

Technical Migration Guide

Strategies to Evolve Your Data Warehouse to the Databricks Lakehouse



Contents

- Lakehouse Architecture 3
- The Databricks Lakehouse Platform 4
- Business Value 5
 - Single source of truth 5
 - Data team 6
 - Future-proof 6
- Migration to Lakehouse 7
 - Overview 7
 - Migration strategy 8
 - Migration planning 9
 - ELT approach 12
 - Agile modernization 15
 - Security and data governance 17
 - Team involvement 19
- Conclusion 19

Lakehouse Architecture

Data warehouses were designed to provide a central data repository with analytic compute capabilities to help business leaders get analytical insights, support decision-making and business intelligence (BI). Legacy on-premises data warehouse architectures are difficult to scale and make it difficult for data teams to keep up with the exponential growth of data. Oftentimes data teams publish and use a subset of well-defined data for development and testing. This slows down both innovation and time to insight.

Cloud data warehouses (CDW) were an attempt to tackle the on-premises data warehouse challenges. CDWs removed the administrative burden of tasks such as setup, upgrades and backups. CDWs also improved scalability and introduced cloud's pay-as-you-go model to reduce cost. CDWs leverage a proprietary data format to achieve cloud-scale and performance; however, this also leads to customers locked into these formats with difficult paths to support use cases outside the data warehouse itself (i.e., machine learning). Customers often find themselves with a bifurcated architecture, which ultimately leads to a more costly and complex data platform over time.

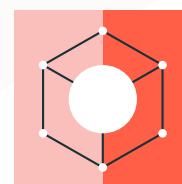
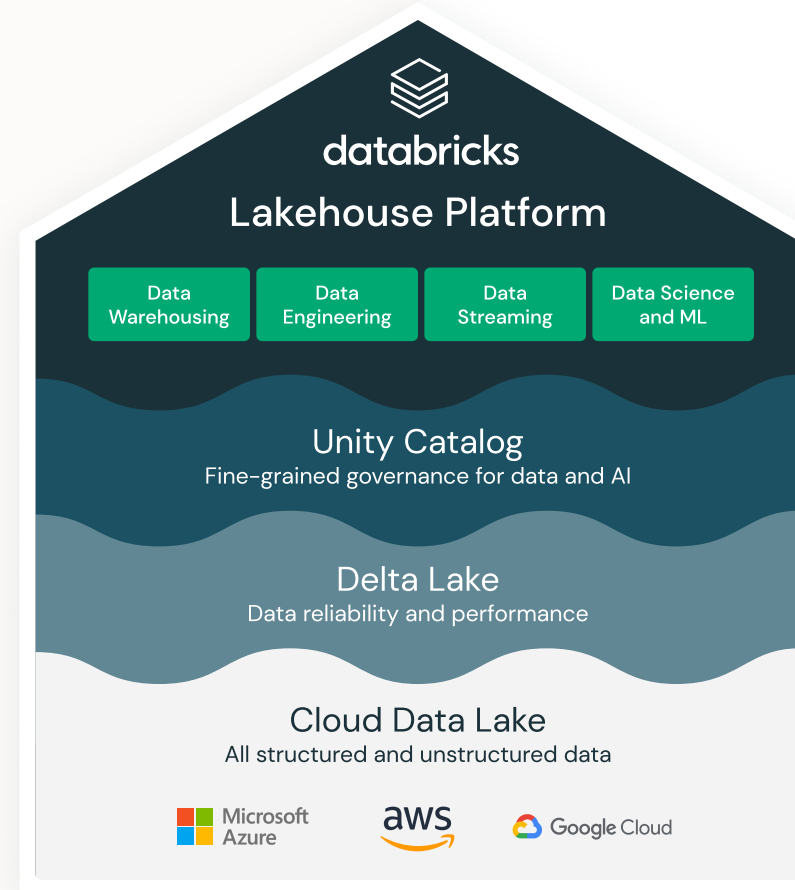
But enterprise data teams don't need a better data warehouse. They need an innovative, simple solution that provides reliable performance, elastic scale and allows self-service to unblock analytics to access all data at a reasonable cost. The answer is the lakehouse.

