



最先端のデータ分析基盤を 実現する AWS のサービスと活用方法

2021年9月28日

アマゾン ウェブ サービス ジャパン株式会社
データベース パートナー ソリューション アーキテクト
吉田 成利



自己紹介

吉田 成利 (Michitoshi Yoshida)

ISV パートナー本部
データベースパートナーソリューションアーキテクト

データベースを専門に、パートナー様の
トレーニングや技術支援を担当

好きな AWS のサービス：

Amazon Relational Database Service

Amazon Aurora, Amazon S3



アジェンダ

- データレイクを中心とした分析環境
- AWS クラウドで実現するデータ活用基盤

データレイクを中心とした 分析環境

データ活用のトレンドと課題



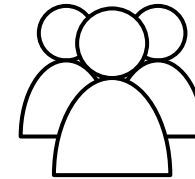
急拡大する
データ量



新しく増える
データソース



多様化するデータ



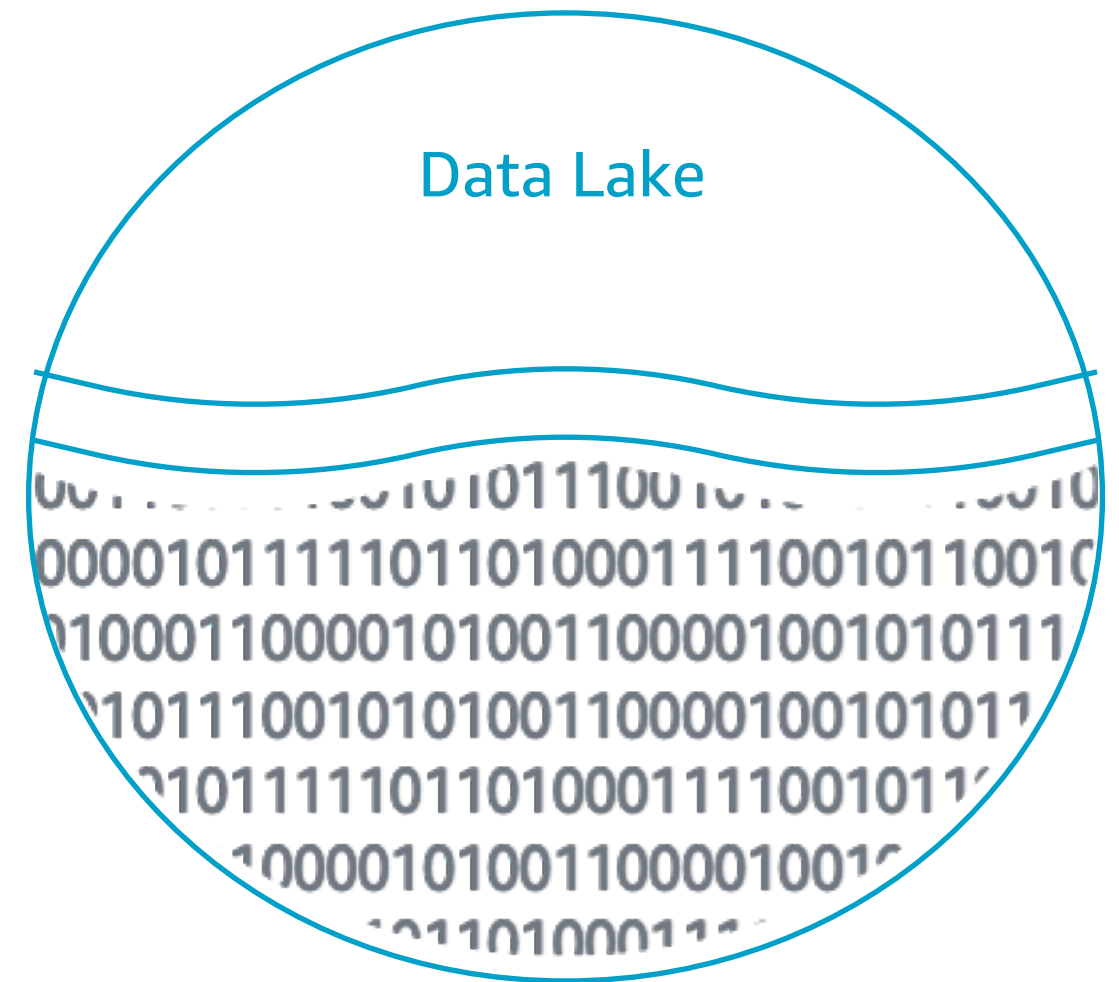
増大する利用者数
広がる利用者層



発展する
ユースケース

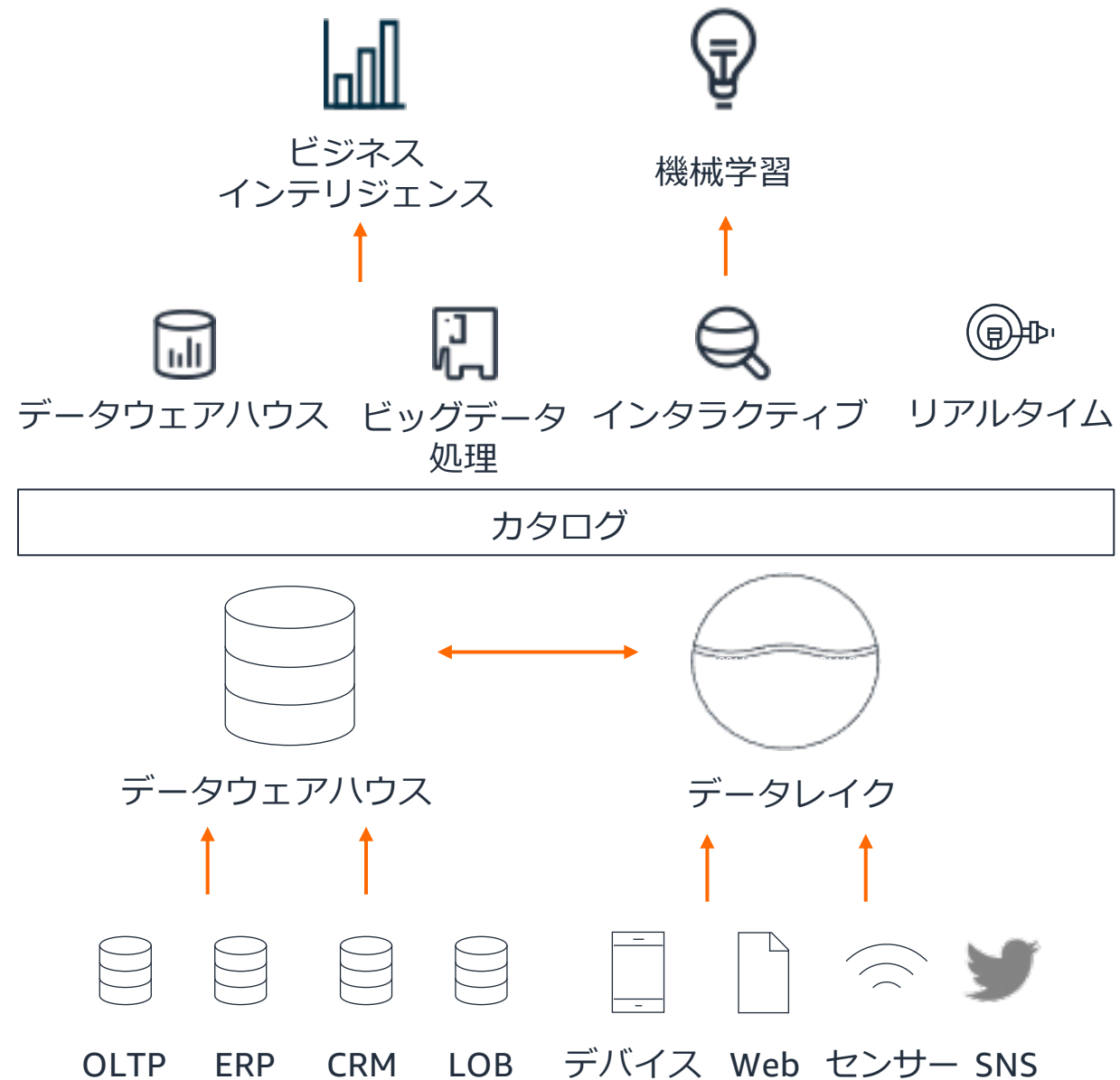
データレイク

データレイクは、規模にかかわらず、すべての構造化データと非構造化データを保存できる一元化されたリポジトリ。
データをそのままの形で保存



<https://aws.amazon.com/jp/big-data/datalakes-and-analytics/what-is-a-data-lake/>

