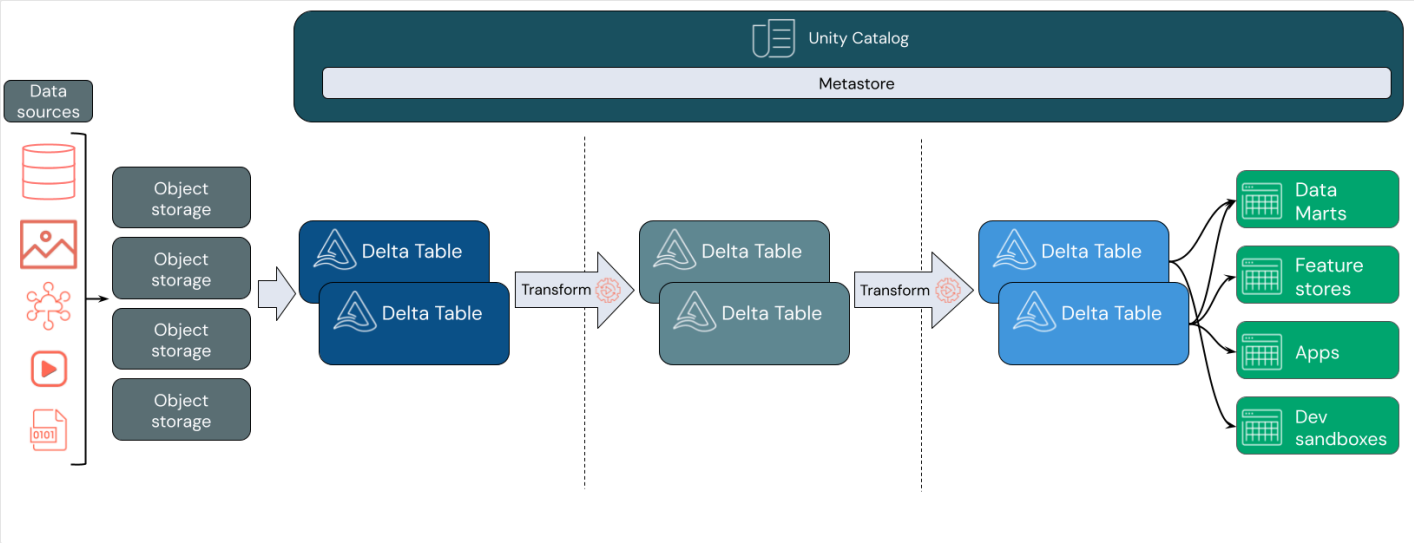
Last updated on ***Oct 1, 2025***

What is a data lakehouse?

A data lakehouse is a data management system that combines the benefits of data lakes and data warehouses. This article describes the lakehouse architectural pattern and what you can do with it on Databricks.



What is a data lakehouse used for?

A data lakehouse provides scalable storage and processing capabilities for modern organizations that want to avoid isolated systems for processing different workloads, like machine learning (ML) and business intelligence (BI). A data lakehouse can help establish a single source of truth, eliminate redundant costs, and ensure data freshness.

Data lakehouses often use a data design pattern that incrementally improves, enriches, and refines data as it moves through layers of staging and transformation. Each layer of the lakehouse can include one or more layers. This pattern is frequently referred to as a medallion architecture. For more information, see What is the medallion lakehouse architecture?

How does the Databricks lakehouse v

Databricks is built on Apache Spark. Apache Spark enables a massively scalable engine that runs on compute resources decoupled from storage. For more information, see Apache Spark overview

The Databricks lakehouse uses two additional key technologies:

- Delta Lake: an optimized storage layer that supports ACID transactions and schema enforcement.
- Unity Catalog: a unified, fine-grained governance solution for data and AI.

