aws SUMMIT

2-H1-1-10 / 2-H1-1-19

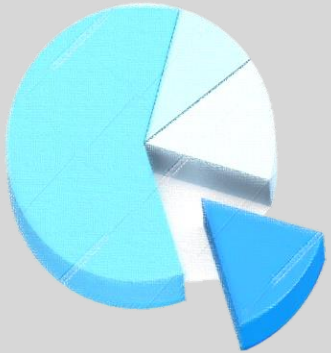
AWS の DataLake を使ったデータ分析入門

アマゾン ウェブ サービス ジャパン 株式会社 技術統括本部 ソリューションアーキテクト

上原 誠

本セッションのFeedbackをお願いします

お手元のサミットガイドブックの表紙に記載している 『QRコード』 からご回答ください。
もれなく素敵なAWSオリジナルグッズをプレゼントします。



プレゼントの引き換えは、パミール3F展示会場内アンケート確認エリア・受付エリアのいずれかにお越し下さい。

名前：上原 誠 (うえはら まこと)

所属：ソリューションアーキテクト

担当：メディア系、アドテクのお客様

最近よく触ってるサービス： AWS Glue

このセッションの対象者

- ビッグデータを活用したいエンジニア
- データ分析 / ML および周辺エコシステムに興味がある
- アーキテクチャの設計に関わるアーキテクト、デベロッパー、マネージャ

このセッションで持ち帰って頂きたいこと

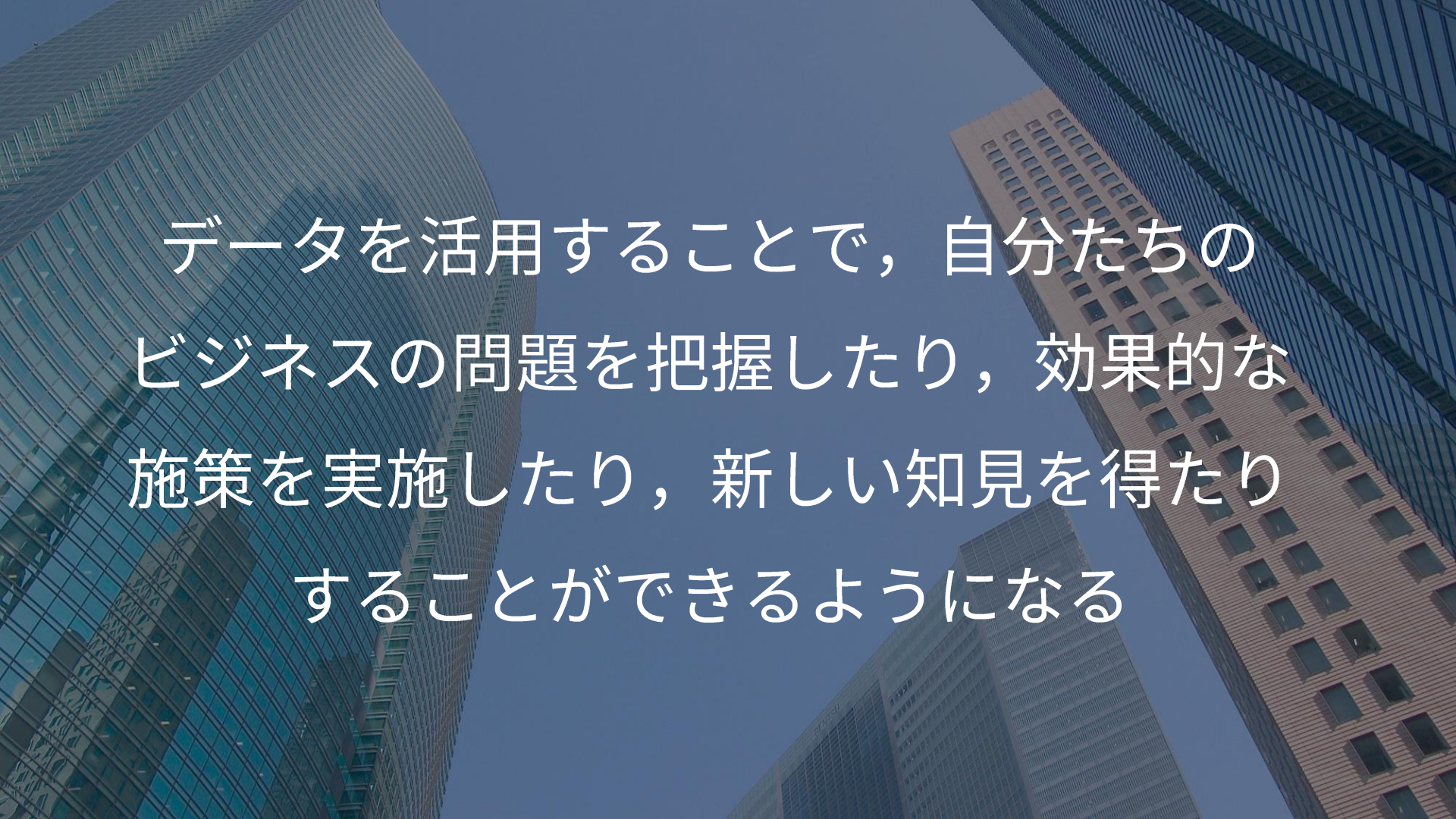
- データ活用の考え方
- データレイクの考え方
- データ活用のための試行錯誤を高速に行うことの必要性とその知識

アジェンダ

1. データ活用の流れ
2. 従来の RDB を中心としたデータ分析
3. データレイクとは
4. データレイクを構成する AWS サービス
5. AWS におけるデータ活用事例
6. まとめ

アジェンダ

1. データ活用の流れ
2. 従来の RDB を中心としたデータ分析
3. データレイクとは
4. データレイクを構成する AWS サービス
5. AWS におけるデータ活用事例
6. まとめ


データを活用することで，自分たちの
ビジネスの問題を把握したり，効果的な
施策を実施したり，新しい知見を得たり
することができるようになる

データ活用の流れ

データを貯める	データを貯める 活用するためのデータがなければ、なにもできない
データを可視化	貯めたデータを適切に加工整形処理して可視化 常に全ての必要なデータを見ることができるようにする
データ 分析 / ML	上記 2 つが回るようになってはじめて、高度な分析 を行いビジネス課題とゴールが明確にできる

ビッグデータ処理のパイプライン

データ分析において唯一の万能なツールはない

各フェーズで必要なツールやリソースを確保

- データ活用の際には、さまざまなツールが必要となる
  - データの取得と保存、加工整形処理を行うためのツール
  - 可視化- BIツール / 基礎集計のための SQL
  - ディープラーニングフレームワークや Hadoop クラスタ、Notebook などの分析ツール
- ツールやモデルによって必要なリソースも異なる
  - CPU / GPU / メモリ / IO
  - 複数インスタンス / クラスタ

高速に試行錯誤を行う

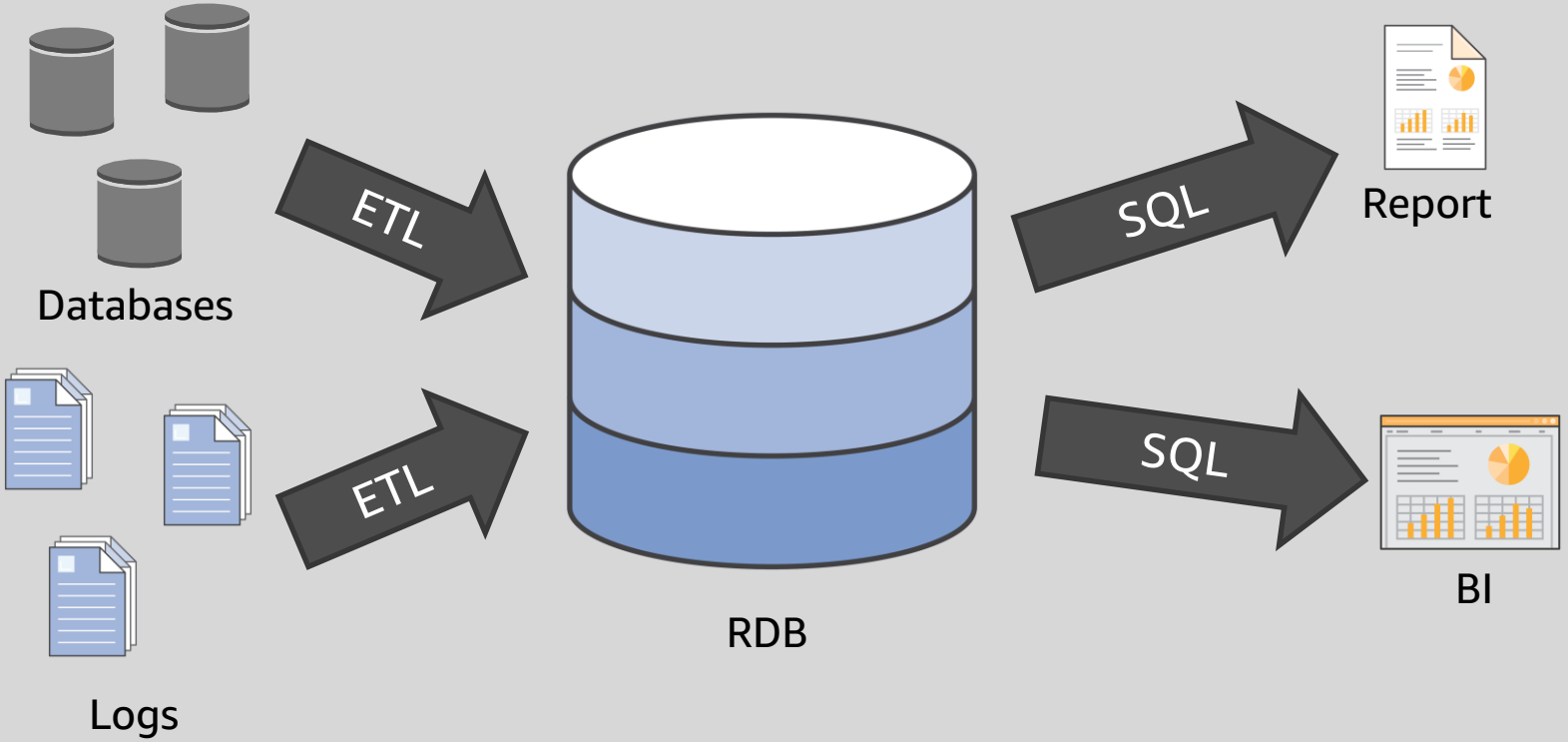
- データ活用は基本的に試行錯誤を伴う
  - データの量や中身の変化への継続的な対応
  - ビジネスの状況変化に伴う、さまざまな指標の定期的な見直し
  - 機械学習モデルの構築・改善サイクル
- この試行錯誤のサイクルをいかに高速に積み重ねるかが、良い結果を導きだすために重要

データ分析をうまく進めるには
適切なツールやリソースを確保し
試行錯誤のサイクルを高速に実施する必要がある

アジェンダ

1. データ活用の流れ
2. 従来の RDB を中心としたデータ分析
3. データレイクとは
4. データレイクを構成する AWS サービス
5. AWS におけるデータ活用事例
6. まとめ

