à un rappel pour indiquer l’heure à laquelle la charge de travail a été examinée. Par exemple, si l’environnement de développement a été vérifié pour la dernière fois il y a des mois, il peut être opportun de le vérifier à nouveau afin de déterminer si de nouveaux services peuvent être adoptés ou si l’environnement est utilisé. Vous pouvez regrouper et étiqueter vos applications avec myApplicationson AWS pour gérer et suivre les métadonnées telles que la criticité, l'environnement, la dernière révision et le centre de coûts. Vous pouvez à la fois suivre le cycle de vie de votre charge de travail et surveiller et gérer le coût, l’état, le niveau de sécurité et les performances de vos applications.

AWS fournit divers services de gestion et de gouvernance que vous pouvez utiliser pour le suivi du cycle de vie des entités. Vous pouvez utiliser AWS Confignotre AWS Systems Manager pour fournir un inventaire détaillé de vos AWS ressources et de votre configuration. Il est recommandé de l’intégrer à vos systèmes de gestion de projets ou ressources existants pour assurer le suivi des projets et produits actifs au sein de votre organisation. La combinaison de votre système actuel avec le riche ensemble d'événements et de mesures fournis par celui-ci vous AWS permet de créer une vue des événements importants du cycle de vie et de gérer les ressources de manière proactive afin de réduire les coûts inutiles.

À l'instar de la gestion du cycle de vie des applications (ALM), le suivi du cycle de vie des projets doit impliquer la collaboration de plusieurs processus, outils et équipes, tels que la conception et le développement, les tests, la production, le support et la redondance de la charge de travail.

En surveillant attentivement chaque phase du cycle de vie d’un projet, les organisations obtiennent des informations cruciales et un meilleur contrôle, ce qui facilite la planification, la mise en œuvre et

la réalisation des projets. Cette surveillance attentive permet de vérifier que les projets répondent non seulement aux normes de qualité, mais également qu'ils sont livrés dans les délais et dans les limites du budget, ce qui favorise la rentabilité globale.

Pour plus d'informations sur la mise en œuvre du suivi du cycle de vie des entités, consultez le livre blanc AWS Well-Architected – Pilier Excellence opérationnelle.

Étapes d'implémentation

- Établissement d'un processus de surveillance du cycle de vie du projet : l'équipe du centre d'excellence du cloud doit établir un processus de surveillance du cycle de vie des projets. Établissez une approche structurée et systématique pour surveiller les charges de travail afin d'améliorer le contrôle, la visibilité et les performances des projets. Rendez le processus de suivi transparent, collaboratif et axé sur l'amélioration continue afin d'en maximiser l'efficacité et la valeur.
- Réalisation d'examens de la charge de travail : comme défini par les stratégies de votre organisation, configurez une cadence régulière pour auditer vos projets existants et effectuer des examens de la charge de travail. Le niveau d'effort consacré à l'audit doit être proportionnel au risque, à la valeur ou au coût approximatif pour l'organisation. Les principaux domaines à inclure dans l'audit sont le risque pour l'organisation d'un incident ou d'une panne, la valeur ou la contribution à l'organisation (mesurée en termes de chiffre d'affaires ou de réputation de la marque), le coût de la charge de travail (mesuré en tant que coût total des ressources et coûts opérationnels) et l'utilisation de la charge de travail (mesurée en nombre de résultats de l'organisation par unité de temps). Si ces domaines changent au cours du cycle de vie, des ajustements de la charge de travail sont nécessaires, tels que la mise hors service complète ou partielle.

Ressources

Documents connexes :

- Conseils pour le marquage AWS
- Qu'est-ce que ALM (la gestion du cycle de vie des applications) ?
- Stratégies gérées par AWS pour les activités professionnelles

Exemples connexes :

- Contrôler l'accès aux IAM politiques Régions AWS d'utilisation

Outils associés

- AWS Config
- AWS Systems Manager
- AWS Budgets
- AWS Organizations
- AWS CloudFormation

COÛT 3. Comment surveillez-vous vos coûts et votre utilisation ?

Définissez des stratégies et des procédures pour surveiller et allouer vos coûts de manière appropriée. Cela vous permet d'évaluer et d'améliorer la rentabilité de cette charge de travail.

Bonnes pratiques

- COST03-BP01 Configurer des sources d'informations détaillées
- COST03-BP02 Ajouter des informations sur l'organisation aux coûts et à l'utilisation
- COST03-BP03 Identifier les catégories d'attribution des coûts
- COST03-BP04 Établir les paramètres de l'organisation
- COST03-BP05 Configuration des outils de facturation et de gestion des coûts
- COST03-BP06 Allouer les coûts en fonction des métriques de charge de travail

COST03-BP01 Configurer des sources d'informations détaillées

Configurez des outils de gestion des coûts et de reporting pour améliorer l'analyse et la transparence des données sur les coûts et l'utilisation. Configurez votre charge de travail pour créer des entrées de journal qui facilitent le suivi et la séparation des coûts et de l'utilisation.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Des informations de facturation détaillées telles que la granularité horaire dans les outils de gestion des coûts permettent aux organisations de suivre leurs consommations plus en détail et les aident à identifier certaines des raisons de l'augmentation des coûts. Ces sources de données offrent la vue la plus précise des coûts et de l'utilisation dans l'ensemble de votre organisation.

Vous pouvez l'utiliser Exportations de données AWS pour créer des exportations du AWS Cost and Usage Report (CUR) 2.0. Il s'agit de la nouvelle méthode recommandée pour recevoir vos données détaillées sur les coûts et l'utilisation auprès de AWS. Il fournit la granularité d'utilisation quotidienne ou horaire, les taux, les coûts et les attributs d'utilisation pour tous les AWS services payants (les mêmes informations queCUR), ainsi que certaines améliorations. Toutes les dimensions possibles sont incluses, CUR telles que le balisage, l'emplacement, les attributs des ressources et le compteIDs.

Il existe trois types d'exportation en fonction du type d'exportation que vous souhaitez créer : une exportation de données standard, une exportation vers un tableau de bord des coûts et de l'utilisation avec QuickSight intégration Amazon, ou une exportation de données existante.

- Exportation de données standard : exportation personnalisée d’une table qui est envoyée à Amazon S3 de manière récurrente.
- Tableau de bord des coûts et de l'utilisation : exportation et intégration QuickSight vers Amazon pour déployer un tableau de bord prédéfini sur les coûts et l'utilisation.
- Exportation de données héritées : exportation de l'ancienne AWS Cost and Usage Report (CUR).

Vous pouvez créer des exportations de données avec les personnalisations suivantes :

- Inclure la ressource IDs
- Fractionnement des données de répartition des coûts
- Granularité horaire
- Gestion des versions
- Type de compression et format de fichier

Pour vos charges de travail qui exécutent des conteneurs sur Amazon ECS ou AmazonEKS, activez les données de répartition des coûts afin de pouvoir répartir les coûts des conteneurs entre les unités commerciales et les équipes individuelles, en fonction de la manière dont vos charges de travail de conteneurs consomment les ressources de calcul et de mémoire partagées. Les données de répartition des coûts fractionnées introduisent des données de coût et d'utilisation pour les nouvelles ressources au niveau du conteneur. AWS Cost and Usage Report Les données de répartition des coûts sont calculées en calculant le coût des différents ECS services et tâches exécutés sur le cluster.

Un tableau de bord des coûts et de l'utilisation exporte régulièrement le tableau du tableau de bord des coûts et de l'utilisation vers un compartiment S3 et déploie un tableau de bord prédéfini sur les coûts et l'utilisation sur Amazon. QuickSight Utilisez cette option si vous souhaitez déployer rapidement un tableau de bord de vos données de coût et d'utilisation sans la possibilité de personnalisation.

Si vous le souhaitez, vous pouvez toujours exporter CUR en mode traditionnel, où vous pouvez intégrer d'autres services de traitement, tels que AWS Gluela préparation des données pour l'analyse et l'exécution d'analyses de données avec Amazon Athena en utilisant SQL pour interroger les données.

Étapes d'implémentation

- Création d'exportations de données : créez des exportations personnalisées avec les données souhaitées et contrôlez le schéma de vos exportations. Créez des exportations de données de facturation et de gestion des coûts à l'aide de BasicSQL, et visualisez vos données de facturation et de gestion des coûts en les intégrant à Amazon QuickSight. Vous pouvez également exporter vos données en mode standard pour les analyser avec d'autres outils de traitement, tels qu'Amazon Athena.
- Configuration du rapport sur les coûts et l'utilisation : à l'aide de la console de facturation, configurez au moins un rapport sur les coûts et l'utilisation. Configurez un rapport avec une granularité horaire qui inclut tous les identifiants et toutes les ressources. IDs Vous pouvez également créer d'autres rapports avec différentes granularités pour fournir des informations récapitulatives générales.
- Configuration de la granularité horaire dans Cost Explorer : pour accéder aux données sur le coût et l'utilisation avec une granularité horaire au cours des 14 derniers jours, pensez à activer les données horaires et au niveau des ressources dans la console de facturation.
- Configuration de la journalisation de l'application : vérifiez que votre application journalise chaque résultat opérationnel qu'elle produit afin de le suivre et le mesurer. Veillez à ce que la granularité de ces données soit au moins horaire pour être mise en correspondance avec les données de coût et d'utilisation. Pour plus de détails sur la journalisation et la surveillance, voir Pilier d'excellence opérationnelle Well-Architected.

Ressources

Documents connexes :

- Exportations de données AWS

- AWS Glue
- Amazon QuickSight
- Tarification de la gestion des coûts AWS
- Balisage de ressources AWS
- Analyse des coûts à l'aide de Cost Explorer
- Gestion de systèmes AWS Cost and Usage Report
- Pilier d'excellence opérationnelle Well-Architected

Exemples connexes :

- Configuration de compte AWS
- Exportations de données pour AWS Billing and Cost Management
- AWS Cost Explorer Cas d'utilisation courants

COST03-BP02 Ajouter des informations sur l'organisation aux coûts et à l'utilisation

Définissez un schéma de balisage en fonction de votre organisation, des attributs de la charge de travail et des catégories de répartition des coûts afin de pouvoir filtrer et rechercher des ressources ou surveiller les coûts et l'utilisation dans les outils de gestion des coûts. Mettez en œuvre un balisage cohérent sur toutes les ressources, dans la mesure du possible, par objectif, équipe, environnement ou tout autre critère pertinent pour votre entreprise.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Implémentez le balisage dans AWS pour ajouter des informations organisationnelles à vos ressources, qui seront ensuite ajoutées à vos informations de coûts et d'utilisation. Une balise est une paire clé-valeur. La clé est définie et doit être unique dans votre organisation, et la valeur est unique à un groupe de ressources. Voici un exemple de paire clé-valeur : la clé est `Environment`, avec une valeur `Production`. Toutes les ressources de l'environnement de production auront cette paire clé-valeur. Le balisage permet de catégoriser et de suivre vos coûts à l'aide d'informations significatives pertinentes sur l'organisation. Vous pouvez appliquer des balises qui représentent des catégories d'organisations (telles que les centres de coûts, les noms d'application, les projets ou les propriétaires), et identifier les charges de travail et leurs caractéristiques (telles que les tests ou la production) pour attribuer vos coûts et votre utilisation dans toute votre organisation.

Lorsque vous appliquez des balises à vos AWS ressources (telles que des Amazon Elastic Compute Cloud instances ou des Amazon Simple Storage Service compartiments) et que vous activez les balises, vous ajoutez ces AWS informations à vos rapports sur les coûts et l'utilisation. Vous pouvez exécuter des rapports et effectuer des analyses sur les ressources balisées et non balisées pour permettre une meilleure conformité avec les politiques internes de gestion des coûts et assurer une attribution précise.

La création et la mise en œuvre d'une norme de AWS balisage pour les comptes de votre entreprise vous aident à gérer et à gouverner vos AWS environnements de manière cohérente et uniforme. Utilisez les politiques relatives AWS Organizations aux balises pour définir les règles relatives à la manière dont les balises peuvent être utilisées sur les AWS ressources de vos comptes dans AWS Organizations. Les politiques de balises vous permettent d'adopter facilement une approche standardisée pour le balisage des ressources AWS

AWS L'éditeur de balises vous permet d'ajouter, de supprimer et de gérer les balises de plusieurs ressources. Avec Tag Editor, vous pouvez rechercher les ressources que vous souhaitez baliser, puis gérer les balises des ressources de vos résultats de recherche.

AWS Cost Categories vous permet d'attribuer une signification organisationnelle à vos coûts, sans avoir à étiqueter les ressources. Vous pouvez associer vos informations de coût et d'utilisation à des structures d'organisation internes uniques. Vous définissez des règles de catégorie pour associer et catégoriser les coûts à l'aide des dimensions de facturation, telles que les comptes et les balises. Cela fournit un autre niveau de fonctionnalité de gestion en plus du balisage. Vous pouvez également associer des comptes et des balises spécifiques à plusieurs projets.

Étapes d'implémentation

- Définition d'un schéma de balisage : réunissez toutes les parties prenantes de votre entreprise pour définir un schéma. Il s'agit généralement de membres du personnel technique, de l'équipe financière et de la direction. Définissez une liste de balises que toutes les ressources doivent avoir, ainsi qu'une liste de balises que des ressources doivent avoir. Veillez à ce que les noms et les valeurs des balises soient cohérents dans l'ensemble de votre organisation.
- Ressources de balises : en utilisant vos catégories de répartition des coûts définies, placez des balises sur toutes les ressources de vos charges de travail en fonction des catégories. Utilisez des outils tels que CLI l'éditeur de balises ou AWS Systems Manager pour augmenter l'efficacité.
- Implémenter les catégories de AWS coûts : vous pouvez créer des catégories de coûts sans implémenter le balisage. Les catégories de coûts utilisent les dimensions de coûts et d'utilisation

existantes. Créez des règles de catégorie à partir de votre schéma et mettez-les en œuvre dans les catégories de coûts.

- Automatisation du balisage : pour veiller à maintenir des niveaux élevés de balisage sur toutes les ressources, automatisez le balisage afin que les ressources soient automatiquement balisées lorsqu'elles sont créées. Utilisez des services tels que AWS CloudFormation pour vérifier que les ressources sont balisées lors de leur création. Vous pouvez également créer une solution pour baliser automatiquement les ressources à l'aide de fonctions Lambda ou utiliser un microservice personnalisé qui analyse régulièrement la charge de travail et supprime toutes les ressources qui ne sont pas balisées, ce qui est idéal pour les environnements de test et de développement.
- Surveillance du balisage et création de rapports associés : pour veiller à maintenir des niveaux élevés de balisage dans votre organisation, surveillez les balises de vos charges de travail et créez des rapports associés. Vous pouvez utiliser AWS Cost Explorer pour afficher le coût des ressources balisées et non balisées, ou recourir à des services tels que Tag Editor. Examinez régulièrement le nombre de ressources non balisées et prenez les mesures nécessaires pour ajouter des balises jusqu'à ce que vous atteigniez le niveau de balisage souhaité.

Ressources

Documents connexes :

- Bonnes pratiques du balisage
- AWS CloudFormation Tag de ressource
- AWS Cost Categories
- Ressources de balisage AWS
- Analyser vos coûts avec AWS les budgets
- Analyse des coûts à l'aide de Cost Explorer
- Gestion des rapports d'utilisation et de coûts AWS

Vidéos connexes :

- Comment puis-je étiqueter mes AWS ressources pour répartir ma facture par centre de coûts ou par projet ?
- Ressources de balisage AWS

COST03-BP03 Identifier les catégories d'attribution des coûts

Identifiez les catégories d’organisation telles que les unités commerciales, les services ou les projets qui pourraient être utilisés pour répartir les coûts au sein de votre organisation entre les entités consommatrices internes. Utilisez ces catégories pour renforcer la responsabilité en matière de dépenses, sensibiliser aux coûts et encourager des comportements de consommation efficaces.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Le processus de catégorisation des coûts est crucial pour la budgétisation, la comptabilité, les rapports financiers, la prise de décision, l’analyse comparative et la gestion de projet. En classant et en catégorisant les dépenses, les équipes peuvent mieux comprendre les types de coûts qu’elles doivent supporter tout au long de leur transition vers le cloud, ce qui les aide à prendre des décisions éclairées et à gérer les budgets de manière efficace.

La responsabilité des dépenses liées au cloud incite fortement à une gestion disciplinée de la demande et des coûts. Il en résulte des économies importantes sur les coûts liés au cloud pour les organisations qui allouent la majeure partie de leurs dépenses en matière de cloud à des unités commerciales ou à des équipes consommatrices. En outre, l’affectation des dépenses liées au cloud aide les organisations à adopter davantage de bonnes pratiques en matière de gouvernance centralisée du cloud.

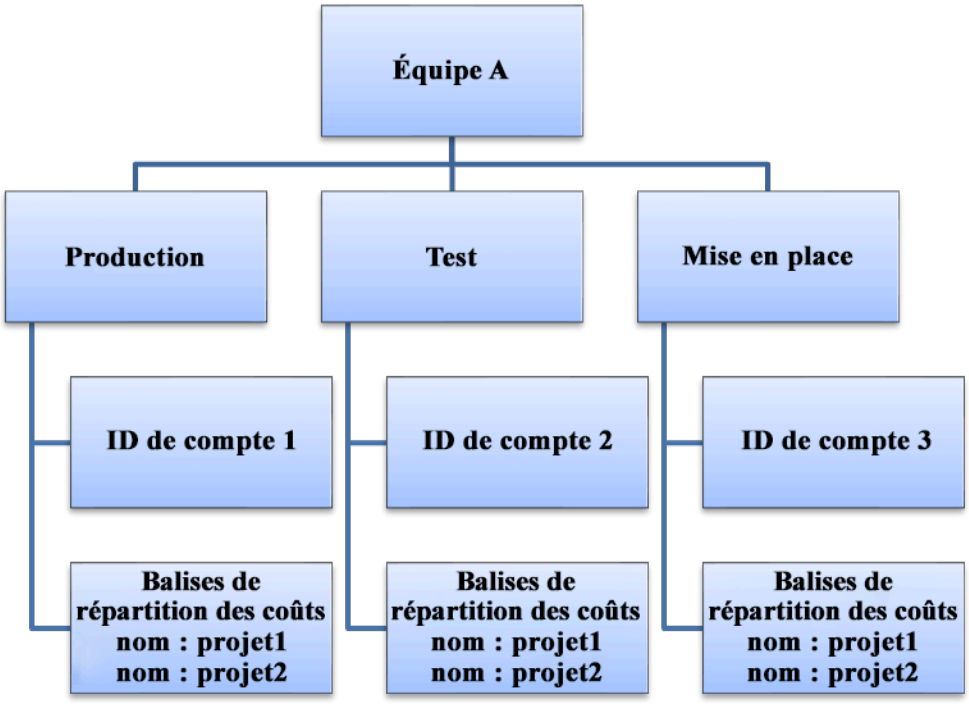
Travaillez avec votre équipe financière et les autres parties prenantes concernées pour comprendre les exigences relatives à la répartition des coûts au sein de votre entreprise lors de vos appels périodiques. Les coûts de la charge de travail doivent être répartis sur l’ensemble du cycle de vie, y compris le développement, les tests, la production et la mise hors service. Vous devez comprendre comment les coûts engagés pour l’apprentissage, le développement du personnel et la création d’idées sont attribués dans l’organisation. Cela peut être utile pour affecter correctement les comptes utilisés à cette fin aux budgets de formation et de développement, au lieu des budgets génériques de coûts informatiques.

Après avoir défini vos catégories d'attribution des coûts avec les parties prenantes de votre organisation, utilisez AWS Cost Categories pour regrouper vos informations sur les coûts et l'utilisation dans des catégories pertinentes AWS Cloud, telles que le coût d'un projet spécifique, ou Comptes AWS pour les départements ou les unités commerciales. Vous pouvez créer des catégories personnalisées et mapper vos informations de coût et d’utilisation dans ces catégories en fonction des règles que vous définissez grâce à différentes dimensions, telles que le compte, la balise,

le service ou le type de frais. Une fois les catégories de coûts définies, vous pouvez afficher vos informations de coût et d'utilisation pour chacune d'entre elles pour permettre à votre organisation de prendre de meilleures décisions stratégiques et d'achat. Ces catégories sont également visibles dans AWS Cost Explorer AWS Budgets AWS Cost and Usage Report , et.

Par exemple, créez des catégories de coûts pour vos unités commerciales (DevOps équipe) et, pour chaque catégorie, créez plusieurs règles (règles pour chaque sous-catégorie) avec plusieurs dimensions (étiquettes de répartition des coûts Comptes AWS, services ou type de frais) en fonction des groupements que vous avez définis. Vous pouvez utiliser les catégories de coûts pour organiser vos coûts à l'aide d'un moteur basé sur des règles. Les règles que vous configurez organisent vos coûts en catégories. Dans le cadre de ces règles, vous pouvez filtrer en utilisant plusieurs dimensions pour chaque catégorie, telles que les spécifiques Comptes AWS, les AWS services ou les types de frais. Vous pouvez utiliser ces catégories pour plusieurs produits dans la console AWS Billing and Cost Management et Cost Management. Cela inclut AWS Cost Explorer AWS Budgets, AWS Cost and Usage Report, et AWS Cost Anomaly Detection.

À titre d'exemple, le diagramme suivant vous montre comment regrouper vos coûts et vos informations d'utilisation dans votre organisation en ayant plusieurs équipes (catégorie de coûts), plusieurs environnements (règles), et chaque environnement ayant plusieurs ressources ou actifs (dimensions).



Graphique des coûts et de l'utilisation de l'organisation

Vous pouvez également regrouper les coûts avec les catégories de coûts. Une fois que vous avez créé les catégories de coûts (jusqu'à 24 heures après la création d'une catégorie de coût peuvent être nécessaires pour que les valeurs soient mises à jour dans vos relevés d'utilisation), elles apparaissent dans AWS Cost Explorer, AWS Budgets, AWS Cost and Usage Report et AWS Cost Anomaly Detection. Dans AWS Cost Explorer et AWS Budgets, une catégorie de coûts apparaît en tant que dimension de facturation supplémentaire. Vous pouvez l'utiliser pour filtrer la valeur de la catégorie de coûts spécifique, ou pour regrouper les valeurs par catégorie de coûts.

Étapes d'implémentation

- Définition des catégories de votre organisation : rencontrez les parties prenantes internes et les unités commerciales pour définir les catégories qui reflètent la structure et les besoins de votre organisation. Ces catégories devraient correspondre directement à la structure des catégories financières existantes, telles que l'unité commerciale, le budget, le centre de coûts ou le service. Examinez les résultats que le cloud apporte à votre entreprise, tels que la formation ou l'éducation, car il s'agit également de catégories organisationnelles.
- Définition de vos catégories fonctionnelles : rencontrez les parties prenantes internes et les unités commerciales pour définir des catégories qui reflètent les fonctions de votre entreprise. Il peut s'agir de la charge de travail ou des noms d'application, ainsi que du type d'environnement, comme la production, les tests ou le développement.
- Définissez les catégories de AWS coûts : créez des catégories de coûts pour organiser vos informations sur les coûts et l'utilisation à l'aide de AWS Cost Categories et cartographiez vos AWS coûts et votre utilisation dans des catégories pertinentes. Plusieurs catégories peuvent être attribuées à une ressource, et une ressource peut se trouver dans plusieurs catégories. Par conséquent, définissez autant de catégories que nécessaire afin de pouvoir gérer vos coûts dans la structure catégorisée à l'aide des catégories de coûts AWS .

Ressources

Documents connexes :

- Balisage de ressources AWS
- Utilisation des balises de répartition des coûts
- Analysez vos coûts avec AWS Budgets
- Analyse des coûts à l'aide de Cost Explorer
- Gestion de systèmes AWS Cost and Usage Report

- AWS Cost Categories
- Gérez vos coûts avec AWS Cost Categories
- Création de catégories de coûts
- Étiquetage des catégories de coûts
- Fractionnement des frais dans les catégories de coûts
- Fonctionnalités des catégories de coûts AWS

Exemples connexes :

- Organisez vos données de coûts et d'utilisation avec AWS Cost Categories
- Gérez vos coûts avec AWS Cost Categories
- Ateliers Well-Architected : visualisation des coûts et de l’utilisation
- Ateliers Well-Architected : catégories de coûts

COST03-BP04 Établir les paramètres de l'organisation

Établissez les métriques de l’organisation qui sont requises pour cette charge de travail. Les rapports des clients produits ou les pages Web diffusées aux clients sont des exemples de métriques d’une charge de travail.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Comprenez comment le rendement de votre charge de travail est mesuré par rapport à la réussite métier. Chaque charge de travail comporte généralement un petit ensemble de résultats majeurs qui indiquent les performances. Si votre charge de travail est complexe et comporte de nombreux éléments, vous pouvez en dresser la liste par ordre de priorité ou définir et suivre les métriques de chaque élément. Travaillez avec vos équipes pour savoir quelles métriques vous devez utiliser. Cette unité sera utilisée pour comprendre l’efficacité de la charge de travail, ou le coût de chaque production commerciale.

Étapes d’implémentation

- Définition de résultats de charge de travail : rencontrez les parties prenantes de l’entreprise et définissez les résultats de la charge de travail. Ces résultats constituent une des mesures principales de l’utilisation des clients. Ils doivent être des métriques économiques, et non

techniques. Il doit exister un petit nombre de métriques générales (moins de cinq) par charge de travail. Si la charge de travail produit plusieurs résultats pour différents cas d'utilisation, regroupez-les dans une seule métrique.

- Définition de résultats de composants de charge de travail : le cas échéant, si la charge de travail est volumineuse et complexe ou que vous pouvez facilement la diviser en composants (tels que des microservices) avec des entrées et des sorties bien définies, définissez des métriques pour chaque composant. L'effort doit refléter la valeur et le coût du composant. Procédez des plus grands aux plus petits composants.

Ressources

Documents connexes :

- Ressources de balisage AWS
- Analyser vos coûts avec AWS les budgets
- Analyse des coûts à l'aide de Cost Explorer
- Gestion des rapports d'utilisation et de coûts AWS

COST03-BP05 Configuration des outils de facturation et de gestion des coûts

Configurez les outils de gestion des coûts conformément aux politiques de votre organisation en matière de gestion et d'optimisation des dépenses dans le cloud. Ils incluent les services, les outils et les ressources pour organiser et suivre les données de coûts et d'utilisation, avoir plus de contrôle par la facturation consolidée et les autorisations d'accès, améliorer la planification via des budgets et des prévisions, recevoir des notifications ou des alertes, et réduire les coûts grâce aux optimisations des ressources et de la tarification.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Pour établir une solide responsabilisation, considérez d'abord la stratégie de votre compte comme faisant partie de votre stratégie de répartition des coûts. Faites les choses correctement et vous n'aurez peut-être pas besoin d'aller plus loin. Sinon, certains points risquent d'être omis et d'autres problèmes pourraient survenir par la suite.

Pour encourager la responsabilisation en matière de dépenses liées au cloud, accordez aux utilisateurs l'accès à des outils qui fournissent une visibilité sur leurs coûts et leur utilisation. AWS vous recommande de configurer toutes les charges de travail et toues les équipes aux fins suivantes :

- Organisation : établissez votre répartition des coûts et votre base de référence de la gouvernance avec votre propre stratégie de balisage et votre propre taxonomie. Créez plusieurs AWS comptes à l'aide d'outils tels que AWS Control Tower « AWS Organisation ». Marquez les AWS ressources prises en charge et classez-les de manière significative en fonction de la structure de votre organisation (unités commerciales, départements ou projets). Marquez les noms de compte pour des centres de coûts spécifiques et associez-les à AWS Cost Categories pour regrouper les comptes des unités commerciales dans leurs centres de coûts afin que le propriétaire de l'unité puisse voir la consommation de plusieurs comptes en un seul endroit.
- Accès : suivez les informations de facturation à l'échelle de l'organisation dans la facturation consolidée. Vérifiez que les parties prenantes et les responsables d'unités commerciales appropriés y ont accès.
- Contrôle : créez des mécanismes de gouvernance efficaces avec les garde-fous appropriés pour éviter les scénarios inattendus lorsque vous utilisez des politiques de contrôle des services (SCP), des politiques de balises, des IAM politiques et des alertes budgétaires. Par exemple, vous pouvez autoriser les équipes à créer des ressources spécifiques dans des régions de prédilection uniquement en utilisant des mécanismes de contrôle efficaces et empêcher la création de ressources sans balise spécifique (comme le centre de coûts).
- État actuel : configurez un tableau de bord affichant les niveaux actuels de coût et d'utilisation. Le tableau de bord doit être disponible dans un endroit hautement visible dans l'environnement de travail, comme un tableau de bord des opérations. Vous pouvez exporter des données et utiliser le tableau de bord des coûts et de l'utilisation à partir du hub d'optimisation des coûts AWS ou de tout autre produit pris en charge pour créer cette visibilité. Vous devrez peut-être créer différents tableaux de bord pour différents profils. Par exemple, le tableau de bord des responsables peut être différent du tableau de bord des ingénieurs.
- Notifications : envoyez des notifications lorsque le coût ou l'utilisation dépassent les limites définies et que des anomalies surviennent lors de la détection AWS des budgets ou des anomalies de AWS coûts.
- Rapports : résumez toutes les informations de coût et d'utilisation. Sensibilisez et responsabilisez les parties prenantes concernant vos dépenses liées au cloud grâce à des données de coûts détaillées et attribuables. Créez des rapports pertinents pour l'équipe qui les utilise et qui contiennent des recommandations.
- Suivi : affiche le coût et l'utilisation actuels par rapport aux objectifs ou cibles configurés.

- Analyse : permettez aux membres de l'équipe d'effectuer une analyse personnalisée et approfondie jusqu'à la granularité horaire, quotidienne ou mensuelle avec différents filtres (ressource, compte, tag, et bien plus encore).
- Inspection : restez à jour avec vos opportunités de déploiement de ressources et d'optimisation des coûts. Recevez des notifications via Amazon CloudWatchSNS, Amazon ou Amazon SES pour les déploiements de ressources au niveau de l'organisation. Passez en revue les recommandations d'optimisation des coûts avec AWS Trusted Advisor ou AWS Compute Optimizer.
- Rapports de tendance : affichez la variabilité des coûts et de l'utilisation sur la période requise avec la granularité nécessaire.
- Prévisions : affichez les coûts futurs prévus, estimez votre utilisation des ressources et dépensez en fonction des tableaux de bord des prévisions que vous créez.

Vous pouvez utiliser le hub d'optimisation des coûts AWS pour comprendre les opportunités potentielles de réduction des coûts consolidées à partir d'un emplacement centralisé et créer des exportations de données à intégrer à Amazon Athena. Vous pouvez également utiliser le hub d'optimisation des AWS coûts pour déployer le tableau de bord des coûts et de l'utilisation, qui utilise Amazon QuickSight pour une analyse interactive des coûts et un partage sécurisé des informations sur les coûts.

Si vous ne disposez pas des compétences ou de la bande passante essentielles dans votre organisation, vous pouvez travailler avec AWS ProServ, AWS Managed Services (AMS) ou AWS des partenaires. Vous pouvez également utiliser des outils tiers mais assurez-vous de valider la proposition de valeur.

Étapes d'implémentation

- Autorisation de l'accès aux outils selon les équipes : configurez vos comptes et créez des groupes ayant accès aux rapports de coûts et d'utilisation requis pour leurs consommations et utilisez AWS Identity and Access Management pour contrôler l'accès à des outils tels que AWS Cost Explorer. Ces groupes doivent inclure des représentants de toutes les équipes qui possèdent ou gèrent une application. Chaque équipe a ainsi accès à ses informations de coût et d'utilisation pour suivre sa consommation.
- Organisation des balises et des catégories de coûts : organisez vos coûts sur l'ensemble des équipes, des unités commerciales, des applications, des environnements et des projets. Utilisez des balises de ressources pour organiser les coûts, par balises de répartition des coûts. Créez des catégories de coûts basées sur des dimensions en utilisant des balises, des comptes, des services, etc. pour mapper vos coûts.

- Configurer AWS les budgets : configurez les AWS budgets sur tous les comptes pour vos charges de travail. Définissez des budgets pour les dépenses globales des comptes et des budgets pour les charges de travail à l'aide de balises et de catégories de coûts. Configurez les notifications dans AWS les budgets pour recevoir des alertes lorsque vous dépassez les montants budgétisés ou lorsque vos coûts estimés dépassent vos budgets.
- Configurez AWS la détection des anomalies de AWS coûts : utilisez la détection des anomalies de coûts pour vos comptes, services principaux ou catégories de coûts que vous avez créés afin de surveiller vos coûts et votre utilisation et de détecter les dépenses inhabituelles. Vous pouvez recevoir des alertes individuellement sous forme de rapports agrégés et recevoir des alertes par e-mail ou dans un SNS sujet Amazon, ce qui vous permet d'analyser et de déterminer la cause première de l'anomalie et d'identifier le facteur à l'origine de l'augmentation des coûts.
- Utilisation d'outils d'analyse des coûts : configurez AWS Cost Explorer pour votre charge de travail et vos comptes afin de visualiser vos données de coût pour une analyse plus approfondie. Créez un tableau de bord pour la charge de travail qui suit les dépenses globales et les principales métriques d'utilisation de la charge de travail, et qui prévoit les futurs coûts en fonction de vos anciennes données de coût.
- Utilisez des outils d'analyse des économies : utilisez le AWS Cost Optimization Hub pour identifier les opportunités d'économies grâce à des recommandations personnalisées, notamment la suppression des ressources inutilisées, le redimensionnement, les plans d'épargne, les réservations et les recommandations relatives aux optimiseurs de calcul.
- Configuration d'outils avancés : vous pouvez éventuellement créer des visuels pour faciliter l'analyse interactive et le partage des informations sur les coûts. Avec Data Exports on AWS Cost Optimization Hub, vous pouvez créer un tableau de bord des coûts et de l'utilisation alimenté par Amazon QuickSight pour votre organisation, qui fournit des détails et une granularité supplémentaires. Vous pouvez également implémenter des fonctionnalités d'analyse avancées en utilisant les exportations de données dans Amazon Athena pour les requêtes avancées, et créer des tableaux de bord sur Amazon. QuickSight Collaborez avec des partenaires AWS pour adopter des solutions de gestion du cloud pour une surveillance et une optimisation consolidées des factures du cloud.

Ressources

Documents connexes :

- Qu'est-ce que AWS Billing and Cost Management la gestion des coûts ?
- Mise en place de votre AWS environnement de bonnes pratiques

- Bonnes pratiques pour le balisage des ressources AWS
- Marquer vos ressources AWS
- AWS Cost Categories
- Analyser vos coûts avec AWS les budgets
- Analysez vos coûts avec AWS Cost Explorer
- Qu'est-ce que AWS les exportations de données ?

Vidéos connexes :

- Deploying Cloud Intelligence Dashboards
- Recevez des alertes sur n'importe quel indicateur d'optimisation des coûts FinOps ou KPI

Exemples connexes :

- Tableau de bord des coûts et de l'utilisation développé par Amazon QuickSight
- Atelier sur la gouvernance des coûts et de l'utilisation AWS

COST03-BP06 Allouer les coûts en fonction des métriques de charge de travail

Répartissez les coûts de la charge de travail en fonction des métriques d'utilisation ou des résultats économiques afin de mesurer la rentabilité de la charge de travail. Mettez en œuvre un processus pour analyser les données de coût et d'utilisation avec les services d'analytique, ce qui peut fournir des informations et des fonctionnalités de refacturation.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

L'optimisation des coûts consiste à obtenir des résultats commerciaux au prix le plus bas, ce qui ne peut être réalisé qu'en répartissant les coûts de la charge de travail en fonction des métriques de la charge de travail (mesurées par l'efficacité de la charge de travail). Surveillez les métriques de charge de travail définies via des fichiers journaux ou une autre surveillance d'application. Combinez ces données avec les coûts de la charge de travail, qui peuvent être obtenus en examinant les coûts avec une valeur de balise spécifique ou un ID de compte spécifique. Effectuez cette analyse au niveau horaire. Votre efficacité change généralement si certains composants de coût sont statiques (par exemple, une base de données dorsale exécutée en permanence) avec un taux de demandes

variable (par exemple, des pics d'utilisation entre neuf heures et dix-sept heures, avec peu de demandes pendant la nuit). La compréhension de la relation entre les coûts statiques et les coûts variables vous aide à cibler vos activités d'optimisation.

La création de métriques de charge de travail pour les ressources partagées peut s'avérer difficile par rapport à des ressources telles que les applications conteneurisées sur Amazon Elastic Container Service ECS (Amazon) et Amazon API Gateway. Cependant, il existe certains moyens de catégoriser l'utilisation et de suivre les coûts. Si vous devez suivre Amazon ECS et les ressources AWS Batch partagées, vous pouvez activer les données de répartition des coûts fractionnées dans AWS Cost Explorer. Grâce au partage des données de répartition des coûts, vous pouvez comprendre et optimiser le coût et l'utilisation de vos applications conteneurisées et répartir les coûts des applications entre les différentes entités commerciales en fonction de la manière dont les ressources de calcul et de mémoire partagées sont consommées.

Étapes d'implémentation

- Allocation de coûts aux métriques de charge de travail : à l'aide des métriques définies et des balises configurées, créez une métrique qui combine la sortie de la charge de travail et son coût. Utilisez des services d'analyse tels qu'Amazon Athena et Amazon QuickSight pour créer un tableau de bord d'efficacité adapté à la charge de travail globale et à tous les composants.

Ressources

Documents connexes :

- Ressources de balisage AWS
- Analyser vos coûts avec AWS les budgets
- Analyse des coûts à l'aide de Cost Explorer
- Gestion des rapports d'utilisation et de coûts AWS

Exemples connexes :

- Améliorez la visibilité des coûts ECS d'Amazon AWS Batch grâce aux AWS données de répartition des coûts

COÛT 4. Comment mettez-vous les ressources hors service ?

Mettez en œuvre le contrôle des modifications et la gestion des ressources depuis le début du projet jusqu'à la fin de vie. Cela garantit que vous arrêtez ou résiliez les ressources inutilisées pour réduire le gaspillage.

Bonnes pratiques

- COST04-BP01 Suivez les ressources tout au long de leur durée de vie
- COST04-BP02 Mettre en œuvre un processus de mise hors service
- COST04-BP03 Ressources de mise hors service
- COST04-BP04 Démanteler automatiquement les ressources
- COST04-BP05 Appliquer les politiques de conservation des données

COST04-BP01 Suivez les ressources tout au long de leur durée de vie

Définissez et mettez en œuvre une méthode pour suivre les ressources et leurs associations avec les systèmes, tout au long de leur durée de vie. Vous pouvez utiliser le balisage pour identifier la charge de travail ou la fonction de la ressource.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Mettez hors service les ressources de charge de travail qui ne sont plus requises. Cela concerne notamment les ressources utilisées pour les tests. Une fois les tests terminés, les ressources peuvent être supprimées. Le suivi des ressources avec des balises (et l'exécution de rapports sur ces balises) peut vous aider à identifier les actifs à mettre hors service, du fait de leur non-utilisation ou de l'expiration de la licence. Les balises constituent un moyen efficace de suivre les ressources, car elles identifient la ressource avec sa fonction ou une date connue à laquelle elle peut être mise hors service. Des rapports sur ces balises peuvent ensuite être exécutés. Les exemples de valeurs pour le balisage de fonctionnalité sont le `feature-X testing` qui permet d'identifier l'objectif de la ressource en matière de cycle de vie de la charge de travail. Un autre exemple consiste à utiliser `LifeSpan` ou `TTL` pour les ressources, telles que le nom et la valeur de la clé de to-be-deleted balise, pour définir la période ou le moment précis de la mise hors service.

Étapes d'implémentation

- Implémentation d'un schéma de balisage : implémentez un schéma de balisage qui identifie la charge de travail à laquelle appartient la ressource, en veillant à ce que toutes les ressources de la charge de travail soient balisées en conséquence. Le balisage vous aide à catégoriser les ressources par objectif, équipe, environnement ou autres critères pertinents pour votre entreprise. Pour plus de détails sur les cas d'utilisation, les stratégies et les techniques de balisage, consultez la section Bonnes pratiques de balisage AWS.
- Mise en œuvre d'une surveillance du débit ou des sorties de la charge de travail : mettez en œuvre une surveillance du débit de la charge de travail ou des alarmes, en lançant des demandes d'entrée ou des achèvements de sortie. Configurez-la pour envoyer des notifications lorsque les demandes ou les réponses de la charge de travail sont nulles, ce qui indique que ses ressources ne sont plus utilisées. Intégrez un facteur temporel si la charge de travail est régulièrement nulle dans des conditions normales. Pour plus de détails sur les ressources inutilisées ou sous-utilisées, consultez la section Contrôles d'optimisation des coûts AWS Trusted Advisor.
- AWS Ressources de groupe : créez des groupes pour les AWS ressources. Vous pouvez l'AWS Resource Groupsutiliser pour organiser et gérer AWS les ressources qui s'y trouvent Région AWS. Vous pouvez ajouter des balises à la plupart de vos ressources afin d'identifier et de trier plus facilement ces dernières au sein de votre organisation. Utilisez Tag Editor pour ajouter des balises aux ressources prises en charge en bloc. Prévoyez d'utiliser AWS Service Catalog pour créer, gérer et distribuer des portefeuilles de produits approuvés aux utilisateurs finaux, ainsi que pour gérer le cycle de vie des produits.

