L'influence de la gravité des données dans les architectures des lacs de données

Cédrine Madera[1], Anne Laurent[2], Thérèse Libourel[3], André Miralles[4].

1. *IBM & LIRMM, Montpellier, France*
   cedrinemadera@fr.ibm.com
2. *Université de Montpellier LIRMM, Montpellier, France*
   anne.laurent@umontpellier.fr
3. *UMR Espace-Dev (UM, IRD, UG, UA, ULR), Université de Montpellier*
   therese.libourel@umontpellier
4. *UMR Tetis/IRSTEA, Maison de la télédétection, Montpellier, France*
   andre.miralles@teledetection.fr

RESUME. La révolution digitale qui met au cœur de sa stratégie la donnée fait émerger le concept de lac de données. Celui-ci devient un composant incontournable pour la découverte de l'information potentiellement enfouie dans les données. Nombre d'industriels qui s'engagent sur cette voie recourent de plus en plus à l'intégration de lacs de données dans leur système d'information et utilisent le plus souvent une plateforme fédératrice, reposant sur la technologie open source « Apache Hadoop ». Cette approche purement industrielle mono technologie commence à trouver ses limites. Dans cet article, nous nous intéressons, d'un point de vue académique, à l'hypothèse de la remise en cause de cette mono technologie par divers facteurs, dont ceux liés à la gravité des données. Nous illustrons notre hypothèse par un cas d'usage en milieu industriel.

ABSTRACT. The digital revolution that puts the data at the heart of its strategy brings out the new concept of data lake. It becomes an essential component for the discovery of information potentially hidden in data. Many practitioners who commit to this pat rely heavily on the integration of data lakes in their information system and most often use a unifying platform, based on an open source technology "Apache Hadoop". This unique industrial approach begins to find its limits. We are interested, from an academic point of view, in the hypothesis of the questioning of this mono technology by various factors, including those related to the data gravity. We illustrate our hypothesis with a use case in industrial environment.

MOTS-CLES : lac de données, gravité des données, architecture informatique, système d'information, déplacement de données, duplication de données

KEYWORDS: data lake, data gravity, architecture, information system, data migration, data duplication

1. Introduction

L'internet des objets associé à la production, sans cesse croissante, de données émises dans les systèmes d'information traditionnels conduit à une accumulation de données disponibles sans précédent. Ces données disponibles, si elles attisent les convoitises en vue d'en tirer une richesse en termes d'information notamment, posent des questions quant à leur conservation, leur pertinence, leur mise en forme, leur gouvernance mais surtout leur capitalisation en vue de valorisation ultérieure.

La capitalisation de cette « richesse » est un enjeu majeur des systèmes d'information d'aujourd'hui mais aussi de demain. Le lac de données (*Data Lake*) est un nouveau composant du système d'information (Madera et Laurent, 2016), il met au cœur de sa conception la donnée et non l'information à délivrer, complétant les autres composants existants, tels que les systèmes décisionnels. Un de leur principal objectif est de permettre l'exploration et l'analyse de sources de données diverses afin de trouver de nouveaux « modèles » (*« pattern »*) d'information.

Dans le contexte industriel, les architectes d'information[1] ont en charge la gouvernance, la conception et l'outillage technologique de ces lacs de données. L'architecture d'information qu'ils doivent mettre en place doit prendre en compte conjointement des contraintes fonctionnelles (*cf.* section 2) mais aussi non fonctionnelles. De par l'important volume de données à traiter, la diversité des formats de ces données mais aussi leur coût considéré comme peu élevé, la technologie de type *Apache Hadoop* s'est imposée comme référente, quasi unique, occultant les discussions sur l'impact des contraintes non fonctionnelles sur ce choix. Cette technologie comme solution unique pour les lacs de données est désormais remise en cause (Russom, 2017) par le monde industriel et des architectures hybrides, avec introduction de technologie complémentaire à *Apache Hadoop*, sont désormais envisagées.

Notre intérêt académique se porte sur les facteurs pouvant influencer cette hybridation technologique mais aussi applicative. Les travaux de (McCrory, 2010 ; Alrehamy et Walker, 2015) ont introduit la notion de gravité des données qui, du point de vue de l'architecture, peut être considérée comme une contrainte non fonctionnelle. Partant de cette idée, un des objectifs de nos recherches en cours consiste à compléter la définition de (McCrory, 2010) afin d'étudier son impact sur les architectures des lacs de données. Nous vérifions nos hypothèses au travers de l'analyse d'un cas réel d'architecture d'un lac de données en milieu industriel.

