

データアーキテクチャシリーズ

オープンデータレイクハウス

Clouderaで最新のデータレイクハウスを構築する



概要

このホワイトペーパーでは、データレイクハウスのアーキテクチャーについて紹介します。データレイクハウスとは何か、データレイクハウスはなぜ作られたのか、そしてデータレイクハウスで対処できる課題について説明し、Cloudera ベースのリファレンスアーキテクチャー、そしてレイクハウスを拡張できる重要な2つの領域を紹介します。

バージョン: 1.0
作成者: Daniel J Hand

目次

概要	2
はじめに	4
対象読者	4
目的	4
推奨資料	4
データレイクハウスアーキテクチャーとは	5
定義	5
起源	5
特性	5
データレイクハウスアーキテクチャーが有用な理由	6
データレイクとデータウェアハウスの制約	6
制約の克服と特性の実現	7
データ重複の回避	7
分析エンジンのオープンエコシステムのサポート	7
柔軟なハイブリッド導入オプション	7
オープンデータレイクハウスの構築	8
Cloudera Data Platform (CDP)	8
Shared Data Experience (SDX)	9
データハブ	9
データサービス	9
Cloudera Data Engineering (CDE)	9
Cloudera Data Warehouse (CDW)	10
Cloudera Machine Learning (CML)	10
データカタログ	10
管理コンソール	10
Apache Iceberg — オープンテーブルフォーマット	10
データ品質	12
データレイクハウスの先へ	12

はじめに

このセクションでは、このホワイトペーパーを作成した理由、対象者、このホワイトペーパーを読むべき理由、推奨資料を簡単にまとめています。

対象読者

このホワイトペーパーは、エンタープライズデータプラットフォームチームのアーキテクト、リーダー、運用リーダー、エンジニアリングリーダー、ビジネスリーダー向けに書かれたものです。また、特に Cloudera の製品やサービスに適用されるデータレイクハウスアーキテクトへの理解を深めたいと考えている最高データ責任者 (CDO) や最高情報責任者 (CIO) にとっても有益な読み物となるでしょう。

目的

データレイクハウスは、データメッシュ、データファブリックと並ぶ新しい3大データアーキテクトの1つです。各アーキテクトの概要、重要性、ハイブリッド環境に大規模に導入する方法をよく理解しておく必要があります。

Cloudera は、データレイクハウスアーキテクトの導入を数年間にわたって支援してきました。最近では、クラウドネイティブなデータサービスとして複数のデータ分析サービスを導入し、新たなテーブルストレージフォーマットを採用したことで、Cloudera はトランザクション、データ/テーブルのバージョンング、スナップショットなど、データウェアハウスの重要な管理機能を完全にサポートできるようになりました。

推奨資料

以下の推奨資料は、このホワイトペーパーの直接的な裏付けとなる情報源だけに限定しています。

- Cloudera の公式ブログ：英語 日本語
- Exploring Lakehouse Architecture & Use cases (英語)、Gartner®、2022年1月 (gartner.com へのアクセス権が必要です)
- Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics (英語)
- Introducing Apache Iceberg in Cloudera Data Platform (英語)
- Cloudera Data Platform (CDP) で Apache Iceberg を使用する5つの理由
- A comparison of Data Lake table formats (英語)

データレイクハウスアーキテクチャーとは

このセクションでは、データレイクハウスのアーキテクチャーについて紹介します。データレイクハウスの起源、データレイクハウスで対処できる課題、変わりゆく一般的な定義、そして改善の余地について考察します。

定義

Gartner® 社は、データレイクハウスのアーキテクチャーまたはパラダイムを次のように定義しています。

「データレイクハウスは、AI、BI、ML、データエンジニアリングを単一のプラットフォームでサポートすることを目的として、データウェアハウスとデータレイクの機能を統合および統一します*」

この定義はこれまで徐々に拡大され、より多くの分析サービスが含まれるようになりました。Cloudera では、この傾向が今後も続き、リアルタイムストリーミング分析や運用データストアの領域も含まれるようになると予想しています。

起源

データレイクハウスという言葉が初めてエンタープライズデータプラットフォームの用語として登場したのは、2017年のことです。Jellyvision が構造化データの処理 (データウェアハウス) とスキーマレスシステム (データレイク) をどのようにして結び付けたのかを説明するために使われました。これら2つのアーキテクチャーパラダイムの統合がデータレイクハウスの誕生へとつながりました。