Ressources

Documents connexes :

- AWS Auto Scaling
- AWS Trusted Advisor
- AWS Trusted Advisor Contrôles d'optimisation des coûts
- Ressources de balisage AWS
- Publication des métriques personnalisées

Vidéos connexes :

- Comment optimiser les coûts en utilisant AWS Trusted Advisor

Exemples connexes :

- Organiser AWS les ressources
- Optimisez les coûts en utilisant AWS Trusted Advisor

COST04-BP02 Mettre en œuvre un processus de mise hors service

Mettez en œuvre un processus pour identifier et mettre hors service les ressources inutilisées.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Mettez en place un processus normalisé dans l'ensemble de votre organisation pour identifier et supprimer les ressources inutilisées. Ce processus doit définir la fréquence à laquelle les recherches sont effectuées, ainsi que les processus de suppression de la ressource, afin de s'assurer que toutes les exigences de l'organisation sont respectées.

Étapes d'implémentation

- Création et implémentation d'un processus de mise en services : en travaillant avec les développeurs et les propriétaires de la charge de travail, créez un processus de mise hors service de la charge de travail et de ses ressources. Le processus doit couvrir la méthode pour vérifier que la charge de travail et chacune de ses ressources sont en cours d'utilisation. Détaillez également les étapes nécessaires à la mise hors service de la ressource, en la supprimant et en garantissant la conformité aux exigences réglementaires. Toutes les ressources associées doivent être incluses, notamment les licences ou le stockage dédié. Informez les propriétaires de la charge de travail que le processus de mise hors service a été lancé.

  Suivez les étapes de mise hors service suivantes pour effectuer une à une les vérifications requises dans le cadre de votre processus :

  - Identification des ressources à mettre hors service : identifiez les ressources éligibles à la mise hors service dans votre AWS Cloud. Enregistrez toutes les informations nécessaires et planifiez la mise hors service. Dans votre chronologie, assurez-vous d'envisager la survenue de problèmes imprévus et d'identifier les étapes les plus propices à cette éventualité au cours du processus.
  - Coordination et communication : travaillez avec les propriétaires de la charge de travail pour confirmer la ressource à mettre hors service

- Enregistrez les métadonnées et créez des sauvegardes : enregistrez les métadonnées (telles que les métadonnées publiquesIPs, régionales, AZVPC, sous-réseaux et groupes de sécurité) et créez des sauvegardes (telles que les instantanés ou prises d'Amazon Elastic Block StoreAMI, l'exportation de clés et l'exportation de certificats) si cela est nécessaire pour les ressources de l'environnement de production ou s'il s'agit de ressources critiques.
- Valider infrastructure-as-code : déterminez si les ressources ont été déployées avec AWS CloudFormation Terraform ou tout autre outil de infrastructure-as-code déploiement afin qu'elles puissent être redéployées si nécessaire. AWS Cloud Development Kit (AWS CDK)
- Prévention d’accès : appliquez des contrôles restrictifs pendant un certain temps, afin d’empêcher l’utilisation des ressources pendant que vous déterminez si la ressource est requise. Vérifiez que l’environnement des ressources peut être rétabli à son état d’origine si nécessaire.
- Suivez votre processus de mise hors service interne : suivez les tâches administratives et le processus de mise hors service de votre organisation, comme la suppression de la ressource du domaine de votre organisation, de l'DNSenregistrement et de la suppression de la ressource de votre outil de gestion de configuration, de votre outil de surveillance, de votre outil d'automatisation et de vos outils de sécurité.

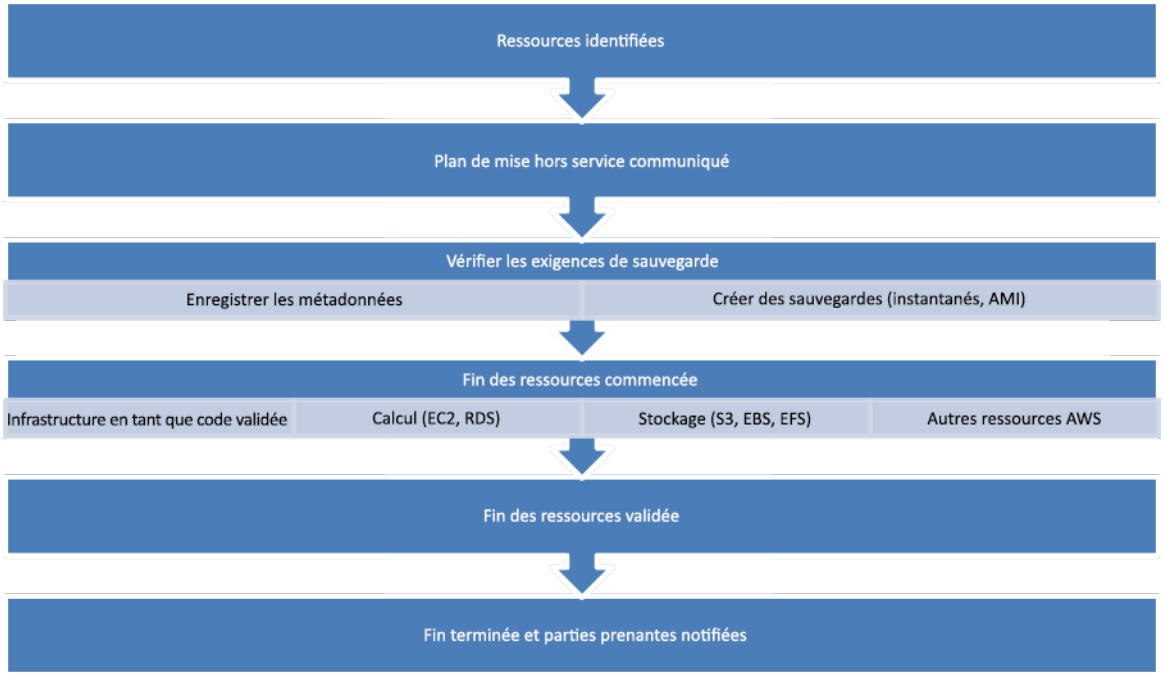
Si la ressource est une EC2 instance Amazon, consultez la liste suivante. Pour plus de détails, consultez Comment supprimer ou résilier mes EC2 ressources Amazon ?

- Arrêtez ou mettez fin à toutes vos EC2 instances Amazon et à tous vos équilibreurs de charge. Les EC2 instances Amazon sont visibles dans la console pendant une courte période après leur résiliation. Vous n’êtes pas facturé pour les instances qui ne sont pas en cours d’exécution
- Supprimez votre infrastructure Autoscaling.
- Libérez tous les hôtes dédiés.
- Supprimez tous les EBS volumes Amazon et les EBS instantanés Amazon.
- Libérez toutes les adresses IP Elastic.
- Désenregistrez toutes les Amazon Machine Images ()AMIs.
- Mettez fin à tous les AWS Elastic Beanstalk environnements.

Si la ressource est un objet du stockage Amazon S3 Glacier et si vous supprimez une archive avant d’atteindre la durée minimale de stockage, nous vous facturerons une taxe de suppression anticipée au prorata. La durée minimale de stockage d’Amazon S3 Glacier dépend de la classe de stockage utilisée. Pour obtenir un résumé de la durée de stockage minimale pour chaque classe de stockage, consultez la section Performances des classes de stockage Amazon S3. Pour en savoir

plus sur le mode de calcul des frais de suppression anticipée, consultez la tarification d'Amazon S3.

L'organigramme suivant du processus de mise hors service simple décrit les étapes de la mise hors service. Avant de mettre hors service des ressources, vérifiez que les ressources identifiées pour la mise hors service ne sont pas utilisées par l'organisation.



Flux de mise hors service des ressources.

Ressources

Documents connexes :

- AWS Auto Scaling
- AWS Trusted Advisor
- AWS CloudTrail

Vidéos connexes :

- Supprimer la CloudFormation pile mais conserver certaines ressources
- Découvrez quel utilisateur a lancé l'EC2instance Amazon

Exemples connexes :

- Supprimer ou résilier des EC2 ressources Amazon
- Découvrez quel utilisateur a lancé une EC2 instance Amazon

COST04-BP03 Ressources de mise hors service

Mettez hors service les ressources déclenchées par des événements tels que les audits périodiques ou les modifications d'utilisation. La mise hors service est généralement effectuée régulièrement et elle peut être manuelle ou automatisée.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

La fréquence et l'effort de recherche des ressources inutilisées doivent refléter les économies potentielles, de sorte qu'un compte ayant un faible coût doit être analysé moins fréquemment qu'un compte ayant des coûts plus importants. Les recherches et les événements de mise hors service peuvent être initiés par des changements d'état dans la charge de travail, comme un produit en fin de vie ou en cours de remplacement. Les recherches et les événements de mise hors service peuvent également être initiés par des événements externes, tels que des changements dans les conditions du marché ou l'arrêt d'un produit.

Étapes d'implémentation

- Mise hors service de ressources : il s'agit de la phase d'amortissement des ressources AWS qui ne sont plus nécessaires ou de la fin d'un contrat de licence. Effectuez toutes les vérifications finales avant de passer à l'étape de l'élimination et de mettre hors service les ressources afin d'éviter toute perturbation indésirable, comme la réalisation d'instantanés ou de sauvegardes. En utilisant le processus dédié, mettez hors service chaque ressource ayant été identifiée comme inutilisée.

Ressources

Documents connexes :

- AWS Auto Scaling
- AWS Trusted Advisor

Exemples connexes :

- Ateliers Well-Architected : mise hors service des ressources (niveau 100)

COST04-BP04 Démanteler automatiquement les ressources

Concevez votre charge de travail de manière à gérer proprement l'arrêt des ressources lorsque vous identifiez et mettez hors service des ressources non critiques, des ressources qui ne sont pas nécessaires ou des ressources peu utilisées.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Utilisez l'automatisation pour réduire ou supprimer les coûts associés au processus de mise hors service. La conception de votre charge de travail pour effectuer une mise hors service automatisée réduira le coût global de la charge de travail pendant sa durée de vie. Vous pouvez utiliser Amazon EC2 Auto Scaling ou Application Auto Scaling pour effectuer le processus de mise hors service. Vous pouvez également implémenter un code personnalisé à l'aide du APIou SDK pour mettre automatiquement hors service les ressources de charge de travail.

Les applications modernes sont conçues d'abord sans serveur, une stratégie qui donne la priorité à l'adoption de services sans serveur. AWS a développé des services sans serveur pour les trois couches de votre stack : calcul, intégration et magasins de données. L'utilisation d'une architecture sans serveur vous permettra de réduire les coûts pendant les périodes de faible trafic avec une augmentation et une réduction automatiques.

Étapes d'implémentation

- Implémentation d'Amazon EC2 Auto Scaling ou d'Application Auto Scaling : pour les ressources prises en charge, configurez-les avec Amazon EC2 Auto Scaling ou Application Auto Scaling. Ces services peuvent vous aider à optimiser votre utilisation et à réduire vos coûts lorsque vous consommez AWS des services. Lorsque la demande baisse, ces services suppriment automatiquement toute capacité de ressource excédentaire pour vous permettre d'éviter les dépenses excessives.
- Configurer CloudWatch pour mettre fin aux instances : les instances peuvent être configurées pour se terminer à l'aide d'CloudWatch alarmes. En utilisant les métriques du processus de mise hors service, mettez en œuvre une alarme avec une action du cloud Amazon Elastic Compute. Veillez à vérifier l'opération dans un environnement hors production avant le déploiement.

- Implémenter le code au sein de la charge de travail : vous pouvez utiliser le AWS SDK ou AWS CLI pour mettre hors service les ressources de charge de travail. Implémentez au sein de l'application un code qui s'intègre aux AWS ressources qui ne sont plus utilisées, les arrête ou les supprime.
- Utilisez des services sans serveur : privilégiez la création d'architectures sans serveur et d'architectures pilotées par les événements AWS pour créer et exécuter vos applications. AWS propose plusieurs services technologiques sans serveur qui, par nature, fournissent automatiquement une utilisation optimisée des ressources et une mise hors service automatisée (mise à l'échelle interne et externe). Avec des applications sans serveur, l'utilisation des ressources est optimisée automatiquement et vous ne payez jamais d'approvisionnement excessif.

Ressources

Documents connexes :

- Amazon EC2 Auto Scaling
- Commencer à utiliser Amazon EC2 Auto Scaling
- Application Autoscaling
- AWS Trusted Advisor
- Sans serveur activé AWS
- Création d'alarmes qui arrêtent, mettent hors service, redémarrent ou récupèrent une instance
- Ajouter des actions de résiliation aux CloudWatch alarmes Amazon

Exemples connexes :

- Planification de la suppression automatique des AWS CloudFormation piles
- Ateliers Well-Architected : mise hors service automatique des ressources (niveau 100)
- Nettoyage AWS automatique de Servian

COST04-BP05 Appliquer les politiques de conservation des données

Définissez des stratégies de conservation des données sur les ressources prises en charge pour traiter la suppression des objets conformément aux exigences de votre organisation. Identifiez et supprimez les ressources et les objets inutiles ou orphelins qui ne sont plus nécessaires.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Utilisez des stratégies de conservation des données et des stratégies de cycle de vie pour réduire les coûts associés au processus de mise hors service et les coûts de stockage des ressources identifiées. La définition de vos stratégies de conservation des données et de cycle de vie pour la migration et la suppression automatisées de la classe de stockage réduira les frais de stockage généraux pendant la durée de vie. Vous pouvez utiliser Amazon Data Lifecycle Manager pour automatiser la création et la suppression de snapshots Amazon Elastic Block Store et d'Amazon Machine Images (AMIs) EBS soutenues par Amazon, et utiliser Amazon S3 Intelligent-Tiering ou une configuration de cycle de vie Amazon S3 pour gérer le cycle de vie de vos objets Amazon S3. Vous pouvez également implémenter du code personnalisé à l'aide du APIou SDK pour créer des politiques de cycle de vie et des règles de politique pour les objets à supprimer automatiquement.

Étapes d’implémentation

- Utiliser Amazon Data Lifecycle Manager : utilisez les politiques de cycle de vie d'Amazon Data Lifecycle Manager pour automatiser la suppression des EBS instantanés Amazon et des produits EBS sauvegardés par AMIs Amazon.
- Configuration du cycle de vie d’un compartiment : utilisez la configuration du cycle de vie d’Amazon S3 sur un compartiment pour définir les actions qu’Amazon S3 doit effectuer pendant le cycle de vie d’un objet, ainsi que la suppression à la fin du cycle de vie de l’objet, en fonction des besoins de votre entreprise.

Ressources

Documents connexes :

- AWS Trusted Advisor
- Amazon Data Lifecycle Manager
- Comment définir une configuration de cycle de vie sur un compartiment Amazon S3

Vidéos connexes :

- Automatisez les EBS instantanés Amazon avec Amazon Data Lifecycle Manager
- Empty an Amazon S3 bucket using a lifecycle configuration rule

Exemples connexes :

- Vidage d’un compartiment Amazon S3 à l’aide d’une règle de configuration de cycle de vie

- Laboratoire Well-Architected : mise hors service automatique des ressources (niveau 100)

Ressources rentables

Questions

- COÛT 5. Comment évaluer les coûts lorsque vous sélectionnez des services ?
- COÛT 6. Comment atteindre les objectifs de coût lorsque vous sélectionnez le type, la taille et le nombre de ressources ?
- COÛT 7. Comment utiliser les modèles de tarification pour réduire les coûts ?
- COÛT 8. Comment planifier les frais de transfert de données ?

COÛT 5. Comment évaluer les coûts lorsque vous sélectionnez des services ?

Amazon EC2, Amazon EBS et Amazon S3 sont les services fondamentaux d'AWS. Les services gérés tels qu'Amazon RDS et Amazon DynamoDB, sont des services AWS de plus haut niveau, ou de niveau application. En sélectionnant les services fondamentaux et les services gérés appropriés, vous pouvez optimiser cette charge de travail en matière de coûts. Par exemple, en utilisant des services gérés, vous pouvez réduire ou supprimer une grande partie de votre traitement administratif et opérationnel, et vous dégagez ainsi du temps pour travailler sur les applications et les activités liées aux activités.

Bonnes pratiques

- COST05-BP01 Identifier les exigences de l'organisation en matière de coûts
- COST05-BP02 Analyser toutes les composantes de la charge de travail
- COST05-BP03 Procéder à une analyse approfondie de chaque composant
- COST05-BP04 Sélectionnez des logiciels offrant des licences économiques
- COST05-BP05 Sélectionner les composants de cette charge de travail pour optimiser les coûts conformément aux priorités de l'organisation
- COST05-BP06 Effectuer une analyse des coûts pour différentes utilisations au fil du temps

COST05-BP01 Identifier les exigences de l'organisation en matière de coûts

Collaborez avec les membres de l'équipe pour définir l'équilibre entre l'optimisation des coûts et les autres piliers, tels que la performance et la fiabilité, pour cette charge de travail.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Dans la plupart des organisations, le département des technologies de l’information (TI) est composé de plusieurs petites équipes, chacune ayant son propre programme et son propre domaine d’intervention. Le tout reflète les spécialités et les compétences des membres de son équipe. Vous devez comprendre les objectifs généraux, les priorités et les buts de votre organisation et la manière dont chaque département ou projet contribue à ces objectifs. La catégorisation de toutes les ressources essentielles, notamment le personnel, les équipements, les technologies, le matériel et les services externes, est cruciale pour atteindre les objectifs de l’organisation et mettre en place une planification budgétaire exhaustive. L’adoption de cette approche systématique de l’identification et de la compréhension des coûts est fondamentale pour établir un plan de coûts réaliste et solide pour l’organisation.

Lorsque vous sélectionnez des services pour votre charge de travail, il est essentiel que vous compreniez les priorités de votre entreprise. Créez un équilibre entre l'optimisation des coûts et les autres piliers du AWS Well-Architected Framework, tels que les performances et la fiabilité. Ce processus doit être mené de manière systématique et régulière afin de refléter l’évolution des objectifs de l’organisation, des conditions du marché et de la dynamique opérationnelle. Une charge de travail entièrement optimisée en matière de coûts est la solution la plus conforme aux besoins de votre organisation, et pas nécessairement la moins coûteuse. Rencontrez toutes les équipes de votre organisation (équipes produits, commerciales, techniques et financières) pour recueillir des informations. Évaluez l’impact des compromis entre des intérêts concurrents ou des approches alternatives pour prendre des décisions éclairées au moment de déterminer où concentrer les efforts ou de choisir une ligne de conduite.

Par exemple, l’accélération de la mise sur le marché de nouvelles fonctionnalités peut être privilégiée par rapport à l’optimisation des coûts, ou vous pouvez choisir une base de données relationnelle pour les données non relationnelles afin de simplifier l’effort de migration d’un système, plutôt que de migrer vers une base de données optimisée pour votre type de données et de mettre à jour votre application.

Étapes d’implémentation

- Identification des exigences de l’organisation en matière de coût : réunissez-vous avec les membres de l’équipe de votre organisation, y compris les personnes chargées de la gestion des produits, les responsables d’application, les équipes de développement et d’exploitation, la direction et les services financiers. Hiérarchisez les piliers Well-Architected de cette charge de

travail et ses composants. Vous devriez obtenir un classement des piliers par ordre de priorité. Vous pouvez également attribuer une pondération à chaque pilier pour indiquer le degré de priorité supplémentaire d'un pilier ou une similarité de priorité entre deux piliers.

- Traitement et documentation de la dette technique : au cours de l'examen de la charge de travail, abordez la dette technique. Documentez un élément en attente pour retenir la charge de travail à l'avenir dans le but de la refactoriser ou de la réorganiser pour l'optimiser davantage. Il est essentiel de communiquer clairement les concessions qui ont été faites aux autres parties prenantes.

Ressources

Bonnes pratiques associées :

- REL11-BP07 Architectez votre produit pour atteindre les objectifs de disponibilité et les accords de niveau de service en matière de disponibilité () SLAs
- OPS01-BP06 Évaluer les compromis

Documents connexes :

- AWS Calculateur du coût total de possession (TCO)
- Classes de stockage Amazon S3
- Produits cloud

COST05-BP02 Analyser toutes les composantes de la charge de travail

Assurez-vous que chaque composant de la charge de travail est analysé, peu importe la taille ou les coûts actuels. L'effort de vérification doit tenir compte des avantages potentiels, tels que les coûts actuels et prévus.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Les composants de la charge de travail, qui sont conçus pour apporter une valeur métier à l'organisation, peuvent englober différents services. Pour chaque composant, on peut choisir des AWS Cloud services spécifiques pour répondre aux besoins de l'entreprise. Cette sélection peut être influencée par des facteurs tels que la connaissance ou l'expérience antérieure de ces services.

Après avoir identifié les exigences de votre organisation comme indiqué dans le document COST05-BP01 Identifier les exigences de l'organisation en matière de coûts, effectuez une analyse approfondie de tous les éléments de votre charge de travail. Analysez chaque composant en tenant compte des coûts et des tailles actuels et prévus. Examinez le coût de l'analyse par rapport aux économies potentielles de la charge de travail au cours de son cycle de vie. L'effort d'analyse de tous les composants de cette charge de travail doit correspondre aux économies ou aux améliorations potentielles escomptées grâce à l'optimisation de ce composant spécifique. Par exemple, si le coût de la ressource proposée est de 10 USD par mois et que les charges prévues ne dépassent pas 15 USD par mois, une journée d'effort pour réduire les coûts de 50 % (5 USD par mois) pourrait dépasser le bénéfice potentiel sur la durée de vie du système. Utilisez une estimation plus rapide et plus efficace basée sur des données pour obtenir le meilleur résultat global pour ce composant.

Les charges de travail peuvent évoluer dans le temps, et un ensemble de services qui est actuellement adapté peut ne pas être optimal si l'architecture ou l'utilisation de la charge de travail évolue. L'analyse pour la sélection des services doit intégrer les états de charge de travail et les niveaux d'utilisation actuels et futurs. La mise en œuvre d'un service pour un état ou un usage futur de la charge de travail peut réduire les coûts globaux en diminuant ou en supprimant l'effort nécessaire pour effectuer des changements futurs. Par exemple, l'utilisation de EMR Serverless peut être le choix approprié au départ. Toutefois, à mesure que la consommation de ce service augmente, la transition vers EMR le service activé EC2 pourrait réduire les coûts liés à cette composante de la charge de travail.

AWS Cost Exploreret le AWS Cost and Usage Report s (CUR) peut analyser le coût d'une preuve de concept (PoC) ou d'un environnement d'exécution. Vous pouvez également utiliser AWS Pricing Calculator pour estimer les coûts de charge de travail.

Rédigez un flux de travail à suivre par les équipes techniques pour vérifier leurs charges de travail. Bien que ce flux de travail doive être simple, couvrez également toutes les étapes nécessaires pour vous assurer que les équipes comprennent chaque composant de la charge de travail et sa tarification. Votre organisation pourra ensuite suivre et personnaliser ce flux de travail en fonction des besoins spécifiques de chaque équipe.

1. Répertoriez chaque service utilisé en fonction de votre charge de travail : c'est un bon point de départ. Identifiez tous les services actuellement utilisés et l'origine des coûts.
2. Compréhension du fonctionnement de la tarification pour ces services : veillez à comprendre le modèle de tarification de chaque service. Les différents AWS services ont des modèles de tarification différents en fonction de facteurs tels que le volume d'utilisation, le transfert de données et la tarification spécifique aux fonctionnalités.

3. Concentrez-vous sur les services qui entraînent des coûts de charge de travail inattendus et qui ne correspondent pas à votre utilisation prévue ni aux résultats commerciaux : identifiez les valeurs aberrantes ou les services dont le coût n'est pas proportionnel à la valeur ou à l'utilisation associée à l'utilisation AWS Cost Explorer ou aux services AWS Cost and Usage Report. Il est important de corréler les coûts aux résultats commerciaux afin de prioriser les efforts d'optimisation.
4. AWS Cost Explorer, CloudWatch Logs, VPC Flow Logs et Amazon S3 Storage Lens pour comprendre la cause première de ces coûts élevés : ces outils jouent un rôle essentiel dans le diagnostic des coûts élevés. Chaque service propose une approche différente pour visualiser et analyser l'utilisation et les coûts. Par exemple, Cost Explorer aide à déterminer les tendances globales en matière de coûts, CloudWatch Logs fournit des informations opérationnelles, VPC Flow Logs affiche le trafic IP et Amazon S3 Storage Lens est utile pour les analyses de stockage.
5. AWS Budgets À utiliser pour établir des budgets pour certains montants pour des services ou des comptes : La définition de budgets est un moyen proactif de gérer les coûts. AWS Budgets À utiliser pour définir des seuils budgétaires personnalisés et recevoir des alertes lorsque les coûts dépassent ces seuils.
6. Configurez les CloudWatch alarmes Amazon pour envoyer des alertes de facturation et d'utilisation : configurez la surveillance et les alertes pour les mesures de coût et d'utilisation. CloudWatch les alarmes peuvent vous avertir lorsque certains seuils sont dépassés, ce qui améliore le temps de réponse des interventions.

Encouragez des améliorations notables et des économies financières au fil du temps grâce à un examen stratégique de tous les composants de la charge de travail, quelles que soient leurs caractéristiques actuelles. L'effort déployé dans ce processus d'évaluation doit être délibéré, et tenir dûment compte des bénéfices potentiels qui pourraient en découler.

Étapes d'implémentation

- Répertorier les composants de la charge de travail : créez une liste des composants de votre charge de travail. Utilisez cette liste pour vérifier que chaque composant a été analysé. L'effort déployé doit refléter la sévérité de la charge de travail telle que définie par les priorités de l'organisation. Regroupez les ressources sur le plan fonctionnel pour améliorer l'efficacité, notamment du stockage des bases de données de production s'il existe plusieurs bases de données.
- Prioriser la liste des composants : prenez la liste des composants et priorisez-la par ordre d'effort. Elle est généralement classée par ordre de coût du composant (du plus cher au moins cher) ou par ordre de criticité (telle qu'elle est définie par les priorités de votre organisation).

- Exécution de l’analyse : pour chaque élément de la liste, examinez les options et les services disponibles et choisissez l’option qui correspond le mieux à vos priorités organisationnelles.

Ressources

Documents connexes :

- AWS Pricing Calculator
- AWS Cost Explorer
- Classes de stockage Amazon S3
- Produits AWS Cloud

Vidéos connexes :

- AWS Série sur l'optimisation des coûts : CloudWatch

COST05-BP03 Procéder à une analyse approfondie de chaque composant

Examinez le coût global de chaque composant pour l’organisation. Calculez le coût total de possession en tenant compte du coût des opérations et de la gestion, en particulier lorsque vous utilisez des services gérés par un fournisseur de cloud. L’effort d’examen doit prendre en compte les avantages potentiels (par exemple, la durée de l’analyse est proportionnelle au coût du composant).

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Tenez compte du gain de temps qui permettra à votre équipe de se concentrer sur le remboursement de la dette technique, l’innovation, les fonctionnalités à valeur ajoutée et la création de votre avantage différentiel. Par exemple, il peut être nécessaire de procéder à un lift-and-shift (également appelé réhébergement) de vos bases de données depuis votre environnement sur site vers le cloud aussi rapidement que possible et de l’optimiser ultérieurement. Il est intéressant d’explorer les économies possibles réalisées en utilisant des services gérés sur AWS pouvant supprimer ou réduire les coûts de licence. Les services gérés AWS éliminent les charges opérationnelles et administratives liées à la maintenance d'un service, telles que l'application de correctifs ou la mise à niveau du système d'exploitation, et vous permettent de vous concentrer sur l'innovation et les affaires.

Étant donné que les services gérés fonctionnent à l'échelle du cloud, ils peuvent réduire le coût par transaction ou par service. Vous pouvez effectuer des optimisations potentielles afin d'obtenir des bénéfices concrets, sans pour autant changer l'architecture de base de l'application. Par exemple, vous souhaitez peut-être réduire le temps que vous consacrez à la gestion des instances de base de données en migrant vers une database-as-a-service plateforme telle qu'Amazon Relational Database Service (RDSAmazon) ou en migrant votre application vers une plateforme entièrement gérée telle que. AWS Elastic Beanstalk

En général, les services gérés ont des attributs que vous pouvez définir pour assurer une capacité suffisante. Vous devez définir et surveiller ces attributs afin que votre capacité excédentaire soit réduite au minimum et que vos performances soient maximisées. Vous pouvez modifier les attributs liés à AWS Managed Services l'utilisation du AWS Management Console ou AWS APIs et SDKs pour adapter les besoins en ressources à l'évolution de la demande. Par exemple, vous pouvez augmenter ou diminuer le nombre de nœuds sur un EMR cluster Amazon (ou un cluster Amazon Redshift) à des fins d'extension ou d'intégration.

Vous pouvez également regrouper plusieurs instances sur une AWS ressource pour activer une utilisation plus dense. Par exemple, vous pouvez mettre en service plusieurs petites bases de données sur une seule instance de base de données Amazon Relational Database Service (RDSAmazon). À mesure que l'utilisation augmente, vous pouvez migrer l'une des bases de données vers une instance de RDS base de données Amazon dédiée à l'aide d'un processus de capture instantanée et de restauration.

Lors de la mise en service de charges de travail sur des services gérés, vous devez connaître les exigences d'ajustement de la capacité du service. Ces exigences sont généralement le temps, l'effort et toute incidence sur le fonctionnement normal de la charge de travail. La ressource allouée doit laisser le temps à tout changement de se produise, en allouant la surcharge requise pour le permettre. Les efforts permanents requis pour modifier les services peuvent être réduits à pratiquement zéro en utilisant APIs et en SDKs intégrant des outils de système et de surveillance, tels qu'Amazon CloudWatch.

Amazon RDS, Amazon Redshift et Amazon ElastiCache fournissent un service de base de données géré. Amazon AthenaEMR, Amazon et Amazon OpenSearch Service fournissent un service d'analyse géré.

AMSest un service qui gère AWS l'infrastructure pour le compte des entreprises clientes et partenaires. Il fournit un environnement sécurisé et conforme sur lequel vous pouvez déployer vos charges de travail. AMSutilise des modèles d'exploitation du cloud d'entreprise avec automatisation

pour vous permettre de répondre aux exigences de votre organisation, de passer plus rapidement au cloud et de réduire vos coûts de gestion permanents.

Étapes d'implémentation

- Réalisation d'une analyse approfondie : à l'aide de la liste des composants, examinez chaque composant de la plus haute priorité à la plus basse. Pour les composants les plus prioritaires et les plus coûteux, effectuez une analyse supplémentaire et évaluez toutes les options disponibles et leur impact sur le long terme. Pour les composants de moindre priorité, évaluez si des changements d'utilisation modifieraient la priorité du composant, puis analysez l'effort approprié.
- Comparez les ressources gérées et non gérées : considérez le coût opérationnel des ressources que vous gérez et comparez-les aux ressources AWS gérées. Par exemple, passez en revue vos bases de données exécutées sur des EC2 instances Amazon et comparez-les avec Amazon RDS Options (un service AWS géré) ou avec Amazon EMR par rapport à l'exécution d'Apache Spark sur AmazonEC2. Lorsque vous passez d'une charge de travail autogérée à une charge de travail AWS entièrement gérée, étudiez attentivement les options qui s'offrent à vous. Les trois facteurs les plus importants à prendre en compte sont le type de service géré que vous souhaitez utiliser, le processus que vous utiliserez pour migrer vos données et la compréhension du modèle de responsabilité partagée AWS.

Ressources

Documents connexes :

- AWS Calculateur du coût total de possession (TCO)
- Classes de stockage Amazon S3
- Produits AWS Cloud
- AWS Modèle de responsabilité partagée

Vidéos connexes :

- Why move to a managed database?
- Qu'est-ce qu'Amazon EMR et comment puis-je l'utiliser pour le traitement des données ?

Exemples connexes :

- Pourquoi passer à une base de données gérée

- Consolidez les données issues de bases de données SQL Server identiques dans une seule base de données Amazon RDS for SQL Server en utilisant AWS DMS
- Fournissez des données à grande échelle à Amazon Managed Streaming for Apache Kafka (Amazon) MSK
- Migrer unASP. NETapplication Web pour AWS Elastic Beanstalk

COST05-BP04 Sélectionnez des logiciels offrant des licences économiques

Les logiciels open source éliminent les coûts de licences logicielles, qui peuvent entraîner des coûts significatifs pour la charge de travail. Lorsqu'un logiciel sous licence est requis, évitez les licences liées à des attributs arbitrairesCPUs, par exemple, recherchez des licences liées à des sorties ou à des résultats. Le coût de ces licences est plus proche de l'avantage qu'elles procurent.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : bas

Directives d'implémentation

L'open source est né dans le contexte du développement de logiciels pour indiquer que le logiciel est conforme à certains critères de distribution gratuite. Les logiciels open source sont composés de code source que tout le monde peut inspecter, modifier et améliorer. En fonction des exigences commerciales, des compétences des ingénieurs, de l'utilisation prévue ou d'autres dépendances technologiques, les entreprises peuvent envisager d'utiliser des logiciels open source AWS afin de minimiser leurs coûts de licence. En d'autres termes, le coût des licences logicielles peut être éliminé grâce à l'utilisation de logiciels open source. Cela peut avoir un impact significatif sur les coûts de charge de travail à mesure que la taille de la charge de travail évolue.

Mesurez les avantages des logiciels sous licence par rapport au coût total pour optimiser votre charge de travail. Modélisez les modifications apportées aux licences et leur impact sur vos coûts de charge de travail. Si un fournisseur modifie le coût de votre licence de base de données, examinez en quoi cela affecte l'efficacité globale de votre charge de travail. Prenez en compte l'historique des annonces de tarification de vos fournisseurs pour connaître les tendances des changements de licence pour leurs produits. Les coûts de licence peuvent également évoluer indépendamment du débit ou de l'utilisation, par exemple pour les licences qui évoluent en fonction du matériel (licences CPU liées). Ces licences doivent être évitées, car les coûts peuvent rapidement augmenter sans résultats correspondants.

Par exemple, l'exploitation d'une EC2 instance Amazon dans us-east-1 avec un système d'exploitation Linux vous permet de réduire les coûts d'environ 45 % par rapport à l'exécution d'une autre instance EC2 Amazon fonctionnant sous Windows.

AWS Pricing CalculatorIl offre un moyen complet de comparer les coûts de différentes ressources avec différentes options de licence, telles que les RDS instances Amazon et les différents moteurs de base de données. En outre, cela AWS Cost Explorer fournit une perspective inestimable sur les coûts des charges de travail existantes, en particulier celles associées à différentes licences. Pour la gestion des licences, AWS License Manager propose une méthode rationalisée pour superviser et gérer les licences logicielles. Les clients peuvent déployer et utiliser leur logiciel open source préféré dans AWS Cloud.

Étapes d'implémentation

- Analyse des options de licence : passez en revue les conditions de licence des logiciels disponibles. Recherchez les versions open source qui ont les fonctionnalités requises et déterminez si les avantages des logiciels sous licence l'emportent sur le coût. Des conditions favorables permettent d'aligner le coût du logiciel sur les avantages qu'il procure.
- Analyse du fournisseur de logiciels : passez en revue les historiques de tarification ou de licence du fournisseur. Recherchez les changements qui ne s'alignent pas sur les résultats, tels que les conditions pénalisantes de l'exécution sur des matériels ou des plateformes spécifiques à un fournisseur. Déterminez également comment ils effectuent les audits et les sanctions qui pourraient être imposées.

Ressources

Documents connexes :

- Open Source sur AWS
- AWS Calculateur du coût total de possession (TCO)
- Classes de stockage Amazon S3
- Produits cloud

Exemples connexes :

- Blogs open source
- AWS Blogs open source
- Évaluation de l'optimisation et des licences

COST05-BP05 Sélectionner les composants de cette charge de travail pour optimiser les coûts conformément aux priorités de l'organisation

Tenez compte du coût lorsque vous sélectionnez tous les composants de votre charge de travail. Cela inclut l'utilisation de services gérés et au niveau des applications ou de services sans serveur, de conteneurs ou d'une architecture axée sur les événements pour réduire le coût global. Réduisez les coûts de licence en utilisant des logiciels open source, des logiciels qui ne comportent pas de frais de licence ou des alternatives pour réduire les dépenses.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Tenez compte du coût des services et des options lorsque vous sélectionnez tous les composants. Cela inclut l'utilisation de services gérés et au niveau de l'application, tels qu'Amazon Relational Database Service (RDSAmazon), Amazon DynamoDB, Amazon Simple Notification Service (Amazon) et SNS Amazon Simple Email Service (SESAmazon) afin de réduire le coût global de l'organisation.

Utilisez des systèmes sans serveur et des conteneurs pour le calcul, comme AWS Lambda et Amazon Simple Storage Service (Amazon S3) pour les sites web statiques. Conteneurisez votre application si possible et utilisez des services de conteneurs AWS gérés tels qu'Amazon Elastic Container Service (AmazonECS) ou Amazon Elastic Kubernetes Service (Amazon). EKS

Réduisez les coûts de licence en utilisant des logiciels open source ou des logiciels qui n'impliquent pas de frais de licence, par exemple, Amazon Linux pour le calcul des charges de travail ou la migration des bases de données vers Amazon Aurora.

Vous pouvez utiliser des services sans serveur ou au niveau des applications tels que Lambda, Amazon SimpleQueue Service (Amazon)SQS, Amazon et Amazon. SNS SES Ces services vous dispensent de gérer une ressource et assurent les fonctions d'exécution de code, de mise en file d'attente et de distribution de messages. L'autre avantage est qu'ils sont mis à l'échelle en termes de performances et de coûts en fonction de l'utilisation, ce qui permet une répartition et une attribution efficace des coûts.

L'utilisation d'une architecture axée sur les événements est également possible avec les services sans serveur. Les architectures axées sur les événements reposent sur la technologie push, ce qui signifie que tout se passe à la demande au fur et à mesure que l'événement se présente dans le routeur. Ainsi, vous ne payez pas pour qu'une interrogation continue vérifie un événement. Cela

signifie moins de consommation de bande passante réseau, moins CPU d'utilisation, moins de capacité de parc inactive et moins de SSL poignées de TLS main.

Pour plus d'informations sur la technologie sans serveur, consultez le livre blanc Well-Architected – Présentation des applications sans serveur.

Étapes d'implémentation

- Sélection de chaque service pour optimiser le coût : à l'aide de votre liste de priorités et d'analyse, sélectionnez chaque option qui correspond le mieux à vos priorités organisationnelles. Au lieu d'augmenter la capacité pour répondre à la demande, envisagez d'autres options qui peuvent vous offrir de meilleures performances à moindre coût. Par exemple, si vous devez examiner le trafic attendu pour vos bases de données AWS, envisagez d'augmenter la taille de l'instance ou d'utiliser les ElastiCache services Amazon (Redis ou Memcached) pour fournir des mécanismes de mise en cache pour vos bases de données.
- Évaluation de l'architecture axée sur les événements : une architecture sans serveur vous permet également de créer une architecture basée sur les événements pour les applications distribuées reposant sur des microservices, ce qui vous aide à créer des solutions évolutives, résilientes, flexibles et rentables.

Ressources

Documents connexes :

- AWS Calculateur du coût total de possession (TCO)
- AWS sans serveur
- Qu'est-ce qu'une architecture axée sur les événements ?
- Classes de stockage Amazon S3
- Produits cloud
- Amazon ElastiCache (RedisOSS)

Exemples connexes :

- Démarrage avec les architectures axées sur les événements
- Architecture basée sur les événements
- Comment Statsig fonctionne 100 fois plus efficacement avec Amazon ElastiCache (Redis) OSS

- Bonnes pratiques d'utilisation des AWS Lambda fonctions

COST05-BP06 Effectuer une analyse des coûts pour différentes utilisations au fil du temps

Les charges de travail peuvent changer au fil du temps. Certains services ou fonctionnalités sont plus rentables à différents niveaux d'utilisation. Si vous effectuez l'analyse de chaque composant au fil du temps et en fonction de l'utilisation prévue, la charge de travail reste rentable pendant toute sa durée de vie.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Au fur AWS et à mesure du lancement de nouveaux services et fonctionnalités, les services les mieux adaptés à votre charge de travail peuvent changer. L'effort requis doit refléter les avantages potentiels. La fréquence de révision de la charge de travail dépend des exigences de votre organisation. S'il s'agit d'une charge de travail d'un coût important, la mise en œuvre de nouveaux services plus tôt permettra de maximiser les économies, de sorte qu'un examen plus fréquent peut être avantageux. Une autre initiation à vérifier est le changement des modèles d'utilisation. D'importants changements d'utilisation peuvent indiquer que d'autres services seraient plus optimaux.

Si vous devez transférer des données AWS Cloud, vous pouvez sélectionner une grande variété d' AWS offres de services et d'outils partenaires pour vous aider à migrer vos ensembles de données, qu'il s'agisse de fichiers, de bases de données, d'images de machines, de volumes de blocs ou même de sauvegardes sur bande. Par exemple, pour déplacer une grande quantité de données vers et depuis AWS ou pour traiter des données en périphérie, vous pouvez utiliser l'un des appareils AWS spécialement conçus pour déplacer des pétaoctets de données hors ligne de manière rentable. Autre exemple : pour des taux de transfert de données plus élevés, un service de connexion directe peut être moins cher VPN qu'un service fournissant la connectivité constante requise pour votre entreprise.

Évaluez votre activité de mise à l'échelle en fonction de l'analyse des coûts pour une utilisation différente au fil du temps. Analysez le résultat pour voir si la stratégie de mise à l'échelle peut être ajustée pour ajouter des instances avec plusieurs types d'instances et d'options d'achat. Vérifiez vos paramètres pour voir si le minimum peut être réduit pour satisfaire les demandes des utilisateurs avec une plus petite taille de flotte et ajouter davantage de ressources pour répondre à la demande élevée attendue.

