Brought to you by:

About Databricks

Databricks is the data and AI company. More than 10,000 organizations

The Data Lakehouse

Databricks Special Edition

by Ari Kaplan and Amit Kara

for dummies®
A Wiley Brand

The Data Lakehouse For Dummies®, Databricks Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN: 978-1-394-30800-2 (pbk); ISBN: 978-1-394-30801-9 (ebk); 978-1-394-30802-6 (ePub).
Some blank pages in the print version may not be included in the ePDF version.

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Project Manager and Development Editor: Carrie Burchfield-Leighton

Sr. Managing Editor: Rev Mengle

Acquisitions Editor: Traci Martin

Client Account Manager: Matt Cox

Table of Contents

Introduction

The data lakehouse architecture enables companies to deliver faster on their data and artificial intelligence (AI) initiatives. It simplifies your data estate, eliminating data silos by combining the best of two worlds: the flexibility and cost-effectiveness of data lakes and the analytic abilities of data warehouses. Lakehouses are built on open source and open standards, simplifying and unifying all your data, business intelligence (BI), data warehousing (DW), machine learning (ML), and AI within one architecture. And throughout it all, governance is the central source of a robust governance framework, providing end-to-end visibility and control of all your data estate.

The data lakehouse reduces costs, scales to trillions of records, unifies all types of data, simplifies workflows, produces faster analytical and AI insights, scales to trillions of records, and democratizes data to everyone. Historically, many individual point solutions addressed each individual need: a database or data warehouse to store structured historical data, and a data lake to store unstructured data such as documents, images, streaming social feeds, and videos.

The lakehouse radically simplifies the enterprise data and AI infrastructure and accelerates innovation in an age when ML and generative AI (GenAI) are disrupting every industry. This architecture supports structured, semi-structured, unstructured, and streaming data in one unified and governed architecture, providing the fuel for the full spectrum of data-driven use cases.

About This Book

The Data Lakehouse For Dummies, Databricks Special Edition, is about using the principles of a well-designed architecture that leverages the scalable resources of the cloud to manage all of an organization's data assets. This book introduces the lakehouse architecture, the limitations of legacy solutions, and the modern data stack. It explains why a lakehouse is a foundation for solving data challenges and forms the basis for more intelligence insights. You also discover how Databricks specifically builds on the open source architecture and what it all actually means for your company.

Icons Used in This Book

We occasionally use special icons to focus attention on important items. Here's what you find:

This icon reminds you of information that's worth recalling.

Expect to find something useful or helpful by way of suggestions, advice, or observations here, leveraging experiences from other implementations.

Warning icons are meant to get your attention to steer you clear of potholes, money pits, and other hazards. Paying extra attention to these parts of the book helps you avoid unnecessary roadblocks.

This icon may be interpreted in one of two ways: Techies zero in on the juicy and significant details that follow; others will happily skip ahead to the next paragraph.

Beyond the Book

This book helps you understand how the lakehouse makes your DM efforts more effective and efficient in your company. However, because this is a relatively short book to data lakehouses, we also recommend checking out the following:

- Lakehouse overview: `www.databricks.com/product/data-lakehouse`
- Demo of the lakehouse architecture: `https://bit.ly/3YslCsa`
- *The Data Intelligence Platform For Dummies,* Databrick Special Edition: `www.databricks.com/resources/ebook/maximize-your-organizations-potential-data-and-ai`

IN THIS CHAPTER

- Explaining data warehouses
- Positioning data warehouses limitations
- Describing the concept of data lakes

Chapter 1
Making the Case for Data Lakehouses

Data management (DM) consists of methods, architectural techniques, and tools for managing data consistently across a company. The purpose of DM on an enterprise-wide scale is to fulfill all types of data and requirements for use cases, applications, and business processes. DM encompasses a wide range of activities, including governance, quality, integration, and security, that support the effective use of data.

Data warehousing is a component of DM that focuses specifically on storing and analyzing structured data. Data lakes support unstructured data through file formats. The need for performing analytics on all types of data across multiple data sources, as well as running end-to-end artificial intelligence (AI) and business intelligence (BI), puts high demands on DM.

This chapter describes how the approach to managing data has evolved over time, how traditional solutions fall short, and why the data lakehouse architecture has emerged as the modern standard for DM and data warehousing.

Exploring Traditional Data Warehouses

In the early days of DM, the relational database was the primary method that companies used to collect and analyze data. Relational databases offer a way for companies to store and analyze highly structured data, such as numbers, dates, and text, by using Structured Query Language (SQL). For many years, relational databases were simple and reliable ways to meet a company's data needs — until the sheer volume of data increased so much that traditional databases could no longer handle it all. Data grew from billions of records to hundreds of billions and even trillions. Costs spiraled out of control, and insights struggled to be generated in near real time.

The rise of social media, Internet data, mobile, the Internet of Things (IoT), and more led to companies drowning in data. To store all these new types and amounts of data, traditional databases were no longer sufficient. Companies, therefore, often had to build multiple disconnected databases organized by lines of business to attempt to hold all the different data, users, and use cases, often failing.