それ以来、この言葉の定義は進化を続けており、機械学習 (ML) などの分析サービスが追加されたほか、従来のデータウェアハウスの管理機能もさらに拡充されています。

特性

最新のデータレイクハウスには、データレイクとデータウェアハウスのメリットを低い TCO で実現することが求められます。そのため、以下のような重要な特性を備えている必要があります。

- オープンで柔軟かつパフォーマンスに優れたファイルおよびテーブルフォーマット (Apache Parquet、Iceberg など)
- ACID トランザクション、テーブルのバージョンニング、スナップショット、ペタバイトレベルの共有
- オープンエコシステムにおける多機能分析
- 強力なデータ管理 (セキュリティ、ガバナンス、リネージ)
- 高いデータ品質と信頼性
- クラス最高の SQL パフォーマンス
- 非 SQL インタラクションのための直接アクセスと宣言型アクセス

* Gartner, Exploring Lakehouse Architecture and Use Cases, Sumit Pal, 11 January 2022 (gartner.com へのアクセス権が必要です) GARTNERは、Gartner Inc.または関連会社の米国およびその他の国における登録商標およびサービスマークであり、同社の許可に基づいて使用しています。All rights reserved.

データレイクハウスアーキテクチャーが有用な理由

このセクションでは、データレイクハウスアーキテクチャーが有用である理由を考察します。データレイクとデータウェアハウスの制約について考察し、データレイクハウスが両方の特性を維持しながらどのようにしてこれらの制約を克服するのかを見ていきます。

データレイクとデータウェアハウスの制約

データレイクハウスアーキテクチャーの人気の高まっている理由を理解するためには、データレイクハウスが取って代わるアーキテクチャーとその制約について考える必要があります。データレイクハウスアーキテクチャーは、ほぼ独立しているデータレイクアーキテクチャーとデータウェアハウスアーキテクチャーに取って代わります。

データはまず、各ソースシステムから ETL 操作によってデータレイクに取り込まれます。従来、これらのソースは主に構造化されたデータを含む運用システムでした。しかし、現在では取り込まれるデータの半数以上が半構造化または非構造化データです。

次に、データは別の ETL 操作によってデータウェアハウスにロードされます。多くの場合、データは基盤となる独自のストレージレイヤで所定の論理データモデルに変換されます。その後、SQL を使用してデータを照会し、トランザクション、テーブルのバージョンニング、スナップショットのサポートといったデータウェアハウスの DBMS 機能を活用できるようになります。

このアーキテクチャーは安価で拡張性に優れたストレージという経済的メリットをデータレイクにもたらす一方で、3つの重要な課題をはらんでいます。

- データの重複: データには複数のコピーが必要です。ソースシステムからデータレイクにデータがコピーされると、そのデータはデータレイクからデータウェアハウスに再びコピーされます。この問題の部分的な解決策として、外部テーブルを使用する方法がありますが、管理 (特に ACID トランザクションのサポート) が限定されるというデメリットがあります。データの重複は3つの課題をもたらします。
 - データの陳腐化 — データウェアハウスのデータは、ほとんどの場合データレイクのデータと同期しておらず、データレイク自体も各ソースシステムのデータと同期していません。
 - データの品質と信頼性 — ソースシステムからデータウェアハウスにデータを取り込む ETL 操作が複数回行われると、障害が発生する可能性が高くなります。異なる処理エンジン間の不整合が品質に影響する可能性があります。
 - コストの増加 — 中間ストレージと ETL 操作の繰り返しによって、それぞれ追加のストレージと処理サイクルが消費されます。
- 分析サービスの限定的なサポート: データウェアハウスは、ビジネスインテリジェンス (BI)、レポート作成、高度な分析のために設計されました。そのため、オープンフレームワークやライブラリを使用した ML には対応していません。ML は一般に大量の構造化データと非構造化データを処理する必要があるため、SQL インターフェースのみを提供するのは非効率です。基盤となるストレージに直接アクセスしたり、データフレームの作業を効率化するために、以前定義したスキーマを活用したりすることはできませんでした。また、リアルタイム分析、運用データストア、一部のデータカテゴリー (時系列など) のサポートも、従来のデータウェアハウスでは大抵不十分です。そのため、複数の分析システムを購入して運用することになり、それぞれがデータ重複の問題を抱えることとなります。
- 柔軟性: 従来のデータウェアハウスは、主にオンプレミスの専用ハードウェアで実行されており、クラウドの選択肢は限られています。大規模な増設では容量を追加購入する必要があり、効率的な拡張ができず、リードタイムも長くなります。

