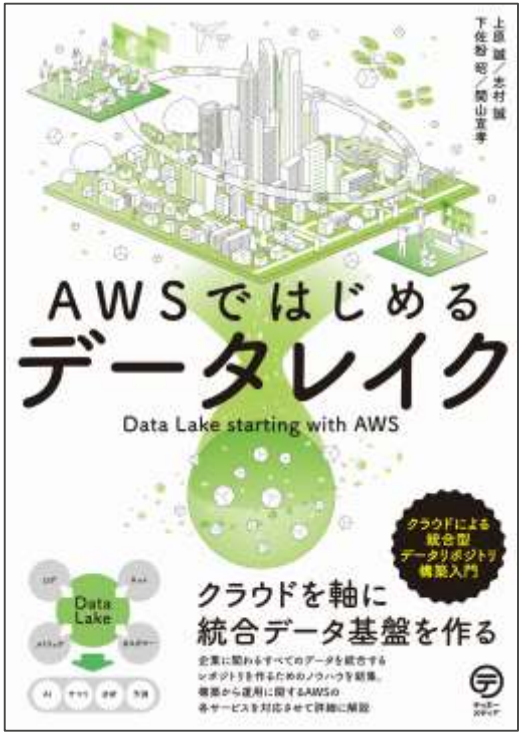
「AWSではじめるデータレイク」出版記念

データレイクはじめの一歩

2020年5月28日
アマゾン ウェブ サービス ジャパン 株式会社
シニアソリューションアーキテクト

下佐粉 昭（しもさこ あきら）

@simosako



2020年6月発売予定！

AWSオンラインセミナーへようこそ

ご質問を受け付けております！

- 書き込んだ質問は主催者にしか見えません
- 最後のQ&A時間で、いただいたご質問からピックアップしてご回答をさせていただきます

① 吹き出しをクリック
② 質問を入力
③ Sendをクリック

終了後にアンケートの記入をお願いいたします

https://bit.ly/2TFPbps

アンケートにお答えいただいた方には本日の資料を後日ご提供させていただきます。

自己紹介

下佐粉 昭（しもさこ あきら）

@simosako

所属：

アマゾン ウェブ サービス ジャパン
シニアソリューションアーキテクト

好きなAWSサービス : QuickSight, Redshift, S3 ...
人間が運用等から解放されて楽になるサービスが好きです

内容

- データレイクはじめの一歩 (40分)
  - データレイクってなに？ RDBやデータウェアハウスとの違い
  - 事例で見るデータレイク on AWS
  - AWSのデータレイク関連サービス

- Q&A (20分)

データレイクってなに？

データレイクはなぜ必要になったのか？

広く普及しているデータ分析用の環境といえば…

- データベース（RDBMS）
- データウェアハウス

これらとデータレイクは何が違うのでしょうか？

データベースとデータウェアハウス①

企業内に多数存在するRDBMS上のデータを統合して分析したいというニーズ
しかし１つのRDBMSでは、企業全体の分析用データを処理しきれません

RDBMS
（例：Amazon RDS）

	汎用RDBMS
操作	SQL標準 & JDBC/ODBC
目的・用途	汎用的に利用可能
同時接続数	大規模（～数万接続） （リードレプリカ等）
データサイズ	数Gバイト～数TB
性能	ランダムIO、シリアルIO 読み書きともにバランス 膨大なデータ集計は不得手

企業に複数あるシステムから
データを統合するには**サイズ、**
性能面で得意領域を外れる

データベースとデータウェアハウス②

データウェアハウスは、用途に特化した作りにすることで
より大規模な演算を実現するための仕組みを持つ

RDBMS
（例：Amazon RDS）

データウェアハウス
（例：Amazon Redshift）

	汎用RDBMS	データウェアハウス
操作	SQL標準 & JDBC/ODBC	SQL標準 & JDBC/ODBC
目的・用途	汎用的に利用可能	大規模演算に特化
同時接続数	大規模（～数万接続） （リードレプリカ等）	限定的
データサイズ	数Gバイト～数TB	数TB～ペタバイト
性能	ランダムIO、シリアルIO 読み書きともにバランス 膨大なデータ集計は不得手	ある程度大きい量の読み取り （シリアルIO）に特化 粒度の細かい更新は不得手