Réalisez une analyse des coûts pour différentes utilisations au fil du temps en discutant avec les parties prenantes de votre organisation et utilisez la fonctionnalité de prévision de AWS Cost Explorer pour prévoir l'impact potentiel des modifications de service. Surveillez le niveau d'utilisation, les lancements AWS Budgets, les alarmes AWS Cost Anomaly Detection de CloudWatch facturation et identifiez et mettez en œuvre plus rapidement les services les plus rentables.

Étapes d'implémentation

- Définition des modèles d'utilisation prévue : en collaboration avec votre organisation, par exemple, les responsables du marketing et les propriétaires de produits, documentez les modèles d'utilisation attendue et prévue de la charge de travail. Discutez avec les parties prenantes de votre entreprise des augmentations de coûts et d'utilisation historiques et prévues et assurez-vous que les augmentations s'alignent sur les exigences de votre entreprise. Identifiez les jours, semaines ou mois civils pendant lesquels vous vous attendez à ce qu'un plus grand nombre d'utilisateurs utilisent vos AWS ressources, ce qui indique que vous devez augmenter la capacité des ressources existantes ou adopter des services supplémentaires pour réduire les coûts et améliorer les performances.
- Réalisation d'une analyse des coûts en fonction de l'utilisation prévue : à l'aide des modèles d'utilisation définis, effectuez une analyse à chacun de ces points. L'effort d'analyse doit refléter le résultat potentiel. Par exemple, si le changement d'utilisation est important, une analyse approfondie doit être effectuée pour vérifier les coûts et les changements éventuels. En d'autres termes, quand les coûts augmentent, l'utilisation de l'entreprise doit également augmenter.

Ressources

Documents connexes :

- AWS Calculateur du coût total de possession (TCO)
- Classes de stockage Amazon S3
- Produits cloud
- Amazon EC2 Auto Scaling
- Migration des données dans le cloud
- AWS Snow Family

Vidéos connexes :

- AWS OpsHub for Snow Family

COÛT 6. Comment atteindre les objectifs de coût lorsque vous sélectionnez le type, la taille et le nombre de ressources ?

Veillez à choisir la taille et le nombre de ressources qui conviennent pour la tâche à accomplir. En choisissant le type, la taille et le nombre les plus rentables, vous réduisez le gaspillage.

Bonnes pratiques

- COST06-BP01 Réaliser une modélisation des coûts
- COST06-BP02 Sélectionnez le type, la taille et le nombre de ressources en fonction des données
- COST06-BP03 Sélectionnez automatiquement le type, la taille et le nombre de ressources en fonction des métriques
- COST06-BP04 Envisager l'utilisation de ressources partagées

COST06-BP01 Réaliser une modélisation des coûts

Identifiez les exigences de l'organisation (telles que les besoins métier et les engagements existants) et réalisez une modélisation des coûts (globaux) de la charge de travail et de chacun de ses composants. Procédez à des évaluations de la charge de travail en fonction de diverses charges prévues et comparez les coûts. L'effort de modélisation doit refléter les avantages potentiels. Par exemple, le temps passé est proportionnel au coût des composants.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

Effectuez une modélisation des coûts de votre charge de travail et de chacun de ses composants, afin de comprendre l'équilibre entre les ressources et de déterminer la taille correcte de chaque ressource dans la charge de travail, compte tenu d'un niveau de performance spécifique. La compréhension des considérations relatives aux coûts peut éclairer le cas d'utilisation et le processus de prise de décision de votre organisation lors de l'évaluation des résultats de réalisation de valeur pour le déploiement d'une charge de travail planifiée.

Procédez à des évaluations de la charge de travail en fonction de diverses charges prévues et comparez les coûts. L'effort de modélisation doit refléter les avantages potentiels. Par exemple, le temps passé est proportionnel au coût des composants ou aux économies prévues. Pour les meilleures pratiques, reportez-vous à la section Révision du pilier de l'efficacité en matière de performance du AWS Well-Architected Framework.

Par exemple, pour créer une modélisation des coûts pour une charge de travail composée de ressources informatiques, AWS Compute Optimizer peut faciliter la modélisation des coûts d'exécution des charges de travail. Il fournit des recommandations de dimensionnement des ressources de calcul basées sur l'utilisation historique. Assurez-vous que les CloudWatch agents sont déployés sur les EC2 instances Amazon afin de collecter des métriques de mémoire qui vous aideront à formuler des recommandations plus précises AWS Compute Optimizer. Il s'agit de la source de données idéale pour les ressources de calcul, car c'est un service gratuit qui utilise le machine learning pour faire plusieurs recommandations en fonction des niveaux de risque.

Il existe plusieurs services que vous pouvez utiliser avec des journaux personnalisés comme sources de données afin de redimensionner les opérations pour d'autres services et composants de la charge de travail, tels qu'AWS Trusted AdvisorAmazon et CloudWatchAmazon CloudWatch Logs. AWS Trusted Advisor vérifie les ressources et signale les ressources peu utilisées, ce qui peut vous aider à bien dimensionner vos ressources et à créer une modélisation des coûts.

Voici des recommandations pour les données et métriques de modélisation des coûts :

- Le suivi doit refléter l'expérience utilisateur avec précision. Choisissez le niveau de précision correct pour la période et choisissez judicieusement le maximum ou le 99e centile au lieu de la moyenne.
- Sélectionnez la granularité appropriée pour la période d'analyse qui couvre tous les cycles de charge de travail. Par exemple, si une analyse de deux semaines est effectuée, vous pourriez négliger un cycle mensuel de forte utilisation, ce qui pourrait conduire à une sous-allocation.
- Choisissez les AWS services adaptés à votre charge de travail planifiée en tenant compte de vos engagements existants, des modèles de tarification sélectionnés pour les autres charges de travail et de votre capacité à innover plus rapidement et à vous concentrer sur votre valeur commerciale principale.

Étapes d'implémentation

- Réalisation d'une modélisation des coûts des ressources : déployez la charge de travail ou une démonstration de faisabilité dans un compte séparé avec les types et tailles de ressources spécifiques à tester. Exécutez la charge de travail avec les données de test et enregistrez les résultats, ainsi que les données de coût pour la période où le test a été effectué. Redéployez ensuite la charge de travail ou modifiez les types et les tailles des ressources et relancez le test. Incluez les frais de licence de tous les produits que vous pourriez utiliser avec ces ressources et les frais d'opérations (main-d'œuvre ou ingénierie) estimés pour le déploiement et la gestion de

ces ressources pendant la création de la modélisation des coûts. Envisagez une modélisation des coûts par période (heure, jour, mois, année ou trois ans).

Ressources

Documents connexes :

- AWS Auto Scaling
- Identification des opportunités à la bonne taille
- CloudWatch Fonctionnalités d'Amazon
- Optimisation des coûts : Amazon EC2 Right Sizing
- AWS Compute Optimizer
- AWS Calculateur de prix

Exemples connexes :

- Réalisation d’une modélisation des coûts basée sur les données
- Estimer le coût des configurations de AWS ressources planifiées
- Choisissez les bons AWS outils

COST06-BP02 Sélectionnez le type, la taille et le nombre de ressources en fonction des données

Sélectionnez la taille ou le type de ressources en fonction des données relatives à la charge de travail et aux caractéristiques des ressources Par exemple, le calcul, la mémoire, le débit ou l’accès intensif en écriture. Cette sélection est généralement effectuée en utilisant une version précédente (sur site) de la charge de travail, en utilisant de la documentation ou d’autres sources d’information sur la charge de travail.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Amazon EC2 propose une large sélection de types d'instances avec différents niveaux de mémoireCPU, de stockage et de capacité réseau pour répondre à différents cas d'utilisation. Ces types d'instances proposent différentes combinaisons de capacités de mémoireCPU, de stockage et de mise en réseau, ce qui vous permet de sélectionner la combinaison de ressources adaptée à vos projets en toute flexibilité. Chaque type d’instance est disponible dans plusieurs tailles afin que vous

puissiez ajuster vos ressources en fonction des exigences de votre charge de travail. Pour déterminer le type d'instance dont vous avez besoin, rassemblez des informations sur la configuration système requise de l'application ou du logiciel que vous envisagez d'exécuter sur votre instance. Ces détails doivent comprendre les éléments suivants :

- Système d'exploitation
- Nombre de CPU cœurs
- GPUnoyaux
- Quantité de mémoire système (RAM)
- Type et espace de stockage
- Exigence de la bande passante du réseau

Identifiez l'objectif des exigences de calcul et l'instance requise, puis explorez les différentes familles d'EC2instances Amazon. Amazon propose les familles de types d'instances suivantes :

- Usage général
- Calcul optimisé
- Mémoire optimisée
- Stockage optimisé
- Calcul accéléré
- HPCOptimisé

Pour mieux comprendre les objectifs spécifiques et les cas d'utilisation qu'une famille d'EC2instances Amazon spécifique peut remplir, consultez la section Types d'AWS instances.

La collecte de la configuration système requise est essentielle pour sélectionner la famille d'instances et le type d'instance les mieux adaptés à vos besoins. Les noms de types d'instances sont composés du nom de famille et de la taille de l'instance. Par exemple, l'instance t2.micro appartient à la famille T2 et a une taille microscopique.

Sélectionnez la taille ou le type de ressources en fonction des caractéristiques de la charge de travail et des ressources (calcul, mémoire, débit ou accès intensif en écriture, par exemple). Cette sélection est généralement effectuée à l'aide d'une modélisation des coûts, d'une version antérieure de la charge de travail (version sur site, par exemple), d'une documentation ou d'autres sources d'informations sur la charge de travail (livres blancs ou solutions publiées). L'utilisation

de calculateurs de AWS prix ou d'outils de gestion des coûts peut aider à prendre des décisions éclairées concernant les types, les tailles et les configurations des instances.

Étapes d’implémentation

- Sélection des ressources en fonction des données : utilisez vos données de modélisation des coûts pour sélectionner le niveau prévu d’utilisation de la charge de travail, ainsi que le type et la taille des ressources spécifiées. Sur la base des données de modélisation des coûts, déterminez le nombre de mémoire virtuelle CPUs totale (GiB), le volume de stockage de l'instance locale (Go), les EBS volumes Amazon et le niveau de performance du réseau, en tenant compte du taux de transfert de données requis pour l'instance. Effectuez toujours vos choix en vous appuyant sur des analyses détaillées et des données précises afin d’optimiser les performances tout en gérant efficacement les coûts.

Ressources

Documents connexes :

- AWS Types d’instances
- AWS Auto Scaling
- CloudWatch Fonctionnalités d'Amazon
- Optimisation des coûts : EC2 bonne taille

Vidéos connexes :

- Sélection de l'EC2instance Amazon adaptée à vos charges de travail
- Right size your service

Exemples connexes :

- Il est désormais plus facile de découvrir et de comparer les types d'EC2instances Amazon

COST06-BP03 Sélectionnez automatiquement le type, la taille et le nombre de ressources en fonction des métriques

Utilisez les métriques de la charge de travail en cours pour sélectionner la taille et le type appropriés afin d’optimiser les coûts. Mettez en service de manière appropriée le débit, le dimensionnement et le

stockage pour les services de calcul, de stockage, de données et de mise en réseau. Pour ce faire, utilisez une boucle de rétroaction, telle que la mise à l'échelle automatique ou du code personnalisé dans la charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Créez une boucle de rétroaction qui utilise des métriques actives de la charge de travail en cours pour apporter des modifications à cette dernière. Vous pouvez utiliser un service géré, tel que AWS Auto Scaling, que vous configurez pour effectuer les opérations de dimensionnement qui vous conviennent. AWS fournit également APIsSDKs, et des fonctionnalités qui permettent de modifier les ressources avec un minimum d'effort. Vous pouvez programmer une charge de travail sur stop-and-start une EC2 instance Amazon pour autoriser le changement de taille ou de type d'instance. De cette manière, vous tirez parti des avantages d'un redimensionnement tout en supprimant presque tous les coûts opérationnels nécessaires pour effectuer la modification.

Certains AWS services intègrent une sélection automatique du type ou de la taille, comme Amazon Simple Storage Service Intelligent-Tiering. Amazon S3 Intelligent-Tiering déplace automatiquement vos données entre deux niveaux d'accès, accès fréquent et accès peu fréquent, en fonction de vos modèles d'utilisation.

Étapes d'implémentation

- Amélioration de votre observabilité en configurant les indicateurs de charge de travail : capturez les métriques clés de la charge de travail. Ces indicateurs fournissent une indication de l'expérience client, telle que le résultat de la charge de travail, et tiennent compte des différences entre les types et les tailles de ressources, telles que CPU l'utilisation de la mémoire. Pour les ressources de calcul, analysez les données de performance afin de dimensionner correctement vos EC2 instances Amazon. Identifiez les instances inactives et celles qui sont sous-utilisées. Les indicateurs clés à rechercher sont CPU l'utilisation et l'utilisation de la mémoire (par exemple, 40 % d'CPUutilisation dans 90 % des cas, comme expliqué dans Rightsizing with AWS Compute Optimizer and Memory Utilization Enabled). Identifiez les instances dont CPU l'utilisation maximale et l'utilisation de la mémoire sont inférieures à 40 % sur une période de quatre semaines. Ce sont les instances dont la taille doit être adaptée pour réduire les coûts. Pour les ressources de stockage telles qu'Amazon S3, vous pouvez utiliser Amazon S3 Storage Lens, qui vous permet de voir 28 métriques réparties dans différentes catégories au niveau du compartiment, et 14 jours de données historiques dans le tableau de bord par défaut. Vous pouvez filtrer votre tableau de bord

Amazon S3 Storage Lens par récapitulatif et optimisation des coûts ou événements pour analyser des métriques spécifiques.

- Afficher les recommandations de redimensionnement : utilisez les recommandations de redimensionnement et l'outil de EC2 redimensionnement Amazon dans AWS Compute Optimizer la console de gestion des coûts, ou passez en revue le dimensionnement correct de vos ressources pour ajuster votre charge AWS Trusted Advisor de travail. Il est important d'utiliser les bons outils pour dimensionner correctement les différentes ressources et de suivre les directives de dimensionnement, qu'il s'agisse d'une EC2 instance Amazon, de classes de AWS stockage ou de types d'instances Amazon. RDS Pour les ressources de stockage, vous pouvez utiliser Amazon S3 Storage Lens qui vous donne une visibilité sur l'utilisation du stockage d'objets et les tendances d'activité en plus de faire des recommandations exploitables afin d'optimiser les coûts et d'appliquer les bonnes pratiques en matière de protection des données. À l'aide des recommandations contextuelles qu'Amazon S3 Storage Lens tire de l'analyse des métriques sur toute votre organisation, vous pouvez prendre des mesures immédiates pour optimiser votre stockage.
- Sélection automatique du type et de la taille des ressources en fonction des métriques : à l'aide des métriques de charge de travail, sélectionnez manuellement ou automatiquement les ressources de votre charge de travail. Pour les ressources de calcul, la configuration d' AWS Auto Scaling ou la mise en œuvre du code dans votre application peut limiter l'effort requis si des changements fréquents sont nécessaires. De plus, la mise en œuvre des modifications peut ainsi survenir de manière plus précoce qu'avec un processus manuel. Vous pouvez lancer et mettre automatiquement à l'échelle une flotte d'instances à la demande et d'instances Spot au sein d'un même groupe Auto Scaling. Outre les remises accordées sur l'utilisation des instances Spot, vous pouvez utiliser des instances réservées ou un Savings Plan afin de bénéficier de réductions sur les tarifs standard des instances à la demande. Tous ces facteurs combinés vous aident à optimiser les économies réalisées sur les EC2 instances Amazon et à déterminer l'échelle et les performances souhaitées pour votre application. Vous pouvez également utiliser une stratégie de sélection du type d'instance basée sur les attributs (ABS) dans Auto Scaling Groups (ASG), qui vous permet d'exprimer les besoins de votre instance sous la forme d'un ensemble d'attributs, tels que vCPU, memory et storage. Vous pouvez utiliser automatiquement les types d'instances de nouvelle génération lorsqu'ils sont publiés et accéder à une gamme de capacités plus étendue avec les instances Amazon EC2 Spot. Amazon EC2 Fleet et Amazon EC2 Auto Scaling sélectionnent et lancent des instances qui correspondent aux attributs spécifiés, éliminant ainsi le besoin de sélectionner manuellement les types d'instances. En ce qui concerne les ressources de stockage, vous pouvez utiliser les fonctionnalités Amazon S3 Intelligent Tiering et Amazon EFS Infrequent Access, qui vous permettent de sélectionner automatiquement des

classes de stockage qui permettent de réaliser des économies automatiques sur les coûts de stockage lorsque les modèles d'accès aux données changent, sans impact sur les performances ni surcharge opérationnelle.

Ressources

Documents connexes :

- AWS Auto Scaling
- AWS Dimensionnement correct
- AWS Compute Optimizer
- CloudWatch Fonctionnalités d'Amazon
- CloudWatchMise en place
- CloudWatchPublication de métriques personnalisées
- Commencer à utiliser Amazon EC2 Auto Scaling
- Amazon S3 Storage Lens
- Amazon S3 Intelligent Tiering
- Accès EFS peu fréquent à Amazon
- Lancez une EC2 instance Amazon à l'aide du SDK

Vidéos connexes :

- Right Size Your Services

Exemples connexes :

- Sélection du type d'instance basée sur les attributs pour Auto Scaling for Amazon EC2 Fleet
- Optimisation du service de conteneur Amazon Elastic pour le coût à l’aide d’une mise à l’échelle planifiée
- Mise à l'échelle prédictive avec Amazon EC2 Auto Scaling
- Optimisation des coûts et amélioration de la visibilité sur l’utilisation avec Amazon S3 Storage Lens
- Ateliers Well-Architected : recommandations de redimensionnement (niveau 100)

COST06-BP04 Envisager l’utilisation de ressources partagées

Pour les services déjà déployés au niveau de l’organisation pour plusieurs unités commerciales, envisagez d’utiliser des ressources partagées afin d’augmenter l’utilisation et de réduire le coût total de possession (TCO). L’utilisation de ressources partagées peut être une option rentable pour centraliser la gestion et les coûts en utilisant des solutions existantes, en partageant des composants, ou les deux. Gérez les fonctions courantes telles que la surveillance, les sauvegardes et la connectivité, soit dans les limites d’un compte, soit dans un compte dédié. Vous pouvez également réduire les coûts en mettant en œuvre la standardisation ainsi qu’en réduisant la duplication et la complexité.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Lorsque plusieurs charges de travail entraînent la même fonction, utilisez les solutions existantes et les composants partagés pour améliorer la gestion et optimiser les coûts. Envisagez d’utiliser les ressources existantes (en particulier les ressources partagées), telles que des serveurs de base de données hors production ou des services d’annuaire, pour réduire les coûts liés au cloud en appliquant les bonnes pratiques de sécurité et la réglementation de l’entreprise. Pour optimiser la réalisation de la valeur et l’efficacité, il est essentiel de réaffecter les coûts (en utilisant le relevé des services rendus et la rétrofacturation) aux domaines pertinents de l’entreprise qui stimulent la consommation.

Le terme de relevé des services rendus fait référence aux rapports qui répartissent les coûts du cloud en catégories attribuables, telles que les consommateurs, les unités commerciales, les comptes du grand livre ou d’autres entités responsables. L’objectif du relevé des services rendus est de montrer aux équipes, aux unités commerciales ou aux individus le coût des ressources cloud qu’ils consomment.

La rétrofacturation consiste à affecter les dépenses du service central aux unités de coûts sur la base d’une stratégie adaptée à un processus de gestion financière spécifique. Pour les clients, la rétrofacturation impute les coûts occasionnés par un compte de services partagés à différentes catégories de coûts financiers adaptées à un processus de signalement des clients. En mettant en place des mécanismes de rétrofacturation, vous pouvez rendre compte des coûts engendrés par les différentes unités commerciales, les produits et les équipes.

Les charges de travail peuvent être classées en deux catégories : les charges critiques et les charges non critiques. Sur la base de ce classement, utilisez des ressources partagées avec des

configurations générales pour les charges de travail moins critiques. Pour optimiser davantage les coûts, affectez des serveurs réservés uniquement pour les charges de travail critiques. Partagez les ressources ou allouez-les sur plusieurs comptes pour les gérer efficacement. Même avec des environnements de développement, de test et de production distincts, le partage sécurisé est possible et ne compromet pas la structure organisationnelle.

Pour améliorer votre compréhension et optimiser les coûts et l'utilisation des applications conteneurisées, utilisez les données de répartition des coûts qui vous aident à répartir les coûts entre les différentes entités commerciales en fonction de la façon dont l'application consomme les ressources de calcul et de mémoire partagées. Les données de répartition des coûts fractionnés vous aident à obtenir un relevé des services rendus et une rétrofacturation au niveau des tâches pour les charges de travail de conteneurs exécutées sur Amazon Elastic Container Service (Amazon ECS) ou Amazon Elastic Kubernetes Service (Amazon EKS).

Pour les architectures distribuées, créez un VPC à services partagés, qui fournit un accès centralisé aux services partagés requis par les charges de travail dans chacun des VPC. Ces services partagés peuvent inclure des ressources telles que des services d'annuaire ou des points de terminaison d'un VPC. Pour réduire les frais généraux administratifs et les coûts, partagez les ressources depuis un emplacement central au lieu de les créer dans chaque VPC.

Lorsque vous utilisez des ressources partagées, vous pouvez économiser sur les coûts opérationnels, optimiser l'utilisation des ressources et améliorer la cohérence. Dans une conception multicompte, vous pouvez héberger certains services AWS de manière centralisée et y accéder à l'aide de plusieurs applications et comptes dans un hub pour réduire les coûts. Vous pouvez utiliser AWS Resource Access Manager (AWS RAM) pour partager d'autres ressources communes, telles que des sous-réseaux VPC et des attachements AWS Transit Gateway, AWS Network Firewall ou des pipelines d'IA Amazon SageMaker. Dans un environnement multicompte, utilisez AWS RAM pour créer une ressource une fois et la partager avec d'autres comptes.

Les organisations doivent baliser les coûts partagés de manière efficace et vérifier qu'aucune partie significative de leurs coûts ne reste non balisée ou non allouée. Si vous ne répartissez pas les coûts partagés de manière efficace et que personne n'assume la responsabilité de la gestion partagée des coûts, les coûts du cloud partagé peuvent monter en flèche. Vous devez savoir où vous avez engagé des coûts au niveau des ressources, de la charge de travail, de l'équipe ou de l'organisation, car ces informations vous permettent de mieux comprendre la valeur fournie au niveau concerné par rapport aux résultats commerciaux obtenus. En fin de compte, les entreprises bénéficient des économies réalisées grâce au partage de l'infrastructure cloud. Encouragez la répartition des coûts sur les ressources cloud partagées afin d'optimiser les dépenses liées au cloud.

Étapes d'implémentation

- Évaluation des ressources existantes : passez en revue les charges de travail existantes qui utilisent des services similaires pour votre charge de travail. En fonction des composants de la charge de travail, considérez les plateformes existantes si la logique métier ou les exigences techniques le permettent.
- Utilisation du partage des ressources en AWS RAM et restriction en conséquence : utilisez la AWS RAM pour partager des ressources avec d'autres comptes AWS au sein de votre organisation. Lorsque vous partagez des ressources, vous n'avez pas besoin de dupliquer les ressources sur plusieurs comptes, ce qui réduit la charge opérationnelle liée à la maintenance des ressources. Ce processus vous aide également à partager en toute sécurité les ressources que vous avez créées avec les rôles et les utilisateurs de votre compte et avec d'autres Comptes AWS.
- Balisage des ressources : balisez les ressources susceptibles d'être concernées par des rapports sur les coûts et classez-les dans des catégories de coûts. Activez ces balises de ressources liées aux coûts pour la répartition des coûts afin de fournir une visibilité sur l'utilisation des ressources AWS. Concentrez-vous sur la création d'un niveau de granularité approprié en ce qui concerne la visibilité des coûts et de l'utilisation, et influencez les comportements de consommation du cloud grâce à des rapports sur la répartition des coûts et au suivi des KPI.

Ressources

Bonnes pratiques associées :

- SEC03-BP08 Partager des ressources en toute sécurité au sein de votre organisation

Documents connexes :

- Présentation de AWS Resource Access Manager
- Services AWS que vous pouvez utiliser avec AWS Organizations
- Ressources AWS partageables
- Requêtes sur le coût et l'utilisation (CUR) d'AWS

Vidéos connexes :

- AWS Resource Access Manager - granular access control with managed permissions
- How to design your AWS cost allocation strategy

- Catégories de coûts AWS

Exemples connexes :

- Comment rétrofacturer des services partagés : exemple AWS Transit Gateway
- Comment créer un modèle de rétrofacturation/relevé des services reçus pour les Savings Plans à l'aide des requêtes sur les coûts et l'utilisation (CUR)
- Utilisation du partage VPC pour une architecture rentable de microservices à plusieurs comptes
- Amélioration de la visibilité des coûts d'Amazon EKS avec les données de répartition des coûts fractionnés AWS
- Amélioration de la visibilité des coûts d'Amazon ECS et de AWS Batch avec les données de répartition des coûts fractionnés AWS

COÛT 7. Comment utiliser les modèles de tarification pour réduire les coûts ?

Utilisez le modèle de tarification qui convient le mieux à vos ressources pour réduire les dépenses.

Bonnes pratiques

- COST07-BP01 Effectuer une analyse du modèle de tarification
- COST07-BP02 Choisissez les régions en fonction du coût
- COST07-BP03 Sélectionnez des accords avec des tiers avec des conditions rentables
- COST07-BP04 Mettre en œuvre des modèles de tarification pour tous les composants de cette charge de travail
- COST07-BP05 Effectuer une analyse du modèle de tarification au niveau du compte de gestion

COST07-BP01 Effectuer une analyse du modèle de tarification

Analysez chaque composant de la charge de travail. Déterminez si le composant et les ressources fonctionneront pendant des périodes prolongées (pour les réductions d'engagement), ou dynamiques et de courte durée (pour les instances Spot ou à la demande). Effectuez une analyse de la charge de travail à l'aide des recommandations des outils de gestion des coûts et appliquez des règles métier à ces recommandations pour obtenir des rendements élevés.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d’implémentation

AWS propose plusieurs modèles de tarification qui vous permettent de payer vos ressources de la manière la plus rentable, en fonction des besoins de votre organisation et du produit. Travaillez avec vos équipes pour déterminer le modèle de tarification le plus approprié. Souvent, votre modèle de tarification consiste en une combinaison de plusieurs options, en fonction de votre disponibilité.

Les instances à la demande vous permettent de payer la capacité de calcul ou de base de données à l’heure ou à la seconde (60 secondes au minimum) en fonction des instances que vous exécutez, sans engagements à long terme ou paiements initiaux.

Les Savings Plans sont un modèle de tarification flexible qui propose des prix bas sur AmazonEC2, Lambda et sur AWS Fargate l'utilisation, en échange d'un engagement à utiliser régulièrement (mesuré en dollars par heure) sur des durées d'un an ou trois ans.

Les instances Spot sont un mécanisme de EC2 tarification d'Amazon qui vous permet de demander de la capacité de calcul supplémentaire à un tarif horaire réduit (jusqu'à 90 % de réduction sur le prix à la demande) sans engagement initial.

Les instances réservées vous permettent de bénéficier d’une réduction allant jusqu’à 75 % en prépayant la capacité. Pour plus de détails, consultez la section Optimisation des coûts avec les réservations.

Vous pouvez choisir d’inclure un Savings Plan pour les ressources associées aux environnements de production, de qualité et de développement. Comme les ressources de l’environnement de test (sandbox) ne sont activées qu’en cas de besoin, vous pouvez également choisir un modèle à la demande pour les ressources de cet environnement. Utilisez les instances Amazon Spot pour réduire les EC2 coûts Amazon ou utilisez Compute Savings Plans pour réduire les coûts liés à AmazonEC2, Fargate et Lambda. L’outil de recommandations AWS Cost Explorer offre des opportunités de remises d’engagement avec les Savings Plans.

Si vous avez acheté des instances réservées pour Amazon EC2 par le passé ou si vous avez établi des pratiques de répartition des coûts au sein de votre organisation, vous pouvez continuer à utiliser les instances EC2 réservées Amazon pour le moment. Cependant, nous recommandons une stratégie visant à utiliser des Savings Plans à l’avenir comme un mécanisme plus flexible de réduction des coûts. Vous pouvez actualiser les recommandations de Savings Plans (SP) AWS Cost Management pour générer de nouvelles recommandations de Savings Plans à tout moment. Utilisez les instances réservées (RI) pour réduire les coûts d'AmazonRDS, Amazon Redshift ElastiCache, Amazon et Amazon OpenSearch Service. Les Savings Plans et les instances réservées sont

disponibles en trois options : paiement intégral à l'avance, avance sur le paiement et aucun paiement initial. Utilisez les recommandations fournies dans les recommandations d'achat du AWS Cost Explorer RI et du SP.

Pour trouver des opportunités de charges de travail Spot, utilisez une vue horaire de votre utilisation globale et recherchez des périodes régulières d'évolution d'utilisation ou d'élasticité. Vous pouvez utiliser des instances Spot pour des applications flexibles et tolérantes aux pannes Les exemples incluent les serveurs Web apatrides, les API points de terminaison, les applications de mégadonnées et d'analyse, les charges de travail conteneurisées, le CI/CD et d'autres charges de travail flexibles.

Analysez vos RDS instances Amazon EC2 et Amazon pour savoir si elles peuvent être désactivées lorsque vous ne les utilisez pas (après les heures de bureau et les week-ends). Cette approche vous permettra de réduire les coûts de 70 % ou plus par rapport à leur utilisation 24 heures sur 24 et 7 jours sur 7. Si vous avez des clusters Amazon Redshift qui ne doivent être disponibles qu'à des moments précis, vous pouvez mettre le cluster en pause et reprendre son utilisation plus tard. Lorsque le cluster Amazon Redshift ou l'RDSinstance Amazon et EC2 Amazon sont arrêtés, la facturation du calcul est interrompue et seuls les frais de stockage s'appliquent.

Notez que les réservations de capacité à la demande (ODCR) ne constituent pas un discount tarifaire. Les réserves de capacité sont facturées au tarif à la demande équivalent, que vous exécutiez des instances dans la capacité réservée ou non. Pensez à cette option lorsque vous devez fournir une capacité suffisante pour les ressources que vous prévoyez d'exploiter. ODCRsne doivent pas nécessairement être liés à des engagements à long terme, car ils peuvent être annulés lorsque vous n'en avez plus besoin, mais ils peuvent également bénéficier des remises proposées par les Savings Plans ou les instances réservées.

Étapes d'implémentation

- Analyse de l'élasticité de la charge de travail : utilisez la granularité horaire dans Cost Explorer ou dans un tableau de bord personnalisé pour analyser l'élasticité de votre charge de travail. Recherchez les modifications régulières du nombre d'instances en cours d'exécution. Les instances de courte durée sont de bonnes candidates pour les instances Spot ou les parcs d'instances Spot.
  - Atelier Well-Architected : explorateur de coûts
  - Atelier Well-Architected : visualisation des coûts
- Passage en revue des contrats de tarification existants : passez en revue les contrats ou les engagements en cours pour les besoins à long terme. Analysez ce dont vous disposez actuellement et le degré d'utilisation de ces engagements. Tirez parti des remises contractuelles

ou des accords d'entreprise préexistants. Les contrats d'entreprise offrent aux clients la possibilité de personnaliser les accords qui répondent le mieux à leurs besoins. Pour les engagements à long terme, envisagez des remises sur les tarifs réservés, des instances réservées ou des Savings Plans pour le type d'instance, la famille d'instances et les zones de disponibilité spécifiques. Région AWS

- Analyse des remises sur les engagements : à l'aide de Cost Explorer dans votre compte, consultez les recommandations relatives aux Savings Plans et aux instances réservées. Pour mettre en œuvre les recommandations correctes avec les réductions et les risques requis, suivez les recommandations des ateliers Well-Architected.

Ressources

Documents connexes :

- Accès aux recommandations d'instances réservées
- Options d'achat d'instance
- AWS Entreprise

Vidéos connexes :

- Save up to 90% and run production workloads on Spot

Exemples connexes :

- Atelier Well-Architected : explorateur de coûts
- Atelier Well-Architected : visualisation des coûts
- Atelier Well-Architected : modèles de tarification

COST07-BP02 Choisissez les régions en fonction du coût

La tarification des ressources peut être différente dans chaque région. Identifiez les différences de coûts entre régions et déployez uniquement dans les régions aux coûts plus élevés afin de répondre aux exigences de latence, de résidence des données et de souveraineté des données. En intégrant le coût de la région, vous payez le prix global le plus bas pour cette charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d’implémentation

L'AWS Cloud infrastructure est mondiale, hébergée sur plusieurs sites dans le monde entier et construite autour de zones de disponibilité Régions AWS, de zones locales, d' AWS Outposts et de zones de longueur d'onde. Une région est un emplacement physique dans le monde et chaque région est une zone géographique distincte AWS dotée de plusieurs zones de disponibilité. Les zones de disponibilité, qui sont plusieurs emplacements isolés dans chaque région, consistent en un ou plusieurs centres de données discrets, chacun disposant d’une alimentation, d’un réseau et d’une connectivité redondants.

Chacune Région AWS fonctionne dans les conditions du marché local, et le prix des ressources est différent dans chaque région en raison des différences dans le coût du terrain, de la fibre, de l'électricité et des taxes, par exemple. Choisissez une région spécifique pour exploiter un composant ou l’ensemble de votre solution afin que vous puissiez fonctionner au prix le plus bas possible au niveau mondial. Utilisez le calculateur AWS pour estimer les coûts de votre charge de travail dans différentes régions en cherchant des services par type d’emplacement (région, zone de longueur d’onde et zone locale) et par région.

Lorsque vous concevez vos solutions, une bonne pratique consiste à placer les ressources de calcul au plus près de l’utilisateur pour fournir une latence plus faible et une importante souveraineté des données. Sélectionner le lieu géographique en fonction de votre entreprise, votre confidentialité des données, vos performances et vos exigences en matière de sécurité. Pour les applications avec utilisateurs finaux internationaux, utilisez plusieurs emplacements.

Utilisez des régions proposant des prix inférieurs pour les AWS services afin de déployer vos charges de travail si vous n'avez aucune obligation en matière de confidentialité des données, de sécurité ou d'exigences commerciales. Par exemple, si votre région par défaut est Asie-Pacifique (Sydney) (`ap-southwest-2`), et s'il n'existe aucune restriction (confidentialité des données, sécurité, par exemple) concernant l'utilisation d'autres régions, le déploiement d'EC2instances Amazon non critiques (développement et test) dans l'est des États-Unis (Virginie du Nord) (`us-east-1`) vous coûtera moins cher.

	Conformité	Latence	Coût	Services/Fonctionnalités
Région 1	✓	15 ms	$$	✓
Région 2	✓	20 ms	$$$	X
Région 3	✓	80 ms	$	✓
Région 4	✓	15 ms	$$	✓
Région 5	✓	20 ms	$$$	X
Région 6	✓	15 ms	$	✓
Région 7	✓	80 ms	$	✓
Région 8	✓	15 ms	$	X

Tableau matriciel des fonctionnalités des régions

Le tableau matriciel précédent nous montre que la région 6 est la meilleure option pour ce scénario donné car la latence y est faible comparé aux autres régions, le service y est disponible et il s'agit de la région la moins chère.

Étapes d'implémentation

- Révision Région AWS des prix : Analysez les coûts de la charge de travail dans la région actuelle. En commençant par les coûts les plus élevés par service et par type d'utilisation, calculez les coûts dans les autres régions disponibles. Si l'économie prévue est supérieure au coût du déplacement du composant ou de la charge de travail, migrez vers la nouvelle région.
- Révision des exigences des déploiements sur plusieurs régions : analysez les exigences et les obligations de votre entreprise (confidentialité des données, sécurité ou performances) pour savoir s'il existe des restrictions vous empêchant d'utiliser plusieurs régions. Si aucune obligation ne vous restreint à utiliser une seule région, alors utilisez-en plusieurs.
- Analyse du transfert de données requises : tenez compte des coûts de transfert de données lors de la sélection des régions. Rapprochez vos données de votre client et des ressources. Choisissez les endroits les moins coûteux Régions AWS où les données circulent et où le transfert de données est minimal. En fonction des besoins de votre entreprise en matière de transfert de données, vous pouvez utiliser Amazon CloudFront, AWS PrivateLink, AWS Direct Connect, et AWS Virtual Private

Networkpour réduire vos coûts de mise en réseau, améliorer les performances et renforcer la sécurité.

Ressources

Documents connexes :

- Accès aux recommandations d'instances réservées
- EC2Tarification Amazon
- Options d'achat d'instance
- Tableau des régions

Vidéos connexes :

- Save up to 90% and run production workloads on Spot

Exemples connexes :

- Présentation des coûts de transfert des données pour les architectures courantes
- Considérations des coûts pour les déploiements mondiaux
- Éléments à prendre en compte lors de la sélection d'une région pour vos charges de travail
- Ateliers Well-Architected : restreindre l'utilisation d'un service par région (niveau 200)

COST07-BP03 Sélectionnez des accords avec des tiers avec des conditions rentables

Les accords et conditions rentables garantissent que le coût de ces services évolue en fonction des avantages qu'ils offrent. Choisissez des accords et une tarification qui évoluent lorsqu'ils apportent des avantages supplémentaires à votre organisation.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

De nombreux produits du marché peuvent vous aider à gérer les coûts de vos environnements cloud. Ils peuvent présenter des différences en matière de fonctionnalités selon les exigences des clients. Certains privilégieront la gouvernance ou la visibilité des coûts et d'autres l'optimisation des coûts, par exemple. L'un des facteurs clés pour une optimisation et une gouvernance efficaces des coûts

consiste à utiliser le bon outil avec les bonnes fonctionnalités et le bon modèle de tarification. Ces produits ont des modèles de tarification différents. Certains correspondent à un certain pourcentage de votre facture mensuelle, et d'autres à un pourcentage des économies réalisées. Idéalement, vous ne devriez payer que ce dont vous avez besoin.

Lorsque vous utilisez des solutions ou des services tiers dans le cloud, il est important que les structures de tarification soient alignées sur les résultats souhaités. La tarification doit se mettre à l'échelle en fonction des résultats et de la valeur qu'elle fournit. Par exemple, dans le cas d'un logiciel facturé à un pourcentage des économies réalisées, plus vous économisez (résultat), plus le logiciel est cher. Les contrats de licence qui prévoient un paiement proportionnel à vos dépenses ne sont pas toujours dans votre intérêt pour optimiser les coûts. Toutefois, si l'éditeur offre des avantages clairs pour toutes les parties de votre facture, ces frais progressifs peuvent être justifiés.

Par exemple, une solution qui fournit des recommandations à Amazon EC2 et facture un pourcentage du montant total de votre facture peut devenir plus onéreuse si vous utilisez d'autres services qui ne présentent aucun avantage. Prenons également l'exemple d'un service géré facturé à un pourcentage du coût des ressources gérées. Une instance de plus grande taille ne nécessite pas nécessairement plus d'efforts de gestion, mais elle peut être facturée plus cher. Vérifiez que ces accords de tarification de service incluent un programme ou des fonctions d'optimisation des coûts dans leur service afin d'améliorer leur rentabilité.

Les clients peuvent trouver ces produits du marché plus avancés ou plus faciles à utiliser. Vous devez prendre en compte le coût de ces produits et réfléchir aux possibilités d'optimisation des coûts à long terme.

Étapes d'implémentation

- Analyse des accords et des conditions des tiers : passez en revue les prix figurant dans les accords avec des tiers. Effectuez une modélisation pour différents niveaux d'utilisation et tenez compte des nouveaux coûts tels que l'utilisation de nouveaux services ou l'augmentation des services actuels en raison de la croissance de la charge de travail. Déterminez si les coûts supplémentaires apportent les avantages requis à votre entreprise.

Ressources

Documents connexes :

- Accès aux recommandations d'instances réservées
- Options d'achat d'instance

Vidéos connexes :

- Save up to 90% and run production workloads on Spot

COST07-BP04 Mettre en œuvre des modèles de tarification pour tous les composants de cette charge de travail

Les ressources fonctionnant en permanence doivent utiliser une capacité réservée telle que des Savings Plans ou des instances réservées. La capacité à court terme est configurée pour utiliser des instances Spot ou un parc d'instances Spot. Les instances à la demande ne sont utilisées que pour les charges de travail de courte durée qui ne peuvent pas être interrompues et qui ne durent pas assez longtemps pour la capacité réservée, entre 25 et 75 % de la période, selon le type de ressource.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

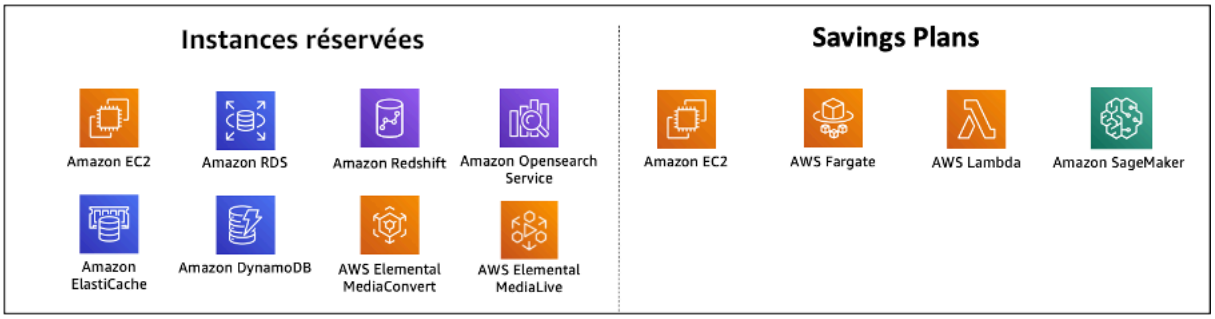
Directives d'implémentation

Pour améliorer la rentabilité, AWS fournit plusieurs recommandations d'engagement basées sur votre utilisation passée. Vous pouvez utiliser ces recommandations pour comprendre les économies que vous pouvez réaliser et comment l'engagement sera utilisé. Vous pouvez utiliser ces services comme On-Demand, Spot, ou vous engager pour une certaine période et réduire vos coûts à la demande grâce aux instances réservées (RIs) et aux Savings Plans (SPs). Vous devez comprendre non seulement chaque composant de la charge de travail et les différents AWS services, mais également les remises d'engagement, les options d'achat et les instances ponctuelles pour ces services afin d'optimiser votre charge de travail.

