これからのデータウェアハウス選択のポイントと Amazon Redshift による解決

大薗 純平
アマゾン ウェブ サービス ジャパン株式会社
アナリティクス スペシャリスト ソリューション アーキテクト

aws SUMMIT ONLINE

自己紹介

大薗 純平 (おおぞの じゅんぺい)

ソリューション アーキテクト

アナリティクス分野の技術支援を担当

好きなサービス :
Amazon Redshift をはじめとした
データ分析サービス

本セッションの対象者とゴール

- 対象者
  - データウェアハウス (DWH) を業務で使用しており、DWH についての知識・理解のある方
- ゴール
  - DWH を運用する上で課題になりがちな点を抑え、DWH に求められるものを認識する
  - Amazon Redshift が上記の課題をどのように解決できるのかについて理解する
- 本セッションでお話しないこと
  - DWH やデータ分析手法そのものに関する基本的な説明
  - Amazon Redshift を含む各種 AWS サービスの細かな仕様や設定などの解説

Agenda

データウェアハウスにまつわる課題

これからのデータウェアハウス選択のポイント

Amazon Redshift による解決

まとめ

データウェアハウスにまつわる課題

aws SUMMIT ONLINE

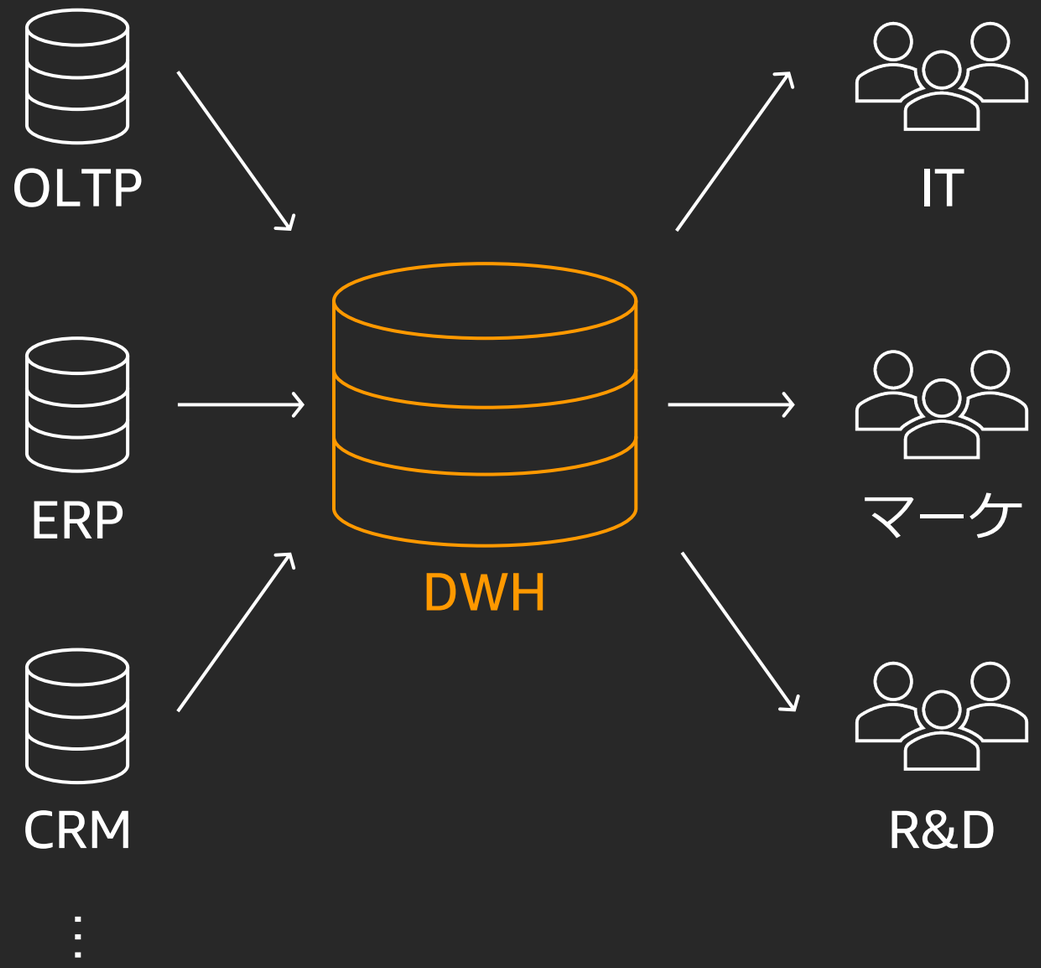
DWH 2 つのかたち

- 統合型
  - DWH 機能を一箇所に集約
  - IT 部門など特定の部署が管理する、エンタープライズ DWH に多い形
  - より全体のガバナンスは効かせやすくなる

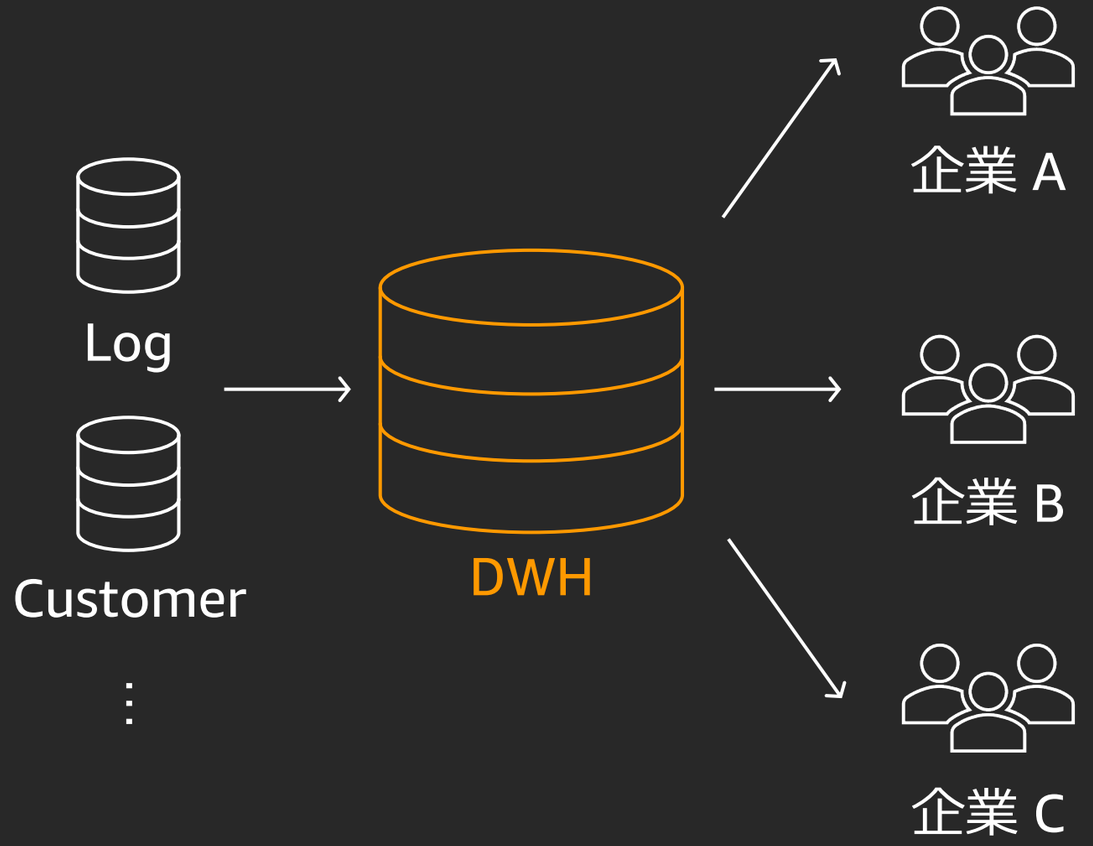
- 目的別
  - 組織や用途に応じて、DWH 機能を分離
  - それぞれの利用部門 / チームが、自分たちの裁量で DWH を管理するケースも増えている
  - 新たな分析の取り組みをスピーディに行いやすい