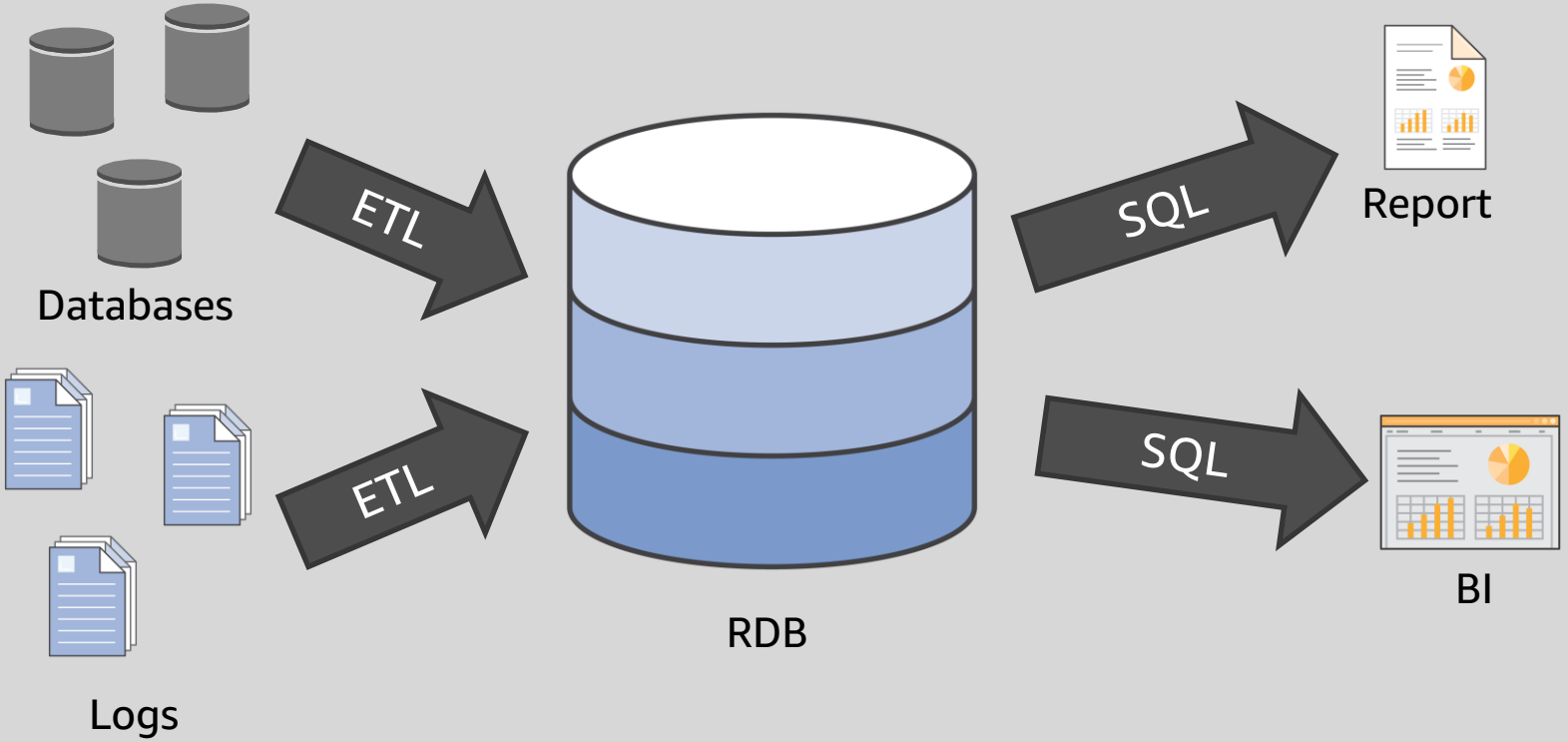
従来の RDB を中心としたデータ分析



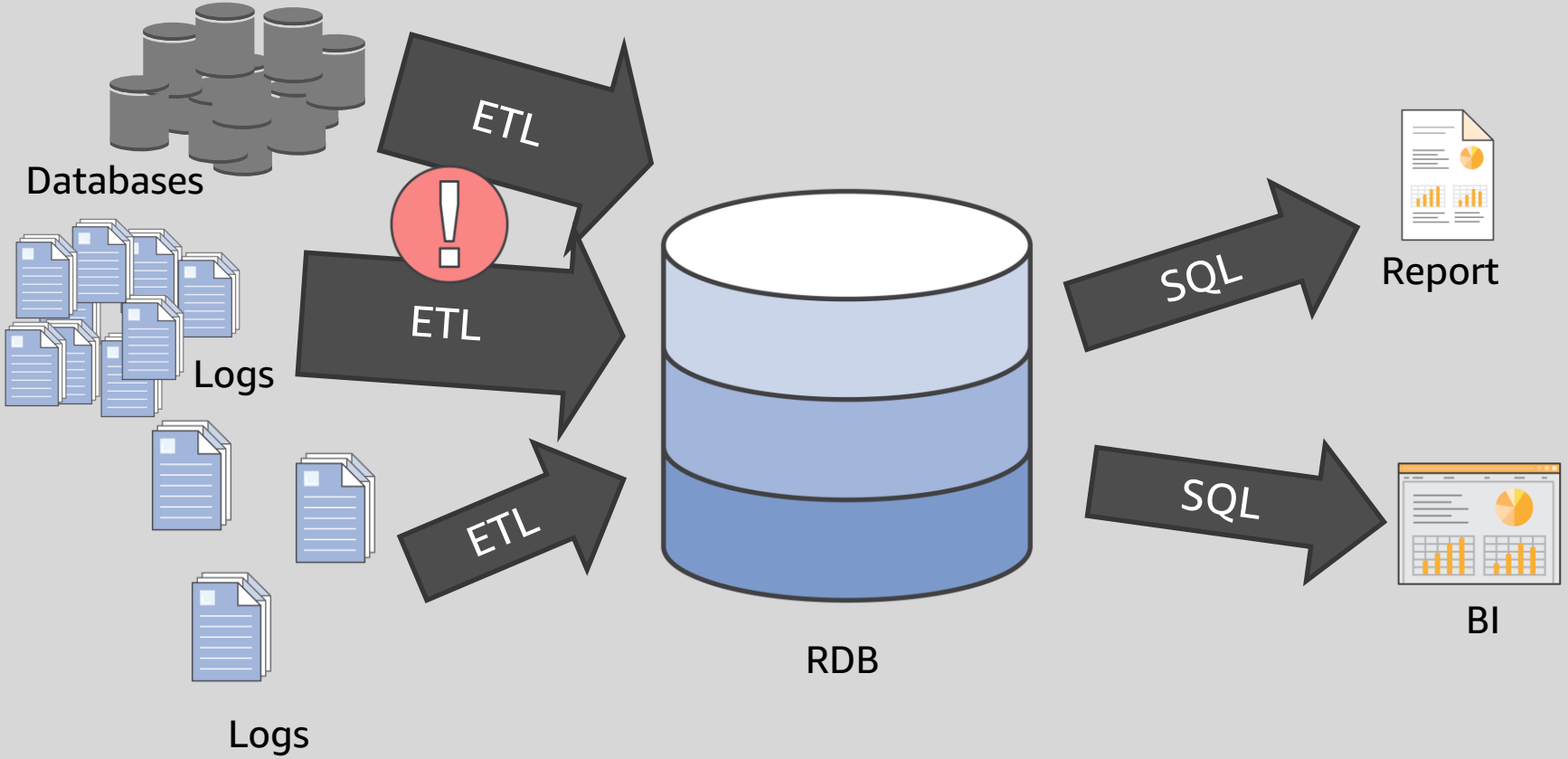
RDB を利用する利点

- スキーマが定義されている
- データの型も強制できる
- アクセスするツールが簡単、エコシステムが安定
- トランザクション

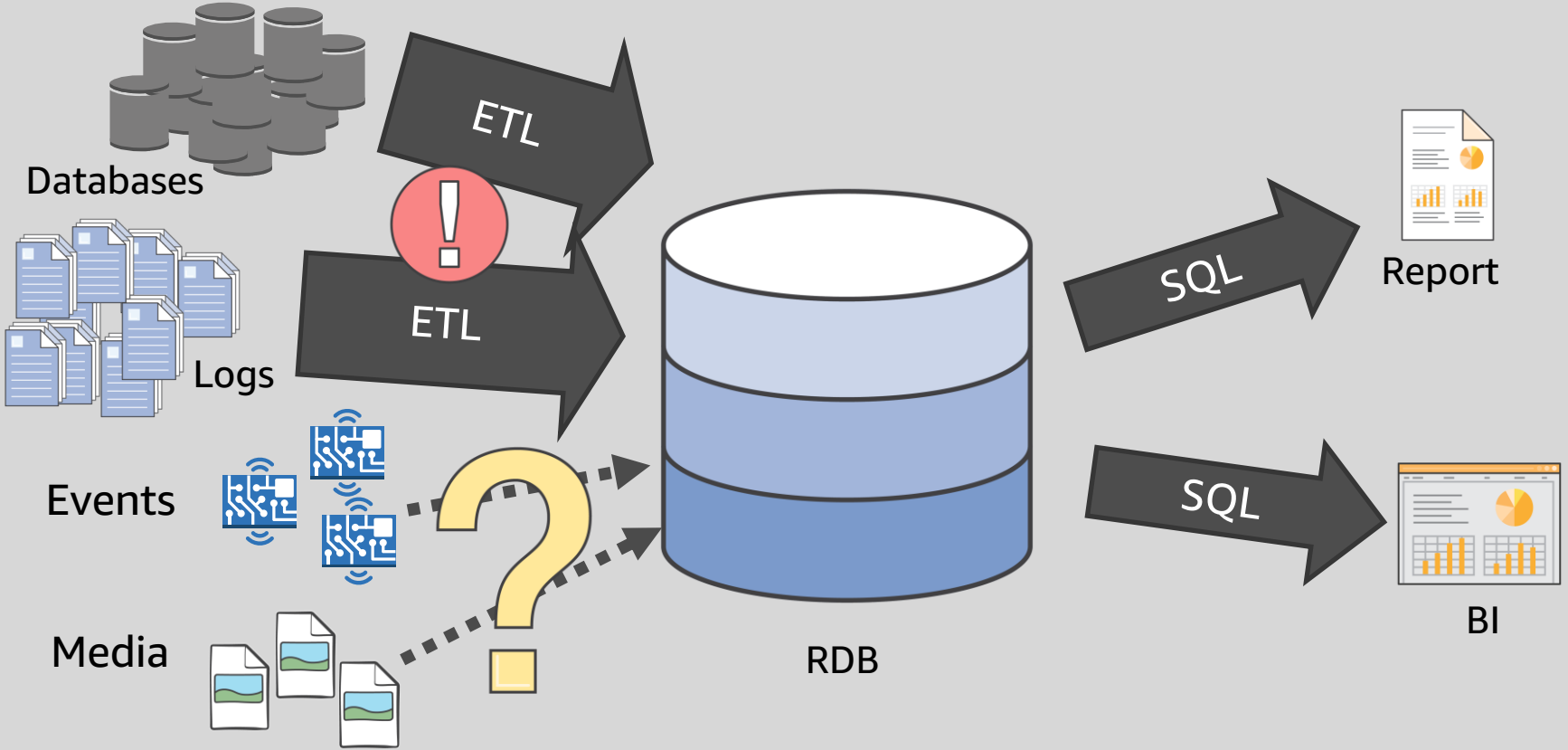
課題1: 入ってくるデータの量と質の変化



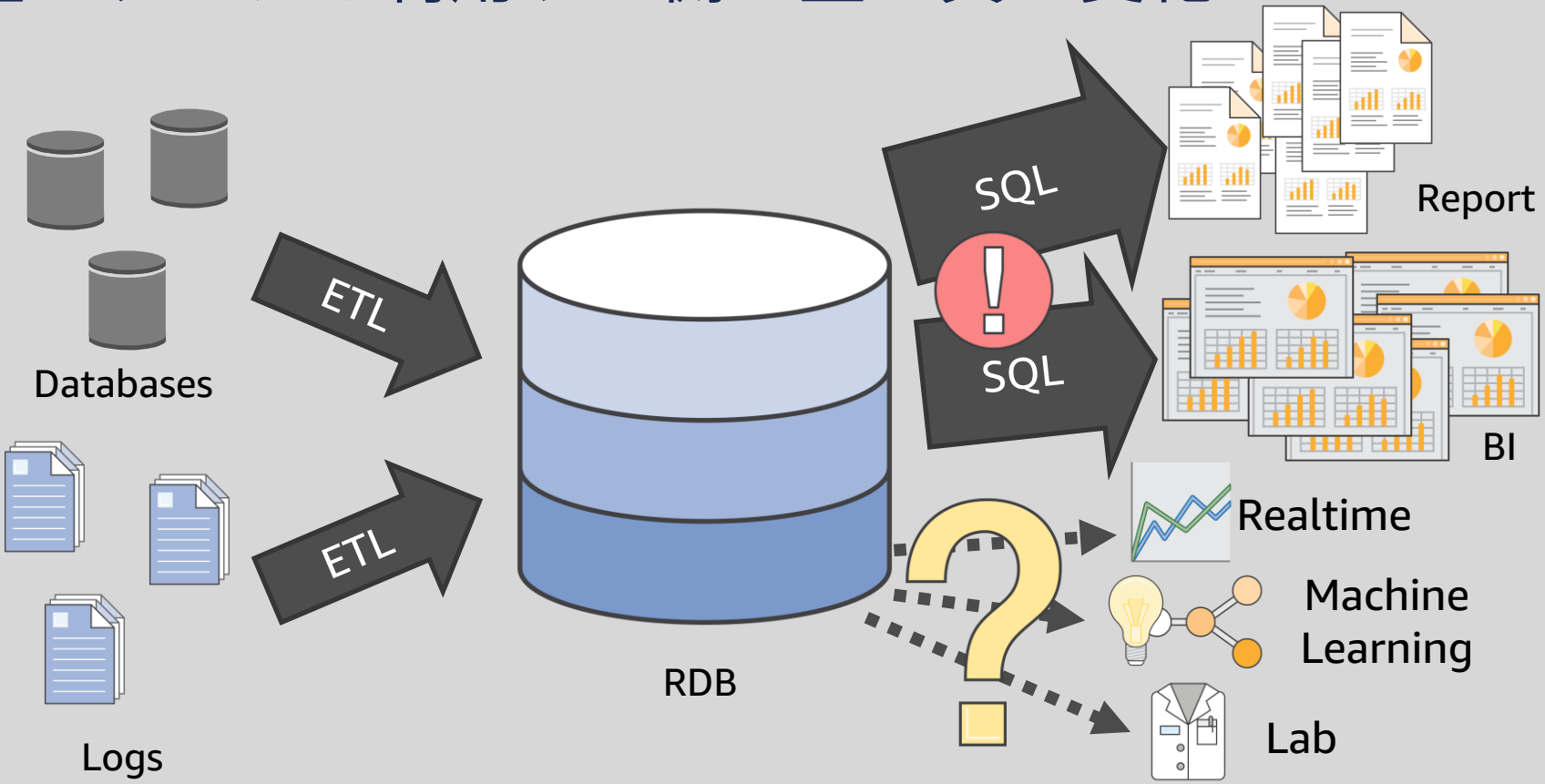
課題1: 入ってくるデータの量と質の変化



課題1: 入ってくるデータの量と質の変化



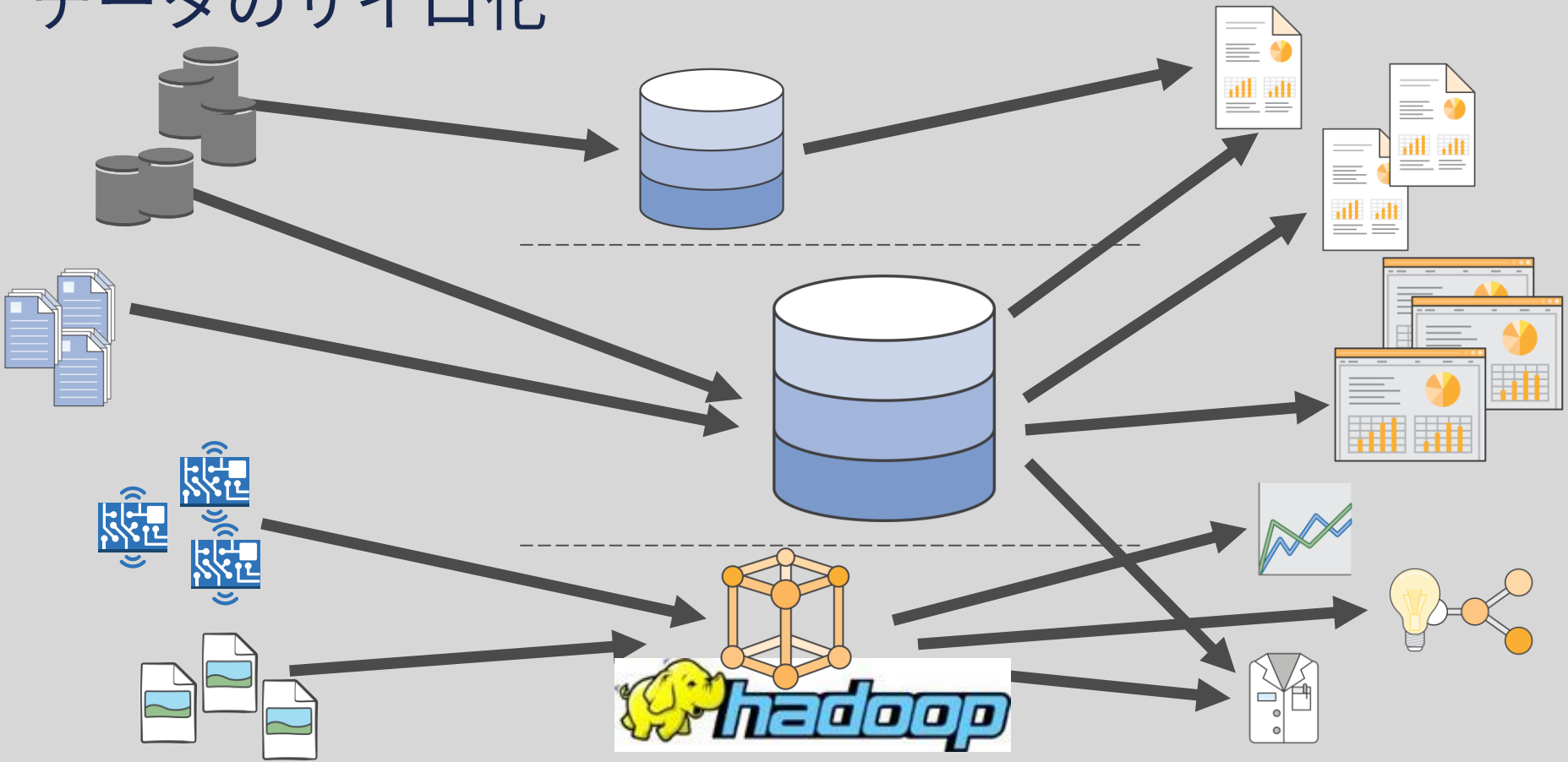
課題2: データを利用する側の量と質の変化



RDB の課題

- スケーラビリティに限界がある
- 非構造化データに対応しずらい
- リアルタイム分析や機械学習などの新たな分析要件に対応しずらい、またはできない

データのサイロ化



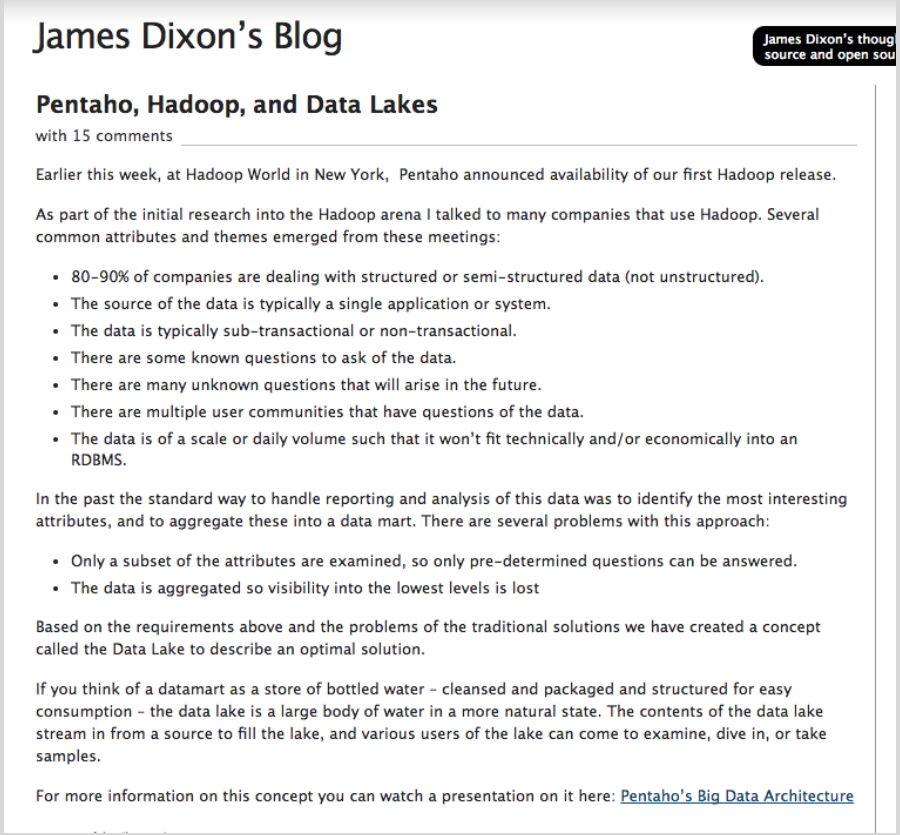
データのサイロ化の課題

- スケーラビリティの課題は解消しない
  - 要求が10倍になったら、サイロ数が10倍になる？？
- どこにどのデータがあるか分からない
  - 途中で場所が変わったら更にカオスに
- サイロをまたいだ分析が困難
  - 同じデータを複数サイロに重複して持つと、無駄や同期が課題に

アジェンダ

1. データ活用の流れ
2. 従来の RDB を中心としたデータ分析
3. データレイクとは
4. データレイクを構成する AWS サービス
5. AWS におけるデータ活用事例
6. まとめ

データレイクとは



James Dixon's Blog

James Dixon's thoug
source and open sou

Pentaho, Hadoop, and Data Lakes

with 15 comments

Earlier this week, at Hadoop World in New York, Pentaho announced availability of our first Hadoop release.

As part of the initial research into the Hadoop arena I talked to many companies that use Hadoop. Several common attributes and themes emerged from these meetings:

- 80-90% of companies are dealing with structured or semi-structured data (not unstructured).
- The source of the data is typically a single application or system.
- The data is typically sub-transactional or non-transactional.
- There are some known questions to ask of the data.
- There are many unknown questions that will arise in the future.