これら3つの大きな制約は、それぞれが複雑さ、信頼性、品質、TCO の増加をもたらします。

制約の克服と特性の実現

データレイクハウスアーキテクチャーは、次の方法によって、これらの制限に対処しながら特性を確保します。

データ重複の回避

ソースシステムからデータレイクにデータをコピーし、さらにデータウェアハウスにコピーする代わりに、データレイクハウスではレイク内の基盤となるデータに抽象化レイヤを提供します。この基盤となるデータレイク上にあるトランザクションメタデータレイヤは、トランザクション、データのバージョンング、スナップショットといったデータウェアハウスの一般的な管理機能をサポートします。データウェアハウスにデータを取り込むための ETL のステップを減らすことで、エラーの可能性を減らし、効率を高め、処理エンジンで不整合が発生する可能性を減らすことができます。

分析エンジンのオープンエコシステムのサポート

BI やレポート作成のために十分な SQL インターフェースを提供することは必要ですが、それでは ML のサポートには不十分であり、かなり限定的です。そのため、データレイクハウスアーキテクチャーを導入するシステムには、レイク内の基盤となるデータに直接アクセスできる機能が必要です。同時に、ML フレームワークには、データサイエンスのパイプライン構築やモデル作成のために、メタデータを活用してデータフレームにデータを取り込むプロセスを簡素化できることが求められます。

柔軟なハイブリッド導入オプション

TCO を大幅に削減するためには、高価な専用ハードウェアから脱却する必要があります。最近のデータレイクハウスアーキテクチャーの導入では、コンピューティングとストレージが分離され、クラウドネイティブなアーキテクチャーが採用されるようになってきました。これにより、それぞれを独立して拡張できるようになり、共有データに対して複数の分析ワークロードを簡単に実行できるようになりました。

データレイクハウスは、オンプレミス、コモディティハードウェア、パブリッククラウドが必要です。オブジェクトストレージやマネージドコンテナサービスを各環境で活用できるクラウドネイティブなハイブリッドソリューションを採用することのメリットがあります。

オープンデータレイクハウスの構築

このセクションでは、Cloudera Data Platform (CDP) について、CDP Public Cloud に焦点を当てて紹介します。次に、Cloudera のオープンデータレイクハウスの基盤となる論理サービスの主なコンポーネントを簡単に説明します。また、Apache Iceberg がどのようにして柔軟で拡張性に優れたテーブルフォーマットを提供し、データライフサイクル全体にわたる複数の分析サービスへのスキーマベースのアクセスをサポートしているかについても説明します。最後に、現在のデータレイクハウスの先を見据え、Cloudera がストリーミング分析をデータレイクハウスにどのように組み込んでいるかについて解説します。

Cloudera Data Platform (CDP)