データウェアハウス用途に特化することの意味

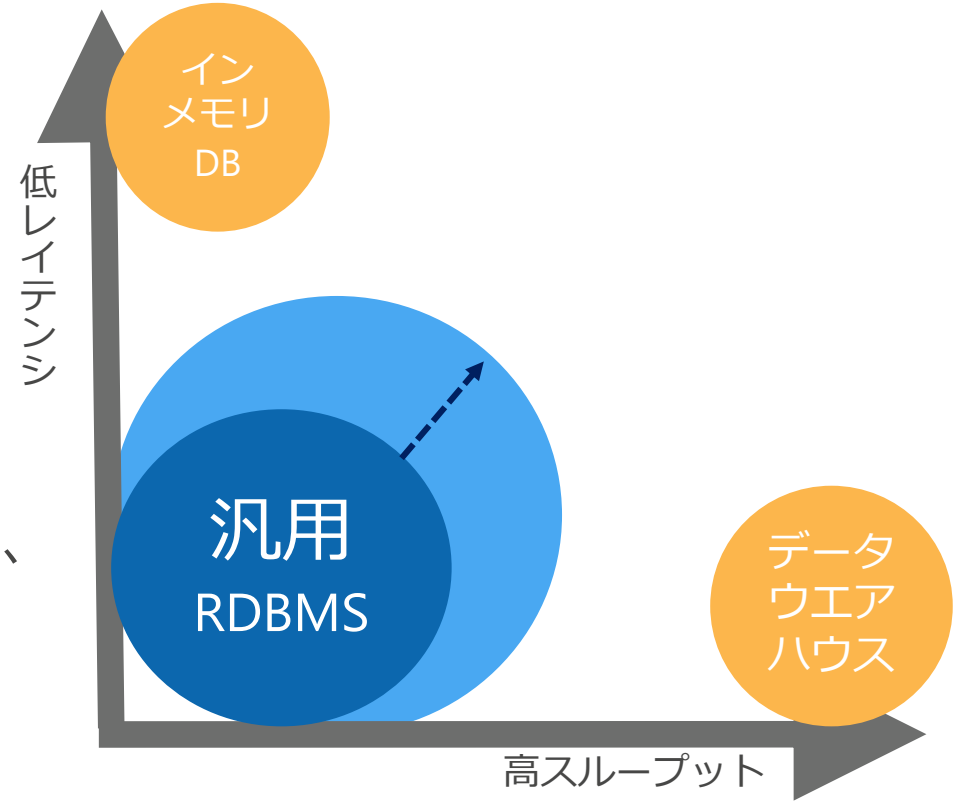
スループットの最大化

シリアルのIO読み取り性能を重視

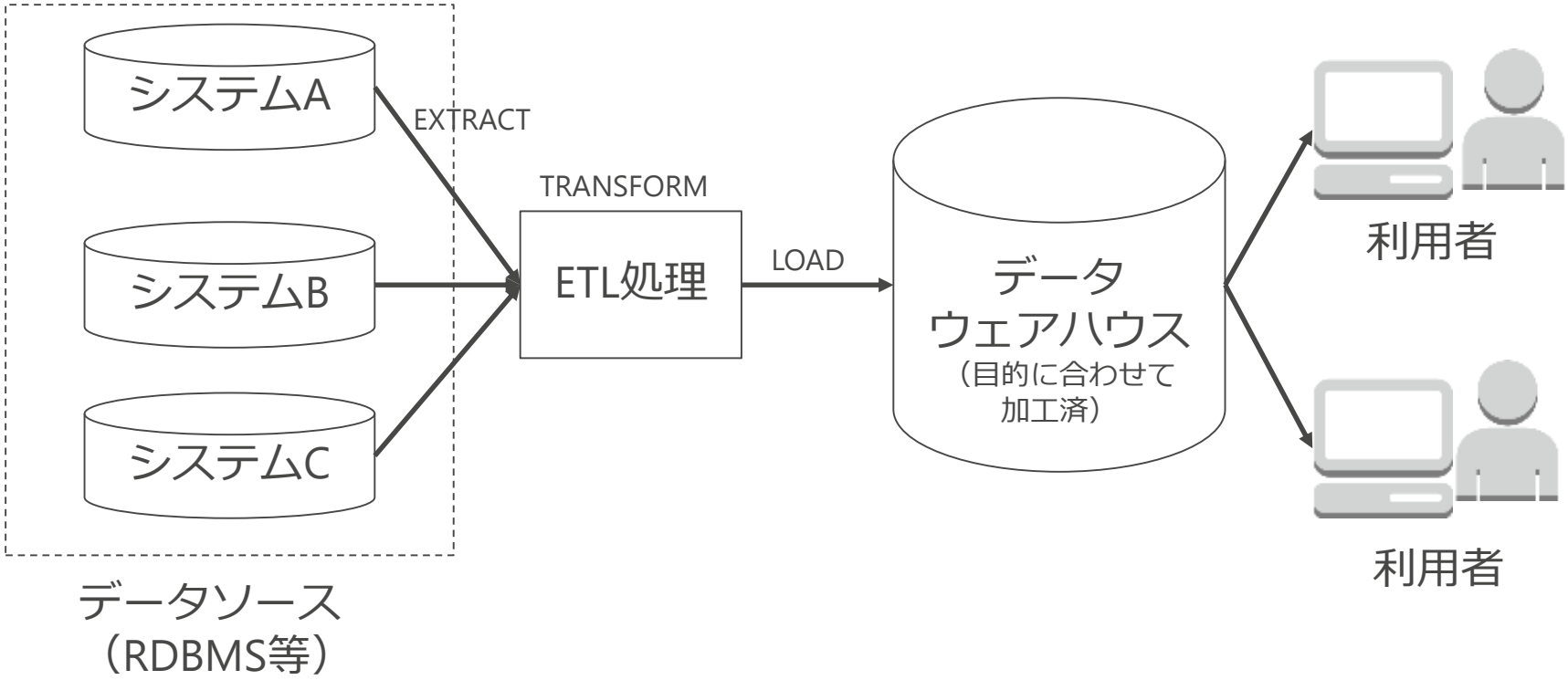
特化したアーキテクチャ

他の性能は「妥当な範囲で」維持する

汎用RDBMSも進歩し続けており、
利用範囲が拡大している

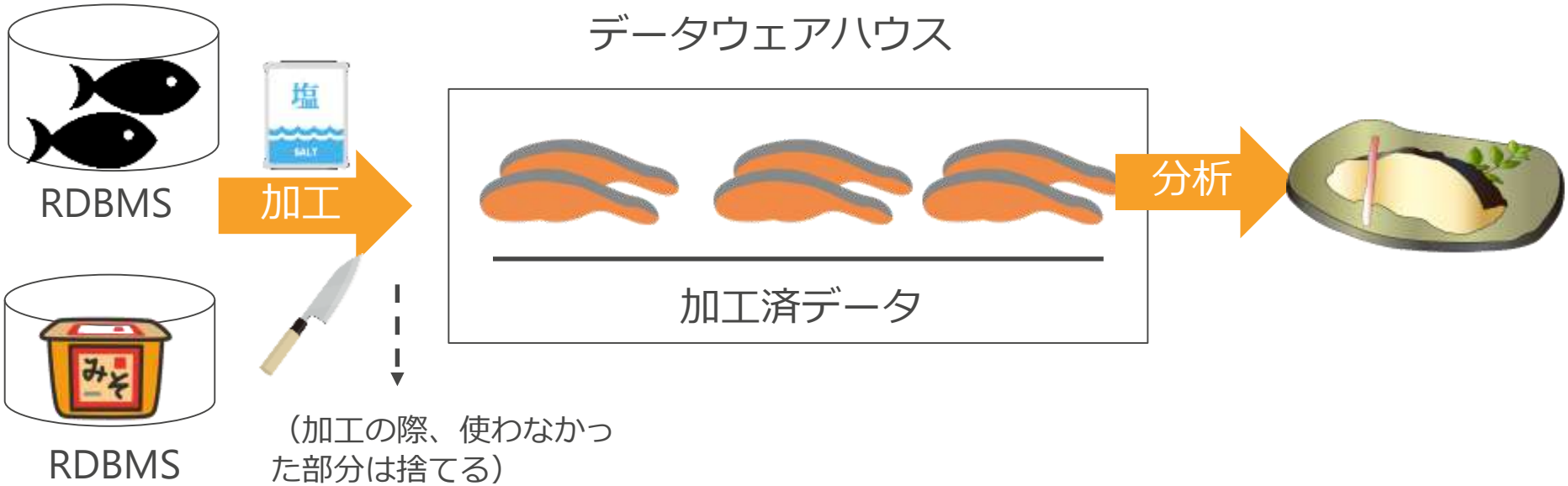


データソース（RDBMS）からETL処理を経て データウェアハウスを構築する構成の確立



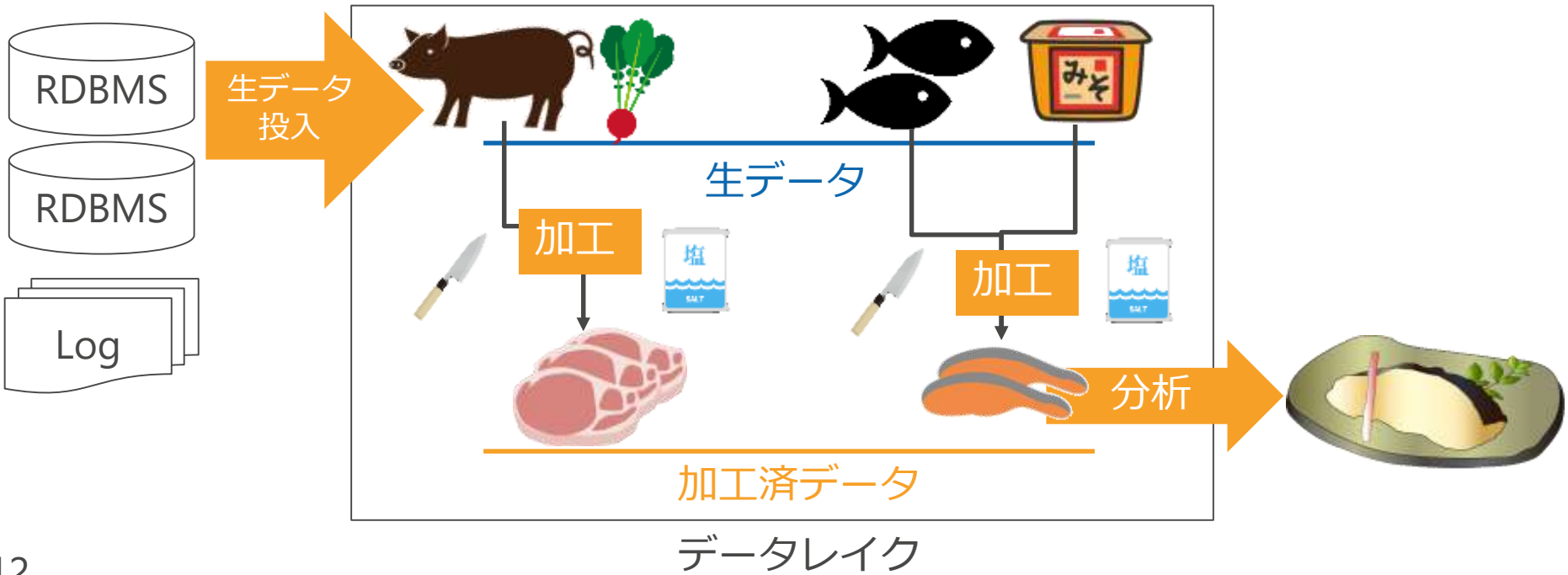
データウェアハウス構成の課題を料理に例えると

データウェアハウスは「目的に合わせた加工済」データを保存
例）西京焼きのため、魚をさばいて、味噌に漬け込んだ形で保存してある
＝＞後から刺し身を作りたくなっても、生魚は加工済…



データレイクによる解決

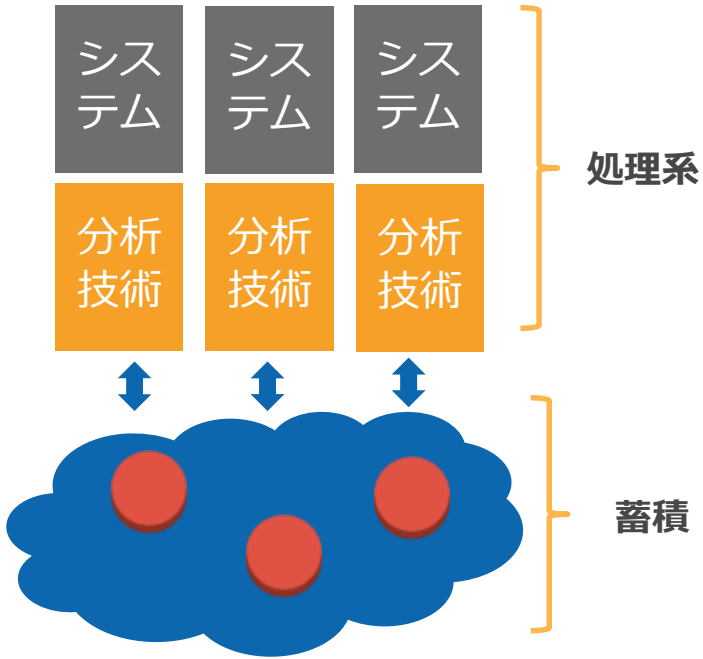
- データを「生」のまま保存し、将来のニーズに備える
- 加工したデータもデータレイクに保存
- 分析処理はデータウェアハウス等、データレイクの外で実現