Sorting out Data Warehouse Limitations

Without a way to centralize and efficiently use their data, companies ended up with decentralized, fragmented stores of data, called *data silos,* across the organization. With so much data stored across different silos, companies needed a way to unify them. *Data warehouses* were born to meet this need and to unite disparate structured databases across the organization.


TECHNICAL STUFF

The concept of data warehousing dates back to the late 1980s and, in essence, was intended to provide an architectural model for the flow of structured data from operational systems to decision-support environments. Early data warehouses were also on-premises, running on hardware fully managed by the company itself. A shift toward cloud data warehousing in the early 2010s had external companies such as Amazon, Google, and Microsoft hosting and managing the hardware that data warehouses ran on.

The shift to cloud-based solutions offered several advantages over traditional on-premises data warehouses. It lowered upfront

costs (operating expenses [OpEx] versus capital expenditures [CapEx]), setup and deployed faster, scaled larger, and improved access across the globe.

WARNING

Traditional data warehouses have inherent limitations that became more prohibitive as data volumes grew significantly larger, and there became a new need to manage unstructured data cost-effectively. These limitations greatly challenged enterprises, which started the push for better, faster, and more flexible DM solutions. The ability to store, manage, and govern a variety of data in a variety of formats had finally arrived.

Diving into Data Lakes

To make analytics possible on a variety of data formats and to address concerns about the cost and vendor lock-in of data warehouses, Apache Spark emerged as the leading open-source distributed data processing technology, replacing Hadoop, which was more limited and cumbersome to manage. These technologies allowed large data sets to be processed with clusters of computers working in parallel.

Listing the Technical Reasons Why a Traditional Data Lake Isn't Enough

While suitable for storing data, data lakes lack some critical features that data warehouses are better for:

- They don't support atomicity, consistency, isolation, and durability (ACID) transactions, which risk corrupting files and data inconsistencies.
- They don't enforce schema or data quality.
- They're inefficient, having to store multiple copies of data; and modifying existing data causes the rewriting of a lot of data when you just want to make short updates.
- Their lack of data consistency and isolation make it almost impossible to simultaneously write and append new data.

- Jobs that fail mid-way lead to data quality issues, are hard to detect, and need to restart from scratch.
- They make it difficult and inefficient to handle large volumes of unstructured data. As the number and size of files increase, performance can degrade, and it gets complex to understand the relationship among your sets of data without predefined schemas.
- Data can proliferate into millions of tiny files or a few gigantic files, often negatively impacting performance.

As the volume and variety of data kept surging the need for a flexible, high-performance DM architecture kept increasing. More than ever, companies require systems for diverse data applications, including SQL analytics, real-time monitoring, data science (DS), machine learning (ML), and AI. Most of the recent advances in GenAI incorporate better models to process unstructured data (text, images, video, audio, and social streaming). Still, these types of data are precisely the types that a data warehouse doesn't support.

WARNING

Without a data lakehouse, multiple solutions must be patched together: several data lakes, data warehouses, ML, and GenAI tools. This introduces additional complexity and cost: Data professionals need to constantly move and copy data among the systems, costing two to three times more to store and maintain all that redundant data. In addition, having all these multiple vendor solutions introduces a lack of unified access control, a lack of one single auditing log, and the cost of multiple vendor contracts.

The Advent of the Data Lakehouse

When data lakehouses came onto the scene, they were a watershed moment because they enabled companies for the first time to analyze massive amounts of both structured and unstructured data together, which before was simply was too costly, too big, too slow, or too complex.

One of the fundamental aspects of a lakehouse is unified data governance that eliminates data silos. Lakehouses unify data warehousing and AI use cases in a single architecture, simplifying the modern data stack for engineering, analytics, BI, data science, ML, and GenAI.

REMEMBER

Open source software including Apache Spark, Delta Lake, and Unity Catalog are the lakehouse's underlying technology that offers many advantages over traditional data lakes and data warehouses:

- Speed through in-memory processing, up to 100 times faster
- Ease of use through the support of Python, R, SQL, and Scala
- Versatility for handling a variety of data processing such as batch and real-time streaming
- Advanced analytics and GenAI
- Fault tolerance to avoid crashing and restarting lengthy processes

What Data Lakehouses Solve for Enterprises

Most organizations struggle to realize a vision that unifies all their data needs. There are so many systems:

- Data warehousing for your structured data and data lakes for unstructured data
- BI platforms to visualize your business insights
- Orchestration and Extract, Transform, Load (ETL) solutions to prepare, merge, filter, and move data
- Real-time systems for streaming use cases
- Data science and ML platforms for advanced use cases such as predictions and classifications
- GenAI and creating AI-driven applications

WARNING

Having all these divergent solutions leads to three groups of problems, as shown in Figure 1-1:

- Enterprises are struggling with the massive sprawl across all these data silos. For each vendor, there are access and security controls, audit trails to keep track of activity, monitoring dashboards, and governance frameworks. This sprawl adds risks, costs, and operational inefficiencies.