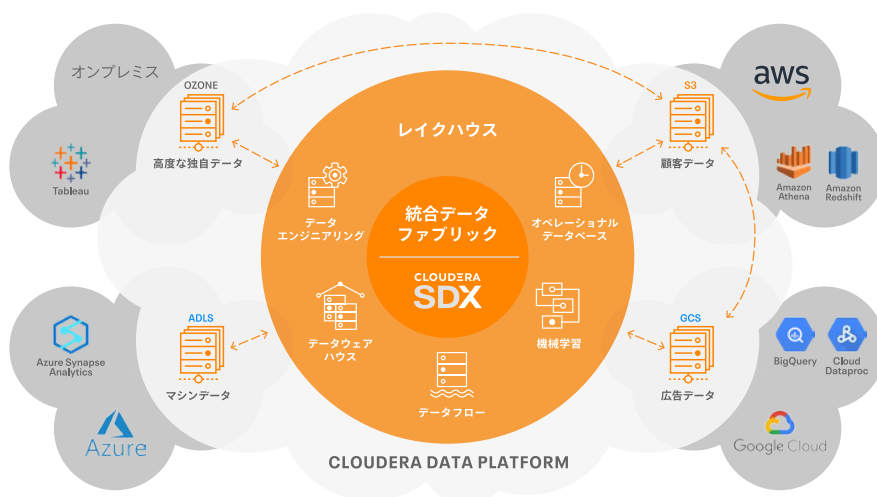


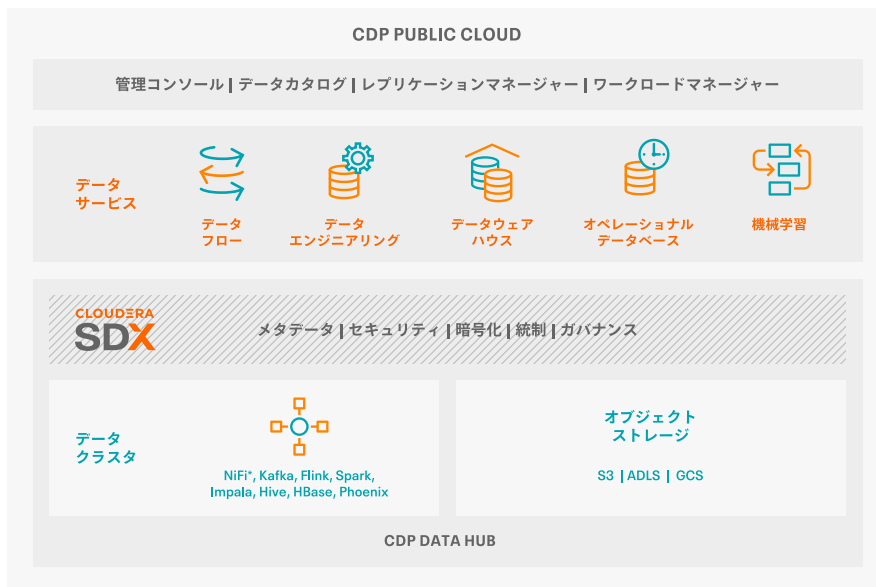
図1 — Cloudera Data Platform (CDP)

Cloudera Data Platform (CDP) は、クラウド、分析、データを自由に選択できるように設計されたハイブリッドデータプラットフォームです。高い拡張性とセキュリティを備え、最適なパフォーマンスであらゆる場所にあるデータに対し、データ管理とデータ分析をすばやく簡単に実行できます。

CDP を利用すると、データがどこに保存されているかに関係なく、データセンターや複数のデータクラウド間でアプリケーション、データ、ユーザーを安全かつ双方向に移動できます。これは、最新の3大データアーキテクチャーを取り入れることで実現できます。

- オープンデータレイクハウスは、ハイブリッドおよびマルチクラウド上のクラウドネイティブなオブジェクトストアで、ストリーミングデータと保存データの両方に対する多機能分析を可能にします。
- 統合データファブリックは、複数のクラウドやオンプレミスで、異なるデータソースに対しインテリジェントかつ安全で一元的なオーケストレーションを行います。
- 拡張性に優れたデータメッシュは、共通のデータインフラストラクチャーを維持しながら、所有権を部門横断型のチームに分散することでデータサイロを解消します。

図2は、CDP Public Cloud を構成するサービスコンポーネントの概要です。ここでは、これらのコンポーネントがどのようにデータレイクハウスアーキテクチャーを支えているのかを見ていきます。



*フロー管理の料金表

図2 — CDP Public Cloud のサービスコンポーネント

Shared Data Experience (SDX)

Cloudera Shared Data Experience (SDX) は、エンタープライズグレードの一元化されたセキュリティ、ガバナンス、リネージ、管理機能と共有メタデータおよびデータカタログを組み合わせて構築されています。クラウドネイティブなオブジェクトストレージにガバナンスレイヤを提供し、データレイクを実現します。

データハブ

データハブを使用すると、ユーザーは分析クラスタをデータライフサイクル全体にわたってエラスティックな IaaS 環境として導入できます。これにより、ハードウェアやインストールされた個々のサービスコンポーネントなどのクラスタ構成を非常に効果的に統制できます。データハブにはクラウドネイティブな設計が採用されており、仮想マシンをコンピューティング単位とすることで、コンピューティングとストレージを分離します。また、環境の変化に応じてリソースを自動拡張することも可能です。