La suite de cet article se présente de la manière suivante. Dans la section 2, nous rappelons les notions de lac de données et de gravité des données. La section 3 est

[1] L'architecte d'information se concentre sur les éléments requis pour structurer les aspects informationnels et données des solutions retenues et pour concevoir, construire, tester, installer, exploiter et maintenir le système d'information de la solution. Pour mener à bien sa mission, l'architecte système d'information doit en premier lieu étudier les besoins fonctionnels, établir une cartographie du système en analysant l'existant, puis proposer un modèle d'architecture et enfin la mettre en œuvre en choisissant une infrastructure matérielle et logicielle.

consacrée l'impact de la prise en compte de la gravité des données sur l'architecture des lacs de données. Dans la section 4 nous présentons notre étude expérimentale, au travers du cas d'usage industriel d'un lac de donnée dédié à la collecte de données de métrologie[2] d'un parc informatique. Nous concluons et présentons quelques perspectives dans la section 5.

2. Introduction des concepts de lac de données et de gravité des données

2.1 Les lacs de données

Nous considérons les lacs de données comme un nouveau composant du système d'information, qui se positionne en complément des systèmes décisionnels existants tels que les entrepôts de données. Leur principal objectif est de permettre la capitalisation des données d'une organisation, dans leur format le plus brut, afin d'en extraire de la valeur et permettre une valorisation du capital données de l'organisation. Dans nos précédents travaux (Madera et Laurent, 2016), nous en avons donné la définition suivante :

Le lac de données est une collection de données, non transformées, de formats non contraints (tous formats acceptés), conceptuellement rassemblées en un endroit unique mais potentiellement non matérialisé, destinées à un/des utilisateurs experts en science de données, munie d'un catalogue de méta-données ainsi que d'un ensemble de règles et méthodes de gouvernance de données.

Les lacs de données sont très souvent associés à la technologie de type *Apache Hadoop* (MarketsAndMarkets, 2016), ce qui, en termes d'architecture, peut limiter les solutions à explorer. En effet, le choix de solutions pour supporter les architectures des lacs de données si il se cantonne au champ des solutions basées sur *Apache Hadoop* peut certes simplifier le champ d'investigation mais il peut aussi occulter certaines problématiques générées par ce choix, voire oublier de répondre à certaines contraintes des lacs de données. Ces contraintes sont notamment la sensibilité des données que le lac de données souhaite collecter. La volumétrie en est une autre tout comme le coût de déplacement de ces données vers le lac. Ces contraintes peuvent remettre en cause une solution de collecte massive de données et imposer plutôt une approche où les données ne sont pas déplacées.

Sans remettre en cause la création d'environnement collecteur de données, nous souhaitons dans cet article étudier quels facteurs pourraient influencer, voire remettre en cause les architectures des lacs de données où toutes les données sont déplacées physiquement vers un ou plusieurs environnements de stockage composant les lacs de données. La gravité est un de ces facteurs.

---

[2] Dans le cadre d'un parc informatique (réseaux, serveurs, baies de stockages, etc.), l'objectif de la métrologie est de connaître et de comprendre le fonctionnement du parc informatique afin de pouvoir, non seulement intervenir dans l'urgence en cas de problème, mais aussi d'améliorer les performances, d'anticiper son évolution et sa planification.

2.2 La gravité des données

Par analogie de raisonnement entre la gravitation en sciences physiques[3], les données s'accumulent avec le temps, et peuvent être considérées comme plus denses ou avoir une plus grande masse. Lorsque la densité ou la masse croissent, l'attraction gravitationnelle des données augmente. Les services et applications ont leur propre masse et, par conséquent, ont leur propre gravité ; mais les données sont beaucoup plus volumineuses et plus denses qu'eux. Ainsi, alors que les données continuent d'augmenter, les services et les applications sont plus susceptibles d'être attirés par les données, plutôt que l'inverse. Cela ressemble, par mimétisme avec la gravité au sens physique, à l'exemple de la pomme qui tombe sur la Terre plutôt que l'inverse parce que la Terre a plus de masse que la pomme