データ活用の全体像



データレイクが提供するもの:

構造化、半構造化、非構造化データの取り扱い

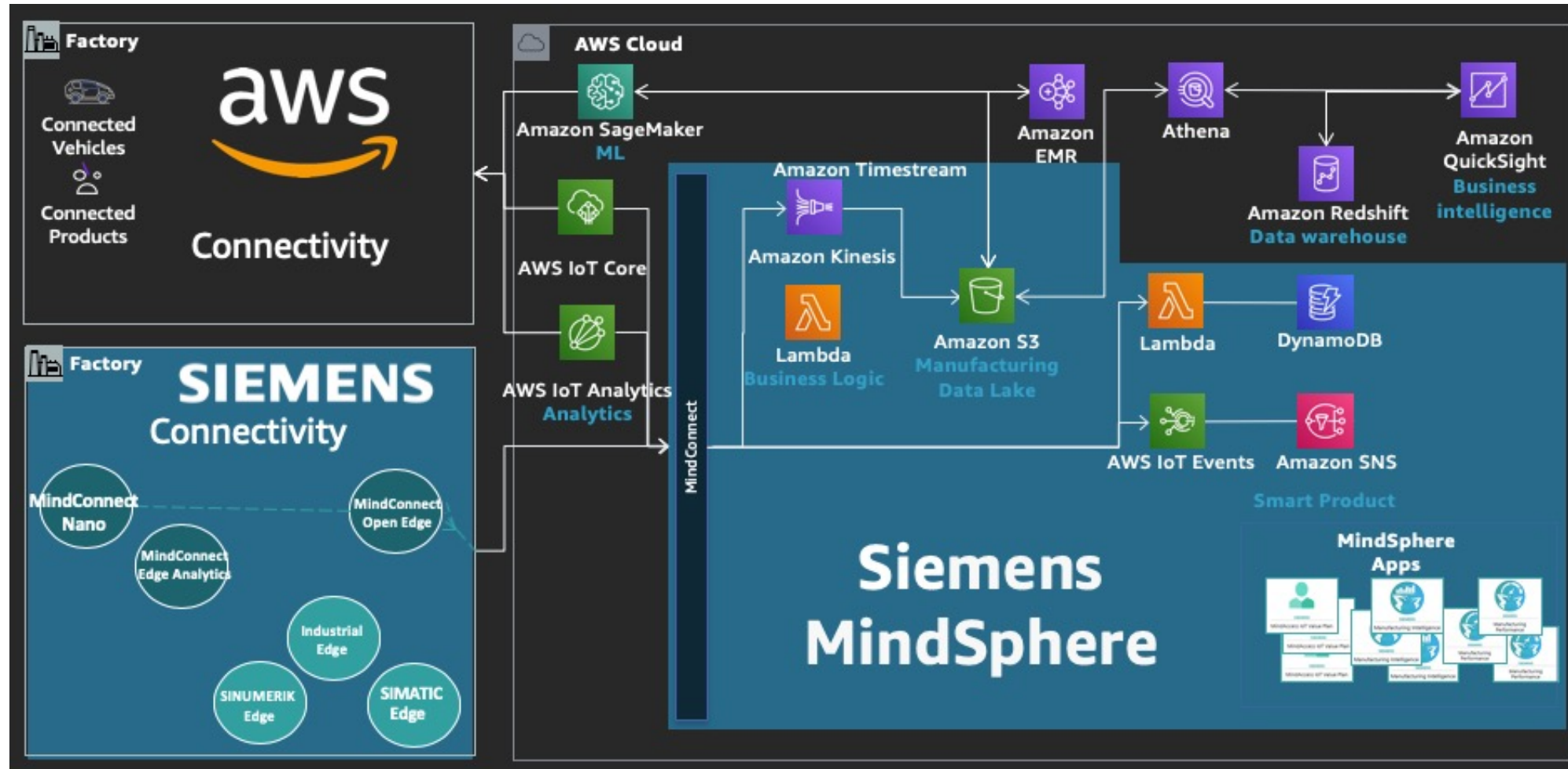
ペタバイト、エクサバイトにわたる拡張性

様々な分析および機械学習ツールとの連携

低コストなデータの保存と分析

Siemens 様：リアルタイムに生産ラインの設備効率を表示・改善

モンテレー工場をインダストリー4.0のモデルとして、リアルタイムに生産ラインの総合設備効率（OEE）を可視化



- 過去の生産データから不具合のあるテストマシンを修正、初回パスの歩留まりを3%向上。
- 8週間未満で構想から完全な生産に移行。
- プラントのOEEは40～50%、計画が工場全体に導入されると、チームの目標は60%に、最終的には、85%到達を目指す。

https://aws.amazon.com/jp/solutions/case-studies/siemens-mindsphere/?trk=manufacturing_case-studies_card