データサービス

データサービスは、動的な拡張と独立したアップグレードが可能なコンテナ化された分析アプリケーションです。Amazon EKS、Microsoft Azure AKS、Google GKE などのクラウドマネージド Kubernetes サービス上にデプロイされたコンテナを使用することで、ユーザーはデータハブと同様のクラスタをデプロイできるほか、それらを PaaS 環境として提供できるというメリットもあります。Cloudera Data Flow (CDF)、Cloudera Data Engineering (CDE)、Cloudera Data Warehousing (CDW)、Cloudera Operational Database (COD)、Cloudera Machine Learning (CML) は、すべて CDP Public Cloud のデータサービスとして利用できます。

Cloudera Data Engineering (CDE)

CDP Data Engineering (CDE) は、クラウドネイティブなデータエンジニアリングデータサービスです。CDE は Apache Spark 上に構築されており、Apache Airflow によるオーケストレーションの自動化、高度なパイプライン監視、ビジュアルトラブルシューティング、包括的な管理ツールにより、エンタープライズ分析チームの ETL プロセスを合理化します。

CDE は CDP と緊密に統合されており、SDX によるエンドツーエンドの把握、セキュリティ、データリネージ、さらには他の CDP サービス (データウェアハウスや機械学習など) とのシームレスな統合が可能です。

Cloudera Data Warehouse (CDW)

Apache HIVE や HIVE ACID を HIVE テーブルフォーマットと組み合わせて使用することで従来のデータウェアハウスの特性の多くを実現できますが、Apache Impala と Apache Iceberg を組み合わせればより広範囲をカバーできます。そのため、データレイクハウスのトランザクションデータウェアハウスエンジンとして、Apache Impala をお勧めします。

現在、Cloudera は Iceberg テーブルの保存と照会をサポートしています。ACID トランザクションのサポートは2022年8月に開始される予定です。HIVE メタストアには Iceberg メタデータが保存され、これにはデータレイク上のテーブルの場所が含まれています。一方、HIVE テーブルフォーマットとは異なり、Iceberg ではデータとメタデータの両方をデータレイクに保存するため、後のセクションでご紹介する多くのメリットがあります。

Cloudera Machine Learning (CML)

Cloudera Machine Learning (CML) は、データサイエンスライフサイクル全体をサポートする機械学習ワークフローソリューションです。CDW と同様に、コンテナを使用してデータエンジニアリングや機械学習のタスクを効率的に実行できるように設計されています。プログラミング言語の Python と R をサポートし、オープンソースの機械学習ライブラリとフレームワークを使用するのが一般的です。

CML は、ML モデルパイプラインの実験とスコアリングをサポートしており、最適な ML アルゴリズムを体系的に選択し、モデルパラメータをチューニングできます。一度トレーニングした ML モデルは、安全な RESTful API を使用してデプロイや管理を行えます。

ML モデルのパフォーマンスを経時的に監視してモデルドリフトを検出できます。パフォーマンスがしきい値を下回った場合は、モデルの再トレーニングと再デプロイを自動的にスケジュールできます。

CML から Iceberg テーブルへのアクセスは、シンプルかつ直感的に行えます。Spark エンジンを使用して、`spark-runtime`、`iceberg-session`、`pluggable-spark-session-catalog` の Iceberg jar を含む接続を作成します。データベースカタログファイルの場所を指定し、種類を「hive」に指定します。これで、Spark SQL を使用してデータベースを操作する準備が整いました。

データカタログ

データカタログは、データレイクハウス全体のデータにアクセスしやすくするための一元的かつ拡張性に優れた手段です。データカタログにより、「どのようなデータがあるのか」「どこにあるのか」「誰が所有しているのか」がわかります。また、データプロファイリング、データリネージ、セキュリティ、分類、監査の機能も備えています。

管理コンソール

管理コンソールを使用すると、CDP Public Cloud、CDP Private Cloud、レガシーバージョンの CDH と HDP を一元的に管理できます。管理コンソールは、各データレイクハウスをサポートするユーザー、環境、分析サービスの管理に対応しています。

