Databricks Data Lakehouse & Data Intelligence Platform

従来のデータプラットフォームの課題と革新的な解決策

2024年6月27日

従来のデータプラットフォームの課題

組織における一般的なデータインフラの問題点

複数ツールの分散

データウェアハウス、ETL、データレイク、オーケストレーション、AIなど異なるツールが分散し、運用が複雑化

統合の困難さ

異なるツール間の連携が難しく、一貫したデータフローの確立に多大な労力が必要

ベンダーロックイン

プロプライエタリ形式によるデータのロックイン、ベンダー依存のリスク増大

データサイロ化

同一データが複数の場所に存在し、整合性の維持や管理が複雑化

これらの課題により、データの価値を最大限に活用できず、コストと管理の負担が増大します。

Databricksの統合ソリューション

データプラットフォームの複雑さを解消するための統合データ環境

統合データプラットフォーム

Databricksはデータプラットフォームに必要なすべてのツールを緊密に統合し、単一プラットフォームで提供します

データウェアハウジング

従来の複数ツールに代わり、単一プラットフォーム上で高性能なデータウェアハウス機能を提供

ETL処理

データ抽出・変換・ロードを同一環境で効率的に実行、シームレスな連携を実現

データレイク管理

大規模データの保存と処理を最適化、オープンフォーマットで柔軟なアクセス

AI/MLソリューション

データサイエンティスト向けの高度な機械学習・AIツールを同一プラットフォーム上で提供

統合ソリューションの主なメリット：複雑な連携作業の排除、一貫したセキュリティポリシー、統合されたデータガバナンス、運用コスト削減

オープンソースソリューションの利点

ベンダーロックインからの解放と柔軟性

ベンダーロックイン解消

プロプライエタリ形式による制限から解放され、自由にデータにアクセス可能

柔軟なクラウド選択

AWS、Azure、GCPなど、任意のクラウドプロバイダーでデータを管理可能

オープンフォーマット

Parquet、CSV等の標準フォーマットでデータを保存し、相互運用性を確保

拡張性と移植性

必要に応じて別のプラットフォームへ移行可能、長期的な技術投資の保護

Delta Lake - オープンソースの中核技術

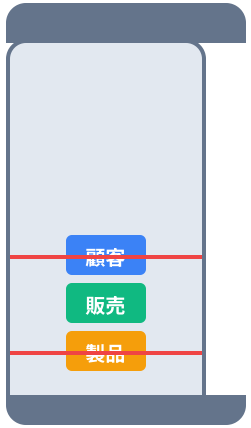
Delta Lakeは、データレイク上にデータウェアハウス機能を実現するオープンソース技術。ACID準拠のトランザクション、スキーマ適用、タイムトラベルなどの機能を提供します。

データサイロの問題

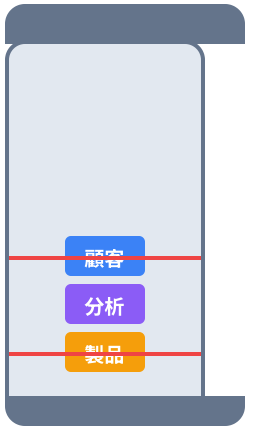
データの分断がもたらす非効率と課題

データサイロとは

データサイロとは、組織内で同じデータが異なるシステムや場所に分散して保存され、相互に連携しにくい状態のことを指します。



データレイク



データウェア
ハウス

主な問題点

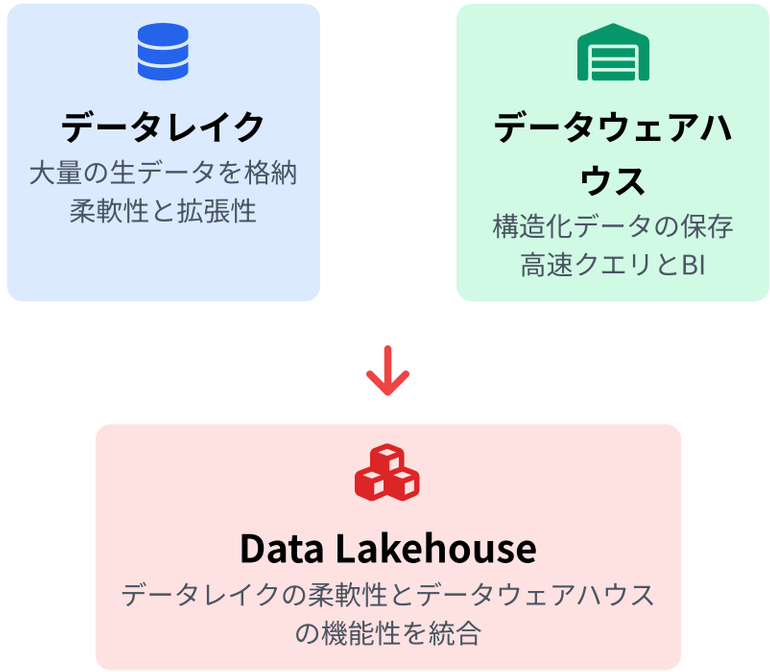
- 同じデータが複数の場所に重複して存在
- データのオーナーシップが不明確
- データ整合性の維持が困難