データウェアハウスを巨大化していくのではなく、

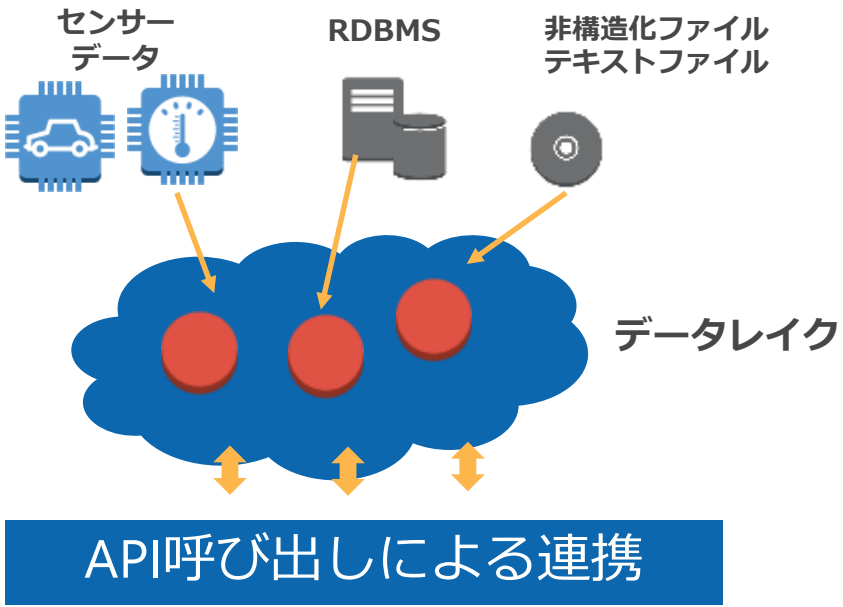
データ蓄積と処理系を分離することのメリット

- 処理と蓄積を分離することで、新しいニーズや技術に対応しやすい環境に
- 蓄積はデータレイクに集中させ、多様なデータを保存
- AWSでは複数の分析環境の構築・運用が容易なため、いつでも切り替え可能



データレイク（蓄積）

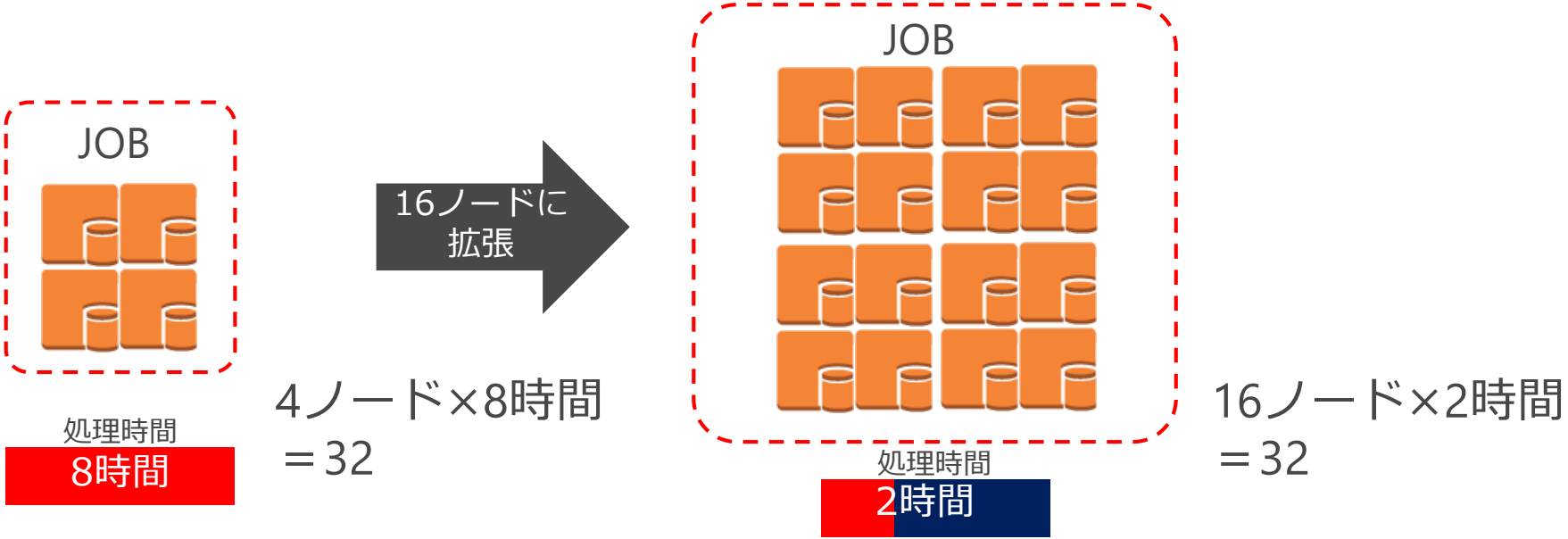
- 唯一「真」のデータ置き場
  - Single Source Of Truth
- データを失わない
- サイズ制限からの開放
- 決められた方法（API）ですぐにアクセスできる



処理系ではスケールアウトを重視する

クラウドではスケールアウトがコスト・時間の両面で効率的

- 必要な時に必要なだけ処理ノードを追加できる
- ノードを増やしても利用時間が短くなればコストは同じ*



*サービスにより料金体系や提供形態が異なるため、全てのサービスでこの考え方が適用できるとは限りません

データレイクを中心とした分析環境 on クラウド

- 生データをデータレイクに集め、将来的なニーズに対応
- 分析や可視化といった活用部分は取り替え可能な構成
- 処理系はスケールアウト可能な技術を活用

データレイクは蓄積だけで機能するか？

狭義のデータレイクはデータの保存場所

…しかし、それだけではデータレイクは使いづらいものに

1）どこに、どういうデータが置いてあるのかの管理

- データレイク内のカタログ（メタデータ管理）が必要
- データソースをから新しいデータや変更を発見・記録

2）生データは、そのままでは分析に向かない

- 分析可能な形への整形、最適化

データレイク側で行う整形

目的に合わせた変形は活用層で自由に実施可能
…しかし統一された形に整形されていないと分析はできない

ビジネス的な最適化
- 文字コードの変換：UTF-8等に統合を検討
- IDの統一：システムで異なるIDを発行している場合等の対応
- 日付、表現の統一：日付の表現方式、タイムゾーン、大文字小文字 等

パフォーマンス面の最適化
- 圧縮：読み取りデータサイズの縮小
- パーティショニング:月・日等単位で分割。アクセスに必要な範囲を限定
- フォーマット変換 : Parquet等へ変形し、列単位でのアクセスの最適化