レイクハウスアーキテクチャ

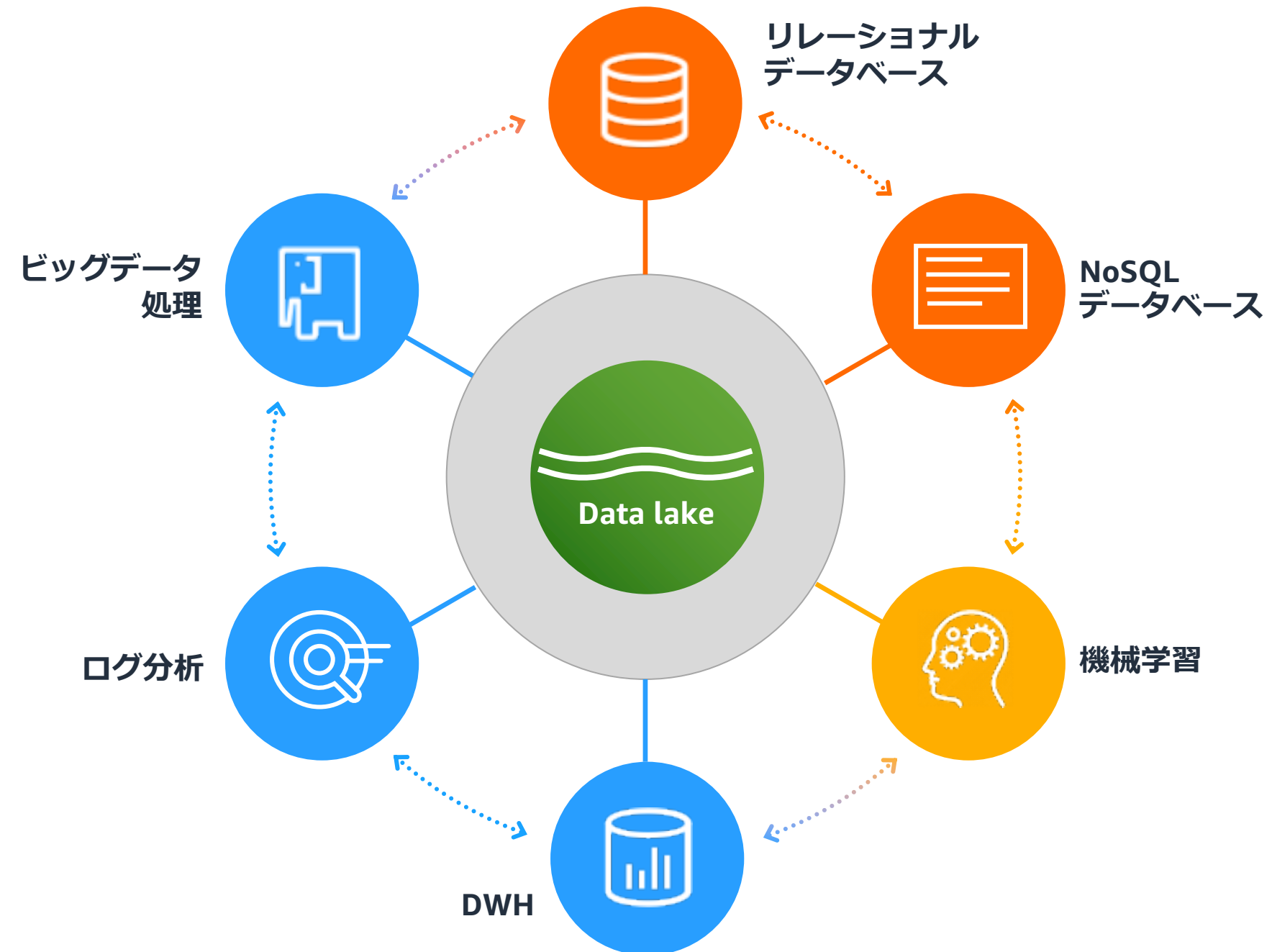
スケーラブルなデータレイク

ユースケースに応じた分析サービス

シームレスなデータ移動

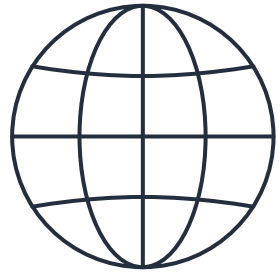
統一されたガバナンス

パフォーマンスと費用対効果



AWS クラウドで実現する データ活用基盤

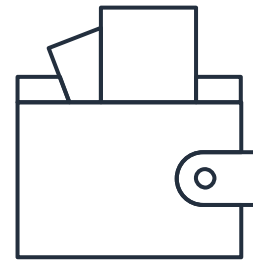
データ分析基盤を AWS クラウドで構築すべき理由



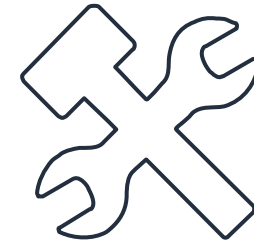
包括的な
ポートフォリオ



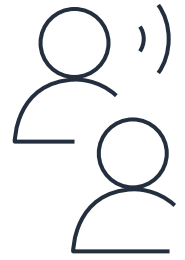
セキュリティ



高い
コスト効果

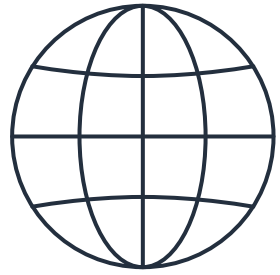


構築の
容易さ



豊富な顧客事例と
パートナー
ソリューション

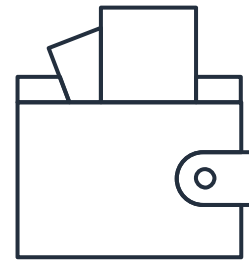
データ分析基盤を AWS クラウドで構築すべき理由



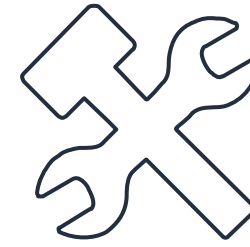
包括的な
ポートフォリオ



セキュリティ



高い
コスト効果

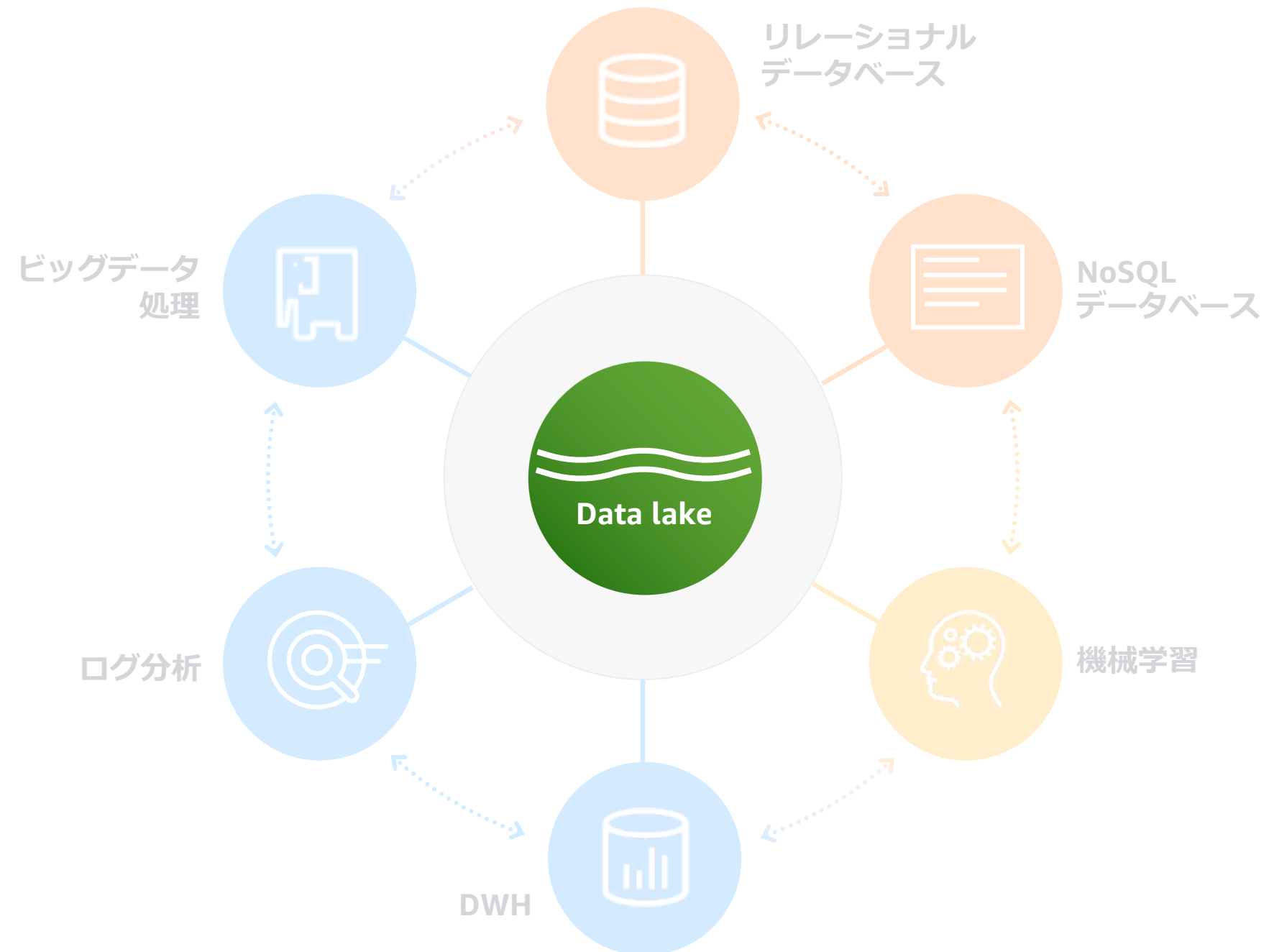