- There are multiple user communities that have questions of the data.
- The data is of a scale or daily volume such that it won't fit technically and/or economically into an RDBMS.

In the past the standard way to handle reporting and analysis of this data was to identify the most interesting attributes, and to aggregate these into a data mart. There are several problems with this approach:

- Only a subset of the attributes are examined, so only pre-determined questions can be answered.
- The data is aggregated so visibility into the lowest levels is lost

Based on the requirements above and the problems of the traditional solutions we have created a concept called the Data Lake to describe an optimal solution.

If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.

For more information on this concept you can watch a presentation on it here: Pentaho's Big Data Architecture

https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

増え続けるデータに対して多くの企業での課題

Hadoop 初期の研究で、Hadoop を使用する多くの企業で共通の課題

- 構造化データと非構造化データの混在
- 分析パターンの把握と整理の難しさ
- ビッグデータへの対処

データレイク

どのような分析を行いたいかを事前に完璧に把握することはできない

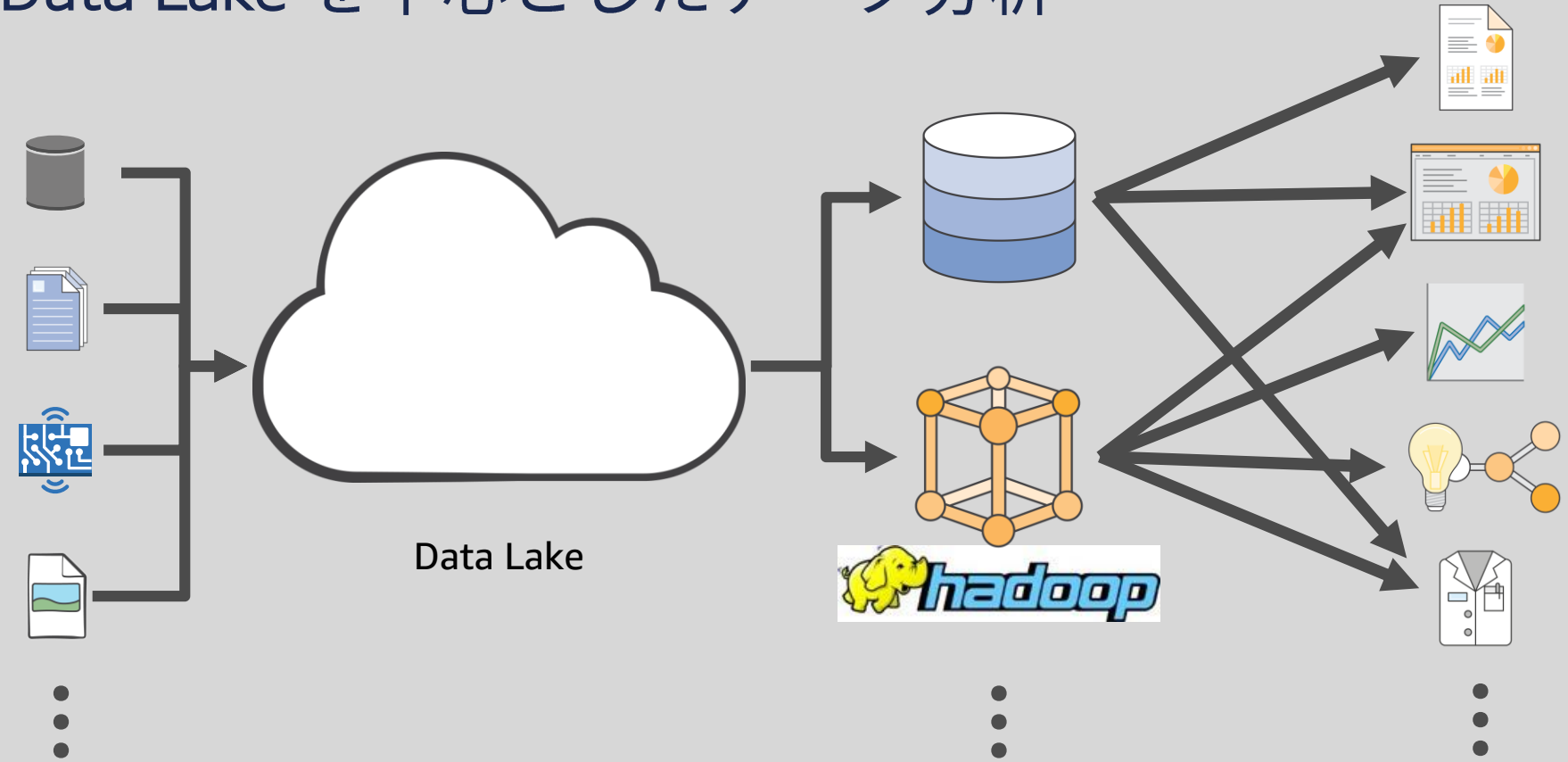
データレイクは、どんな要望がでてきても対処できるよう

全てのデータを生のフォーマットで安全に保存し

データ活用時に必要な構造で取り出すという考え方

データレイクは未来の不確実性に対処する術

Data Lake を中心としたデータ分析

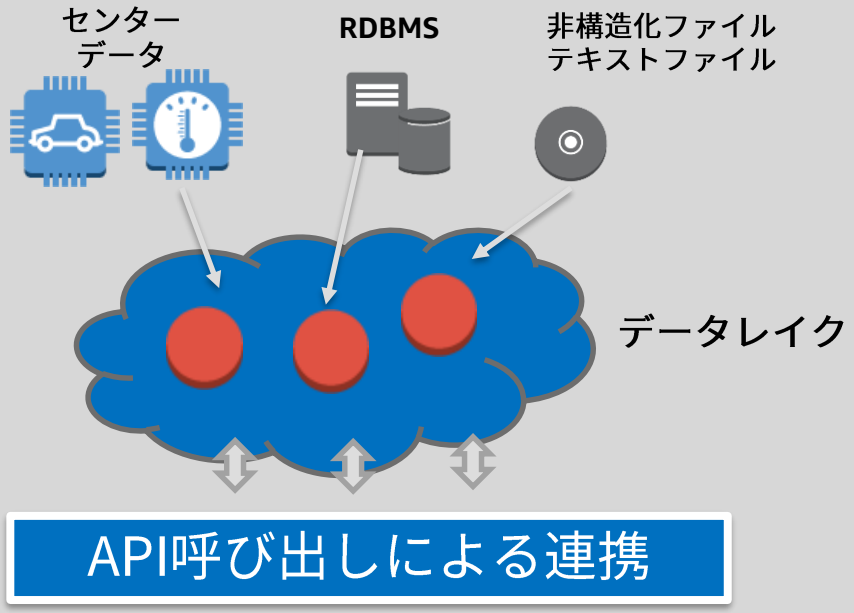


データレイクのメリット

- ストレージと計算処理の分離
  - それぞれ独立してスケールできるので最適化しやすい
- Single Source of Truth (SSOT)
  - データレイクにあるものを正とすれば良い
- 様々な input/output 手法に対応
  - in/out が独立、ETL も独立できるので、後からの拡張がスムーズ