統合型 DWH の例

組織横断の社内 DWH

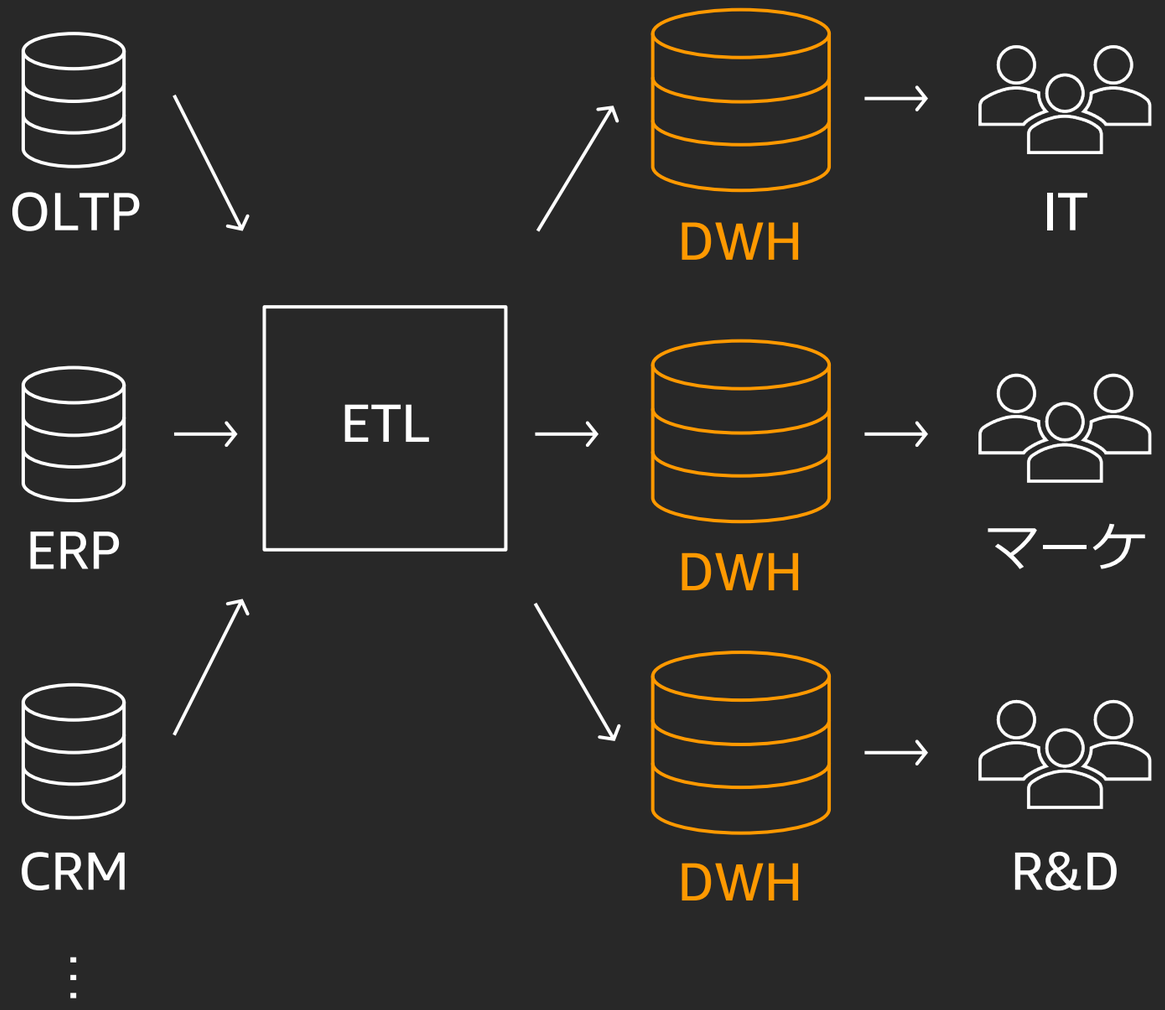


マルチテナント型サービス提供

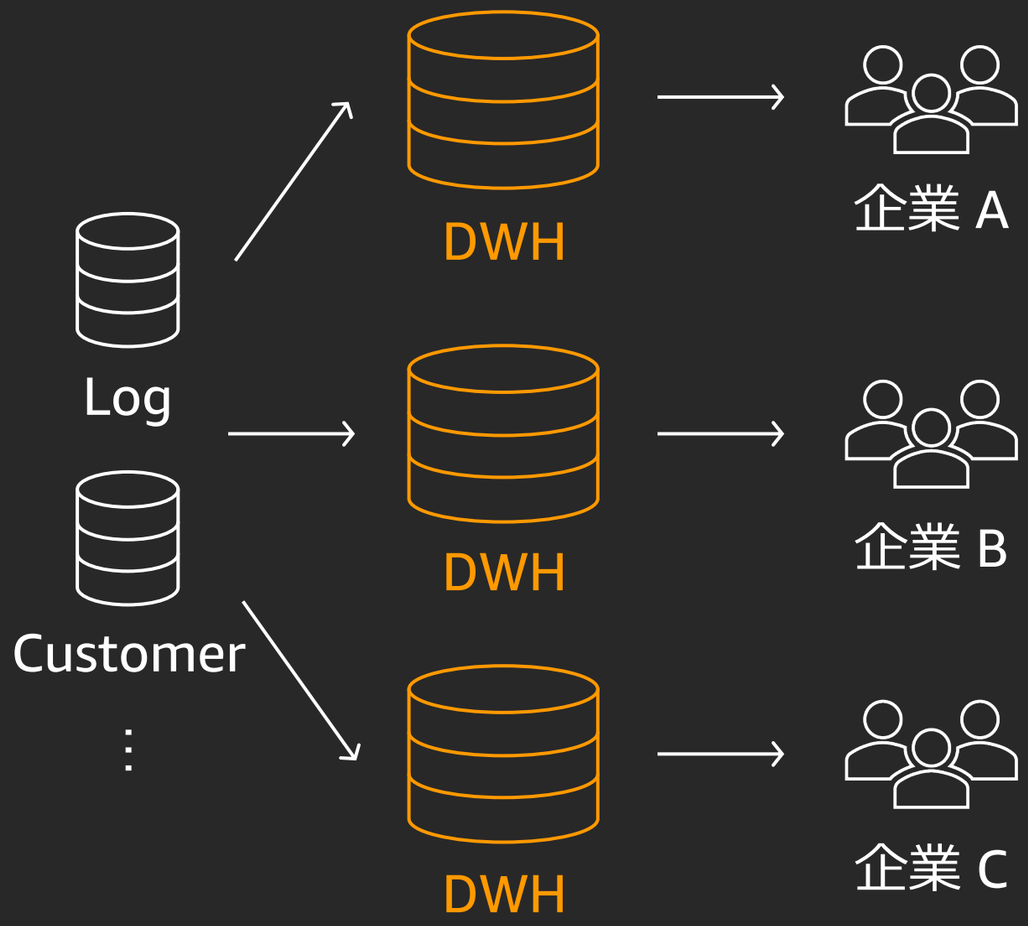


目的別 DWH の例

部署やサービスごとの社内 DWH

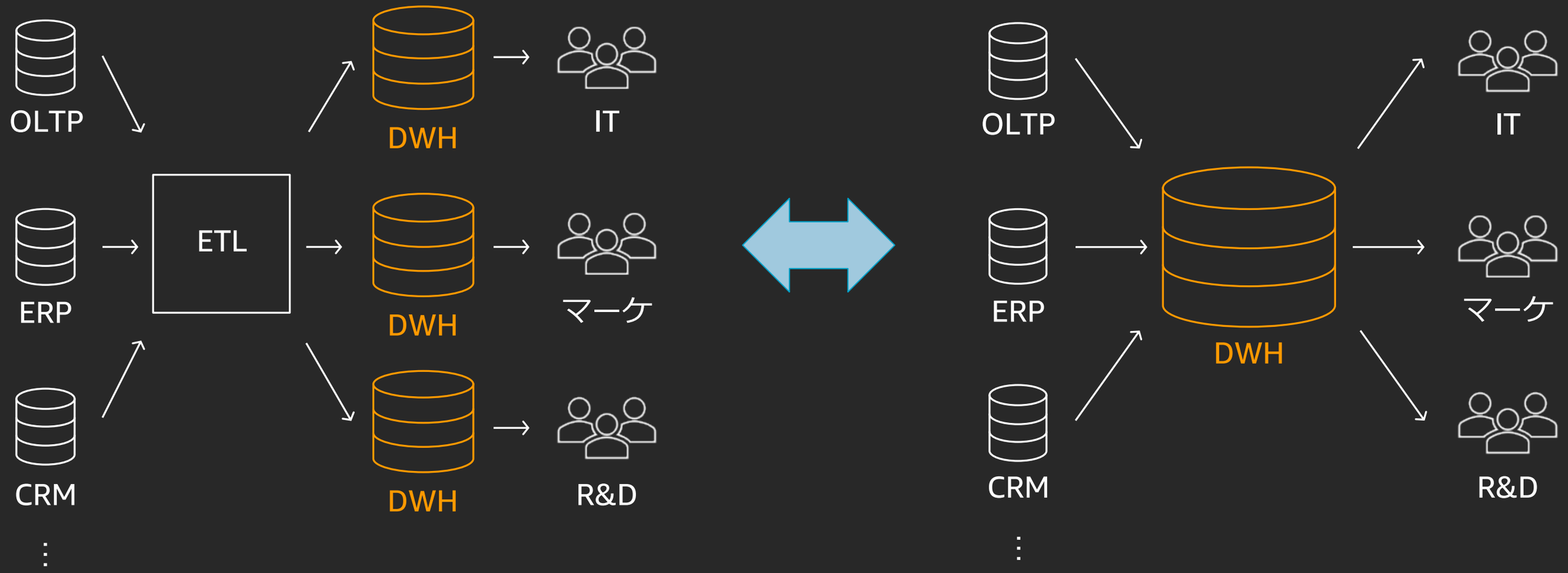


シングルテナント型サービス提供

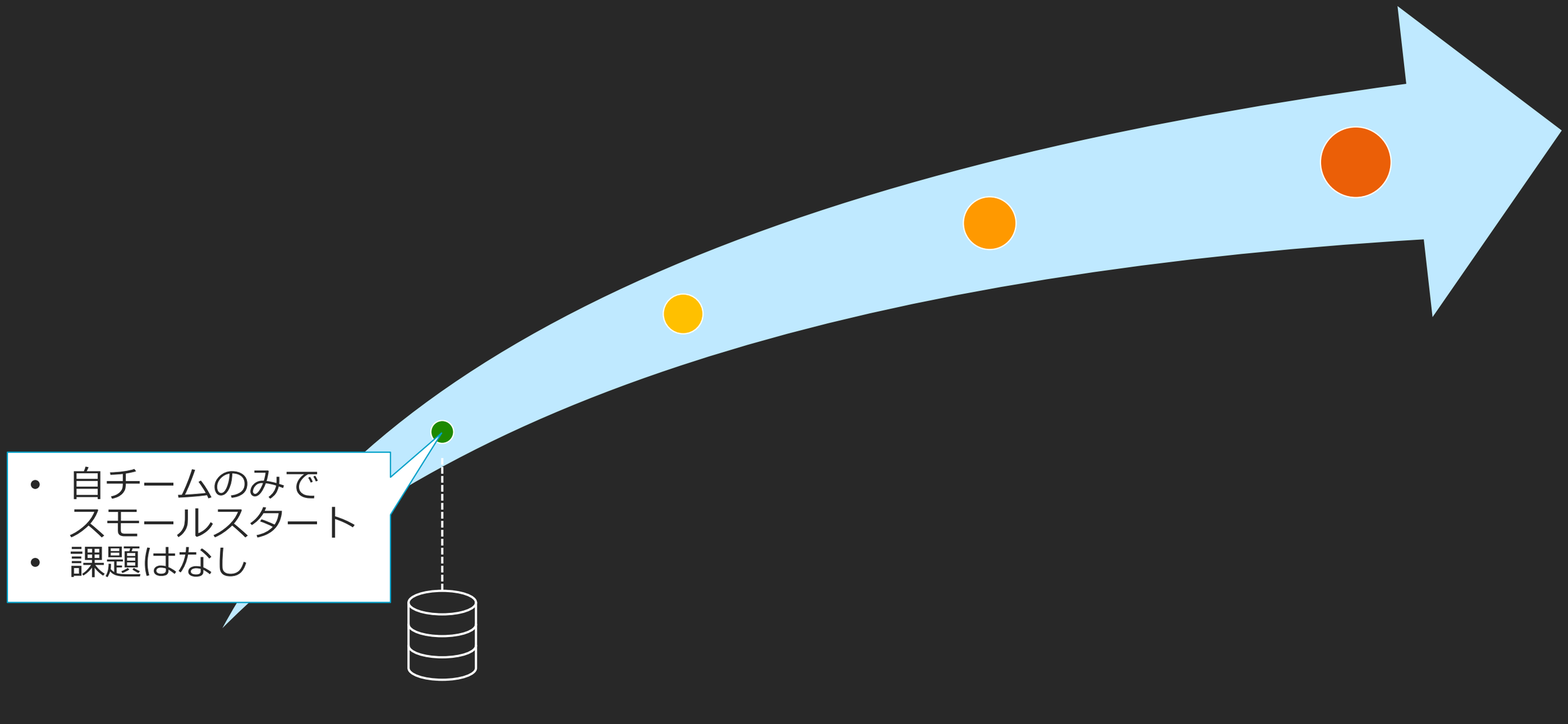


DWH のかたちは一定ではない

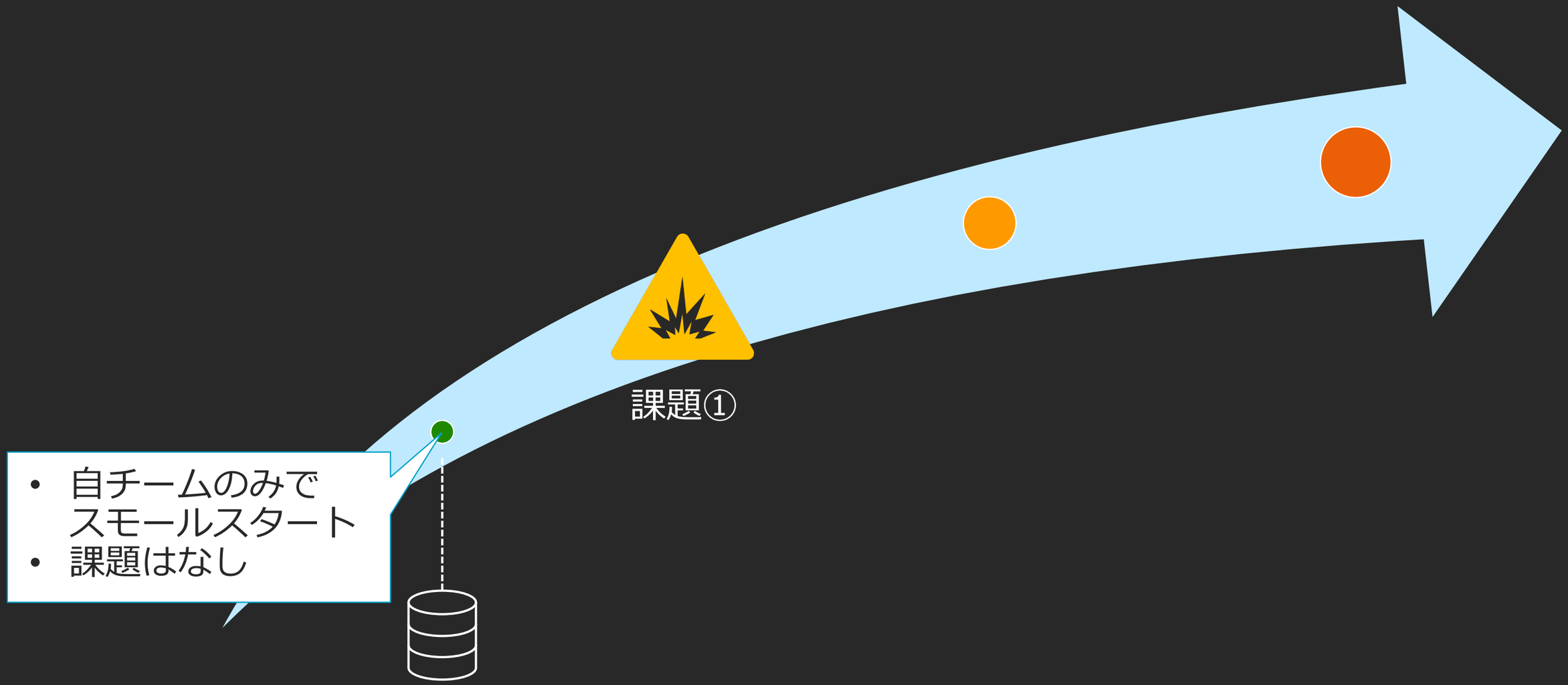