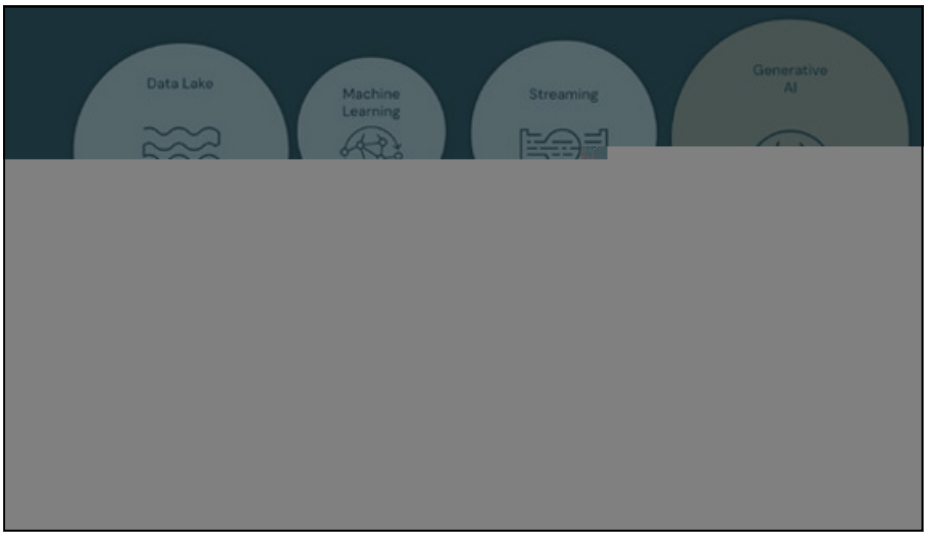

FIGURE 1-1: Age-old challenges that data lakehouses solve.

- Data privacy and control is a massive issue across data silos. With GenAI, a bright light is being shined on being able to transparently understand and manage both the data inputs and the outputs from AI. Having one architecture unifies governance, reducing risks.
- There's a lack of technically skilled employees who can make sense of all these disparate solutions, and they become a bottleneck. Your company relies on them to derive business insights. Having one architecture that democratizes managing data and getting business insights improves your business like never before. Even if you solve the prior problems, most of your company relies on your technical team to create data products.

IN THIS CHAPTER

- Looking at the data and AI maturity curve
- Delineating the technical concepts
- Understanding what lakehouses give you
- Resolving challenges by adopting a lakehouse

Chapter 2
Explaining Data Lakehouses

The data lakehouse architecture combines the best elements of data lakes (for artificial intelligence [AI]) and data warehouses (for business intelligence [BI]) to help you reduce costs and deliver your data and AI initiatives faster. Built on open source and open standards, a lakehouse simplifies your data estate by eliminating the silos that historically complicate data and AI.

Open data lakehouses are underpinned by widely adopted open source projects such as Apache Spark for processing, Delta Lake for storage, MLflow to manage the machine learning (ML) life cycle, and Delta Sharing to securely share live data from your lakehouse to any computing platform without replication and complicated Extract, Transform, Load (ETL).

The best lakehouses are flexible to run on all major cloud providers. They provide the ability for Python, SQL, R, and Scala to run on all your unified data. Lakehouses form the foundations for data intelligence platforms, which open up a whole new world of possibilities to democratize data and AI across an organization. Data intelligence platforms use GenAI with an intelligence engine to understand the semantics of your data and use that across the platform (see Chapter 4).

REMEMBER

Data lakehouses are unified, open, and scalable. They combine the best of data lakes and data warehouses to remove data silos, bring all types of data together in one platform, provide a single unified governance, and simplify it all. This enables your business to deliver data and AI initiatives much more quickly, with more intelligence, transparency, and trust. At the same time, lakehouses reduce operational costs, enable collaboration among all personas, and improve business intelligence, streaming, data science (DS), AI, data warehouse, and orchestration.

In this chapter, you discover all you need to know about lakehouses, including what types of problems this architecture helps to overcome and why this is significantly different from other data warehousing solutions.

Following the Data and AI Maturity Curve

Enabling data intelligence is a journey companies take to enable their companies to be truly data-driven for the best business decisions and outcomes. In order to become a modern data-driven organization, companies typically move along the data and AI maturity curve shown in Figure 2-1.