Tenez compte des exigences des composants de votre charge de travail et maîtrisez les différents modèles de tarification de ces services. Définissez les besoins de disponibilité de ces composants. Déterminez s'il existe plusieurs ressources indépendantes qui remplissent la fonction dans la charge de travail, et quelles sont les exigences de la charge de travail au fil du temps. Comparez le coût des ressources à l'aide du modèle de tarification à la demande par défaut et à celui des autres modèles applicables. Tenez compte de toute modification éventuelle des ressources ou des éléments de la charge de travail.

Par exemple, examinons cette architecture d'application Web sur AWS. Cet exemple de charge de travail comprend plusieurs AWS services, tels qu'Amazon Route 53, Amazon AWS WAF, les EC2 instances Amazon CloudFront, les RDS instances Amazon, les équilibreurs de charge, le stockage Amazon S3 et Amazon Elastic File System (AmazonEFS). Vous devez passer en revue chacun de

ces services et identifier les opportunités de réduction de coûts des différents modèles de tarification. Certains d'entre eux peuvent être éligibles RIs ouSPs, tandis que d'autres ne sont disponibles que sur demande. Comme le montre l'image suivante, certains AWS services peuvent être validés à l'aide de RIs ouSPs.



AWS services engagés à l'aide d'instances réservées et de plans d'épargne

Étapes d’implémentation

- Implémentation de modèles de tarification : à l’aide des résultats de vos analyses, achetez des Savings Plans, des instances réservées ou implémentez des instances Spot. S'il s'agit de votre premier achat sous engagement, choisissez les cinq ou dix meilleures recommandations de la liste, puis surveillez et analysez les résultats au cours des deux prochains mois. AWS Cost Management Console vous guide tout au long du processus. Consultez les recommandations RI ou SP de la console, personnalisez les recommandations (type, paiement et durée), passez en revue l’engagement horaire (par exemple, 20 USD/heure), puis ajoutez-les au panier. Les remises s’appliquent automatiquement à l’utilisation éligible. Achetez régulièrement un petit nombre d’engagements avec remise, par exemple toutes les deux semaines ou tous les mois. Mettez en œuvre des instances Spot pour les charges de travail qui peuvent être interrompues ou qui sont sans état. Enfin, sélectionnez des EC2 instances Amazon à la demande et allouez des ressources pour les besoins restants.
- Cycle de vérification de la charge de travail : mettez en œuvre un cycle de vérification de la charge de travail, qui analyse spécifiquement la couverture du modèle de tarification. Une fois que la charge de travail dispose de la couverture requise, achetez des engagements avec remise supplémentaires régulièrement (tous les deux ou trois mois) ou en fonction de l’évolution de la consommation de votre organisation.

Ressources

Documents connexes :

- Compréhension des recommandations de vos Savings Plans
- Accès aux recommandations d’instances réservées
- Comment acheter des instances réservées
- Options d’achat d’instance
- Instances Spot
- Modèles de réservation pour d'autres AWS services
- Services pris en charge par les Savings Plans

Vidéos connexes :

- Save up to 90% and run production workloads on Spot

Exemples connexes :

- Ce que vous devez prendre en compte avant de souscrire des Savings Plans
- Comment utiliser Cost Explorer pour analyser mes dépenses et mon utilisation ?

COST07-BP05 Effectuer une analyse du modèle de tarification au niveau du compte de gestion

Vérifiez les outils de facturation et de gestion des coûts et consultez les remises recommandées avec les engagements et les réservations pour mener une analyse régulière au niveau du compte de gestion.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d’implémentation

La modélisation régulière des coûts vous aide à mettre en œuvre les possibilités d’optimisation sur plusieurs charges de travail. Par exemple, si plusieurs charges de travail utilisent des instances à la demande au niveau agrégé, le risque de changement est moindre, et la mise en œuvre d’une réduction basée sur un engagement permet d’obtenir un coût global plus faible. Il est recommandé d’effectuer les analyses selon des cycles réguliers de deux semaines à un mois. Cela vous permet de faire de petits achats d’ajustement, de sorte que la couverture de vos modèles de tarification continue à évoluer en fonction de l’évolution de vos charges de travail et de leurs composants.

Utilisez l’outil de recommandations AWS Cost Explorer pour identifier des opportunités de remises sur engagement dans votre compte de gestion. Les recommandations au niveau du compte de

gestion sont calculées en tenant compte de l'utilisation de tous les comptes de votre organisation AWS qui ont des instances réservées (RI) ou des Savings Plans (SP). Elles sont également calculées lorsque le partage des remises est activé afin de recommander un engagement qui maximise les économies sur tous les comptes.

Bien que l'achat au niveau du compte de gestion optimise les économies maximales dans de nombreux cas, il peut arriver que vous envisagiez d'acheter SPs au niveau du compte associé, par exemple lorsque vous souhaitez que les remises s'appliquent d'abord à l'utilisation de ce compte associé en particulier. Les recommandations pour les comptes des membres sont calculées au niveau du compte individuel, afin de maximiser les économies pour chaque compte isolé. Si votre compte contient à la fois des engagements RI et SP, ils seront appliqués dans cet ordre :

1. RI zonale
2. RI standard
3. RI convertible
4. Instance Savings Plan
5. Compute Savings Plan

Si vous achetez un SP au niveau du compte de gestion, les économies seront appliquées en fonction du pourcentage de remise du plus élevé au plus bas. SPsau niveau du compte de gestion, examinez tous les comptes associés et appliquez les économies là où la réduction sera la plus élevée. Si vous souhaitez limiter les domaines dans lesquels les économies sont appliquées, vous pouvez souscrire à un Savings Plan au niveau du compte associé. Dans ce cas, chaque fois que ce compte utilisera des services de calcul éligibles, la réduction sera appliquée en premier sur ce compte. Lorsque le compte n'exécute pas de services informatiques éligibles, la réduction est partagée entre les autres comptes liés sous le même compte de gestion. Le partage des remises est activé par défaut, mais il peut être désactivé si nécessaire.

Dans une famille de facturation consolidée, les Savings Plans s'appliquent d'abord à l'utilisation du compte du propriétaire, puis à l'utilisation des autres comptes. Cela se produit uniquement si le partage est activé. Vos Savings Plans sont d'abord appliqués à votre pourcentage d'économies le plus élevé. S'il existe plusieurs utilisations avec des pourcentages d'économies identiques, les Savings Plans sont appliqués à la première utilisation avec le taux de Savings Plans le plus bas. Les Savings Plans continuent de s'appliquer jusqu'à ce qu'il n'y ait plus d'utilisations restantes ou que votre engagement soit épuisé. Toute utilisation restante est facturée aux taux à la demande. Vous pouvez actualiser les recommandations de plans d'épargne dans AWS Cost Management pour générer de nouvelles recommandations de plans d'épargne à tout moment.

Après avoir analysé la flexibilité des instances, choisissez un niveau d'engagement selon les recommandations. Créez une modélisation des coûts en analysant les coûts à court terme de la charge de travail avec différentes options de ressources potentielles, en analysant les modèles de AWS tarification et en les alignant sur les exigences de votre entreprise afin de déterminer le coût total de possession et les opportunités d'optimisation des coûts.

Étapes d'implémentation

Analyse des remises sur les engagements : à l'aide de Cost Explorer dans votre compte, consultez les recommandations relatives aux Savings Plans et aux instances réservées. Assurez-vous de comprendre les recommandations du Savings Plan et estimez vos dépenses et les économies que vous réalisez chaque mois. Examinez les recommandations au niveau du compte de gestion, qui sont calculées en tenant compte de l'utilisation de tous les comptes membres de votre organisation AWS qui comportent des instances réservées (RI) ou des Savings Plans avec le partage des remises activé. Ainsi, vous réaliserez un maximum d'économies sur tous les comptes. Vous pouvez confirmer que vous avez mis en œuvre les bonnes recommandations avec les remises et les risques requis en suivant les ateliers Well-Architected.

Ressources

Documents connexes :

- Comment fonctionne la AWS tarification ?
- Options d'achat d'instance
- Présentation du Savings Plan
- Recommandations en matière de Savings Plan
- Accès aux recommandations d'instances réservées
- Compréhension de la recommandation de vos Savings Plans
- Comment les Savings Plans s'appliquent à votre AWS consommation
- Savings Plans avec facturation consolidée
- Activation des remises sur les Savings Plans et sur instances réservées partagées

Vidéos connexes :

- Save up to 90% and run production workloads on Spot

Exemples connexes :

- Atelier Well-Architected AWS : modèles de tarification (niveau 200)
- Ateliers Well-Architected AWS : analyse des modèles de tarification (niveau 200)
- Que dois-je prendre en considération avant de souscrire un Savings Plan ?
- Comment puis-je utiliser le déploiement des Savings Plans pour réduire le risque lié à l'engagement ?
- Quand utiliser les instances Spot ?

COÛT 8. Comment planifier les frais de transfert de données ?

Veillez à planifier et à surveiller les frais de transfert de données afin de pouvoir prendre des décisions architecturales pour minimiser les coûts. Une modification architecturale minime, mais efficace, peut réduire de façon spectaculaire vos coûts d'exploitation au fil du temps.

Bonnes pratiques

- COST08-BP01 Effectuer une modélisation du transfert de données
- COST08-BP02 Sélection des composants pour optimiser les coûts de transfert de données
- COST08-BP03 Mettre en œuvre des services pour réduire les coûts de transfert de données

COST08-BP01 Effectuer une modélisation du transfert de données

Recueillez les exigences de l'organisation et procédez à la modélisation du transfert de données de la charge de travail et de chacun de ses composants. Vous identifiez ainsi le coût le plus bas pour ses besoins de transfert de données actuels.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

L'habitude de concevoir une architecture utilisant des centres de données sur site ou le manque de connaissances peut conduire à négliger les frais de transfert de données lors de la conception d'une solution dans le cloud. Les frais de transfert de données AWS entrants sont déterminés en fonction de la source, de la destination et du volume du trafic. La prise en compte de ces frais lors de la phase de conception peut permettre de réaliser des économies. Il est très important de comprendre dans quelle mesure le transfert de données intervient dans votre charge de travail, le coût du transfert et les avantages associés pour estimer avec précision le coût total de possession (TCO). Cela vous permet de prendre une décision avisée pour modifier ou accepter la décision architecturale. Par

exemple, vous pouvez avoir une configuration à plusieurs zones de disponibilité dans laquelle vous répliquez les données entre les zones de disponibilité.

Vous modélisez les composants des services qui transfèrent les données de votre charge de travail, et décidez qu'il s'agit d'un coût acceptable (semblable au paiement du calcul et du stockage dans les deux zones de disponibilité) pour atteindre la fiabilité et la résilience requises. Modélisez les coûts sur différents niveaux d'utilisation. L'utilisation de la charge de travail peut changer dans le temps, et différents services peuvent être plus rentables à différents niveaux.

Lorsque vous modélisez votre transfert de données, réfléchissez au volume de données ingérées et à leur provenance. Tenez également compte de la quantité de données traitées et de la capacité de stockage ou de calcul requise. Lors de la modélisation, suivez les bonnes pratiques de mise en réseau pour l'architecture de votre charge de travail afin d'optimiser vos coûts potentiels de transfert de données.

Ils AWS Pricing Calculator peuvent vous aider à voir les coûts estimés pour des AWS services spécifiques et le transfert de données attendu. Si une charge de travail est déjà en cours d'exécution (à des fins de test ou dans un environnement de pré-production), utilisez AWS Cost Explorer or AWS Cost and Usage Report (CUR) pour comprendre et modéliser vos coûts de transfert de données. Configurez une preuve de concept (PoC) ou testez votre charge de travail et exécutez un test avec une charge simulée réaliste. Vous pouvez modéliser vos coûts selon différentes demandes de charge de travail.

Étapes d'implémentation

- Identification des exigences : quels sont l'objectif principal et les exigences commerciales du transfert de données prévu entre la source et la destination ? Quel est le résultat commercial attendu ? Recueillez les besoins de l'entreprise et définissez le résultat attendu.
- Identifier la source et la destination : Quelles sont la source et la destination des données pour le transfert de données Régions AWS, par exemple au sein, vers AWS des services ou vers Internet ?
  - Transfert de données au sein d'un Région AWS
  - Transfert de données entre Régions AWS
  - Transfert de données vers Internet
- Identification des classifications de données : quelle est la classification des données pour ce transfert de données ? De quel type de données s'agit-il ? Quelle est la taille des données ? À quelle fréquence les données doivent-elles être transférées ? Les données sont-elles sensibles ?

- Identifier les AWS services ou les outils à utiliser : Quels sont les AWS services utilisés pour ce transfert de données ? Est-il possible d'utiliser un service déjà provisionné pour une autre charge de travail ?
- Calcul des coûts de transfert des données : utilisez la tarification AWS du modèle de transfert de données que vous avez créé précédemment pour calculer les coûts de transfert de données de la charge de travail. Calculez les coûts de transfert de données à différents niveaux d'utilisation, tant pour l'augmentation que pour la réduction de la charge de travail. Lorsqu'il existe plusieurs options pour l'architecture de la charge de travail, calculez le coût de chaque option à titre de comparaison.
- Association des coûts aux résultats : pour chaque coût de transfert de données, précisez le résultat qu'il permet d'atteindre pour la charge de travail. S'il s'agit d'un transfert entre composants, ce peut être pour le découplage. S'il s'agit d'un transfert entre zones de disponibilité, ce peut être pour la redondance.
- Création d'une modélisation du transfert de données : après avoir rassemblé toutes les informations, créez une modélisation conceptuelle du transfert de données de base pour plusieurs cas d'utilisation et différentes charges de travail.

Ressources

Documents connexes :

- AWS solutions de mise en cache
- AWS Tarification
- EC2Tarifs Amazon
- VPCTarification Amazon
- Compréhension des frais de transfert de données

Vidéos connexes :

- Monitoring and Optimizing Your Data Transfer Costs
- S3 Transfer Acceleration

Exemples connexes :

- Présentation des coûts de transfert des données pour les architectures courantes
- AWS Conseils prescriptifs pour la mise en réseau

COST08-BP02 Sélection des composants pour optimiser les coûts de transfert de données

Tous les composants sont sélectionnés, et l’architecture est conçue pour réduire les coûts de transfert des données. Cela inclut l'utilisation de composants tels que l'optimisation wide-area-network (WAN) et les configurations de zones de multidisponibilité (AZ)

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

L’architecture pour le transfert de données minimise les coûts de transfert de données. Cela peut impliquer l’utilisation de réseaux de diffusion de contenu pour localiser les données plus près des utilisateurs, ou l’utilisation de liaisons réseau dédiées depuis vos sites vers AWS. Vous pouvez également utiliser WAN l'optimisation et l'optimisation des applications pour réduire la quantité de données transférées entre les composants.

Lorsque vous transférez des données vers ou à l'intérieur du AWS Cloud, il est essentiel de connaître la destination en fonction des différents cas d'utilisation, de la nature des données et des ressources réseau disponibles afin de sélectionner les AWS services appropriés pour optimiser le transfert de données. AWS propose une gamme de services de transfert de données adaptés aux diverses exigences en matière de migration de données. Sélectionnez les options appropriées de stockage de données et de transfert de données en fonction des besoins commerciaux de votre organisation.

Lorsque vous planifiez ou passez en revue l’architecture de votre charge de travail, tenez compte des points suivants :

- Utilisez des VPC points de terminaison à l'intérieur AWS : les VPC points de terminaison permettent d'établir des connexions privées entre vous VPC et les services pris en charge AWS . Cela vous évite d’utiliser l’Internet public, qui peut engendrer des coûts de transfert de données.
- Utiliser une NAT passerelle : utilisez une NATpasserelle pour que les instances d'un sous-réseau privé puissent se connecter à Internet ou à des services extérieurs au vôtreVPC. Vérifiez si les ressources situées derrière la NAT passerelle qui envoient le plus de trafic se trouvent dans la même zone de disponibilité que la NAT passerelle. Si ce n'est pas le cas, créez de nouvelles NAT passerelles dans la même zone de disponibilité que la ressource afin de réduire les frais de transfert de données entre AZ.
- AWS Direct Connect AWS Direct Connect Use contourne l'Internet public et établit une connexion privée directe entre votre réseau local et. AWS Cela peut être plus rentable et plus cohérent que de transférer de gros volumes de données sur Internet.

- Évitez de transférer des données au-delà des frontières régionales : les transferts de données entre Régions AWS (d'une région à l'autre) entraînent généralement des frais. La décision de poursuivre dans une voie multirégionale doit être mûrement réfléchie. Pour plus de détails, consultez la section Scénarios multirégionaux.
- Surveillez le transfert de données : utilisez Amazon CloudWatch et les journaux de VPC flux pour recueillir des informations sur votre transfert de données et l'utilisation du réseau. Analysez les informations de trafic réseau capturées dans votre VPCs ordinateur, telles que l'adresse IP ou la plage de données à destination et en provenance des interfaces réseau.
- Analysez l'utilisation de votre réseau : utilisez des outils de mesure et de reporting tels que des CUDOS tableaux de bord AWS Cost Explorer, ou CloudWatch pour comprendre le coût de transfert de données de votre charge de travail.

Étapes d’implémentation

- Sélection de composants de transfert de données : en utilisant la modélisation du transfert de données expliqué dans COST08-BP01 Effectuer une modélisation du transfert de données, concentrez-vous sur les coûts de transfert de données les plus importants ou sur ce qu’ils seraient si l’utilisation de la charge de travail changeait. Recherchez d’autres architectures ou des composants supplémentaires qui suppriment ou réduisent la nécessité d’un transfert de données, ou en diminuent le coût.

Ressources

Bonnes pratiques associées :

- COST08-BP01 Effectuer une modélisation du transfert de données
- COST08-BP03 Mettre en œuvre des services pour réduire les coûts de transfert de données

Documents connexes :

- Migration des données dans le cloud
- Solutions de mise en cache AWS
- Diffusez du contenu plus rapidement avec Amazon CloudFront

Exemples connexes :

- Présentation des coûts de transfert des données pour les architectures courantes
- AWS Conseils d'optimisation du réseau
- Optimisez les performances et réduisez les coûts d'analyse du réseau avec VPC Flow Logs au format Apache Parquet

COST08-BP03 Mettre en œuvre des services pour réduire les coûts de transfert de données

Mettez en œuvre des services pour réduire le transfert de données. Par exemple, utilisez des emplacements périphériques ou des réseaux de diffusion de contenu (CDN) pour diffuser du contenu aux utilisateurs finaux, créez des couches de mise en cache devant vos serveurs d'applications ou vos bases de données, et utilisez des connexions réseau dédiées plutôt que VPN pour la connectivité au cloud.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Il existe différents AWS services qui peuvent vous aider à optimiser l'utilisation du transfert de données sur votre réseau. En fonction des composants de votre charge de travail, du type et de l’architecture du cloud, ces services peuvent vous aider à la compression, à la mise en cache, ainsi qu’au partage et à la distribution de votre trafic sur le cloud.

- Amazon CloudFront est un réseau mondial de diffusion de contenu qui fournit des données avec une faible latence et des vitesses de transfert élevées. Il place les données en cache au niveau des emplacements périphériques dans le monde entier, ce qui réduit la charge sur vos ressources. En utilisant CloudFront, vous pouvez réduire les efforts administratifs liés à la diffusion de contenu à un grand nombre d'utilisateurs dans le monde entier avec une latence minimale. Le pack d'économies en matière de sécurité peut vous aider à économiser jusqu'à 30 % sur votre CloudFront consommation si vous prévoyez d'augmenter votre consommation au fil du temps.
- AWS Direct Connect vous permet de mettre en place d’une connexion réseau dédiée depuis vos sites vers AWS. Cela peut réduire les coûts de réseau, augmenter la bande passante et fournir une expérience réseau plus constante que les connexions Internet.
- Le AWS VPN vous permet d’établir une connexion sécurisée et privée entre votre réseau privé et le réseau mondial AWS . Il est idéal pour les petits bureaux ou les partenaires commerciaux, car il fournit une connectivité simplifiée, et il s’agit d’un service entièrement géré et élastique.
- VPCLes points de terminaison permettent la connectivité entre les AWS services via un réseau privé et peuvent être utilisés pour réduire les coûts de transfert de données publiques et de

NATpasserelle. Les VPCpoints de terminaison Gateway sont gratuits et prennent en charge Amazon S3 et Amazon DynamoDB. Les VPCpoints de terminaison d'interface sont fournis par AWS PrivateLinkun tarif horaire et un coût d'utilisation par Go.

- NATles passerelles fournissent une évolutivité et une gestion intégrées pour réduire les coûts, par opposition à une instance autonome. NAT Placez les NAT passerelles dans les mêmes zones de disponibilité que les instances à fort trafic et envisagez d'utiliser des VPC points de terminaison pour les instances qui ont besoin d'accéder à Amazon DynamoDB ou Amazon S3 afin de réduire les coûts de transfert et de traitement des données.
- Utilisez AWS Snow Familydes appareils dotés de ressources informatiques pour collecter et traiter les données à la périphérie. AWS Snow Family les appareils (Snowcone, Snowball et Snowmobile) vous permettent de transférer des pétaoctets de données de manière rentable et hors ligne. AWS Cloud

Étapes d’implémentation

- Mettre en œuvre les services : sélectionnez les services AWS réseau applicables en fonction du type de charge de travail de votre service en utilisant la modélisation du transfert de données et en consultant les journaux de VPC flux. Regardez où se situent les coûts les plus élevés et les flux les plus importants. Passez en revue les AWS services et déterminez s'il existe un service qui réduit ou supprime le transfert, en particulier la mise en réseau et la diffusion de contenu. Recherchez également les services de mise en cache où il existe une répétition d’accès aux données, ou de grands volumes de données.

Ressources

Documents connexes :

- AWS Direct Connect
- AWS Découvrez nos produits
- Solutions de mise en cache AWS
- Amazon CloudFront
- AWS Snow Family
- Offre groupée Amazon CloudFront Security Savings

Vidéos connexes :

- Monitoring and Optimizing Your Data Transfer Costs
- AWS Série sur l'optimisation des coûts : CloudFront
- Comment puis-je réduire les frais de transfert de données pour ma NAT passerelle ?

Exemples connexes :

- Comment rétrofacturer des services partagés : exemple AWS Transit Gateway
- Comprenez en profondeur les détails du transfert de AWS données à partir du rapport sur les coûts et l'utilisation à l'aide de la requête Athena et QuickSight
- Présentation des coûts de transfert des données pour les architectures courantes
- Utilisation de AWS Cost Explorer pour analyser les coûts de transfert de données
- Optimisation des coûts de vos AWS architectures en utilisant les fonctionnalités d'Amazon CloudFront
- Comment puis-je réduire les frais de transfert de données pour ma NAT passerelle ?

Gestion de la demande et offre de ressources

Question

- COÛT 9. Comment gérer la demande et offrir des ressources ?

COÛT 9. Comment gérer la demande et offrir des ressources ?

Pour une charge de travail dont les dépenses et les performances sont équilibrées, assurez-vous que tout ce que vous payez est utilisé et évitez une sous-utilisation importante des instances. Une métrique d'utilisation faussée dans un sens ou dans l'autre a un impact négatif sur votre organisation, que ce soit en termes de coûts d'exploitation (dégradation des performances due à une sur-utilisation) ou de gaspillage de dépenses AWS (en raison d'une sur-allocation).

Bonnes pratiques

- COST09-BP01 Effectuer une analyse de la demande de charge de travail
- COST09-BP02 Implémenter un tampon ou un accélérateur pour gérer la demande
- COST09-BP03 Fournir des ressources de manière dynamique

COST09-BP01 Effectuer une analyse de la demande de charge de travail

Analysez la demande de la charge de travail au fil du temps. Veillez à ce que l'analyse couvre les tendances saisonnières et représente avec précision les conditions d'exploitation pendant toute la durée de la charge de travail. L'effort d'analyse doit refléter les avantages potentiels : par exemple, le temps passé est proportionnel au coût de la charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : élevé

Directives d'implémentation

L'analyse de la demande de charge de travail pour le cloud computing implique de comprendre les modèles et les caractéristiques des tâches informatiques qui sont lancées dans l'environnement du cloud. Cette analyse aide les utilisateurs à optimiser l'affectation des ressources, à gérer les coûts et à vérifier que les performances sont conformes aux niveaux requis.

Ayez connaissance des exigences de la charge de travail. Les exigences de votre organisation doivent indiquer les délais de réponse de la charge de travail aux demandes. Le temps de réponse peut être utilisé pour déterminer si la demande est gérée ou si l'offre de ressources doit changer pour répondre à la demande.

L'analyse doit inclure la prévisibilité et la répétabilité de la demande ainsi que le taux et l'ampleur de variation de la demande. Effectuez l'analyse sur une période suffisamment longue pour intégrer toute variation saisonnière, telle que le end-of-month traitement ou les pics pendant les fêtes.

L'effort d'analyse doit refléter les avantages potentiels de la mise à l'échelle. Examinez le coût total attendu du composant, ainsi que les augmentations ou diminutions d'utilisation et de coût au cours de la durée de vie de la charge de travail.

Voici quelques aspects clés dont il faut tenir compte lors de l'analyse de la demande de charge de travail pour le cloud computing :

1. Indicateurs d'utilisation des ressources et de performance : analysez la manière dont les AWS ressources sont utilisées au fil du temps. Déterminez les schémas d'utilisation en période de pointe et en période creuse afin d'optimiser l'affectation des ressources et les stratégies de mise à l'échelle. Surveillez les métriques de performance telles que les temps de réponse, la latence, le débit et les taux d'erreur. Ces métriques permettent d'évaluer l'état et l'efficacité globales de l'infrastructure cloud.
2. Comportement de mise à l'échelle des utilisateurs et des applications : comprenez le comportement des utilisateurs et son impact sur la charge de travail. L'examen des schémas de

trafic des utilisateurs permet d'améliorer la diffusion du contenu et la réactivité des applications. Analysez l'évolution des charges de travail en fonction de l'augmentation de la demande. Déterminez si les paramètres d'autoscaling sont configurés correctement et efficacement pour gérer les fluctuations de charge.

3. Types de charges de travail : identifiez les différents types de charges de travail s'exécutant dans le cloud, comme le traitement par lots, le traitement des données en temps réel, les applications web, les bases de données ou le machine learning. Chaque type de charge de travail peut avoir des besoins en ressources et des profils de performance différents.
4. Contrats de niveau de service (SLAs) : comparez les performances réelles SLAs pour garantir la conformité et identifier les domaines à améliorer.

Vous pouvez utiliser Amazon CloudWatch pour collecter et suivre les métriques, surveiller les fichiers journaux, définir des alarmes et réagir automatiquement aux modifications de vos AWS ressources. Vous pouvez également utiliser Amazon CloudWatch pour obtenir une visibilité à l'échelle du système sur l'utilisation des ressources, les performances des applications et la santé opérationnelle.

Avec AWS Trusted Advisor, vous pouvez provisionner vos ressources en suivant les bonnes pratiques pour améliorer les performances et la fiabilité du système, renforcer la sécurité et rechercher des possibilités d'économies. Vous pouvez également désactiver les instances hors production et utiliser Amazon CloudWatch et Auto Scaling pour répondre aux augmentations ou aux baisses de la demande.

Enfin, vous pouvez utiliser AWS Cost ExplorerAmazon QuickSight avec le fichier AWS Cost and Usage Report (CUR) ou les journaux de votre application pour effectuer une analyse avancée de la demande de charge de travail.

Globalement, une analyse complète de la demande de charge de travail permet aux organisations de prendre des décisions éclairées sur le provisionnement, la mise à l'échelle et l'optimisation des ressources, ce qui se traduit par une amélioration des performances, de la rentabilité et de la satisfaction des utilisateurs.

Étapes d'implémentation

- Analyse des données de la charge de travail existante : analysez les données de la charge de travail existante, des versions précédentes de la charge de travail ou des modèles d'utilisation prévus. Utilisez Amazon CloudWatch, les fichiers journaux et les données de surveillance pour mieux comprendre comment la charge de travail a été utilisée. Analysez un cycle complet de la charge de travail et collectez des données pour détecter les changements saisonniers tels

que end-of-month end-of-year les événements. L'effort reflété dans l'analyse doit refléter les caractéristiques de la charge de travail. L'effort le plus important doit porter sur les charges de travail à forte valeur ajoutée qui subissent les plus grandes variations dans la demande. Le moindre effort doit porter sur les charges de travail de faible valeur ajoutée qui subissent des variations minimes dans la demande.

- Prévision de l'influence extérieure : rencontrez les membres des équipes de toute l'organisation qui peuvent influencer ou modifier la demande dans la charge de travail. Les équipes communes sont celles des ventes, du marketing ou du développement commercial. Collaborez avec elles pour connaître les cycles qu'elles appliquent et déterminer s'il existe des événements susceptibles de modifier la demande de la charge de travail. Prévoyez la demande de la charge de travail à l'aide de ces données.

Ressources

Documents connexes :

- Amazon CloudWatch
- AWS Trusted Advisor
- AWS X-Ray
- AWS Auto Scaling
- AWS Instance Scheduler
- Commencer à utiliser Amazon SQS
- AWS Cost Explorer
- Amazon QuickSight

Vidéos connexes :

Exemples connexes :

- Surveillance, suivi et analyse pour l'optimisation des coûts
- Recherche et analyse des connexions CloudWatch

COST09-BP02 Implémenter un tampon ou un accélérateur pour gérer la demande

La mise en mémoire tampon et la limitation modifient la charge de travail en atténuant les pics éventuels. Mettez en œuvre une limitation lorsque vos clients effectuent de nouveaux essais. Mettez

en œuvre une mémoire tampon pour stocker la demande et reporter le traitement. Veillez à ce que vos limitations et mémoires tampons soient conçues de manière à ce que les clients reçoivent une réponse dans les délais requis.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Dans le cloud computing, la mise en place d'une réserve ou d'une limitation est cruciale pour gérer la demande et réduire la capacité allouée requise pour votre charge de travail. Pour des performances optimales, il est essentiel d'évaluer la demande totale, y compris les pics, le taux de variation des demandes et le temps de réponse nécessaire. Lorsque les clients ont la possibilité de renvoyer leurs demandes, il devient pratique d'appliquer la limitation. À l'inverse, pour les clients qui ne disposent pas de fonctionnalités de nouvelle tentative, l'approche idéale consiste à mettre en œuvre une mémoire tampon. Ces mémoires tampons rationalisent l'afflux de demandes et optimisent l'interaction des applications avec des vitesses opérationnelles variées.

Courbe de demande avec deux pics distincts nécessitant une capacité allouée élevée

Prenons l'exemple d'une charge de travail dont la courbe de demande est représentée dans l'image précédente. Cette charge de travail a deux pics, et pour gérer ces pics, la capacité des ressources comme indiqué par la ligne orange est allouée. Les ressources et l'énergie utilisées pour cette charge de travail ne sont pas indiquées par la zone sous la courbe de la demande, mais par la zone sous la ligne de la capacité allouée, car cette dernière est nécessaire pour gérer ces deux pics. L'aplatissement de la courbe de demande de la charge de travail peut vous aider à réduire la capacité allouée pour une charge de travail et à réduire son impact environnemental. Pour atténuer le pic, envisagez de mettre en œuvre une limitation ou une mise en mémoire tampon.

Pour mieux les comprendre, examinons les notions de limitation et de mise en mémoire tampon.

Limitation : si la source de la demande a la capacité de réessayer, alors vous pouvez mettre en place une limitation. La limitation indique à la source qu'elle doit réessayer ultérieurement si elle ne peut répondre à la demande actuellement. La source attend un certain temps, puis relance la demande. L'implémentation de la limitation a l'avantage de limiter la quantité maximale de ressources et les coûts maximaux de la charge de travail. Dans AWS, vous pouvez utiliser Amazon API Gateway pour implémenter la régulation.

Basée sur la mémoire tampon : une approche basée sur la mémoire tampon utilise des producteurs (composants qui envoient des messages à la file d'attente), des consommateurs (composants qui

reçoivent des messages de la file d'attente) et une file d'attente (qui contient les messages) pour stocker les messages. Les messages sont lus par les consommateurs et traités, ce qui permet aux messages de fonctionner au rythme qui répond aux besoins commerciaux des consommateurs. À l'aide d'une mémoire tampon, les messages des producteurs sont hébergés dans des files d'attente ou des flux, prêts à être consultés par les consommateurs en fonction de leurs besoins opérationnels.

Dans AWS, vous pouvez choisir parmi plusieurs services pour implémenter une approche de mise en mémoire tampon. Amazon Simple Queue Service (AmazonSQS) est un service géré qui fournit des files d'attente permettant à un seul consommateur de lire des messages individuels. Amazon Kinesis fournit un flux de données qui permet à de nombreux consommateurs de lire les mêmes messages.

La mise en mémoire tampon et la limitation peuvent atténuer les pics éventuels en modifiant la sollicitation de votre charge de travail. Utilisez la limitation lorsque les clients retentent des actions, et la mise en mémoire tampon pour conserver la demande et la traiter ultérieurement. Si vous utilisez une mise en mémoire tampon, créez votre charge de travail de manière à ce qu'elle réponde à la demande dans les délais requis et assurez-vous que vous êtes en mesure de traiter les demandes de travail en double. Analysez la demande globale, le taux de variation et le temps de réponse requis pour dimensionner correctement la limitation ou le tampon nécessaire.

Étapes d'implémentation

- Analyse des demandes des clients : analysez les demandes des clients afin de déterminer s'ils sont capables d'effectuer de nouveaux essais. Pour les clients qui ne peuvent pas effectuer de nouveaux essais, des mémoires tampon doivent être mises en œuvre. Analysez la demande globale, le taux de variation et le temps de réponse requis pour déterminer la taille de limitation ou de mémoire tampon nécessaire.
- Implémentation d'une mémoire tampon ou d'une limitation : implémentez une mémoire tampon ou une limitation dans la charge de travail. Une file d'attente telle qu'Amazon Simple Queue Service (AmazonSQS) peut fournir une mémoire tampon aux composants de votre charge de travail. Amazon API Gateway peut fournir une régulation pour les composants de votre charge de travail.

Ressources

Bonnes pratiques associées :

- SUS02-BP06 Mettre en œuvre la mise en mémoire tampon ou la régulation pour aplatir la courbe de demande
- REL05-BP02 Demandes d'accélérateur

Documents connexes :

- AWS Auto Scaling
- AWS Instance Scheduler
- APIPasserelle Amazon
- Amazon Simple Queue Service
- Commencer à utiliser Amazon SQS
- Amazon Kinesis

Vidéos connexes :

- Choosing the Right Messaging Service for Your Distributed App

Exemples connexes :

- Gestion et surveillance de la API régulation de vos charges de travail
- Limiter à grande échelle un système multi-tenant hiérarchisé à l'aide REST API de Gateway API
- Activation de la hiérarchisation et de la régulation dans une EKS solution Amazon SaaS multi-locataires à l'aide d'Amazon Gateway API
- Intégration d’applications à l’aide de files d’attente et de messages

COST09-BP03 Fournir des ressources de manière dynamique

Les ressources sont allouées de façon planifiée. Cela peut reposer sur la demande, par exemple, via une mise à l’échelle automatique, ou sur le temps, lorsque la demande est prévisible et que les ressources sont fournies en fonction de la durée. Ces méthodes permettent de réduire au minimum la surallocation ou la sous-allocation.

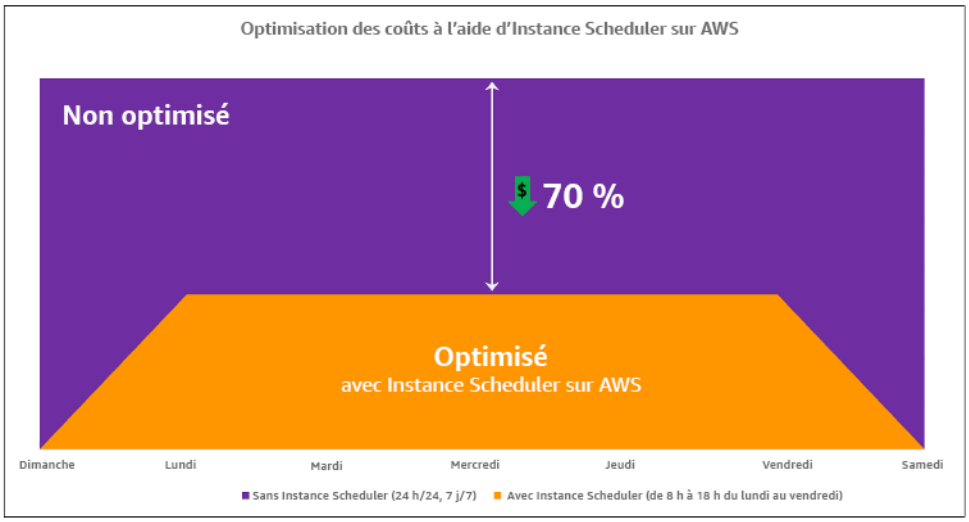
Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d’implémentation

Les AWS clients disposent de plusieurs moyens pour augmenter les ressources disponibles pour leurs applications et fournir des ressources pour répondre à la demande. L'une de ces options consiste à utiliser AWS Instance Scheduler, qui automatise le démarrage et l'arrêt des instances Amazon Elastic Compute Cloud (AmazonEC2) et Amazon Relational Database Service (Amazon). RDS L'autre option consiste à utiliser AWS Auto Scaling, ce qui vous permet de dimensionner

automatiquement vos ressources informatiques en fonction de la demande de votre application ou de votre service. La fourniture de ressources en fonction de la demande vous permettra de payer uniquement les ressources que vous utilisez, de réduire les coûts en lançant des ressources lorsqu'elles sont nécessaires et d'y mettre fin lorsqu'elles ne le sont pas.

AWS Instance Scheduler vous permet de configurer l'arrêt et le démarrage de vos RDS instances Amazon et EC2 Amazon à des heures définies afin de pouvoir répondre à la demande pour les mêmes ressources dans un calendrier cohérent, par exemple chaque jour, les utilisateurs accèdent aux EC2 instances Amazon à huit heures du matin dont ils n'ont pas besoin après six heures du soir. Cette solution permet de réduire les coûts opérationnels en arrêtant des ressources qui ne sont pas utilisées et en les redémarrant quand il le faut.



Optimisation des coûts avec AWS Instance Scheduler.

Vous pouvez également configurer facilement les plannings de vos EC2 instances Amazon sur l'ensemble de vos comptes et de vos régions à l'aide d'une interface utilisateur (UI) simple utilisant la configuration AWS Systems Manager rapide. Vous pouvez planifier des RDS instances Amazon EC2 ou Amazon avec AWS Instance Scheduler et arrêter et démarrer des instances existantes. Cependant, vous ne pouvez pas arrêter et démarrer des instances qui font partie de votre groupe Auto Scaling (ASG) ou qui gèrent des services tels qu'Amazon Redshift ou Amazon OpenSearch Service. Les groupes Auto Scaling ont leur propre planification pour les instances du groupe et ces instances sont créées.

AWS Auto Scaling vous permet d'ajuster votre capacité pour maintenir des performances stables et prévisibles au coût le plus bas possible. Il s'agit d'un service gratuit et entièrement géré destiné à augmenter la capacité de votre application qui s'intègre aux EC2 instances Amazon et aux flottes

Spot, à AmazonECS, à Amazon DynamoDB et à Amazon Aurora. L'autoscaling permet de découvrir automatiquement les ressources de votre charge de travail qui peuvent être configurées. Le service est doté de stratégies de mise à l'échelle intégrées pour optimiser les performances, les coûts ou un équilibre entre les deux et offre une mise à l'échelle prédictive pour faire face aux pics réguliers.

Plusieurs options sont disponibles pour mettre à l'échelle votre groupe Auto Scaling :

- Maintenir les niveaux d'instance actuels à tout moment
- Mise à l'échelle manuelle
- Mise à l'échelle selon un calendrier
- Mise à l'échelle en fonction de la demande
- Utiliser la mise à l'échelle prédictive

Les stratégies d'autoscaling diffèrent et peuvent être classées dans la catégorie des stratégies de mise à l'échelle dynamiques et planifiées. Les stratégies dynamiques sont une mise à l'échelle manuelle ou dynamique, une mise à l'échelle planifiée ou prédictive. Vous pouvez utiliser des stratégies de mise à l'échelle pour une mise à l'échelle dynamique, planifiée et prédictive. Vous pouvez également utiliser les métriques et les alarmes d'Amazon CloudWatch pour déclencher des événements de dimensionnement adaptés à votre charge de travail. Nous vous recommandons d'utiliser des modèles de lancement qui vous permettent d'accéder aux fonctionnalités et améliorations les plus récentes. Toutes les fonctionnalités d'autoscaling ne sont pas disponibles lorsque vous utilisez des configurations de lancement. Par exemple, vous ne pouvez pas créer un groupe Auto Scaling qui lance à la fois des instances Spot et des instances à la demande, ou qui spécifie plusieurs types d'instance. Vous devez utiliser un modèle de lancement pour configurer ces fonctions. Lorsque vous utilisez des modèles de lancement, nous vous recommandons de créer une version pour chacun d'entre eux. Avec la gestion des versions des modèles de lancement, vous pouvez créer un sous-ensemble de l'ensemble complet de paramètres. Ensuite, vous pouvez le réutiliser pour créer d'autres versions du même modèle de lancement.

