Brought to you by:

databricks

The Data Lakehouse

for dummies®

A Wiley Brand

Unify and govern all your data and AI

Use the lakehouse for analytics and AI

Support data, AI, and BI workloads

2nd Databricks Special Edition

Ari Kaplan
Amit Kara

About Databricks

Databricks is the data and AI company. More than 20,000 organizations worldwide — including Block, Comcast, Condé Nast, Rivian, Shell and over 60 percent of the Fortune 500 — rely on the Databricks Data Intelligence Platform to take control of their data and put it to work with AI. Databricks is headquartered in San Francisco, with offices around the globe, and was founded by the original creators of Lakehouse, Apache Spark™, Delta Lake, and MLflow. To learn more, follow Databricks on social media:

 www.linkedin.com/company/databricks

 https://x.com/databricks

 www.instagram.com/databricksinc

 www.facebook.com/databricksinc

The Data Lakehouse

2nd Databricks Special Edition

by Ari Kaplan and Amit Kara

for dummies®
A Wiley Brand

The Data Lakehouse For Dummies®, 2nd Databricks Special Edition

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

ISBN 978-1-394-39663-4 (pbk); ISBN 978-1-394-39664-1 (ebk); ISBN 978-1-394-39665-8 (ebk)

Publisher's Acknowledgments

Acquisitions Editor: Traci Martin
Senior Managing Editor: Rev Mengle
Development Editor: Jen Bingham
Sales Manager: Molly Daugherty
Content Refinement Specialist: Umeshkumar Rajasekhar

Table of Contents

Introduction

The data lakehouse 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 capabilities of data warehouses. Lakehouses are built on open source and open standards, which unify and simplify your data management. They support all your data needs — including business intelligence (BI), data warehousing (DW), online transaction processing (OLTP), machine learning (ML), and AI — enabling you to quickly build secure data and AI apps. This enables businesses to collaboratively build more intelligent applications on all of their data and workloads. And throughout it all, governance is central for providing end-to-end visibility and control of all your data estate.

The data lakehouse reduces costs, 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, semistructured, 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

This book introduces the data lakehouse to manage and govern all of your organization's data assets, and explains the limitations of legacy solutions. It explains why a lakehouse is a foundation for solving data challenges and forms the basis for data intelligence platforms. 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

This book occasionally uses 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 and 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 data management efforts more effective and efficient in your company. However, because this is a relatively short book on data lakehouses, we also recommend checking out the following:

- Lakehouse overview: `databricks.com/product/data-lakehouse`
- Video of the lakehouse architecture: `youtube.com/watch?v=13TownvHT7w`
- *The Data Intelligence Platform For Dummies,* Databricks Special Edition: `databricks.com/resources/ebook/maximize-your-organizations-potential-data-and-ai`

IN THIS CHAPTER

- Explaining data warehouses
- Positioning data warehouse and data lake limitations
- Describing the concept of data lakehouses

Chapter 1
Making the Case for Data Lakehouses

Every company today aims to be a data and artificial intelligence (AI)-driven organization. This was once a contentious idea, but now it's widely accepted. This approach is no longer just about data — AI is essential to success. But finding this type of success at scale is difficult for most organizations that need high-quality data that is both secure and consumable across the organization.

The current state of data and AI systems is highly fragmented and extremely complex — a nightmare of high costs and proprietary formats. Organizations often have multiple data warehouses due to acquisitions, independent business units, and legacy systems. Many companies have accumulated a patchwork of data environments over time. These silos create inefficiencies and prevent organizations from leveraging data effectively. Consolidation is critical for cost reduction and agility. That's where data lakehouses come in.

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

Exploring Traditional Data Warehouses

DM enables companies to corral their data across the whole company by using consistent methods, techniques, and tools. The purpose of DM on an enterprise-wide scale is to fulfill all types of requirements for use cases, applications, and business processes. Simply put, DM supports the effective use of data, encompassing governance, quality, integration, and security.

Data warehousing is one component of DM that focuses specifically on storing and analyzing structured data in rows and columns. Data lakes are another type of DM that supports 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 AI and business intelligence (BI), puts high demands on DM.


TECHNICAL STUFF

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, 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.

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]), set up and deployed faster, scaled larger, and improved access across the globe.

Traditional data warehouses have inherent limitations that became more prohibitive as data volumes grew significantly larger, and a need arose 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

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

- They don't support atomic, consistent, isolated, and durable (ACID) transactions, which risk corrupting files and causing 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 makes it almost impossible to simultaneously write and append new data.
- Jobs that fail midway 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 generative (Gen)AI 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 a 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 technology because they enabled companies to analyze massive amounts of both structured and unstructured data together for the first time, which before was simply 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, MLflow, Delta Lake, Apache Iceberg, and Unity Catalog) is the lakehouse's underlying technology — it offers many advantages over traditional data lakes and data warehouses:

- Speed through in-memory processing, often 100 times faster
- Ease of use through the support of Python, R, SQL, and Scala
- Versatility for handling a variety of data processing methods such as batch and real-time streaming
- GenAI and advanced analytics
- 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 for creating AI-driven applications and productivity agents

WARNING

Having all these divergent solutions without a unified data lakehouse leads to many 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 track activity, monitoring dashboards, and governance frameworks. This sprawl adds risks, costs, and operational inefficiencies.
- Data privacy and control are massive issues when attempting to apply them across data silos. With GenAI, a bright light is being trained on the ability 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 this lack can 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.