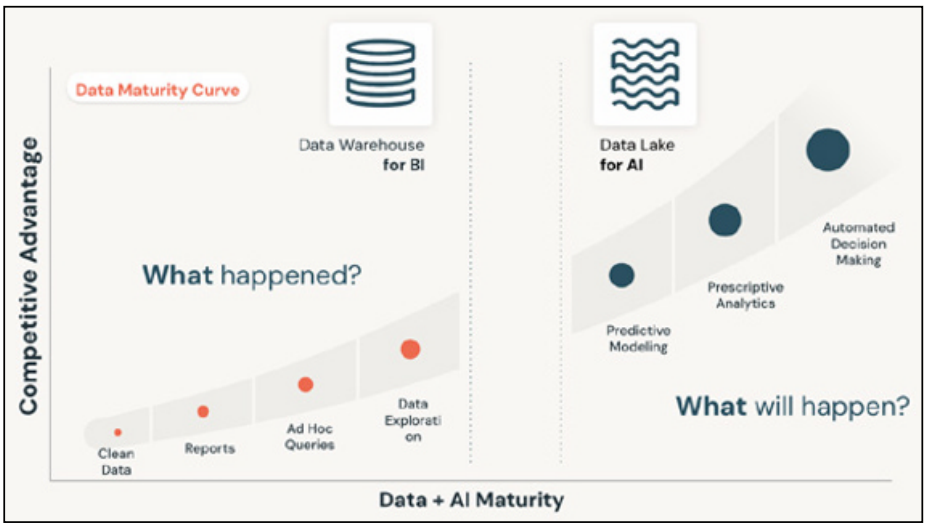

FIGURE 2-1: The data and AI maturity curve.

At the beginning of the journey, companies use databases and data warehouses to see what happened in the past, such as historical sales transactions and activity logs. They obtain structured data, explore it, and provide pre-canned reports and ad hoc queries. As

companies mature, they add data lakes to perform predictive analytics on what may happen in the future. They collect unstructured data such as documents, social media, images, and videos to help them make more intelligent decisions on a variety of data. They want prescriptive analytics to guide them on the best courses of action. The most mature companies go beyond traditional ML by incorporating GenAI on their own proprietary data and automating the decision-making where beneficial. And good news — the lakehouse enables all of this.

Sorting Out the Technical Concepts of a Lakehouse

Data lakehouses take an innovative approach by combining the data warehousing attributes of reliability, performance, and quality with the openness and scale of data lakes. A lakehouse has the following key features:

- **Openness:** The underlying technology leverages open source solutions, which offers benefits, such as lower cost, transparency, flexibility, and avoiding vendor lock-in. Lakehouses leverage open storage formats such as the popular Delta Lake and Iceberg; Unity Catalog (UC) open-source software (OSS) for governance; and MLflow for streamlining the ML life cycle with experiment tracking, model packing and deployment.
- **Decoupled storage and compute:** This separation enables more cost-efficient and scalable systems, unleashing massive amounts of data and concurrent users.
- **Unified governance for data and AI assets:** UC OSS is the central source of a robust governance framework for data in lakehouses. It provides end-to-end visibility and control through audit trails, credential management for different users on different sources of data, transparency such as lineage, data discovery, and data sharing.
- **AI:** Lakehouses support AI — both GenAI and ML — with unified data management so models can use all your corporate data assets. They facilitate ML operations (MLOps) through MLflow to develop, test, and deploy AI models. They provide compliance and governance for models and notebooks themselves — as well as the underlying data. And they enable

collaboration among the data scientists, data engineers, and business analysts into one platform, for shared innovation.

- **BI support**: BI has been the most common way for business workers to get their insights. Lakehouses enable BI tools to directly access the source data, reducing staleness, latency and cost. Instead of needing to maintain multiple copies of data (in a data lake and a warehouse), it can now be stored singularly in the lakehouse.
- **Atomicity, consistency, isolation, durability (ACID) support:** Support for ACID transactions ensures consistency as multiple parties concurrently read and write data, typically using SQL.
- **Diverse data types:** The best business insights often come from a variety of data types such as structured and unstructured: images, video, audio, semi-structured data, and text. Lakehouses support this multimodal approach.
- **Diverse workloads:** Support includes DS, ML, and SQL and analytics. Multiple tools may be needed to support all these workloads, but they all rely on the same data repository.
- **Batch and real-time streaming:** Lakehouses support batch processing, which is e f cient for managing large volumes of data by processing groups of transactions collectively. Lakehouses also support streaming data — from social media to the Internet of Things (IoT) — to be ingested and analyzed as soon as it's received.

Knowing What Data Lakehouses Bring to the Table

In the past, decision-making was mainly based on structured data. Today's DM systems must be much more flexible and also support unstructured data in any format, enabling advanced AI techniques.

REMEMBER

With the lakehouse approach, this flexibility is achieved by deeply simplifying the data infrastructure to accelerate innovation. This is especially important because AI is revolutionizing all industries and demands an elastic infrastructure that supports speed and operational efficiency.

Multimodal support of a variety of data types

One significant difference among approaches is the variety of data being managed. While a data warehouse only handles structured data for most modern analysis and reporting, it's essential to incorporate all types of data: structured, unstructured, batch, and real time.

Lowers overall costs and avoids vendor lock-in

Vendor lock-in happens when a customer becomes dependent on a particular vendor for its solutions and services, making the customer unable to use another vendor solution without substantial costs to switch. This issue can lead to companies paying many license fees and being forced to pay for creating multiple data copies and writing custom code to make data accessible across third-party systems. This doesn't work for making your architecture future-proof.

WARNING

Legacy data warehouses come with significant operational costs and vendor lock-in, which make solutions inflexible and less cost-efficient than the lakehouse approach, which comes with low operational costs and no vendor lock-in, making the data architecture future-proof.