構築の
容易さ



豊富な顧客事例と
パートナー
ソリューション

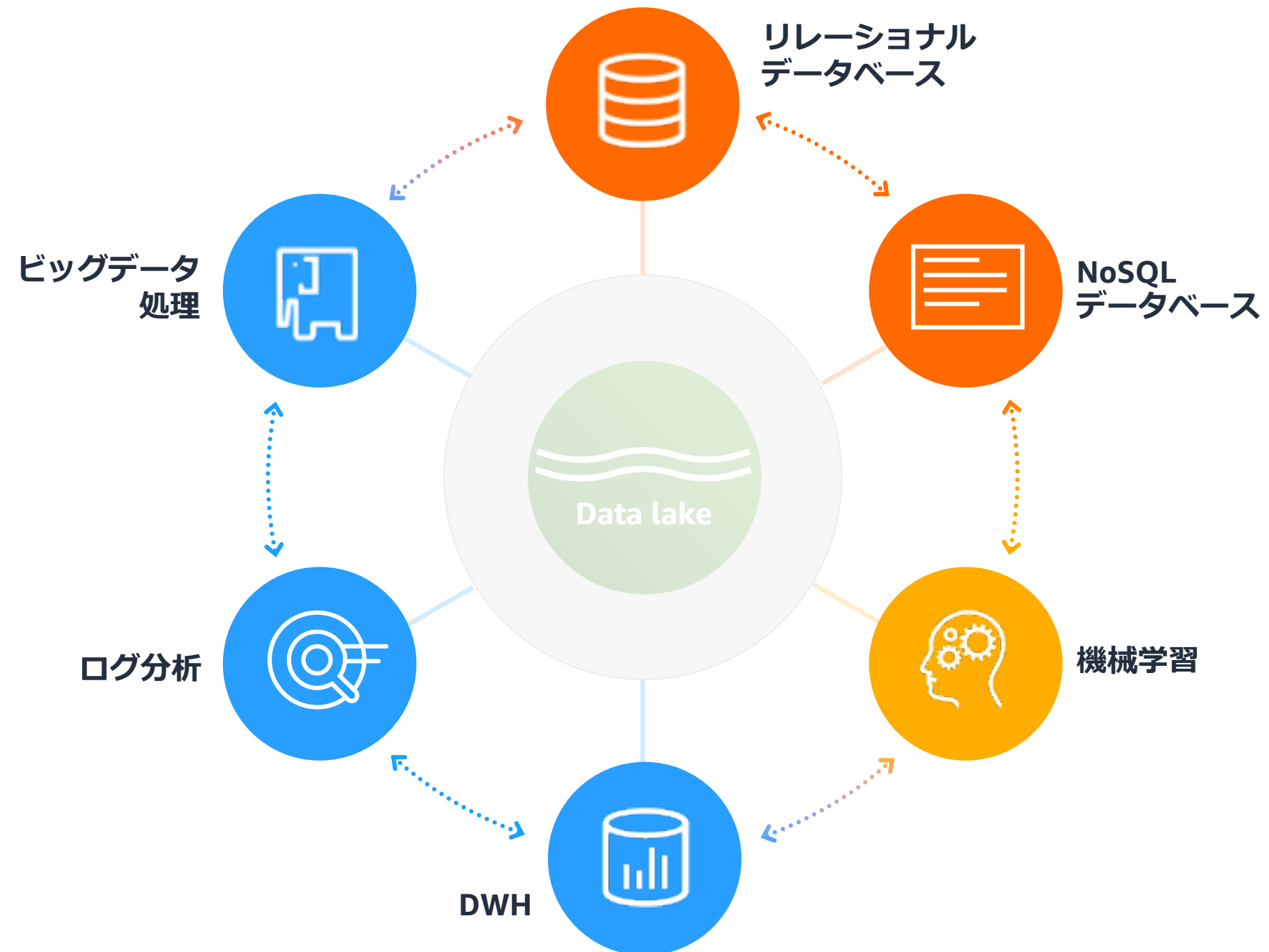
スケーラブルな データレイク



AWS のデータレイク = Amazon S3



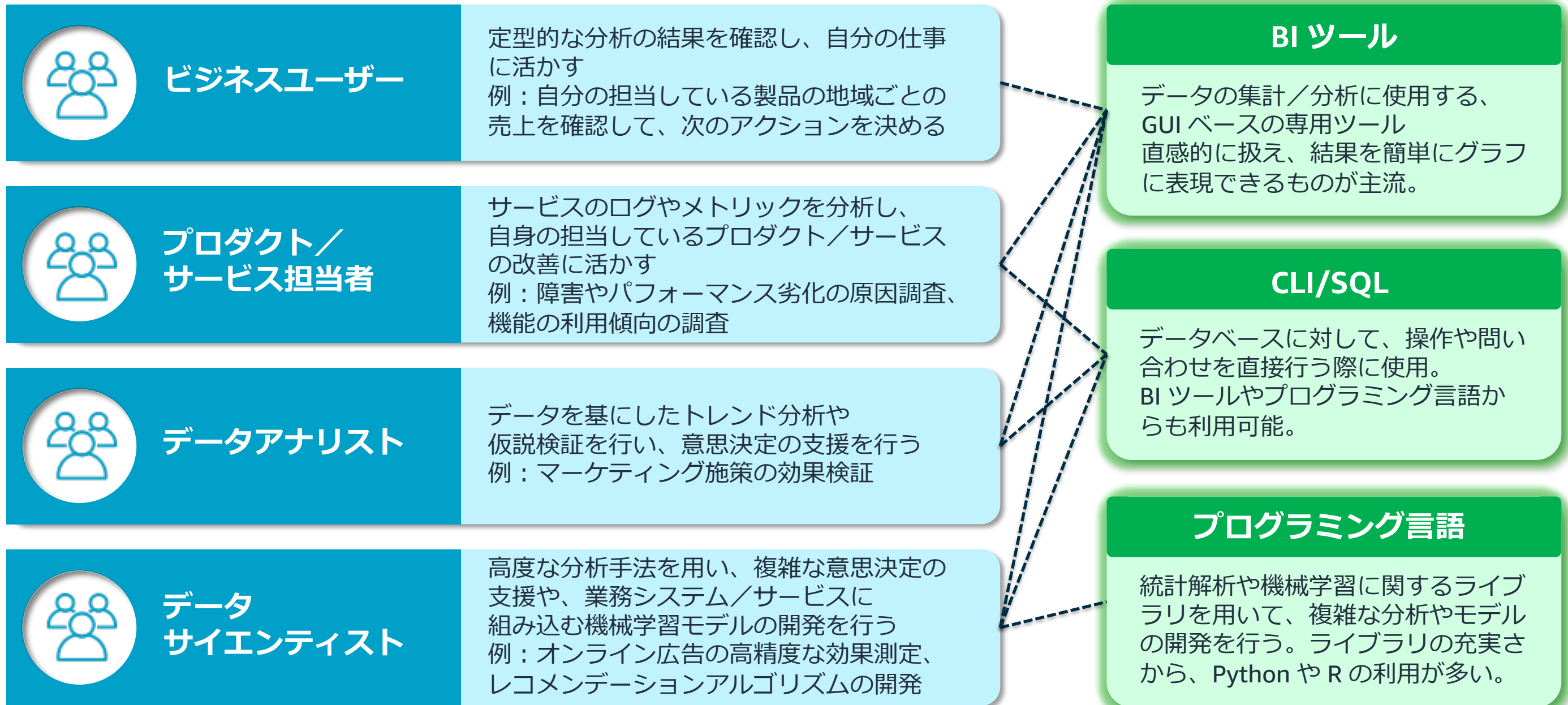
ユースケースに応じた分析サービス



データ分析の代表的なユースケースと利用ツール

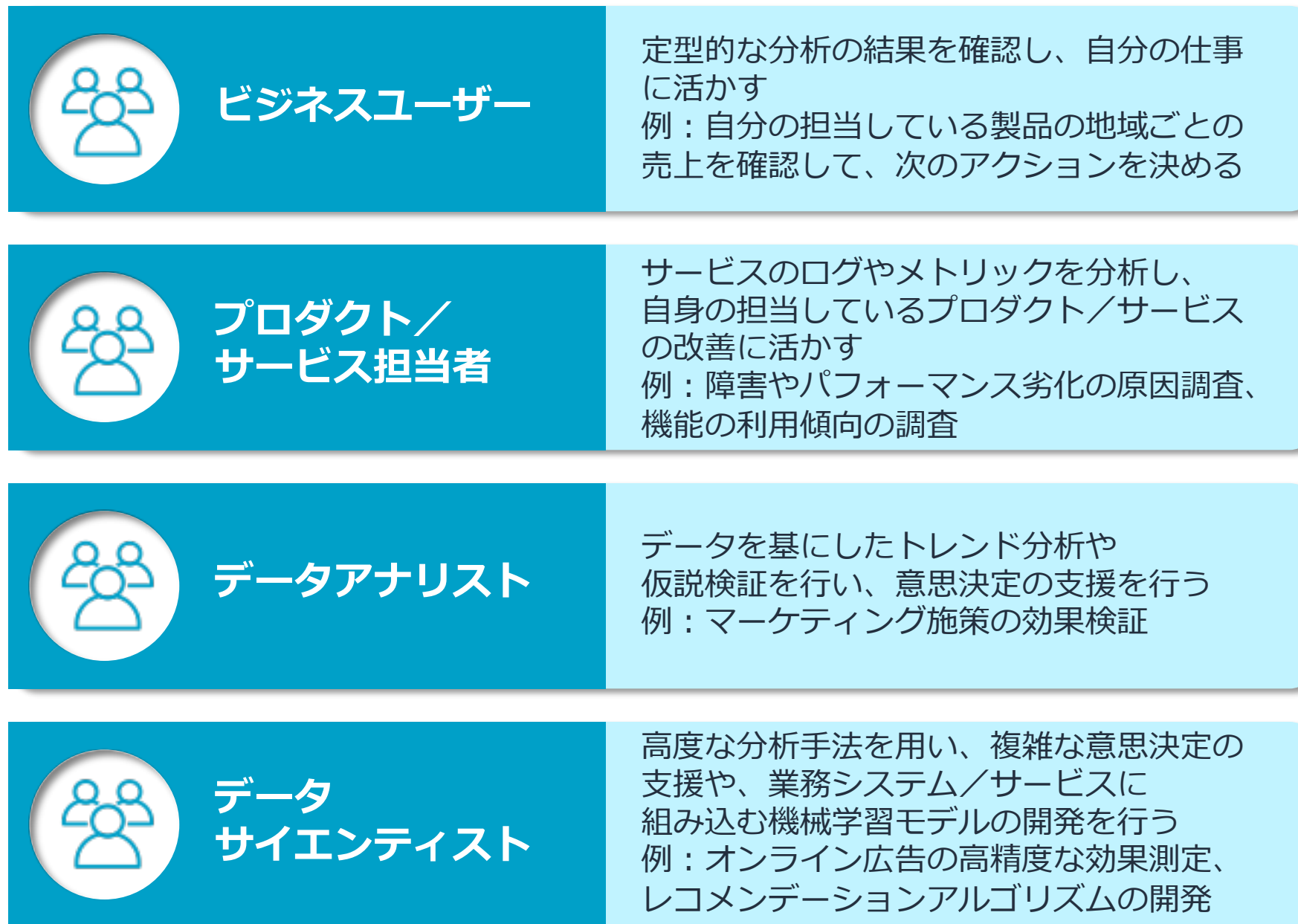
ユースケース

利用ツール

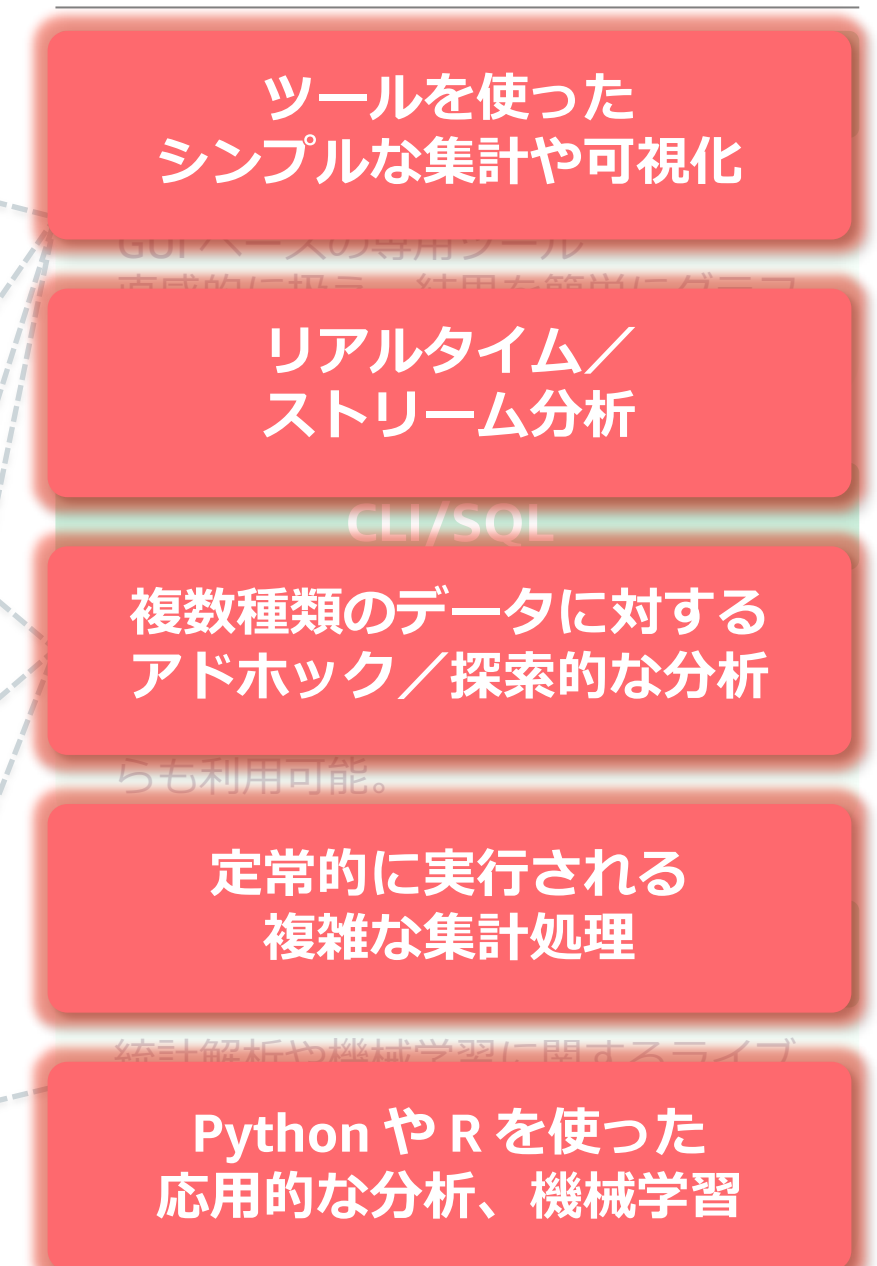


データ分析基盤で実行されるワークロード

ユースケース



分析基盤で実行されるワークロード



ユースケースに合わせた最適な分析サービス

パフォーマンス、コスト、スケール、ユースケースに沿った最適な分析サービスを選択



Amazon EMR

ビッグデータ処理



Amazon Redshift

データウェアハウス



Amazon Athena

対話的なクエリ

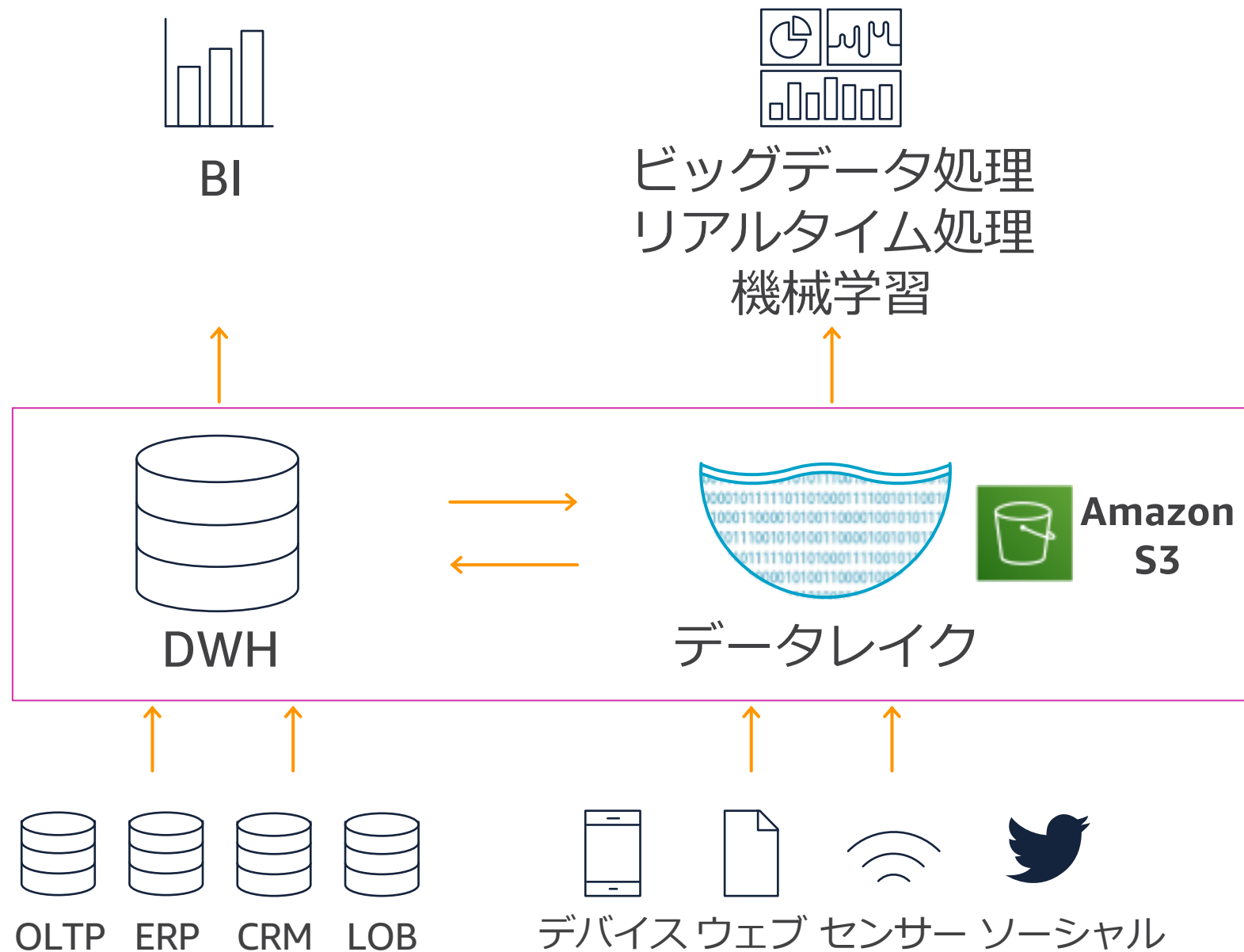


Amazon Kinesis
and Amazon MSK

リアルタイム分析

マネージドサービス

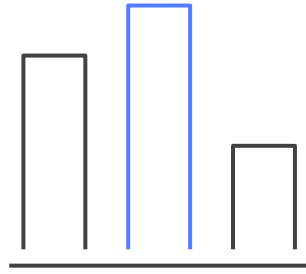