データレイクとは (ここまでのまとめ)

- 多様なデータを一箇所に保存
- 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
- システム全体のハブ



アジェンダ

1. データ活用の流れ
2. 従来の RDB を中心としたデータ分析
3. データレイクとは
4. データレイクを構成する AWS サービス
5. AWS におけるデータ活用事例
6. まとめ

データ処理のパイプライン

クラウドで変わるデータ分析：
データは加工せず全期間を残す

これまで：

1. ディスクの容量に上限がある
2. データはサマリーだけ、もしくは期間限定で保存
3. 処理できる内容は固定的

インフラ管理者の仕事：
ストレージを溢れさせず、時間内に処理が終るようにサイズや処理内容を調整する

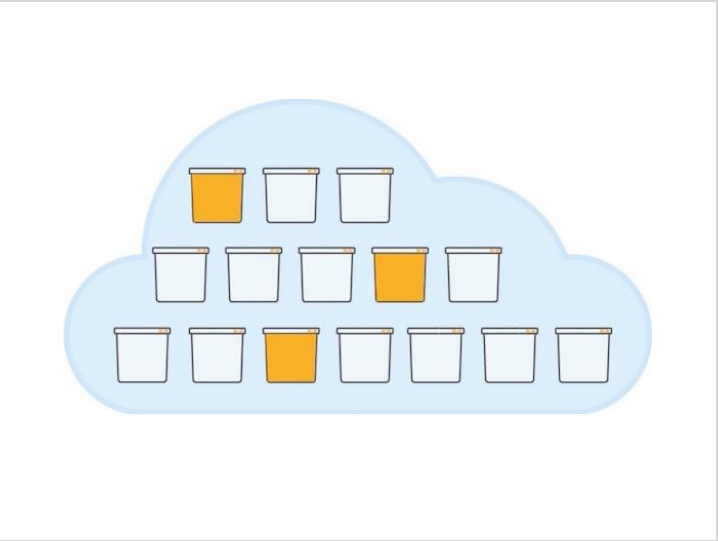
On AWS：

1. 安価・上限無しのストレージ
2. オリジナルデータを全て残す
3. 処理対象・処理内容はビジネスに合わせて変わる

インフラ管理者の仕事：
データを活用して新しい課題に素早く対応できるインフラを用意する。個別リクエストへの対応

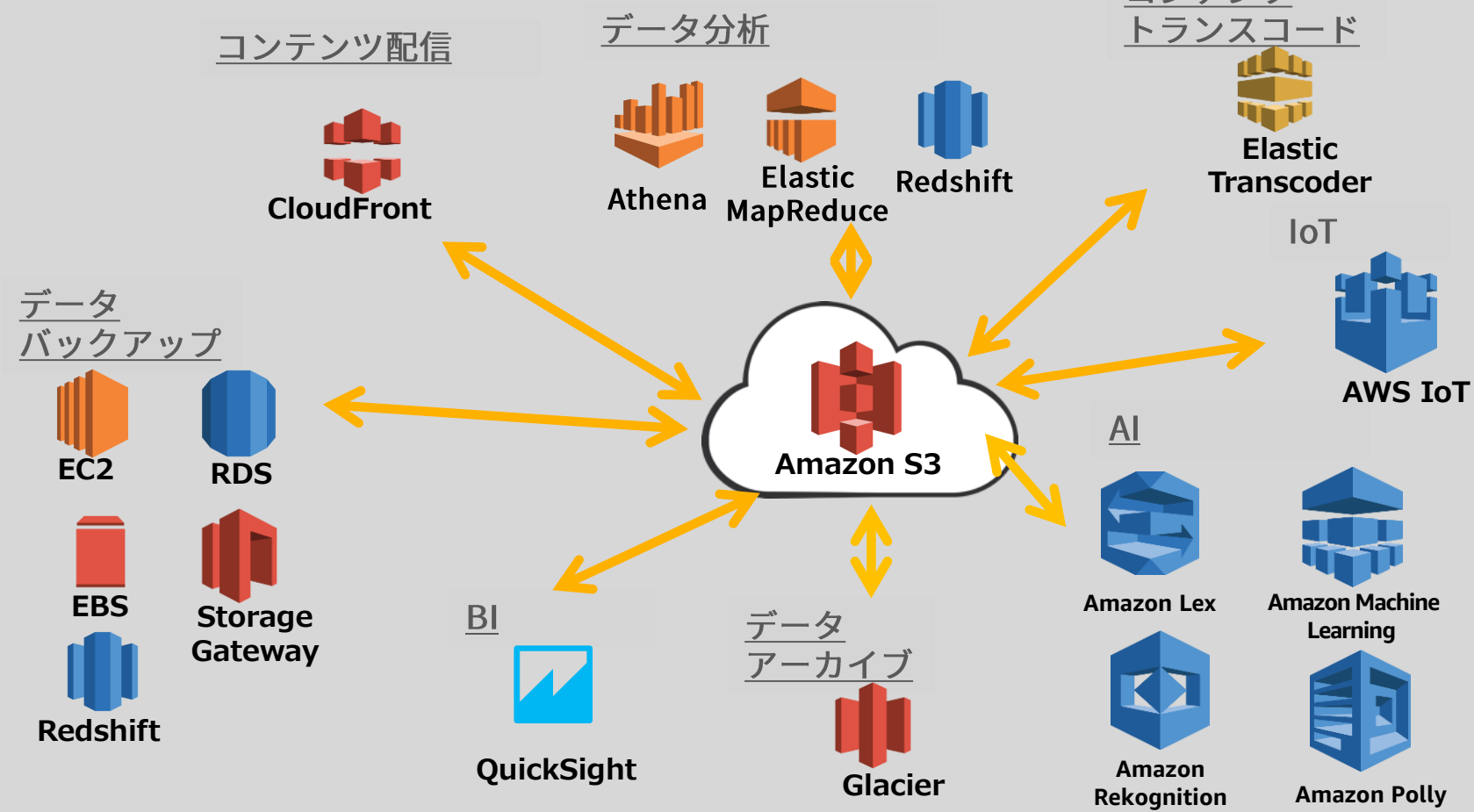
Amazon S3

高い耐久性と可用性を持つスケーラブルなオブジェクトストレージ



- 容量無制限
- 99.999999999%の耐久性
- 利用した分だけの課金
- API操作可能 多くの AWS サービスと連携

AWSのデータレイク ＝ Amazon S3



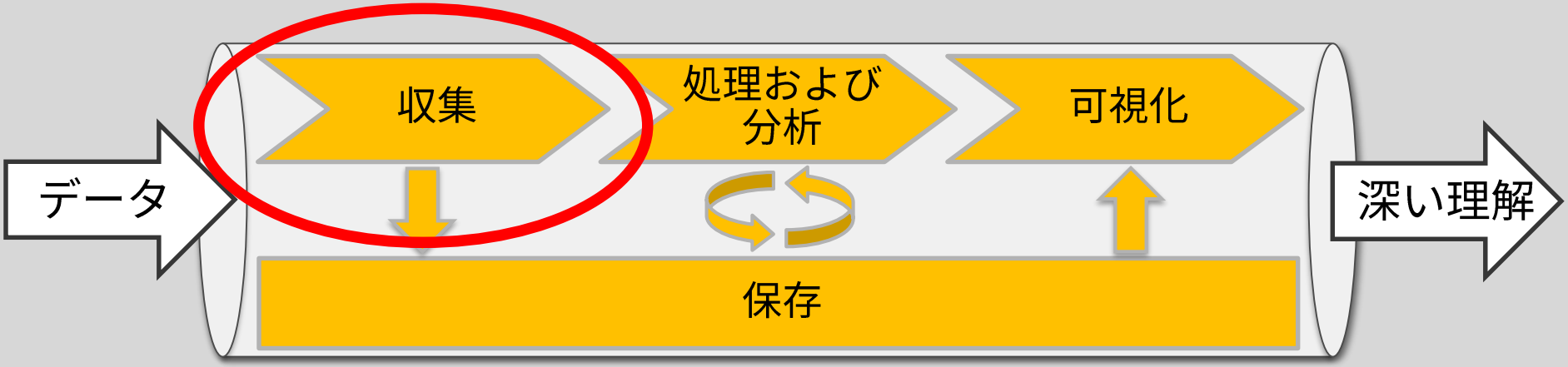
AWSのデータレイク ＝ Amazon S3
様々なアクセスに対する S3 のセキュリティ

- IAM を使ったきめ細かなアクセス制御

- CloudTrail を使った S3 操作ログの記録と監査

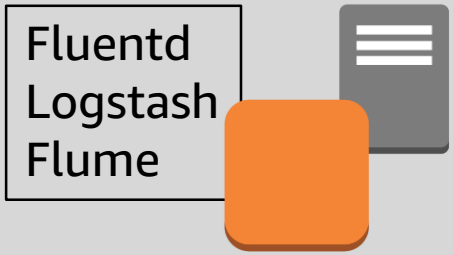
- S3 へのアクセスログの記録と監査
- バケットポリシーによるアクセス制御
- オブジェクトの暗号化 (CSE, SSE)