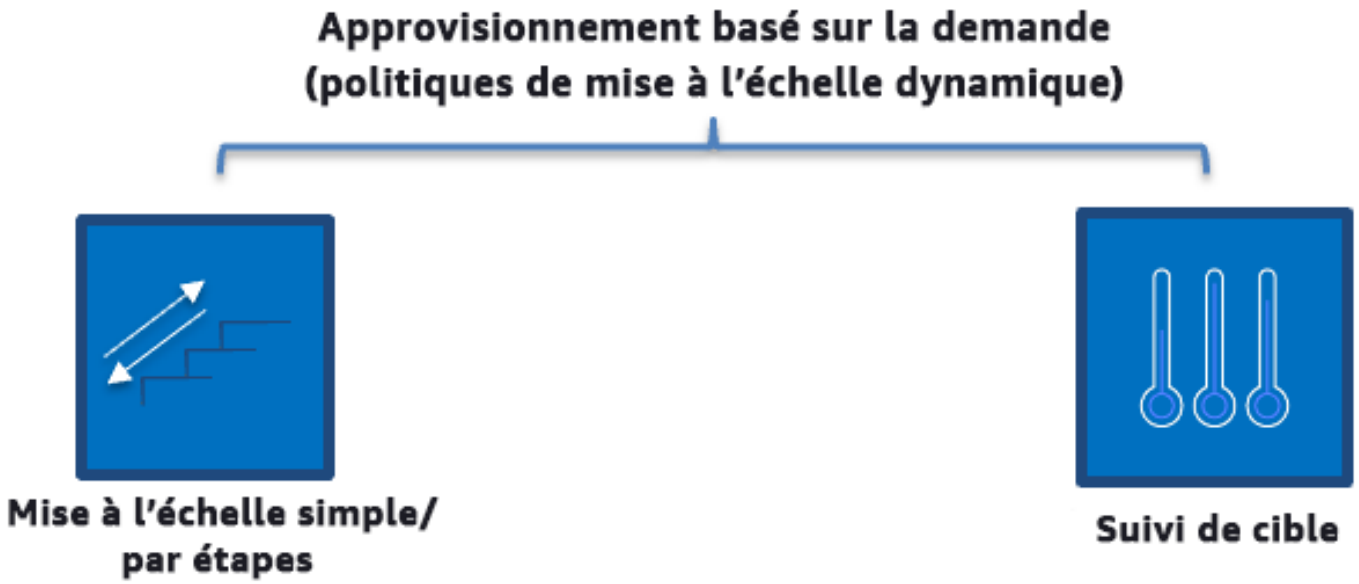
Vous pouvez utiliser AWS Auto Scaling ou intégrer la mise à l'échelle dans votre code avec AWS APIsou SDKs. Cela réduit le coût global de votre charge de travail en supprimant le coût opérationnel lié à la modification manuelle de votre environnement et les modifications peuvent être réalisées beaucoup plus rapidement. Cela adapte également les ressources de votre charge de travail à votre demande à tout moment. Afin de suivre ces bonnes pratiques et de fournir des ressources de manière dynamique à votre organisation, vous devez comprendre la mise à l' AWS Cloudéchelle horizontale et verticale ainsi que la nature des applications exécutées sur les EC2 instances Amazon.

Il est préférable que votre équipe de gestion financière du cloud travaille avec les équipes techniques afin de suivre cette bonne pratique.

Elastic Load Balancing (ELB) vous aide à effectuer une mise à l'échelle en répartissant la demande sur plusieurs ressources. Grâce à ASG Elastic Load Balancing, vous pouvez gérer les demandes entrantes en acheminant le trafic de manière optimale afin qu'aucune instance ne soit submergée dans un groupe Auto Scaling. Les demandes seraient réparties entre toutes les cibles d'un groupe cible selon une procédure circulaire sans tenir compte de la capacité ou de l'utilisation.

Les métriques typiques peuvent être des EC2 métriques Amazon standard, telles que CPU l'utilisation, le débit du réseau et la latence observée par Elastic Load Balancing pour les demandes et les réponses. Dans la mesure du possible, vous devez utiliser une métrique qui indique l'expérience du client, généralement une métrique personnalisée qui peut provenir du code d'application au sein de votre charge de travail. Pour expliquer comment répondre à la demande de manière dynamique dans ce document, nous allons regrouper l'autoscaling en deux catégories, à savoir les modèles d'offre basés sur la demande et les modèles d'offre basés sur le temps, puis approfondir chacune d'entre elles.

Offre basée sur la demande : tirez parti de l'élasticité du cloud pour fournir les ressources nécessaires à l'évolution de la demande en vous appuyant sur l'état de la demande en temps quasi réel. Pour les fonctionnalités d'approvisionnement, d'utilisation APIs ou de service basées sur la demande afin de faire varier par programmation la quantité de ressources cloud dans votre architecture. Cela vous permet de mettre à l'échelle les composants de votre architecture, d'augmenter le nombre de ressources pendant les pics de demande pour maintenir les performances, et de diminuer la capacité lorsque la demande diminue pour réduire les coûts.

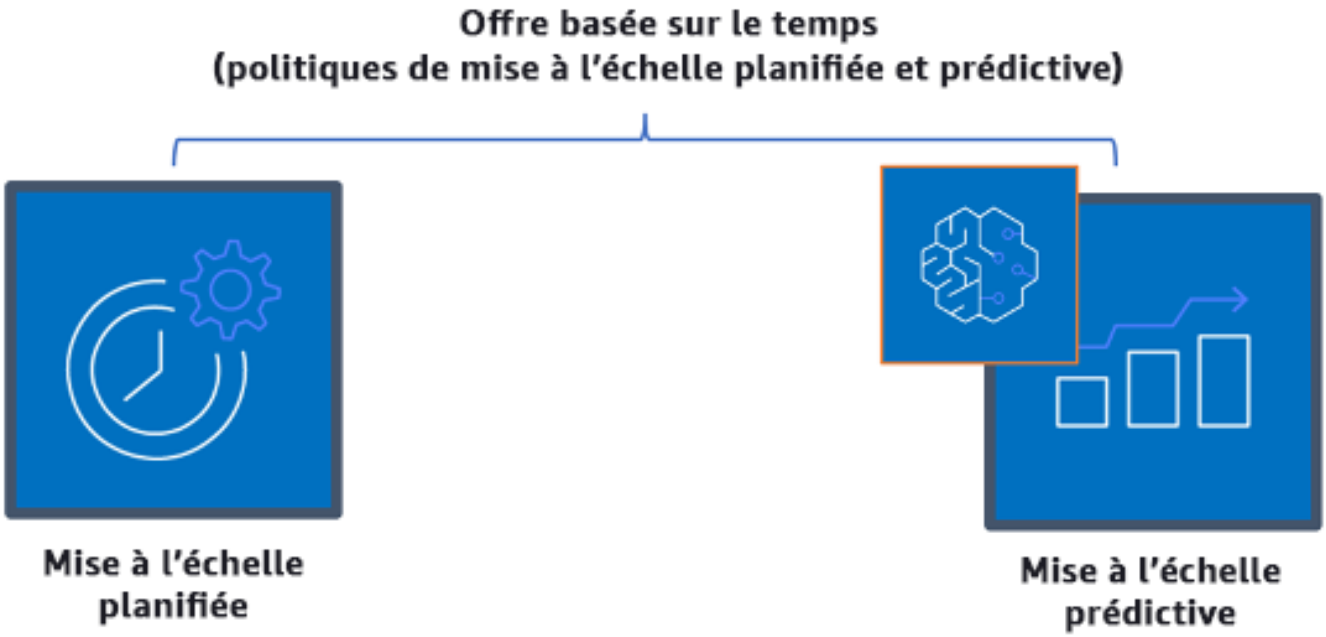


Stratégies de mise à l'échelle dynamique basées sur la demande

- Mise à l'échelle simple/par étape : surveille les métriques et ajoute/supprime des instances selon des étapes définies manuellement par les clients.
- Suivi des cibles : mécanisme de contrôle semblable à un thermostat qui ajoute ou supprime automatiquement des instances afin de maintenir les métriques à une cible définie par le client.

Lorsque vous concevez une architecture en adoptant une approche basée sur la demande, gardez à l'esprit deux considérations clés. Premièrement, vous devez comprendre à quelle vitesse vous devez allouer de nouvelles ressources. Deuxièmement, vous devez comprendre que l'importance de la marge entre l'offre et la demande variera. Vous devez être prêt à faire face au taux de variation de la demande, ainsi qu'aux défaillances de ressources.

Offre basée sur le temps : une approche fondée sur le temps permet d'aligner la capacité des ressources sur la demande qui est prévisible ou bien définie par le temps. Cette approche ne dépend généralement pas des niveaux d'utilisation des ressources. Une approche basée sur le temps garantit que les ressources sont disponibles au moment précis où elles sont nécessaires et peuvent être fournies sans aucun retard dû à des procédures de démarrage et aux vérifications du système ou de la cohérence. Grâce à une approche basée sur le temps, vous pouvez fournir des ressources supplémentaires ou augmenter la capacité pendant les périodes de pointe.



Stratégies de mise à l'échelle basées sur le temps

Vous pouvez utiliser l'autoscaling planifié pour mettre en place une approche basée sur le temps. Les charges de travail peuvent être programmées de manière à être réduites ou augmentées horizontalement à des moments définis (par exemple, au début des heures de travail), ce qui rend les ressources disponibles lorsque les utilisateurs arrivent ou que la demande augmente. La mise à l'échelle prédictive utilise des modèles pour augmenter horizontalement, tandis que la mise à l'échelle planifiée utilise des heures prédéfinies pour augmenter horizontalement. Vous pouvez également utiliser la stratégie de sélection du type d'instance (ABS) basée sur les attributs dans les groupes Auto Scaling, qui vous permet d'exprimer les besoins de votre instance sous la forme d'un ensemble d'attributs, tels que vCPU, memory et storage. Cela vous permet également d'utiliser automatiquement des types d'instances de nouvelle génération lors de leur sortie et d'accéder à une gamme plus large de capacités avec les instances Amazon EC2 Spot. Amazon EC2 Fleet et Amazon EC2 Auto Scaling sélectionnent et lancent des instances qui correspondent aux attributs spécifiés, ce qui évite de devoir sélectionner manuellement les types d'instances.

Vous pouvez également tirer parti du AWS APIsand SDKs et AWS CloudFormationpour approvisionner et mettre hors service automatiquement des environnements complets selon vos besoins. Cette approche est idéale pour les environnements de développement ou de test qui s'exécutent uniquement pendant des heures ou des périodes de travail définies. Vous pouvez

l'utiliser APIs pour redimensionner la taille des ressources au sein d'un environnement (mise à l'échelle verticale). Par exemple, vous pouvez monter en charge une charge de travail de production en modifiant la taille ou la catégorie d'instance. Cela peut être réalisé en arrêtant et en redémarrant l'instance, puis en sélectionnant une taille ou une catégorie différente. Cette technique peut également être appliquée à d'autres ressources, telles qu'Amazon EBS Elastic Volumes, qui peuvent être modifiées pour augmenter la taille, ajuster les performances (IOPS) ou modifier le type de volume en cours d'utilisation.

Lorsque vous concevez une architecture en adoptant une approche basée sur le temps, gardez à l'esprit deux considérations clés. Premièrement, dans quelle mesure le modèle d'utilisation est-il cohérent ? Deuxièmement, quel est l'impact d'un changement de modèle ? Vous pouvez augmenter la précision des prédictions en surveillant vos charges de travail et en utilisant l'informatique décisionnelle. Si vous constatez des modifications importantes dans le modèle d'utilisation, vous pouvez ajuster les heures pour vous assurer que la couverture est fournie.

Étapes d'implémentation

- Configuration d'une mise à l'échelle planifiée : pour des changements prévisibles de la demande, une mise à l'échelle temporelle peut fournir le nombre correct de ressources en temps utile. Elle est également utile si la création et la configuration des ressources ne sont pas assez rapides pour répondre à l'évolution de la demande. À l'aide de l'analyse de la charge de travail, configurez la mise à l'échelle programmée via AWS Auto Scaling. Pour configurer la planification basée sur le temps, vous pouvez utiliser le dimensionnement prédictif du dimensionnement planifié pour augmenter à l'avance le nombre d'EC2instances Amazon dans vos groupes Auto Scaling en fonction des changements de charge attendus ou prévisibles.
- Configurer le dimensionnement prédictif : le dimensionnement prédictif vous permet d'augmenter le nombre d'EC2instances Amazon dans votre groupe Auto Scaling avant les tendances quotidiennes et hebdomadaires des flux de trafic. Si vous avez des pics de trafic réguliers et des applications lentes au démarrage, vous devez envisager la mise à l'échelle prédictive. La mise à l'échelle prédictive vous permet de mettre à l'échelle le système plus rapidement en initialisant de la capacité avant d'atteindre la charge projetée par comparaison avec la mise à l'échelle dynamique seule, qui est réactive par nature. Par exemple, si les utilisateurs commencent à utiliser votre charge de travail au début des heures de bureau mais pas pendant les heures qui suivent, la mise à l'échelle prédictive peut ajouter de la capacité avant le début des heures de bureau, ce qui supprime le retard lié au fait d'attendre que la mise à l'échelle dynamique réagisse au changement de trafic.

- Configuration de la mise à l'échelle automatique dynamique : pour configurer la mise à l'échelle en fonction des mesures de la charge de travail active, utilisez l'autoscaling Utilisez l'analyse et configurez l'autoscaling pour déclencher les bons niveaux de ressources, et vérifiez que la charge de travail est mise à l'échelle dans les délais requis. Vous pouvez lancer et mettre automatiquement à l'échelle une flotte d'instances à la demande et d'instances Spot au sein d'un même groupe Auto Scaling. Outre les remises accordées sur l'utilisation des instances Spot, vous pouvez utiliser des instances réservées ou un Savings Plan afin de bénéficier de réductions sur les tarifs standard des instances à la demande. Tous ces facteurs combinés vous aident à optimiser les économies réalisées sur les EC2 instances Amazon et à obtenir l'échelle et les performances souhaitées pour votre application.

Ressources

Documents connexes :

- AWS Auto Scaling
- AWS Instance Scheduler
- Mettre la taille de votre groupe Auto Scaling à l'échelle
- Commencer à utiliser Amazon EC2 Auto Scaling
- Commencer à utiliser Amazon SQS
- Dimensionnement planifié pour Amazon EC2 Auto Scaling
- Scalage prédictif pour Amazon EC2 Auto Scaling

Vidéos connexes :

- Target Tracking Scaling Policies for Auto Scaling
- AWS Planificateur d'instances

Exemples connexes :

- Sélection du type d'instance basée sur les attributs pour Auto Scaling for Amazon EC2 Fleet
- Optimisation des coûts du service de conteneur Amazon Elastic à l'aide d'une mise à l'échelle planifiée
- Mise à l'échelle prédictive avec Amazon EC2 Auto Scaling

- Comment utiliser le planificateur d'instance pour planifier AWS CloudFormation des instances Amazon EC2 ?

Optimisation au fil du temps

Questions

- COÛT 10. Comment évaluer les nouveaux services?
- COÛT 11. Comment évaluer le coût de l’effort ?

COÛT 10. Comment évaluer les nouveaux services?

À mesure qu’AWS publie de nouveaux services et de nouvelles fonctionnalités, une bonne pratique consiste à vérifier vos choix architecturaux existants afin d’être sûr qu’ils continuent à être les plus rentables.

Bonnes pratiques

- COST10-BP01 Élaborer un processus de révision de la charge de travail
- COST10-BP02 Revoir et analyser régulièrement cette charge de travail

COST10-BP01 Élaborer un processus de révision de la charge de travail

Développez un processus qui définit les critères et le processus de révision de la charge de travail. La révision doit refléter les bénéfices potentiels. Par exemple, les charges de travail principales ou celles qui représentent plus de 10 % de la facture sont révisées chaque trimestre ou semestre, tandis que les charges de travail qui comptent pour moins de 10 % des frais sont révisées une fois par an.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : élevé

Directives d’implémentation

Pour avoir la charge de travail la plus rentable, vous devez régulièrement la vérifier pour déterminer s’il existe des possibilités de mettre en œuvre de nouveaux services, fonctionnalités et composants. Pour réduire globalement les coûts, le processus doit être proportionnel au montant potentiel des économies. Par exemple, les charges de travail qui représentent 50 % de vos dépenses totales doivent être examinées plus régulièrement et plus en profondeur que les charges de travail qui représentent 5 % de vos dépenses totales. Facteur dans tous les facteurs externes ou volatilité. Si

la charge de travail dessert une région géographique ou un segment de marché spécifique, et que l'on prévoit des changements dans ce domaine, des révisions plus fréquentes peuvent permettre de réaliser des économies. Un autre facteur à prendre en compte est l'effort de mise en œuvre des modifications. Si des coûts importants sont associés au test et à la validation des modifications, les révisions doivent être moins fréquentes.

Tenez compte du coût à long terme de la maintenance des composants et ressources obsolètes et hérités ainsi que de l'impossibilité d'y intégrer de nouvelles fonctionnalités. Le coût actuel des tests et de la validation peut dépasser l'avantage proposé. Toutefois, au fil du temps, le coût du changement peut augmenter de manière significative, car l'écart entre la charge de travail et les technologies actuelles s'accroît, ce qui entraîne des coûts encore plus élevés. Par exemple, le coût du passage à un nouveau langage de programmation peut ne pas être actuellement rentable. Toutefois, en cinq ans, le coût des personnes compétentes dans ce langage pourrait augmenter et, en raison de l'accroissement de la charge de travail, vous transféreriez un système encore plus important vers le nouveau langage, ce qui nécessiterait encore plus d'efforts qu'auparavant.

Décomposez votre charge de travail en composants, attribuez le coût du composant (une estimation est suffisante), puis énumérez les facteurs (par exemple, l'effort et les marchés extérieurs) à côté de chaque composant. Utilisez ces indicateurs pour déterminer une fréquence de révision pour chaque charge de travail. Par exemple, vous pouvez avoir des serveurs Web représentant un coût élevé, un faible effort de changement et des facteurs externes élevés, ce qui entraîne une fréquence moyenne de révision. Une base de données centrale peut être un coût moyen, impliquer un effort de modification élevé et représenter des facteurs externes faibles, ce qui se traduit par une fréquence de révision moyenne.

Définissez un processus d'évaluation de nouveaux services, des modèles de conception, des types de ressources et des configurations pour optimiser le coût de votre charge de travail au fur et à mesure que ces éléments deviennent disponibles. À l'instar des processus de révision des piliers de performance et de révision des piliers de fiabilité, identifiez, validez et hiérarchisez les activités d'optimisation et d'amélioration ainsi que la résolution des problèmes et intégrez-les à votre backlog.

Étapes d'implémentation

- Définition de la fréquence de révision : définissez la fréquence à laquelle la charge de travail et ses composants doivent être révisés. Allouez du temps et des ressources pour une amélioration continue et une fréquence de vérification afin d'améliorer l'efficacité et l'optimisation de votre charge de travail. Il s'agit d'une combinaison de facteurs qui peut varier en fonction de la charge de travail dans votre organisation et d'un composant à l'autre dans la charge de travail. Les facteurs courants incluent l'importance pour l'organisation mesurée en termes de chiffre d'affaires ou de

marque, le coût total d'exécution de la charge de travail (y compris les coûts d'exploitation et des ressources), la complexité de la charge de travail, la facilité de mise en œuvre d'un changement, les contrats de licence de logiciel et l'augmentation significative des coûts de licences pénalisants en cas de changement. Les composants peuvent être définis de manière fonctionnelle ou technique, tels que les serveurs Web et les bases de données, ou les ressources de calcul et de stockage. Équilibrez les facteurs en conséquence et développez une période pour la charge de travail et ses composants. Vous pouvez décider de réviser la charge de travail complète tous les 18 mois, les serveurs Web tous les 6 mois, la base de données tous les 12 mois, le stockage de calcul et de courte durée tous les 6 mois et le stockage de longue durée tous les 12 mois.

- Définition de la minutie des révisions : définissez les efforts consacrés à la révision de la charge de travail ou des composants de la charge de travail. Similaire à la fréquence de vérification, il s'agit d'un équilibre entre plusieurs facteurs. Évaluez et hiérarchisez vos possibilités d'amélioration afin de concentrer les efforts là où ils permettent d'obtenir les plus grands avantages, tout en estimant la quantité d'efforts nécessaire pour ces activités. Si les résultats attendus sont en deçà des objectifs et que les efforts requis sont plus coûteux, itérez alors avec d'autres plans d'action. Vos processus de vérification doivent dédier du temps et des ressources pour permettre d'effectuer des améliorations progressives continues. Par exemple, vous pouvez décider d'analyser le composant de base de données pendant une semaine, les ressources de calcul pendant une semaine et le stockage pendant quatre heures.

Ressources

Documents connexes :

- AWS Blog d'actualités
- Types de cloud computing
- Nouveautés d' AWS

Exemples associés :

- AWS Support : services proactifs
- Révisions régulières de la charge de travail pour les charges SAP de travail

COST10-BP02 Revoir et analyser régulièrement cette charge de travail

Les charges de travail existantes sont régulièrement révisées sur la base de chaque processus défini afin de déterminer si de nouveaux services peuvent être adoptés, si les services existants peuvent être remplacés ou si les charges de travail peuvent être repensées.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

AWS ajoute constamment de nouvelles fonctionnalités afin que vous puissiez expérimenter et innover plus rapidement avec les dernières technologies. AWS What's New explique comment procéder AWS et fournit un bref aperçu des AWS services, des fonctionnalités et des annonces d'expansion régionale au fur et à mesure de leur publication. Vous pouvez explorer en profondeur les lancements annoncés et les utiliser pour réviser et analyser vos charges de travail existantes. Pour tirer parti des avantages des nouveaux AWS services et fonctionnalités, vous devez évaluer vos charges de travail et implémenter de nouveaux services et fonctionnalités selon les besoins. Cela signifie que vous devrez peut-être remplacer les services existants que vous utilisez pour votre charge de travail ou moderniser votre charge de travail pour adopter ces nouveaux AWS services. Par exemple, vous pouvez réviser vos charges de travail et remplacer le composant de messagerie par Amazon Simple Email Service. Cela élimine le coût d'exploitation et de maintenance d'une flotte d'instances, tout en fournissant toutes les fonctionnalités à un coût réduit.

Pour analyser votre charge de travail et mettre en évidence les opportunités potentielles, vous devez envisager non seulement de nouveaux services mais aussi de nouvelles façons de construire des solutions. Consultez les vidéos This is My Architecture ci-dessous AWS pour en savoir plus sur les conceptions architecturales d'autres clients, leurs défis et leurs solutions. Consultez la série All-In pour découvrir les applications réelles des AWS services et les témoignages de clients. Vous pouvez également regarder la série de vidéos Retour aux fondamentaux qui explique, examine et détaille les bonnes pratiques en matière de modèles d'architecture cloud de base. Une autre source est les vidéos How to Build This, conçues pour aider les personnes à avoir de grandes idées sur la manière de donner vie à leur produit minimum viable (MVP) à l'aide de AWS services. C'est un moyen pour les constructeurs du monde entier qui ont une idée forte d'obtenir des conseils architecturaux auprès d'architectes de AWS solutions expérimentés. Enfin, vous pouvez consulter les ressources documentaires de mise en route, qui contiennent des didacticiels étape par étape.

Avant de réviser votre architecture, suivez les exigences de votre entreprise en matière de charge de travail, de sécurité et de confidentialité des données afin d'utiliser un service ou une région spécifique, et les exigences de performance tout en déroulant votre processus d'examen.

Étapes d'implémentation

- Révision régulière de la charge de travail : à l'aide de votre processus défini, effectuez des révisions à la fréquence spécifiée. Veillez à faire l'effort approprié sur chaque composant. Ce processus est similaire au processus de conception initial dans lequel vous avez sélectionné des services pour l'optimisation des coûts. Analysez les services et les avantages qu'ils apporteraient, cette fois-ci en tenant compte du coût du changement, et non seulement des avantages à long terme.
- Mise en œuvre de nouveaux services : si le résultat de l'analyse est de mettre en œuvre des modifications, effectuez d'abord une analyse de base de la charge de travail pour connaître le coût actuel de chaque sortie. Mettez en œuvre les modifications, puis effectuez une analyse pour vérifier le nouveau coût de chaque sortie.

Ressources

Documents connexes :

- AWS Blog d'actualités
- Nouveautés avec AWS
- AWS Documentation
- AWS Commencer
- AWS Ressources générales

Vidéos connexes :

- AWS - C'est mon architecture
- AWS - Retour à l'essentiel
- AWS - Série All-In
- Comment créer ceci

COÛT 11. Comment évaluer le coût de l'effort ?

Bonnes pratiques

- COST11-BP01 Automatiser les opérations

COST11-BP01 Automatiser les opérations

Évaluez les coûts d'exploitation sur le cloud, en vous concentrant sur la quantification du temps et des efforts que l'automatisation permet d'économiser dans les tâches administratives, les déploiements, l'atténuation du risque d'erreurs humaines, la conformité et d'autres opérations. Évaluez le temps et les coûts associés nécessaires aux efforts opérationnels et automatisez les tâches administratives afin de minimiser les efforts manuels dans la mesure du possible.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

L'automatisation des opérations réduit la fréquence des tâches manuelles, améliore l'efficacité et profite aux clients en offrant une expérience cohérente et fiable lors du déploiement, de l'administration ou de l'exploitation des charges de travail. Vous pouvez libérer les ressources d'infrastructure des tâches opérationnelles manuelles et les utiliser pour des tâches et des innovations de plus grande valeur, améliorant ainsi la valeur métier. Les entreprises ont besoin d'un moyen éprouvé et testé pour gérer leurs charges de travail dans le cloud. Cette solution doit être sûre, rapide et rentable, avec un minimum de risques et une fiabilité maximale.

Commencez par hiérarchiser vos activités opérationnelles en fonction de l'effort requis en examinant le coût global des opérations. Par exemple, combien de temps faut-il pour déployer de nouvelles ressources dans le cloud, apporter des modifications d'optimisation aux ressources existantes ou mettre en œuvre les configurations nécessaires ? Examinez le coût total des actions humaines en tenant compte du coût des opérations et de la gestion. Privilégiez les automatisations des tâches administratives afin de réduire l'effort humain.

L'effort de révision doit refléter le bénéfice potentiel. Par exemple, examinez le temps passé à effectuer des tâches manuellement plutôt qu'automatiquement. Donnez la priorité à l'automatisation des activités répétitives, de grande valeur, chronophages et complexes. Les activités de grande valeur ou qui présentent un risque élevé d'erreur humaine sont généralement celles qu'il vaut mieux commencer à automatiser, car le risque représente souvent un coût opérationnel supplémentaire non souhaité (par exemple, l'équipe chargée des opérations fait des heures supplémentaires).

Utilisez des outils d'automatisation tels que AWS Systems Manager ou AWS Config pour rationaliser les opérations, la conformité, le suivi, le cycle de vie et les processus de résiliation. Grâce AWS aux services, aux outils et aux produits tiers, vous pouvez personnaliser les automatisations que vous mettez en œuvre pour répondre à vos besoins spécifiques. Le tableau suivant présente certaines des fonctions et des fonctionnalités d'exploitation de base que vous pouvez réaliser avec des services AWS pour automatiser l'administration et l'exploitation :

- AWS Audit Manager: Auditez en permanence votre AWS utilisation pour simplifier l'évaluation des risques et de la conformité
- AWS Backup : gérez et automatisez la protection des données de manière centralisée.
- AWS Config : configurez les ressources de calcul, évaluez, auditez et estimez les configurations et l'inventaire des ressources.
- AWS CloudFormation : lancez des ressources à haute disponibilité avec l'infrastructure en tant que code.
- AWS CloudTrail : gestion du changement informatique, conformité et contrôle.
- Amazon EventBridge Programmez des événements et déclenchez AWS Lambda pour passer à l'action.
- AWS Lambda: Automatisez les processus répétitifs en les déclenchant par des événements ou en les exécutant selon un calendrier fixe avec AWS EventBridge.
- AWS Systems Manager : démarrer et arrêter les charges de travail, appliquer des correctifs aux systèmes d'exploitation, automatiser la configuration et assurer la gestion continue.
- AWS Step Functions : planifiez les tâches et automatisez les flux de travail.
- AWS Service Catalog : consommation de modèles, infrastructure sous forme de code avec conformité et contrôle.

Si vous souhaitez adopter des automatisations immédiatement en utilisant des AWS produits et services et si vous ne disposez pas des compétences nécessaires dans votre organisation, contactez AWS Managed Services (AMS), des services AWS professionnels ou des AWS partenaires pour favoriser l'adoption de l'automatisation et améliorer votre excellence opérationnelle dans le cloud.

AWS Managed Services (AMS) est un service qui gère AWS l'infrastructure pour le compte des entreprises clientes et partenaires. Il fournit un environnement sécurisé et conforme sur lequel vous pouvez déployer vos charges de travail. AMSutilise des modèles d'exploitation du cloud d'entreprise avec automatisation pour vous permettre de répondre aux exigences de votre organisation, de passer plus rapidement au cloud et de réduire vos coûts de gestion permanents.

AWS Les services professionnels peuvent également vous aider à atteindre les résultats commerciaux souhaités et à automatiser les opérations avec AWS. Elle aide les clients à déployer des activités informatiques automatisées, robustes et agiles, ainsi que des fonctionnalités de gouvernance optimisées pour le cloud. Pour des exemples de surveillance détaillés et les bonnes pratiques recommandées, consultez le livre blanc sur le pilier de l'excellence opérationnelle.

Étapes d’implémentation

- Créez une seule fois et déployez-en plusieurs : utilisez infrastructure-as-code cette CloudFormation option AWS SDK ou AWS CLI déployez-la une seule fois et utilisez-la plusieurs fois pour des environnements similaires ou pour des scénarios de reprise après sinistre. Utilisez des balises lors du déploiement pour suivre votre consommation comme défini dans d’autres bonnes pratiques. AWS Launch WizardÀ utiliser pour réduire le temps de déploiement de nombreuses charges de travail d'entreprise courantes. AWS Launch Wizard vous guide dans le dimensionnement, la configuration et le déploiement des charges de travail d'entreprise conformément aux AWS meilleures pratiques. Vous pouvez également utiliser le Service Catalog, qui vous permet de créer et de gérer des modèles infrastructure-as-code approuvés à utiliser AWS afin que tout le monde puisse découvrir des ressources cloud approuvées en libre-service.
- Automatisation de la conformité continue : envisagez d’automatiser l’évaluation et la correction des configurations enregistrées par rapport à des normes prédéfinies. Lorsque vous AWS Organizations associez les fonctionnalités de AWS Config et AWS CloudFormation, vous pouvez gérer et automatiser efficacement la conformité des configurations à grande échelle pour des centaines de comptes membres. Vous pouvez consulter les modifications apportées aux configurations et aux relations entre les AWS ressources et vous plonger dans l'historique d'une configuration de ressources.
- La fonction d’automatisation des tâches de surveillance d’ AWS fournit différents outils que vous pouvez utiliser pour surveiller les services. Vous pouvez configurer ces outils pour automatiser les tâches de surveillance. Créez et mettez en œuvre un plan de surveillance qui collecte les données de surveillance de toutes les parties de votre charge de travail afin de pouvoir déboguer plus facilement une panne multipoint si elle se produit. Par exemple, vous pouvez utiliser les outils de surveillance automatisés pour observer Amazon EC2 et vous signaler tout problème concernant les vérifications de l'état du système, les vérifications du statut des instances et les CloudWatch alarmes Amazon.
- Automatisation de la maintenance et des opérations : exécutez automatiquement les opérations de routine sans intervention humaine. À l'aide de AWS services et d'outils, vous pouvez choisir les AWS automatisations à mettre en œuvre et à personnaliser en fonction de vos besoins spécifiques. Par exemple, utilisez EC2Image Builder pour créer, tester et déployer des images de machines virtuelles et de conteneurs à utiliser sur site AWS ou pour appliquer des correctifs à vos EC2 instances. AWS SSM Si l'action souhaitée ne peut pas être réalisée avec les AWS services ou si vous avez besoin d'actions plus complexes avec des ressources de filtrage, automatisez vos opérations à l'aide de using AWS Command Line Interface(AWS CLI) ou AWS SDK d'outils. AWS CLI permet d'automatiser l'ensemble du processus de contrôle et de gestion des AWS services

à l'aide de scripts sans utiliser le AWS Management Console. Sélectionnez votre préférence AWS SDKs pour interagir avec les AWS services. Pour d'autres exemples de code, consultez le référentiel d'exemples de AWS SDK code.

- Création d'un cycle de vie continu grâce aux automatisations : il est important d'établir et de préserver des politiques de cycle de vie matures, non seulement pour des raisons de réglementation ou de redondance, mais également pour optimiser les coûts. Vous pouvez l'utiliser AWS Backup pour gérer et automatiser de manière centralisée la protection des données des magasins de données, tels que vos compartiments, vos volumes, vos bases de données et vos systèmes de fichiers. Vous pouvez également utiliser Amazon Data Lifecycle Manager pour automatiser la création, la conservation et la suppression de EBS snapshots et EBS de sauvegardes. AMIs
- Supprimer les ressources inutiles : il est assez courant d'accumuler des ressources inutilisées dans le sandbox ou le développement Comptes AWS. Les développeurs créent et testent divers services et ressources dans le cadre du cycle de développement normal, puis ils ne suppriment pas ces ressources lorsqu'elles ne sont plus nécessaires. Les ressources inutilisées peuvent entraîner des coûts inutiles et parfois élevés pour l'organisation. La suppression de ces ressources contribue à réduire les coûts d'exploitation de ces environnements. Assurez-vous que vos données ne sont pas nécessaires ou qu'elles sont sauvegardées en cas de doute. Vous pouvez utiliser AWS CloudFormation pour nettoyer les piles déployées, ce qui supprime automatiquement la plupart des ressources définies dans le modèle. Vous pouvez également créer une automatisation pour la suppression de AWS ressources à l'aide d'outils tels que aws-nuke.

Ressources

Documents connexes :

- Modernisation des opérations dans le AWS Cloud
- Services AWS pour l'automatisation
- Infrastructure et automatisation
- AWS Systems Manager Automation
- Surveillance automatique et surveillance manuelle
- AWS automatisations pour SAP l'administration et les opérations
- AWS Managed Services
- AWS Professional Services

Vidéos connexes :

- Automatisez la conformité continue à grande échelle dans AWS
- AWS Backup Démo : Backup entre comptes et entre régions
- Appliquer des correctifs pour vos instances Amazon EC2

Exemples connexes :

- Réinvention des opérations automatisées (partie I)
- Réinvention des opérations automatisées (partie II)
- Automatisez la suppression des AWS ressources en utilisant aws-nuke
- Supprimez les EBS volumes Amazon inutilisés en utilisant AWS Config et AWS SSM
- Automatisez la conformité continue à grande échelle dans AWS
- Automatisations informatiques avec AWS Lambda

Durabilité

Lorsque des charges de travail sont créées dans le cloud, le pilier Durabilité consiste à comprendre les impacts des services utilisés, à mesurer les impacts tout au long du cycle de vie de la totalité de la charge de travail et à appliquer des principes de conception et les bonnes pratiques afin de réduire ces impacts. Vous trouverez des recommandations sur l'implémentation dans le livre blanc Pilier Durabilité.

Domaines de bonnes pratiques

- Sélection d'une région
- Alignement à la demande
- Logiciels et architecture
- Données
- Matériel et services
- Processus et culture

Sélection d'une région

Question

SUS 1 Comment sélectionner les régions pour votre charge de travail ?

Le choix de la région en fonction de votre charge de travail influe considérablement sur ses indicateurs de performance clés, y compris les performances, les coûts et l'empreinte carbone. Pour améliorer efficacement ces indicateurs de performance clés, vous devez choisir les régions pour votre charge de travail en fonction des exigences et des objectifs de durabilité de votre entreprise.

Bonnes pratiques

SUS01-BP01 Choisir une région en fonction des exigences et des objectifs de durabilité de l'entreprise

Choisissez une région pour votre charge de travail en fonction des exigences et des objectifs de durabilité de votre entreprise afin d'optimiser ses KPI, dont les performances, les coûts et l'empreinte carbone.

Anti-modèles courants :

- Vous sélectionnez la région de la charge de travail en fonction de votre propre emplacement.
- Vous regroupez toutes les ressources de charge de travail dans un seul emplacement géographique.

Avantages liés au respect de cette bonne pratique : le fait de placer une charge de travail à proximité des projets d'énergie renouvelable d'Amazon ou des régions affichant une faible intensité de carbone publiée peut contribuer à réduire l'empreinte carbone d'une charge de travail dans le cloud.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Le AWS Cloud est un réseau en constante expansion de régions et de points de présence (PoP), avec une infrastructure de réseau mondiale les reliant entre eux. Le choix de la région en fonction de votre charge de travail influe considérablement sur ses indicateurs de performance clés, y compris les performances, les coûts et l'empreinte carbone. Pour améliorer efficacement ces indicateurs

de performance clés, vous devez choisir les régions pour votre charge de travail en fonction des exigences et des objectifs de durabilité de votre entreprise.

Étapes d'implémentation

- Présélectionner les régions potentielles : suivez ces étapes pour évaluer et présélectionner les régions potentielles pour votre charge de travail en fonction des exigences de votre entreprise, notamment de la conformité, des fonctionnalités disponibles, du coût et de la latence :
  - Confirmez que ces régions sont conformes aux réglementations locales en vigueur (par exemple, la souveraineté des données).
  - Utilisez les listes de services régionaux AWS pour vérifier si les régions disposent des services et des fonctionnalités dont vous avez besoin pour gérer votre charge de travail.
  - Calculez le coût de la charge de travail pour chaque région à l'aide du AWS Pricing Calculator.
  - Testez la latence du réseau entre les emplacements des utilisateurs finaux et chaque Région AWS.
- Choisir des régions : choisissez des régions proches de projets Amazon d'énergie renouvelable et des régions où le réseau a une intensité de carbone publiée inférieure à celle d'autres sites (ou régions).
  - Identifiez vos directives de durabilité pertinentes pour suivre et comparer les émissions de carbone d'une année à l'autre sur la base du protocole sur les gaz à effet de serre (méthodes basées sur le marché et basées sur la localisation).
  - Choisissez une région en fonction de la méthode que vous utilisez pour suivre les émissions de carbone. Pour plus de détails sur le choix d'une région en fonction de vos directives en matière de développement durable, consultez Comment sélectionner une région pour votre charge de travail en fonction des objectifs de durabilité.

Ressources

Documents connexes :

- Comprendre les estimations de vos émissions de carbone
- Amazon à travers le monde
- Méthodologie de l'énergie renouvelable
- Éléments à prendre en compte lors de la sélection d'une région pour vos charges de travail

Vidéos connexes :

- AWS re:Invent 2023 - Sustainability innovation in AWS Global Infrastructure
- AWS re:Invent 2023 – Architecture durable : passé, présent et future
- AWS re:Invent 2022 – Fournir des architectures durables et performantes
- AWSre:Invent 2022 – L'architecture de manière durable et réduisez votre AWSempreinte carbone
- AWSre:Invent 2022 – Durabilité dans AWS les infrastructures mondiales

Alignement à la demande

Question

- SUS 2 Comment aligner les ressources du cloud sur votre demande ?

SUS 2 Comment aligner les ressources du cloud sur votre demande ?

La façon dont les utilisateurs et les applications consomment vos charges de travail et d'autres ressources peut vous aider à identifier les améliorations nécessaires pour atteindre vos objectifs de durabilité. Mettez à l'échelle l'infrastructure pour répondre en permanence à la demande et vérifiez que vous n'utilisez que les ressources minimales requises pour prendre en charge vos utilisateurs. Alignez les niveaux de service sur les besoins des clients. Positionnez des ressources afin de limiter le réseau nécessaire aux utilisateurs et aux applications pour les consommer. Supprimez les ressources inutilisées. Fournissez aux membres de votre équipe des appareils qui répondent à leurs besoins et minimisent leur impact en matière de durabilité.

Bonnes pratiques

- SUS02-BP01 Mettre à l'échelle l'infrastructure de la charge de travail de façon dynamique
- SUS02-BP02 S'aligner sur les objectifs SLAs de durabilité
- SUS02-BP03 Arrêter la création et la maintenance des actifs inutilisés
- SUS02-BP04 Optimiser le placement géographique des charges de travail en fonction de leurs exigences en matière de réseau
- SUS02-BP05 Optimiser les ressources des membres de l'équipe pour les activités réalisées
- SUS02-BP06 Mettre en œuvre la mise en mémoire tampon ou la régulation pour aplatir la courbe de demande

SUS02-BP01 Mettre à l’échelle l’infrastructure de la charge de travail de façon dynamique

Utilisez l’élasticité du cloud et mettez à l’échelle votre infrastructure de façon dynamique afin de rapprocher l’offre de ressources cloud de la demande et d’éviter de sur provisionner une capacité dans votre charge de travail.