ビジネスの成長や組織の変化に応じて変わりうる



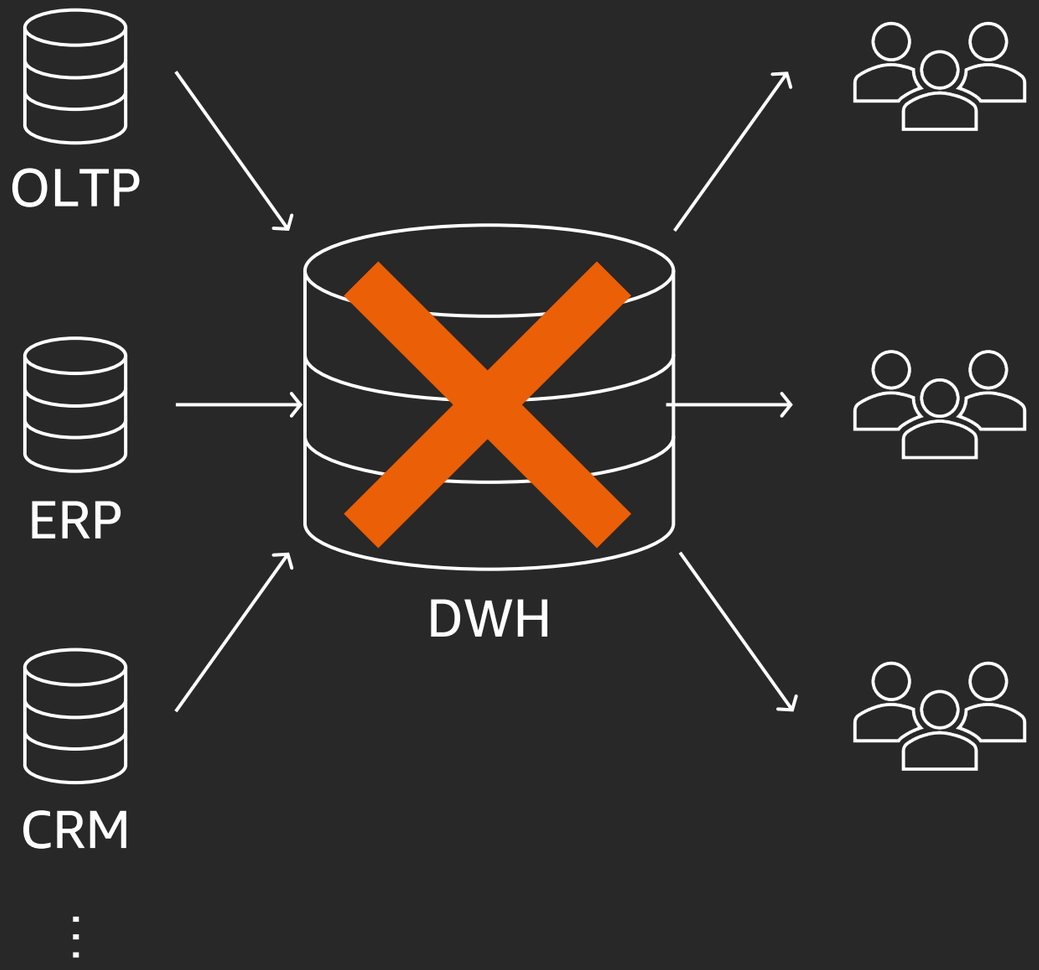
DWH の成長に伴う課題の例



DWH の成長に伴う課題の例

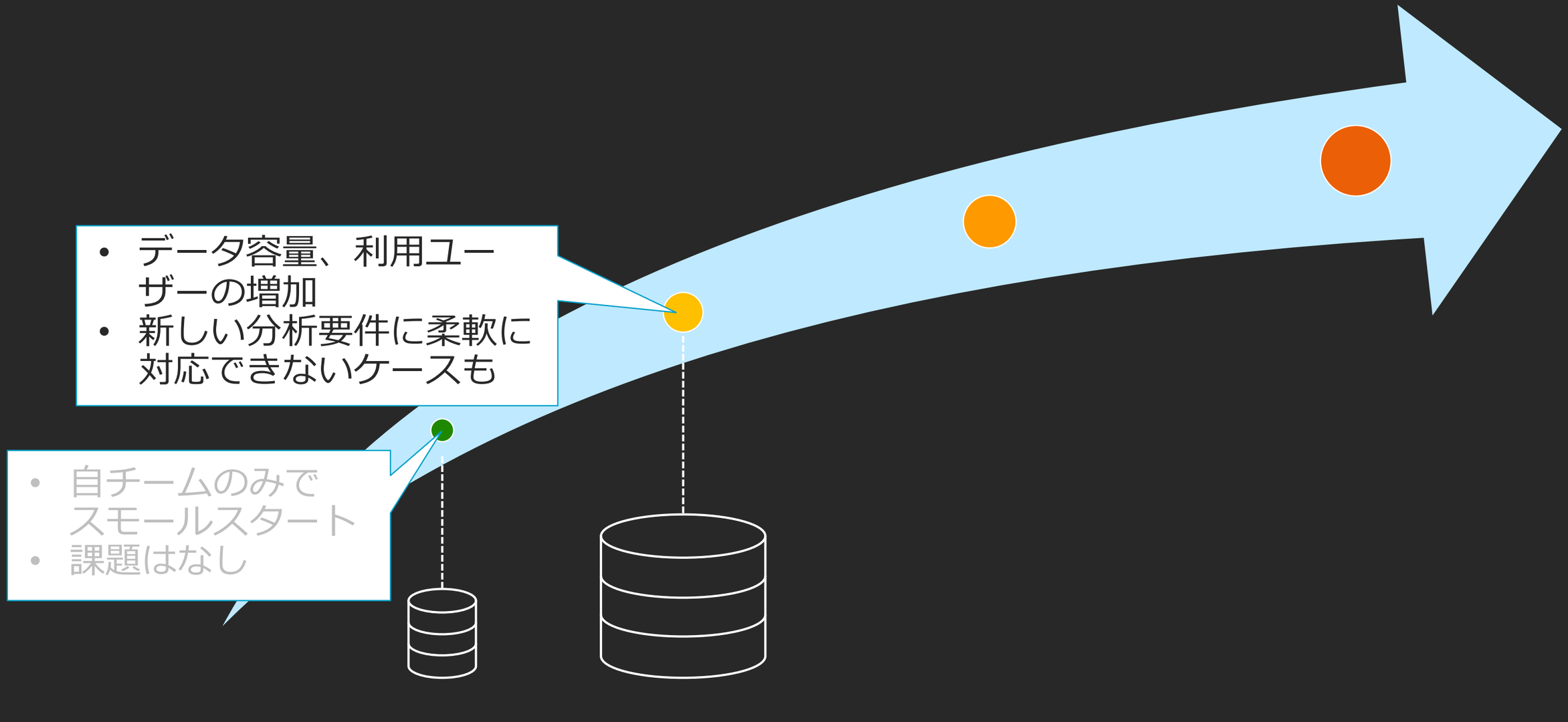


課題①：ビジネス変化への対応が困難

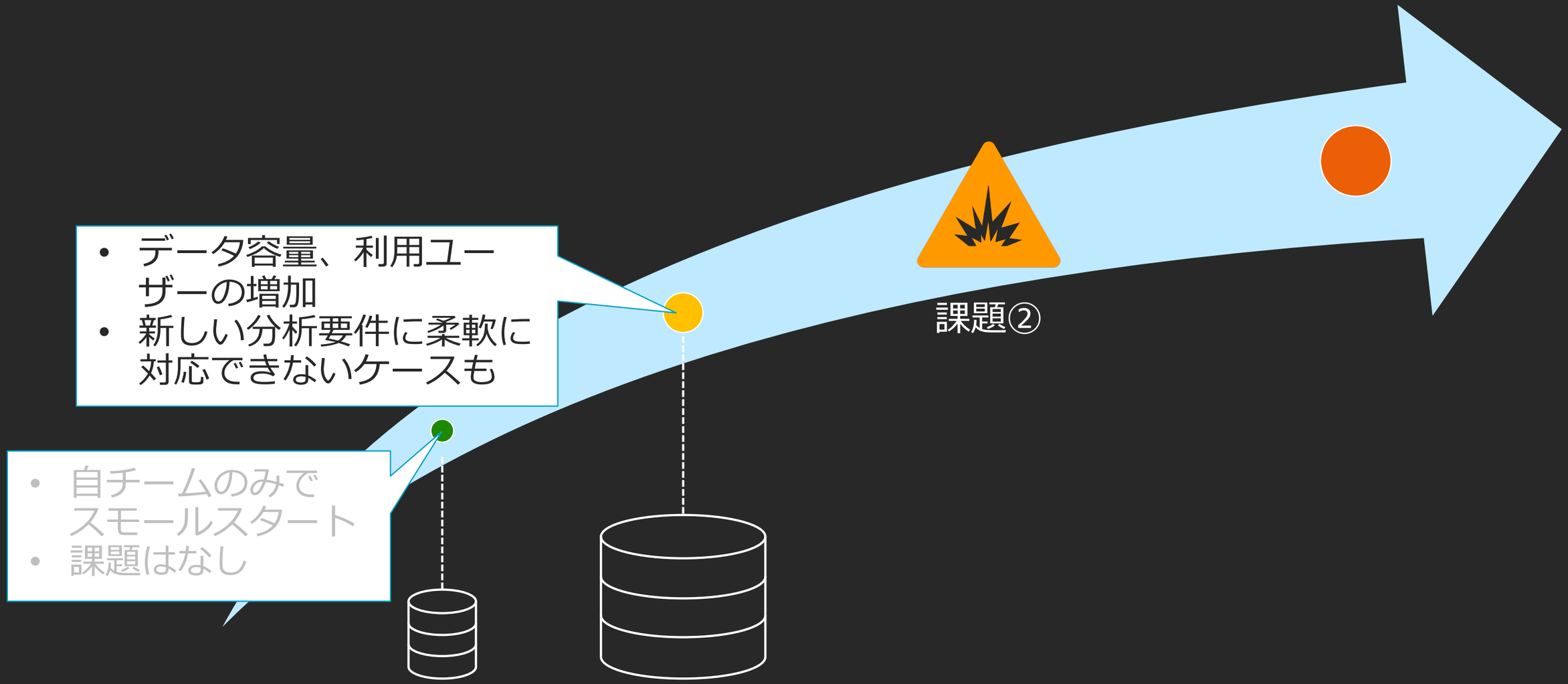


- データが予想以上に増えた場合、取り込みデータの制限や過去データの削除で対応
- 新たな分析要件の取り込みも困難
- 結果として**ビジネスの変化に柔軟に対応できない**