ビジネスへの影響

- データ管理コストの増大
- 分析の遅延とデータの鮮度低下
- データドリブンな意思決定の妨げ

Data Lakehouseとは

データレイクとデータウェアハウスの利点を兼備した次世代データプラットフォーム



Data Lakehouseの主な特徴

構造化・非構造化データの統合
あらゆる種類のデータを一元管理し、統合分析が可能

高性能クエリエンジン
データウェアハウス並みの分析性能とBI機能を提供

オープン形式でデータ保存
ベンダーロックインを回避し、柔軟なデータ活用

トランザクションのサポート
ACID準拠のトランザクション、バージョニング機能

Data Lakehouseの登場により、データコピーの削減、管理の簡素化、コスト削減、そして分析とAIの統合が実現しました。

Delta Lakeの利点

Data Lakehouseを実現する基盤技術

ACIDトランザクション

RDBMSのようなトランザクション機能をサポートし、データ整合性を保証

バージョニング

データの変更履歴を保持し、いつでも過去のバージョンにアクセス可能

タイムトラベル

過去の任意の時点のデータを復元可能、障害発生時のデータ回復にも有効

監査履歴

誰がいつどのようにデータを変更したかの完全な記録を提供

スキーマ強制と進化

データ構造を確実に保護しながら、スキーマの柔軟な進化をサポート

統合アーキテクチャ

バッチ処理とストリーミング処理を統合し、同一データに対する様々な分析を可能に

Delta Lakeはオープンソースプロジェクトとして提供され、Databricks Data Lakehouseプラットフォームの中核技術として機能しています。

Databricks Lakehouse Platform

統合データプラットフォームのアーキテクチャ

主な特徴:

- 複数クラウド対応
- オープンソース技術
- 統合ガバナンス
- 複数ペルソナ向けツール

Databricks Data Lakehouse & Data Intelligence Platform

レイヤー構成アーキテクチャ

Databricksプラットフォームは複数の統合レイヤーで構成されています

データツール層

- **データエンジニアリング**
  ジョブ、ワークフロー、ノートブック
- **データアナリスト**
  SQL、ダッシュボード
- **データサイエンティスト**
  MLソリューション

データインテリジェンスエンジン

高度な分析と洞察を提供する統合エンジン

Unity Catalog

統合ガバナンス、セキュリティ管理、メタデータ管理

Delta Lake

データレイクハウスエンジン - オープンソースのトランザクションストレージ

データレイク

オープンフォーマット（Parquet, CSV等）でのデータ保存

クラウドプラットフォーム

複数クラウドプロバイダーに対応

Azure　GCP　AWS

Data Intelligence Platformとは

次世代のデータ活用基盤の全体像

企業データからの高度なインサイト

自然言語クエリによるデータ探索と分析を可能にし、専門知識がなくてもデータからインサイトを抽出できます

コード生成と自動化

AIによるコード提案や自動生成により、データエンジニアリング・分析作業を効率化

統合プラットフォーム

データエンジニア、アナリスト、データサイエンティストなど全ペルソナのための単一環境を提供

ガバナンスとセキュリティ

Unity Catalogによる一元的なガバナンス体制でデータの安全性と管理性を確保

Databricksの「Data Intelligence Platform」は、データレイクの柔軟性、データウェアハウスの管理性、そして最新のAI技術を組み合わせた次世代のデータプラットフォームです。

まとめと次回予告

主なポイント

Databricksは統合データソリューション
複数ツールの分散問題を解消

オープンソース技術の活用
ベンダーロックインからの解放

Data Lakehouse
データレイクとデータウェアハウスの統合

Data Lakehouseのメリット

- データの一元管理による重複排除
- オープンフォーマットによるデータアクセス性向上
- トランザクションサポート、バージョニング機能
- AIと分析ワークロードの統一プラットフォーム
- Delta Lakeによるデータ整合性の保証

→ 次回予告

次回はDatabricksのアーキテクチャの詳細について解説します。各コンポーネントの役割と連携についてより深く学びましょう。