Anti-modèles courants :

- Vous ne mettez pas à l’échelle votre infrastructure avec la charge de l’utilisateur.
- Vous mettez à l’échelle manuellement votre infrastructure en permanence.
- Vous conservez une capacité accrue après un événement de mise à l’échelle au lieu de la réduire.

Avantages liés au respect de cette bonne pratique : la configuration et le test de l’élasticité de la charge de travail permettent de faire correspondre efficacement l’offre de ressources cloud à la demande et d’éviter le sur provisionnement de capacité. Vous pouvez profiter de l’élasticité du cloud pour mettre à l’échelle automatiquement la capacité pendant et après les pics de demande, afin d’utiliser uniquement le bon nombre de ressources nécessaires pour répondre aux exigences de votre entreprise.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Le cloud vous apporte la flexibilité dont vous avez besoin pour développer ou réduire vos ressources de manière dynamique via une grande variété de mécanismes afin de répondre aux fluctuations de la demande. Rapprocher de façon optimale l’offre de la demande a le plus faible impact environnemental pour une charge de travail.

La demande peut être fixe ou variable, ce qui nécessite des métriques et une automatisation pour que la gestion ne devienne pas contraignante. Les applications peuvent se mettre à l’échelle de façon verticale (dans les deux sens) en modifiant la taille de l’instance, de façon horizontale (dans les deux sens) en modifiant le nombre d’instances, ou une combinaison des deux.

Vous pouvez utiliser plusieurs approches pour rapprocher l’offre de ressources de la demande.

- Approche de suivi des objectifs : surveillez votre métrique de capacité de mise à l’échelle et augmentez ou réduisez automatiquement votre capacité selon vos besoins.
- Mise à l’échelle prédictive : mettez à l’échelle en prévision des tendances quotidiennes et hebdomadaires.

- Approche basée sur le calendrier : définissez votre propre calendrier de mise à l'échelle en fonction de changements de charge prévisibles.
- Mise à l'échelle des services : choisissez des services (tels que les services sans serveur) qui sont évolutifs de manière native dès leur conception ou qui proposent une mise à l'échelle automatique en tant que fonctionnalité.

Identifiez les périodes d'utilisation faible ou nulle, et mettez vos ressources à l'échelle afin de supprimer toute capacité excédentaire et améliorer l'efficacité.

Étapes d'implémentation

- L'élasticité correspond à l'offre de ressources dont vous disposez et à la demande pour ces ressources. Les instances, les conteneurs et les fonctions fournissent les mécanismes pour l'élasticité, soit en combinaison avec la mise à l'échelle automatique, soit en tant que fonction du service. AWS fournit une gamme de mécanismes de mise à l'échelle automatique pour veiller à ce que les charges de travail puissent réduire verticalement avec rapidité et facilité pendant les périodes de faible charge utilisateur. Voici des exemples de mécanismes de mise à l'échelle automatique :

Mécanisme de mise à l'échelle automatique	Où utiliser
Amazon EC2 Auto Scaling	Utilisez-le pour vous assurer que vous disposez du nombre adéquat d'instances Amazon EC2 disponibles pour gérer la charge utilisateur de votre application.
Application Autoscaling	Permet de mettre à l'échelle automatiquement les ressources pour des services AWS individuels au-delà d'Amazon EC2, tels que les fonctions Lambda ou les services Amazon Elastic Container Service (Amazon ECS).
Outil Cluster Autoscaler de Kubernetes	Utilisez-le pour mettre à l'échelle automatiq uement les clusters Kubernetes sur AWS.

- La mise à l'échelle est souvent abordée pour les services de calcul, tels que les instances Amazon EC2 ou les fonctions AWS Lambda. Envisagez la configuration de services non informatiques tels

que les unités de capacité de lecture et d'écriture Amazon DynamoDB ou les partitions Amazon Kinesis Data Streams pour répondre à la demande.

- Vérifiez que les métriques de l'augmentation ou de la diminution sont validées par rapport au type de charge de travail déployée. Si vous déployez une application de transcodage vidéo, une utilisation de 100 % du processeur est attendue. N'en faites pas votre métrique principale. Le cas échéant, vous pouvez utiliser une métrique personnalisée (telle que l'utilisation de la mémoire) pour votre politique de mise à l'échelle. Pour choisir les bonnes métriques, tenez compte des conseils suivants pour Amazon EC2 :
  - La métrique doit être une métrique d'utilisation valide et décrire à quel point l'instance est occupée.
  - La valeur de métrique doit augmenter ou diminuer en proportion du nombre d'instances présentes dans le groupe Auto Scaling.
- Utilisez une mise à l'échelle dynamique plutôt qu'une mise à l'échelle manuelle pour votre groupe Auto Scaling. Nous vous recommandons également d'utiliser des politiques de dimensionnement pour le suivi des cibles dans votre dimensionnement dynamique.
- Vérifiez que les déploiements de charges de travail peuvent gérer à la fois les événements d'augmentation et de diminution des charges de travail. Créez des scénarios de test pour les événements de diminution afin de vérifier que la charge de travail se comporte comme prévu et n'a aucun impact sur l'expérience utilisateur (comme la perte de sessions permanentes). Vous pouvez utiliser l'historique des activités pour vérifier une activité de mise à l'échelle dans un groupe Auto Scaling.
- Évaluez votre charge de travail pour les modèles prédictifs et mettez-la à l'échelle de manière proactive pour anticiper les changements prévisibles et prévus de la demande. Avec la mise à l'échelle prédictive, vous pouvez supprimer le besoin de surprovisionner de la capacité. Pour en savoir plus, reportez-vous à Mise à l'échelle prédictive avec Amazon EC2 Auto Scaling.

Ressources

Documents connexes :

- Mise en route avec Amazon EC2 Auto Scaling
- Mise à l'échelle prédictive pour EC2 alimentée par le machine learning
- Analyser le comportement des utilisateurs avec Amazon OpenSearch Service, Amazon Data Firehose et Kibana
- Qu'est-ce qu'Amazon CloudWatch ?

- Surveillance de la charge de la base de données avec Performance Insights sur Amazon RDS
- Présentation de la prise en charge native pour la mise à l'échelle prédictive avec Amazon EC2 Auto Scaling
- Présentation de Karpenter, un Cluster Autoscaler de Kubernetes hautement performant et open source
- Présentation approfondie d'Amazon ECS Cluster Auto Scaling

Vidéos connexes :

- AWSre:Invent 2023 - Mise à l'échelle AWS des 10 premiers millions d'utilisateurs
- AWS re:Invent 2023 - Architecture durable : passé, présent et future
- AWS re:Invent 2022 - Build a cost-, energy-, and resource-efficient compute environment
- AWS re:Invent 2022 - Scaling containers from one user to millions
- AWS re:Invent 2#023 - Scaling FM inference to hundreds of models with Amazon SageMaker AI
- AWS re:Invent 2023 - Harness the power of Karpenter to scale, optimize & upgrade Kubernetes

Exemples connexes :

- Autoscaling

SUS02-BP02 S'aligner sur les objectifs SLAs de durabilité

Passez en revue et optimisez les accords de niveau de service relatifs à la charge de travail (SLA) en fonction de vos objectifs de durabilité afin de minimiser les ressources nécessaires pour soutenir votre charge de travail tout en continuant à répondre aux besoins de l'entreprise.

Anti-modèles courants :

- La charge SLAs de travail est inconnue ou ambiguë.
- Vous définissez votre objectif SLA uniquement en fonction de la disponibilité et des performances.
- Vous utilisez le même modèle de conception (comme une architecture multi-AZ) pour toutes vos charges de travail.

Avantages de l'établissement de cette meilleure pratique : L'alignement sur les objectifs SLAs de durabilité permet une utilisation optimale des ressources tout en répondant aux besoins de l'entreprise.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d’implémentation

SLAsdéfinissez le niveau de service attendu d'une charge de travail dans le cloud, tel que le temps de réponse, la disponibilité et la conservation des données. Ils influent sur l’architecture, l’utilisation des ressources et l’impact environnemental d’une charge de travail sur le cloud. À un rythme régulier, passez en revue SLAs et faites des compromis qui réduisent considérablement l'utilisation des ressources en échange de baisses acceptables des niveaux de service.

Étapes d’implémentation

- Comprenez les objectifs de durabilité : identifiez les objectifs de durabilité de votre organisation, tels que la réduction des émissions de carbone ou l’amélioration de l’utilisation des ressources.
- Révision SLAs : Évaluez vos SLAs produits pour déterminer s'ils répondent aux besoins de votre entreprise. Si vous dépassezSLAs, effectuez un examen plus approfondi.
- Comprenez les compromis : déterminez les compromis entre la complexité de votre charge de travail (comme le volume élevé d’utilisateurs simultanés), les performances (comme la latence) et l’impact sur le développement durable (comme les ressources requises). Généralement, la priorisation de deux des facteurs se fait au détriment du troisième.
- Ajustez SLAs : ajustez le vôtre SLAs en faisant des compromis qui réduisent de manière significative les impacts sur le développement durable en échange de baisses acceptables des niveaux de service.
  - Durabilité et fiabilité : les charges de travail hautement disponibles ont tendance à consommer davantage de ressources.
  - Durabilité et performance : l’utilisation de plus de ressources pour améliorer les performances pourrait avoir un impact environnemental plus important.
  - Durabilité et sécurité : des charges de travail trop sécurisées peuvent avoir un impact environnemental plus important.
- Définissez la durabilité SLAs si possible : incluez la durabilité SLAs pour votre charge de travail. Par exemple, définissez un niveau d'utilisation minimal afin de garantir la durabilité SLA de vos instances de calcul.

- Utilisez des modèles de conception efficaces : utilisez des modèles de conception tels que les microservices AWS qui donnent la priorité aux fonctions critiques pour l'entreprise et permettent de réduire les niveaux de service (tels que les objectifs de temps de réponse ou de temps de reprise) pour les fonctions non critiques.
- Communiquez et responsabilisez : partagez-les SLAs avec toutes les parties prenantes concernées, y compris votre équipe de développement et vos clients. Utilisez les rapports pour suivre et surveiller lesSLAs. Attribuez la responsabilité d'atteindre les objectifs de durabilité de votre entrepriseSLAs.
- Utilisez des incitations et des récompenses : utilisez des incitations et des récompenses pour atteindre ou SLAs dépasser les objectifs de durabilité.
- Révisez et répétez : révisez et ajustez régulièrement vos objectifs SLAs pour vous assurer qu'ils correspondent à l'évolution des objectifs de durabilité et de performance.

Ressources

Documents connexes :

- Comprenez les modèles de résilience et les compromis pour concevoir une architecture efficace dans le cloud
- Importance du contrat de niveau de service pour les fournisseurs de SaaS

Vidéos connexes :

- AWS re:Invent 2023 - Capacité, disponibilité, rentabilité : choisissez trois
- AWS re:INVENT 2023 - Architecture durable : passé, présent et futur
- AWS re:Invent 2023 - Modèles d'intégration avancés et compromis pour les systèmes faiblement couplés
- AWS re:Invent 2022 - Fournir des architectures durables et performantes
- AWS re:Invent 2022 - Créez un environnement informatique économe en termes de coûts, d'énergie et de ressources

SUS02-BP03 Arrêter la création et la maintenance des actifs inutilisés

Mettez hors service les ressources inutilisées de votre charge de travail afin de réduire le nombre de ressources cloud nécessaires pour répondre à votre demande et minimiser le gaspillage.

Anti-modèles courants :

- Vous n'analysez pas votre application pour détecter les ressources redondantes ou qui ne sont plus nécessaires.
- Vous ne supprimez pas les ressources redondantes ou qui ne sont plus nécessaires.

Avantages de l'établissement de cette meilleure pratique : la suppression des actifs inutilisés libère des ressources et améliore l'efficacité globale de la charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Les ressources inutilisées consomment les ressources du cloud telles que l'espace de stockage et la puissance de calcul. En identifiant et en éliminant ces ressources, vous pouvez les libérer, ce qui se traduit par une architecture cloud plus efficace. Analysez régulièrement les ressources de l'application telles que les rapports pré-compilés, les jeux de données, les images statiques et les modèles d'accès aux ressources pour identifier des redondances, une sous-utilisation et d'éventuelles cibles de mise hors service. Supprimez ces ressources redondantes pour réduire le gaspillage de ressources dans votre charge de travail.

Étapes d'implémentation

- Réaliser un inventaire : réalisez un inventaire complet pour identifier tous les actifs relevant de votre charge de travail.
- Analyser l'utilisation : la surveillance continue vous permet d'identifier les actifs statiques qui ne sont plus nécessaires.
- Supprimer les actifs inutilisés : élaborez un plan pour supprimer les actifs dont vous n'avez plus besoin.
  - Avant de supprimer une ressource, évaluez l'impact de sa suppression sur l'architecture.
  - Consolidez les ressources générées qui se chevauchent afin de supprimer tout traitement redondant.
  - Mettez à jour vos applications pour ne plus produire et stocker les ressources qui ne sont pas nécessaires.
- Communiquer avec des tiers : demandez aux tiers d'arrêter de produire et de stocker les ressources gérées en votre nom qui ne sont plus nécessaires. Demandez à consolider les actifs redondants.

- Utilisez des politiques de cycle de vie : utilisez des politiques de cycle de vie pour supprimer automatiquement les actifs inutilisés.
  - Vous pouvez utiliser Amazon S3 Lifecycle afin de gérer vos objets au cours de leur cycle de vie.
  - Vous pouvez utiliser Amazon Data Lifecycle Manager pour automatiser la création, la conservation et la suppression des EBS instantanés Amazon et des produits EBS sauvegardés par AMIs Amazon.
- Examiner et optimiser : examinez régulièrement votre charge de travail pour identifier et supprimer les ressources inutilisées.

Ressources

Documents connexes :

- Optimisation de votre AWS infrastructure pour la durabilité, partie II : stockage
- Comment puis-je résilier les ressources actives dont je n'ai plus besoin sur mon ordinateur Compte AWS ?

Vidéos connexes :

- AWS re:INVENT 2023 - Architecture durable : passé, présent et futur
- AWS re:Invent 2022 - Préserver et optimiser la valeur des actifs multimédias numériques à l'aide d'Amazon S3
- AWS re:Invent 2023 - Optimisez les coûts dans vos environnements multi-comptes

SUS02-BP04 Optimiser le placement géographique des charges de travail en fonction de leurs exigences en matière de réseau

Pour votre charge de travail, sélectionnez un emplacement et des services cloud qui réduisent la distance que le trafic réseau doit parcourir et diminuent les ressources réseau totales requises pour gérer votre charge de travail.

Anti-modèles courants :

- Vous sélectionnez la région de la charge de travail en fonction de votre propre emplacement.
- Vous regroupez toutes les ressources de charge de travail dans un seul emplacement géographique.

- Tout le trafic passe par vos centres de données existants.

Avantages liés au respect de cette bonne pratique : placer une charge de travail à proximité de ses clients fournit une faible latence, tout en réduisant les mouvements de données sur le réseau ainsi que l'impact sur l'environnement.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

L' AWS Cloud infrastructure est construite autour d'options de localisation telles que les régions, les zones de disponibilité, les groupes de placement et les emplacements périphériques tels que AWS Outpostsles zones AWS locales. Ces options d'emplacement sont responsables de la connectivité entre les composants d'application, les services cloud, les réseaux périphériques et les centres de données sur site.

Analysez les modèles d'accès au réseau dans votre charge de travail pour identifier comment utiliser ces options de localisation dans le cloud et réduire la distance que le trafic réseau doit parcourir.

Étapes d'implémentation

- Analysez les modèles d'accès au réseau dans votre charge de travail afin d'identifier comment les utilisateurs utilisent votre application.
  - Utilisez des outils de surveillance, tels qu'Amazon CloudWatch AWS CloudTrail, pour recueillir des données sur les activités du réseau.
  - Analysez les données pour identifier le modèle d'accès au réseau.
- Choisissez les régions pour votre déploiement de charge de travail en fonction des éléments clés suivants :
  - Votre objectif en matière de développement durable : comme expliqué dans Sélection de la région.
  - Lieu de stockage de vos données : pour les applications utilisant de grandes quantités de données (telles que le big data et le machine learning). Le code de l'application doit s'exécuter aussi près que possible des données.
  - Lieu de stockage de vos données : pour les applications orientées utilisateur, choisissez une région (ou des régions) proche des utilisateurs de votre charge de travail.

  - Autres contraintes : tenez compte des contraintes telles que le coût et la conformité, comme expliqué dans la section Éléments à prendre en compte lors de la sélection d'une région pour vos charges de travail.
- Utilisez des solutions de mise en cache locale ou proposées par AWS pour les ressuorces fréquemment utilisées afin d'améliorer les performances, de réduire les déplacements de données et de diminuer l'impact environnemental.

Service	Utilisation
Amazon CloudFront	Utilisez-le pour mettre en cache du contenu statique tel que des images, des scripts et des vidéos, ainsi que du contenu dynamique tel que API des réponses ou des applications Web.
Amazon ElastiCache	Permet de mettre en cache du contenu pour les applications Web.
DynamoDB Accelerator	Permet d'ajouter une accélération en mémoire à vos tables DynamoDB.

- Utilisez des services capables de vous aider à exécuter du code plus proche des utilisateurs de votre charge de travail :

Service	Utilisation
Lambda@Edge	Destiné aux opérations exigeantes en puissance de calcul qui sont lancées lorsque des objets ne sont pas dans le cache.
CloudFront Fonctions Amazon	À utiliser pour des cas d'utilisation simples tels que HTTP les manipulations de demandes ou de réponses qui peuvent être initiées par des fonctions de courte durée.

Service	Utilisation
AWS IoT Greengrass	Permet d'exécuter du calcul local, une messagerie et une mise en cache de données pour les appareils connectés.

- Utilisez le regroupement de connexions afin de pouvoir réutiliser les connexions et réduire les ressources nécessaires.
- Utilisez des magasins de données distribués qui ne s'appuient pas sur des connexions persistantes ni sur des mises à jour synchrones pour des raisons de cohérence afin de servir les populations régionales.
- Remplacez la capacité du réseau statique pré-allouée par une capacité dynamique partagée, et partagez l'impact en matière de durabilité de la capacité du réseau avec d'autres abonnés.

Ressources

Documents connexes :

- Optimisation de votre AWS infrastructure pour la durabilité, partie III : mise en réseau
- ElastiCache Documentation Amazon
- Qu'est-ce qu'Amazon CloudFront ?
- CloudFront Principales fonctionnalités d'Amazon
- AWS Infrastructure mondiale
- AWS Zones locales et AWS Outposts choix de la technologie adaptée à votre charge de travail périphérique
- Groupes de placement
- AWS Zones Locales
- AWS Outposts

Vidéos connexes :

- Démystifier le transfert de données sur AWS
- Augmenter les performances du réseau sur les instances Amazon EC2 de nouvelle génération
- AWS Vidéo explicative sur les Zones Locales
- AWS Outposts: Overview and How it Works

- AWS re:Invent 2023 - Une stratégie de migration pour les charges de travail en périphérie et sur site
- AWS re:INVENT 2021 - AWS Outposts : Apporter l' AWS expérience sur site
- AWS re:Invent 2020 - AWS Wavelength : Exécutez des applications avec une latence extrêmement faible à la périphérie de la 5G
- AWS re:Invent 2022 - Zones AWS locales : création d'applications pour une périphérie distribuée
- AWS re:Invent 2021 - Création de sites Web à faible latence avec Amazon CloudFront
- AWS re:Invent 2022 - Améliorez les performances et la disponibilité avec AWS Global Accelerator
- AWS re:Invent 2022 - Construisez votre réseau étendu mondial en utilisant AWS
- AWS re:Invent 2020 : gestion du trafic mondial avec Amazon Route 53

Exemples connexes :

- AWS Ateliers de réseautage
- Architecting for sustainability - Minimize data movement across networks

SUS02-BP05 Optimiser les ressources des membres de l'équipe pour les activités réalisées

Optimisez les ressources fournies aux membres de l’équipe pour réduire l’impact sur la durabilité environnementale tout en répondant à leurs besoins.

Anti-modèles courants :

- Vous ignorez l’impact des appareils utilisés par les membres de votre équipe sur l’efficacité globale de votre application cloud.
- Vous gérez et mettez à jour manuellement les ressources utilisées par les membres de l’équipe.

Avantages liés au respect de cette bonne pratique : l'optimisation des ressources des membres de l'équipe améliore l'efficacité globale des applications compatibles avec le cloud.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d’implémentation

Comprenez les ressources que les membres de votre équipe utilisent pour consommer vos services, leur cycle de vie prévu et l’impact financier et sur la durabilité. Mettez en œuvre des stratégies pour

optimiser ces ressources. Par exemple, effectuez des opérations complexes, telles que le rendu et la compilation, sur une infrastructure évolutive hautement utilisée plutôt que sur des systèmes mono-utilisateurs puissants et sous-utilisés.

Étapes d’implémentation

- Utilisez des postes de travail économes en énergie : fournissez aux membres de l'équipe des postes de travail et des périphériques économes en énergie. Utilisez des fonctionnalités de gestion de l’alimentation efficaces (comme le mode faible consommation) sur ces appareils afin de réduire leur consommation d’énergie.
- Adopter la virtualisation : utilisez des bureaux virtuels et le streaming d'applications pour limiter les exigences liées aux mises à niveau et aux appareils.
- Encouragez la collaboration à distance : encouragez les membres de l'équipe à utiliser des outils de collaboration à distance tels qu'Amazon Chime ou AWS Wickr à réduire les déplacements et les émissions de carbone associées.
- Utilisez des logiciels économes en énergie : fournissez aux membres de l'équipe des logiciels économes en énergie en supprimant ou en désactivant les fonctionnalités et les processus inutiles.
- Gérer les cycles de vie : évaluez l'impact des processus et des systèmes sur le cycle de vie de votre appareil et choisissez des solutions qui réduisent au minimum le besoin de remplacer celui-ci tout en répondant aux exigences de l'entreprise. Entretenez et mettez à jour régulièrement les postes de travail et les logiciels afin de maintenir et d’améliorer l’efficacité.
- Gestion des appareils à distance : intégrez la gestion à distance des appareils afin de réduire les déplacements professionnels nécessaires.
  - AWS Systems Manager Fleet Manager est une interface utilisateur (UI) unifiée qui vous permet de gérer à distance vos nœuds exécutés sur site AWS ou sur site.

Ressources

Documents connexes :

- Qu'est-ce qu'Amazon WorkSpaces ?
- Optimiseur de coûts pour Amazon WorkSpaces
- Documentation Amazon AppStream 2.0
- NICE DCV

Vidéos connexes :

- Gestion des coûts pour Amazon WorkSpaces sur AWS

SUS02-BP06 Mettre en œuvre la mise en mémoire tampon ou la régulation pour aplatir la courbe de demande

La mise en mémoire tampon et la limitation aplatissent la courbe de la demande et réduisent la capacité provisionnée requise pour votre charge de travail.

Anti-modèles courants :

- Vous traitez les demandes des clients immédiatement alors que ce n'est pas nécessaire.
- Vous n'analysez pas les exigences des demandes des clients.

Avantages liés au respect de cette bonne pratique : l'aplatissement de la courbe de demande réduit la capacité provisionnée requise pour la charge de travail. En réduisant la capacité provisionnée, on réduit la consommation d'énergie et l'impact environnemental.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

L'aplatissement de la courbe de demande de la charge de travail peut vous aider à réduire la capacité provisionnée pour une charge de travail et à réduire son impact environnemental. Prenons l'exemple une charge de travail dont la courbe de demande est représentée dans la figure ci-dessous. Cette charge de travail a deux pics, et pour gérer ces pics, la capacité des ressources comme indiqué par la ligne orange est provisionnée. Les ressources et l'énergie utilisées pour cette charge de travail ne sont pas indiquées par la zone sous la courbe de la demande, mais par la zone sous la ligne de la capacité provisionnée, car la capacité provisionnée est nécessaire pour gérer ces deux pics.

Courbe de demande avec deux pics distincts nécessitant une capacité allouée élevée

Vous pouvez utiliser la mise en mémoire tampon ou la limitation pour modifier la courbe de la demande et lisser les pics, ce qui signifie moins de capacité provisionnée et moins d'énergie consommée. Mettez en œuvre la limitation lorsque vos clients peuvent effectuer de nouvelles tentatives. Mettez en œuvre une mémoire tampon pour stocker la demande et reporter le traitement.

Effet de la limitation sur la courbe de demande et la capacité provisionnée.

Étapes d’implémentation

- Analysez les demandes des clients pour déterminer comment y répondre. Les questions à se poser sont les suivantes :
  - Cette demande peut-elle être traitée de manière asynchrone ?
  - Le client a-t-il la possibilité de lancer de nouvelles tentatives ?
- Si le client a la possibilité de lancer de nouvelles tentatives, vous pouvez mettre en œuvre un système de limitation, qui indique à la source que si elle ne peut pas répondre à la demande au moment même, elle doit réessayer plus tard.
  - Vous pouvez utiliser Amazon API Gateway pour implémenter la régulation.
- Pour les clients qui ne peuvent pas effectuer de nouvelles tentatives, il faut mettre en œuvre un tampon pour aplanir la courbe de demande. Un tampon diffère le traitement des demandes, ce qui permet aux applications qui s’exécutent à différents débits de communiquer efficacement. Une approche basée sur la mémoire tampon utilise une file d’attente ou un flux pour accepter les messages des producteurs. Les messages sont lus par les consommateurs et traités, ce qui permet aux messages de fonctionner au rythme qui répond aux besoins des entreprises.
  - Amazon Simple Queue Service (AmazonSQS) est un service géré qui fournit des files d'attente permettant à un seul consommateur de lire des messages individuels.
  - Amazon Kinesis fournit un flux de données qui permet à de nombreux consommateurs de lire les mêmes messages.
- Analysez la demande globale, le taux de variation et le temps de réponse requis pour dimensionner correctement la limitation ou le tampon nécessaire.

Ressources

Documents connexes :

- Commencer à utiliser Amazon SQS
- Intégration d’applications à l’aide de files d’attente et de messages
- Gestion et surveillance de la API régulation de vos charges de travail
- Limiter à grande échelle un système multi-tenant hiérarchisé à l'aide REST API de Gateway API
- Intégration d’applications à l’aide de files d’attente et de messages

Vidéos connexes :

- AWS re:Invent 2022 - Modèles d'intégration d'applications pour les microservices
- AWS re:Invent 2023 - Économies intelligentes : stratégies d'optimisation des coûts d'Amazon EC2
- AWS re:Invent 2023 - Modèles d'intégration avancés et compromis pour les systèmes faiblement couplés

Logiciels et architecture

Question

- SUS 3 Comment tirer parti des modèles de logiciels et d’architecture afin de soutenir vos objectifs de durabilité ?

SUS 3 Comment tirer parti des modèles de logiciels et d’architecture afin de soutenir vos objectifs de durabilité ?

Mettez en œuvre des modèles permettant de lisser les charges et de conserver une haute utilisation constante des ressources déployées afin de réduire les ressources consommées. Les composants peuvent devenir inactifs s’ils ne sont pas utilisés à la suite de changements de comportement des utilisateurs dans le temps. Révisez les modèles et l’architecture afin de consolider les composants sous-utilisés et d’augmenter l’utilisation globale. Mettez hors service les composants qui ne sont plus nécessaires. Comprenez les performances des composants de vos charges de travail et optimisez les composants qui consomment le plus de ressources. Soyez conscient des appareils que vos clients utilisent pour accéder à vos services et mettez en œuvre des modèles qui réduisent le besoin de mises à niveau des appareils.

Bonnes pratiques

- SUS03-BP01 Optimiser le logiciel et l'architecture pour les tâches asynchrones et planifiées
- SUS03-BP02 Supprimer ou refactoriser les composants de charges de travail faiblement utilisés ou inutilisés
- SUS03-BP03 Optimiser les sections de votre code qui consomment le plus de temps ou de ressources
- SUS03-BP04 Optimiser l'impact sur les appareils et équipements
- SUS03-BP05 Utiliser des modèles logiciels et des architectures qui soutiennent au mieux l’accès aux données et les modèles de stockage.

SUS03-BP01 Optimiser le logiciel et l'architecture pour les tâches asynchrones et planifiées

Utilisez des modèles d’architecture et de logiciels efficaces comme ceux axés sur les files d’attente afin de maintenir une utilisation élevée et constante des ressources déployées.

Anti-modèles courants :

- Vous mettez en service trop de ressources dans votre charge de travail cloud pour répondre aux pics imprévus de la demande.
- Votre architecture ne découple pas les expéditeurs et les destinataires de messages asynchrones par un composant de messagerie.

Avantages liés au respect de cette bonne pratique :

- Des modèles de logiciels et d’architecture efficaces réduisent les ressources inutilisées dans votre charge de travail et améliorent l’efficacité globale.
- Vous pouvez mettre à l’échelle le traitement indépendamment de la réception de messages asynchrones.
- Par le biais d’un composant de messagerie, vous avez assoupli les exigences de disponibilité auxquelles vous pouvez répondre avec moins de ressources.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Utilisez des modèles d'architecture efficaces, tels que l’architecture axée sur les événements, qui permettent une utilisation uniforme des composants et minimisent le surprovisionnement de votre charge de travail. L’utilisation de modèles d’architecture efficaces réduit au minimum les ressources inutilisées en raison des changements de la demande au fil du temps.

Comprenez les exigences des composants de votre charge de travail et adoptez des modèles d’architecture qui augmentent l’utilisation globale des ressources. Mettez hors service les composants qui ne sont plus nécessaires.

Étapes d’implémentation

- Analysez la demande pour votre charge de travail afin de déterminer comment y répondre.
- Pour les demandes ou les tâches qui ne nécessitent pas de réponses synchrones, utilisez des architectures axées sur les files d’attente et des agents de travail de mise à l’échelle automatique

afin de maximiser l'utilisation. Voici quelques exemples de situations où vous pourriez envisager une architecture axée sur les files d'attente :

Mécanismes de mise en file d'attente	Description
AWS Batch files d'attente pour les emplois	AWS Batch les tâches sont soumises à une file d'attente où elles résident jusqu'à ce qu'elles puissent être planifiées pour être exécutées dans un environnement informati que.
Amazon Simple Queue Service et instances Amazon EC2 Spot	Associer des instances Amazon SQS et Spot pour créer une architecture efficace et tolérante aux pannes.

- Pour les demandes ou les tâches qui peuvent être traitées à tout moment, utilisez les mécanismes de planification afin de traiter les tâches par lots pour plus d'efficacité. Voici quelques exemples de mécanismes de planification sur AWS :

Mécanismes de planification	Description
Amazon EventBridge Scheduler	Une fonctionnalité d'Amazon EventBridge qui vous permet de créer, d'exécuter et de gérer des tâches planifiées à grande échelle.
AWS Glue calendrier basé sur le temps	Définissez un calendrier basé sur le temps pour vos robots d'exploration et vos tâches dans. AWS Glue
Tâches planifiées d'Amazon Elastic Container Service (AmazonECS)	Amazon ECS prend en charge la création de tâches planifiées. Les tâches planifiées utilisent EventBridge les règles d'Amazon pour exécuter des tâches selon un calendrier ou en réponse à un EventBridge événement.
Instance Scheduler	Configurez les plannings de démarrage et d'arrêt pour vos EC2 instances Amazon et Amazon Relational Database Service.

- Si vous utilisez des mécanismes d’interrogation et de webhooks dans votre architecture, remplacez-les par des événements. Utilisez des architectures axées sur les événements pour créer des charges de travail hautement efficaces.
- Tirez parti du mode sans serveur AWS pour éliminer l'infrastructure surprovisionnée.
- Dimensionnez les composants individuels afin d’éviter les ressources inactives attendant une entrée.
  - Vous pouvez utiliser les recommandations de redimensionnement dans AWS Cost Explorer ou AWS Compute Optimizer pour identifier les opportunités de redimensionnement.
  - Pour en savoir plus, se reporter à Dimensionnement adéquat : dimensionnement des instances en fonction des charges de travail.

Ressources

Documents connexes :

- Qu'est-ce qu'Amazon Simple Queue Service ?
- Qu'est-ce qu'Amazon MQ ?
- Mise à l'échelle basée sur Amazon SQS
- Qu'est-ce que c'est AWS Step Functions ?
- Qu'est-ce que c'est AWS Lambda ?
- Utilisation AWS Lambda avec Amazon SQS
- Qu'est-ce qu'Amazon EventBridge ?
- Gestion des flux de travail asynchrones à l'aide d'un REST API

Vidéos connexes :

- AWS re:Invent 2023 - Sur la voie de l'architecture événementielle sans serveur
- AWS re:Invent 2023 - Utilisation du mode sans serveur pour l'architecture axée sur les événements et la conception axée sur le domaine
- AWS re:Invent 2023 - Modèles avancés basés sur les événements avec Amazon EventBridge
- AWS re:INVENT 2023 - Architecture durable : passé, présent et futur
- Modèles de messages asynchrones | Événements AWS

Exemples connexes :

- Architecture axée sur les événements avec processeurs AWS Graviton et instances Amazon Spot EC2

SUS03-BP02 Supprimer ou refactoriser les composants de charges de travail faiblement utilisés ou inutilisés

Supprimez les composants inutilisés et devenus inutiles, et refactorisez les composants peu utilisés afin de minimiser le gaspillage dans votre charge de travail.

Anti-modèles courants :

- Vous ne vérifiez pas régulièrement le niveau d'utilisation des différents composants de votre charge de travail.
- Vous ne vérifiez pas les recommandations des outils de redimensionnement AWS tels que AWS Compute Optimizer.

Avantages liés au respect de cette bonne pratique : la suppression des composants inutilisés minimise le gaspillage et améliore l'efficacité globale de votre charge de travail dans le cloud.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les composants inutilisés ou sous-utilisés d'une charge de travail dans le cloud consomment des ressources de calcul, de stockage ou de réseau inutiles. Supprimez ou refactorisez ces composants pour réduire directement le gaspillage et améliorer l'efficacité globale d'une charge de travail dans le cloud. Il s'agit d'un processus d'amélioration itératif qui peut être lancé par l'évolution de la demande ou la publication d'un nouveau service cloud. Par exemple, une baisse significative de la durée d'exécution d'une fonction AWS Lambda peut indiquer la nécessité de réduire la taille de la mémoire. De plus, à mesure que AWS publie de nouveaux services et de nouvelles fonctionnalités, les services et l'architecture optimaux pour votre charge de travail peuvent changer.

Surveillez en permanence l'activité de la charge de travail et recherchez les possibilités d'améliorer le niveau d'utilisation des différents composants. En supprimant les composants inutilisés et en effectuant des activités de redimensionnement, vous répondez aux besoins de votre entreprise avec le moins de ressources cloud possible.

Étapes d'implémentation

- Inventorier vos ressources AWS : créez un inventaire de vos ressources AWS. Dans AWS, vous pouvez activer Explorateur de ressources AWS pour explorer et organiser vos ressources AWS. Pour en savoir plus, consultez AWSre:Invent 2022 – Comment gérer les ressources et les applications à grande échelle AWS.
- Surveiller l'utilisation : surveillez et capturez les métriques d'utilisation des composants critiques de votre charge de travail (comme l'utilisation du processeur, l'utilisation de la mémoire ou le débit du réseau dans les métriques Amazon CloudWatch).
- Identifier les composants inutilisés : identifiez les composants inutilisés ou sous-utilisés dans votre architecture.
  - Pour des charges de travail stables, vérifiez les outils de redimensionnement AWS, par exemple AWS Compute Optimizer à intervalles réguliers, pour identifier les composants inactifs, inutilisés ou sous-utilisés.
  - Pour les charges de travail éphémères, évaluez les métriques d'utilisation pour identifier les composants inactifs, inutilisés ou sous-utilisés.
- Supprimer les composants inutilisés : retirez les composants et les ressources associées (comme les images Amazon ECR) qui ne sont plus nécessaires.
  - Nettoyage automatisé des images inutilisées dans Amazon ECR
  - Supprimez des volumes Amazon Elastic Block Store (Amazon EBS) inutilisés en utilisant AWS Config et AWS Systems Manager
- Refactoriser les composants sous-utilisés : refactorisez ou consolidez les composants sous-utilisés avec d'autres ressources pour améliorer l'efficacité de l'utilisation. Par exemple, vous pouvez provisionner plusieurs petites bases de données sur une seule instance de base de données Amazon RDS au lieu d'exécuter des bases de données sur des instances sous-utilisées individuelles.
- Évaluer les améliorations : identifiez les ressources provisionnées par votre charge de travail pour mener à bien une unité de travail. Utilisez ces informations pour évaluer les améliorations obtenues en supprimant ou en refactorisant des composants.
  - Mesure et suivi de l'efficacité du cloud à l'aide de métriques proxy de durabilité, partie I : Que sont les métriques proxy ?
  - Mesure et suivi de l'efficacité du cloud à l'aide de métriques proxy de durabilité, partie II : Établissement d'un pipeline de métriques

Ressources

Documents connexes :

- AWS Trusted Advisor
- Qu’est-ce qu’Amazon CloudWatch ?
- Dimensionnement adéquat : dimensionnement des instances en fonction des charges de travail
- Optimisation de vos coûts avec les recommandations de dimensionnement adéquat

Vidéos connexes :

- AWSre:Invent 2023 - Capacity, availability, cost efficiency: Pick three

Exemples connexes :

- Optimiser les modèles matériels et observer les indicateurs clés de performance de durabilité

SUS03-BP03 Optimiser les sections de votre code qui consomment le plus de temps ou de ressources

Optimisez votre code qui s’exécute dans les différents composants de votre architecture afin de minimiser l’utilisation des ressources tout en maximisant les performances.

Anti-modèles courants :

- Vous ignorez l’optimisation de votre code pour l’utilisation des ressources.
- Vous répondez généralement aux problèmes de performance en augmentant les ressources.
- Votre processus de révision et de développement du code ne permet pas de suivre les variations de performance.

Avantages liés au respect de cette bonne pratique : l’utilisation d’un code efficace minimise l’utilisation des ressources et améliore les performances.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d'implémentation

Il est essentiel d'examiner chaque domaine fonctionnel, y compris le code d'une application conçue dans le cloud, pour optimiser l'utilisation des ressources et les performances. Surveillez en permanence les performances de votre charge de travail dans les environnements de construction et de production et identifiez les possibilités d'améliorer les extraits de code qui utilisent particulièrement bien les ressources. Adoptez un processus de révision régulier pour identifier les bogues ou les anti-modèles dans votre code qui utilisent les ressources de manière inefficace. Exploitez des algorithmes simples et efficaces qui produisent les mêmes résultats pour votre cas d'utilisation.

Étapes d'implémentation

- Utiliser un langage de programmation efficace : utilisez un système d'exploitation et un langage de programmation efficaces pour la charge de travail. Pour en savoir plus sur les langages de programmation économes en énergie (y compris Rust), reportez-vous à Sustainability with Rust.
- Utiliser un compagnon de codage basé sur l'IA : envisagez d'utiliser un compagnon de codage basé sur l'IA tel qu'Amazon Q Developer pour écrire efficacement du code.
- Automatiser les révisions de code : pendant le développement de vos charges de travail, adoptez un processus de révision automatique du code pour améliorer la qualité et identifier les bogues et les anti-modèles.
  - Automatisez les révisions de code avec Amazon CodeGuru Reviewer
  - Détecter les bogues de concurrence avec Amazon CodeGuru
  - Améliorer la qualité du code des applications Python grâce à Amazon CodeGuru
- Utiliser un profileur de code : utilisez un profileur de code pour identifier les sections du code les plus longues ou qui consomment le plus de ressources dans le but de les optimiser.
  - Réduire l'empreinte carbone de votre organisation avec Amazon CodeGuru Profiler
  - Comprendre l'utilisation de la mémoire dans votre application Java avec Amazon CodeGuru Profiler
  - Améliorer l'expérience client et réduire les coûts avec Amazon CodeGuru Profiler
- Surveiller et optimiser : utilisez des ressources de surveillance continue pour identifier les composants nécessitant des ressources élevées ou présentant une configuration sous-optimale.
  - Remplacez les algorithmes à forte intensité de calcul par des versions plus simples et plus efficaces qui produisent le même résultat.
  - Supprimez le code inutile tel que le tri et le formatage.

- Utilisez la refactorisation ou la transformation du code : découvrez les possibilités de transformation du code Amazon Q pour la maintenance et les mises à niveau des applications.
  - Mettez à niveau les versions linguistiques avec Amazon Q Code Transformation
  - AWS re:Invent 2023 - Automate app upgrades & maintenance using Amazon Q Code Transformation

Ressources

Documents connexes :

- Présentation d'Amazon CodeGuru Profiler ?
- Instances FPGA
- Les kits de développement logiciel (SDK) AWS sur les outils pour créer sur AWS

Vidéos connexes :

- Améliorez l'efficacité du code à l'aide d'Amazon CodeGuru Profiler
- Automate Code Reviews and Application Performance Recommendations with Amazon CodeGuru

Exemples connexes :

- Optimisation du code avec Amazon CodeGuru

SUS03-BP04 Optimiser l'impact sur les appareils et équipements

Comprenez les appareils et les équipements utilisés dans votre architecture et employez des stratégies pour réduire leur utilisation. Cela peut minimiser l'impact environnemental global de votre charge de travail dans le cloud.

