

最新の DWH および データレイク動向について

鈴木 浩之

技術統括本部 ソリューションアーキテクト
Amazon ウェブ サービス ジャパン合同会社

自己紹介

鈴木 浩之
Hiroyuki Suzuki

技術統括本部 ソリューションアーキテクト
アマゾンウェブサービスジャパン合同会社



- 通信業界のエンタープライズ企業のご支援
- 様々な業界の Analytics 案件のご支援
- 好きなAWSサービス： Amazon Redshift、 AWS Glue

本セッションでお話しすること

- ・ モダンデータ戦略と目的別分析サービス
- ・ 進化したデータウェアハウス Amazon Redshift
- ・ まとめ

データ量の増加傾向

増え続けるデータはチャンスでありチャレンジ

20 年前の
1 年間に
生成されるデータ量



現在の
1 時間あたりに
生成されるデータ量

*Source: The Seagate Rethink Data Survey, IDC, January 2020

データ分析に関する様々な課題



データのサイロ化

- 利用したいデータを探せない
- 連携が煩雑ですぐに分析できない
- データの標準化ができていない
- 重複持ちなど無駄も多い
- アクセス管理が複雑化



分析ニーズの多様化

- 役割やスキルセットが多様化
- 使いたいツールや方法が違う
- リアルタイムに分析したい

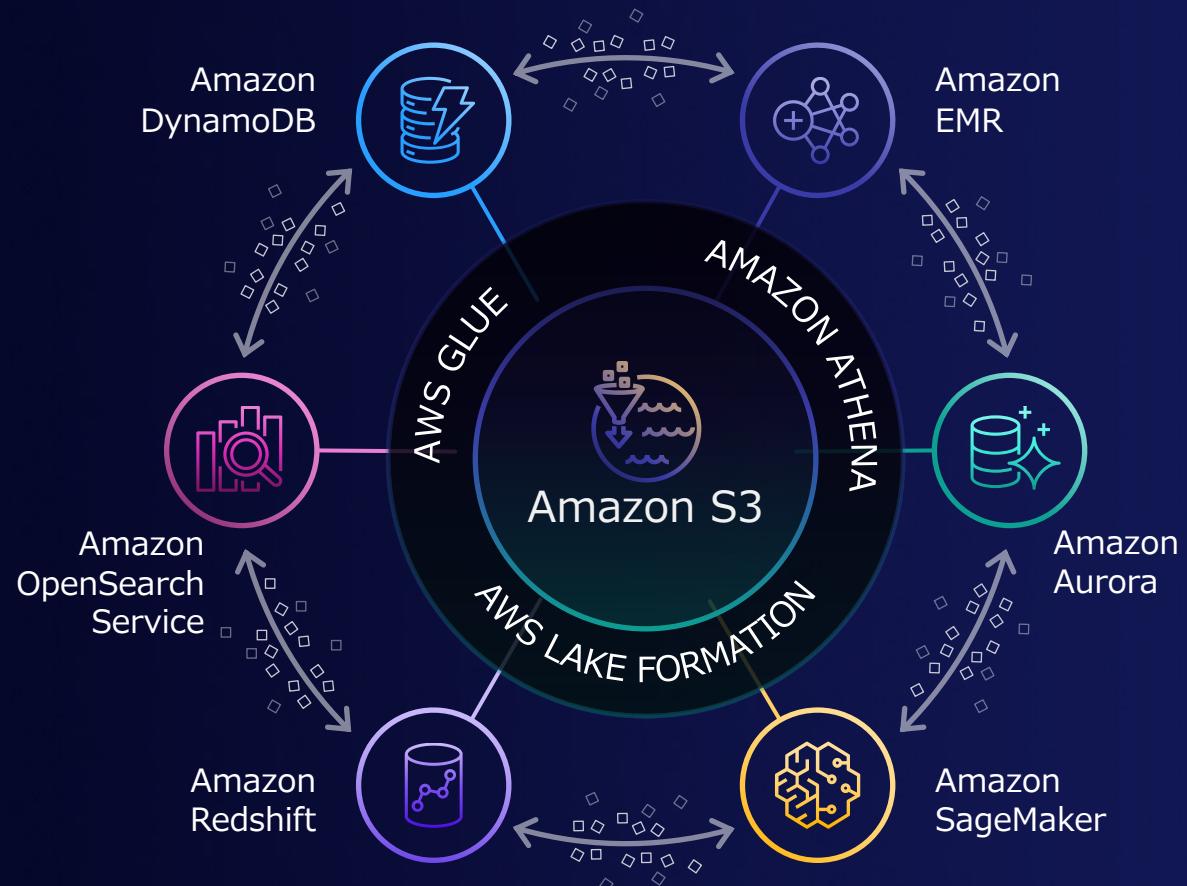


拡張性とコスト

- 増え続けるデータと負荷への追従が困難
- 業務とシステムの成長とともに管理も複雑化し、TCO も増加

モダンデータ戦略

大規模に、誰でも簡単に、様々なデータを分析



- スケーラブルなデータレイク
- ニーズに最適化された分析サービス
- 簡単に使える
- 統合アクセスとガバナンス
- ビルトイン機械学習

様々なニーズに最適化された分析サービス 例

ユーザー層

手段/IF

最適な分析サービス

エグゼクティブ
ビジネス部門

データアナリスト

サイエンティスト

データエンジニア

インフラエンジニア



GUI / BI / Dashboard

SQL (簡易)

SQL (複雑で高度)

Spark / Hive / Presto
Notebook etc.

Python / R
Notebook / IDE

可視化 (ログ)



Amazon QuickSight
BI・ダッシュボード



Amazon Athena
インタラクティブクエリ



Amazon Redshift
データ ウェアハウス



Amazon EMR
ビッグデータフレームワーク



Amazon SageMaker
機械学習



Amazon OpenSearchService
検索・可視化

サービスが増えると
セットアップや管理、コストが気になる。。。。



サーバーレスで解決

サーバーレスとは？



サーバーがない？

サーバーの存在を意識しない

- ✓ 自動化されたスケーリング（拡張と縮退）
- ✓ アイドル時間のリソース自動解放（コスト最適化）
- ✓ ビルトインされた高可用性
- ✓ 自動化されたメンテナンス