データ処理のパイプライン



要件に合わせた収集のツールやサービスを使う

OSS on EC2



OSS のログ収集ツールで豊富なプラグインで様々なデータのインアウトが可能

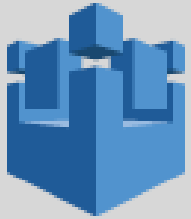
Amazon Kinesis

リアルタイムなストリーミング処理のたの安全なストレージ

AWS Snowball

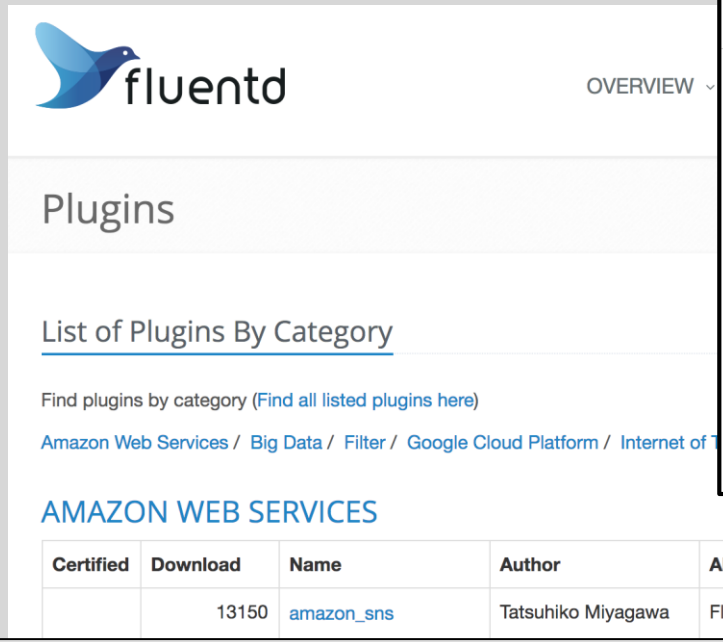
物理デバイスでオンプレミスのペタバイト規模のデータをS3にセキュアに転送

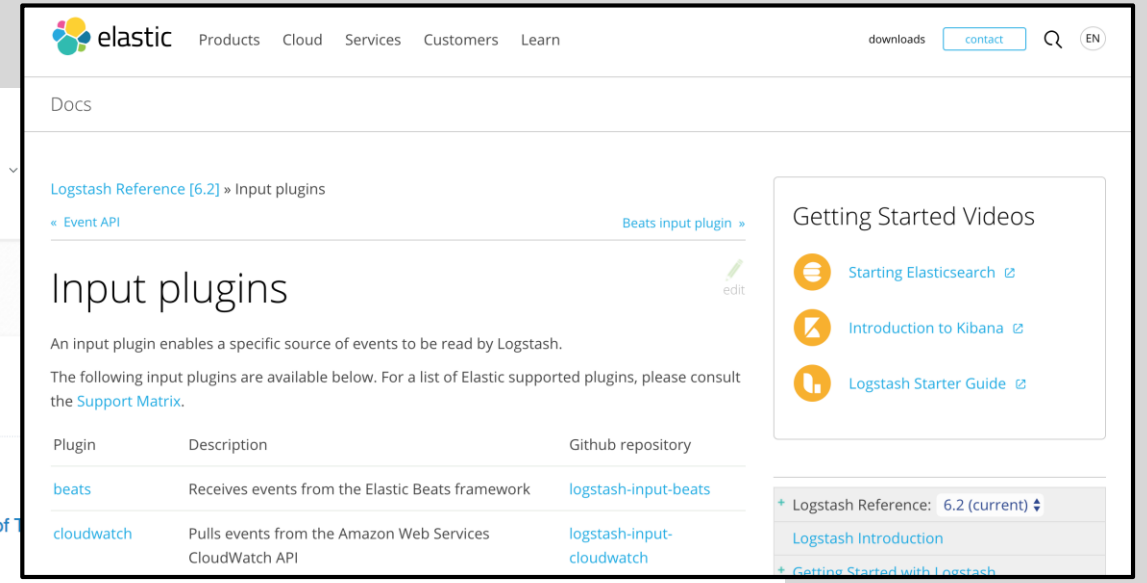
AWS IoT



MQTTやHTTPSによる数十万規模のデバイスからのセキュアなデータ収集

広く使われる OSS

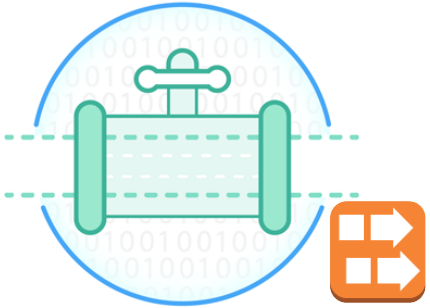




725	kinesis	Amazon Web Services	Fluentd output plugin that sends events to Amazon Kinesis.
608	kinesis-aggregation	Atlassian	Fluentd output plugin that sends KPL style aggregated events to Amazon Kinesis.

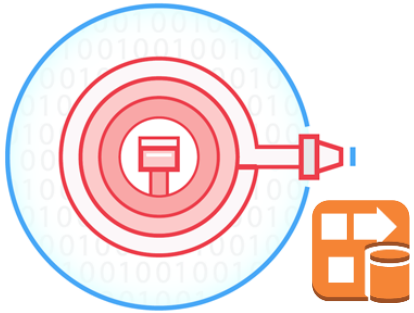
Amazon Kinesis

ストリームデータを収集・処理・配信するためのフルマネージドサービス群



Kinesis Data Streams

ストリームデータを処理・分析するためのデータを格納



Kinesis Data Firehose

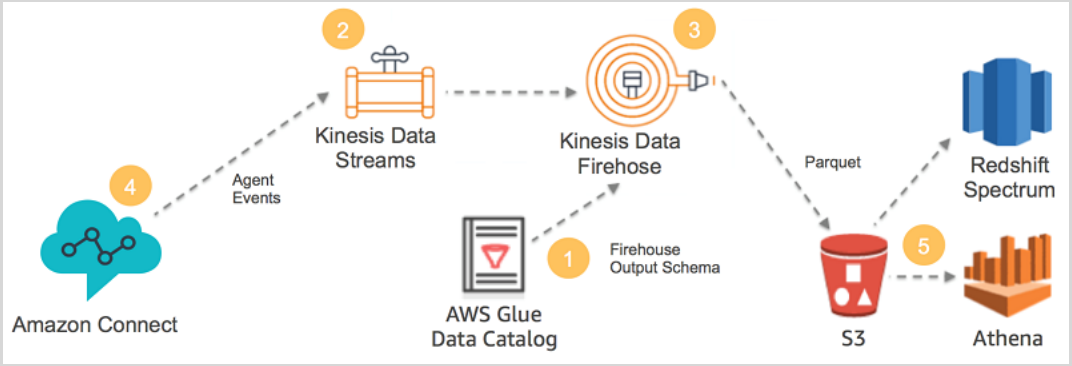
ストリームデータを S3, Redshift, Amazon ES, Splunk に簡単にロード

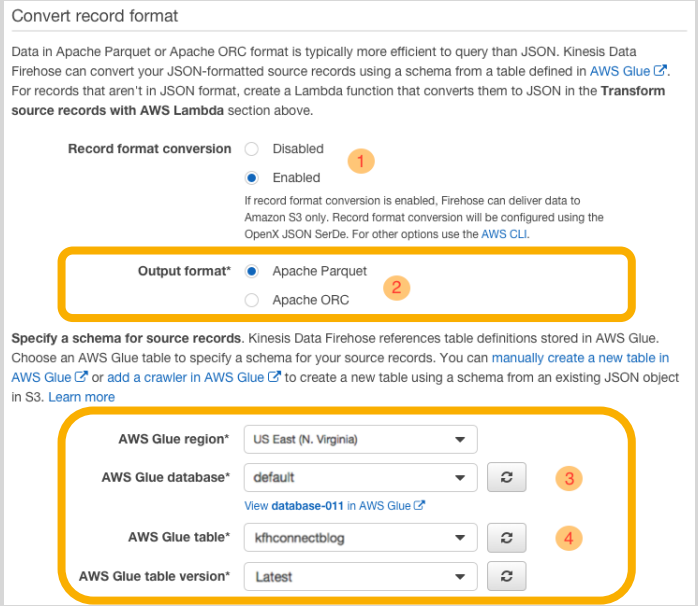
Kinesis Data Analytics

ストリーミングデータを標準的な SQL クエリで簡単に分析

Kinesis Data Firehose から AWS Athena へ連携

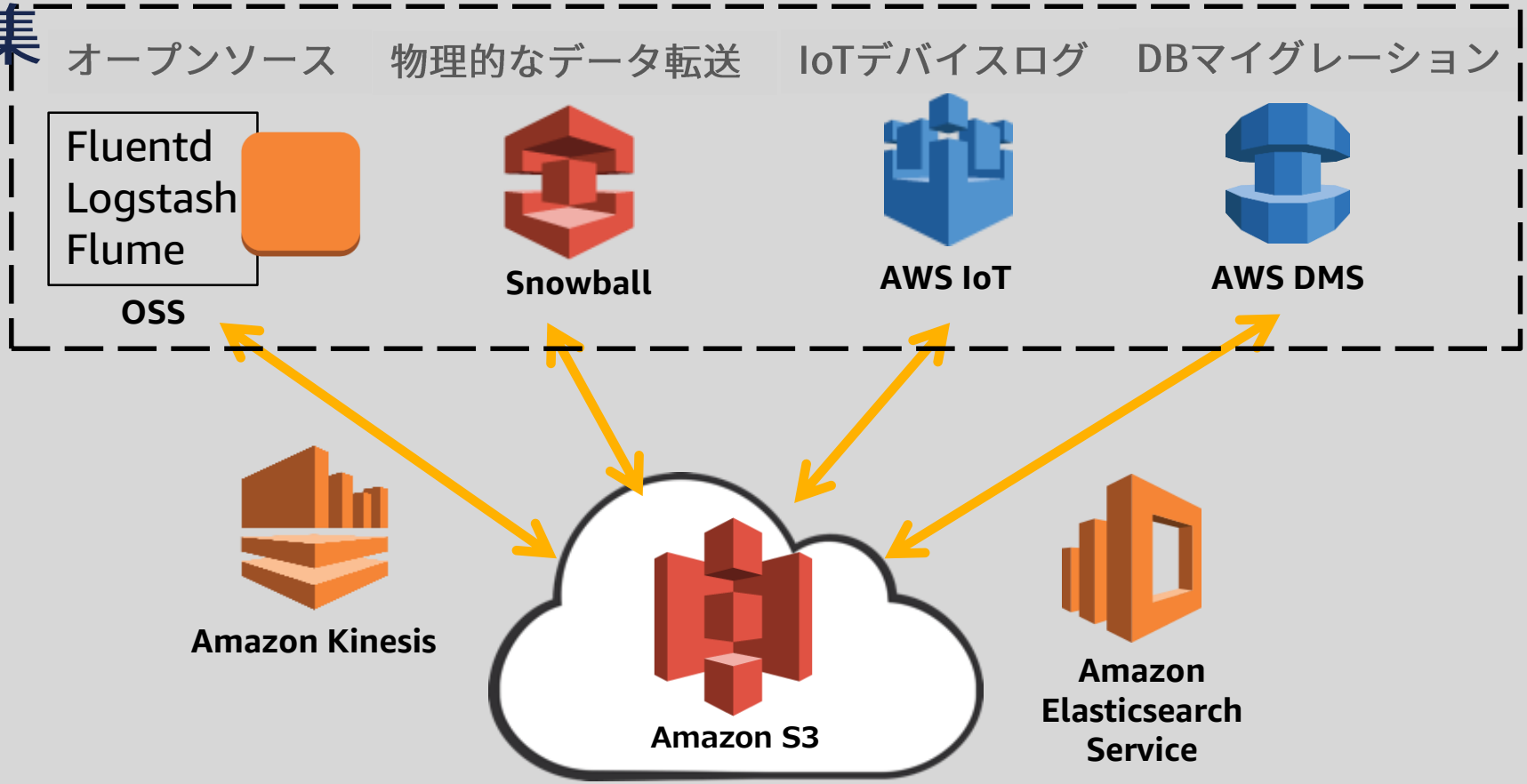
2018.5.10 GA





https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/

要件に合わせたツールでデータレイクにデータ収集



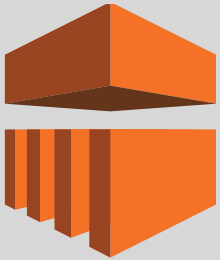
データ処理のパイプライン

要件に合わせた分析・処理サービスを使う

Amazon Redshift

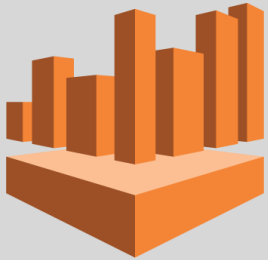
高速DWHとしての機能＋S3上への全データへのアクセス

Amazon EMR



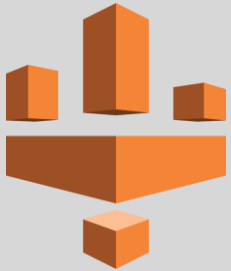
自由度が高く、高いスケーラビリティを誇るプログラム実行基盤

Amazon Athena



S3上のデータにアクセス。サーバ管理や準備不要でアドホックなクエリに強い

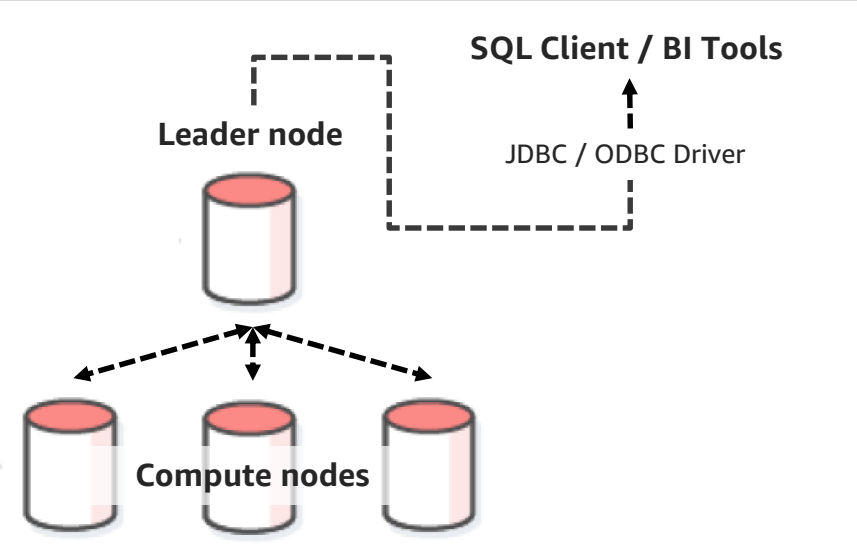
AWS Glue



サーバーレスなETL処理と共通のスキーマ情報のクローリングと管理

Amazon Redshift

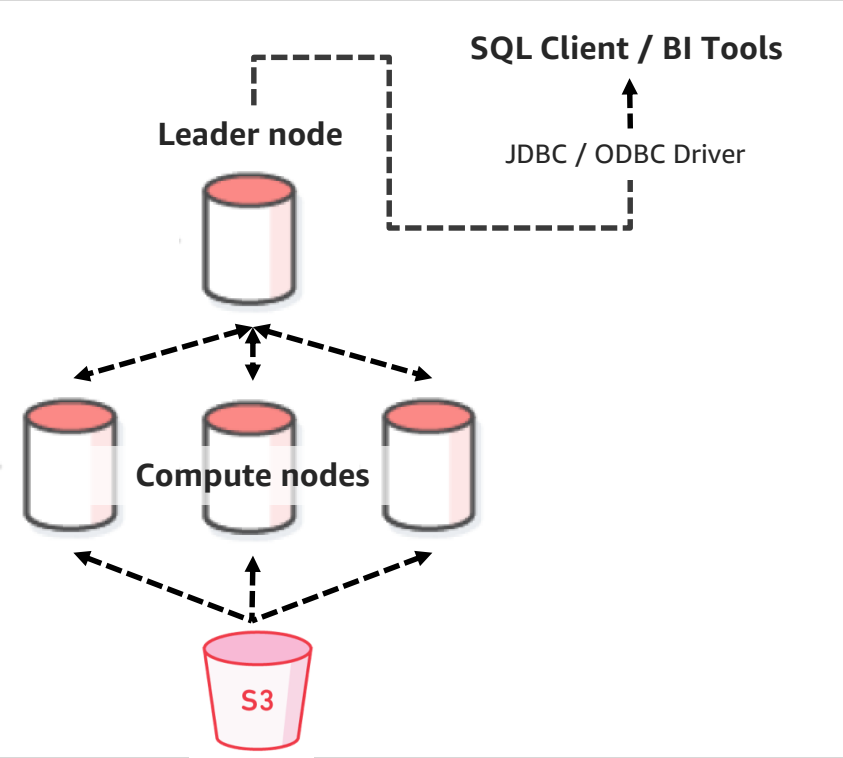
フルマネージドでスケーラブルなデータウェアハウスサービス



- MPP(Massively Parallel Processing) アーキテクチャとカラムナのデータ格納により、スケーラブルで高速なクエリが実行可能
- データストアを最大 2PB まで拡張
- JDBC/ODBC 経由でさまざまな BI ツールと連携