Data ingestion

At the ingestion layer, batch or streaming data arrives from a variety of sources and in a variety of formats. This first logical layer provides a place for that data to land in its raw format. As you convert those files to Delta tables, you can use the schema enforcement capabilities of Delta Lake to check for missing or unexpected data. You can use Unity Catalog to register tables according to your data governance model and required data isolation boundaries. Unity Catalog allows you to track the lineage of your data as it is transformed and refined, as well as apply a unified governance model to keep sensitive data private and secure.

Data processing, curation, and integration

Once verified, you can start curating and refining your data. Data scientists and machine learning practitioners frequently work with data at this stage to start combining or creating new features and complete data cleansing. Once your data has been thoroughly cleansed, it can be integrated and reorganized into tables designed to meet your particular business needs.

A schema-on-write approach, combined with Delta schema evolution capabilities, means that you can make changes to this layer without necessarily having to rewrite the downstream logic that serves data to your end users.

Data serving

The final layer serves clean, enriched data to end users. The final tables should be designed to serve data for all your use cases. A unified governance model means you can track data lineage back to your single source of truth. Data layouts, optimized for different tasks, allow end users

to access data for machine learning applications, data engineering, and business intelligence and reporting.

To learn more about Delta Lake, see What is Delta Lake in Databricks? To learn more about Unity Catalog, see What is Unity Catalog?

Capabilities of a Databricks lakehouse

A lakehouse built on Databricks replaces the current dependency on data lakes and data warehouses for modern data companies. Some key tasks you can perform include:

- **Real-time data processing:** Process streaming data in real-time for immediate analysis and action.
- **Data integration:** Unify your data in a single system to enable collaboration and establish a single source of truth for your organization.
- **Schema evolution:** Modify data schema over time to adapt to changing business needs without disrupting existing data pipelines.
- **Data transformations:** Using Apache Spark and Delta Lake brings speed, scalability, and reliability to your data.
- **Data analysis and reporting:** Run complex analytical queries with an engine optimized for data warehousing workloads.
- **Machine learning and AI:** Apply advanced analytics techniques to all of your data. Use ML to enrich your data and support other workloads.
- **Data versioning and lineage:** Maintain version history for datasets and track lineage to ensure data provenance and traceability.
- **Data governance:** Use a single, unified system to control access to your data and perform audits.
- **Data sharing:** Facilitate collaboration by allowing the sharing of curated data sets, reports, and insights across teams.
- **Operational analytics:** Monitor data quality metrics, model quality metrics, and drift by using Data Quality Monitoring.

Lakehouse vs Data Lake vs Data Warehouse

Data warehouses have powered business intelligence (BI) decisions for about 30 years, having evolved as a set of design guidelines for systems controlling the flow of data. Enterprise data warehouses optimize queries for BI reports, but can take minutes or even hours to generate results. Designed for data that is unlikely to change with high frequency, data warehouses seek to prevent conflicts between concurrently running queries. Many data warehouses rely on proprietary formats, which often limit support for machine learning. Data warehousing on Databricks leverages the capabilities of a Databricks lakehouse and Databricks SQL. For more information, see Data warehousing on Databricks.

Powered by technological advances in data storage and driven by exponential increases in the types and volume of data, data lakes have come into widespread use over the last decade. Data lakes store and process data cheaply and efficiently. Data lakes are often defined in opposition to data warehouses: A data warehouse delivers clean, structured data for BI analytics, while a data lake permanently and cheaply stores data of any nature in any format. Many organizations use data lakes for data science and machine learning, but not for BI reporting due to its unvalidated nature.

The data lakehouse combines the benefits of data lakes and data warehouses and provides:

- Open, direct access to data stored in standard data formats.
- Indexing protocols optimized for machine learning and data science.
- Low query latency and high reliability for BI and advanced analytics.

By combining an optimized metadata layer with validated data stored in standard formats in cloud object storage, the Data Lakehouse allows you to work from the same data and in the same platform across different use cases.

Next step

To learn more about the principles and best practices for implementing and operating a lakehouse using Databricks, see Introduction to the well-architected data lakehouse