サーバーレスがもたらす効果



- 複雑なセットアップや設計不要
- 小規模から大規模まで自動スケール
- より少ない人数で構築・運用
- 使った時だけ課金
- 分析に集中

サーバーレスの選択肢の拡大

大規模に、誰でも簡単に、様々なデータを分析

収集



保存



変換



分析



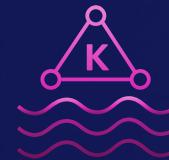
サーバーレスの選択肢の拡大

大規模に、誰でも簡単に、様々なデータを分析



Amazon Kinesis Data Streams On-Demand

- ストリーミングデータサービス
- オンデマンドモードで、ワークロードに応じて自動スケール



プレビュー

Amazon Managed Streaming For Apache Kafka Serverless

- フルマネージド Apache Kafka



プレビュー

Amazon EMR Serverless

- Apache Spark, Apache Hiveなどのビッグデータプラットフォーム



プレビュー

Amazon Redshift Serverless

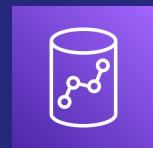
- 大規模データクエリのためのクラウドデータウェアハウス

ユースケース別 サーバーレスで実現するデータ分析基盤 例

サーバーレスで実現するデータ分析基盤 例

少人数でも簡単に大規模分析

分析者



Amazon Redshift
Serverless



Amazon S3

データレイク

サーバーレスで実現するデータ分析基盤 例

ニーズに最適なサービスで

分析者



目的別分析サービス



Amazon EMR
Serverless



Amazon Redshift
Serverless



Amazon
QuickSight

その他



Amazon S3



AWS
Lake Formation

データレイク

サーバーレスで実現するデータ分析基盤 例

様々なデータソースと連携

分析者

データ
ソース

データ
収集

目的別分析サービス



Amazon
AppFlow



AWS Glue



Amazon Kinesis
Family

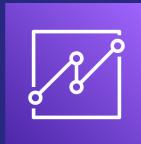
その他



Amazon EMR
Serverless



Amazon Redshift
Serverless



Amazon
QuickSight



Amazon S3



AWS
Lake Formation

データレイク



サーバーレスで実現するデータ分析基盤 例

データメッシュによる連携

分析者



データソース

データ収集



Amazon AppFlow



AWS Glue



Amazon Kinesis Family

その他

目的別分析サービス



Amazon EMR
Serverless



Amazon Redshift
Serverless



Amazon
QuickSight

その他



Amazon S3



AWS
Lake
Formation

データレイク

他ドメインのデータレイク



大規模データクエリのための目的別分析サービス Amazon Redshift

大規模に、誰でも簡単に、様々なデータを分析



増え続けるデータはチャンスでありチャレンジ



Amazon
Redshift

誰でも簡単に
分析できる

インフラストラクチャを気にすることなく、分析に集中

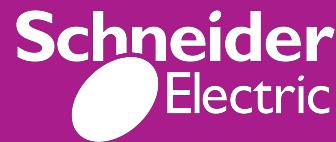
すべてのデータを
分析する

様々なデータソースに対してリアルタイムにインサイトを取得。

大規模環境での
パフォーマンス

動的にスケーリングして複雑で重要なワークロードのクエリ速度を向上

毎日何万ものお客様が Amazon Redshift で エクサバイトのデータを処理しています



Redshift の同時実行スケーリングと RA3 ノードを通じて数万人のユーザーをサポートし、グリーンな未来をサポート



Amazon S3 と Amazon Redshift の柔軟性とスケラビリティにより、1 日あたり 300 億件のレコードから 700 億件のレコードへの急増に対応