Ability to scale and manage all types of workloads

Lakehouse architecture provides nearly limitless scalability because it decouples the storage and compute, meaning you can scale one without necessarily needing to pay for the other. Scalable solutions refer to systems being able to greatly grow the amount of data and increase workloads. This contribution is essential to a company's competitiveness, quality, and reputation. Lakehouses can also handle all types of data workloads: big and small; long-running and quick-retrieval; batch processing, real-time analytics, and ML/AI. Organizations value this versatility, better pricing structure, and easier management over traditional data warehouses.

Data lakehouses also bring serverless compute to the table. This feature allows workloads to be run automatically without the need for humans to pre-provision and manage the underlying infrastructure and workloads. It enables people to automate the time-consuming server management tasks and instead focus on their more important tasks. It simplifies complex cloud policies to on-demand, quicker deployment of compute and workflows, leading to efficiencies and more optimal resource allocation.

Support for BI, ML, and AI

Lakehouses support BI reporting and dashboards through modern data warehousing and are future-proof with support for ML and AI, real-time (streaming) data, and managing raw data in many major formats.

REMEMBER

The key benefit of the lakehouse is that it allows you to unify all your data and run all your analytics and AI in a single place.

Solving Problems with a Lakehouse

A lakehouse enables business analytics and AI at a massive scale. A lakehouse approach can solve many challenges:

- **Unifying data teams:** All your data teams are unified on one architecture.
- **Breaking data silos:** These are broken by managing all your data in a centralized and governed platform. This enables everyone in your organization to access and manage both structured and unstructured data.
- **Preventing data from becoming stale:** Lakehouses can process both batch and streaming data, such as IoT and social media, continually updating tables in near real time so your data is always generating value, staying updated, and never becoming stale.
- **Reducing the risk of vendor lock-in:** The lakehouse approach uses open formats and open standards that allow your data to be stored independently of the tools you currently use to process it. This makes it easier to migrate your data to a different vendor or technology at any time.

IN THIS CHAPTER

- Recognizing data and AI benefits
- Addressing data reliability
- Getting value from a governed lakehouse
- Using lakehouses for BI and DW activities
- Supporting your ML and AI efforts

Chapter 3
Understanding the Underlying Technology

This chapter covers the technology foundations of a well-architected lakehouse on Databricks, focusing on Delta Lake for data management and Unity Catalog (UC) for governance. It also explores how lakehouses support data intelligence platforms for machine learning (ML) and artificial intelligence (AI).

Looking into the Data and AI Benefits

Without a proper data lakehouse strategy, data reliability is a big hindrance to extracting value from data across the enterprise — from raw data, batch, and real-time streaming all the way through ETL to be consumed downstream by BI, DW, ML, DS, and AI. Failed jobs can corrupt and duplicate data with partial writes. Multiple data pipelines reading and writing concurrently to your data lake can compromise data integrity. Many companies end up with their data pipeline efforts being too complex, coordinating among redundant systems with significant operational challenges to process both batch and streaming data jobs. This often results in unreliable data processing jobs that require manual cleanup and

reprocessing after failed jobs, which in turn causes a lot of lead time delay.

You need a well-architected lakehouse to solve these issues. Take a look at Figure 3-1. This open data lakehouse, based on Databricks, addresses these challenges.

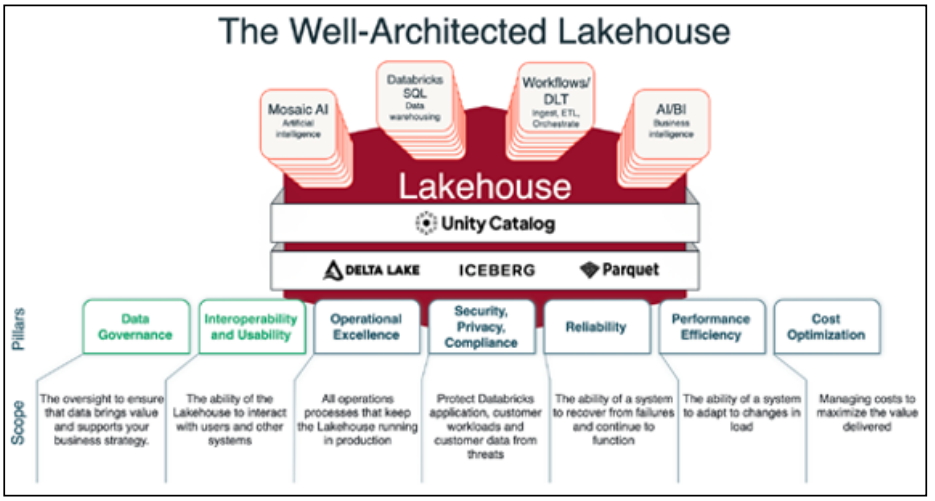

FIGURE 3-1: An example of a well-architected data lakehouse as the foundation for complete data intelligence.

The pillars in Figure 3-1 are selected to specifically improve building and operating a lakehouse architecture, with multi-cloud support.