Les travaux de McCrory (McCrory, 2010) sont les premiers à avoir exposés cette l'analogie entre la gravité de la donnée et la gravité au sens physique, en définissant la force d'attraction entre données et traitements. Dans cette analogie interviennent la masse des données, la vitesse de déplacement de ces données et les traitements/services qui y sont associés. La loi de la gravité stipule que l'attraction entre les objets est directement proportionnelle à leur masse. Dave McCrory (McCrory, 2014) a réutilisé le terme gravité des données pour décrire le phénomène dans lequel le nombre ou la quantité et la vitesse à laquelle les services, les applications, et même les clients sont attirés par les données, augmentent à mesure que la masse des données augmente. Le phénomène de gravitation peut alors être appliqué. Les données qui voient leur gravité augmenter vont attirer les traitements à elles. La force d'attraction exercée par les données sur les traitements ouvre la porte à d'autres paramètres pouvant influencer cette gravité tels que la sensibilité, le trafic du réseau, le coût, etc.

(Alrehamy et Walker, 2015), s'appuyant sur les travaux (McCrory, 2010), mettent en exergue cette gravité des données dans leur lac de données fonctionnellement dédié aux données personnelles. Cependant, leur évaluation de la gravité des données, dans ce cas d'étude où la masse des données s'avère peu importante, les amène à penser, que le paramètre le plus influent n'est pas la masse mais la sensibilité des données ; c'est elle qui va « peser » le plus dans l'évaluation de la gravité des données. Dans leur lac de données dédié aux données personnelles, celles-ci ont une sensibilité si forte que, ce lac attire à lui les traitements devant manipuler ces données. Dans ces travaux, c'est la sensibilité, un des paramètres que les auteurs ont inclus dans la gravité des donnés, qui influence l'attraction du traitement vers les données.

Ces premiers travaux intégrant la prise en compte de la gravité des données via volume (ou la masse) et sensibilité des données tendent à prouver l'influence qui peut s'exercer sur la relation donnée-traitement.

---

[3] En sciences physiques, la gravitation désigne la force qui fait que deux masses s'attirent mutuellement, comme la Terre et le Soleil. La gravité en est le résultat. C'est ce qui fait tomber les objets, comme la pomme tombée d'un arbre observée par Newton.

Il convient donc de regarder, en se basant sur ces travaux et cette analogie physique-gravitation, quelle pourrait être l'influence de la gravité des données dans les architectures des lacs de données.

Les architectures des lacs de données sont basées sur l'acquisition de données, sur lesquelles les traitements d'exploration et d'analyse (par exemple) vont s'appliquer. Il n'est pas envisagé, *a priori*, que les traitements des lacs de données se déplacent vers la donnée et que les données ne migrent pas, au sens technique du terme, vers la plateforme du lac de données.

Notre hypothèse est que lorsque l'on prend en compte la gravité des données, le traitement des données du lac peut être amené à se déplacer là où résident les données et non pas le contraire. En architecture, les paramètres qui composent cette gravité, tels que le volume (ou la masse) et la sensibilité, sont considérés comme des éléments de contraintes non fonctionnelles. Cela nous amène à considérer la gravité des données comme étant une contrainte non fonctionnelle[4] à l'évaluer lors de la conception des lacs de données. Par habitude de conception, fortement liée à des contraintes technologiques dans les systèmes d'information classiques, les données sont toujours déplacées vers les traitements qui les utilisent, ceci afin de protéger notamment les performances des systèmes opérationnels les émettant. La vision précédente de la gravité des données peut remettre en cause ce postulat. Sur cette voie, nous nous sommes donc attachés à définir quels paramètres non fonctionnels sont pertinents dans les lacs de données relativement au problème de la corrélation entre gravité et transferts données-traitements. Trois ont retenu notre attention : le volume (ou la masse), le coût et la sensibilité.