Redshift ML による予測分析により、将来の医薬品コストを予測し、傾向を特定



ETL のパフォーマンスが 2 倍に向上し、5.3 TB を超える日次ゲームデータを処理するようにスケーリングされました



20~30億件の求人検索の推奨に対して15~20分でモデル推論を実行
追加費用なしで2~3時間短縮



Easy analytics for everyone



誰でも簡単に
分析できる

お客様の声



Easy analytics for everyone

誰でも簡単に
分析できる

| インフラを気にせず
大規模な分析をしたい

| 自分に合った分析ツール・方法を
サクッと使いたい

| 高度な分析を
誰でも簡単にできるようにしたい

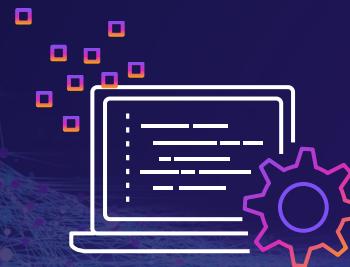


Amazon Redshift Serverless

データウェアハウスのインフラストラクチャの管理不要で
データからインサイトを数秒で取得



データからインサイトを
より簡単に取得

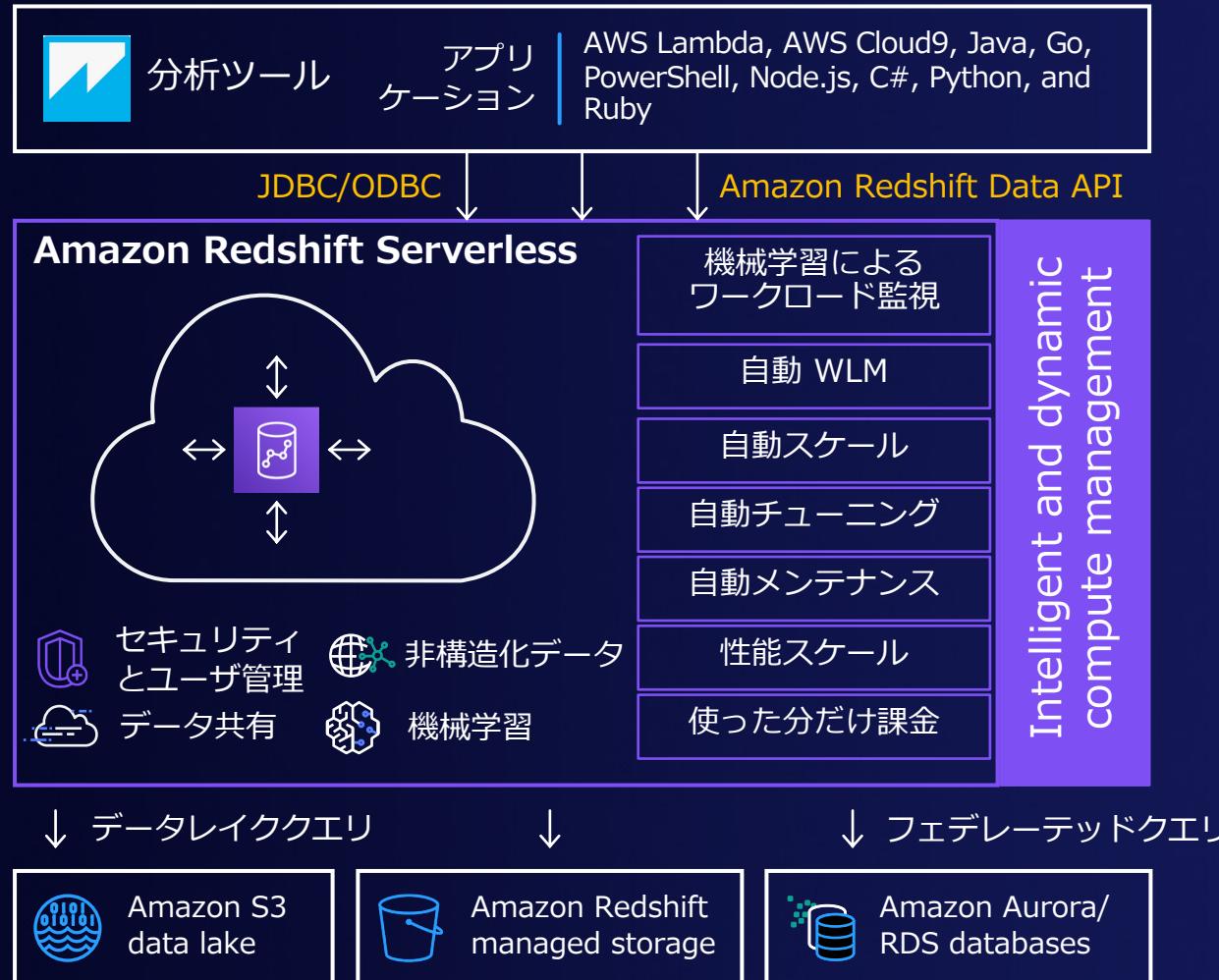


一貫して高い
パフォーマンスを実現



コストを最適化

Amazon Redshift Serverless



- シンプルにクエリ実行のためのエンドポイントへ接続するだけ
- 基本的に従来のプロビジョン型 Redshift の機能を踏襲
- 自動スケールと自動チューニングで事前の設計やサーバ運用管理が不要
- 費用は使った分だけ
 - ✓ クエリワークロード実行時のコンピューティング (RPU) 利用時間単位 + 蓄積データ量 (GB) の時間単位
 - ✓ 1 RPU = 2 vCPU, 16 GiB メモリで、RPU は 32 ~ 512 までスケール可能

各種クエリサービスの使い分け



Amazon
Athena



Amazon
Redshift
Serverless

NEW



Amazon
Redshift
Provisioned

用途

簡易な処理

簡易～複雑な処理

(例：複数の結合やサブクエリ)

簡易～複雑な処理

(例：複数の結合やサブクエリ)

規模

低頻度または予測不能
ワークロード
(小～中規模)

低頻度または予測不能
ワークロード
(小～大規模)
最大 RPU 512

予測可能
ワークロード
(中～超大規模)
最大 ra3.16xlarge 128 ノード

基盤管理

自動

自動

プロビジョニング
詳細なチューニング可

費用

クエリスキャン量
(別途ストレージサービス使用料)

ワークロード実行時間
RMS ストレージ使用量

クラスタ稼働時間
(リザーブドインスタンス適応可能)
RMS ストレージ使用量
クエリスキャン量
(Spectrum 利用時)



お客様の声



Easy analytics for everyone

誰でも簡単に
分析できる

| インフラを気にせず
大規模な分析をしたい

| 自分に合った分析ツール・方法を
サクッと使いたい

| 高度な分析を
誰でも簡単にできるようにしたい

セットアップ不要のRedshift用クエリエディタ

誰でも簡単に分析できる

Amazon Redshift Query Editor v2

The screenshot shows the Amazon Redshift Query Editor v2 interface. On the left, there's a sidebar with navigation links for Database, Queries, Notebooks (Preview), and Charts. The main area has tabs for Cluster, Serverless (admin), Database, and sample_data_dev. A query editor window displays a complex SQL query involving multiple tables and subqueries. Below the query is a results table titled "Result 1 (9)" showing two rows of data. A context menu is open over the results table, with "CSV" highlighted. To the right of the results is a pie chart with various categories and percentages.

- セットアップ不要 (サーバーレス)
- 各種シングルサインオンプロバイダと統合したダイレクト接続
※マネージメントコンソールログイン不要
- データの直接インポート (CSV) / エクスポート (CSV / JSON) 可能
- クエリエディタ利用のための追加料金はなし

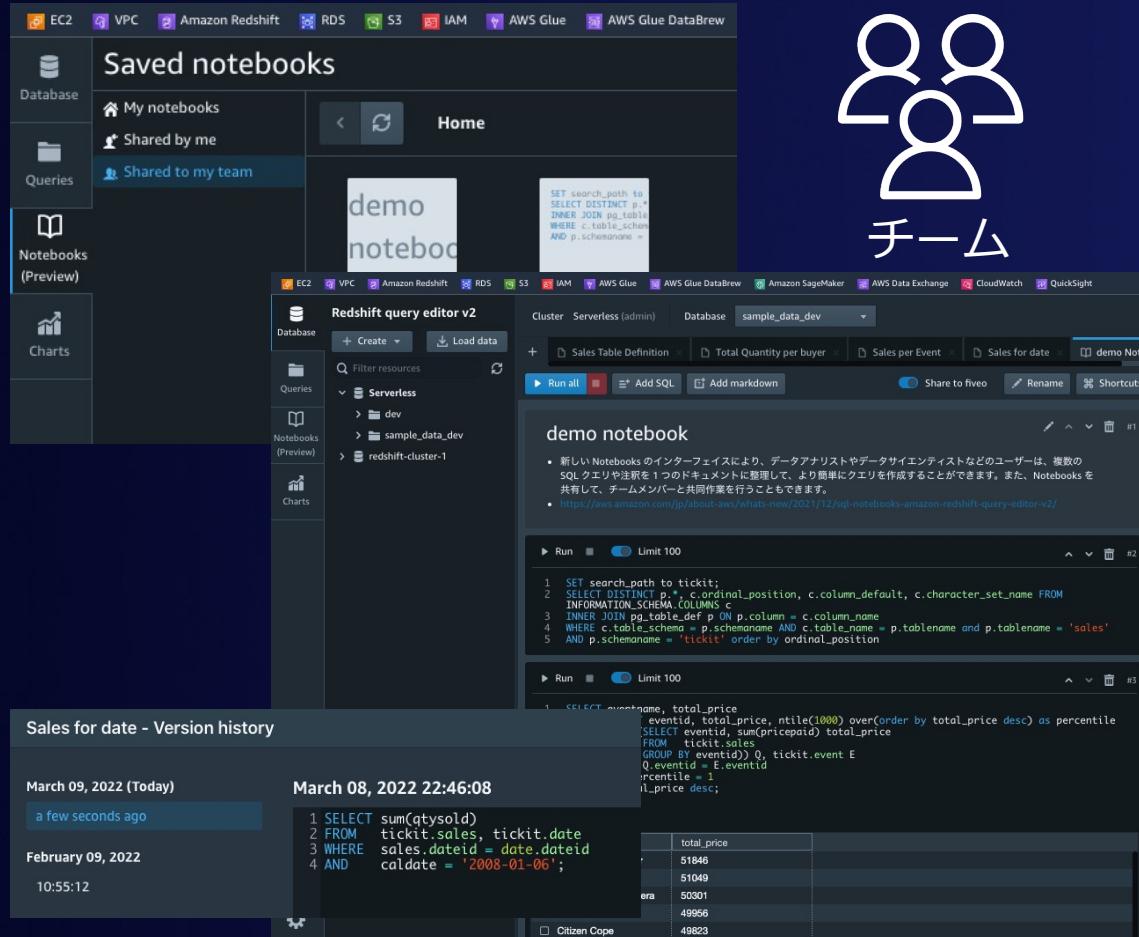
An Overview of Amazon Redshift Query Editor V2
<https://www.youtube.com/watch?v=IwZNIroJUnc>