Anti-modèles courants :

- Vous ignorez l'impact environnemental des appareils utilisés par vos clients.
- Vous gérez et mettez à jour manuellement les ressources utilisées par les clients.

Avantages liés au respect de cette bonne pratique : la mise en œuvre de modèles logiciels et de fonctionnalités optimisés pour les appareils du client peut réduire l'impact environnemental global de la charge de travail dans le cloud.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

La mise en œuvre de modèles et de fonctionnalités logicielles optimisés pour les appareils des clients peut réduire l'impact environnemental de plusieurs façons :

- La mise en œuvre de nouvelles fonctionnalités qui sont rétrocompatibles peut réduire le nombre de remplacements de matériel.
- L'optimisation d'une application pour qu'elle fonctionne efficacement sur les appareils peut contribuer à réduire leur consommation d'énergie et à prolonger leur durée de vie (s'ils sont alimentés par une batterie).
- L'optimisation d'une application pour les appareils peut également réduire le transfert de données sur le réseau.

Comprenez les appareils et les équipements utilisés dans votre architecture, leur cycle de vie prévu et l'impact du remplacement de ces composants. Mettez en œuvre des modèles et des fonctionnalités logicielles qui minimisent la consommation d'énergie de l'appareil, réduisent la nécessité pour les clients de remplacer l'appareil et aussi de le mettre à niveau manuellement.

Étapes d'implémentation

- Réaliser un inventaire : dressez l'inventaire des appareils utilisés dans votre architecture. Les appareils peuvent être des mobiles, des tablettes, IOT des appareils, des lampes intelligentes ou même des appareils intelligents en usine.
- Utilisez des appareils économes en énergie : envisagez d'utiliser des appareils économes en énergie dans votre architecture. Utilisez les configurations de gestion de l'alimentation sur les appareils pour passer en mode faible consommation lorsqu'ils ne sont pas utilisés.
- Exécutez des applications efficaces : optimisez l'exécution de l'application sur les appareils :
  - utilisez des stratégies telles que l'exécution de tâches en arrière-plan pour réduire leur consommation d'énergie.

  - Prenez en compte la bande passante et la latence du réseau lorsque vous créez des charges utiles et intégrez des capacités qui aident vos applications à fonctionner correctement sur des liens à faible bande passante et à latence élevée.
  - Convertissez les charges utiles et les fichiers dans les formats optimisés requis par les appareils. Par exemple, vous pouvez utiliser Amazon Elastic Transcoder ou AWS Elemental MediaConvert pour convertir des fichiers multimédias numériques volumineux haute qualité en formats que les utilisateurs peuvent lire sur des appareils mobiles, des tablettes, des navigateurs Web et des téléviseurs connectés.
  - Réalisez des activités gourmandes en calcul côté serveur (comme le rendu d'images) ou utilisez le streaming d'applications pour améliorer l'expérience utilisateur sur des appareils plus anciens.
  - Segmentez et paginez la sortie, en particulier, pour les séances interactives, afin de gérer les charges utiles et limiter les exigences en matière de stockage local.
- Impliquer les fournisseurs : collaborez avec des fournisseurs d'appareils qui utilisent des matériaux durables et assurent la transparence de leurs chaînes d'approvisionnement et de leurs certifications environnementales.
- Utiliser les mises à jour over-the-air (OTA) : utilisez le mécanisme automatisé over-the-air (OTA) pour déployer les mises à jour sur un ou plusieurs appareils.
  - Vous pouvez utiliser un pipeline CI/CD pour mettre à jour les applications mobiles.
  - Vous pouvez utiliser AWS IoT Device Management pour gérer à distance les appareils connectés à grande échelle.
- Utiliser des parcs d'appareils gérés : pour tester les nouvelles fonctionnalités et les mises à jour, utilisez Device Farm avec des ensembles représentatifs de matériel et itérez le développement pour maximiser les dispositifs pris en charge. Pour en savoir plus, consultez SUS06-BP05 Utiliser des tests Device Farms gérés.
- Continuer à surveiller et à améliorer : suivez la consommation d'énergie des appareils pour identifier les domaines à améliorer. Utilisez les nouvelles technologies ou les bonnes pratiques pour améliorer les impacts environnementaux de ces appareils.

Ressources

Documents connexes :

- Qu'est-ce que c'est AWS Device Farm ?
- AppStream 2.0 Documentation
- NICE DCV

- OTAtutoriel pour la mise à jour du firmware sur les appareils fonctionnant sous Free RTOS
- Optimisation de vos appareils IoT pour la durabilité environnementale

Vidéos connexes :

- AWS re:Invent 2023 - Améliorez la qualité de vos applications mobiles et Web en utilisant AWS Device Farm

SUS03-BP05 Utiliser des modèles logiciels et des architectures qui soutiennent au mieux l'accès aux données et les modèles de stockage.

Comprenez comment les données sont utilisées au sein de votre charge de travail, et comment elles sont consommées par vos utilisateurs, transférées et stockées. Utilisez des modèles et des architectures logicielles qui prennent le mieux en charge l'accès et le stockage des données afin de minimiser les ressources de calcul, de mise en réseau et de stockage nécessaires pour supporter la charge de travail.

Anti-modèles courants :

- Vous partez du principe que toutes les charges de travail ont des modèles de stockage de données et d'accès similaires.
- Vous n'utilisez qu'un seul niveau de stockage, partant du principe que toutes les charges de travail s'intègrent dans ce niveau.
- Vous partez du principe que les modèles d'accès aux données n'évolueront pas dans le temps.
- Votre architecture prend en charge un potentiel pic important d'accès aux données, ce qui fait que les ressources restent inactives la plupart du temps.

Avantages liés au respect de cette bonne pratique : la sélection et l'optimisation de votre architecture en fonction des modèles d'accès aux données et de stockage contribueront à réduire la complexité du développement et à augmenter l'utilisation globale. Savoir quand utiliser les tables globales, le partitionnement des données et la mise en cache vous aidera à réduire les frais généraux opérationnels et à vous mettre à l'échelle en fonction des besoins de votre charge de travail.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Pour améliorer la durabilité à long terme de la charge de travail, utilisez des modèles d'architecture qui prennent en charge les caractéristiques d'accès aux données et de stockage pour votre charge de travail. Ces modèles vous aident à récupérer et à traiter efficacement les données. Par exemple, vous pouvez utiliser une architecture de données moderne sur AWS avec des services spécifiques optimisés pour vos cas d'utilisation analytiques uniques. Ces modèles d'architecture permettent un traitement efficace des données et réduisent l'utilisation des ressources.

Étapes d'implémentation

- Comprendre les caractéristiques des données : analysez les caractéristiques de vos données et les modèles d'accès afin d'identifier la bonne configuration pour vos ressources cloud. Les caractéristiques clés à prendre en considération sont les suivantes :
  - Type de données : structurées, semi-structurées, non structurées
  - Croissance des données : limitée, illimitée
  - Durabilité des données : persistantes, éphémères, temporaires
  - Modèles d'accès : lectures ou écritures, fréquence de mise à jour, irrégulière ou cohérente.
- Utiliser les modèles d'architecture optimaux : utilisez les modèles d'architecture qui prennent le mieux en charge les modèles d'accès et de stockage des données.
  - Modèles permettant la persistance des données
  - Let's Architect! Architectures de données modernes
  - Bases de données AWS : l'outil adapté à la tâche
- Utiliser des services spécialement conçus : utilisez des technologies adaptées à vos besoins.
  - Utilisez des technologies qui peuvent fonctionner en natif avec les données compressées.
    - Formats de fichiers pour prendre en charge la compression Athena
    - Options de formatage pour les entrées et sorties ETL dans AWS Glue
    - Chargement de fichiers de données comprimés d'Amazon S3 avec Amazon Redshift
  - Utilisez des services d'analytique spécialement conçus pour le traitement des données dans votre architecture. Pour en savoir plus sur les services analytiques de AWS, reportez-vous à AWS re:Invent 2022 - Building modern data architectures on AWS.
  - Utilisez le moteur de base de données qui prend le mieux en charge votre modèle de requête dominant. Gérez vos index de base de données pour une bonne efficacité des requêtes. Pour en savoir plus, reportez-vous à Bases de données AWS et AWS re:Invent 2022 - Modernize apps with purpose-built databases.

- Minimiser le transfert des données : sélectionnez des protocoles réseaux qui réduisent la quantité de capacité réseau consommée dans votre architecture.

Ressources

Documents connexes :

- COPIE de formats de données en colonnes avec Amazon Redshift
- Conversion de votre format d'enregistrement d'entrée dans Firehose
- Améliorer la performance des requêtes sur Amazon Athena grâce à une conversion en formats de colonnes
- Surveillance de la charge de base de données avec Performance Insights sur Amazon Aurora
- Surveillance de la charge de la base de données avec Performance Insights sur Amazon RDS
- Classe de stockage Amazon S3 Intelligent-Tiering
- Créez un magasin d'événements CQRS avec Amazon DynamoDB

Vidéos connexes :

- AWS re:Invent 2022 - Building data mesh architectures on AWS
- AWS re:Invent 2023 - Deep dive into Amazon Aurora and its innovations
- AWS re:Invent 2023 - Improve Amazon EBS efficiency and be more cost-efficient
- AWS re:Invent 2023 - Optimizing storage price and performance with Amazon S3
- AWSre:Invent 2023 : création et optimisation d'un lac de données sur Amazon S3
- AWS re:Invent 2023 - Advanced event-driven patterns with Amazon EventBridge

Exemples connexes :

- AWS Atelier sur les bases de données sur mesure
- Journée d'immersion dans l'architecture moderne des données AWS
- Créez un maillage de données sur AWS

Données

Question

SUS 4 Comment tirer parti des stratégies et des modèles de gestion des données pour soutenir vos objectifs de durabilité ?

Mettez en œuvre des pratiques de gestion des données afin de réduire le stockage alloué nécessaire pour assurer votre charge de travail et les ressources nécessaires à son utilisation. Veillez à bien connaître vos données et utilisez des technologies et des configurations de stockage qui soutiennent plus efficacement la valeur métier des données et leur utilisation. Adoptez un cycle de vie des données offrant un stockage plus efficace et moins performant quand les exigences baissent et supprimez les données qui ne sont plus nécessaires.

Bonnes pratiques

- SUS04-BP01 Mettre en œuvre une politique de classification des données
- SUS04-BP02 Utiliser des technologies qui prennent en charge les modèles d'accès et de stockage des données
- SUS04-BP03 Utiliser des politiques pour gérer le cycle de vie de vos ensembles de données
- SUS04-BP04 Utiliser l'élasticité et l'automatisation pour étendre le stockage par blocs ou le système de fichiers
- SUS04-BP05 Supprimer les données inutiles ou redondantes
- SUS04-BP06 Utiliser des systèmes de fichiers partagés ou le stockage pour accéder aux données courantes
- SUS04-BP07 Réduire le mouvement des données entre les réseaux
- SUS04-BP08 Sauvegarder des données uniquement lorsqu’elles sont difficiles à recréer

SUS04-BP01 Mettre en œuvre une politique de classification des données

Classifiez les données pour identifier leur criticité vis-à-vis des résultats économiques, et choisissez le niveau de stockage économe en énergie approprié pour stocker les données.

Anti-modèles courants :

- Vous n’identifiez pas les ressources de données actuellement traitées ou stockées ayant des caractéristiques similaires (comme la sensibilité, la criticité métier ou les exigences réglementaires).

- Vous n'avez pas implémenté de catalogue de données pour inventorier vos ressources de données.

Avantages liés au respect de cette bonne pratique : la mise en œuvre d'une politique de classification des données vous permet de déterminer le niveau de stockage le plus économe en énergie pour les données.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

La classification des données implique d'identifier les types de données actuellement traitées ou stockées dans un système d'information détenu ou exploité par une organisation. Elle implique également de déterminer la criticité des données et l'impact possible d'une compromission, d'une perte ou d'une mauvaise utilisation de ces données.

Mettez en œuvre la politique de classification des données en partant de l'utilisation contextuelle des données et en créant un schéma de catégorisation qui prend en compte le niveau de criticité d'un jeu de données déterminé vis-à-vis des opérations d'une organisation.

Étapes d'implémentation

- Réaliser un inventaire des données :procédez à l'inventaire des différents types de données qui existent pour votre charge de travail.
- Données du groupe :déterminez la criticité, la confidentialité, l'intégrité et la disponibilité des données en fonction du risque vis-à-vis de l'organisation. Prenez en compte ces exigences pour regrouper les données dans l'un des niveaux de classification des données que vous adoptez. À titre d'exemple, consultez Quatre étapes simples pour classer vos données et sécuriser votre start-up.
- Définissez les niveaux et les politiques de classification des données : pour chaque groupe de données, définissez le niveau de classification des données (par exemple, public ou confidentiel) et les politiques de gestion. Balisez les données en conséquence. Pour en savoir plus sur les catégories de classification des données, consultez le livre blanc Classification des données.
- Révision périodique : passez régulièrement en revue et auditez votre environnement pour détecter les données non étiquetées et non classifiées. Utilisez l'automatisation pour identifier ces données, puis classez et balisez les données de manière appropriée. À titre d'exemple, consultez le catalogue de données et les robots dans AWS Glue.

- Établissez un catalogue de données : établissez un catalogue de données qui fournit des fonctionnalités d'audit et de gouvernance.
- Documentation : Documentez les politiques de classification des données et les procédures de traitement pour chaque classe de données.

Ressources

Documents connexes :

- Utilisation de l'effet de levier AWS Cloud tpour soutenir la classification des données
- Marquer les politiques de AWS Organizations

Vidéos connexes :

- AWS re:Invent 2022 - Favoriser l'agilité grâce à la gouvernance des données activée AWS
- AWS re:Invent 2023 - Protection des données et résilience grâce au stockage AWS

SUS04-BP02 Utiliser des technologies qui prennent en charge les modèles d'accès et de stockage des données

Utilisez les technologies de stockage qui prennent le mieux en charge l'accès à vos données et leur stockage pour limiter le provisionnement de ressources tout en soutenant votre charge de travail.

Anti-modèles courants :

- Vous partez du principe que toutes les charges de travail ont des modèles de stockage de données et d'accès similaires.
- Vous n'utilisez qu'un seul niveau de stockage, partant du principe que toutes les charges de travail s'intègrent dans ce niveau.
- Vous partez du principe que les modèles d'accès aux données n'évolueront pas dans le temps.

Avantages liés au respect de cette bonne pratique : en choisissant et en optimisant vos technologies de stockage en fonction des modèles d'accès aux données et de stockage, vos besoins métier demanderont moins de ressources cloud et vous améliorerez l'efficace globale de votre charge de travail cloud.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d’implémentation

Choisissez la solution de stockage la mieux adaptée à vos modèles d’accès ou envisagez de changer vos modèles d’accès en fonction de la solution de stockage pour optimiser les performances.

Étapes d’implémentation

- Évaluer les caractéristiques des données et de l’accès : évaluez les caractéristiques de vos données et le modèle d’accès afin de recueillir les caractéristiques clés de vos besoins en matière de stockage. Les caractéristiques clés à prendre en considération sont les suivantes :
  - Type de données : structurées, semi-structurées, non structurées
  - Croissance des données : limitée, illimitée
  - Durabilité des données : persistantes, éphémères, temporaires
  - Modèles d’accès : lectures ou écritures, fréquence de mise à jour, irrégulière ou cohérente.
- Choisir la bonne technologie de stockage : migrer les données vers la technologie de stockage appropriée qui prend en charge les caractéristiques des données et le modèle d’accès. Voici quelques exemples de technologies de AWS stockage et leurs principales caractéristiques :

Type	Technologie	Principales caractéristiques
Stockage d’objets	Amazon S3	Service de stockage d’objets offrant une capacité de mise à l’échelle illimitée, une haute disponibilité et plusieurs options d’accessibilité. Le transfert et l’accès à des objets à l’intérieur et à l’extérieur d’Amazon S3 peuvent utiliser un service, tel que Transfer Acceleration ou Access Points, pour répondre à votre localisation, à vos besoins en matière de sécurité et à vos modèles d’accès.

Type	Technologie	Principales caractéristiques
Archivage et stockage	Amazon S3 Glacier	Classe de stockage d'Amazon S3 conçue pour l'archivage de données.
Système de fichiers partagé	Amazon Elastic File System (AmazonEFS)	Système de fichiers montable auquel plusieurs types de solutions informatiques peuvent accéder. Amazon augmente et réduit EFS automatiquement le stockage et optimise les performan ces pour garantir de faibles latences constantes.
Système de fichiers partagé	Amazon FSx	Construit sur les dernières solutions AWS informati ques pour prendre en charge quatre systèmes de fichiers couramment utilisés : Open NetApp ONTAPZFS, Windows File Server et Lustre. FSxLa latence, le débit et le débit d'Amazon IOPS varient selon le système de fichiers et doivent être pris en compte lors de la sélection du système de fichiers adapté à vos besoins en matière de charge de travail.

Type	Technologie	Principales caractéristiques
Stockage en mode bloc	Boutique Amazon Elastic Block (AmazonEBS)	Service de stockage par blocs évolutif et performant conçu pour Amazon Elastic Compute Cloud (AmazonEC 2). Amazon EBS inclut le stockage SSD sauvegardé pour les charges de travail transactionnelles intensives et le stockage HDD sauvegard é pour les charges de travail IOPS gourmandes en débit.
Base de données relationn elle	Amazon Aurora, Amazon RDS, Amazon Redshift	Conçu pour prendre en charge ACID (atomicit é, cohérence, isolation, durabilité) les transactions et maintenir l'intégrité référenti elle et la forte cohérence des données. De nombreuses applications traditionnelles, systèmes de planification des ressources d'entrepr ise (ERP), de gestion de la relation client (CRM) et de commerce électroni que utilisent des bases de données relationnelles pour stocker leurs données.

Type	Technologie	Principales caractéristiques
Base de données clé-valeur	Amazon DynamoDB	Optimisées pour les modèles d’accès courants, généralement pour stocker et récupérer de gros volumes de données. Les applications Web à trafic élevé, les systèmes d’e-commerce et les applications de jeu sont des cas d’utilisation typiques pour les bases de données clé-valeur.

- Automatisez l'allocation de stockage : pour les systèmes de stockage de taille fixe, tels qu'Amazon EBS ou AmazonFSx, surveillez l'espace de stockage disponible et automatisez l'allocation de stockage lorsqu'un seuil est atteint. Vous pouvez tirer parti CloudWatch d'Amazon pour collecter et analyser différentes statistiques pour Amazon EBS et Amazon FSx.
- Choisissez la bonne classe de stockage : choisissez la classe de stockage adaptée à vos données.
  - Les classes de stockage Amazon S3 peuvent être configurées au niveau de l’objet. Un compartiment unique peut contenir les objets stockés dans toutes les classes de stockage.
  - Vous pouvez utiliser les stratégies de cycle de vie Amazon S3 pour faire passer automatiquement des objets d’une classe de stockage à une autre ou supprimer des données sans aucune modification au niveau de l’application. Ces mécanismes de stockage vous imposent généralement de faire un compromis entre l’efficacité des ressources, la latence d’accès et la fiabilité.

Ressources

Documents connexes :

- Types de EBS volumes Amazon
- Boutique d'EC2instances Amazon
- Amazon S3 Intelligent Tiering
- Caractéristiques d'Amazon EBS I/O
- Utilisation des classes de stockage Amazon S3
- Qu’est-ce qu’Amazon S3 Glacier ?

Vidéos connexes :

- AWS re:Invent 2023 - Améliorez EBS l'efficacité d'Amazon et soyez plus rentable
- AWS re:Invent 2023 - Optimisation du prix et des performances du stockage avec Amazon S3
- AWS re:Invent 2023 - Création et optimisation d'un lac de données sur Amazon S3
- AWS re:Invent 2022 - Création d'architectures de données modernes sur AWS
- AWS re:Invent 2022 - Modernisez les applications avec des bases de données spécialement conçues
- AWS re:Invent 2022 - Création d'architectures de maillage de données sur AWS
- AWS re:Invent 2023 - Découvrez Amazon Aurora et ses innovations
- AWS re:Invent 2023 - Modélisation avancée des données avec Amazon DynamoDB

Exemples connexes :

- Exemples Amazon S3
- AWS Atelier sur les bases de données spécialement conçues
- Bases de données pour développeurs
- AWS Journée d'immersion dans l'architecture de données moderne
- Créez un maillage de données sur AWS

SUS04-BP03 Utiliser des politiques pour gérer le cycle de vie de vos ensembles de données

Gérez le cycle de vie de toutes vos données et appliquez automatiquement la suppression pour réduire au minimum le stockage total requis pour votre charge de travail.

Anti-modèles courants :

- Vous supprimez manuellement les données.
- Vous ne supprimez aucune donnée de vos charges de travail.
- Vous ne déplacez pas les données vers des niveaux de stockage plus écoénergétiques en fonction de leurs exigences de conservation et d’accès.

Avantages liés au respect de cette bonne pratique : l’utilisation de politiques de cycle de vie des données garantit un accès et une rétention efficaces des données dans une charge de travail.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les exigences en matière de conservation et d'accès des jeux de données varient généralement au cours de leur cycle de vie. Par exemple, votre application peut nécessiter un accès fréquent à certains jeux de données pendant une période limitée. Après cela, ces jeux de données sont rarement consultés. Pour améliorer l'efficacité du stockage de données et du calcul au fil du temps, mettez en œuvre des politiques de cycle de vie, qui sont des règles qui définissent la manière dont les données sont traitées au fil du temps.

Avec les règles de configuration du cycle de vie, vous pouvez demander au service de stockage spécifique de transférer un jeu de données vers des niveaux de stockage plus écoénergétiques, de l'archiver ou de le supprimer. Cette pratique minimise le stockage et l'extraction actifs des données, ce qui entraîne une réduction de la consommation d'énergie. En outre, des pratiques telles que l'archivage ou la suppression de données obsolètes soutiennent la conformité réglementaire et la gouvernance des données.

Étapes d'implémentation

- Utiliser la classification des données : classez les jeux de données dans votre charge de travail.
- Définir des règles de traitement : définissez des procédures de traitement pour chaque classe de données.
- Activer l'automatisation : définissez des politiques de cycle de vie automatisées pour appliquer des règles de cycle de vie. Voici quelques exemples de la configuration des politiques de cycle de vie automatisé pour différents services de stockage AWS :

Service de stockage	Comment définir des politiques de cycle de vie automatisées
Amazon S3	Vous pouvez utiliser Amazon S3 Lifecycle afin de gérer vos objets au cours de leur cycle de vie. Si vos schémas d'accès sont inconnus, changeants ou imprévisibles, vous pouvez utiliser Amazon S3 Intelligent-Tiering, qui surveille les schémas d'accès et déplace automatiquement les objets qui n'ont pas été accédés vers des niveaux d'accès moins

Service de stockage	Comment définir des politiques de cycle de vie automatisées
	coûteux. Vous pouvez tirer parti des métriques Amazon S3 Storage Lens pour identifier les opportunités d'optimisation et les lacunes dans la gestion du cycle de vie.
Amazon Elastic Block Store	Vous pouvez utiliser Amazon Data Lifecycle Manager pour automatiser la création, la rétention et la suppression des instantanés EBS et des AMI basées sur EBS.
Amazon Elastic File System	La fonction de gestion du cycle de vie Amazon EFS gère automatiquement le stockage de vos systèmes de fichiers.
Amazon Elastic Container Registry	Les politiques de cycle de vie d'Amazon ECR automatisent le nettoyage des images de conteneur en faisant expirer des images selon l'ancienneté ou le décompte.
AWS Elemental MediaStore	Vous pouvez utiliser une politique de cycle de vie des objets qui régit la durée de stockage des objets dans le conteneur MediaStore.

- Supprimer les ressources inutilisées : supprimez les volumes, les instantanés et les données inutilisés dont la période de conservation est dépassée. Utilisez des fonctionnalités de service natives telles qu'Amazon DynamoDB Time To Live ou la conservation des journaux Amazon CloudWatch pour la suppression.
- Regrouper et compresser : regroupez et compressez les données le cas échéant en fonction des règles de cycle de vie.

Ressources

Documents connexes :

- Optimisez vos règles de cycle de vie Amazon S3 grâce à l'analyse des classes de stockage Amazon S3

- Évaluation des ressources avec AWS Config Rules

Vidéos connexes :

- AWS re:Invent 2021 - Amazon S3 Lifecycle best practices to optimize your storage spend
- AWS re:Invent 2023 - Optimizing storage price and performance with Amazon S3
- Simplifiez le cycle de vie de vos données et optimisez les coûts de stockage avec Amazon S3 Lifecycle
- Réduisez vos coûts de stockage en utilisant Amazon S3 Storage Lens

SUS04-BP04 Utiliser l'élasticité et l'automatisation pour étendre le stockage par blocs ou le système de fichiers

Utilisez l’élasticité et l’automatisation pour étendre le stockage par blocs ou le système de fichiers au fur et à mesure que le volume de données augmente afin de minimiser le stockage total provisionné.

Anti-modèles courants :

- Vous provisionnez un grand bloc de stockage ou un grand système de fichiers pour vos besoins futurs.
- Vous surprovisionnez les opérations d'entrée et de sortie par seconde (IOPS) de votre système de fichiers.
- Vous ne contrôlez pas l’utilisation de vos volumes de données.

AAvantages liés au respect de cette bonne pratique : la réduction du surprovisionnement du système de stockage réduit les ressources inactives et améliore l’efficacité globale de votre charge de travail.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Créez des systèmes de stockage par blocs et des systèmes de fichiers avec une allocation de taille, un débit et une latence adaptés à votre charge de travail. Utilisez l’élasticité et l’automatisation pour étendre le stockage par blocs ou le système de fichiers en fonction de la croissance des données sans avoir à provisionner ces services de stockage de manière excessive.

Étapes d’implémentation

- Pour le stockage de taille fixe tel qu'Amazon EBS, vérifiez que vous surveillez la quantité de stockage utilisée par rapport à la taille de stockage globale et créez une automatisation, si possible, pour augmenter la taille de stockage lorsque vous atteignez un seuil.
- Utilisez des volumes Elastic et des services de données par bloc gérés pour automatiser l’allocation de stockage supplémentaire à mesure que vos données persistantes augmentent. Par exemple, vous pouvez utiliser Amazon EBS Elastic Volumes pour modifier la taille ou le type de volume ou ajuster les performances de vos EBS volumes Amazon.
- Choisissez la bonne classe de stockage, le bon mode de performance et le mode de débit adapté à votre système de fichiers afin de répondre aux besoins de votre entreprise, sans les dépasser.
  - EFSPerformances d'Amazon
  - Performances des EBS volumes Amazon sur les instances Linux
- Définissez des niveaux cibles d’utilisation des volumes de données et redimensionnez les volumes en dehors des plages attendues.
- Dimensionnez correctement les volumes en lecture seule en fonction des données.
- Migrez les données vers des magasins d’objets pour éviter d’allouer la capacité excédentaire des tailles de volume fixes vers le stockage par bloc.
- Examinez régulièrement les volumes Elastic et les systèmes de fichiers pour mettre fin aux volumes inutilisés et réduire les ressources surprovisionnées pour les adapter à la taille actuelle des données.

Ressources

Documents connexes :

- Étendre le système de fichiers après le redimensionnement d'un volume EBS
- Modifier un volume à l'aide d'Amazon EBS Elastic Volumes
- Documentation Amazon FSx
- Qu’est ce qu’Amazon Elastic File System ?

Vidéos connexes :

- Présentation approfondie d'Amazon EBS Elastic Volumes

- Stratégies d'optimisation d'Amazon EBS et de Snapshot pour de meilleures performances et des économies
- Optimisation d'Amazon EFS en termes de coûts et de performances, en utilisant les meilleures pratiques

SUS04-BP05 Supprimer les données inutiles ou redondantes

Supprimez les données inutiles ou redondantes pour minimiser les ressources de stockage requises pour stocker vos jeux de données.

Anti-modèles courants :

- Vous dupliquez des données qui peuvent être facilement obtenues ou recréées.
- Vous sauvegardez toutes les données sans tenir compte de leur criticité.
- Vous ne supprimez les données que de façon irrégulière, sur les événements opérationnels ou pas du tout.
- Vous stockez les données de manière redondante, quelle que soit la durabilité du service de stockage.
- Vous activez la gestion des versions sans aucune justification commerciale.

Avantages liés au respect de cette bonne pratique : la suppression des données inutiles réduit la taille de stockage requise pour votre charge de travail et son impact environnemental.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

La suppression des jeux de données inutiles et redondants permet de réduire les coûts de stockage et l’empreinte environnementale. Cette pratique peut également rendre l’informatique plus efficace, car les ressources de calcul traitent uniquement des données importantes au lieu de données inutiles. Automatisez la suppression des données inutiles. Utilisez des technologies qui dédupliquent les données au niveau du fichier et du bloc. Utilisez les fonctionnalités des services pour la réplication et la redondance des données natives.

Étapes d'implémentation

- Évaluer les jeux de données publics : déterminez si vous pouvez éviter de stocker des données en utilisant des jeux de données existants publiquement accessibles dans AWS Data Exchange et les données ouvertes sur AWS.
- Dédupliquer les données : utilisez des mécanismes qui peuvent dédupliquer les données au niveau du bloc et de l'objet. Voici quelques exemples de déduplication des données sur AWS :

Service de stockage	Mécanismes de déduplication
Amazon S3	Utilisez AWS Lake Formation FindMatches pour rechercher les enregistrements correspondants dans un jeu de données (y compris ceux sans identifiant) à l'aide de la nouvelle transformation FindMatches ML.
Amazon FSx	Utilisez la déduplication des données sur Amazon FSx for Windows
Instantanés volumes Amazon Elastic Block Store	Les instantanés sont des sauvegardes incrémentielles, ce qui signifie que seuls les blocs de l'appareil qui ont changé depuis l'instantané le plus récent sont enregistrés.

- Utiliser des politiques de cycle de vie : utilisez des politiques de cycle de vie pour automatiser la suppression des données inutiles. Tirez parti de fonctionnalités de service natives telles qu'Amazon DynamoDB Time To Live, Amazon S3 Lifecycle ou la rétention des journaux Amazon CloudWatch pour la suppression.
- Utiliser la virtualisation des données : utilisez les capacités de virtualisation des données sur AWS afin de maintenir les données à leur source et d'éviter leur duplication.
  - Virtualisation des données natives dand le cloud sur AWS
  - Optimiser le modèle de données à l'aide du partage de données Amazon Redshift
- Utiliser la sauvegarde incrémentielle : utilisez une technologie de sauvegarde capable d'effectuer des sauvegardes incrémentielles.
- Utiliser la durabilité native : tirez parti de la durabilité d'Amazon S3 et de la réplication d'Amazon EBS pour atteindre vos objectifs de durabilité au lieu de recourir à des technologies autogérées (telles qu'un réseau redondant de disques indépendants (RAID)).

- Utiliser une journalisation efficace : centralisez les données de journaux et de suivi, dédupliquez les entrées de journaux identiques et établissez des mécanismes pour ajuster le niveau de détail, si nécessaire.
- Utiliser une mise en cache efficace : préremplissez les caches uniquement lorsque cela est justifié.
- Établissez la surveillance et l'automatisation des caches pour redimensionner correctement les caches.
- Supprimer les ressources de version antérieure : supprimez les déploiements et les ressources obsolètes des magasins d'objets et des caches périphériques lors de la transmission des nouvelles versions de votre charge de travail.

Ressources

Documents connexes :

- Modification de la conservation des données de journaux dans CloudWatch Logs
- Déduplication des données sur Amazon FSx for Windows File Server
- Les fonctions d'Amazon FSx for ONTAP incluent la déduplication des données
- Invalider des fichiers sur Amazon CloudFront
- Utilisation d'AWS Backup pour la sauvegarde et la restauration des systèmes de fichiers Amazon EFS
- Qu'est-ce qu'Amazon CloudWatch Logs ?
- Utilisation de sauvegardes sur Amazon RDS
- Intégrez et dédupliquez des ensembles de données à l'aide de AWS Lake Formation

Vidéos connexes :

- Cas d'utilisation du partage de données pour Amazon Redshift

Exemples connexes :

- Comment analyser les journaux d'accès au serveur Amazon S3 à l'aide d'Amazon Athena ?

SUS04-BP06 Utiliser des systèmes de fichiers partagés ou le stockage pour accéder aux données courantes

Adoptez des systèmes de fichiers ou de stockage partagés pour éviter la duplication des données et permettre une infrastructure plus efficace pour votre charge de travail.

Anti-modèles courants :

- Vous mettez en service le stockage pour chaque client individuel.
- Vous ne détachez pas le volume de données des clients inactifs.
- Vous ne fournissez pas d'accès au stockage pour les plateformes et les systèmes.

Avantages liés au respect de cette bonne pratique : 'utilisation de systèmes de fichiers ou de stockage partagés permet de partager des données avec un ou plusieurs consommateurs sans avoir à les copier. Cela permet de réduire les ressources de stockage nécessaires à la charge de travail.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Si plusieurs utilisateurs ou applications accèdent aux mêmes jeux de données, l'utilisation de la technologie de stockage partagé est cruciale pour avoir une infrastructure efficace pour votre charge de travail. La technologie de stockage partagé fournit un emplacement central pour stocker et gérer les jeux de données et éviter la duplication des données. Elle assure également la cohérence des données entre les différents systèmes. En outre, la technologie de stockage partagé permet d'utiliser plus efficacement la puissance de calcul, car plusieurs ressources informatiques peuvent accéder aux données et les traiter simultanément en parallèle.

Ne récupérez les données de ces services de stockage partagé qu'en fonction des besoins et détachez les volumes inutilisés pour libérer des ressources.

Étapes d'implémentation

- Utiliser un stockage partagé : migrez les données vers le stockage partagé lorsque les données ont plusieurs consommateurs. Voici quelques exemples de technologie de stockage partagé sur AWS :

Option de stockage	Utilisation
Amazon EBS Multi-Attach	Amazon EBS Multi-Attach vous permet d'attacher un volume SSD IOPS provisionnés

Option de stockage	Utilisation
	(io1 ou io2) à plusieurs instances basées sur Nitro situées dans la même zone de disponibilité.
Amazon EFS	Reportez-vous à Quand choisir Amazon EFS.
Amazon FSx	Reportez-vous à Choisir un système de fichiers Amazon FSx.
Amazon S3	Les applications qui ne nécessitent pas de structure de système de fichiers et qui sont conçues pour fonctionner avec le stockage d'objet peuvent utiliser Amazon S3 comme une solution de stockage d'objet massivement évolutive, durable et peu coûteuse.

- Récupérer les données requises : copiez des données vers ou récupérez des données depuis des systèmes de fichiers partagés uniquement si nécessaire. Par exemple, vous pouvez créer un système de fichiers Amazon FSx pour Lustre soutenu par Amazon S3 et charger uniquement le sous-ensemble de données requis pour le traitement des tâches sur Amazon FSx.
- Supprimer les données inutiles : supprimez les données selon vos modèles d'utilisation comme indiqué dans SUS04-BP03 Utiliser des politiques pour gérer le cycle de vie de vos ensembles de données.
- Détacher les clients inactifs : détachez les volumes des clients qui ne les utilisent pas activement.

Ressources

Documents connexes :

- Liaison de votre système de fichiers à un compartiment Amazon S3
- Utilisation d'Amazon EFS pour AWS Lambda dans vos applications sans serveur
- Amazon EFS Intelligent-Tiering optimise les coûts liés aux charges de travail en fonction de l'évolution des modèles d'accès
- Utilisation d'Amazon FSx avec votre référentiel de données sur site

Vidéos connexes :

- Optimisation des coûts de stockage avec Amazon EFS
- AWS re:Invent 2023 - What's new with AWS file storage
- AWS re:Invent 2023 - File storage for builders and data scientists on Amazon Elastic File System

SUS04-BP07 Réduire le mouvement des données entre les réseaux

Utilisez des systèmes de fichiers partagés ou un stockage objet pour accéder aux données communes et minimiser les ressources réseau totales requises pour prendre en charge le déplacement des données de votre charge de travail.

Anti-modèles courants :

- Vous stockez toutes les données dans la même Région AWS, indépendamment de l'endroit où se trouvent les utilisateurs des données.
- Vous n'optimisez ni la taille ni le format des données avant de les déplacer sur le réseau.

Avantages liés au respect de cette bonne pratique : l'optimisation du déplacement des données sur le réseau réduit les ressources réseau totales nécessaires à la charge de travail et diminue son impact environnemental.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Le déplacement des données dans votre entreprise nécessite des ressources de calcul, de réseau et de stockage. Utilisez des techniques pour minimiser les déplacements de données et améliorer l'efficacité globale de votre charge de travail.

Étapes d'implémentation

- Utiliser la proximité : tenez compte de la proximité des données ou des utilisateurs comme facteur de décision lors de la sélection d'une région pour votre charge de travail.
- Partitionner les services : partitionnez les services consommés par région afin que les données spécifiques à une région soient stockées dans la région où elles sont consommées.
- Utiliser des formats de fichiers efficaces : utilisez des formats de fichiers efficaces (tels que Parquet ou ORC) et compressez les données avant de les déplacer sur le réseau.

- Minimiser le mouvement des données : ne déplacez pas les données inutilisées. Voici quelques exemples qui peuvent vous aider à éviter de déplacer des données inutilisées :
  - Réduisez les réponses de l'API aux seules données pertinentes.
  - Agrégez les données lorsqu'elles sont détaillées (les informations au niveau de l'enregistrement ne sont pas requises).
  - Reportez-vous à Atelier Well-Architected : optimiser le modèle de données à l'aide du partage de données Amazon Redshift.
  - Envisagez le partage de données entre comptes dans AWS Lake Formation.
- Utilisez des services de périphérie : utilisez des services qui peuvent vous aider à exécuter du code au plus près des utilisateurs de votre charge de travail.

Service	Utilisation
Lambda@Edge	Utilisez ce service pour les opérations exigeantes en puissance de calcul qui sont exécutées lorsque des objets ne sont pas dans le cache.
Fonctions CloudFront	Utilisez ce système pour des cas d'utilisa tion simples tels que les manipulations de requêtes/réponses HTTP(s) qui peuvent être lancées par des fonctions de courte durée.
AWS IoT Greengrass	Exécutez du calcul local, une messagerie et une mise en cache de données pour les appareils connectés.

Ressources

Documents connexes :

- Optimisation de votre infrastructure AWS pour la durabilité, partie 3 : mise en réseau
- Infrastructure mondiale AWS
- Fonctions clés d'Amazon CloudFront, y compris le réseau périphérique mondial CloudFront
- Compression des requêtes HTTP dans Amazon OpenSearch Service

- Compression intermédiaire de données avec Amazon EMR
- Chargement de fichiers de données comprimés d'Amazon S3 vers Amazon Redshift
- Diffusion de fichiers compressés avec Amazon CloudFront

Vidéos connexes :

- Demystifying data transfer on AWS

Exemples connexes :

- Architecting for sustainability - Minimize data movement across networks

SUS04-BP08 Sauvegarder des données uniquement lorsqu'elles sont difficiles à recréer

Évitez de sauvegarder les données qui n'ont aucune valeur commerciale afin de minimiser les besoins en ressources de stockage pour votre charge de travail.

Anti-modèles courants :

- Vous n'avez aucune stratégie de sauvegarde en place pour vos données.
- Vous sauvegardez des données qui peuvent être facilement recréées.

Avantages liés au respect de cette bonne pratique : éviter de sauvegarder des données non critiques réduit les ressources de stockage requises pour la charge de travail et réduit son impact environnemental.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Le fait d'éviter la sauvegarde de données inutiles peut contribuer à réduire les coûts et les ressources de stockage utilisées par la charge de travail. Sauvegardez uniquement les données ayant une valeur opérationnelle ou nécessaires pour répondre aux exigences en matière de conformité. Examinez les politiques de sauvegarde et excluez tout magasin éphémère n'apportant aucune valeur dans un scénario de récupération.

Étapes d'implémentation

- Classer les données : mettre en œuvre la politique de classification des données telle que décrite dans SUS04-BP01 Mettre en œuvre une politique de classification des données.
- Concevoir une stratégie de sauvegarde : tirez parti de l'importance de votre classification des données et concevez une stratégie de sauvegarde en fonction de vos objectif de délai de reprise (RTO) et objectif de point de reprise (RPO). Évitez de sauvegarder les données non critiques.
  - Excluez les données qui peuvent être facilement recréées.
  - Excluez les données éphémères de vos sauvegardes.
  - Excluez les copies locales des données, sauf si le temps nécessaire pour restaurer ces données à partir d'un emplacement commun dépasse vos contrats de niveau de service (SLA).
- Utiliser une sauvegarde automatisée : utilisez une solution automatisée ou un service géré pour sauvegarder les données essentielles à l'entreprise.
  - AWS Backup est un service entièrement géré qui vous permet de facilement centraliser et automatiser la protection des données sur les services AWS dans le cloud et sur site. Pour obtenir des conseils pratiques sur la façon de créer des sauvegardes automatisées à l'aide de AWS Backup, consultez la section Test de la sauvegarde et de la restauration de données.
  - Automating backups and optimizing backup costs for Amazon EFS à l'aide de AWS Backup.