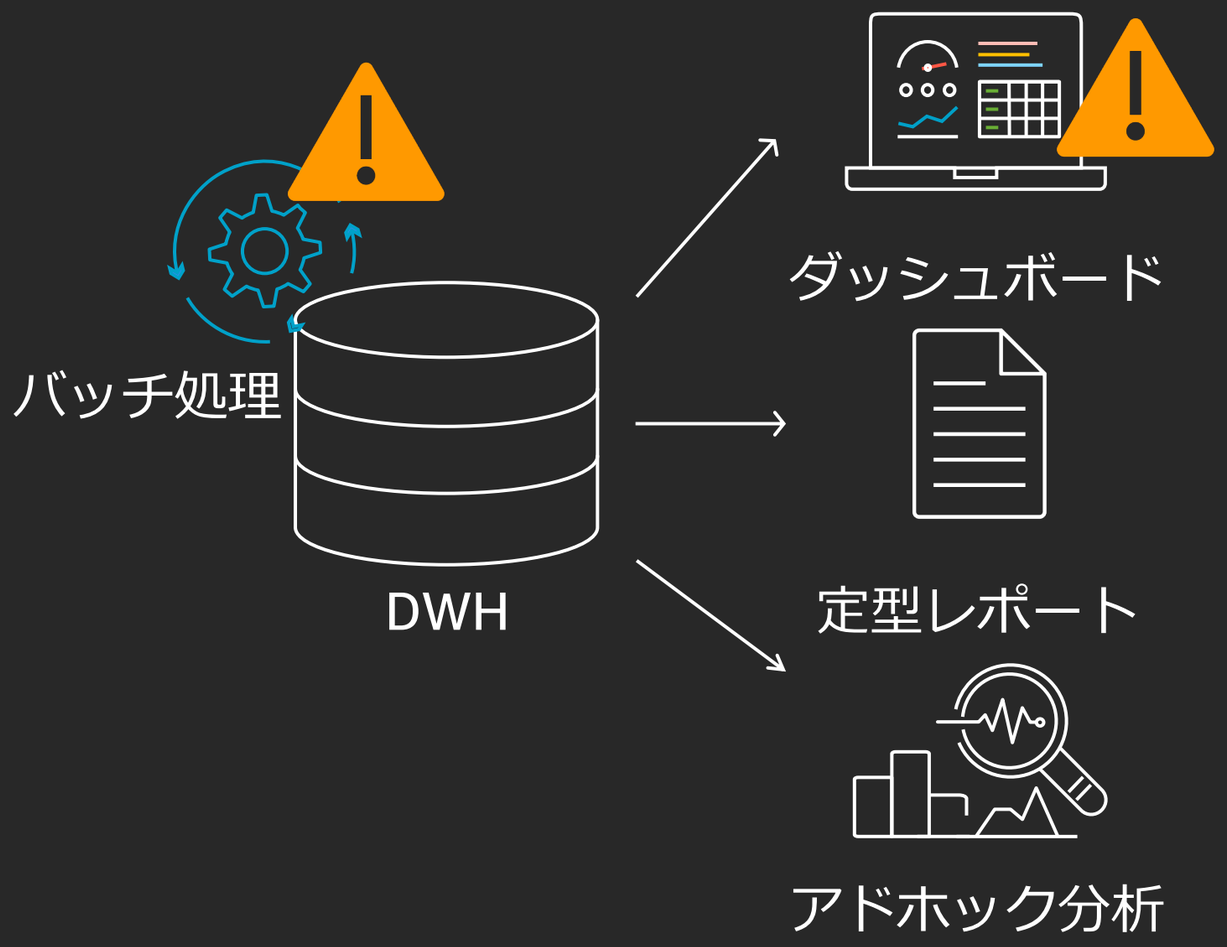
課題①：ビジネス変化への対応が困難



課題①：ビジネス変化への対応が困難

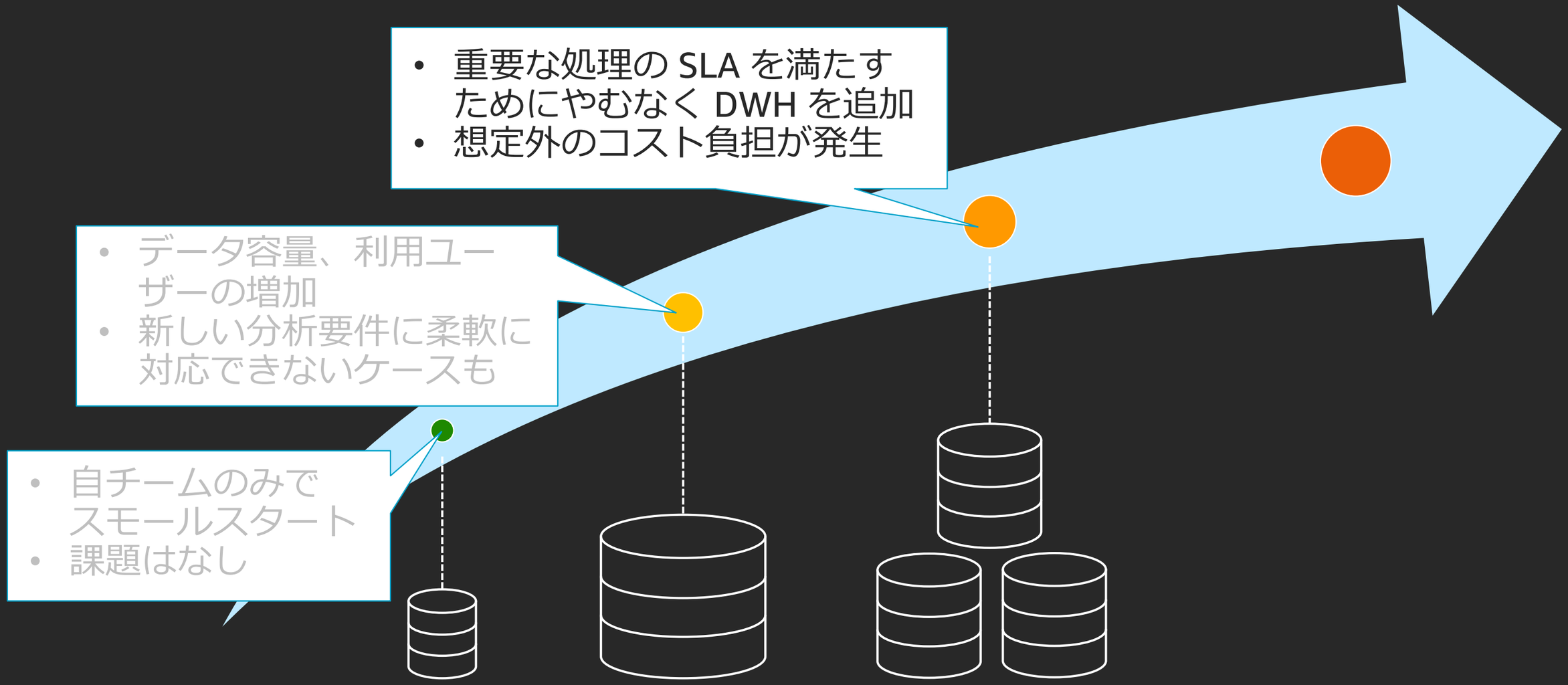


課題②：パフォーマンス SLA の維持

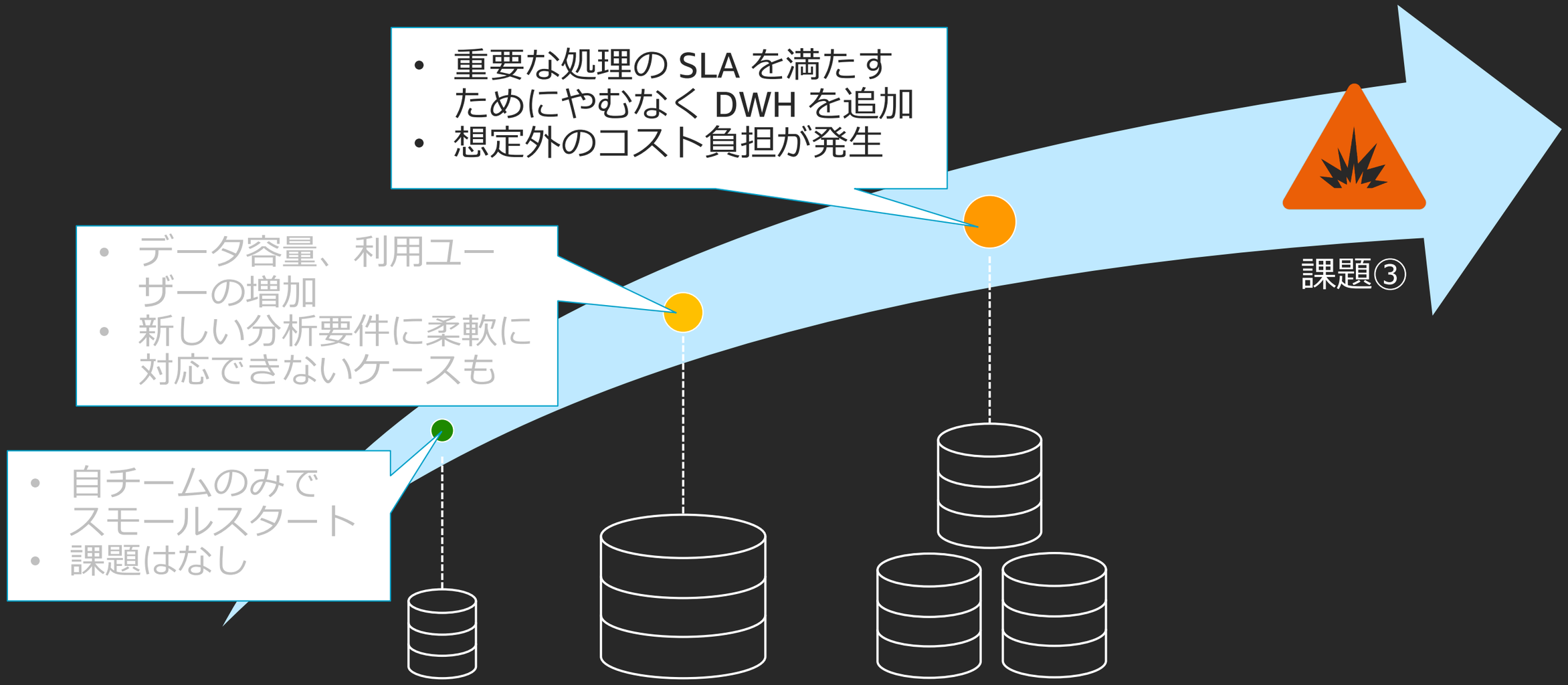


- それぞれの処理に定義された**パフォーマンスの SLA を守ることが困難**に
- チューニングの術がない場合、DWH リソースを追加する形となり、**想定外の追加コストが発生**

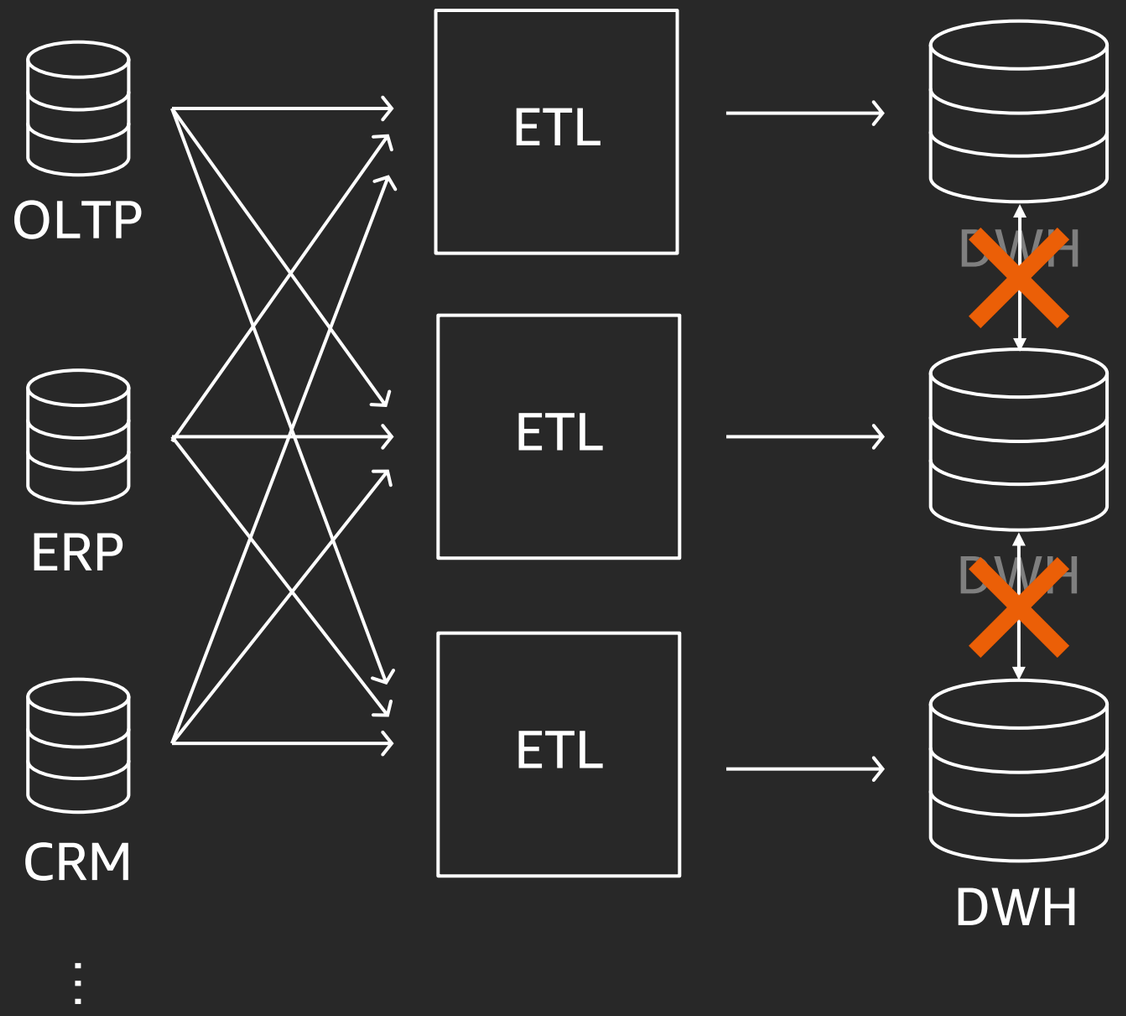
課題②：パフォーマンス SLA の維持



課題②：パフォーマンス SLA の維持



課題③：データのサイロ化と権限管理



- DWH 間でデータの共有はされず、データの重複や値の不一致がありいわゆる**データのサイロ化が発生**
- あとから横断的な分析のために**権限管理を実施するのが困難**