セットアップ不要のRedshift用クエリエディタ

チームで共有

Amazon Redshift Query Editor v2

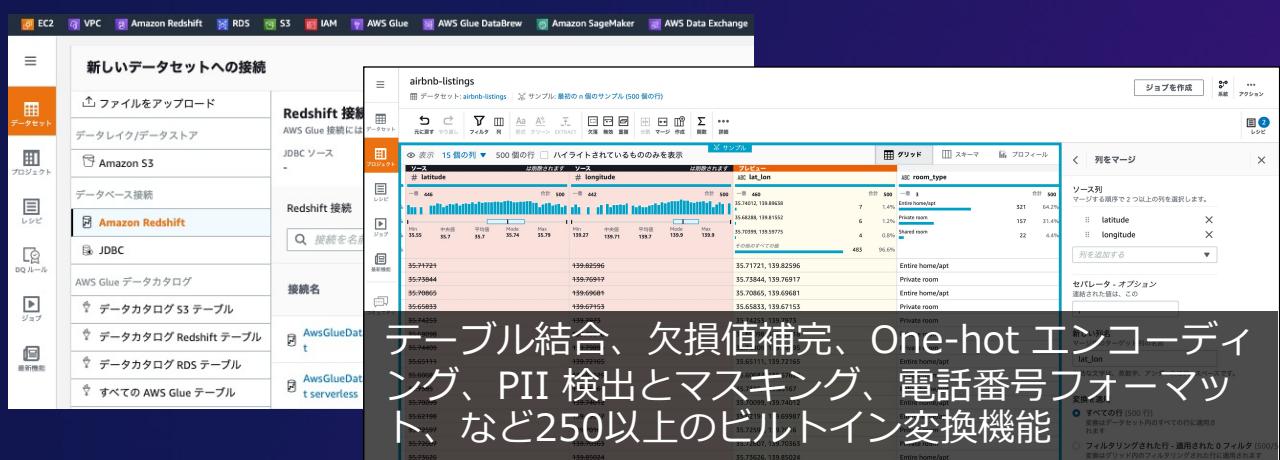
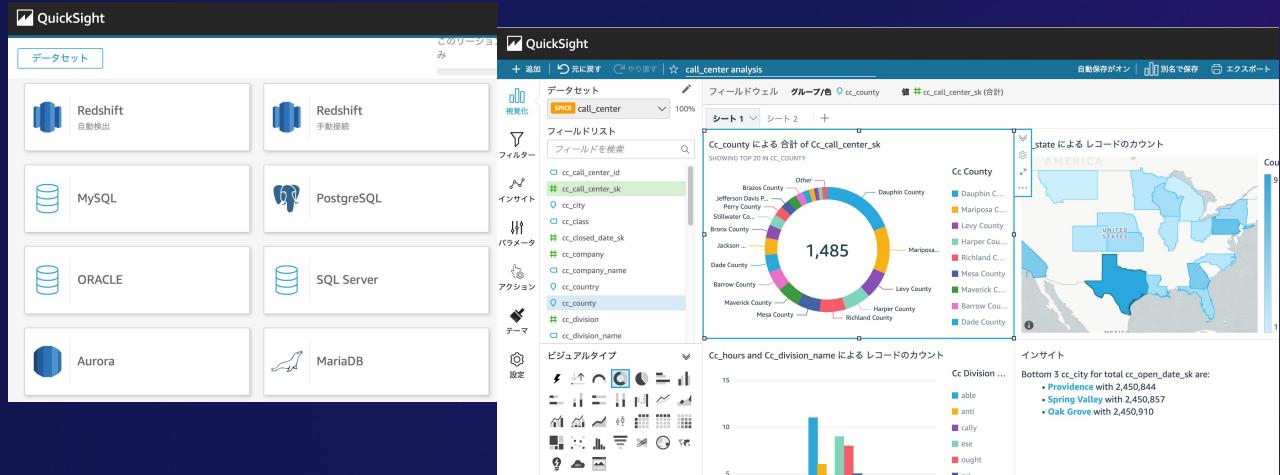
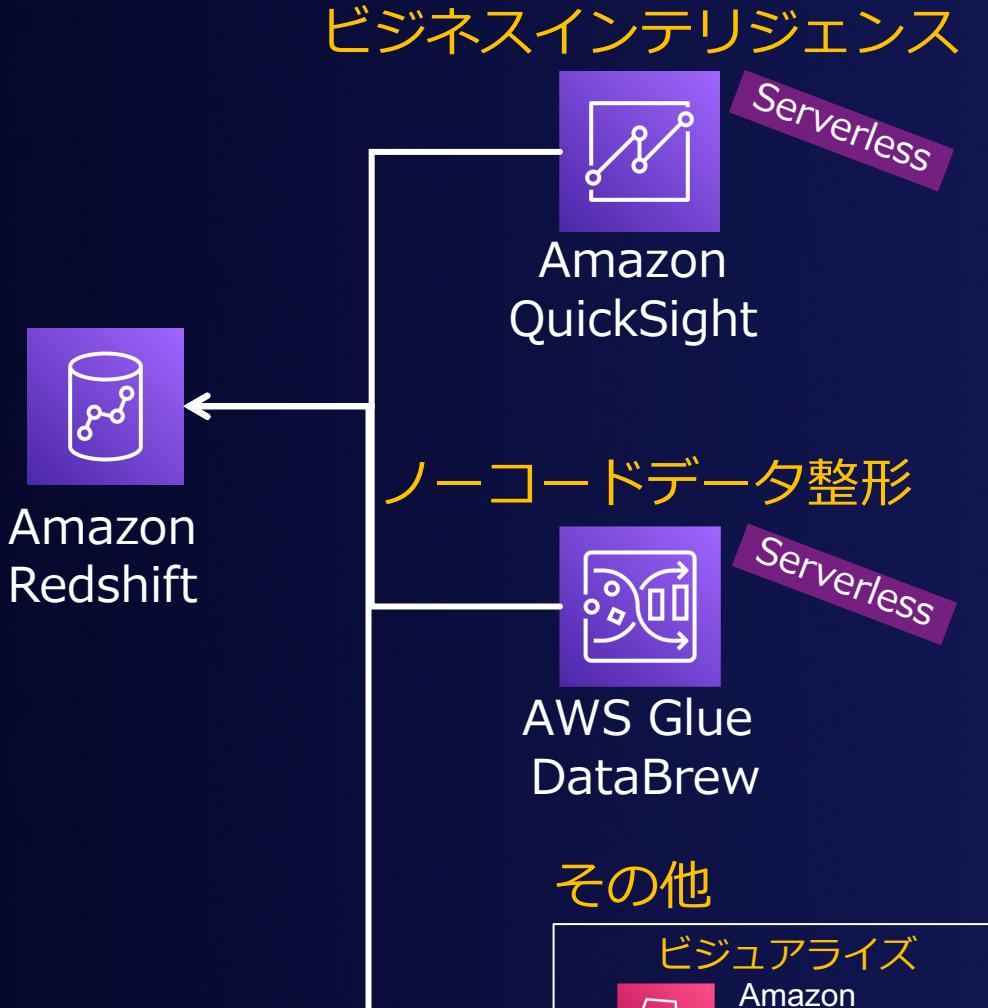