Apache Iceberg — オープンテーブルフォーマット

Apache Iceberg は、Apache 財団のオープンソースプロジェクトです。2018年に Netflix 社によってオープンソース化された後、[コントリビュータ企業の強力なコミュニティ](#)によって有数のオープンテーブルフォーマットへと成長しました。コントリビュータ企業には、Tabular 社、Apple 社、Netflix 社、LinkedIn 社、複数のパブリッククラウドベンダー、そしてもちろん Cloudera もその名を連ねています。このコミュニティが一丸となり、Iceberg の迅速なイノベーションを実現すると共に、オープンスタンダード、さらにはオープンエコシステムの実現に向けて取り組んでいます。現在、Iceberg はサードパーティー製エンジンによる **非常に幅広い運用**をサポートしています。

前述したとおり、データレイクハウスにはいくつかの特性があります。これらの特性は、データレイクとデータウェアハウスの特性を組み合わせたものです。処理エンジンと柔軟なテーブルフォーマットを組み合わせただけでは、データレイクハウスを導入したとは言えません。さらに、データレイクの特性も組み込む必要があります。

Iceberg は、ペタバイト規模のテーブルをサポートする柔軟かつオープンなストレージフォーマットを提供します。図3に示すように、Iceberg はデータとメタデータの両方をデータレイクに保存することでこれを実現します。通常、データは Apache Parquet 形式で、関連するメタデータは Apache Avro 形式で保存されます。その後、データカタログ内のエンタリはデータレイク上のマニフェストファイルのポインタとなります。