課題③：データのサイロ化と権限管理

- 企業全体で DWH が乱立し、データのサイロ化が発生
- 新たに発生するセキュリティ & ガバナンスの要件

- 重要な処理の SLA を満たすためにやむなく DWH を追加
- 想定外のコスト負担が発生

- データ容量、利用ユーザーの増加
- 新しい分析要件に柔軟に対応できないケースも

- 自チームのみでスモールスタート
- 課題はなし

DWH の成長に伴う課題の例

- 企業全体で DWH が乱立し、データのサイロ化が発生
- 新たに発生するセキュリティ & ガバナンスの要件

- 重要な処理の SLA を満たすためにやむなく DWH を追加
- 想定外のコスト負担が発生

- データ容量、利用ユーザーの増加
- 新しい分析要件に柔軟に対応できないケースも

- 自チームのみでスモールスタート
- 課題はなし

データウェアハウスにまつわる課題

1 ビジネス変化への対応が困難

ビジネスが変化すればワークロードも変化する

システム側が変化に柔軟に対応できる構成になっていない

2 パフォーマンス SLA の維持

DWH は常に成長し続けるものであり、データもユーザーも分析ニーズも増える

それでも守るべき SLA は変わりづらい

3 データのサイロ化と権限管理

個別最適を求めるとデータのサイロ化が発生しやすくなる

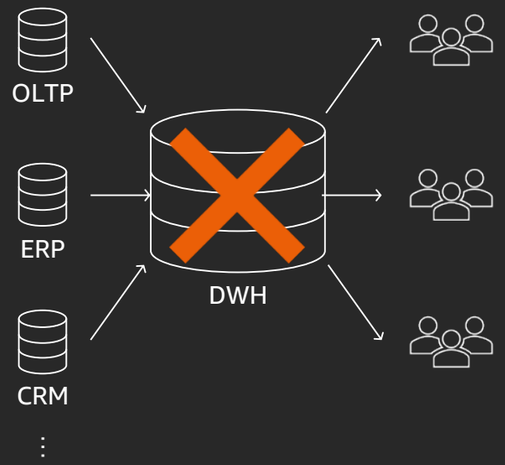
ビジネスの成長や外部環境の変化に対応できる権限管理の仕組みになっていない

これからのデータウェアハウス選択のポイント

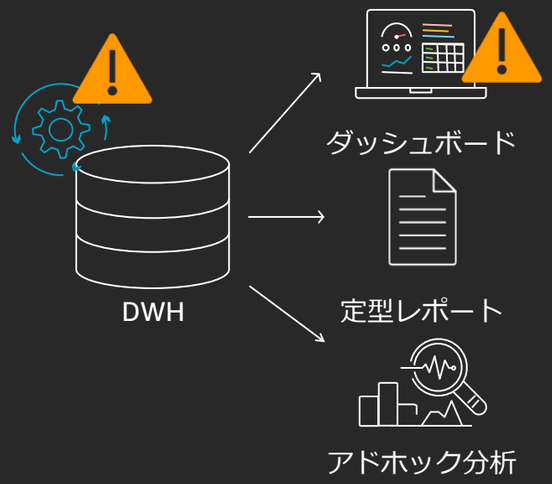
aws SUMMIT ONLINE

データウェアハウスにまつわる課題 (再掲)

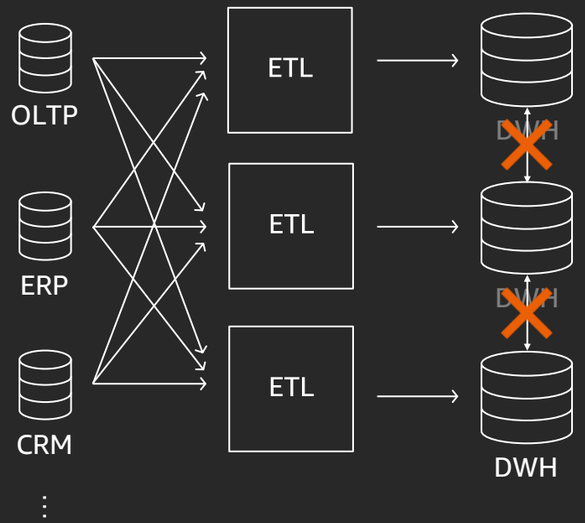
1 ビジネス変化への対応が困難



2 パフォーマンス SLA の維持



3 データのサイロ化と権限管理



このような課題は DWH だけで
解決することが難しくなっている

DWH を取り巻く環境の変化

データ容量/種類
の爆発的増加

あらゆるデータを
繋げて洞察を得たい

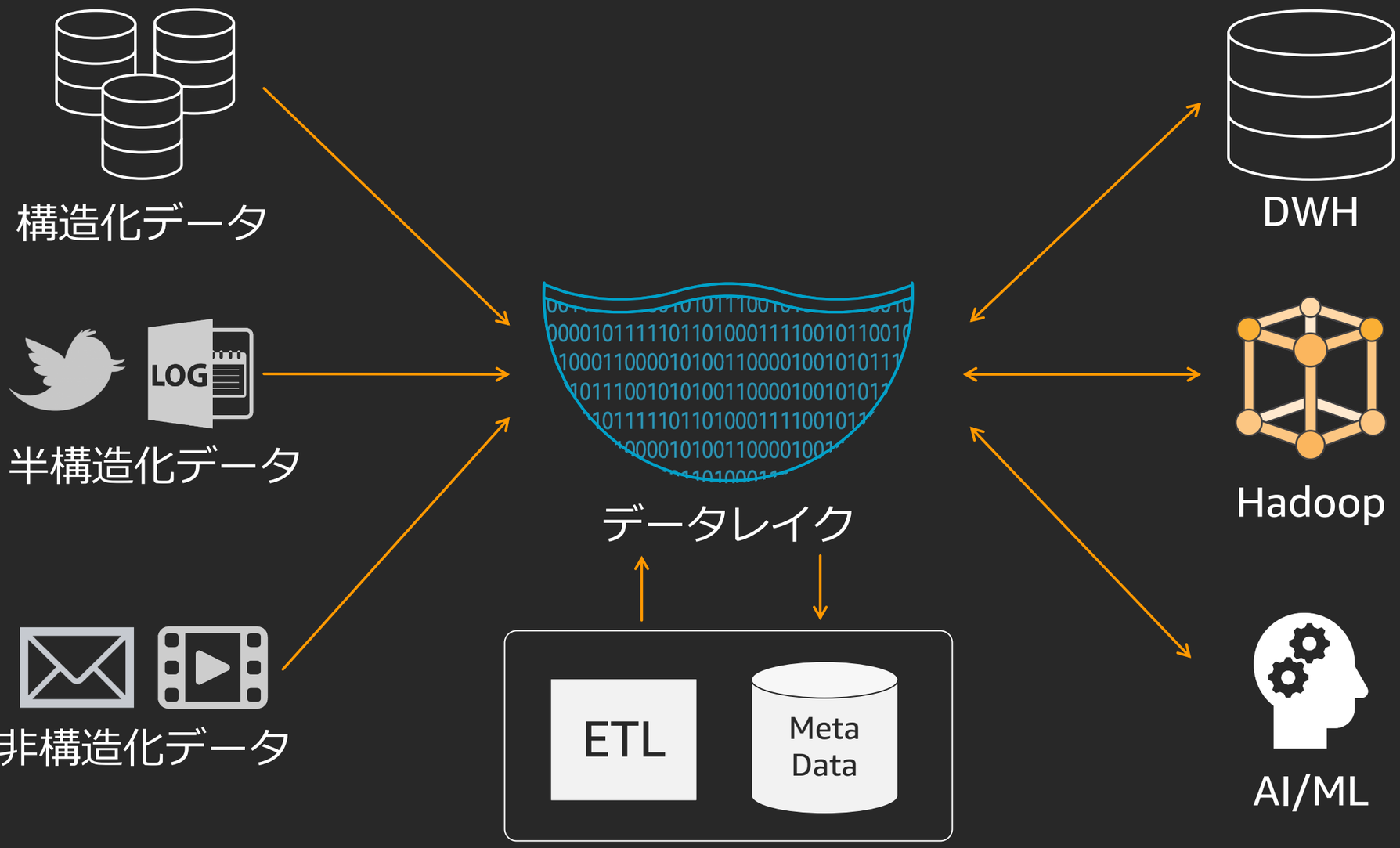
クラウドへの
移行の加速

データレイクという考え方

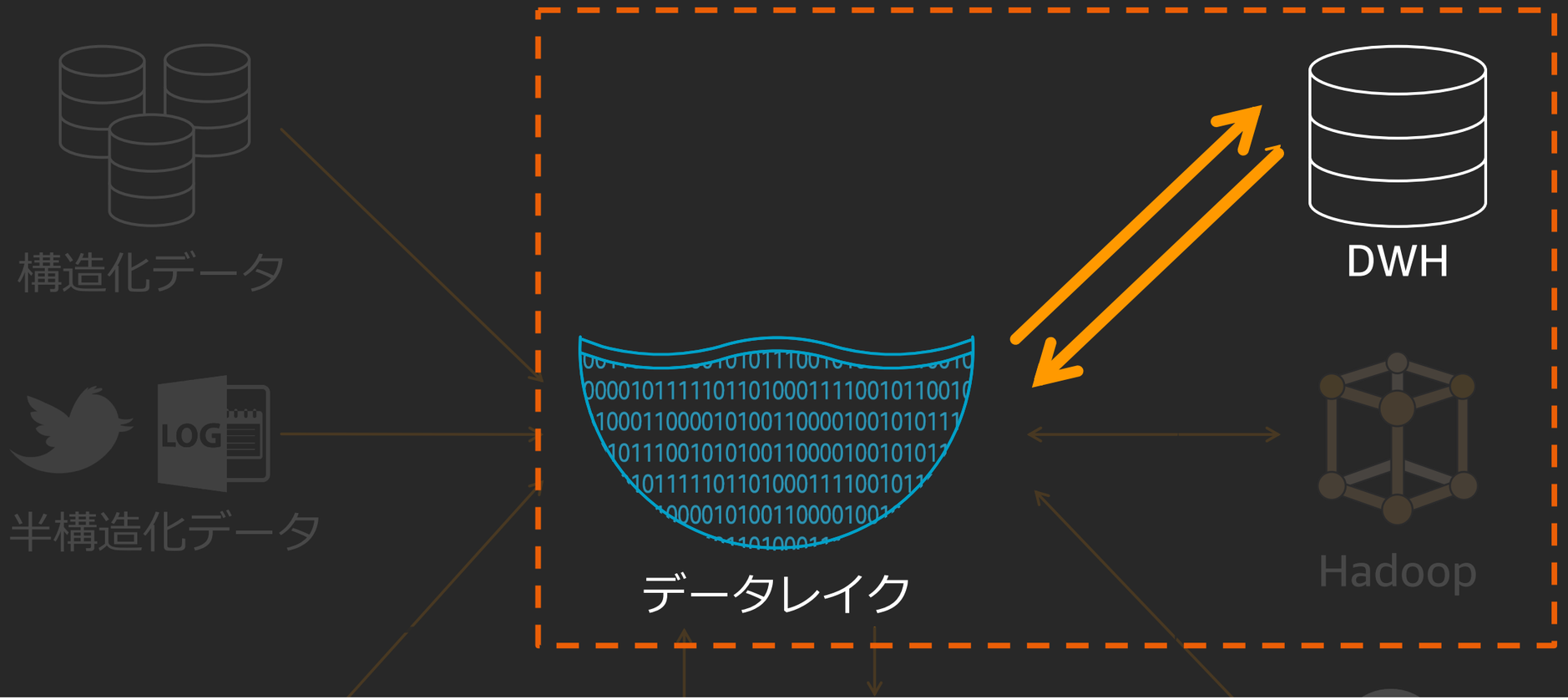
- すべてのデータを、スケールする形で、かつ低コストで一箇所に集めて保存するためのもの
- 適切にアクセスコントロールを定義したうえで、データをセキュアに保持することができるもの
- 「民主化」した形での組織内のデータアクセスを可能とするもの
- 新しいデータ分析を、素早くかつ簡単に実行することができるもの

データレイク

データレイクの典型的な構成



DWH はデータレイクの一部へ



1. 全データをデータレイクに集約することで**データのサイロ化を解消**
2. DWH には必要なデータのみ連携し、ビジネスの変化や時間の経過と共に必要に応じて**データレイクに書き戻す**