チーム

- 標準タブのクエリエディタの他に、複数の SQL や Markdown を含めた Notebook ◀ プレビュー
- 各ユーザが作成したクエリやチャート(グラフ)、Notebook は特定のチーム内で共有できる
- クエリ資産は変更履歴を自動管理できる



最適化された専用サービスと連携 例



3rdパーティ サービス

などなど

お客様の声



Easy analytics for everyone

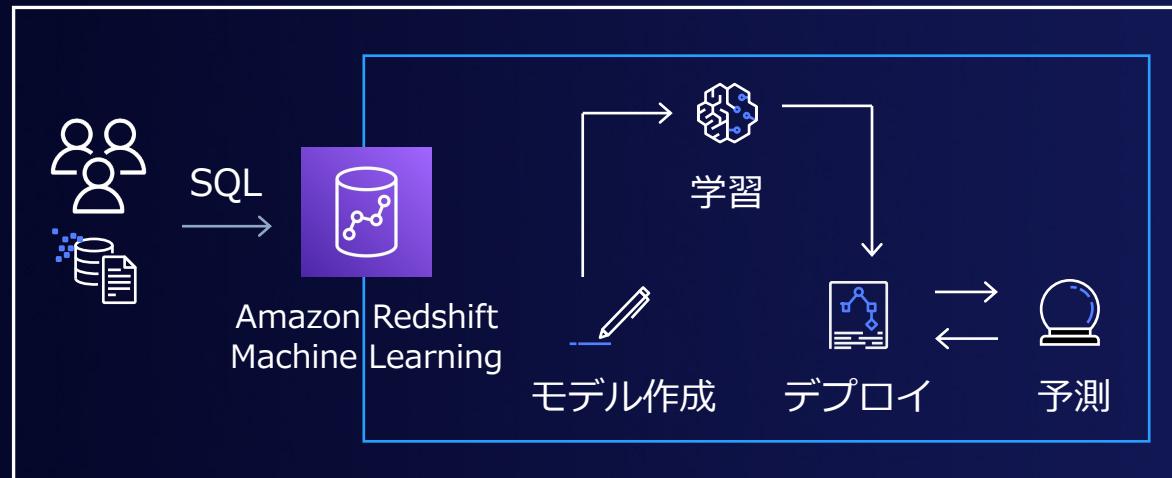
誰でも簡単に
分析できる

| インフラを気にせず
大規模な分析をしたい

| 自分に合った分析ツール・方法を
サクッと使いたい

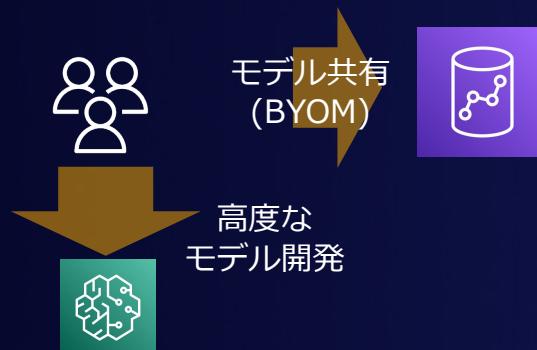
| 高度な分析を
誰でも簡単にできるようにしたい

AI の民主化を加速：使い慣れた SQL で機械学習



ユースケース例

データサイエンティスト



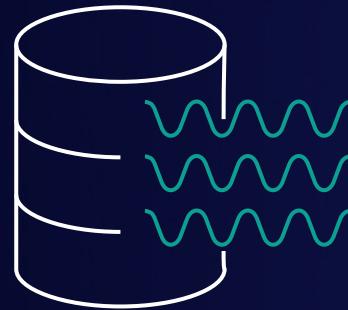
Amazon SageMaker

データアナリスト



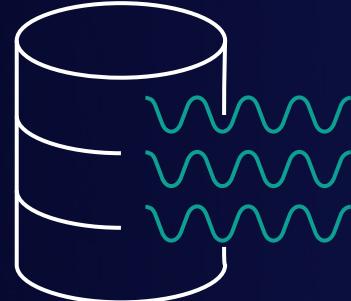
- Amazon Redshift のインターフェースから SQL のみで機械学習モデルの作成・学習・予測可能
- 教師あり学習(回帰・バイナリ分類・マルチクラス分類)と、教師なし学習(K-Means クラスタリングアルゴリズム)をサポート
- Amazon SageMaker Autopilot と統合された Auto ML 機能
- 独自のモデルをインポートして利用も可能