Data Reliability with Delta Lake UniForm

Delta Lake UniForm is an open source storage layer that brings data reliability to your existing data lake by providing the ability to get the fast performance of data warehouses on the data lake while also guaranteeing data consistency through atomicity, consistency, isolation, durability (ACID) transactions. Delta Lake adds data reliability guarantees across both batch and streaming. This results in data continuously flowing through your data lake and providing end-users with the most complete, reliable, and up-to-date data available. Companies also like Delta Lake because it is extremely cost-effective, eliminating the need to manage and pay for multiple versions of the same data.

Seeing the Value of a Governed Lakehouse Architecture with UC OSS

The governance of a lakehouse architecture is paramount. Today, many organizations are dealing with the challenge of managing tons of data and AI assets across different systems — everything from structured and unstructured data to notebooks and AI models.

What if you could centralize all of this into one unified platform to streamline governance, improve visibility, and reduce compliance risks? UC open-source software (OSS) offers a unified governance layer for data and AI assets on the lakehouse architecture. With UC, organizations can seamlessly govern all of their structured and unstructured data, ML models, GenAI assets, tables, notebooks, dashboards, and files on any major cloud platform. Data teams can use UC to securely discover, tag, access, and collaborate on trusted data and AI assets to boost productivity and unlock the full potential of the lakehouse architecture. This unified approach to governance accelerates the data and AI initiatives while simplifying regulatory compliance. Plus, it helps teams collaborate better, cut costs, and innovate faster. In short, UC makes managing data and AI simpler and more efficient, so your organization can scale confidently.

UC helps enterprises address the key challenges by providing a unified visibility into data and AI with lineage and auditing, a single permission model for access control of data and AI, AI-powered monitoring for observability and discovery, and open data sharing.

Seeing Why Lakehouse Architectures Are Best for BI and DW Workloads

BI is pervasively used across business analytics, whether through SQL queries or through BI dashboards such as Power BI, Tableau, and Databricks AI/BI. BI on traditional, non-lakehouse warehouses has several inherent issues. They're slow and inefficient to process large amounts of data; they have complex infrastructure with disjointed architectures; they have fragmented and incomplete governance across multiple domains.

These inherent DW issues invariably lead to negative business outcomes:

- **High costs:** Budget overruns, which can be severe, from the cost of processing data as volumes continue to grow with no end in sight. There is budget pressure to just "keep the lights on" without room to adequately fund new development and new projects.
- **Lack of agility:** Fewer resources are available to address new business requirements because the engineering and IT staff spend so much time and effort navigating the complex interdependencies among systems.
- **Data breaks down at scale:** Silos, out-of-sync copies of data, and multiple development environments make operations become untenable to meet business needs and SLAs.
- **Operational risks:** Decision-making based on stale, incomplete, or low-quality data leads to operational risks.
- **Compliance and governance risks:** Limited governance architecture makes it impossible to securely democratize data access. Also non-unified audit data leads to avoid regulatory violations.

So, how does the lakehouse architecture solve these challenges with legacy architectures?

- **Unified platform:** Data warehousing built on one platform for all your data, analytics, and AI workloads, on top of your existing data lake and data warehouse with all of your data
- **Price/performance is AI-optimized:** Built-in intelligence that learns over time and improves performance for your workloads
- **Lower cost:** A best-in-class price/performance at scale, next-gen engine for the lakehouse, greatly accelerating computing speed
- **Unified governance:** Fine-grained governance, security, data lineage and monitoring for all your data assets, including tables, dashboards and models
- **Data sharing:** Allows companies to securely share their data and AI assets with external users and organizations, regardless of which data platform or cloud they're using through the open-sourced Delta Sharing and secured through UC

» **Query federation:** Enables users to run queries across multiple data sources and clouds, eliminating the costly need to migrate and ingest data into one platform, maintaining robust governance through UC

WARNING

Built-in complexities and costs are associated with transferring data from data lakes to data warehouses for ETL workloads, and proprietary data formats prevent direct data access with other tools and increase the risk of vendor lock-in. There are also increased cost and governance challenges associated with managing multiple copies of data and security models across your infrastructure.

Describing the Payoff for Exploratory GenAI, Data Science, and ML

Data scientists face numerous challenges along each step of data science workflows, hindering productivity. As organizations continue to become more AI-driven, a collaborative environment for easier access and visibility into all data, models trained on the data, reproducibility and transparency of the results, and insights uncovered within the data are critical. However, this collaborative environment hasn't always been easy to achieve.

Barriers to ML and AI productivity

Lakehouse architecture is essential for ML and AI use cases, like predictions, classifications, LLMs, and vector indexes. However, ML and AI efforts are often costly and complex, with companies spending excessive time on infrastructure and DevOps just to start analysis. Many data assets are untagged and unsearchable, leaving over 90 percent of unstructured data unused. Performing AI on structured and unstructured data is challenging without a unified lakehouse.

The lack of technical data scientists drives the need to democratize tools for non-technical users. Companies often piece together open-source libraries, leading to error-prone handoffs between data engineering and data science. These silos, disparate tools, and processes make ML life cycle management difficult, because tracking experiments, models, and dependencies is complex and limits reproducibility and efficiency.