データレイク内のデータ保存戦略

オリジナルデータは全て残す

　= 加工しても消さず、加工結果は別に保存

例）三段階に構成

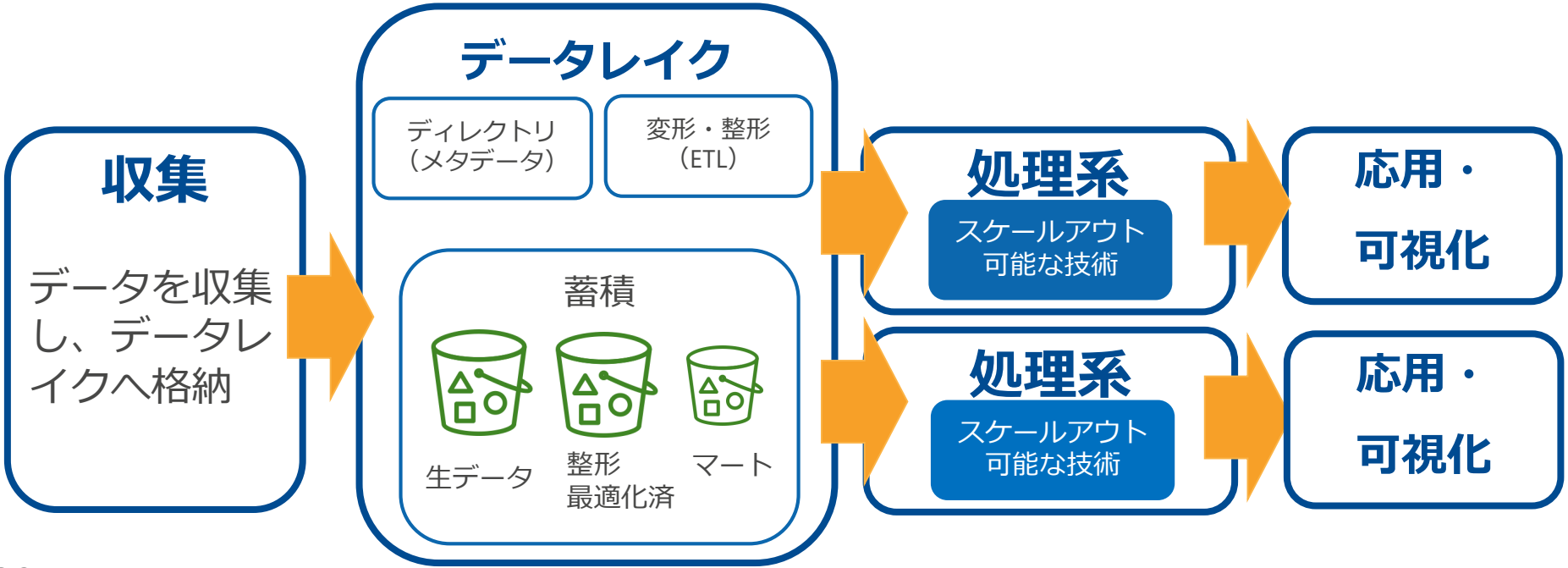
オリジナルデータ
（生データそのまま）

整形＋クエリ最適化

特定分析用データ
（データマート等）

データレイクを中心とした分析環境

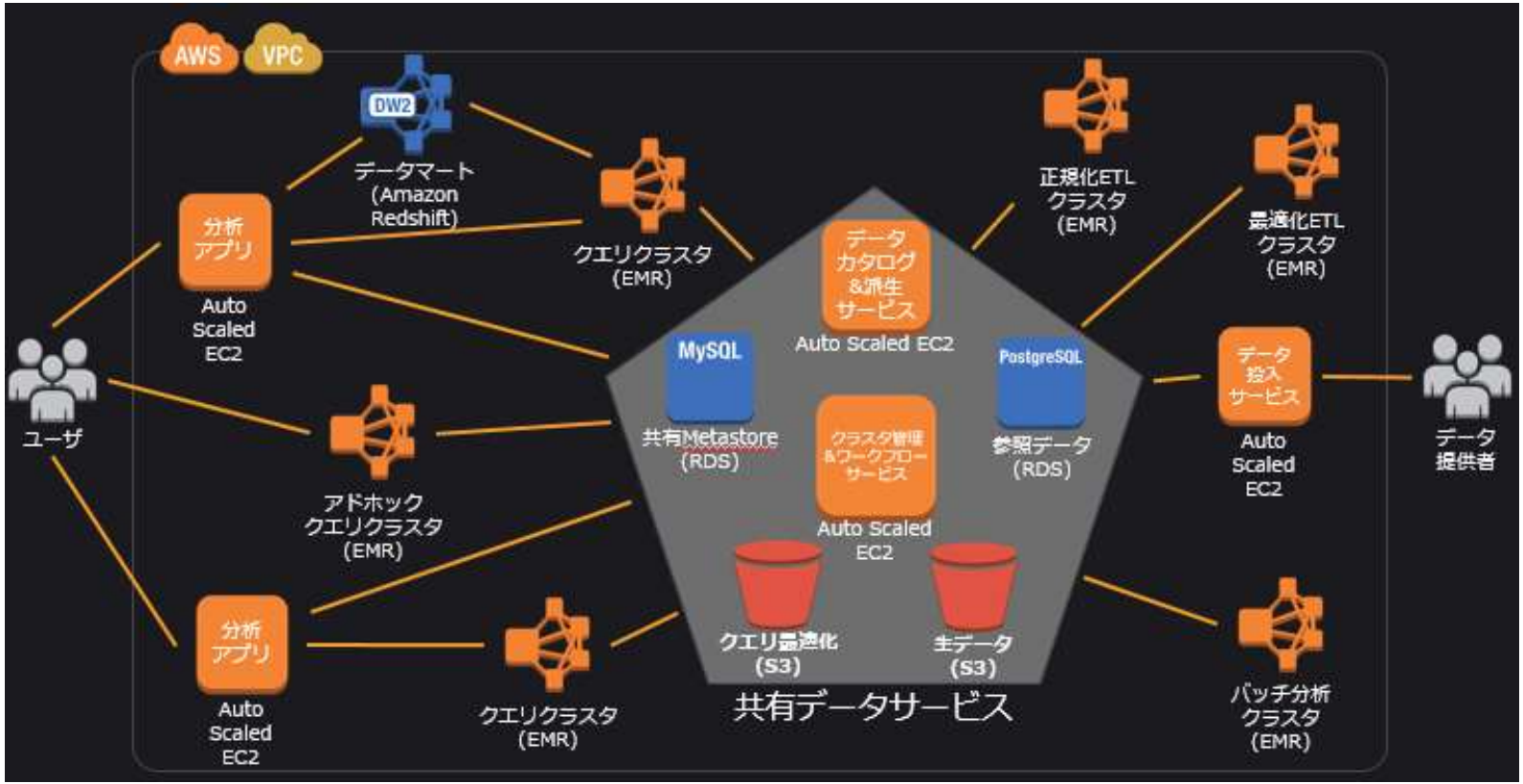
データレイク側で蓄積、メタデータ管理、整形を持つ
変形・整形の機能も「取り替え可能」な構成を維持



事例で見るデータレイク on AWS

事例：Finra様 750億イベント/日の処理基盤

- S3をデータ共有サービスとして定義し、EMRやRedshiftからアクセス



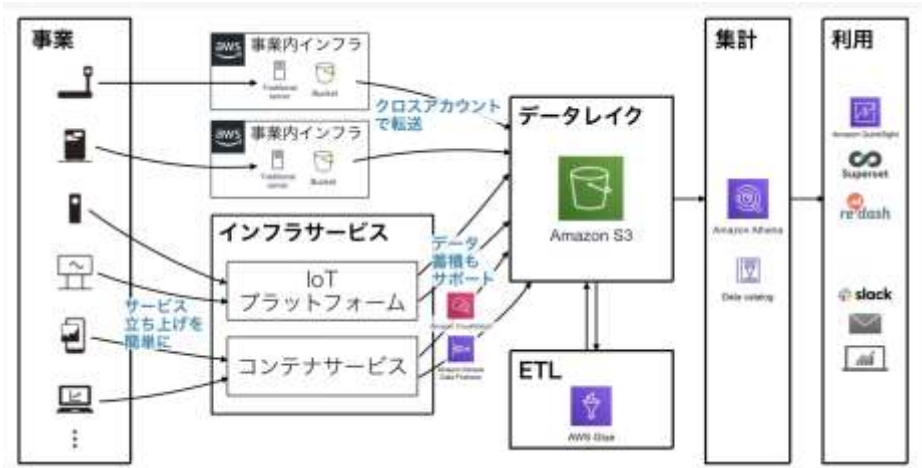
re:Invent 2015発表資料 BDT305 「Amazon EMR Deep Dive & Best Practices」より引用

事例：リコー様 共通データ基盤の構築とデータ活用の促進

状況
- 2015年頃から AWS の活用が本格化
- データ活用の環境が各事業部で個別に拡大

課題
- 事業を跨いだデータ活用が出来ない
- 何のデータがあるか分からず活用が進まない

解決法
- データを集め、活用するための仕組みを提供
  - データを貯めたらすぐ可視化する仕組み
  - データを簡単に貯められる仕組み
- データを民主化するための活動を実施