Amazon Redshift



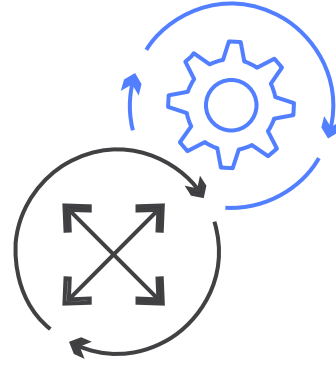
Amazon Redshift

高速、スケーラブルで
費用対効果の高い
データウェアハウス
およびデータレイク
分析マネージドサービス

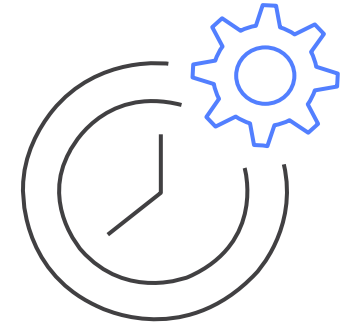
Amazon Redshift の特長



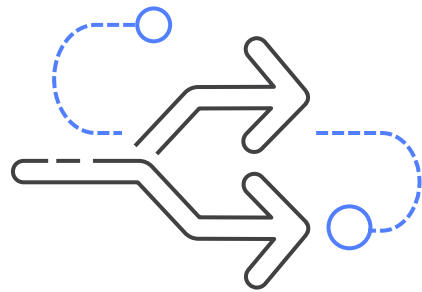
ハイパフォーマンス



拡張性 & 柔軟性



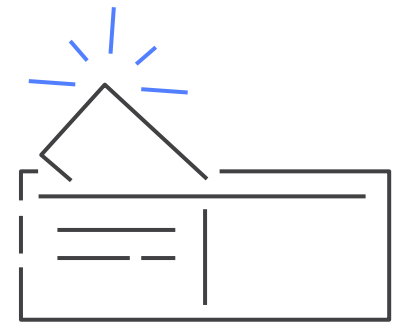
データレイク &
AWS サービスとの親和性



フルマネージド



セキュリティ &
コンプライアンス

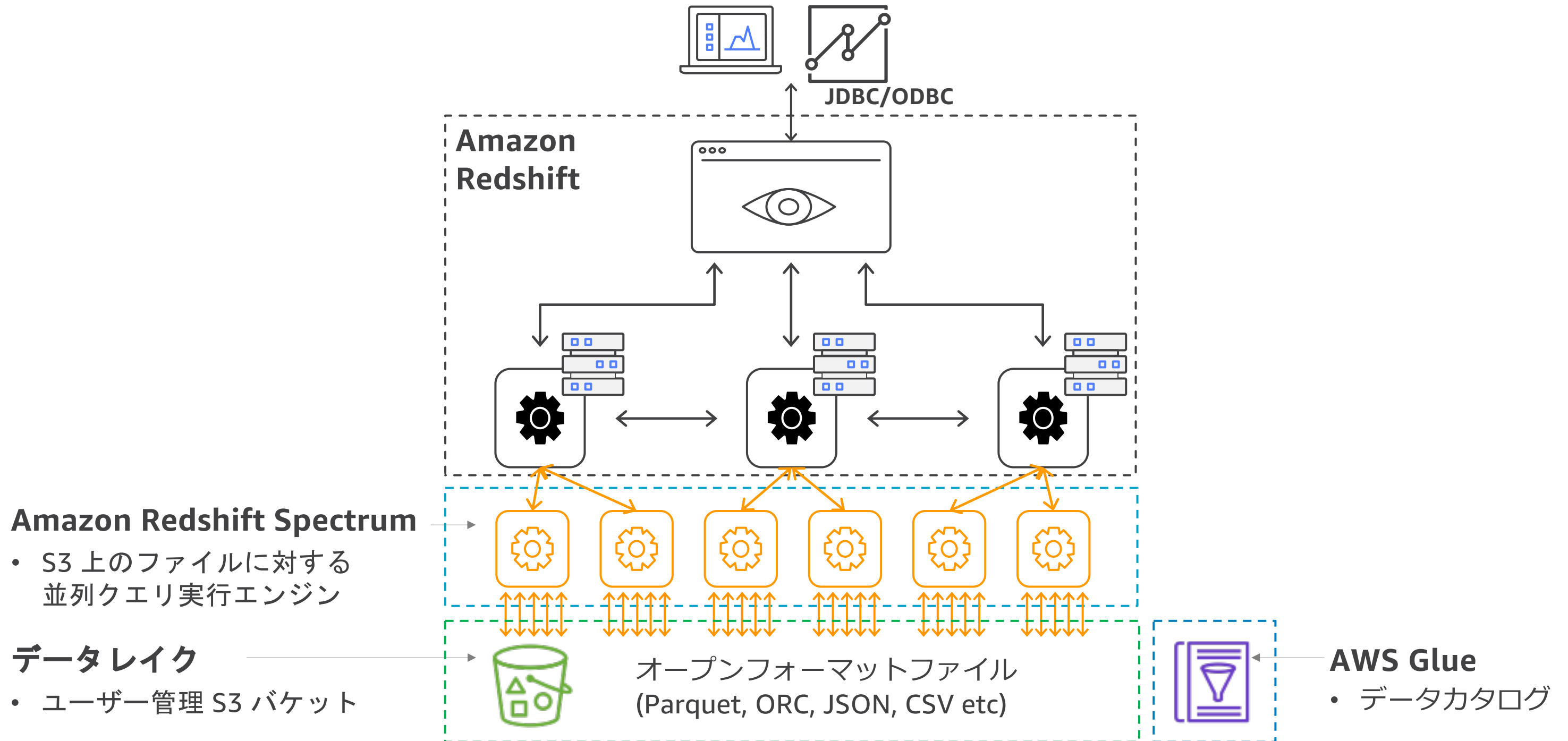


高いコスト効果

Redshift は後方互換性を維持しながら進化し続ける

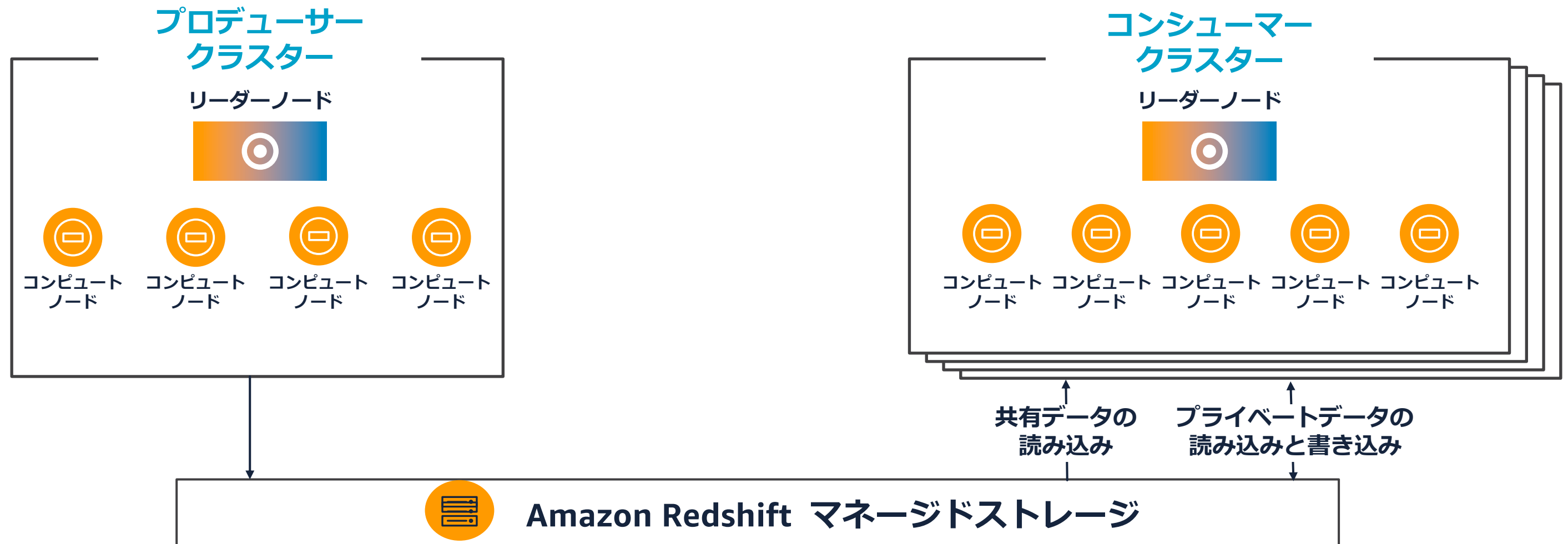


Redshift Spectrum でアーキテクチャをデータレイクに拡張



Amazon Redshift Data Sharing

Redshift クラスター間でセキュアに簡単にデータを共有することが可能



共有データにアクセスするプロデューサー/コンシューマーそれぞれのワークロードを分離