これからのデータウェアハウスに求められるもの

1. ビジネス変化への対応が困難

柔軟性とスケーラビリティ

事業のフェーズやビジネスの変化に応じて
コスト効果高く、**柔軟に構成を変えたり、**
必要なときに必要なリソースを拡張できること

2. パフォーマンス SLA の維持

高いベースラインパフォーマンスと
運用管理機能

デフォルトで高いパフォーマンスを出しつつ、
ユーザーによるチューニング余地も持てること

3. データのサイロ化と権限管理

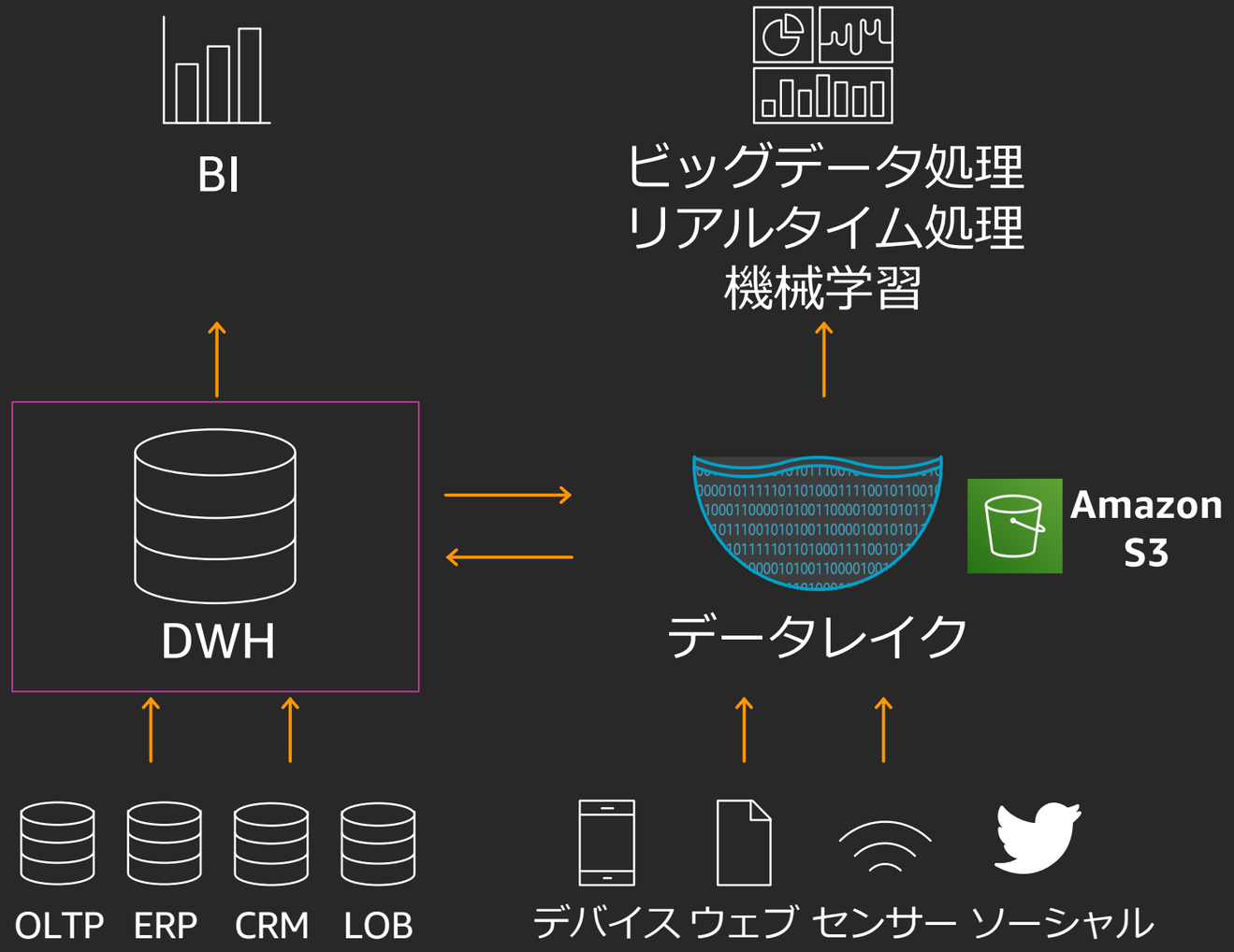
データレイク連携とセキュリティ

DWH をデータレイクにまで簡単に拡張し、
コストを抑えた形でデータを分析可能にし、
さらに**きめ細やかな権限の管理ができる**こと

Amazon Redshift による解決

aws SUMMIT ONLINE

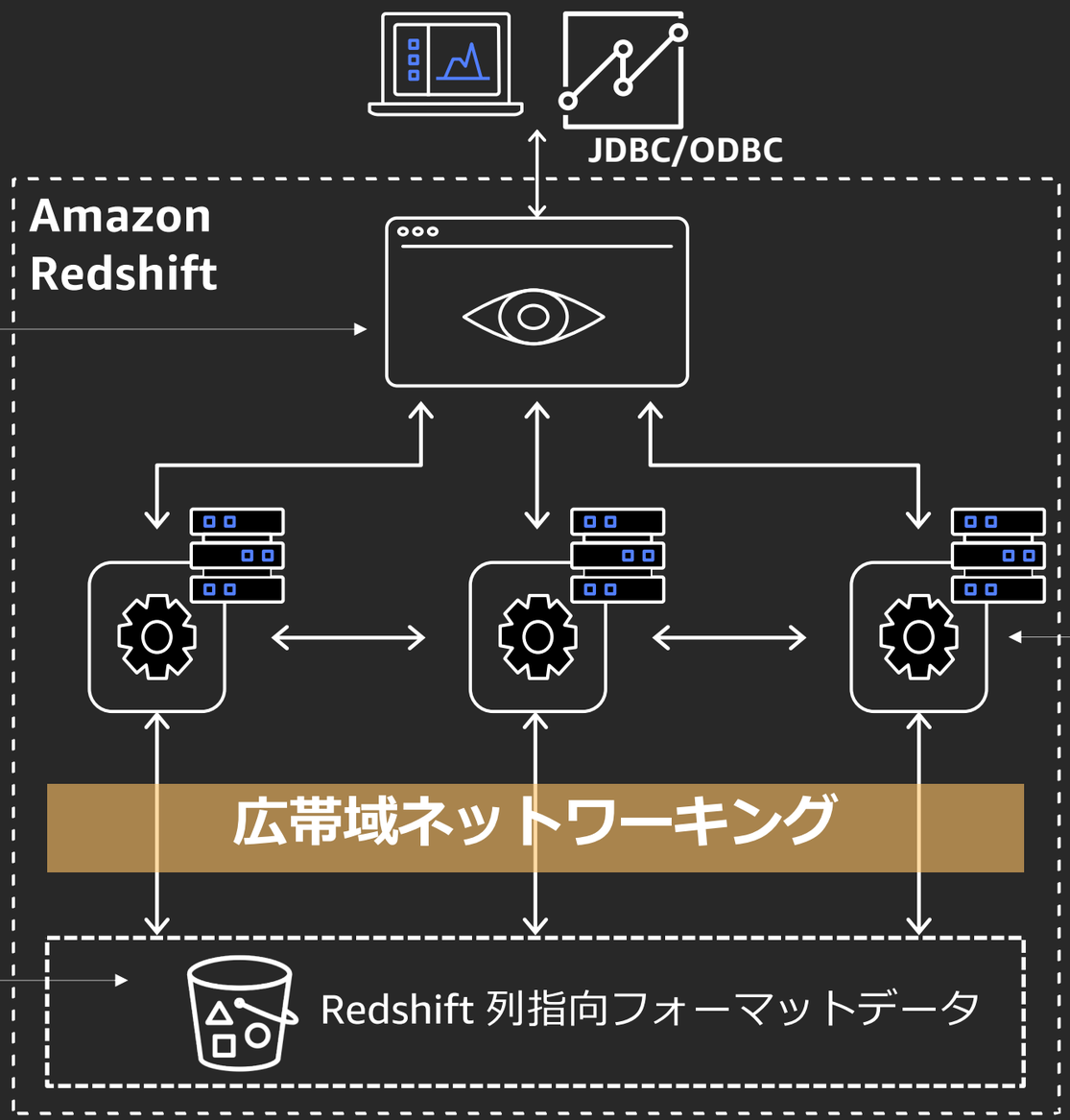
Amazon Redshift とは



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 のアーキテクチャ (RA3)



リーダーノード

- クエリのエンドポイント
- クエリオプティマイザが SQL 処理コードを生成し、コンピュートノードに展開

コンピュートノード

- 高速ローカル SSD キャッシュ＋大容量 RAM 搭載＋広帯域ネットワーキング
- クエリの並列実行エンジン

マネージドストレージ

- Redshift 管理 S3 バケット
- データの永続ストレージ

これからのデータウェアハウスに求められるもの

1 **柔軟性とスケーラビリティ**

事業のフェーズやビジネスの変化に応じてコスト効果高く、

柔軟に構成を変えたり、必要なときに必要なリソースを拡張できること

2 **高いベースラインパフォーマンスと運用管理機能**

デフォルトで高いパフォーマンスを出しつつ、

ユーザーによるチューニング余地も持てること

3 **データレイク連携とセキュリティ**

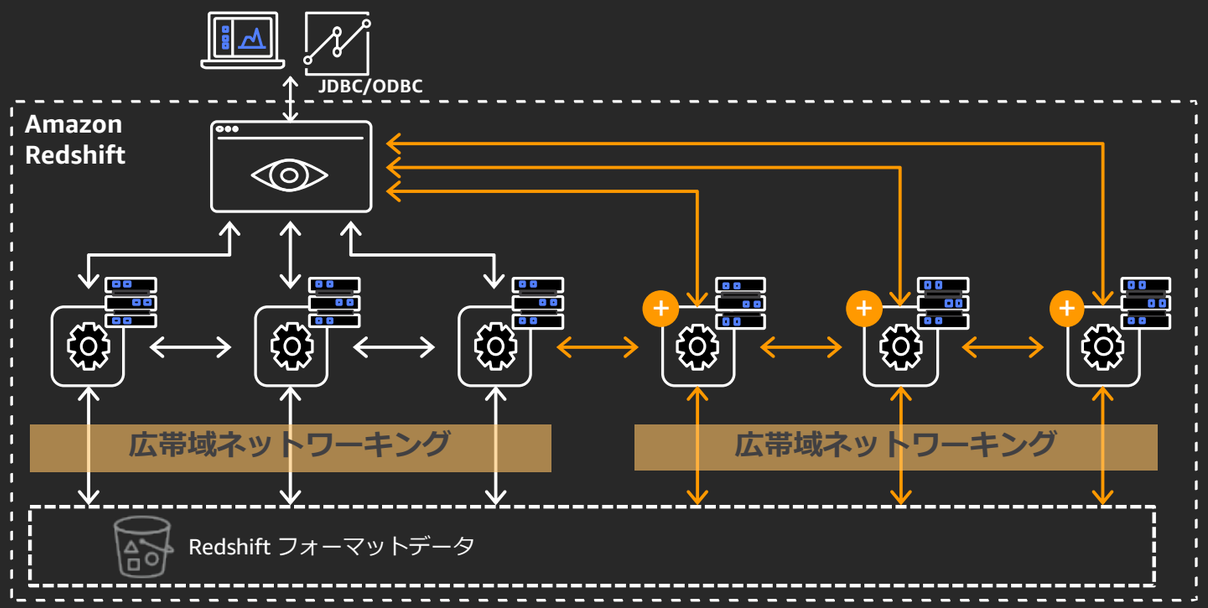
DWH をデータレイクにまで簡単に拡張し、コストを抑えた形でデータを分析可能にし、