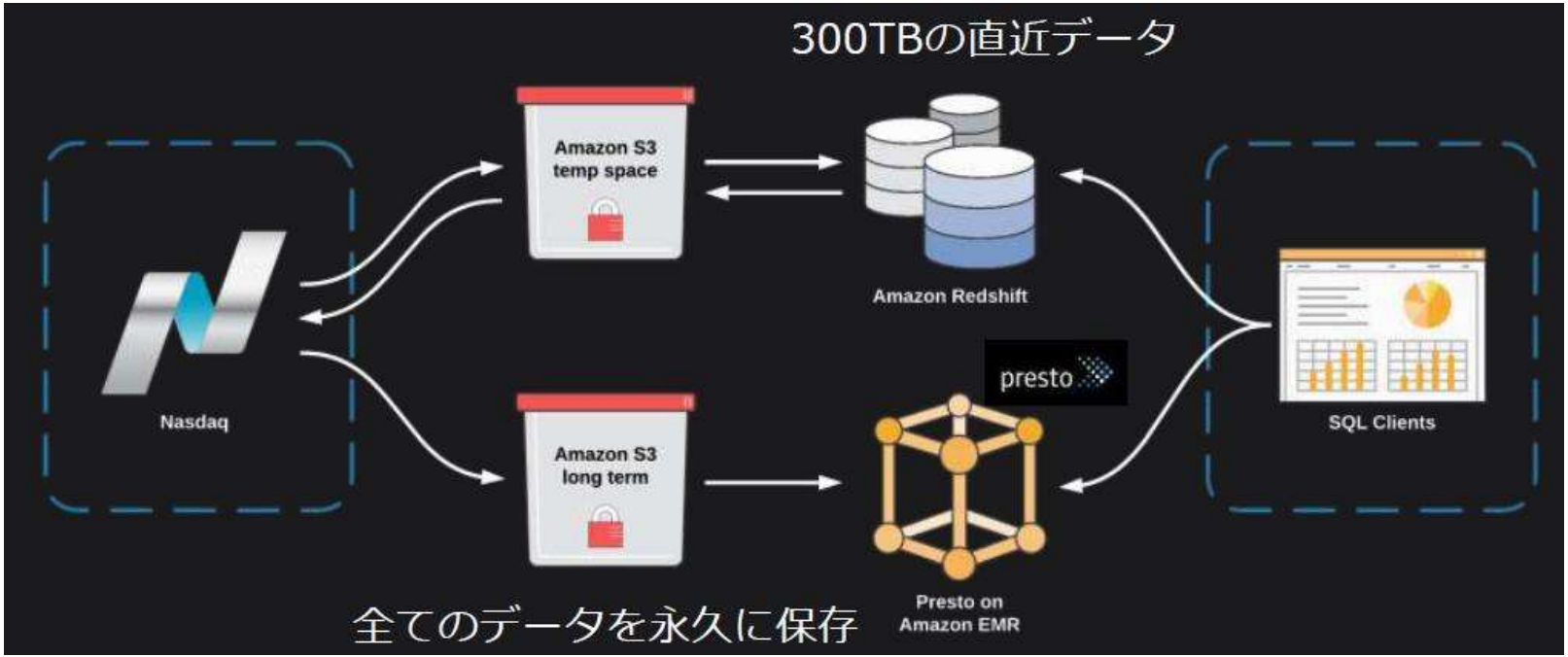


https://pages.awscloud.com/rs/112-TZM-766/images/L3-02.pdf

事例：Nasdaq様 Redshift/EMRを使い分け

- Redshift:300TB分の直近データ
- EMR+Presto+S3:全期間データ

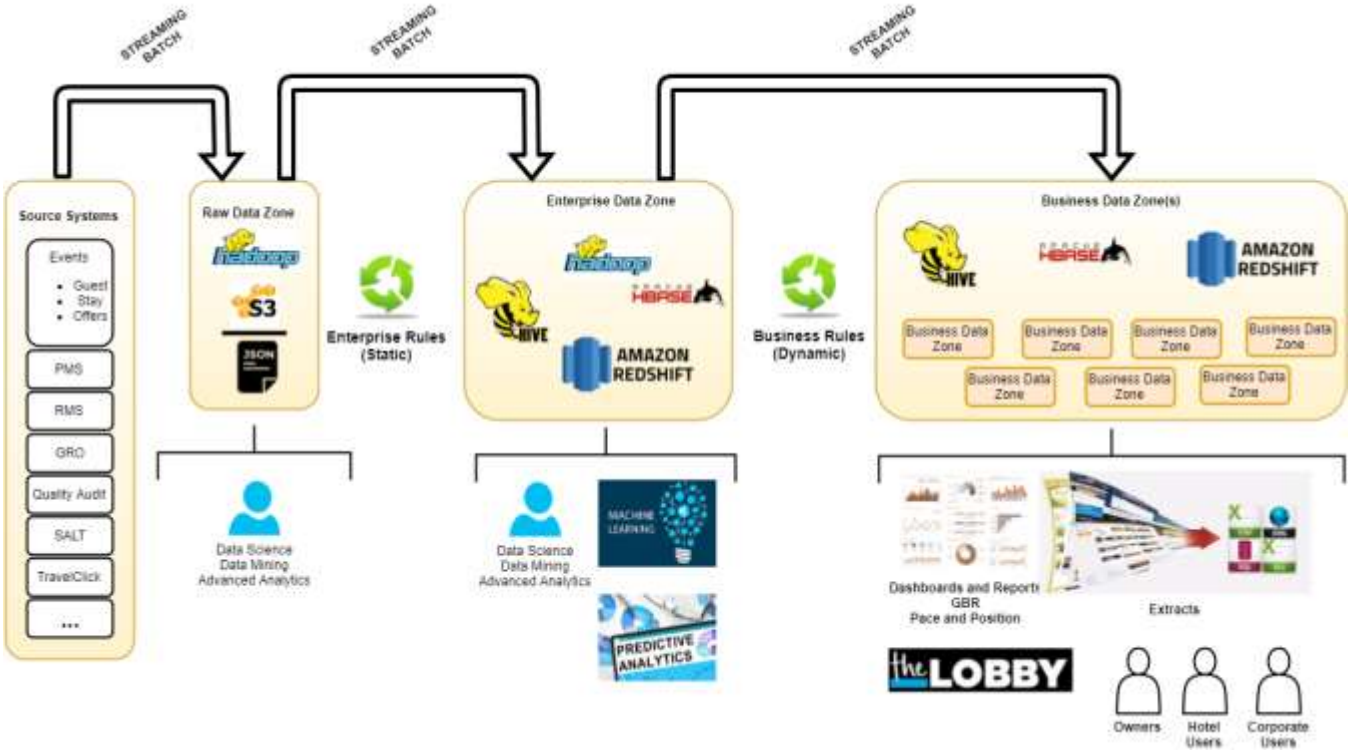
} 共通のSQLでアクセス



re:Invent 2015発表資料 BDT314 「A Big Data & Analytics App on Amazon EMR & Amazon Redshift 」より引用

事例：ヒルトン様 データレイクを核にEIM基盤

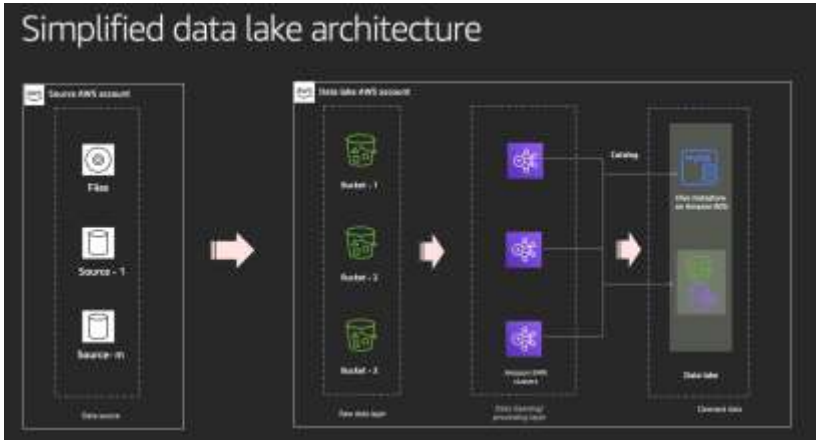
3ゾーン構成のデータレイクでEIM (Enterprise Information Management) 基盤を実現

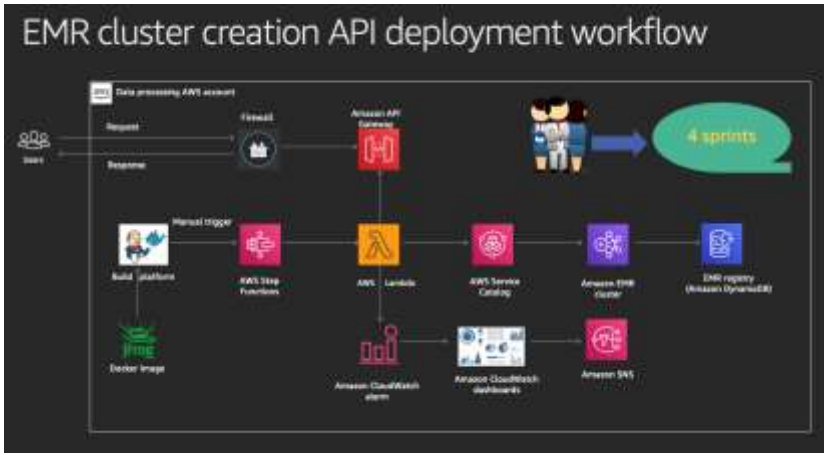


https://www.slideshare.net/AmazonWebServices/ask-an-amazon-redshift-customer-anything-ant389-aws-reinvent-2018