The lakehouse pattern represents a paradigm shift from traditional on-premises data warehouse systems that are expensive and complex to manage. It uses an open data management architecture that combines the flexibility, cost-efficiency and scale of data lakes with the data management and ACID semantics of data warehouses. A lakehouse pattern enables data transformation, cleansing and validation to support both business intelligence and machine learning (ML) users on all data. Lakehouse is cloud-centric and unifies a complete up-to-date data set for teams, allowing collaboration across an organization.

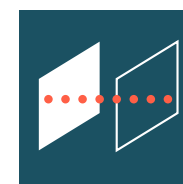
The Databricks Lakehouse Platform

The Databricks Lakehouse Platform is **simple**; it unifies your data, governance, analytics and AI on one platform. It's **open** — the open source format Delta Lake unifies your data ecosystem with open standards and data formats. Databricks is **multicloud** — delivering one **consistent experience across all clouds** so you don't need to reinvent the wheel for every cloud platform that you're using to support your data and AI efforts.

Databricks SQL stores and processes data using Delta Lake to simplify and enhance data warehousing capabilities. Analysts can use their favorite language, SQL, popular transformation tools such as dbt, and preferred BI tools like Power BI and Tableau to analyze data. The built-in query editor reduces contextual switching and improves productivity. Administrators enjoy simplified workload management via serverless compute and auto-scaling to meet high-concurrency workload needs. All this at a fraction of the cost of traditional data warehouses.



Simple



Open



Multicloud

Business Value

Single source of truth

Databricks Delta Lake leverages cloud-based blob storage to provide an infinitely scalable storage layer where you can store all your data, including raw and historical data, alongside structured data tables in the data warehouse. The lakehouse pattern avoids data silos and shares the same elastic scale and governance across all use cases: BI, data engineering, streaming and AI/ML. This means that data engineering teams don't have to move data to a proprietary data warehouse for business analysts or create a separate data store to support data science.

Instead, data teams can access the open format Delta tables directly and combine data sets in the lakehouse, as needed. Data scientists can also work collaboratively on common data with access to versioned history to facilitate repeatable experiments. A single source of truth facilitates moving from descriptive to predictive analytics.



Data team

With central data governance and fine-grained access control capabilities to secure the lakehouse, you can enable self-service SQL analytics for everyone on the Databricks Lakehouse Platform. This allows each team to be more agile and innovate faster.



Data Analysts — Using the Databricks SQL editor or their tools of choice (DBT, Power BI, Tableau), SQL analysts can leverage familiar toolsets.



Data Engineers — Utilizing Delta Lake as a unified storage layer, data engineering teams can eliminate duplicate data and ETL jobs that move data across various systems. Databricks supports both batch and streaming workloads to reduce bottlenecks and serve the most up-to-date data to downstream users and applications.



Administrators — The pay-as-you-go, decentralized compute resource allows each team to run their workload in isolated environments without worrying about contention. Serverless SQL endpoint frees your team from infrastructure management challenges.

The Databricks Lakehouse Platform provides a reliable ETL and data management framework to simplify ETL pipelines. Data teams can build end-to-end data transformations in a single pipeline instead of many small ETL tasks. Databricks supports data quality enforcement to ensure reliability with auto-scalable infrastructure. Your teams can onboard new data sources quickly to power new use cases with fresh data. This not only allows your team to efficiently and reliably deliver high-quality data in a timely manner, it also reduces ETL workload cost significantly.

Future-proof

Unlike CDWs that lock customers in, Databricks offers an open platform with open standards, open protocols and open data formats. It supports a full range of popular languages (SQL, Python, R, Scala) and popular BI tools. You can leverage the performant and low-cost distributed compute layer for data processing — or use a variety of tools and engines to efficiently access the data via Databricks APIs. Databricks also allows data consumption with a rich partner ecosystem. Teams can handle all existing BI and AI use cases with the flexibility to support future use cases as they emerge.