さらに**きめ細やかな権限の管理ができる**こと

柔軟性とスケーラビリティ

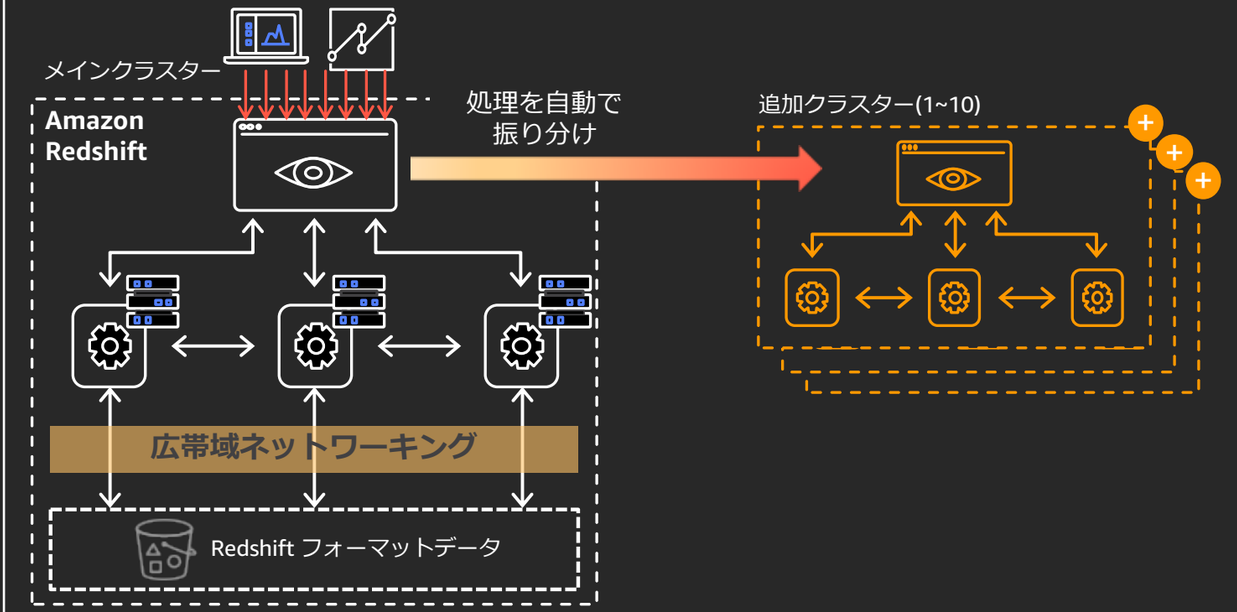
Elastic Resize

必要なときに必要な分のリソースを自動的にスケールアップ/アウト



Concurrency Scaling

同時アクセスが急増した場合でも、クラスターを自動追加して並列処理



これからのデータウェアハウスに求められるもの

1 **柔軟性とスケーラビリティ**

事業のフェーズやビジネスの変化に応じてコスト効果高く、
柔軟に構成を変えたり、必要なときに必要なリソースを拡張できること

2 **高いベースラインパフォーマンスと運用管理機能**

デフォルトで高いパフォーマンスを出しつつ、
ユーザーによるチューニング余地も持てること

3 **データレイク連携とセキュリティ**

DWH をデータレイクにまで簡単に拡張し、コストを抑えた形でデータを分析可能にし、
さらに**きめ細やかな権限の管理ができる**こと

高いベースラインパフォーマンスと運用管理機能

分析に特化したアーキテクチャ

- 列指向ストレージとクエリオプティマイザ
- SQL の並列分散実行
- 結果キャッシュ

継続的なパフォーマンス最適化

- 継続的パフォーマンス改善
- テーブルメンテナンスやワークロード管理の自動化
- 実行時間の短いクエリのスループットを高速化

運用管理機能を活用したチューニング

- ワークロード管理 (WLM) とクエリ優先度
- マテリアライズドビュー
- チューニングアドバイザ

高いベースラインパフォーマンスと運用管理機能

分析に特化したアーキテクチャ

- 列指向ストレージとクエリオプティマイザ
- SQL の並列分散実行
- 結果キャッシュ

継続的なパフォーマンス最適化

- **継続的パフォーマンス改善**
- テーブルメンテナンスやワークロード管理の自動化
- 実行時間の短いクエリのスループットを高速化

運用管理機能を活用したチューニング

- **ワークロード管理 (WLM) とクエリ優先度**
- マテリアライズドビュー
- チューニングアドバイザ

ベースラインパフォーマンスが半年で 2 倍以上に

継続的なパフォーマンス最適化

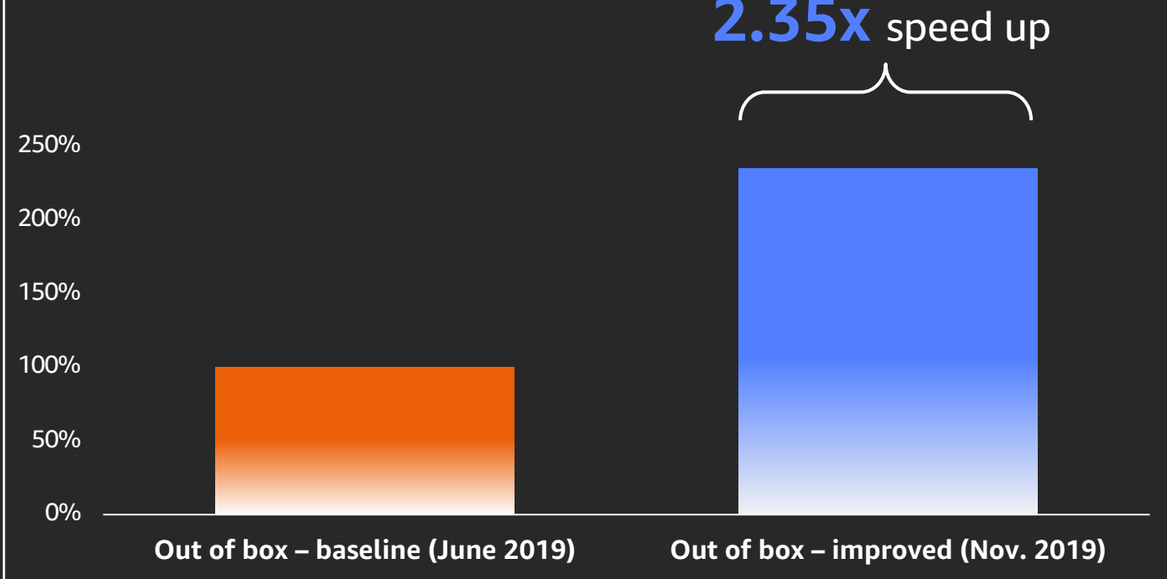
- AZ64 エンコーディング
- JOIN 時のブロードキャストや
  コロケーションのための Bloom filter
- モダン HW の CPU/ネットワークに
  対応したプランナーの改良
- HLL (HyperLogLog) での
  Count Distinct 高速化
- キャッシュ最適化した集計/JOIN処理

等

Cloud DW 30TB benchmark*

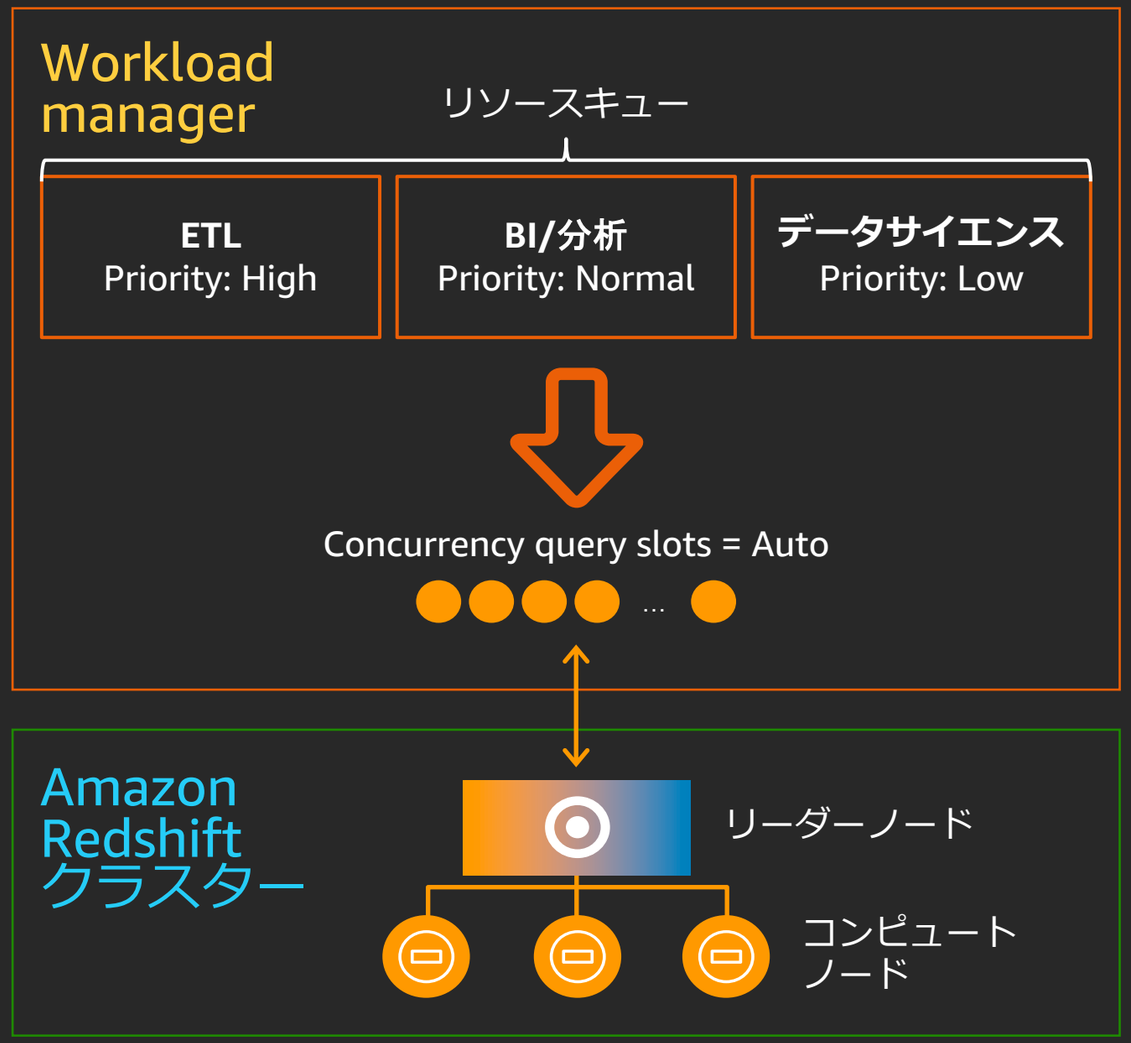
Normalized queries/hour (QPH)

(値が大きい方が高性能)