La ***masse*** des données disponibles devient de plus en plus importante, et une solution basique comme augmenter simplement la capacité de stockage ne suffit plus à répondre à cette problématique. Le ***coût***, lié la plupart du temps aux problématiques de réplication, d'acquisition, de sécurité mais aussi d'extension de capacité de stockage, doit être désormais évalué lors de la conception des architectures des lacs de données. La ***sensibilité*** des données entraîne une gestion spécifique. S'il n'existe pas de définition légale pour les données dites sensibles, les nouvelles règlementations sur la donnée personnelle RGPD[5] par exemple ou bien la cybersécurité et la Loi de Programmation Militaire (LPM) étendent celle déjà délivrée par la CNIL[6]. Chaque organisation peut, en plus de ces obligations réglementaires définir sa propre classification de données dites sensibles. Afin d'englober toutes ces notions, nous

[4] On appelle contrainte non fonctionnelle, les contraintes auxquelles sont soumises les architectures pour délivrer un fonctionnement correct, telles que la performance, le volume, la sécurité, l'évolution d'échelle, la disponibilité, la fiabilité, etc.

[5] Règlement General de la Protection des Données, une nouvelle réglementation européenne qui entrera en vigueur le 25 mai 2018

[6] http://www.cil.cnrs.fr/CIL/spip.php?rubrique300, Les données sensibles sont celles qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou sont relatives à la santé ou à la vie sexuelle de celles-ci.

définirons qu'une donnée est dite sensible au regard du degré de sécurité criticité/précaution que nécessitent son utilisation et son traitement. Cela va donc dépendre de l'évaluation par le propriétaire de ces données.

Ces trois paramètres constituent des contraintes qui peuvent avoir des impacts très forts sur la conception des composants des systèmes d'information et peuvent réduire les choix d'architecture possibles voire remettre en cause un choix existant.

Nous proposons d'enrichir la vision des travaux de recherche précédents et d'affiner le concept de gravité de la donnée en le fondant sur les trois paramètres volume/masse, coût et sensibilité.

3. Impact de la gravité de la donnée sur les architectures des lacs de données

Dans le cadre des lacs de données, il semble que le postulat de déplacement de la donnée vers son traitement (stipulé dans la section 2.1) ait été adopté comme une pratique par défaut, ce qui se traduit par à une duplication, systématique, de toutes les données que l'on veut analyser et explorer. Or nous pensons que, dans ce postulat, la notion de gravité des données n'est pas étudiée, lors de la définition de l'architecture d'un lac de données. L'approche de duplication systématique de toutes les sources de données ne doit pas être l'approche de référence. Dans cet article, nous souhaitons étayer ce point et démontrer que la gravité des données peut jouer un rôle important dans ces architectures et doit être évaluée dès la conception.

3.1 L'impact du volume sur les lacs de données

L'augmentation du volume des données produites et donc de leur masse est l'un des paramètres de la gravité de la donnée. Si cette masse devient trop importante, d'après (McCrory, 2010), la gravité de la donnée va être telle que le traitement des données va être attiré vers elles et donc va donc devoir être déplacé. Le volume est intégré et pris en compte au niveau de la gouvernance du lac de données (cycle de vie des données) et doit donc être évalué finement lors de la conception de l'architecture fonctionnelle du lac de donnée. L'évaluation doit prendre en compte non seulement le volume des données intégrées dans le lac mais aussi prévoir une augmentation ultérieure de celui-ci. En effet, le lac de données a pour vocation de stocker les données le plus brutes possibles mais aussi, en fonctionnement courant, des données préparées, agrégées, archivées, etc. Ces différents états des données vont eux aussi influencer le volume du lac de données, et donc la masse. Les lacs de données doivent dès la conception intégrer cette notion fondamentale de cycle de vie des données qui est l'une des principales fonctionnalités de la gouvernance des données. Il peut être décidé que le volume de données généré va être tel qu'elles ne peuvent pas être déplacées physiquement ou dupliquées mais être seulement accessibles.

3.2 L'impact de la sensibilité sur les lacs de données

La notion de donnée sensible est elle aussi une contrainte qu'il faut évaluer dans les lacs de données. En effet certaines données au regard de leur « sécurité », de leur « sensibilité » ne peuvent pas être dupliquées ou déplacées. L'anonymisation ou la pseudonymisation des données sont des techniques qui permettent de manipuler et déplacer ces données en respectant leur sensibilité. Cependant ces techniques peuvent faire perdre la valeur même des données qui ne sont alors plus exploitables. L'encryptage est aussi une technique pour permettre le déplacement des données sensibles. Il permet de sécuriser le déplacement par exemple mais la donnée devra être déplacée vers un système offrant une continuité de cet encryptage. Le niveau de sensibilité va lui aussi nécessiter un niveau de protection de la plateforme qui héberge ces données, de très haut niveau, impliquant un certain coût. Déplacer la donnée pour lui faire subir un traitement sur un autre environnement peut impliquer un risque élevé et donc bloquer le déplacement de la donnée. Cette problématique est présente dans le monde industriel soumis à d'importantes normes de conformité, directives et réglementation où la protection de la donnée est exigée. La sensibilité est donc un élément crucial dans le choix du transfert données-traitements. Il doit, au même titre que le volume être intégré par conception et par défaut dans les architectures des lacs de données.