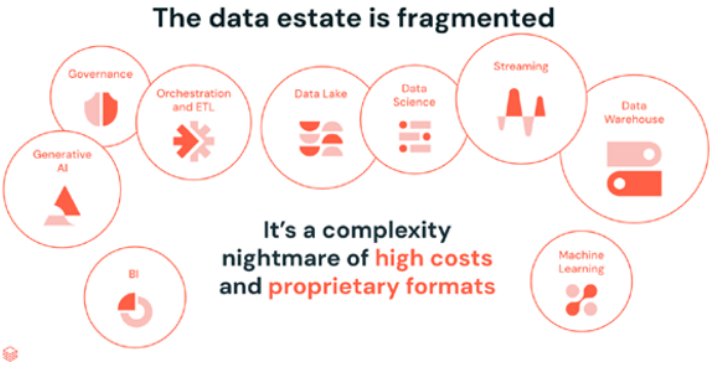


FIGURE 1-1: Age-old challenges that data lakehouses solve.

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

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. 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; lakehouse storage such as Delta Lake and Iceberg; MLflow to manage the machine learning (ML) lifecycle; Delta Sharing to securely share live data from your lakehouse to any computing platform without replication; and Spark Declarative Pipelines to simplify complicated Extract, Transform, Load (ETL) processes using a declarative approach.

The best lakehouses are flexible to run on all major cloud providers. They allow 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 for democratizing data and AI across an organization. Data intelligence platforms use generative (Gen)AI to better understand the semantics of your data and use that across the platform (see Chapter 4).

In this chapter, you will 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.

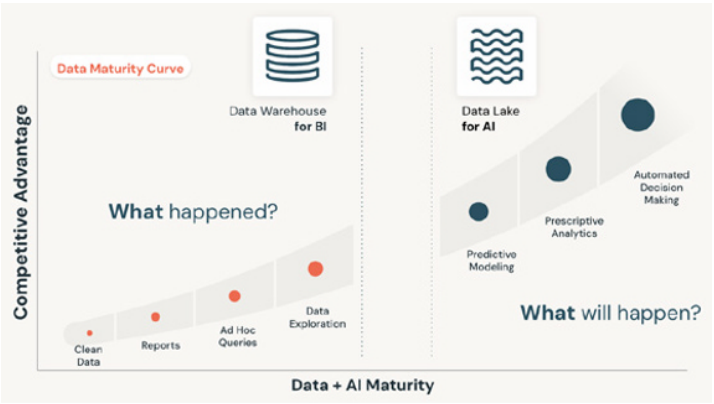


FIGURE 2-1: The data and AI maturity curve.

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 precanned 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 offer 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; Spark Declarative Pipelines for building robust and reliable data pipelines that simplify the development of both batch and streaming ETL; and MLflow for streamlining the ML lifecycle 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.

- **Data warehousing (DW)/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 in DW and beyond, 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.
- **Applications:** Build, deploy, and govern secure applications on your company's data estate, empowering teams to quickly build with the assistance of AI. With a unified lakehouse for data and AI, you can build secure apps directly on that foundation, running alongside your data and models.
- **Diverse data types:** The best business insights often come from a variety of data types such as structured and unstructured: images, video, audio, semistructured data, and text. Lakehouses support this multimodal approach.
- **Diverse workloads:** Support includes AI, DS, ML, BI, online transaction processing (OLTP), and SQL. 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 efficient 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's 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 not cost-efficient. The lakehouse approach 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 can grow the amount of data and increase workloads, contributing 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 offer serverless compute. This feature allows workloads to run automatically without the need for humans to preprovision 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.

REMEMBER

The key benefit of the lakehouse is that it allows you to unify and govern 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. It unifies your data teams on one architecture. It reduces data silos so everyone in your organization can access and process all data types: batch and streaming, structured and unstructured. And it reduces the risk of vendor lock-in by using open formats.

IN THIS CHAPTER

- Recognizing data and AI benefits
- Addressing data reliability
- 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 Extract, Transform, Load (ETL) to be consumed downstream by business intelligence (BI), data warehousing (DW), online transaction processing (OLTP), 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.