プロデューサー/コンシューマーごとに個別に料金をお支払い

リアルタイム分析：Amazon Kinesis, Amazon MSK

ビデオやストリームデータをリアルタイムで簡単に収集、処理、分析



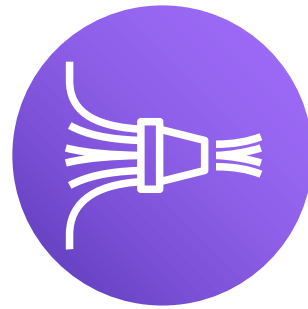
Amazon Kinesis Video Streams

メディアからの
ストリームデータを
キャプチャ、処理、保存し、
再生や分析、機械学習に活用



Amazon Kinesis Data Streams

ストリーミングデータを
スケーラブルに収集、
保存してリアルタイム
分析に活用



Amazon Kinesis Data Firehose

リアルタイムにデータ
ストリームを収集して、
データストアと
分析サービスにロード



Amazon Kinesis Data Analytics

SQL, Apache Flink,
Apache Beamを
使用してストリーム
データを分析



Amazon MSK

Apache Kafkaを使用して
ストリーミングデータを
処理するアプリケーション
を構築、実行

データ分析基盤を AWS クラウドで構築すべき理由



包括的な
ポートフォリオ



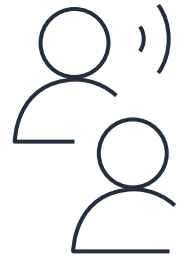
セキュリティ



高い