Gains for developing ML and AI applications


TIP

With a data lakehouse architecture, a key benefit is that you gain quick access to clean and reliable data for downstream analytics and get immediate access to pre-configured serverless clusters to be used by ML and AI workspaces. Lakehouses also

- Provide the foundational architecture for data intelligence platforms, which embed intelligence in every aspect of your data needs. This includes developing GenAI applications, code assistance, intelligent searches of data assets, and democratization of asking questions through natural language with Databricks Genie spaces.
- Enable a unified approach to streamline the end-to-end ML workflow from data preparation to modeling and insights sharing.
- Facilitate tasks of preparing datasets, training models with extremely large datasets, and tracking data versions used to build and manage models through MLflow.
- Allow you to bring your own environment and multi-language support for maximum flexibility.
- Migrate or execute your code remotely on pre-configured and customizable clusters.
- Give you one-click access to ready-to-use, optimized, and scalable ML environments across the life cycle.
- Simplify handoffs among teams along each step in the ML/AI life cycle. Lakehouses offer one architecture for data ingest, feature development, model building, tuning, and deploying models in production, as well as monitoring models in production as data drifts.
- Track experiments, code, results, and artifacts, and manage models in one central hub.

IN THIS CHAPTER

- Understanding data intelligence
- Looking into the Databricks Data Intelligence Platform

Chapter 4
Bringing Data Intelligence to the Data Lakehouse

Data lakehouses unify and govern all data assets in open formats. This reduces data silos and supports workloads spanning from business intelligence (BI) to artificial intelligence (AI). Despite these advancements, companies still see significant challenges with data lakehouses. People are bottlenecked by needing to go through technical staff to build dashboards and reports. It's a struggle to find accurate data, necessitating extensive curation and planning. The rise of generative AI (GenAI) has amplified concerns around security and privacy of large language models (LLMs).

These challenges stem from data platforms' lack of fundamental understanding of organizational data and its usage. Fortunately, GenAI has emerged as a potent new tool to tackle these precise challenges.

Introducing Data Intelligence

GenAI has fundamentally driven companies to become data and AI-centric organizations at their core. To maximize their impact, companies seek to democratize their data and AI and integrate intelligence into all facets of their operations.

Data intelligence revolutionizes data management by employing AI models to deeply understand the semantics of your enterprise data, which is managed and governed by your data. It automatically analyzes the data, optimizing all workflows, and upgrades use cases with entirely new capabilities. Through this deep understanding of data, data intelligence enables the following:

- **The use of natural language:** By leveraging GenAI, data intelligence lets you just converse with your data.
- **Semantic cataloging and discovery:** GenAI understands each organization's data model, metrics, and key performance indicators (KPIs) to offer unparalleled discovery features or automatically identify discrepancies in data use.
- **Automated management and optimization:** AI models can optimize data layout, partitioning and indexing based on data usage, which reduces the need for manual tuning and knob configuration.
- **Enhanced governance and privacy:** DI can automatically detect, classify, and prevent misuse of sensitive data while simplifying management using natural language.
- **First-class support for AI workloads:** DI can enhance any enterprise AI application by allowing it to connect to the relevant business data and leverage the semantics learned by the data intelligence (metrics, KPIs, and so on) to deliver accurate results. AI application developers no longer have to hack intelligence together through brittle prompt engineering.

The Databricks Data Intelligence Platform

The Databricks Data Intelligence Platform is built on top of the data lakehouse and offers the possibilities of AI in data platforms as individual features are added. Databricks builds on the existing

unique capabilities of the Databricks Lakehouse, which offers a unified governance layer across data and AI, a unified, open, format-agnostic storage layer, and a single unified query engine that spans Extract, Transform, Load (ETL), SQL, machine learning (ML), AI, and BI.

In addition, Databricks leveraged Mosaic AI capabilities to generate and leverage AI models in a Data Intelligence Engine, which fuels all parts of the platform. The Data Intelligence Engine is deeply integrated into various layers of the Databricks Platform, including

- **AI/BI Genie:** Create spaces for technical and non-technical to converse with their own data, in their own jargon, and human-based guidance and reinforced learning for most accurate and informed insights.
- **Platform optimization:** Automatically adjusts settings like column indexing and partition layout, strengthening the lakehouse foundation for better performance and lower costs.
- **Enhanced governance:** Improves Unity Catalog (UC) by auto-generating descriptions and tags for all data assets such as tables and columns, enabling better semantic search, AI assistant quality, and governance across the platform.
- **AI assistant:** Enhances Python and SQL code generation and debugging for text-to-SQL and text-to-Python capabilities.
- **Query performance:** Boosts query speed by using data predictions for optimal query planning that provides extremely fast query performance at a low cost.
- **Efficient scaling:** Optimizes ETL and orchestration by predicting workload needs for optimal autoscaling and cost reduction.