Analyze all your data
すべてのデータを
分析する
～シームレスなデータ連携

お客様の声



Analyze all your data

すべてのデータを
分析する

～シームレスなデータ連携

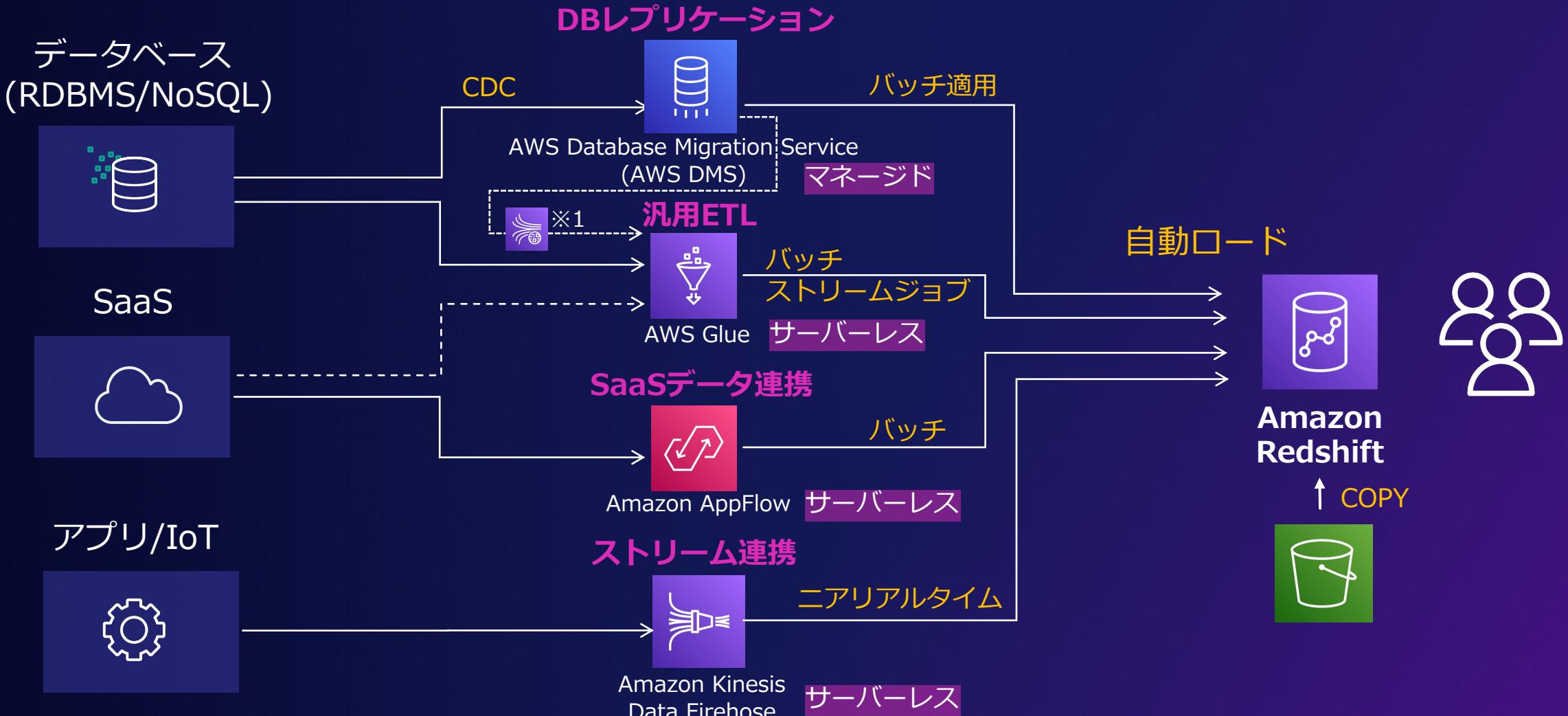
| 様々なデータをラクに取り込みたい

| 外部データに直接アクセスしたい
※データ連携を待てない

| そもそもデータを持っていない

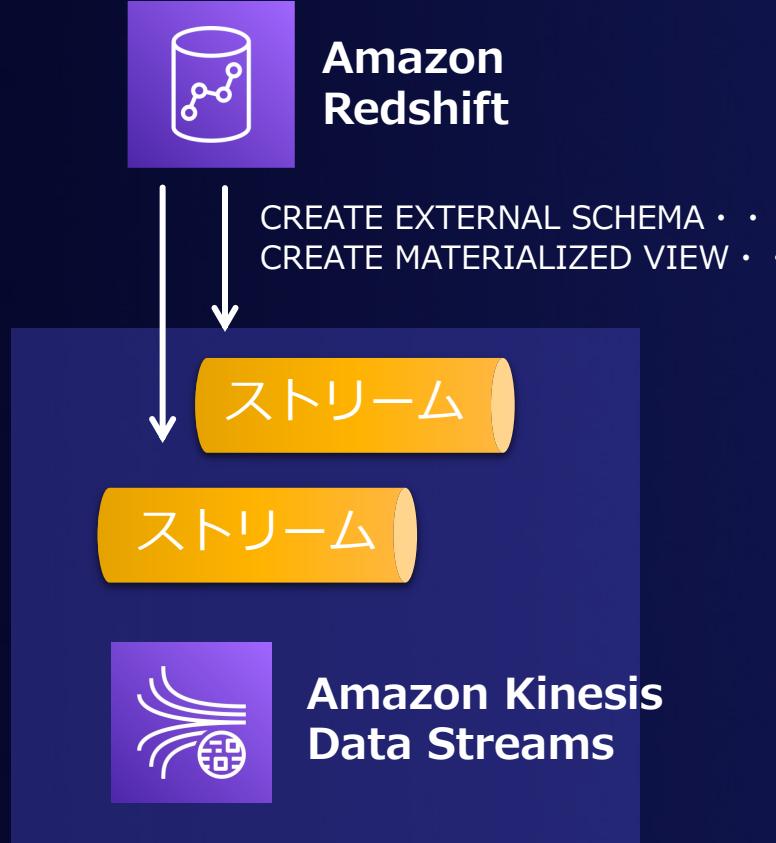
他の様々なデータソースからの取り込み

AWS サービス以外のデータソースからもサーバーレスサービスで簡単に



リアルタイムデータ連携

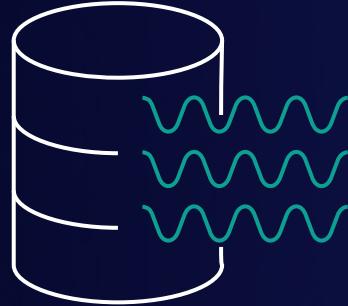
Kinesis データストリームのストリーミング取り込み



- 1 秒あたり数百 MB のストリーミングデータを取り込みながら、低レイテンシークエリを実現
- 従来の S3 COPY 経由をベースとした Kinesis Data Firehose などのニアリアルタイム方式より、**簡易的かつ低遅延な連携**
- ストリームの上にマテリアライズドビューを直接作成できるようにすることで、**データパイプラインを簡素化**
- 複数の Kinesis データストリームに同時に接続してデータを直接取り込み可能

<https://aws.amazon.com/jp/about-aws/whats-new/2022/02/amazon-redshift-public-preview-streaming-ingestion-kinesis-data-streams/>
<https://aws.amazon.com/jp/blogs/big-data/integrate-etl-with-amazon-redshift-streaming-ingestion-preview-to-make-data-available-in-seconds/>