Amazon Redshift Spectrum

Redshift から S3 上のデータに直接クエリできる拡張機能

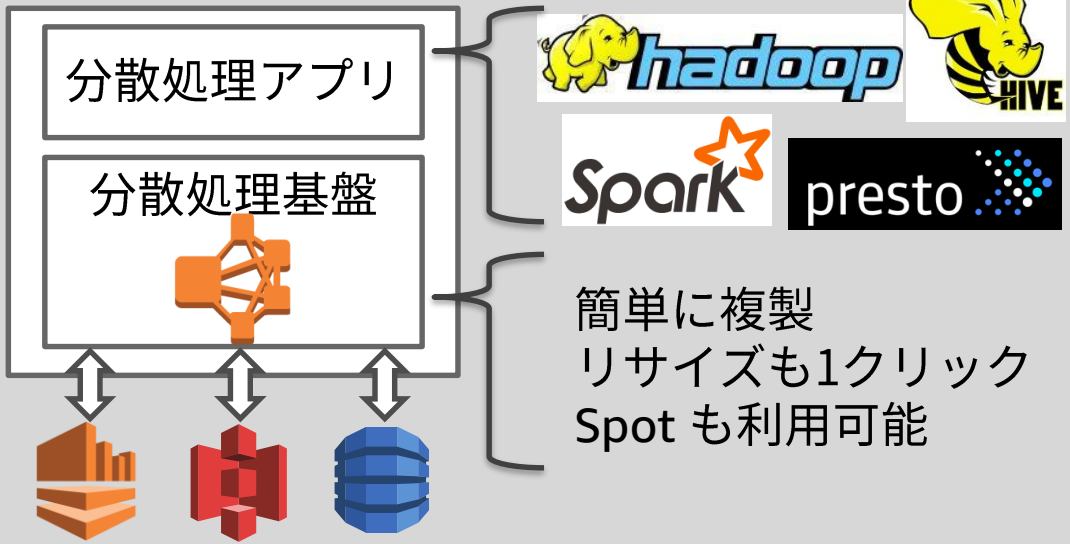


- Redshift クラスタから、直接 S3 上のデータに対してクエリを実行
- Redshift 内のデータと JOIN することも可能
- 利用頻度の低いコールドデータを S3 に保存して Spectrum 経由でアクセスし、ホットデータのみ Redshift 内にロードしておく
- 複数の Redshift クラスタから同じ S3 上のデータを読み込む

Amazon EMR

大規模データ処理を Hadoop/Spark などの分散処理フレームワークを使って効率的に処理

Amazon EMR クラスタ



AWS 上の分散処理サービス
- 簡単かつ安全に Big Data を処理
- 多数のアプリケーションサポート

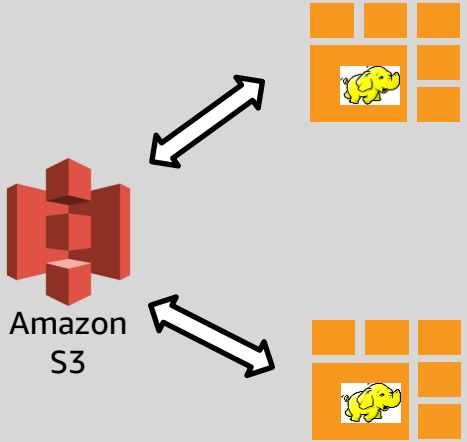
簡単スタート
- 数クリックでセットアップ完了
- 分散処理アプリも簡単セットアップ

低コスト
- ハードウェアへの投資不要
- 従量課金制
- 処理の完了後、クラスタ削除
- Spot インスタンスの活用

EMRFS: S3 を HDFS の様に扱う

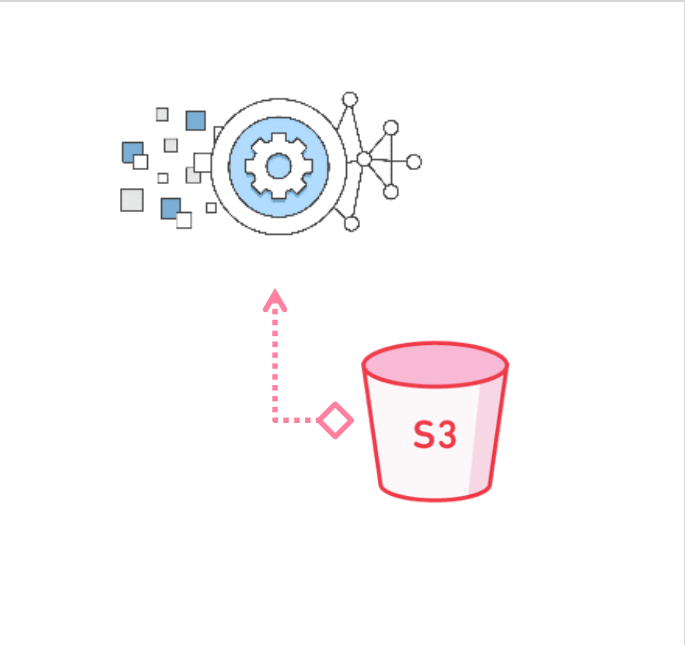
“s3://” と指定するだけでHDFSと同様にS3にアクセス

- 計算資源とストレージを分離できる
  - コスト面でもメリット大
- クラスタの削除が可能
  - クラスタを消してもデータをロストしない
- 複数クラスタ間でデータ共有が簡単
- データは耐久性の高い S3 に配置



Amazon Athena

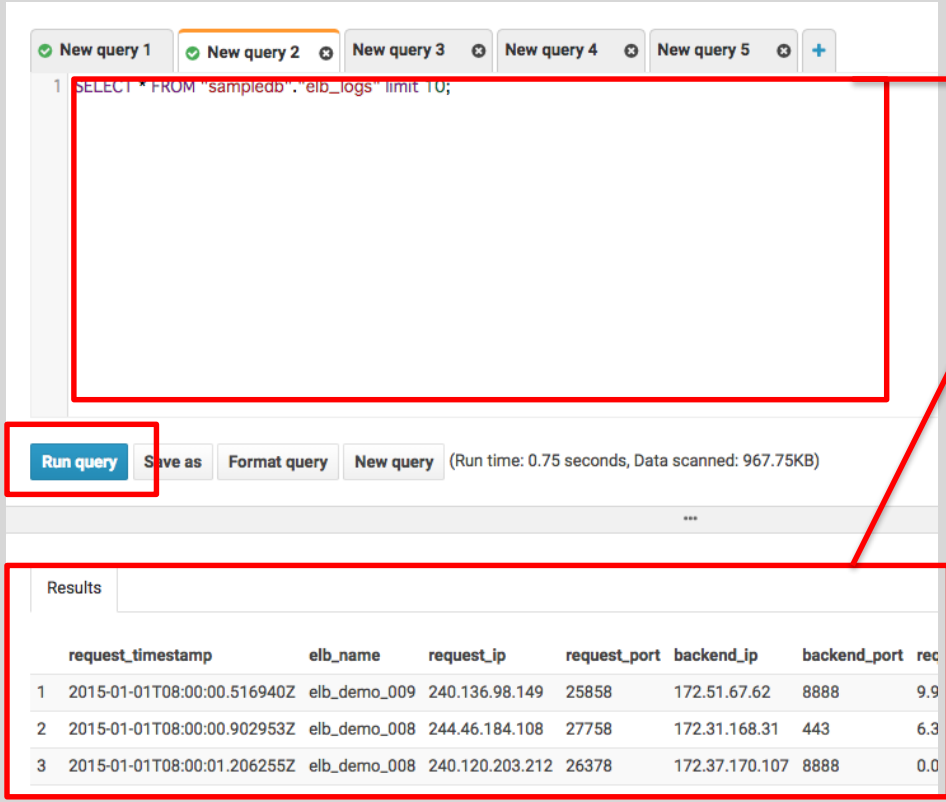
サーバーレスのインタラクティブなクエリサービス



- サーバーレス検索サービス
- S3 のデータに対して直接クエリできる
- Presto ベースで標準 SQL が実行可能
- 実行したクエリの容量ぶんの従量課金
- スキャンされたデータ1TBあたり5$

Amazon Athena

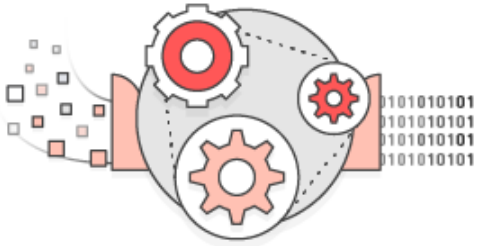
S3 に保存したファイルに直接クエリー



- クエリエディタにSQLを記述しクエリを実行
- コンソールに結果が表示
- 過去のクエリ結果は History からダウンロード可能
- クエリ結果は S3 に自動保存

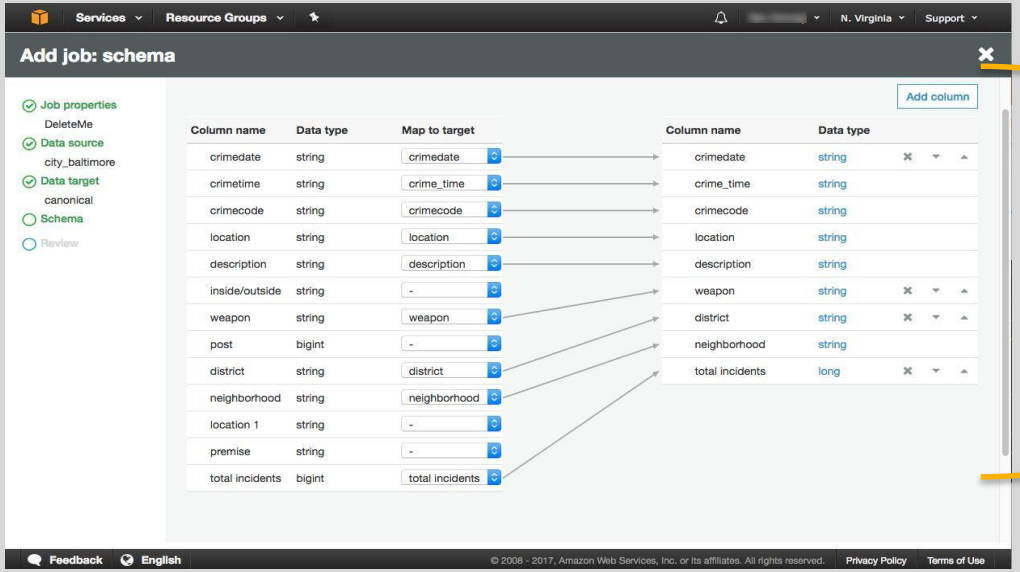
AWS Glue

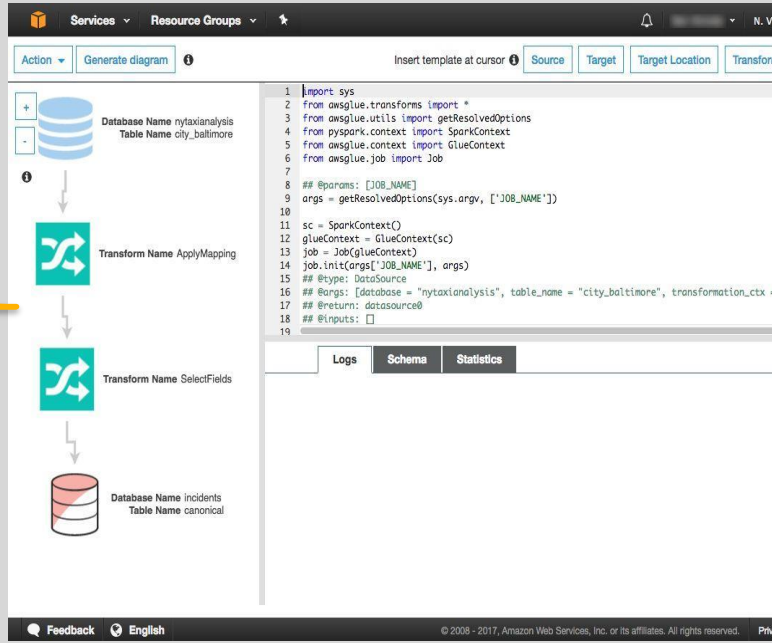
データソースの把握・準備・データ格納を簡単で確実に



- フルマネージドの**ETL サービス + データカタログ**
- GUI 上で作成した ETL 処理フローから **PySpark** または**Scala**コードが生成され必要に応じて編集可能
- データカタログは、**Athena, Redshift Spectrum, EMR 上の Spark, Hive** からも利用可能