Data intelligence platforms are key enablers in simplifying the development of enterprise AI applications, especially in helping to deploy agent systems. These systems combine the data lakehouse with AI agents so the AI understands your data and can solve customer and domain-specific use cases. Mosaic AI provides a unified platform to build agent systems and supports

- **Agents that reason over your data:** Databricks provides an e f cient and secure way to connect your enterprise data to AI agents. With the AI platform built on the lakehouse, there's no need to duplicate data. Instead, you can automatically generate

vector indexes and ML features from your production data. This makes it easy to customize AI models with your data, enabling you to build RAG apps, fine-tune open-source LLMs, and train both custom LLMs and classical ML models.

- **Custom evaluation for your use cases:** Mosaic AI offers built-in evaluation for agents. You can evaluate and use any combination of open source and commercial GenAI models, as well as ML models for your agent system. Mosaic AI helps you measure the output quality of the agents through AI-assisted judges that grade responses and allow human experts to give peer feedback. If quality issues are found, you can trace the root cause, evaluate fixes, and redeploy quickly.
- **Unified governance:** Only Mosaic AI provides end-to-end governance for agents. Customers can govern and apply guardrails across all AI models, including the ones hosted outside of Databricks. Through UC, Mosaic AI automatically enforces proper access controls, sets rate limits to manage costs, prevents harmful content, and tracks lineage throughout the entire AI workflow from data to models.

These AI agent systems can outperform traditional single LLM models by combining many different AI models together (LLMs, classical ML models, and tools), retrievers, vector databases, and tools for evaluation, monitoring, security, and governance. These multiple interacting components offer much higher quality outputs than a single model, allowing organizations to deliver more accurate, safe, and governed AI applications efficiently.

Figure 4-1 shows how Databricks leverages Mosaic AI in its unified platform that builds and manages compound AI systems.

FIGURE 4-1: Building AI agent systems on the Databricks Data Intelligence Platform.

Take a look at how to build AI systems using Mosaic AI:

- **Prepare data.** Use tools like LakeFlow to ingest and prepare your data for AI applications. With the AI platform built on the lakehouse, you don't need to duplicate data; instead, automatically generate vector indexes and ML features from your production data.
- **Build agents:** Use existing models, train new ones, or serve models through Mosaic AI's tools.
- **Deploy agents:** Deploy your models and AI apps at scale using MLflow and the Mosaic AI Agent Framework.
- **Evaluate agents:** Both human and machine-based evaluations ensure the quality and performance of your AI systems by using Mosaic AI Agent Evaluation and Lakehouse Monitoring.

REMEMBER

Unified governance keeps your data and AI assets secure and compliant with tools like Unity Catalog and the Mosaic AI Gateway, ensuring centralized control throughout the process.

IN THIS CHAPTER

- Eliminating data silos
- Allowing open governance
- Enabling BI and ML on all your data
- Reducing costs by consolidating systems

Chapter 5
Ten Reasons Why You Need a Data Lakehouse

Data lakehouses have been implemented by almost every enterprise because of their many benefits. We discuss these benefits throughout this book, but here is an overview:

- **Eliminates data silos:** All your data estate is centralized and unified across the architecture including structured, semi-structured, streaming, and unstructured data.
- **Unifies the best of data warehousing (DW), business intelligence (BI), machine learning (ML), and artificial intelligence (AI):** Data lakehouses are the foundation for all types of workloads, combining the best of two worlds: the flexibility and cost-effectiveness of data lakes and the analytic abilities of data warehouses.
- **Unified and open governance:** This is essential for securing and managing all data and AI assets across various formats and data sources. Governance unifies access management, auditing monitoring, and lineage, allowing for easy discovery, access, and sharing of trusted data across any tool, engine, or cloud platform.
- **Increases data and AI team efficiency and collaboration:** Lakehouses combine the best features of a data warehouse

with the low cost of data lakes. More personas across the business can work together to move faster and simpler, have more scalability, and be more cost-effective.

- **Reduces costs:** Data lakehouses eliminate the costly need to create and move redundant copies of data and reduce costly license fees from multiple vendors.
- **Simplifies data engineering:** You can easily ingest and transform both batch and streaming data with a data lakehouse without worrying about managing the underlying infrastructure. You can make your team's job easier with an AI-powered data intelligence engine that helps you understand your data and pipelines.
- **Scales:** A data lakehouse scales to sizes far beyond legacy solutions — trillions of records — because it decouples the storage and compute while being highly performative and with lower latency.
- **Removes data redundancy:** Lakehouses use a single architecture to remove data redundancy, which happens when you have data on multiple tools for processing, cleaning, and Extract, Transform, Load (ETL).
- **Open source:** Lakehouses leverage open source at every layer to prevent vendor lock-in and provide transparency. Get unified multi-cloud storage for data reliability (Delta Lake Uniform), processing (Apache Spark), managing the ML life cycle (MLflow), and securely sharing live data (Delta Sharing).
- **Serves as the foundation for data intelligence:** Data intelligence better understands the semantics of your data lakehouse and uses for more intelligent searches, code assistance, automation of scaling up and down compute as needed, visualizations with Databricks AI/BI, and democratization by non-technical people to query their own data by using natural language.

databricks

The Databricks Data Intelligence Platform

Unify and govern all your data

While many companies are adopting data lakehouses, understanding and

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.