お客様の声



Analyze all your data
**すべてのデータを
分析する**
～シームレスなデータ連携

| 様々なデータをラクに取り込みたい

| 外部データに直接アクセスしたい
※データ連携を待てない

| そもそもデータを持っていない

他の様々なデータソースへのクエリ

事前のデータ移動不要でライブデータへアクセス

リレーショナルデータベース



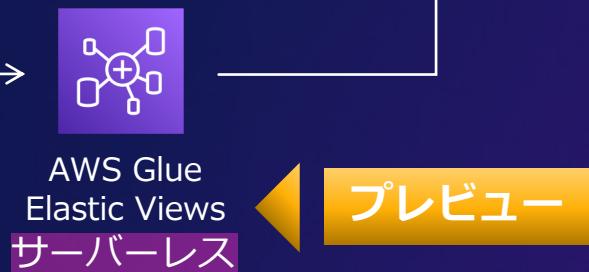
データレイク



NoSQLデータベース

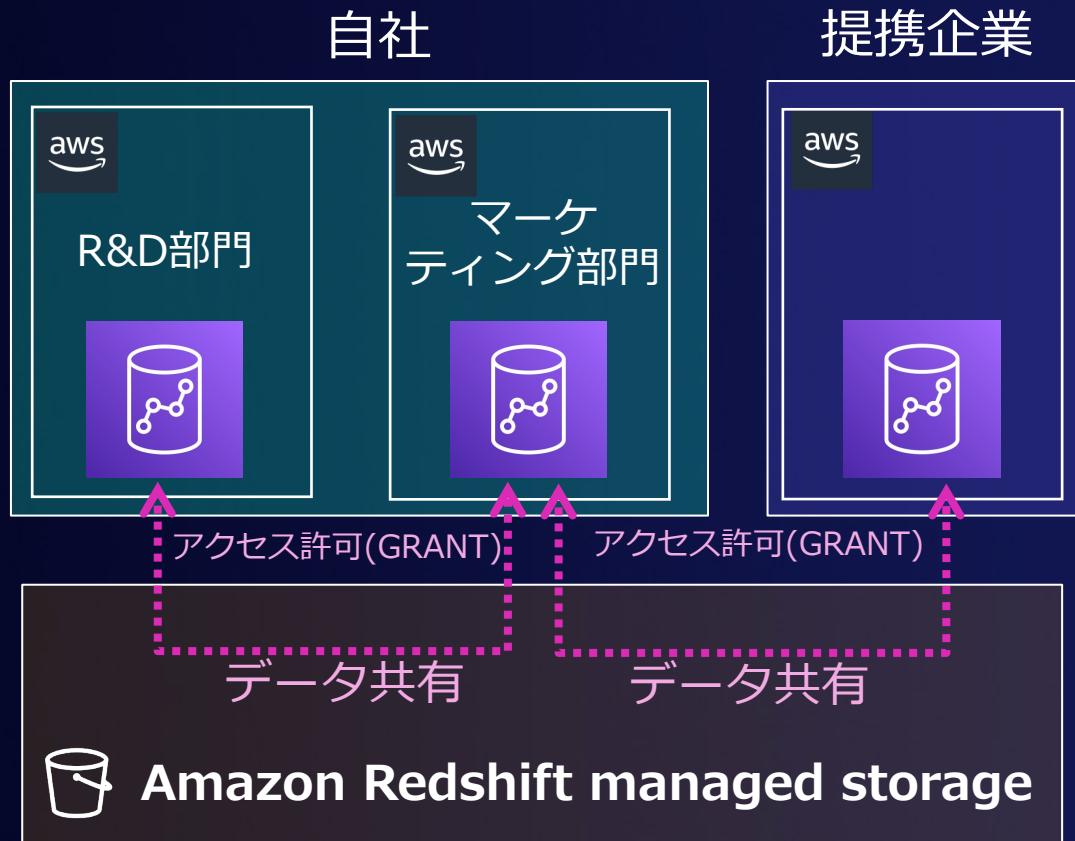


マテリアライズドビュー



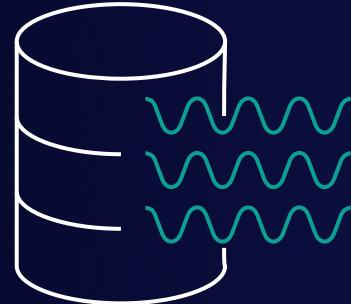
組織や企業を跨いだデータ共有

事前のデータ移動不要でライブデータへアクセス



- Redshift 同士のデータを、データコピー不要で相互に共有可能（データ重複持ちや分割損の排除）
- スキーマ単位、テーブル単位、UDF 単位で共有アクセス制御
- Serverless 型と Provisioned 型(RA3)でサポート
- アカウント間、リージョン間の共有
- データ共有のパフォーマンス強化※

お客様の声



Analyze all your data
**すべてのデータを
分析する**
～シームレスなデータ連携

| 様々なデータをラクに取り込みたい

| 外部データに直接アクセスしたい
※データ連携を待てない

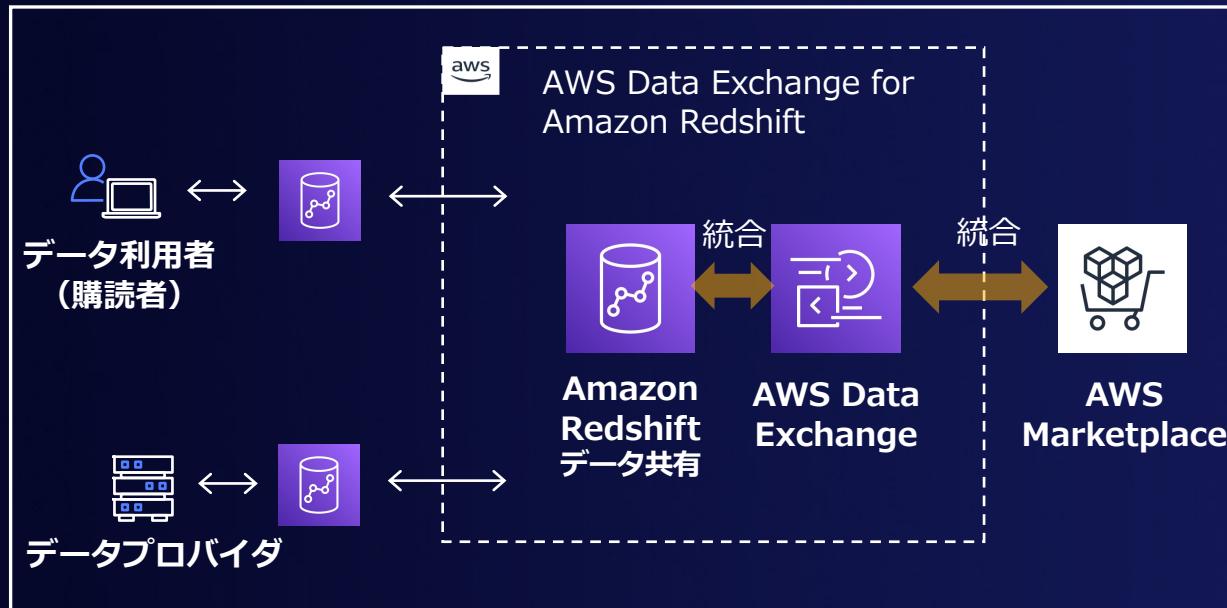
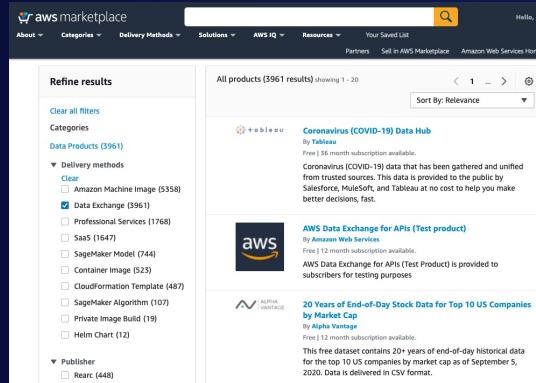
| そもそもデータを持っていない