Glue : GUI での ETL コード雛形生成





- GUIでデータソース、ターゲット、列のマッピングを設定することで雛形が生成される
- CSV→Parquetのフォーマット変換だけといった処理であればコード作成無しで実現

Glue : ETL コードの作成

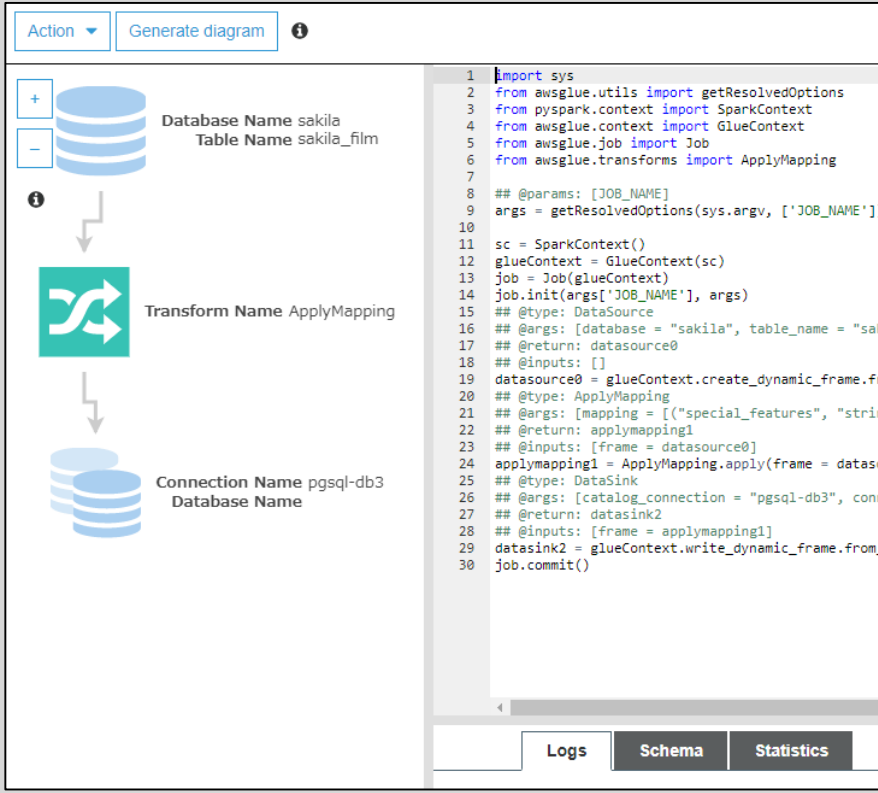
ジョブ＝ETL処理を実行する単位

- **PySpark, Scala** で記述
- Extract（抽出）や、Load（取り込み）は抽象化されているため、**主にTransform（変換）を既述**する

■サンプルスクリプト集

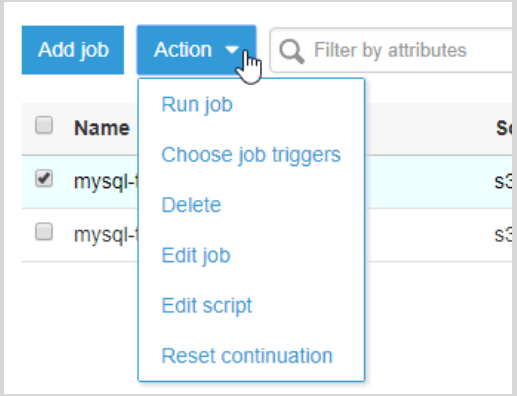
https://github.com/awslabs/aws-glue-samples

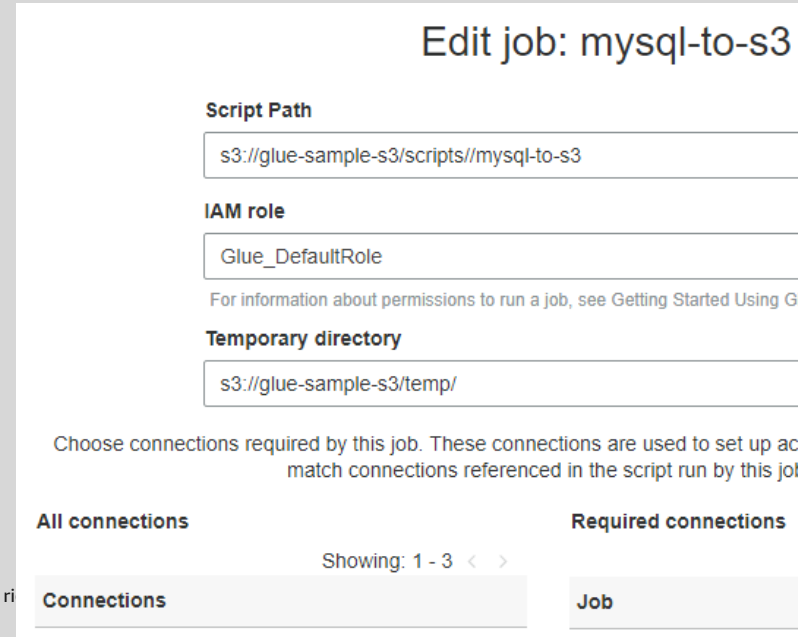
（※同じサイトにFAQもあり、こちらも必読）



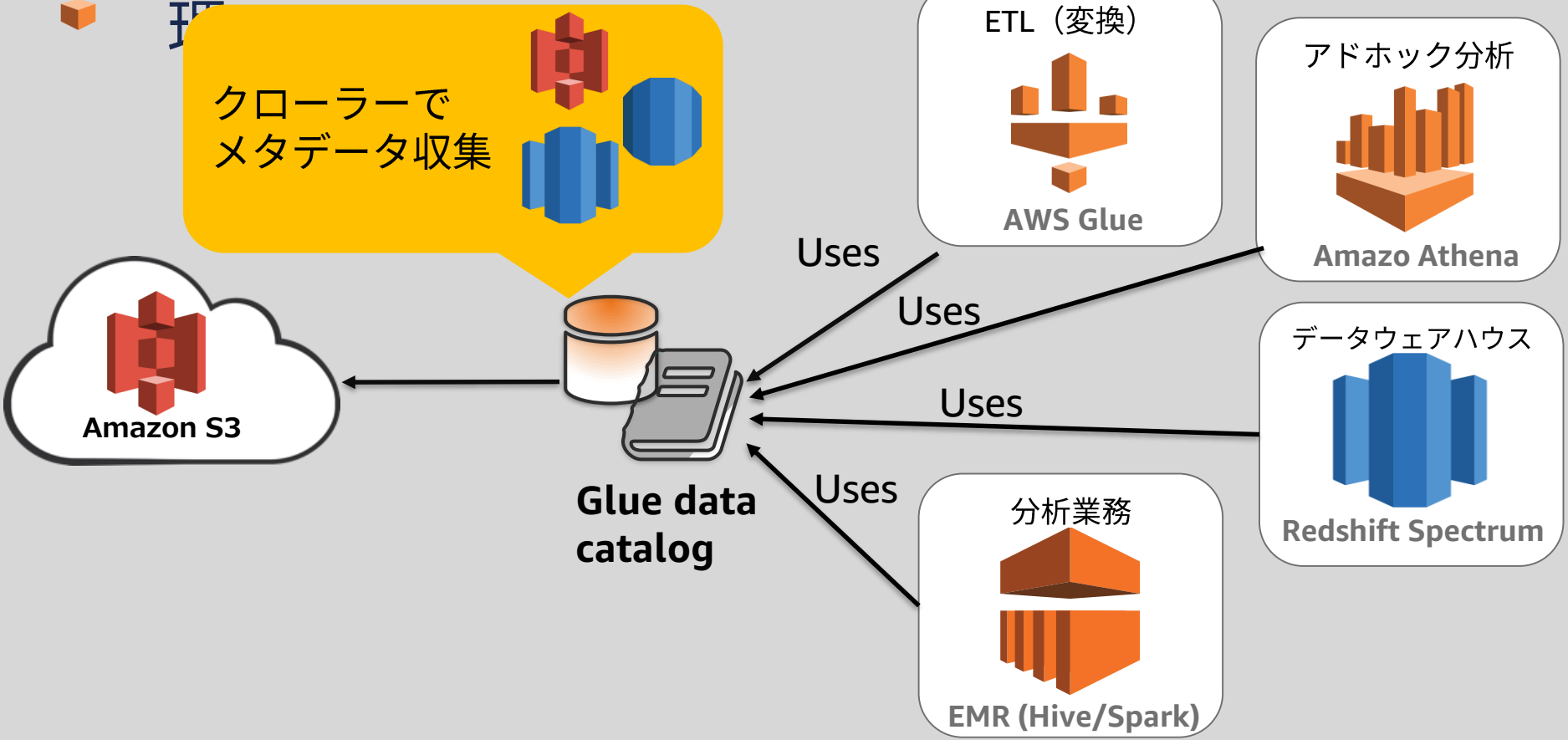
Glue : ジョブの定義と実行

- 作成したコードを読み込んで実行
  - IAM ロールで権限を設定
- ジョブの実行開始方法
  - API コール（手動）
  - トリガー （スケジュール実行可能）
- リトライ制限の指定や、パラメータを渡すことが可能
- 実行ログは CloudWatch Logs に出力