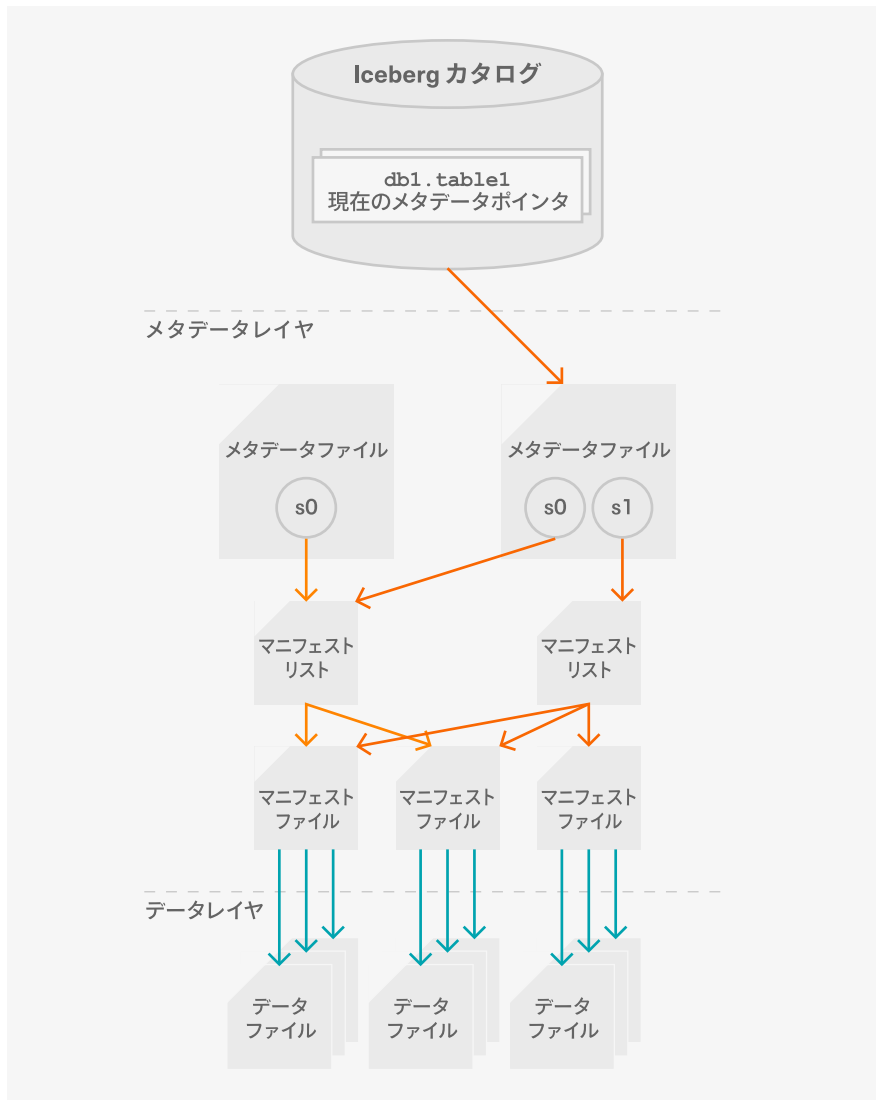


図3 — Apache Iceberg のテーブルアーキテクチャー

Iceberg は、従来のデータウェアハウスの多くの管理機能もサポートしています。これには、トランザクション、データのバージョンング、スナップショットなどが含まれます。Iceberg は、新しいデータのマージ、既存の行の更新、ターゲット削除を実行するための柔軟な SQL コマンドをサポートしています。タイムトラベルを使用すると、全く同じテーブルスナップショットを使用した再現可能なクエリを実行したり、ユーザーが変更を簡単に確認したりできます。バージョンロールバックは、テーブルを以前の既知の状態にリセットすることで、ユーザーが問題を迅速に修正できるようにします。

CDP Public Cloud 上のデータサービスにおける Iceberg のサポートは、2022年6月に開始されました。CDP Private Cloud でも間もなく利用できるようになります。

Cloudera について

Cloudera は、データの力によって、今日不可能なことで明日には実現できると信じています。Cloudera は、複雑なデータを明確で実践的な洞察に転換する力の人々に与えます。Cloudera は、エッジから AI に至るまで、あらゆる場所のあらゆるデータに対応することが可能なエンタープライズデータクラウドを提供します。Cloudera は、オープンソースコミュニティの絶え間ない革新を原動力に、世界最大規模の企業のデジタルトランスフォーメーションを推進していきます。

詳しくは、jp.cloudera.com をご覧ください。

Cloudera とつながる

Cloudera について:

jp.cloudera.com/about.html

Cloudera のブログ:

blog.cloudera.jp

Twitter:

twitter.com/clouderajp

Facebook:

facebook.com/ClouderaJP

YouTube:

youtube.com/c/ClouderaInc

Cloudera コミュニティ:

community.cloudera.com

お客様事例のご紹介:

jp.cloudera.com/about/customers

データ品質

従来のデータウェアハウスでは、3つのステージを経て、データ品質を徐々に向上させていくのが一般的です。このステージは一般に、ランディング、リファイン、プロダクション、あるいはブロンズ、シルバー、ゴールドと呼ばれます。ランディングステージでは、データは csv 形式などの未加工の形式です。データの変換とキュレーションを行いながら、形式を Parquet などに変更し、データモデリングを適用し、効率的な分析に備えて Iceberg テーブルにデータを保存します。この変換によってデータはリファインステージへと移行します。リファインからプロダクションへの最後の移行では、本番環境での使用に備えてデータを最適化する必要があります。これには、データのクレンジングや正規化などが含まれる場合があります。



図4-データレイクハウスの簡略化されたシステムビュー

Cloudera のデータサービスには、Iceberg のクライアントライブラリが付属しています。これにより、3つの品質ステージ間でデータを移行するための変換が可能になります。Iceberg はオープンソースであるため、サードパーティーの製品やサービスと統合してデータ品質操作を行うのも容易です。

データレイクハウスの先へ

前述したとおり、データレイクハウスの定義は、当初はデータレイク上での BI のサポートでしたが、現在では AI、BI、ML、データエンジニアリングを単一のプラットフォームでサポートするものへと着実に進化しています。「データレイクハウスアーキテクチャーとは」のセクションでは、すべてのデータレイクハウスに共通する7つの特性をご紹介しました。さらなる拡張が可能な特性として、分析サービスの追加サポートが考えられます。このように、Cloudera はリアルタイム分析や運用データストアなど、サポートされる分析サービスの拡張に懸命に取り組んでいます。

Cloudera では、単一の処理エンジンをサポートするだけでなく、オープンデータレイクハウスをさらにその先へと拡張する必要があると考えています。現在、Cloudera は Apache Spark、Apache Hive、Apache Impala による Iceberg の実行をサポートしています。これらすべてが、データエンジニアリング、データウェアハウス、機械学習においてデータレイクハウスアーキテクチャーをサポートします。将来的には、リアルタイム分析エンジンの Apache Flink、データフロー管理エンジンの Apache Nifi、Apache HBase を活用した運用データストアなどもサポートしていく予定です。これにより、エッジから AI まで、データライフサイクル全体をカバーする次世代のデータレイクハウスの基盤を実現できます。