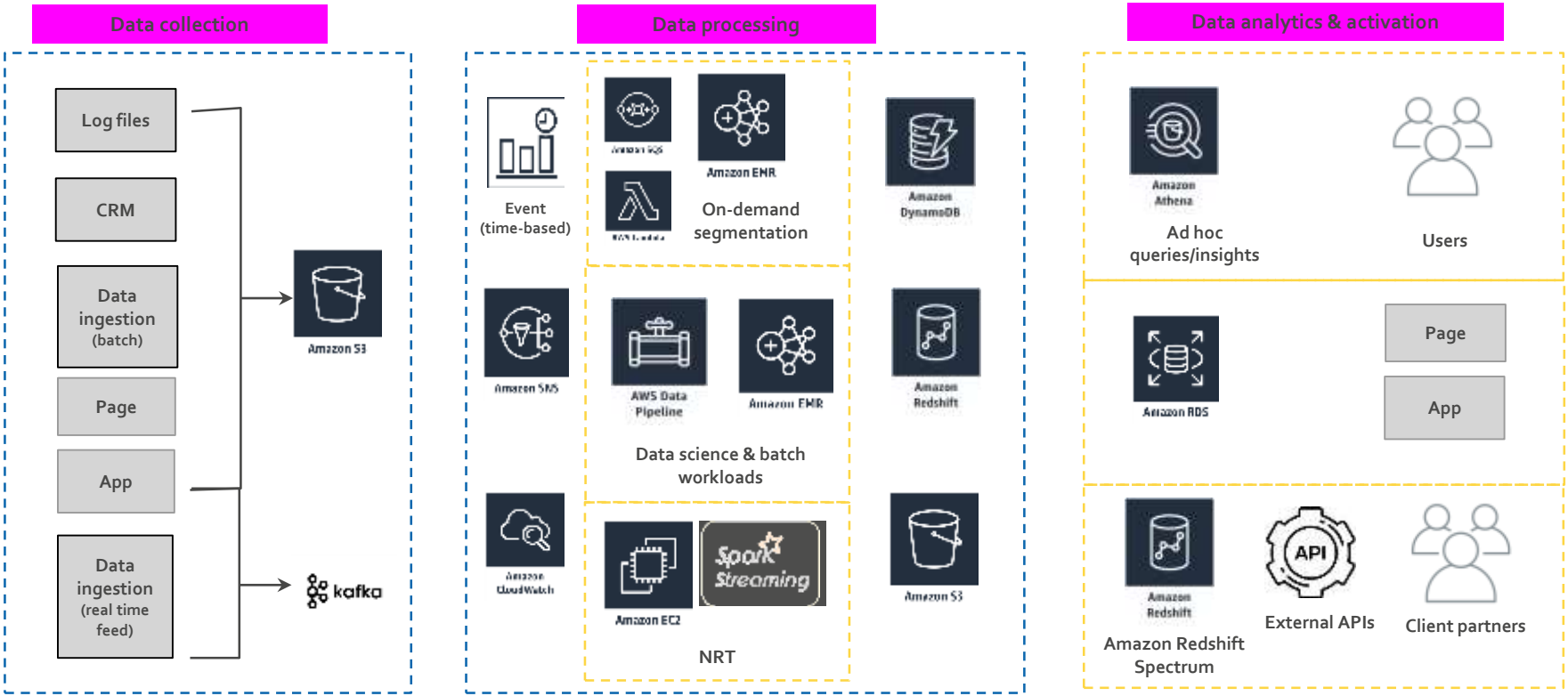
事例: Intuit様 オンプレミスからの移行

- Hadoop処理系が中心の大規模データレイク
- オンプレミスでデータレイクを構築した後に、AWSへの移行を決断
- 理由：大規模になるにつれSLA担保が困難、24x7サポートの負荷、アップデートの負担…
- AWS化で運用管理負担の低減、より進んだ自動化を実現





事例：セールスフォース・ドットコム様 – データレイクDMP
ラムダアーキテクチャを採用：リアルタイム、バッチ双方の処理を実現



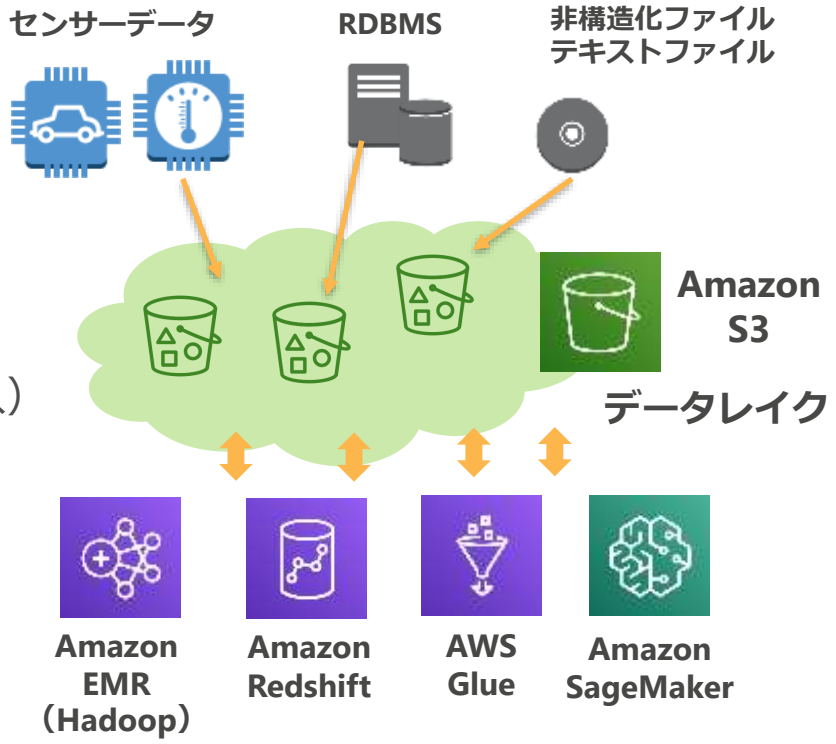
AWSのデータレイク周辺サービス

データレイクのコア = Amazon S3

S3によるデータレイク（蓄積）実現のメリット

- 上限無し：サイジング不要
- 高い耐久性：99.999999999%
- 安価：
  - $0.025/GB/月*（スタンダード）
  - $0.019/GB/月*（標準-低頻度アクセス）