コスト効果



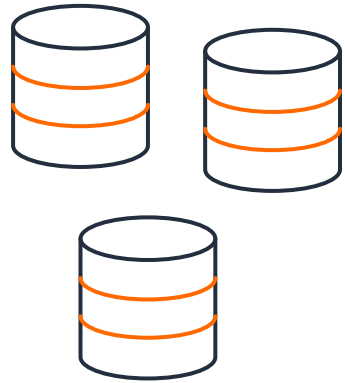
構築の
容易さ



豊富な顧客事例と
パートナー
ソリューション

複数データストア間でのデータ移動、データ統合の課題

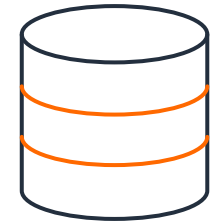
データストア



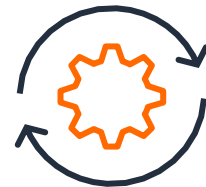
データパイプライン



ターゲット



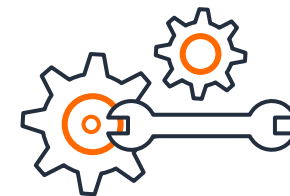
複雑なアプリケーション
コーディング



リトライ処理の考慮

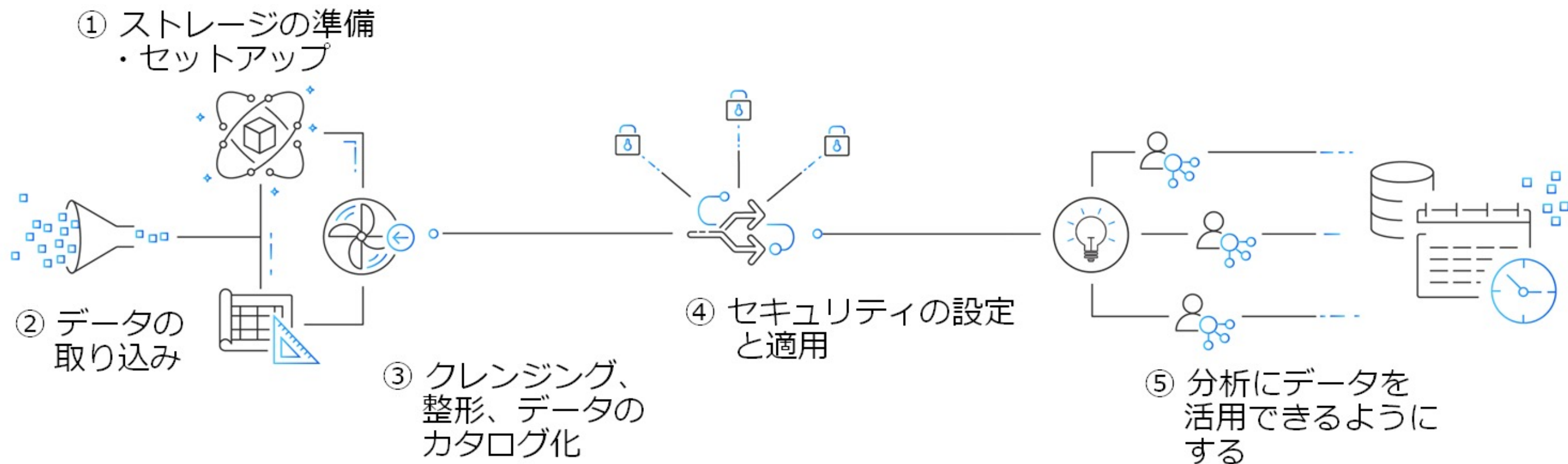


データパイプライン
の管理



ETLの専門家が必要

典型的なデータレイク構築までの流れ



多くの作業ステップが必要

ガバナンスとセキュリティにおける課題

データ資産の管理

ユーザが安心してデータを活用するためのメタデータ／品質管理

データに関する質問に回答できるか：

- ・ XXXのデータは**どこにある**？
- ・ このデータは**どこで生まれたのか**？
- ・ このデータは**信頼できるのか**？
- ・ 規制上、このデータは**そのまま利用して良い**のだろうか？

データセキュリティ

機密データのマスキングや暗号化処理