データカタログで AWS 上のメタデータを集中管理

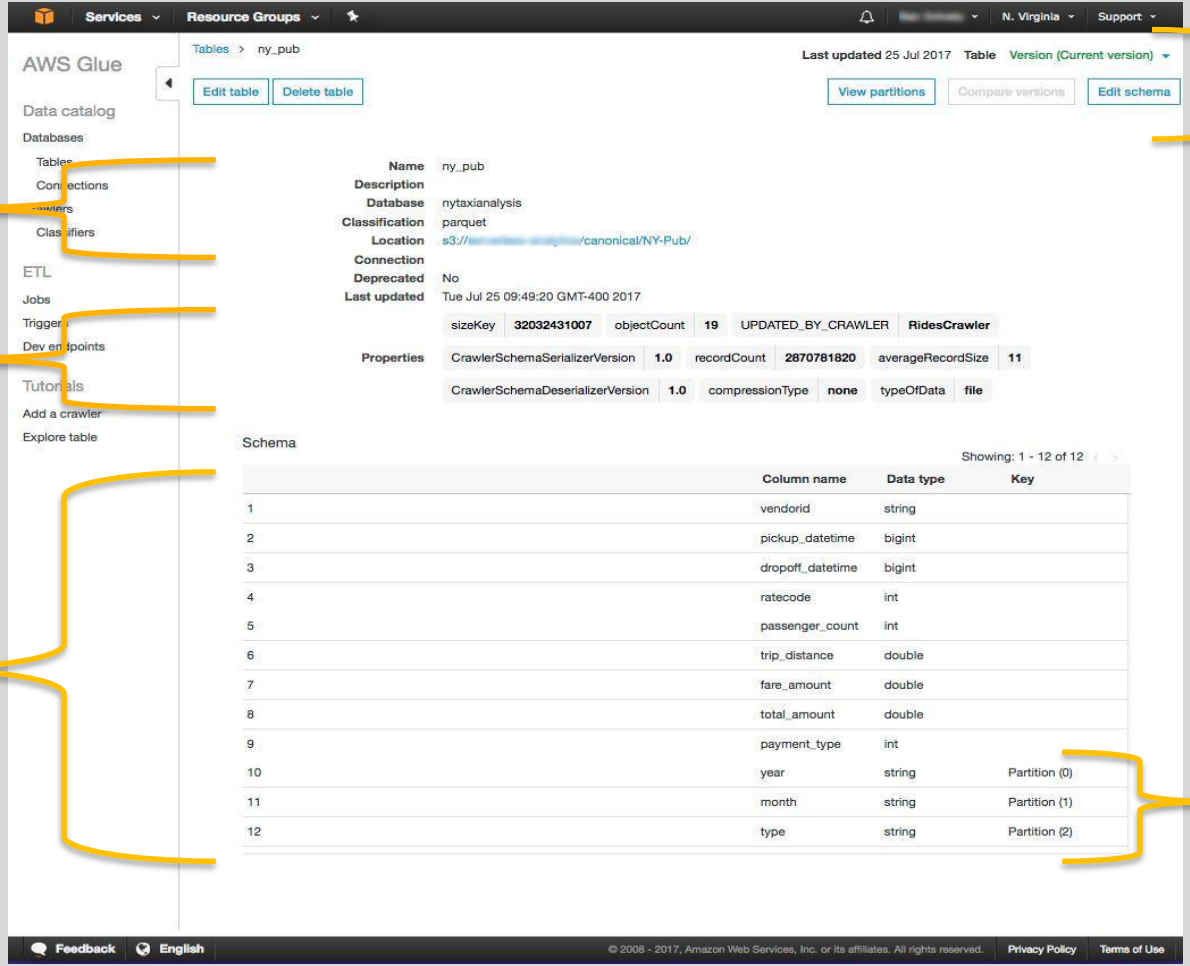


メタデータ（定義）の例

テーブルの
主要情報

テーブルの
プロパティ

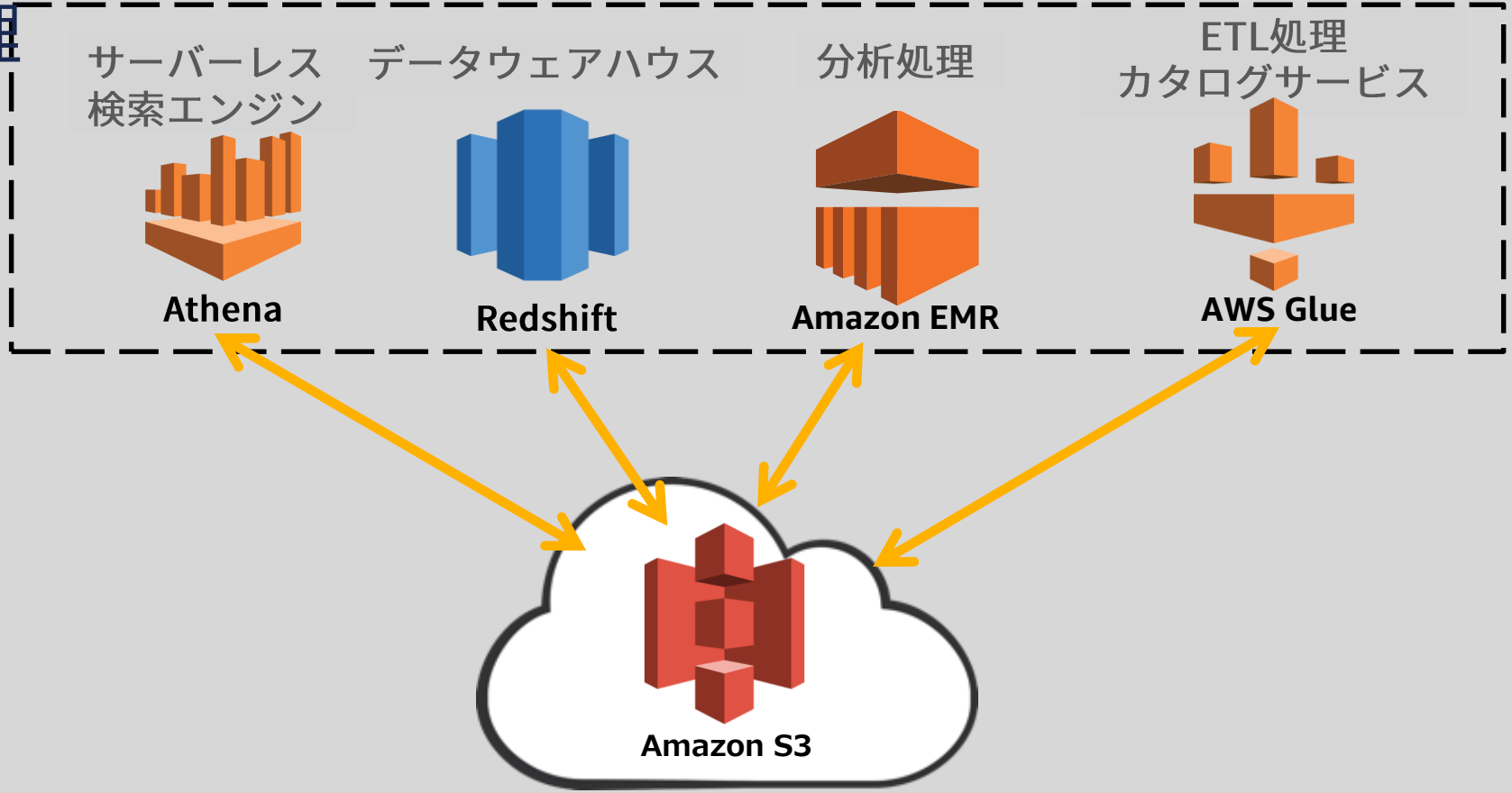
テーブル
スキーマ



テーブル
バージョン

テーブル
パーティション

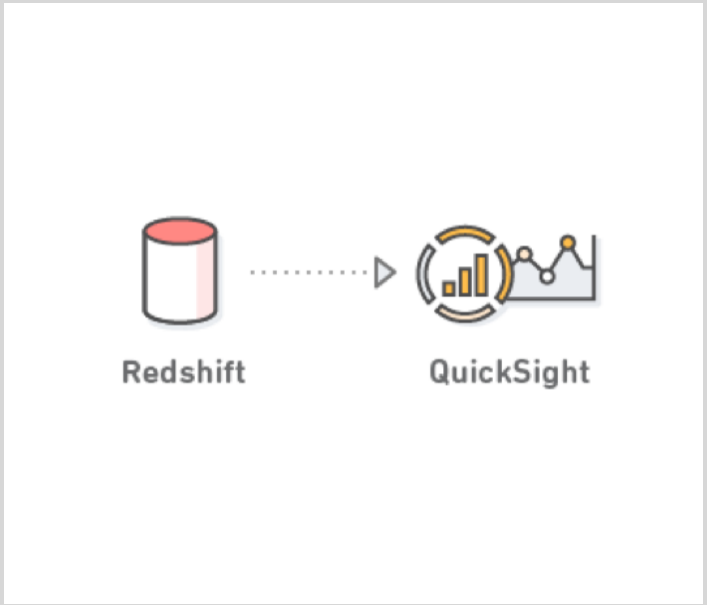
要件に合わせたサービスでデータレイクを分析処理



データ処理のパイプライン

Amazon QuickSight

フルマネージドな BI サービス



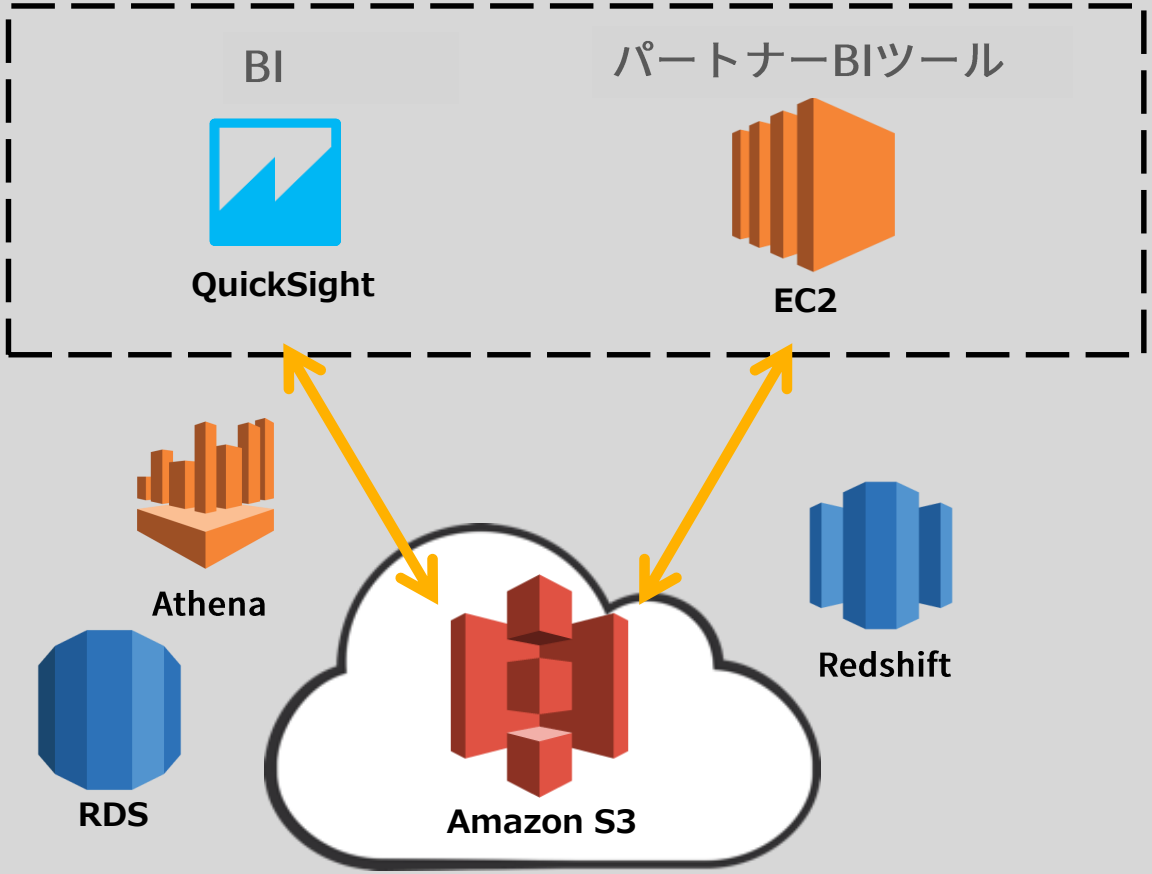
- １ユーザあたり$9/月からの安価な費用
- サーバレスでデータ分析
- ハイパフォーマンスなデータ処理エンジン SPICE を持ち、高速に分析が可能
- Redshift, RDS, S3, Athena, Salesforce, ローカルファイルなどさまざまなソースに接続可能
- AWS のデータソースをすばやく可視化して、組織で共有

パートナー様ソリューション

EC2+BIツール

- 多彩なパートナーソリューション・OSSをEC2上で活用

要件に合わせたツールでデータレイクを可視化

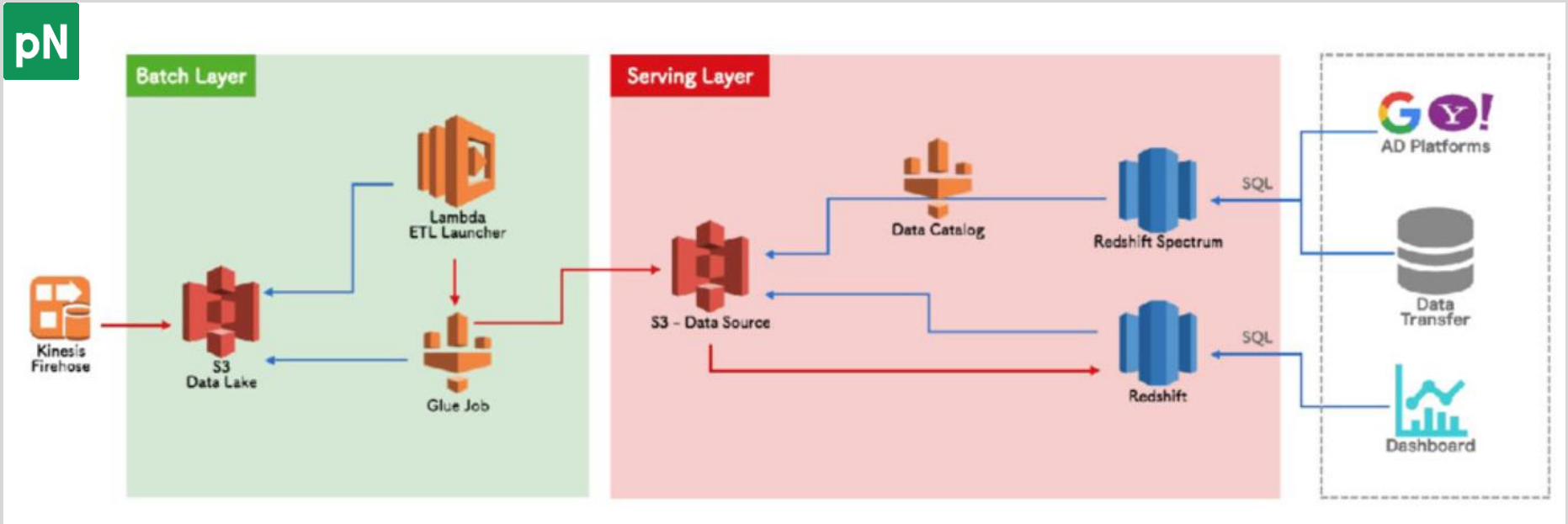


データ処理のパイプライン

アジェンダ

1. データ活用の流れ
2. 従来の RDB を中心としたデータ分析
3. データレイクとは
4. データレイクを構成する AWS サービス
5. AWS におけるデータ活用事例
6. まとめ

事例：primeNumber 様による DMP (Data Management Platform)



https://speakerdeck.com/hiro_koba_jp/aws-etlji-ri-aws-gluehuo-yong-shi-li-at-primenumber

アジェンダ

1. データ活用の流れ
2. 従来の RDB を中心としたデータ分析
3. データレイクとは
4. データレイクを構成する AWS サービス
5. AWS におけるデータ活用事例
6. まとめ

まとめ

- データは加工せずデータレイクに全て残す
- 全てを叶える万能なツールは存在しない
- AWS の様々なサービスを活用し、ストレージと計算処理の分離することで、データ活用に必要なツールやリソースの柔軟な確保を行い、高速な試行錯誤のサイクルを回すことができる

aws SUMMIT

Thank you !