3.3 L'impact du coût sur les lacs de données

La duplication des sources de données pose la problématique non seulement au niveau de la qualité mais aussi du coût de l'extraction multiple d'une même donnée et son impact sur le système où elle est émise ou déplacée. Au niveau de la gouvernance des données, multiplier les copies d'une même donnée peut entraîner une dégradation de sa valeur, engendrer des versions différentes difficiles à gérer, rendre complexe sa traçabilité et donc impacter la qualité globale du système. La mise à disposition de données ou sources de donnée a un coût sur le système émetteur ou hébergeur de cette donnée : au niveau de son extraction, du stockage même temporaire mais aussi au niveau des capacités physiques (mémoire, processeurs, etc.). La multiplication de sollicitations trop importantes peut être un frein à la mise à disposition de copie de données. Le volume de données à dupliquer et extraire, ainsi que la fréquence de ces copies peut aussi accentuer cet impact. Un autre effet de la duplication de la donnée est le coût associé à sa traçabilité. En effet pour répondre à certaines réglementations (précédemment citées), les données doivent être tracées, leurs accès et traitements subis conservés, en vue d'un audit par exemple. Cette traçabilité fait grossir les volumes des *logs* des différents serveurs, augmentant ainsi les coûts de traitement et de stockage. La duplication de données, un des principes utilisés des lacs de données, génère un coût qu'il convient d'évaluer finement dès la conception.

Nous avons donc établi que les trois paramètres volume-masse, sensibilité et coût du déplacement des données inclus dans la gravité des données peuvent remettre en cause la relation données-traitements au sein des lacs de données.

Si nous appliquons l'analogie de la gravité des données avec la vision physique, ce sont les traitements qui utilisent ces données qui seront attirés à elles et pourront donc être déportés à l'endroit où elles se situent et non pas le contraire. C'est donc le traitement qui va aller vers la donnée et non plus la donnée que l'on va déplacer vers le traitement. La gravité des données impacte donc l'architecture applicative des lacs de données mais aussi leur architecture technique. Il faut donc explorer les possibilités techniques d'amener les traitements où résident les données désormais et envisager des solutions alternatives à la structure physique unique comme réceptacle des lacs de données. La prochaine section étudie l'impact de la gravité des données sur les architectures applicatives et techniques des lacs de données au travers d'un cas réel.

4. Etude de cas : La gravité des données sur un lac de données industriel

4.1 Description de l'étude de cas industriel

L'étude de cas industriel dans le domaine financier est celui d'un lac de données dédié à la collecte de données provenant de tout un parc informatique composé de différents types de serveurs, réseaux et baies de stockage. L'objectif de ce lac est d'améliorer la connaissance de ce parc pour en améliorer e pilotage. Le lac de données est basé sur de la technologie *Apache Hadoop* et une suite d'outils d'aide à la manipulation, l'exploration et l'administration des données : *HortonWorks Data Plateforme* (HDP). Les données émises par les serveurs et autres sont poussées en temps réel dans le lac de données HDP et explorées par les utilisateurs du lac de données.

4.2 Evaluation de la gravité des données sur le lac de données métrologie

L'observation de chaque paramètre a été effectuée sur un mois afin d'intégrer les pics d'activités de l'industriel et être représentative.

4.2.1. Le volume

Nous avons recueilli le volume moyen émis par minute et par type de serveurs (mainframe, x86, Unix.) afin d'estimer le volume journalier que le lac de données aurait à intégrer. Pour chaque type de serveurs[7] (qui représentent nos sources de données dans le lac de donnée), nous pouvons calculer le volume journalier attendu dans le lac de données métrologie, soit :

Ej (serveurs) = Nb de serveurs x 24 (heures) x 60 (minutes) x **Ev** (Petabytes)

Où **Ej** est l'estimation moyenne journalière par type de serveur (Petabyte) et **Ev** est l'estimation volume / minute / serveur (Gigabyte).