  例）10TBの保存で約2.1万円/月**
- APIアクセス
  - 多様な言語にライブラリを提供
  - AWS各種サービスと連携



* 費用は2020年5月時点での東京リージョンでの価格です
** 1USドル = 110円で、標準-低頻度アクセスでの試算

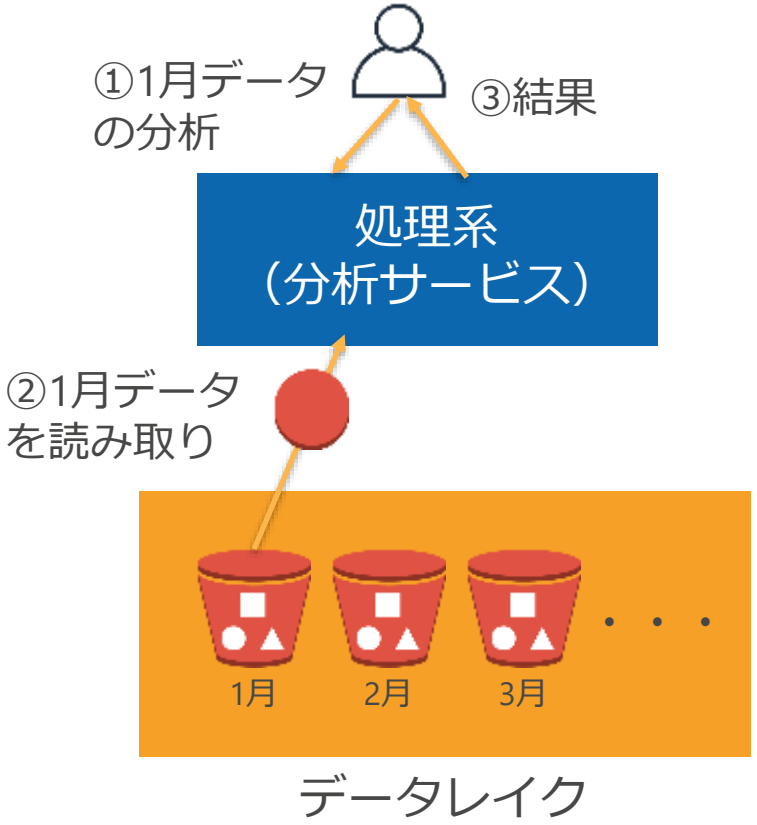
データレイクと処理系との連携方法①

リアルタイムに連携

- 処理系がS3上のファイルの必要な部分にリアルタイムにアクセスして演算

メリット

- リアルタイム性
- ロード時間不要



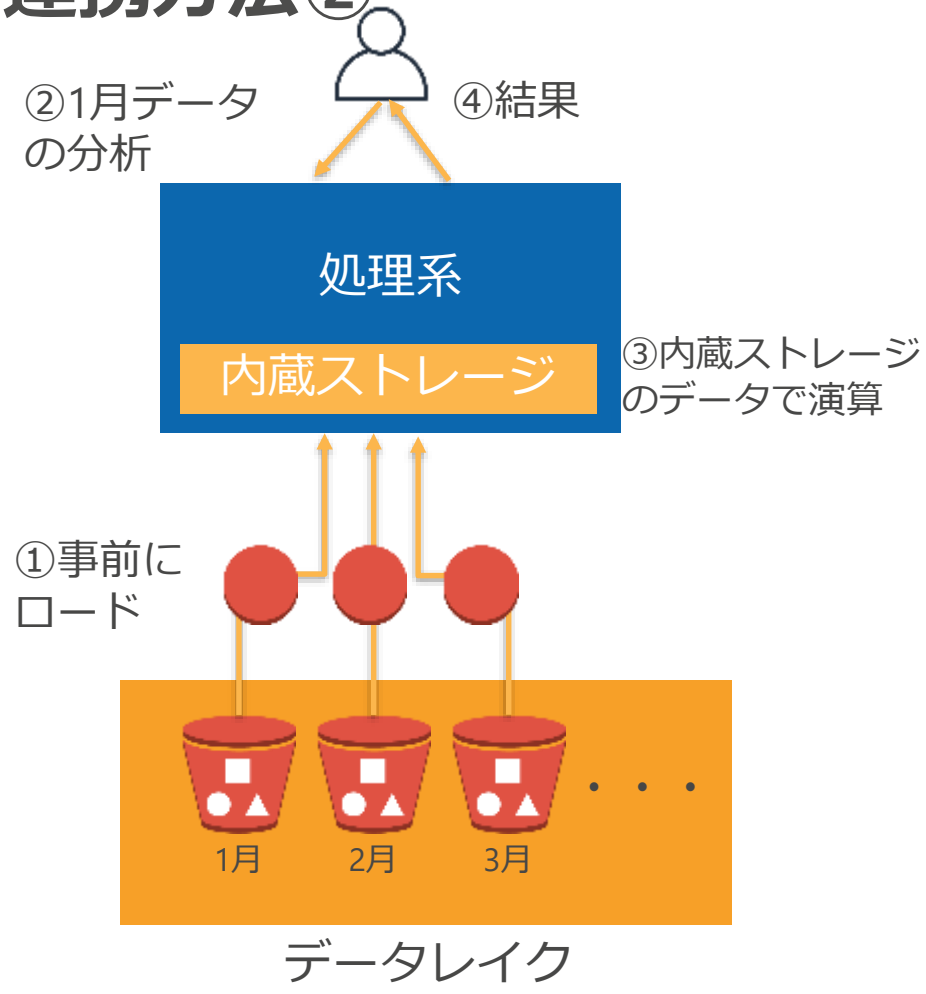
データレイクと処理系との連携方法②

ロード&エクスポート

- ファイルを内蔵ストレージに事前にロードして演算
- エクスポートで書き出し

メリット

- アクセスレイテンシの小ささ

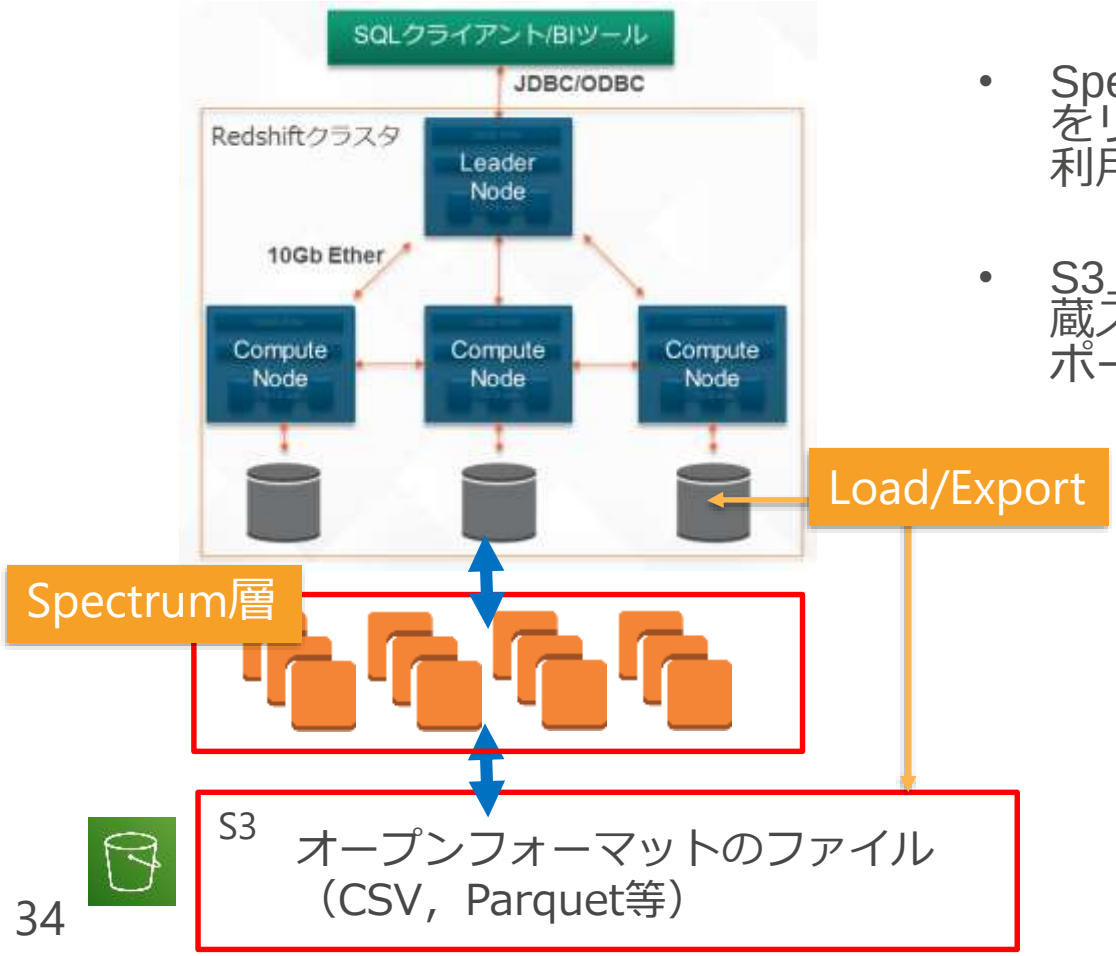


データレイクと連携するAWSの分析サービス（一部）

	Amazon Redshift	Amazon EMR	Amazon Athena
標準・デファクトスタンダード技術	SQL標準	Hadoop/Spark（デファクト） SQLアクセスも可能	SQL標準
分散処理・スケールアウト	ユーザ操作でスケールアウト	ユーザ操作でスケールアウト	自動的な分散処理（設定無し）
S3へ透過的アクセス	可能（Spectrum）	可能（EMRFS）	可能
S3とのロード/エクスポート	可能	可能 (EMRFS経由で読み取り、HDFSへ)	N/A



例：Amazon Redshiftの場合



- Spectrum層により、S3上のファイルをリアルタイム参照し、外部表として利用可能（透過的）
- S3上のデータの高速ロード、および内蔵ストレージ内のデータを高速エクスポート（エクスポート/ロード）