* Cloud DW benchmark は TPC-DS (v2.10) に基づいており、クエリの修正はしていません

効率的な自動ワークロード管理とクエリ優先度



ビジネスの優先度に応じてリソースキューを分け、優先度を設定

優先度の高い処理には、自動的に多くの CPU, IOリソースを割り当て

簡単な設定で、
高い SLA が求められる処理を
優先的に処理することが可能

これからのデータウェアハウスに求められるもの

1 **柔軟性とスケーラビリティ**

事業のフェーズやビジネスの変化に応じてコスト効果高く、
柔軟に構成を変えたり、必要なときに必要なリソースを拡張できること

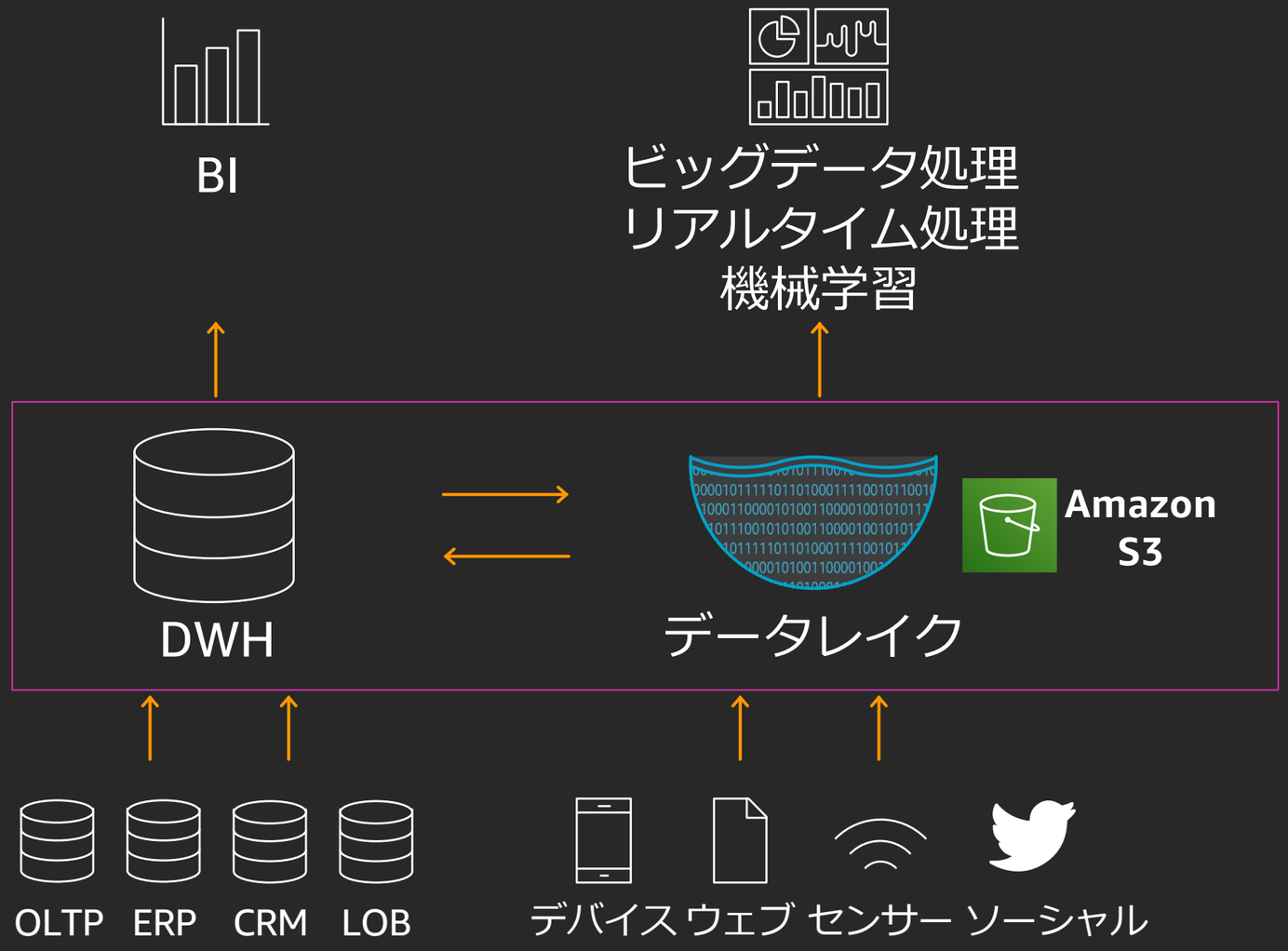
2 **高いベースラインパフォーマンスと運用管理機能**

デフォルトで高いパフォーマンスを出しつつ、
ユーザーによるチューニング余地も持てること

3 **データレイク連携とセキュリティ**

DWH をデータレイクにまで簡単に拡張し、コストを抑えた形でデータを分析可能にし、
さらに**きめ細やかな権限の管理ができる**こと

Amazon Redshift とは



Amazon Redshift

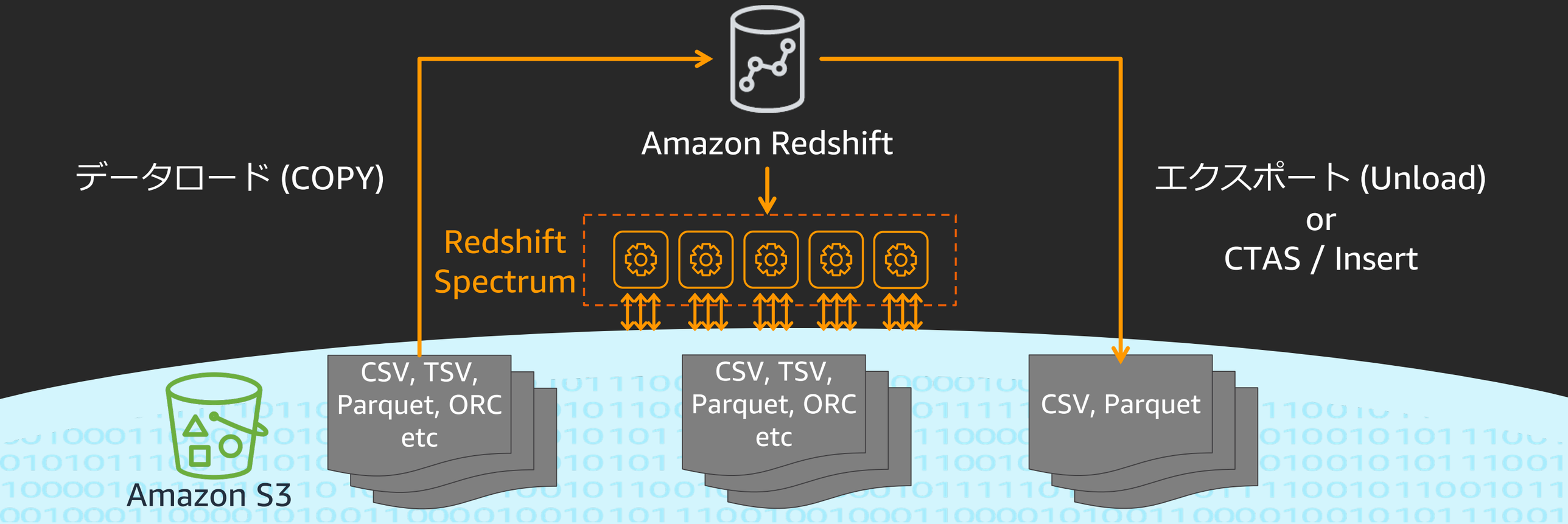
高速、スケーラブルで
費用対効果の高い
データウェアハウス
およびデータレイク
分析マネージドサービス

データレイク連携

Redshift から S3 データレイクに**直接、高速にクエリ可能**

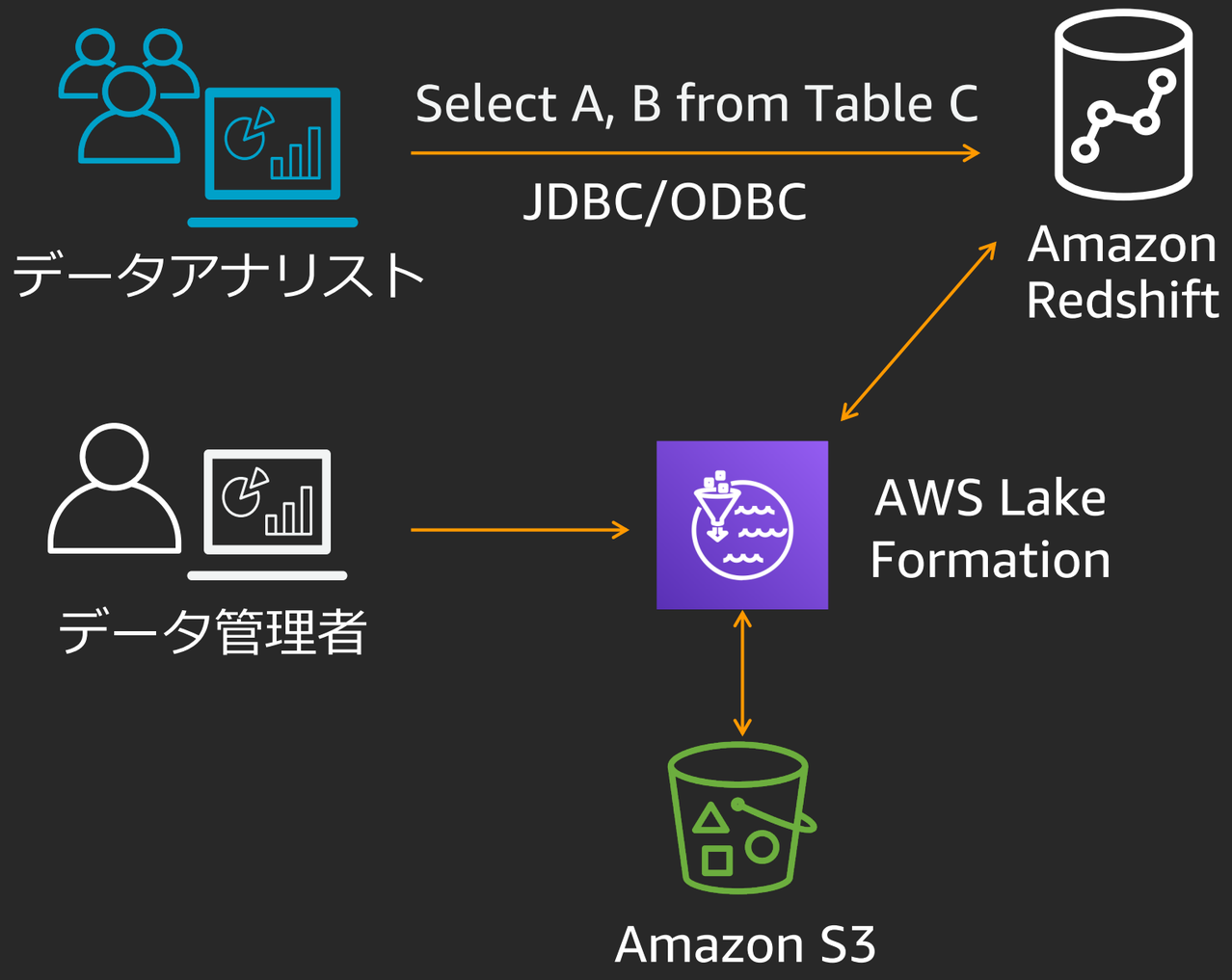
スケールアウト型のコンピュートリソース Redshift Spectrum が自動的に利用される

S3 データレイク上の**データのインポート/エクスポート**も可能



データレイク連携とセキュリティ

AWS Lake Formation と Redshift Spectrum の連携



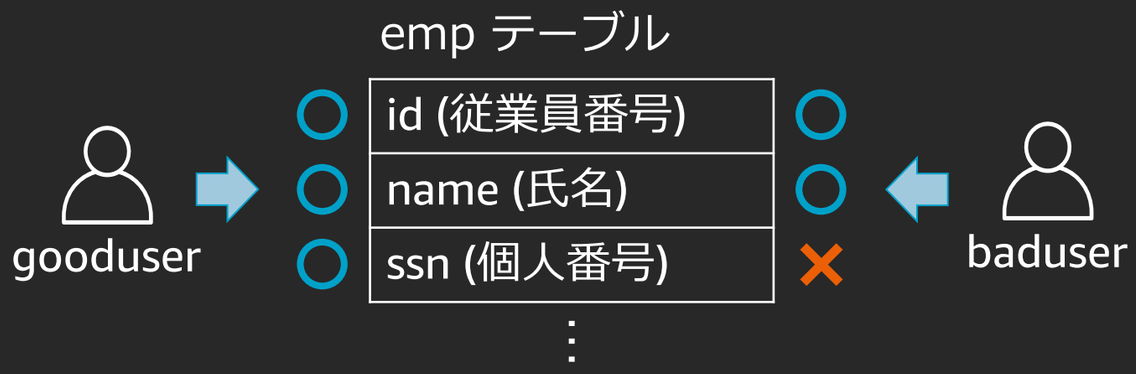
AWS Lake Formation と Amazon Redshift を連携し、**権限管理の一元化**が可能

Lake Formation 経由で Redshift Spectrum がアクセスするオブジェクトに対し、**テーブルや列レベルでのアクセスコントロール**が可能

Amazon Redshift のアクセスコントロール

- 幅広い粒度でアクセスコントロール設定可能
  - データベース、スキーマ、テーブル、列、関数、プロシージャ

- **列レベル**アクセスコントロールの設定
  - 既存のオブジェクトの構造を変えずにセンシティブデータのセキュリティを保護




```
GRANT select(id, name, ssn) on emp to gooduser;
REVOKE select(ssn) on emp from baduser;
```

Amazon Redshift による解決

柔軟性とスケーラビリティ → **Elastic Resize**
Concurrency Scaling

高いベースラインパフォーマンス
と運用管理機能 → **分析に特化したアーキテクチャ**
継続的なパフォーマンス最適化
運用管理機能を活用したチューニング

データレイク連携とセキュリティ → **Redshift Spectrum**
S3 データの Load/Unload
簡単できめ細やかな権限管理機能

まとめ

aws SUMMIT ONLINE

まとめ

従来の DWH におけるよくある課題には以下のようなものがあった

1. ビジネス変化への対応が困難
2. パフォーマンス SLA の維持
3. データのサイロ化と権限管理

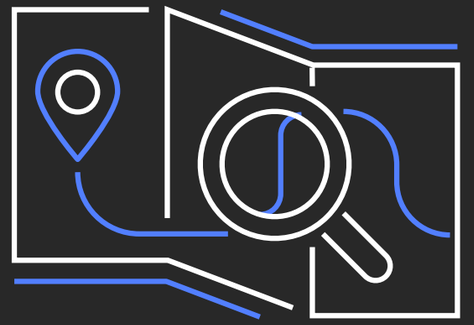
システムのライフサイクルや制限にビジネス要件を合わせるのではなく、
ビジネスの変化に対応しやすい DWH を選択することが重要

データレイクを中心とした形で Amazon Redshift を活用することにより、
上記の課題を解決し、新しい形の分析環境を作っていくことが可能

Thank you!

aws SUMMIT ONLINE

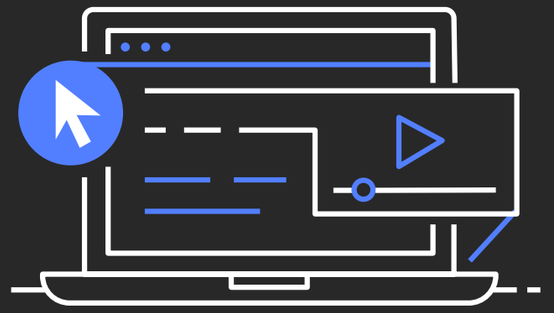
AWS トレーニングと認定



クラウド人材の育成

AWS トレーニングを活用し、ビジネスを牽引する人材の育成と組織作りを促進する

AWS トレーニング活用事例 »



自習コンテンツの活用

ウェビナーやのデジタルトレーニングを受講して、個人のスキルアップを目指す

AWS デジタルトレーニング »

AWS 認定取得を目指す

認定取得を目指して知識を底上げし、AWS の経験とスキルを証明する

AWS 認定の詳細 »

学習パスをお探しの方に

日本語版ランプアップガイドを公開しました。AWS ウェブページ、無料のデジタルトレーニング、クラスルームコース、動画、ホワイトペーパー、認定等を含んだ、9 種の役割別学習ガイドをご覧いただけます。 詳細を見る »

aws.amazon.com/training