[7] Serveurs de type x86 : 18000 ; Serveurs de type Unix : 30 ; Serveurs de type Mainframe : 6 ; Baies de stockage : 50 ; Réseaux : 3 types LAN, MAN, WAN.

Cela donne pour les sources de données de notre lac de données, un volume estimé de 330 Petabytes environ par jour pour tous les serveurs à intégrer. Dans le cadre de la gouvernance de la donnée, il a été décidé une conservation des historiques de données de 30 jours, ce qui implique que la conservation des données augmente le volume dans le lac de données, soit :

Ej x 30 = 9 900 Petabytes soit environ 1 Exabyte de données, dédié à l'historique.

En y ajoutant le 1 exabyte de données produites en 30 jours, nous avons donc un volume de 2 exabytes de données , au minima dans le lac de données.

Le Tableau 1 récapitule les données moyennes mesurées sur les serveurs existants.

Tableau 1. Volume par type de serveur

	Serveur X86	**Serveurs Unix**	**Serveurs Mainframe**	**Baies de stockage**	**Réseaux**
Ev	12	20	1	20	900
Ej	311	0,86	0,00864	14	3,89

Si les 2 exabytes de données représentent un volume important ce dernier ne peut être considéré comme une contrainte assez forte pour empêcher le déplacement des données. Cependant la gestion du cycle de vie de la donnée a été imposée seulement à un mois de conservation d'historique, ce qui explique ce chiffre de 2 exabytes. Ce choix ne nous semble pas réaliste et surtout ne tient pas compte des accroissements de volume généré par les analyses et explorations faites, ni des profondeurs d'historique nécessaires lors du travail en analyse prédictive, où souvent plusieurs mois ou années sont nécessaires. Une profondeur d'historique de seulement un an, entraînerait un volume d'au moins 24 Exabytes de données, ce qui pourrait alors modifier la vision de son déplacement.

A ce stade de l'étude, le volume seul n'est cependant pas jugé assez influant pour bloquer le déplacement des données mais nous émettons une alerte sur l'estimation qui en faite.

4.2.2. La sensibilité

La sensibilité des données a été considérée comme peu influente lors de la conception de l'architecture fonctionnelle. Or l'organisme financier est soumis à la Loi de Programmation Militaire (LPM) et certaines données transitant par ses réseaux doivent être protégées car jugées sensibles. Les données de métrologie provenant notamment des serveurs de type mainframe ont été classifiées hautement sensibles, car les applications critiques de l'industriel sont opérées via ces serveurs. De plus, le cas de la métrologie n'est pas représentatif, chez cet industriel, de la réelle évaluation de la sensibilité, pour les futurs autres lacs de données, notamment celui des données clients qui va être soumis à la règlementation européenne RGPD. Ce facteur n'a donc pas été évalué correctement lors de l'architecture fonctionnelle et peut remettre en

cause le déplacement de certaines données. Le Tableau 2 réévalue la sensibilité des données selon leur provenance.

Tableau 2. Evaluation de la sensibilité des données selon leur provenance

	Serveur X86	**Serveurs Unix**	**Serveurs Mainframe**	**Baies de stockage**	**Réseaux**
Sensibilité	2	6	10	8	9
Évaluation	Faible	Moyenne	Haute	Haute	Haute

Le paramètre de sensibilité doit donc être approfondi notamment pour les données provenant des serveurs mainframe, car ils contiennent les applications et données stratégiques de l'industriel. C'est pour cela que nous avons concentré l'étude de coût du déplacement des données de ce type de serveur (mainframe).

4.2.3. Le coût

Nous avons évalué le coût de déplacement de 1 TB de données par jour du serveur mainframe vers un autre serveur. Ce coût se mesure en « *million instructions per second* » (MIPS) qui est l'unité de facturation d'un serveur. Ce coût est composé des éléments suivants :

– L'utilisation de 4 cœurs de processeurs sur un mainframe de type z13 (taux de charge 85 %) : cela correspond en unité de mesure mainframe à 519 MIPS par jour ; le coût journalier est donc de 6756,4 $ (prix moyen observé pour 519 MIPS) ; sur une année, le coût est estimé à 2 466 103 $;

– à ce premier coût, il faut ajouter ceux d'administration et de maintenance du serveur ; une étude fait état d'un coût moyen de 98 482 $ par an.