Migration to Lakehouse

Overview

A lakehouse is the ideal data architecture for data-driven organizations. It combines the best qualities of data warehouses and data lakes to provide a single solution for all major data workloads and supports use cases from streaming analytics to BI, data science and AI. The Databricks Lakehouse Platform leverages low-cost, durable cloud storage and only consumes (charges for) compute resources when workloads are running. This pay-as-you-go model means compute resources are automatically shut down if no processing is needed. Data teams can use small clusters that can power individual workloads they plan to migrate. They can make the choice to leverage serverless SQL endpoints and completely free data teams from infrastructure capacity planning and cluster maintenance. The auto-scaling, elastic nature of Databricks clusters leads to significant savings on infrastructure cost and maintenance. Organizations typically achieve 50% TCO savings compared to other cloud data warehouses.

Data warehouse migration is never an easy task. Databricks aims to mitigate the things that can go wrong in these demanding migration projects. The Databricks Lakehouse Platform provides many out-of-the-box features to mitigate migration risks.



CUSTOMER STORY

Building the Lakehouse at Atlassian

[Watch now →](#)



CUSTOMER STORY

Driving Freight Transportation Into the Future

[Read more →](#)

Migration strategy

Migration is a huge effort and very expensive. Yet, almost every enterprise has to migrate to new platforms every 3–5 years because the old platform cannot support new use cases, catch up with data growth or meet scaling needs. To get better ROI on migration, implement a migration strategy that can reduce future re-platform needs and extend to your future data and AI strategy.

Use the opportunity of a data migration to standardize your data in open Delta format to allow existing and future tools to access it directly without moving or converting it. Merge your siloed data warehouses into the unified storage layer in the Databricks Lakehouse Platform — without worrying about storage capacity.

The unified storage layer allows your team to deploy a unified data governance on top to secure all data access consistently. Simplify your data governance story with Databricks Unity Catalog.

Move toward a single, consistent approach to data pipelining and refinement. Merge batch and streaming into a single end-to-end pipeline to get fresher data and provide more real-time decisions. Take a metadata-driven approach to align the dataflow with business processes and have data validation and quality check built-in. Through a series of curation and refinement steps, the output results in highly consumable and trusted data for downstream use cases.

The lakehouse architecture makes it possible for the organization to create “data assets” by taking a stepwise approach to improving data and serving all essential use cases. Encourage your BI/analyst team to leverage Databricks serverless endpoints for self-serve and agility. Each team can evaluate their top priority workloads and migrate them in parallel to speed up migration.

Take advantage of Databricks’ rich partner ecosystem. Your favorite partners are likely already integrated via Partner Connect and can be set up with a few clicks. There are also many ISV and SI consulting partners who can help your migration journey.

Migration planning

Migrating a data warehouse to the cloud can be time consuming and challenging for your data teams. It's important to agree on the data architecture, migration strategy and process/frameworks to be used before undertaking a data migration. Databricks provides Migration Assessment and Architecture Review sessions to develop a joint migration roadmap. This process is designed to help organizations to successfully migrate to a lakehouse architecture. Based on information collected and business objectives, the Databricks team will work with customers to propose a target architecture and provide a tailored migration roadmap.

These assessments help get a full picture of current data systems and the future vision. They clarify what you are migrating and do proper use case discovery. This includes identifying workloads and data source dependency, for example:

Sample migration assessment checklist:

- ☐ Identify upstream data sources and workload dependencies
- ☐ Identify active/inactive data sets and database objects
- ☐ Identify downstream application dependencies and data freshness requirements
- ☐ Define a cost-tracking mechanism, such as tag rules for chargeback and cost attribution
- ☐ Define security requirements and data governance
- ☐ Clarify access management need, document needed permissions per user/group
- ☐ Outline current tooling (ingestion, ETL and BI) and what's needed