3rd パーティーデータをマーケットプレイスから

事前のデータ移動不要でライブデータへアクセス



AWS Marketplace



- AWS Marketplace で他社のデータをサブスクリプション、または自社のデータを販売
- 決済と課金制御は自動化
- Amazon Redshift データ共有に対応したデータセットは、サブスクリプション開始後、データの移動や事前コピー不要ですぐに利用可能



Best price performance at any scale
**大規模環境での
パフォーマンス**

自動化されたパフォーマンスチューニング



Automatic
vacuum
delete



Automatic
distribution keys



Automatic
sort keys



Auto
workload
manager



Automatic
table sort



Automatic column
compression
encoding



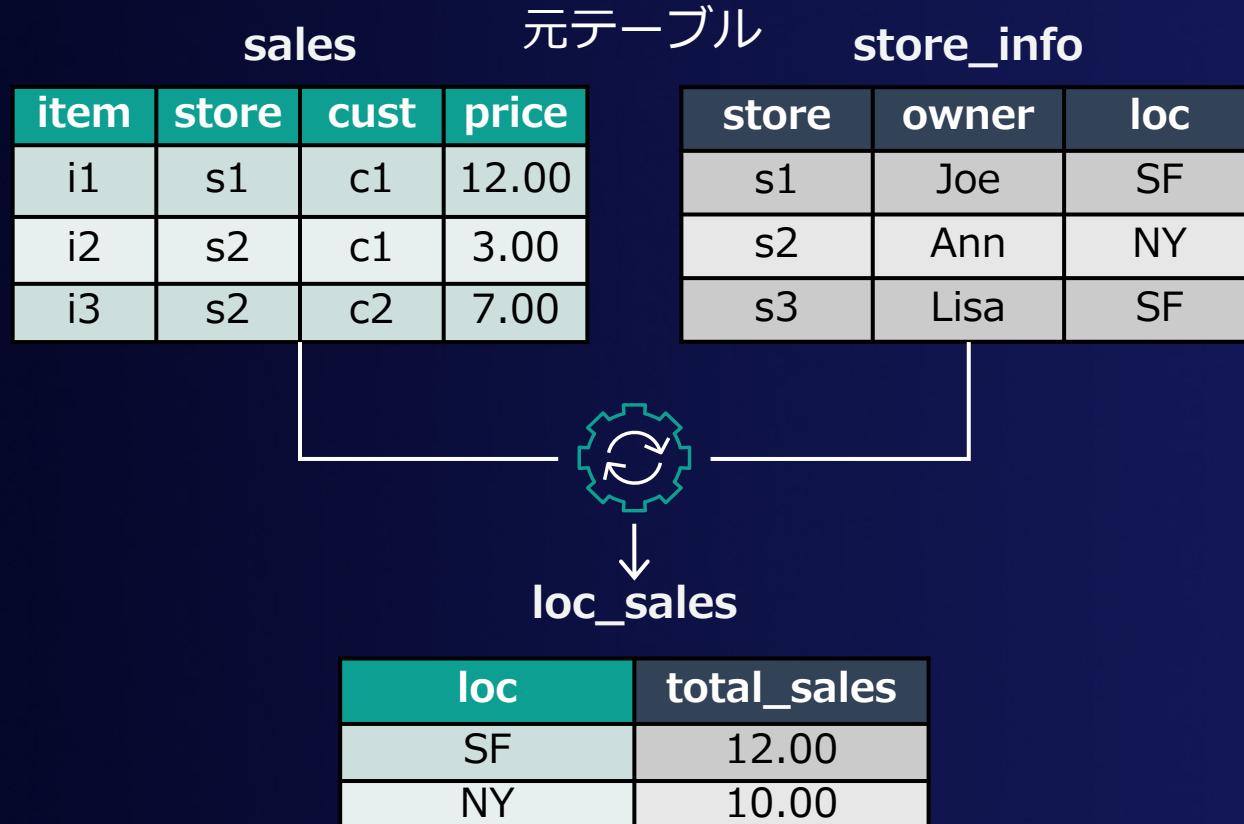
Auto Analyze



Auto refresh & rewrite
Materialized Views

- 物理データ配置やその最適化を自動化
 - ✓ インデックスの作成と運用不要。分散キー・ソートキー・圧縮エンコードの自動最適化。
 - ✓ VACUUM DELETE・ANALYZE・テーブルソートの自動実行
- 機械学習アルゴリズムを用いてクエリを実行するために必要なりソースを動的に管理：ワークロード管理 (WLM)

自動マテリアライズドビュー



自動生成されるマテリアライズドビュー

- 頻繁に実行されるクエリの自動的なパフォーマンス向上
- 継続的にワークフローをモニタし、機械学習によりパフォーマンス面で最適なマテリアライズドビューを自動的に生成・管理
- クエリリライトと自動リフレッシュ機能と合わせ、ユーザは意識せずに最も高速なクエリ結果を得られる

その他 分析に役立つ様々な機能アップデート

空間データ (spatial data)

- GEOMETRY 型と GEOGRAPHY 型の主要な空間データ型と豊富な空間関数で、緯度経度情報を使用した超大容量データの高速な分析が可能

空間関数

<https://docs.aws.amazon.com/redshift/latest/dg/geospatial-functions.html>

PIVOT / UNPIVOT

- 簡単な SQL 操作でテーブルのピボット操作（縦持ち横持ち変換）が可能

PIVOT と UNPIVOT 例

https://docs.aws.amazon.com/ja_jp/redshift/latest/dg/r_FROM_clause-pivot-unpivot-examples.html

SUPER型：半構造化データ

- JSON などの半構造化データを事前のパース処理不要で取り込み、高速な分析が可能。SQL 互換のクエリ言語 PartiQL により、スキーマレスのネストされたデータへも効率的アクセス可能

SUPER型

<https://aws.amazon.com/blogs/big-data/work-with-semistructured-data-using-amazon-redshift-super/>

スカラーLambda UDF

- AWS Lambda で定義されたカスタム関数を SQL クエリの一部として使用可能。Redshift の標準機能で実現できないことも SQL で呼び出す関数として拡張可能

UDFユースケース

<https://docs.aws.amazon.com/redshift/latest/dg/udf-example-uses.html>



まとめ

- モダンデータ戦略と、様々なニーズに最適化された目的別分析サービスを簡単に組み合わせて柔軟に拡張
- Analytics サービスへのサーバーレスの選択肢の拡大。運用管理から解放され、分析に集中できる
- Amazon Redshift で誰でも簡単に、様々なデータを、大規模にスケールして分析できる

Thank you!

鈴木 浩之 (Hiroyuki Suzuki)

ソリューションアーキテクト
アマゾンウェブサービスジャパン合同会社



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.