Ressources

Bonnes pratiques associées :

- REL09-BP01 Identifier et sauvegarder toutes les données qui doivent être sauvegardées, ou reproduire les données à partir de sources
- REL09-BP03 Effectuer automatiquement la sauvegarde des données
- REL13-BP02 Utiliser des stratégies de reprise définies pour répondre aux objectifs de reprise

Documents connexes :

- Utilisation d'AWS Backup pour la sauvegarde et la restauration des systèmes de fichiers Amazon EFS
- Instantanés Amazon EBS
- Utilisation de sauvegardes sur Amazon Relational Database Service
- Partenaire APN : partenaires pouvant faciliter la sauvegarde

- AWS Marketplace : produits pouvant être utilisés pour la sauvegarde
- Sauvegarde d'Amazon EFS
- S'auvegarde d'Amazon FSx for Windows File Server
- Backup et restauration d'Amazon ElastiCache (Redis OSS)

Vidéos connexes :

- AWS re:Invent 2023 - Backup and disaster recovery strategies for increased resilience
- AWS re:Invent 2023 - What's new with AWS Backup
- AWS re:Invent 2021 - Backup, disaster recovery, and ransomware protection with AWS

Exemples connexes :

- Well-Architected Lab : données de sauvegarde

Matériel et services

Question

- SUS 5 Comment sélectionner et utiliser le matériel et les services du cloud dans votre architecture pour soutenir vos objectifs de durabilité ?

SUS 5 Comment sélectionner et utiliser le matériel et les services du cloud dans votre architecture pour soutenir vos objectifs de durabilité ?

Recherchez des possibilités de réduire les impacts en matière de durabilité de la charge de travail en modifiant vos pratiques de gestion du matériel. Réduisez la quantité de matériel nécessaire à allouer et à déployer, et sélectionnez le matériel et les services les plus efficaces pour votre charge de travail individuelle.

Bonnes pratiques

- SUS05-BP01 Utilisez le minimum de matériel pour répondre à vos besoins
- SUS05-BP02 Utiliser des types d'instance ayant le moins d'impact
- SUS05-BP03 Utiliser des services gérés
- SUS05-BP04 Optimiser votre utilisation des accélérateurs de calcul matériels

SUS05-BP01 Utilisez le minimum de matériel pour répondre à vos besoins

Utilisez la quantité minimale de matériel pour votre charge de travail afin de répondre efficacement aux besoins de votre entreprise.

Anti-modèles courants :

- Vous ne surveillez pas l'utilisation des ressources.
- Vous disposez de ressources avec un faible niveau d'utilisation dans votre architecture.
- Vous n'examinez pas l'utilisation du matériel statique pour déterminer s'il doit être redimensionné.
- Vous ne définissez pas d'objectifs d'utilisation du matériel pour votre infrastructure informatique en fonction de votre activitéKPIs.

Avantages liés au respect de cette bonne pratique : la rationalisation de vos ressources cloud permet de réduire l'impact environnemental d'une charge de travail, d'économiser de l'argent et de maintenir les repères de performance.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Sélectionnez de manière optimale le nombre total de composants matériels requis pour votre charge de travail afin d'améliorer son efficacité globale. AWS Cloud Il offre la flexibilité nécessaire pour étendre ou réduire le nombre de ressources de manière dynamique grâce à divers mécanismes, tels que AWS Auto Scaling, et pour répondre à l'évolution de la demande. Il fournit également APIset SDKs permet de modifier les ressources avec un minimum d'effort. Utilisez ces capacités pour apporter des modifications fréquentes à vos mises en œuvre de charges de travail. En outre, utilisez les directives de redimensionnement issues des AWS outils pour exploiter efficacement vos ressources cloud et répondre aux besoins de votre entreprise.

Étapes d'implémentation

- Choisissez le type d'instance : choisissez le type d'instance qui répond le mieux à vos besoins. Pour déterminer comment choisir des instances Amazon Elastic Compute Cloud et utiliser des mécanismes tels que la sélection d'instances basée sur les attributs, consultez les rubriques suivantes :
  - Comment choisir le type d'EC2instance Amazon adapté à ma charge de travail ?
  - Sélection du type d'instance basée sur les attributs pour Amazon EC2 Fleet.

    - Créer un groupe Auto Scaling en utilisant la sélection du type d'instance basée sur des attributs.
- Mettre à l'échelle : diminuez les charges de travail variables par petits paliers.
- Utilisez plusieurs options d'achat de calcul : équilibrez la flexibilité, la capacité de mise à l'échelle et les économies de coûts des instances grâce à plusieurs options d'achat de calcul.
    - Les instances EC2 Amazon On-Demand sont parfaitement adaptées aux nouvelles charges de travail dynamiques et complexes qui ne peuvent pas être liées au type d'instance, à l'emplacement ou à la durée.
    - Les instances Amazon EC2 Spot constituent un excellent moyen de compléter les autres options pour les applications tolérantes aux pannes et flexibles.
    - Tirez parti des Compute Savings Plans pour des charges de travail stables qui offrent de la flexibilité si vos besoins (tels que l'AZ, la région, les familles d'instances ou les types d'instances) changent.
- Tirez parti de la diversité des instances et des zones de disponibilité : optimisez la disponibilité des applications et tirez parti de la capacité excédentaire en diversifiant vos instances et vos zones de disponibilité.
- Instances à la bonne taille : utilisez les recommandations de redimensionnement des AWS outils pour ajuster votre charge de travail. Pour en savoir plus, consultez Optimiser vos coûts avec les recommandations de dimensionnement et Dimensionnement adéquat : Dimensionnement des instances en fonction des charges de travail
    - Utilisez des recommandations de redimensionnement dans AWS Cost Explorer ou AWS Compute Optimizerpour identifier les opportunités de redimensionnement.
- Négocier des accords de niveau de service (SLAs) : négociez pour SLAs permettre de réduire temporairement la capacité pendant que l'automatisation déploie des ressources de remplacement.

Ressources

Documents connexes :

- Optimisation de votre AWS infrastructure pour la durabilité, partie I : calcul
- Sélection du type d'instance basée sur les attributs pour Auto Scaling for Amazon Fleet EC2
- AWS Compute Optimizer Documentation
- Utilisation de Lambda : optimisation de la performance
- Documentation sur la scalabilité automatique

Vidéos connexes :

- AWS re:Invent 2023 - Nouveautés d'Amazon EC2
- AWS re:Invent 2023 - Économies intelligentes : stratégies d'optimisation des coûts d'Amazon Elastic Compute Cloud
- AWS re:Invent 2022 - Optimisation d'Amazon Elastic Kubernetes Service en termes de performances et de coûts AWS
- AWS re:Invent 2023 - Informatique durable : réduction des coûts et des émissions de carbone grâce à AWS

SUS05-BP02 Utiliser des types d’instance ayant le moins d’impact

Contrôlez et utilisez en permanence de nouveaux types d’instances pour tirer parti des améliorations de l’efficacité énergétique.

Anti-modèles courants :

- Vous n’utilisez qu’une seule famille d’instances.
- Vous n’utilisez que des instances x86.
- Vous spécifiez un type d’instance dans votre configuration Amazon EC2 Auto Scaling.
- Vous utilisez des instances AWS de manière non conforme à leur utilisation prévue (par exemple, vous utilisez des instances optimisées pour le calcul pour une charge de travail exigeante en mémoire).
- Vous n’évaluez pas régulièrement de nouveaux types d’instance.
- Vous ne vérifiez pas les recommandations des outils de redimensionnement AWS tels que AWS Compute Optimizer.

Avantages liés au respect de cette bonne pratique : en utilisant des instances économes en énergie et dimensionnées, vous pouvez grandement réduire l’impact sur l’environnement et le coût de votre charge de travail.

Niveau de risque encouru si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

L’utilisation d’instances efficaces dans les charges de travail du cloud est cruciale pour réduire l’utilisation des ressources et pour une meilleure rentabilité. Contrôlez de façon continue le lancement

de nouveaux types d'instances et profitez d'améliorations de l'efficacité énergétique, dont ces types d'instances conçus pour soutenir des charges de travail spécifiques comme l'entraînement et l'inférence du machine learning et le transcodage vidéo.

Étapes d'implémentation

- Découvrez et explorez les types d'instances : découvrez les types d'instances qui peuvent réduire l'impact environnemental de votre charge de travail.
  - Abonnez-vous à What's New with AWS pour rester au fait des dernières technologies et instances AWS.
  - Découvrez les différents types d'instance AWS.
  - Découvrez les instances AWS Graviton qui offrent les meilleures performances par watt d'énergie consommée dans Amazon EC2 en regardant re:Invent 2020 - Deep dive on AWS Graviton2 processor-powered Amazon EC2 instances et Deep dive into AWS Graviton3 and Amazon EC2 C7g instances.
- Utiliser les types d'instances ayant le moins d'impact : planifiez et migrez votre charge de travail vers les types d'instance avec le moins d'impact.
  - Définissez un processus pour évaluer les nouvelles fonctionnalités ou instances pour votre charge de travail. Profitez de l'agilité du cloud pour tester rapidement en quoi les nouveaux types d'instance peuvent améliorer la durabilité environnementale de votre charge de travail. Utilisez des métriques de proxy pour mesurer le nombre de ressources nécessaires pour mener à bien une unité de travail.
  - Si possible, modifiez votre charge de travail pour qu'elle fonctionne avec différents nombres de processeurs et différentes quantités de mémoire afin de maximiser votre choix de type d'instance.
  - Envisagez de migrer votre charge de travail vers des instances basées sur Graviton pour améliorer l'efficacité des performances de votre charge de travail. Pour en savoir plus sur le transfert de charges de travail vers AWS Graviton, consultez AWS Graviton Fast Start et considérations relatives à la transition de charges de travail vers des instances Amazon Elastic Compute AWS Cloud basées sur Graviton.
  - Envisagez de sélectionner l'option AWS Graviton lorsque vous utilisez des services gérés AWS
  - Migrez votre charge de travail vers des régions qui offrent des instances ayant le moindre impact en matière de durabilité et qui répondent à vos exigences métier.
  - Pour les charges de travail de machine learning, tirez parti d'un matériel conçu spécialement pour votre charge de travail, par exemple AWS, Trainum, AWS Inferentia et Amazon EC2 DL1.

AWS Les instances Inferentia telles que les instances Inf2 offrent des performances par watt jusqu'à 50 % supérieures à celles des instances comparables.

- Utilisez Amazon SageMaker AI Inference Recommender pour dimensionner correctement le point de terminaison d'inférence ML.
- Pour les charges de travail en dents de scie (charges de travail dont les besoins en capacité supplémentaire sont peu fréquents), il convient d'utiliser des instances de performance en rafale.
- Pour les charges de travail sans état et tolérantes aux pannes, utilisez les instances spot Amazon EC2 pour augmenter l'utilisation globale du nuage et réduire l'impact des ressources inutilisées sur le développement durable.

- Exploiter et optimiser : exploitez et optimisez votre instance de charge de travail.
  - Pour les charges de travail éphémères, évaluez les métriques d'instance Amazon CloudWatch telles que `CPUUtilization` pour identifier si l'instance est inactive ou sous-utilisée.
  - Pour les charges de travail stables, vérifiez les outils de redimensionnement AWS tels qu'AWS Compute Optimizer à intervalles réguliers pour identifier les opportunités d'optimisation et de redimensionnement des ressources de calcul. Pour d'autres exemples et recommandations, consultez les ateliers suivants :
    - Atelier Well-Architected : recommandations de redimensionnement
    - Well-Architected Lab - Rightsizing avec Compute Optimizer
    - Atelier Well-Architected : optimiser les modèles matériels et observer les indicateurs de performance clés de durabilité

Ressources

Documents connexes :

- Optimisation de votre infrastructure AWS pour la durabilité, partie 1 : calcul
- AWS Graviton
- Amazon EC2 DL1
- Flotte de réserve de capacité Amazon EC2
- Parc d'instances Spot Amazon EC2
- Fonctions : configuration des fonctions Lambda
- Sélection de type d'instance basée sur des attributs pour la flotte d'Amazon EC2
- Création d'applications durables, efficaces et optimisées en matière de coûts sur AWS

- Comment le tableau de bord de durabilité de Contino aide les clients à optimiser leur empreinte carbone

Vidéos connexes :

- AWS re:Invent 2023 - AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 - New Amazon Elastic Compute Cloud generative AI capabilities in AWS Management Console
- AWS re:Invent 2023 = What's new with Amazon Elastic Compute Cloud
- AWS re:Invent 2023 - Smart savings: Amazon Elastic Compute Cloud cost-optimization strategies
- AWS re:Invent 2021 - Deep dive into AWS Graviton3 and Amazon EC2 C7g instances
- AWS re:Invent 2022 - Build a cost-, energy-, and resource-efficient compute environment

Exemples connexes :

- Solution : conseils pour l'optimisation des charges de travail de deep learning pour atteindre la durabilité sur AWS
- Migration des bases de données Amazon Relational Database Service vers Graviton

SUS05-BP03 Utiliser des services gérés

Utilisez les services gérés pour fonctionner plus efficacement dans le cloud.

Anti-modèles courants :

- Vous utilisez des EC2 instances Amazon à faible taux d'utilisation pour exécuter vos applications.
- Votre équipe interne ne fait que gérer la charge de travail, sans avoir le temps de se concentrer sur l'innovation ou les simplifications.
- Vous déployez et maintenez des technologies pour des tâches qui peuvent être exécutées plus efficacement sur des services gérés.

Avantages liés au respect de cette bonne pratique :

- L'utilisation de services gérés transfère la responsabilité à AWS, qui dispose d'informations sur des millions de clients qui peuvent contribuer à de nouvelles innovations et à des gains d'efficacité.

- Le service géré répartit l'impact environnemental du service entre de nombreux utilisateurs grâce aux plans de contrôle multi-réseaux.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les services gérés transfèrent la responsabilité du maintien d'un taux d'utilisation élevé et de l'optimisation de la durabilité du matériel déployé. AWS Les services gérés suppriment également la charge opérationnelle et administrative liée à la maintenance d'un service, ce qui permet à votre équipe de disposer de plus de temps et de se concentrer sur l'innovation.

Passez en revue votre charge de travail pour identifier les composants qui peuvent être remplacés par des services AWS gérés. Par exemple, Amazon RDS, Amazon Redshift et Amazon ElastiCache fournissent un service de base de données géré. Amazon AthenaEMR, Amazon et Amazon OpenSearch Service fournissent un service d'analyse géré.

Étapes d'implémentation

1. Faites l'inventaire de votre charge de travail : dressez l'inventaire de votre charge de travail pour les services et les composants.
2. Identifier les candidats : évaluez et identifiez les composants qui peuvent être remplacés par des services gérés. Voici quelques exemples de situations dans lesquelles vous pourriez envisager de recourir à un service géré :

Tâche	Ce qu'il faut utiliser sur AWS
Hébergement d'une base de données	Utilisez des instances Amazon Relational Database Service (RDSAmazon) gérées au lieu de gérer vos propres instances Amazon RDS sur Amazon Elastic Compute Cloud (EC2Amazon).
Héberger une charge de travail en conteneur	Utilisez AWS Fargate, au lieu de mettre en œuvre votre propre infrastructure de conteneurs.
Hébergement d'applications Web	Utilisez l'hébergement AWS Amplify en tant que service CI/CD et d'hébergement

Tâche	Ce qu'il faut utiliser sur AWS
	entièrement géré pour les sites Web statiques et les applications Web rendues côté serveur.

3. Créez un plan de migration : identifiez les dépendances et créez un plan de migration. Mettez à jour les runbooks et les playbooks en conséquence.
   - AWS Application Discovery Service rassemble et présente automatiquement les informations sur les dépendances et l'utilisation des applications pour vous aider à prendre des décisions en connaissance de cause pour votre programme de migration.
4. Effectuer des tests : testez le service avant de migrer vers le service géré.
5. Remplacez les services autohébergés : utilisez votre plan de migration pour remplacer les services autohébergés par des services gérés.
6. Contrôler et ajuster : surveillez continuellement le service une fois la migration terminée afin d'apporter les modifications nécessaires et d'optimiser le service.

Ressources

Documents connexes :

- AWS Cloud Produits
- AWS Calculateur du coût total de possession (TCO)
- Amazon DocumentDB
- Amazon Elastic Kubernetes Service () EKS
- Amazon Managed Streaming pour Apache Kafka (Amazon) MSK

Vidéos connexes :

- AWS re:Invent 2021 - Des opérations cloud à grande échelle avec AWS Managed Services
- AWS re:Invent 2023 - Meilleures pratiques pour opérer sur AWS

SUS05-BP04 Optimiser votre utilisation des accélérateurs de calcul matériels

Optimisez votre utilisation des instances de calcul accéléré pour réduire les exigences d'infrastructure physique de votre charge de travail.

Anti-modèles courants :

- Vous ne surveillez pas l'utilisation du GPU.
- Vous utilisez une instance à usage général pour la charge de travail alors qu'une instance spécialement conçue peut fournir des performances supérieures, des coûts plus faibles et de meilleures performances par watt.
- Vous utilisez des accélérateurs de calcul matériels pour les tâches où ils sont plus efficaces en utilisant des alternatives basées sur l'UC.

Avantages liés au respect de cette bonne pratique : en optimisant l'utilisation des accélérateurs matériels, vous pouvez réduire les exigences de votre charge de travail en matière d'infrastructure physique.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Si vous avez besoin d'une capacité de traitement élevée, vous pouvez bénéficier de l'utilisation d'instances de calcul accéléré, qui vous donnent accès à des accélérateurs de calcul matériels tels que des unités de traitement graphique (GPU) et des matrices de portes programmables sur site (FPGA). Ces accélérateurs matériels exécutent certaines fonctions comme le traitement graphique ou la correspondance de modèles de données plus efficacement que les alternatives basées sur l'UC. De nombreuses charges de travail accélérées, telles que le rendu, le transcodage et le machine learning, sont très variables en matière d'utilisation des ressources. Exécutez ce matériel uniquement pendant le temps nécessaire et mettez-le hors service grâce à l'automatisation lorsque vous n'en avez plus besoin afin de limiter les ressources consommées.

Étapes d'implémentation

- Explorer les accélérateurs de calcul : identifiez les instances de calcul accéléré qui peuvent répondre à vos besoins.
- Utiliser du matériel conçu spécialement : pour les charges de travail de machine learning, tirez parti d'un matériel conçu spécialement pour votre charge de travail, tel qu'AWS Trainum, AWS Inferentia et Amazon EC2 DL1. AWS Les instances Inferentia telles que les instances Inf2 offrent des performances par watt jusqu'à 50 % supérieures à celles des instances comparables.
- Surveiller les métriques d'utilisation : collectez des métriques d'utilisation pour vos instances de calcul accéléré. Par exemple, vous pouvez utiliser l'agent CloudWatch pour collecter des métriques

telles que `utilization_gpu` et `utilization_memory` pour vos GPU, comme indiqué dans Collecter les métriques des GPU NVIDIA avec Amazon CloudWatch.

- Redimensionner : optimisez le code, le fonctionnement du réseau et les paramètres des accélérateurs matériels pour veiller à ce que le matériel sous-jacent soit pleinement utilisé.
  - Optimiser les paramètres GPU
  - Surveillance et optimisation des GPU dans l'AMI Deep Learning
  - Optimisation des E/S pour le réglage des performances de GPU pour l'entraînement du deep learning dans l'IA Amazon SageMaker
- Maintenir à jour : utilisez les derniers pilotes GPU et bibliothèques à hautes performances.
- Libérer les instances non requises : utilisez l'automatisation pour libérer les instances GPU lorsqu'elles ne sont pas utilisées.

Ressources

Documents connexes :

- Calcul accéléré
- Let's Architect! Architecture avec des puces personnalisées et des accélérateurs
- Comment choisir le type d'instance EC2 approprié pour ma charge de travail ?
- Instances Amazon EC2 VT1
- Choisissez le meilleur accélérateur d'IA et la meilleure compilation de modèles pour l'inférence de vision par ordinateur avec l'IA Amazon SageMaker

Vidéos connexes :

- AWS re:Invent 2021 - How to select Amazon EC2 GPU instances for deep learning
- AWS Online Tech Talks - Deploying Cost-Effective Deep Learning Inference
- AWS re:Invent 2023 - Cutting-edge AI with AWS and NVIDIA
- AWS re:Invent 2022 - [NEW LAUNCH!] Introducing AWS Inferentia2-based Amazon EC2 Inf2 instances
- AWS re:Invent 2022 - Accelerate deep learning and innovate faster with AWS Trainium
- AWS re:Invent 2022 - Deep learning on AWS with NVIDIA: From training to deployment

Processus et culture

Question

- SUS 6 Comment vos processus organisationnels soutiennent-ils vos objectifs de durabilité ?

SUS 6 Comment vos processus organisationnels soutiennent-ils vos objectifs de durabilité ?

Recherchez des opportunités de réduire votre impact en matière de durabilité modifiant vos pratiques de développement, de test et de déploiement.

Bonnes pratiques

- SUS06-BP01 Communiquer et répercuter en cascade vos objectifs de durabilité
- SUS06-BP02 Adopter des méthodes qui peuvent rapidement introduire des améliorations en matière de durabilité
- SUS06-BP03 Maintenir à jour votre charge de travail
- SUS06-BP04 Augmenter l'utilisation des environnements de génération
- SUS06-BP05 Utiliser des tests Device Farms gérés

SUS06-BP01 Communiquer et répercuter en cascade vos objectifs de durabilité

La technologie est un élément clé de la durabilité. Les équipes informatiques jouent un rôle crucial dans la mise en œuvre de changements significatifs pour atteindre les objectifs de durabilité de votre organisation. Ces équipes doivent comprendre clairement les objectifs de durabilité de l'entreprise et s'efforcer de communiquer et de répercuter ces priorités dans l'ensemble des activités.

Anti-modèles courants :

- Vous ne connaissez pas les objectifs de durabilité de votre organisation ni comment ils s'appliquent à votre équipe.
- Vous n'êtes pas suffisamment sensibilisé et formé à l'impact environnemental des charges de travail dans le cloud.
- Vous n'êtes pas sûr des domaines spécifiques à privilégier.
- Vous n'impliquez pas vos employés et vos clients dans vos initiatives de durabilité.

Avantages liés au respect de cette bonne pratique : de l'optimisation de l'infrastructure et des systèmes à l'utilisation de technologies innovantes, les équipes informatiques peuvent réduire les émissions de carbone de l'organisation et minimiser la consommation de ressources. La communication des objectifs de durabilité peut permettre aux équipes informatiques de s'améliorer et de s'adapter en permanence à l'évolution des défis en matière de durabilité. De plus, ces optimisations durables se traduisent aussi souvent par des économies de coûts, ce qui renforce l'argumentaire.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Les principaux objectifs de durabilité des équipes informatiques devraient être d'optimiser les systèmes et les solutions afin d'accroître l'efficacité des ressources et de minimiser l'empreinte carbone de l'organisation et son impact global sur l'environnement. Les services partagés et les initiatives telles que les programmes de formation et les tableaux de bord opérationnels, peuvent aider les organisations à optimiser leurs opérations informatiques et à élaborer des solutions susceptibles de contribuer à réduire de manière significative l'empreinte carbone. Le cloud offre l'opportunité non seulement de transférer les responsabilités en matière d'infrastructure physique et d'approvisionnement en énergie au fournisseur cloud, mais également d'optimiser en permanence l'efficacité des ressources des services basés sur le cloud.

Lorsque les équipes utilisent le modèle d'efficacité et de responsabilité partagée inhérent au cloud, elles peuvent réduire de manière significative l'impact de l'organisation sur l'environnement. Cela peut à son tour contribuer aux objectifs globaux de durabilité de l'organisation et démontrer la valeur de ces équipes en tant que partenaires stratégiques sur la voie d'un futur plus durable.

Étapes d'implémentation

- Définir des buts et des objectifs : établissez des objectifs bien définis pour votre programme informatique. Cela implique de recueillir l'avis des parties prenantes responsables de différents départements tels que l'informatique, la durabilité et les finances. Ces équipes doivent définir des objectifs mesurables qui s'alignent sur les objectifs de durabilité de votre organisation, notamment dans des domaines tels que la réduction des émissions carbone et l'optimisation des ressources.
- Comprendre les limites de la comptabilisation du carbone dans votre entreprise : découvrez comment les méthodes de comptabilisation du carbone telles que le protocole sur les gaz à effet de serre (GES) sont liées à vos charges de travail dans le cloud (pour plus de détails, consultez Durabilité du cloud).

- Utiliser des solutions cloud pour la comptabilisation du carbone : utilisez des solutions cloud telles que les solutions de comptabilisation du carbone sur AWS pour suivre les émissions de GES de première, deuxième et troisième catégories dans l'ensemble de vos opérations, portefeuilles et chaînes de valeur. Grâce à ces solutions, les organisations peuvent rationaliser l'acquisition de données sur les émissions de GES, simplifier les rapports et obtenir des informations pour étayer leurs stratégies climatiques.
- Surveiller l'empreinte carbone de votre portefeuille informatique : suivez et signalez les émissions de carbone de vos systèmes informatiques. Utilisez l'outil d'empreinte carbone du clientAWS pour suivre, mesurer, réviser et prédire les émissions de carbone générées par votre utilisation d'AWS.
- Communiquer à vos équipes l'utilisation des ressources via des métriques proxy : suivez et signalez votre utilisation de ressources via des métriques proxy. Dans les modèles de tarification à la demande du cloud, l'utilisation des ressources est liée au coût, qui est une métrique généralement facile à comprendre. Au minimum, utilisez le coût comme une métrique proxy pour communiquer l'utilisation des ressources et les améliorations apportées par chaque équipe.
  - Activer la granularité horaire dans Cost Explorer et créer un rapport d'utilisation et de coût (CUR) : le rapport CUR fournit une granularité d'utilisation journalière ou horaire, des tarifs, des coûts et des attributs d'utilisation pour tous les services AWS. Utilisez Cloud Intelligence Dashboards et son tableau de bord des métriques proxy de durabilité comme point de départ pour le traitement et la visualisation des données basées sur les coûts et l'utilisation. Pour plus de détails, consultez les documents suivants :
  - Mesure et suivi de l'efficacité du cloud à l'aide de métriques proxy de durabilité, partie I : Que sont les métriques proxy ?
  - Mesure et suivi de l'efficacité du cloud à l'aide de métriques proxy de durabilité, partie II : Établissement d'un pipeline de métriques
- Optimiser et évaluer en permanence : utilisez un processus d'amélioration pour optimiser en permanence vos systèmes informatiques, y compris la charge de travail dans le cloud pour des raisons d'efficacité et de durabilité. Surveillez l'empreinte carbone avant et après la mise en œuvre de la stratégie d'optimisation. Utilisez la réduction de l'empreinte carbone pour évaluer l'efficacité.
- Favoriser une culture de la durabilité : utilisez des programmes de formation (tels que AWS Skill Builder) pour sensibiliser vos employés à la durabilité. Impliquez-les dans des initiatives de durabilité. Partagez et célébrez leurs témoignages de réussite. Utilisez des incitations pour les récompenser s'ils atteignent leurs objectifs de durabilité.

Ressources

Documents connexes :

- Comprendre les estimations de vos émissions de carbone

Vidéos connexes :

- AWS re:Invent 2023 - Accelerate data-driven circular economy initiatives with AWS
- AWS re:Invent 2023 - Sustainability innovation in AWS Global Infrastructure
- AWS re:Invent 2023 - Sustainable architecture: Past, present, and future
- AWS re:Invent 2022 - Delivering sustainable, high-performing architectures
- AWSre:Invent 2022 – L'architecture de manière durable et réduisez votre AWSempreinte carbone
- AWS re:Invent 2022 - Sustainability in AWS global infrastructure

Exemples connexes :

- Well-Architected Lab – Transformer les rapports sur les coûts et l'utilisation en rapports d'efficacité

Formations associées :

- Transformation en matière de durabilité sur AWS
- SimuLearn – Rapports sur la durabilité
- Décarbonisation avec AWS

SUS06-BP02 Adopter des méthodes qui peuvent rapidement introduire des améliorations en matière de durabilité

Adoptez des méthodes et des processus pour valider les améliorations potentielles, minimiser les coûts des tests et apporter de petites améliorations.

Anti-modèles courants :

- L'examen de la durabilité de votre application est une tâche qui n'est effectuée qu'une seule fois au début d'un projet.
- Votre charge de travail est devenue obsolète, car le processus de lancement est trop lourd pour introduire des changements mineurs dans un souci d'efficacité des ressources.

- Vous ne disposez pas de mécanismes pour améliorer votre charge de travail afin d'atteindre davantage de durabilité.

Avantages liés au respect de cette bonne pratique : en établissant un processus pour introduire et suivre les améliorations en matière de durabilité, vous serez en mesure d'adopter en permanence de nouvelles fonctionnalités et capacités, de résoudre les problèmes et d'améliorer l'efficacité de la charge de travail.

Niveau de risque encouru si cette bonne pratique n'est pas respectée : moyen

Directives d'implémentation

Testez et validez les améliorations potentielles en matière de durabilité avant de les déployer en production. Tenez compte du coût des tests lors du calcul des avantages futurs potentiels d'une amélioration. Développez des méthodes d'essai à faible coût pour apporter de petites améliorations.

Étapes d'implémentation

- Comprenez et communiquez les objectifs de durabilité de votre organisation : comprenez les objectifs de durabilité de votre organisation, tels que la réduction des émissions de carbone ou la gestion de l'eau. Traduisez ces objectifs en exigences de durabilité pour vos charges de travail cloud. Communiquez ces exigences aux principales parties prenantes.
- Ajoutez des exigences de durabilité à votre carnet de commandes : ajoutez des exigences relatives à l'amélioration de la durabilité à votre carnet de développement.
- Itérer et améliorer : utilisez un processus d'amélioration itératif pour identifier, évaluer, prioriser, tester et déployer ces améliorations.
- Test à l'aide d'un produit minimum viable (MVP) : développez et testez des améliorations potentielles en utilisant le minimum de composants représentatifs viables afin de réduire le coût et l'impact environnemental des tests.
- Rationaliser le processus : améliorez et rationalisez en permanence vos processus de développement. À titre d'exemple, automatisez votre processus de livraison de logiciels en utilisant des pipelines d'intégration et de livraison continues (CI/CD) pour tester et déployer les améliorations potentielles afin de réduire le niveau d'effort et de limiter les erreurs causées par les processus manuels.
- Formation et sensibilisation : organisez des programmes de formation pour les membres de votre équipe afin de les sensibiliser au développement durable et à l'impact de leurs activités sur les objectifs de durabilité de votre organisation.

- Évaluer et ajuster : évaluez en permanence l'impact des améliorations et procédez aux ajustements nécessaires.

Ressources

Documents connexes :

- AWS active des solutions de durabilité

Vidéos connexes :

- AWS re:Invent 2023 - Sustainable architecture: Past, present, and future
- AWS re:Invent 2022 - Delivering sustainable, high-performing architectures
- AWSre:Invent 2022 - L'architecture de manière durable et réduisez votre AWSempreinte carbone
- AWS re:Invent 2022 - Sustainability in AWS global infrastructure
- AWS re:Invent 2023 - What's new with AWS observability and operations

Exemples connexes :

- Well-Architected Lab – Transformer les rapports sur les coûts et l'utilisation en rapports d'efficacité

SUS06-BP03 Maintenir à jour votre charge de travail

Maintenez votre charge de travail à jour pour adopter des fonctionnalités efficaces, supprimer les problèmes et améliorer l'efficacité globale de votre charge de travail.

Anti-modèles courants :

- Vous supposez que votre architecture actuelle est statique et ne sera pas mise à jour au fil du temps.
- Vous ne disposez pas de systèmes ou de rythme régulier pour évaluer la compatibilité des packages et des logiciels mis à jour avec votre charge de travail.

Avantages liés au respect de cette bonne pratique : en mettant en place un processus pour garder votre charge de travail à jour, vous pourrez adopter de nouvelles fonctionnalités et capacités, résoudre les problèmes et améliorer l'efficacité de la charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

Des systèmes d'exploitation, des moteurs d'exécution, des logiciels intermédiaires, des bibliothèques et des applications à jour peuvent améliorer l'efficacité de la charge de travail et faciliter l'adoption de technologies plus efficaces. Les logiciels à jour peuvent également inclure des fonctions permettant de mesurer plus précisément l'impact en matière de durabilité de votre charge de travail, car les fournisseurs proposent des fonctions pour atteindre leurs propres objectifs de durabilité. Adoptez une cadence régulière pour maintenir votre charge de travail à jour avec les dernières fonctionnalités et versions.

Étapes d'implémentation

- Définir un processus : utilisez un processus et un calendrier pour évaluer les nouvelles fonctionnalités ou instances pour votre charge de travail. Profitez de l'agilité du cloud pour tester rapidement en quoi les nouvelles fonctionnalités peuvent permettre à votre charge de travail de :
  - Réduire les impacts sur la durabilité.
  - Gagner en efficacité de la performance.
  - Supprimer les obstacles à une amélioration planifiée.
  - Améliorer votre capacité à mesurer et à gérer les impacts en matière de durabilité.
- Réaliser un inventaire : établissez l'inventaire de votre logiciel de charge de travail et de l'architecture, et identifiez les composants pouvant être mis à jour.
  - Vous pouvez utiliser la fonctionnalité AWS Systems Manager Inventory pour collecter les métadonnées du système d'exploitation (OS), de l'application et de l'instance à partir de vos instances Amazon EC2 et comprendre rapidement quelles instances exécutent les logiciels et les configurations requis par votre politique logicielle et quelles instances doivent être mises à jour.
- Découvrez la procédure de mise à jour : comprenez comment mettre à jour les composants de votre charge de travail.

Composant de charge de travail	Comment mettre à jour
Images de machine	Utilisez EC2 Image Builder pour gérer les mises à jour des images de serveur Amazon Machine Images (AMI) pour Linux ou Windows.

Composant de charge de travail	Comment mettre à jour
Images de conteneur	Utilisez Amazon Elastic Container Registry (Amazon ECR) avec votre pipeline existant pour gérer des images Amazon Elastic Container Service (Amazon ECS).
AWS Lambda	AWS Lambda comprend les fonctionnalités de gestion des versions.

- Utiliser l'automatisation : utilisez l'automatisation pour le processus de mise à jour afin de réduire le niveau d'effort nécessaire au déploiement des nouvelles fonctionnalités et de limiter les erreurs causées par les processus manuels.
  - Vous pouvez utiliser CI/CD pour mettre automatiquement à jour les AMI, les images de conteneurs et d'autres artefacts liés à votre application cloud.
  - Vous pouvez utiliser des outils tels que le gestionnaire de correctifs AWS Systems Manager pour automatiser le processus de mise à jour du système et planifier l'activité à l'aide des fenêtres de maintenance AWS Systems Manager.

Ressources

Documents connexes :

- Centre d'architecture AWS
- Nouveautés avec AWS
- Outils pour développeurs AWS

Vidéos connexes :

- AWS re:Invent 2022 - Optimize your AWS workloads with best-practice guidance
- All Things Patch : AWS Systems Manager

Exemples connexes :

- Well-Architected Labs - Inventory and Patch Management

- Atelier : AWS Systems Manager

SUS06-BP04 Augmenter l’utilisation des environnements de génération

Augmentez l’utilisation des ressources pour développer, tester et construire vos charges de travail.

Anti-modèles courants :

- Vous provisionnez ou résiliez manuellement vos environnements de construction.
- Vous faites fonctionner vos environnements de construction indépendamment des activités de test, de construction ou de lancement (par exemple, en faisant fonctionner un environnement en dehors des heures de travail des membres de votre équipe de développement).
- Vous provisionnez trop de ressources pour vos environnements de construction.

Avantages liés au respect de cette bonne pratique : en augmentant l’utilisation des environnements de création, vous pouvez améliorer l’efficacité globale de votre charge de travail dans le cloud tout en allouant les ressources aux concepteurs pour qu’ils puissent développer, tester et créer efficacement.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : faible

Directives d’implémentation

Exploitez l’automatisation et l’infrastructure en tant que code pour mettre en place des environnements de construction lorsque cela est nécessaire et les arrêter lorsqu’ils ne sont pas utilisés. Un modèle courant consiste à planifier des périodes de disponibilité qui coïncident avec les heures de travail des membres de votre équipe de développement. Vos environnements de test doivent ressembler de près à la configuration de production. Toutefois, recherchez les possibilités d’utilisation des types d’instance avec une capacité de débordement, des instances Amazon EC2 Spot, des services de base de données à mise à l’échelle automatique, des conteneurs et des technologies sans serveur pour aligner la capacité de développement et de test sur l’utilisation. Limitez le volume de données pour répondre aux exigences du test. Si vous utilisez des données de production dans les tests, étudiez les possibilités de partager les données de production et de ne pas déplacer les données à un autre emplacement.

Étapes d’implémentation

- Utiliser l’infrastructure en tant que code : utilisez l’infrastructure en tant que code pour provisionner vos environnements de construction.

- Utilisez l’automatisation : utilisez l’automatisation pour gérer le cycle de vie de vos environnements de développement et de test et maximiser l’efficacité de vos ressources de construction.
- Maximiser l’utilisation : utilisez des stratégies pour optimiser l’utilisation des environnements de développement et de test.
  - Utilisez des environnements représentatifs viables minimum pour développer et tester les améliorations potentielles.
  - Utilisez les technologies sans serveur si possible.
  - Utilisez des instances à la demande pour compléter les appareils de vos développeurs.
  - Utilisez des types d’instance à capacité de débordement, des instances Spot et d’autres technologies pour harmoniser la capacité de création et l’utilisation.
  - Adoptez des services natifs du cloud pour l’accès à un shell d’instance sécurisé plutôt que de déployer des flottes d’hôtes bastion.
  - Mettez automatiquement à l’échelle vos ressources de construction en fonction de vos tâches de construction.

Ressources

Documents connexes :

- Gestionnaire de session AWS Systems Manager
- Instances de performance à capacité extensible Amazon EC2
- Présentation de AWS CloudFormation
- Présentation de AWS CodeBuild
- Instance Scheduler sur AWS

Vidéos connexes :

- AWS re:Invent 2023 - Continuous integration and delivery for AWS

SUS06-BP05 Utiliser des tests Device Farms gérés

Utilisez les Device Farms gérés pour tester efficacement une nouvelle fonctionnalité sur un ensemble représentatif de matériel.

Anti-modèles courants :

- Vous testez et déployez manuellement votre application sur des appareils physiques individuels.
- Vous n'utilisez pas le service de test d'applications pour tester et interagir avec vos applications (par exemple, les applications Android, iOS et Web) sur des appareils physiques réels.

Avantages liés au respect de cette bonne pratique : l'utilisation de batteries d'appareils gérés pour tester des applications compatibles avec le cloud présente de nombreux avantages :

- La solution comprend des fonctionnalités plus efficaces pour tester l'application sur de nombreux appareils différents.
- Elle élimine la nécessité d'une infrastructure interne pour les essais.
- Elle permet l'utilisation de divers types d'appareils, y compris des matériels plus anciens et moins populaires, ce qui élimine le besoin de mises à niveau inutiles des appareils.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : faible

Directives d'implémentation

L'utilisation de Device Farms gérés peut vous aider à rationaliser le processus de test des nouvelles fonctionnalités sur un ensemble représentatif de matériel. Les tests Device Farms gérés proposent divers types d'appareils, notamment du matériel plus ancien et moins courant, et permettent d'éviter que les mises à niveau inutiles d'appareils affectent la durabilité des clients.

Étapes d'implémentation

- Définir les exigences en matière de tests : définissez vos exigences et votre plan de test (comme le type de test, les systèmes d'exploitation et le calendrier des tests).
  - Vous pouvez utiliser Amazon CloudWatch RUM pour collecter et analyser les données côté client et élaborer votre plan de test.
- Sélectionnez un parc d'appareils gérés : sélectionnez un parc d'appareils gérés capable de répondre à vos exigences en matière de tests. Par exemple, vous pouvez utiliser AWS Device Farm pour tester et comprendre l'impact de vos modifications sur un ensemble représentatif de matériel.
- Utiliser l'automatisation : utilisez l'intégration continue/le déploiement continu (CI/CD) pour programmer et exécuter vos tests.
  - Intégration d'AWS Device Farm à votre pipeline CI/CD pour exécuter des tests Selenium sur plusieurs navigateurs

  - Création et test d’applications iOS et iPadOS avec AWS DevOps et les services mobiles
- Examiner et ajuster : examinez continuellement les résultats de vos tests et apportez les améliorations nécessaires.

Ressources

Documents connexes :

- Liste d’appareils AWS Device Farm
- Affichage du tableau de bord CloudWatch RUM

Vidéos connexes :

- AWS re:Invent 2023 - Improve your mobile and web app quality using AWS Device Farm
- AWS re:Invent 2021 - Optimize applications through end user insights with Amazon CloudWatch RUM

Exemples connexes :

- Exemple d’application AWS Device Farm pour Android
- Exemple d’application AWS Device Farm pour iOS
- Tests Web Appium pour AWS Device Farm

Avis

Il incombe aux clients de procéder à une évaluation indépendante des informations contenues dans le présent document. Ce document : (a) est fourni à titre informatif uniquement, (b) représente les offres de produits et les pratiques AWS actuelles, qui sont sujettes à modification sans préavis, et (c) ne donne lieu à aucun engagement ni aucune assurance de la part d'AWS et de ses sociétés apparentées, fournisseurs ou concédants de licence. Les produits ou services AWS sont fournis « tels quels » sans garantie, représentation ou condition d'aucune sorte, tant expresse qu'implicite. Les responsabilités et obligations d'AWS vis-à-vis de ses clients sont régies par les contrats AWS. Le présent document ne fait partie d'aucun et ne modifie aucun contrat entre AWS et ses clients.

AWS Glossaire

Pour la AWS terminologie la plus récente, consultez le AWS glossaire dans la Glossaire AWS référence.

Les traductions sont fournies par des outils de traduction automatique. En cas de conflit entre le contenu d'une traduction et celui de la version originale en anglais, la version anglaise prévaudra.