ETL&カタログサービス – AWS Glue

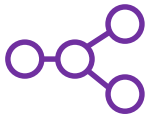
AWS Glue

サーバーレス

スケジューラーと
ワークフロー

コードに集中

データソースの
メタデータ管理



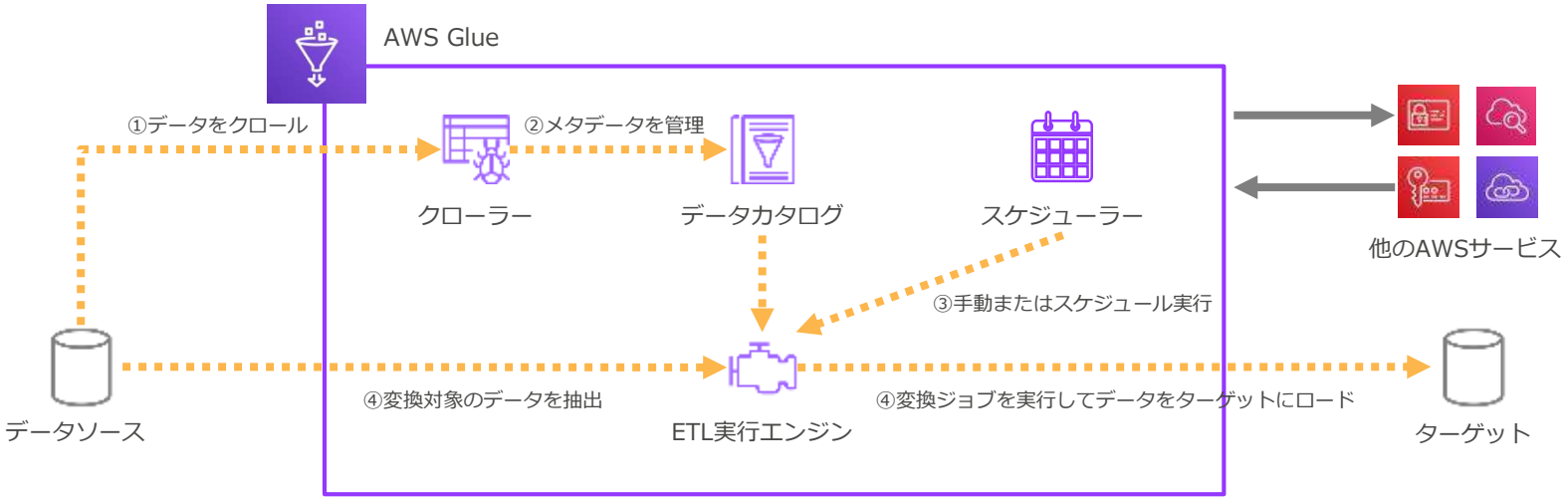
VPC内からのアクセス

他のAWSサービスと
容易に連携

セキュア

Notebookでの開発

AWS Glueの全体像



概要

①クローラーにてデータソースのメタデータをクロールして、データカタログに登録・更新

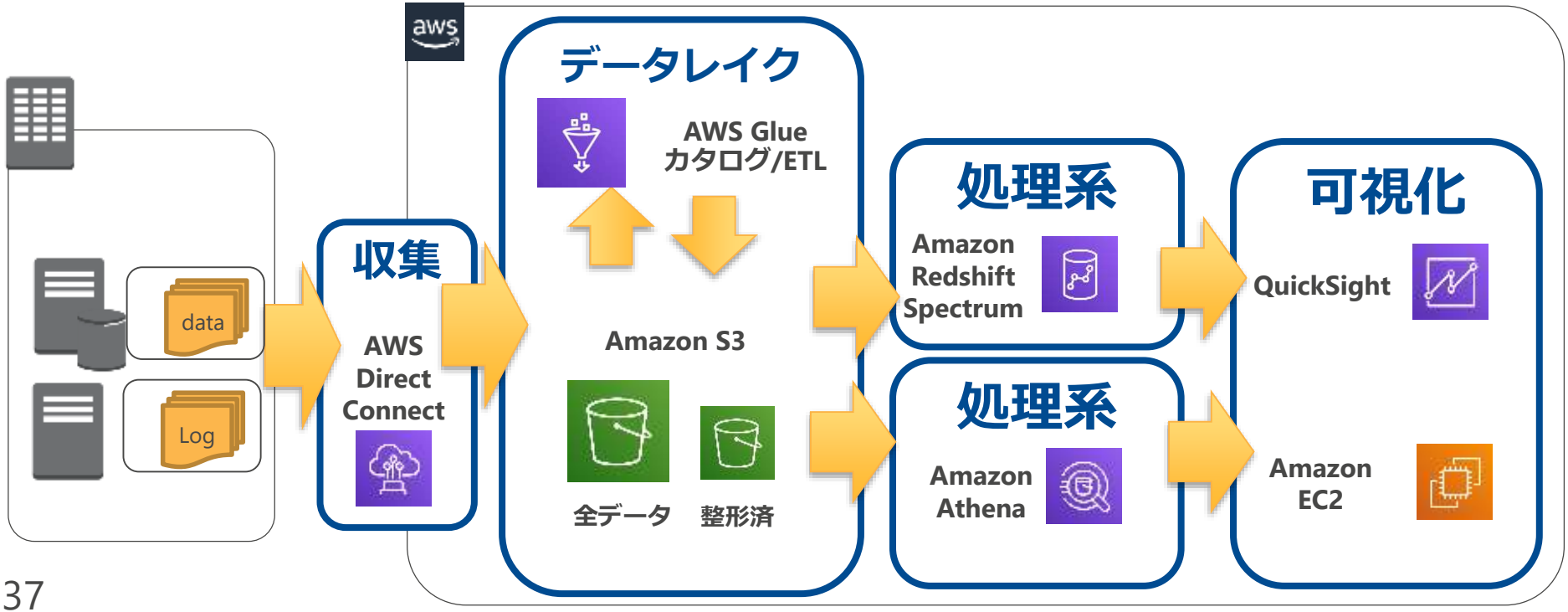
②データカタログにてメタデータを管理

③スケジューラーにてジョブの実行タイミングを定義

④データソースからデータを抽出し、ETL実行エンジンにてジョブをサーバーレスで実行
(ジョブはSpark(PySpark、Scala)またはPython Shellを選択)

全体図（組み合わせ例）

- S3に蓄積、Glueでディレクトリと整形
- RedshiftをDWHとし、SpectrumでS3データにも直アクセス
- アドホックなクエリはAthenaで対応



まとめ：変化を織り込んだビッグデータ処理基盤

データはオリジナルを残す（データレイク）

- 蓄積と処理系を分離し、将来に備える
- S3はデータレイク（蓄積）に最適

ニーズに応じて処理系・テクノロジーを選択

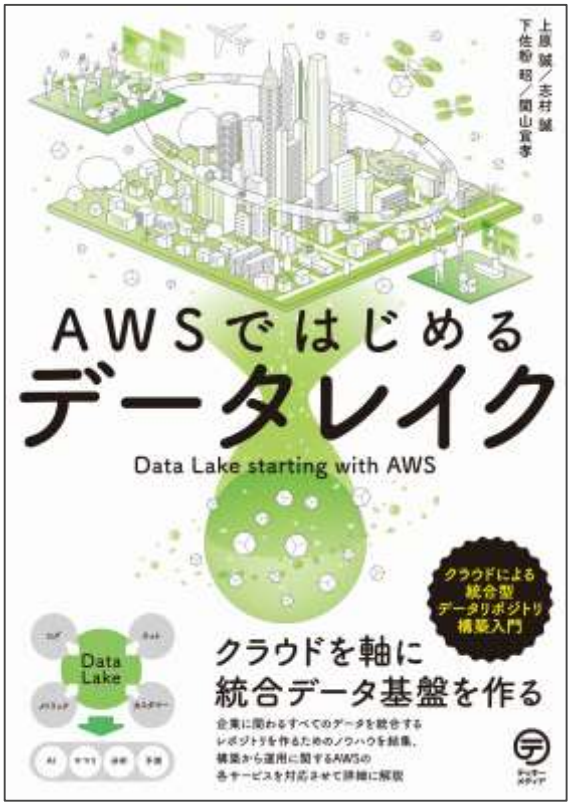
- スケールアウト可能なテクノロジーを選択
- AWSでは各種サービスがS3と連携可能

データレイクの書籍が発売になります！

2020年6月発売予定

http://techiemedia.co.jp/

データレイクの基本概念から
関連AWSサービスの説明、
構築方法まで



関連セミナーのご案内

- **【オンラインセミナー】第2回「AWSではじめるデータレイク」出版記念 - データレイクはじめの一歩　～データレイクの構築と蓄積されたデータの活用手法～**
  - データレイクのアーキテクチャとAWSサービス
  - 2020 年 6 月 4 日（木）　開演：11:00 ～ 12:00
  - https://aws-seminar.smktg.jp/public/seminar/view/2608

- **【オンラインセミナー】第3回「AWSではじめるデータレイク」出版記念 - データレイクはじめの一歩　～データレイクの活用に欠かせない運用のポイント～**
  - データレイクの運用
  - 2020 年 6 月 11 日（木）　開演：11:00 ～ 12:00
  - https://aws-seminar.smktg.jp/public/seminar/view/2641

- **【オンラインセミナー】Amazon Redshift事例祭り（移行編）～Let's Modernize Our Data Warehouses!**
  - 2020 年 6 月 4 日（木）　開演：14:00 ～ 17:00
  - SOMPOひまわり生命様や、Amazon Japanによる事例紹介
  - https://aws-seminar.smktg.jp/public/seminar/view/2344

アンケート記入のお願い

https://bit.ly/2TFPbps
もしくは
https://amazonmr.au1.qualtrics.com/jfe/form/SV_7WgkEcM0STyLlCl

アンケートにお答えいただいた方には
本日の資料を後日ご提供させていただきます

内容についての注意点

- 本資料では2020年5月28日時点のサービス内容および価格についてご説明しています。最新の情報はAWS公式ウェブサイト(http://aws.amazon.com)にてご確認ください。

- 資料作成には十分注意しておりますが、資料内の価格とAWS公式ウェブサイト記載の価格に相違があった場合、AWS公式ウェブサイトの価格を優先とさせていただきます。

- 価格は税抜表記となっています。日本居住者のお客様には別途消費税をご請求させていただきます。

- AWS does not offer binding price quotes. AWS pricing is publicly available and is subject to change in accordance with the AWS Customer Agreement available at http://aws.amazon.com/agreement/. Any pricing information included in this document is provided only as an estimate of usage charges for AWS services based on certain information that you have provided. Monthly charges will be based on your actual use of AWS services, and may vary from the estimates provided.