La réplication de 1 TB de données revient donc à 2,55 M$ par an. Comme le volume estimé par jour de données pour la métrologie à répliquer est estimé à 8,6 TB (cf. Tableau 1), le **coût total sur une année représente plus de 22 M$.** *Ce calcul a été validé par une étude interne à IBM réalisée en laboratoire.*

Le déplacement des données des serveurs mainframe a donc un coût très important, dont le concepteur du lac de données n'a pas tenu compte lors de sa création.

4.3 Conclusion du cas d'étude de lac de données métrologie

Nous avons évalué au travers de trois paramètres (volume-masse, sensibilité et coût) l'impact potentiel de la gravité sur l'architecture du lac de données industriel. Si le volume n'a pas eu d'impact significatif, l'évaluation du coût et de la sensibilité sur certains serveurs (les mainframes) impose que la relation données-traitement soit

revue. Un mode d'accès en fédération et non en réplication doit être mis en place pour les données provenant de ce type de serveurs.

5. Conclusions et perspectives

Le principal objectif poursuivi dans cet article est celui de répondre à la question : qu'est-ce qui peut remettre en cause le choix d'une architecture fédératrice mono-technologique des lacs de données ?

Une partie de la réponse peut être faite en prenant en compte de façon systématique la gravité des données et en évaluant les éléments que sont : le volume, la sensibilité et le coût de déplacement des données vers le lac de données.

Cela ouvre la porte à des solutions alternatives d'architectures de lac de données hybrides, composées de données dupliquées mais aussi de données seulement référencées, accédées en mode *fédération* et dans lesquelles les zones de stockage des données sont elles aussi différentes en termes de d'architectures techniques.

Dans le cadre de notre étude nous avons évalué l'impact de la gravité des données sur un lac de données « *in situ* ». Une perspective à nos travaux de recherche est d'étudier l'impact de ce facteur dans la décision de positionner un lac de données « in situ » versus dans les « nuages ». Dans ce cas, la sensibilité des données personnelles en particulier va nécessiter d'aborder les aspects de confidentialité via à vis du prestataire mais aussi du fournisseur d'accès. Cela va poser des problématiques supplémentaires et générer un coût de gestion. Le transfert des données est une opération qui peut se révéler rapidement très couteuse comme on vient de la voir en section 4.2.3.

Bibliographie

Alrehamy H. et Walker C. (2015). Personal Data Lake With Data Gravity Pull. Proceedings 2015 Ieee Fifth International Conference on Big Data and Cloud Computing Bdcloud 2015, p. 160-167.

Gartner. (2015, September 15). Gartner Says Business Intelligence and Analytics Leaders Must Focus on Mindsets and Culture to Kick Start Advanced Analytics. from http://www.gartner.com/newsroom/id/3130017

Hortonworks. (2017). Hortonworks. from https://fr.hortonworks.com

Lenovo. (2016). Lenovo Big Data Reference Architecture for Hortonworks Data Platform. 4. https://cloud.kapostcontent.net/pub/9b91ad01-2f63-4c7b-ac2d-c0b5bb2af9e5/lenovo-big-data-ra-for-hortonworks-data-platform-1.pdf ? ?kui=dk4jpyPfd3pe6YP6Adkgfg

Madera C. et Laurent A. (2016). The Next Information Architecture Evolution: The Data Lake Wave. Proceedings of the 8th International Conference on Management of Digital Ecosystems (Medes 2016), p. 174-180. doi: 10.1145/3012071.3012077

MarketsAndMarkets. (2016). Data Lakes Market worth 8.81 Billion USD by 2021. from http://www.marketsandmarkets.com/PressReleases/data-lakes.asp

(McCrory D. (2010, December 07). Data Gravity – in the Clouds. from https://blog.mccrory.me/2010/12/07/data-gravity-in-the-clouds/

McCrory D. (2014, March 1). Data Gravity. from https://datagravity.org/

Russom P. (2017). Best Practices Report | Data Lakes: Purposes, Practices, Patterns, and Platforms. March 29, 2017.

Servigne S. (2010). Conception, architecture et urbanisation des systèmes d'information. Encyclopædia Universalis, p. 1-15.

IT Glossary Gartner. (2014) . https://www.gartner.com/it-glossary/data-lake