GDPR などコンプライアンス要件への対応

アクセスコントロール

データへのきめ細やかなアクセス制御
(どのユーザーにどのデータへのアクセスを許可するか)

行レベルセキュリティ

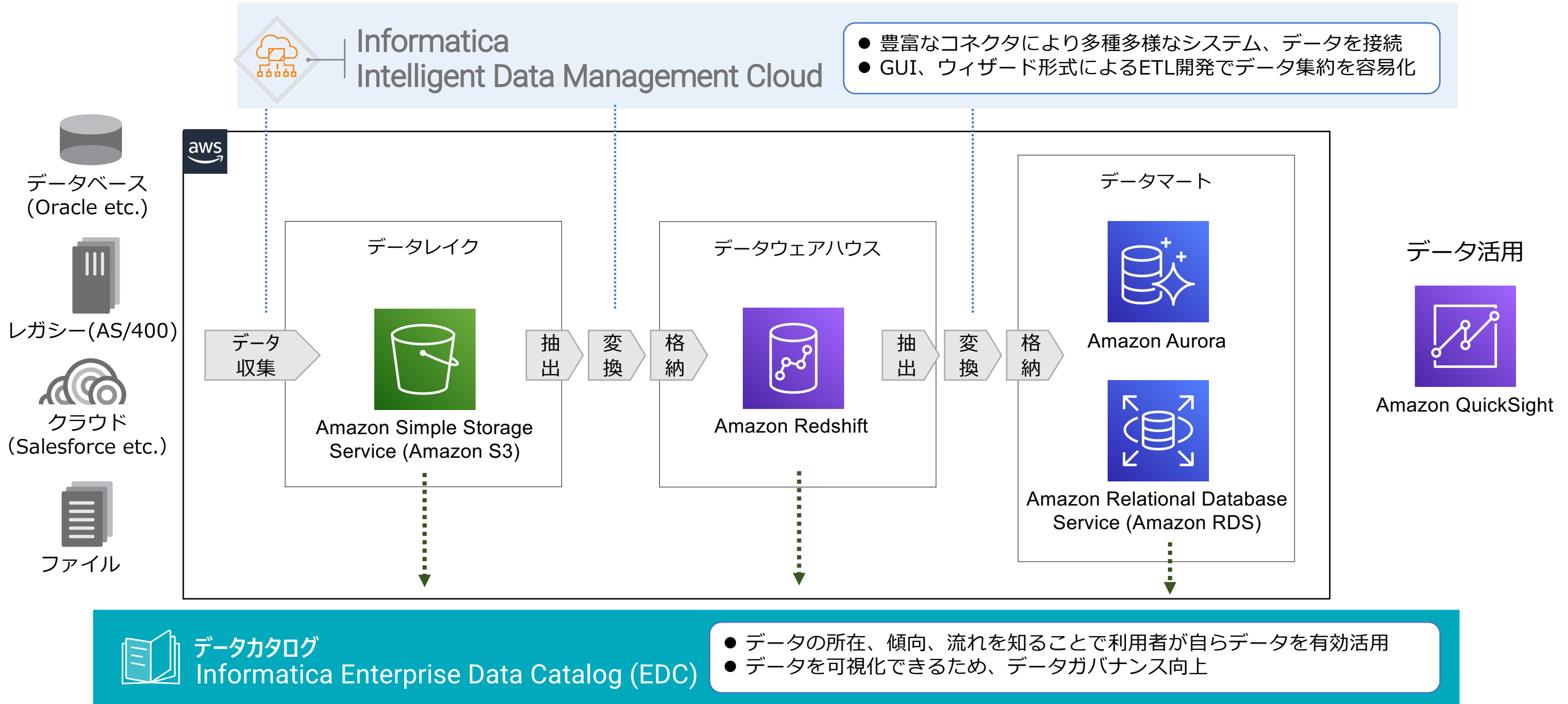
監査

ログ記録

証跡管理

SCSK × インフォマティカによるクラウドデータ分析基盤

データ統合、資産管理を実現する Intelligent Data Management Cloud と Enterprise Data Catalog



まとめ

- レイクハウスアーキテクチャ: データレイクを中心に据え、分析サービスを用いて、データを循環させる
 - 必要な時に必要な分のデータを取り出し、分析を行う
- ユースケースに応じて、目的に合わせた分析サービスを選択することが重要
 - データを安全に、容量制限なく保存できる **Amazon S3** をベースに、AWS、パートナーが提供する分析サービスを組み合わせる
- “使われる” データ分析基盤を実現するには、組織の誰もが、**快適に**、そして何よりも**安全にデータを利活用できる**状態を作ることが重要
 - データの所在や来歴、傾向を可視化し、正しく管理し続ける必要がある
 - AWS クラウド上でのデータ統合、資産管理を支援する SCSK、インフォマティカのソリューションを使って、データ利活用を加速



Thank you!