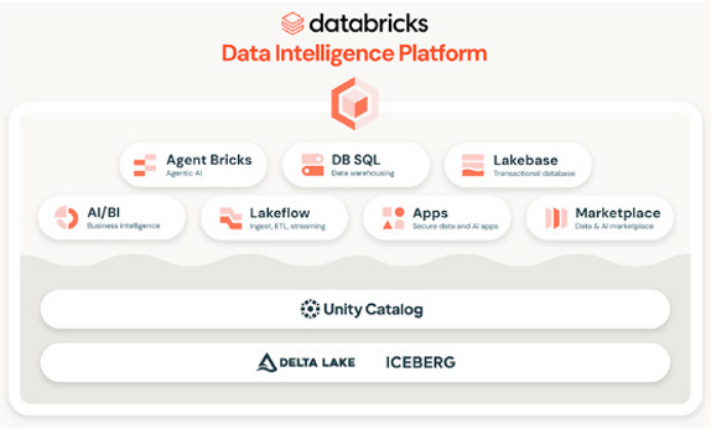


FIGURE 3-1: An example of a well-architected data lakehouse as the foundation for complete data intelligence.

The pillars in Figure 3-1 specifically improve building and operating a lakehouse architecture with multicloud support.

Data Reliability and Governance with Lakehouses

Lakehouse storage is an extremely reliable, cost-effective, scalable, and open format for all lakehouse data needs. It brings data-reliability guarantees to your existing data lake and provides the fast performance of data warehouses. Lakehouse storage also guarantees data consistency through atomicity, consistency, isolation, durability (ACID) transactions. It adds data reliability guarantees across both batch and streaming. This results in data continuously flowing through your lakehouse data lake and providing end-users with the most complete, reliable, and up-to-date data.

TECHNICAL STUFF

There are several leading open standard data storage formats including Delta Lake and Iceberg. Their features are fairly similar and companies can choose which format they prefer. Both have their own reliable and cost-effective ways to manage data, and sit on top of popular data files such as Parquet. Each has a different metadata format, and Delta Lake has schema enforcement and lineage tracking.

The governance of a lakehouse is paramount. Today, many organizations face challenges of managing large amounts of data and AI assets across different systems — everything from structured and unstructured data to notebooks and AI models.

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, AI models, generative (Gen)AI 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.

REMEMBER

UC is a unified governance that accelerates data and AI initiatives, simplifies regulatory compliance, and helps teams collaborate better, cut costs, and innovate faster.

Seeing Why Lakehouses 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 at processing large amounts of data, have complex infrastructure with disjointed architectures, and 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 service level agreements (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, nonunified audit data leads to avoiding regulatory violations.

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

- **Unified platform:** Built on one platform on top of your existing data lake and data warehouse with all of your data analytics, BI, DW, and AI workloads
- **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

Building Transactional Applications on the Lakehouse

Building apps on the lakehouse means you no longer need one system for transactions and another for analytics. Databricks Lakebase gives you a PostgreSQL-compatible database that runs side by side with your analytics and AI. Data stays in open formats with full reliability, so it connects straight into dashboards, AI tools, and models without messy ETL. This makes apps faster, simpler, and more dependable.

Lakebase comes with features that make life easier. You can branch databases like code, roll back with time travel, and scale up or down automatically. Governance from Unity Catalog keeps security consistent across all your data.

Databricks Apps adds the front end. It lets you build and run apps directly on the lakehouse using tools like Streamlit or Gradio. You don't need a separate stack or to move data around. Together, Lakebase and Apps give you one platform to build secure, real-time apps that work with analytics and AI out of the box.

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 costs and governance challenges associated with managing multiple copies of data and security models across your infrastructure.

Describing the Payoff for AI

AI practitioners — whether focusing on data science, ML, classical AI, or generative (Gen)AI — face numerous challenges along each step of their workflow, hindering productivity. As organizations continue to become more AI-driven, a collaborative environment is critical 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. However, this collaborative environment hasn't always been easy to achieve.

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 preconfigured 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 the democratization of asking questions through natural language.
- Enable a unified approach to streamline end-to-end ML/AI workflows from data preparation to modeling and insights sharing.
- Build AI agents that can reason on data from across your data estate. Build, evaluate, deploy, and govern agents that reason across every enterprise system, with support for current and future foundational models.
- Facilitate tasks of preparing datasets, training models with extremely large datasets, and tracking data versions used to build and manage models through MLflow.
- Manage the entire ML lifecycle — developing, experiment tracking, testing, deploying, and monitoring — with Machine Learning Operations (MLOps).
- Fully track your development, training, judging, and operationalizing of your LLMs with humans in the loop — through LLM Operations (LLMOps).
- Give you one-click access to ready-to-use, optimized, and scalable AI environments across the lifecycle.
- Simplify handoffs among teams along each step in the ML/AI lifecycle. Lakehouses offer a unified architecture for data ingestion, 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 solve organizational challenges by unifying and governing all data assets in open formats, reducing data silos and supporting all types of workloads from business intelligence (BI) to artificial intelligence (AI). Now, almost every company has a lakehouse. Despite these advancements, companies still see significant challenges with data lakehouses. People are bottlenecked by needing to go through technical staff to build dashboards when they simply want to talk with their data. It's a struggle to search and discover which of tens of thousands of tables and columns contain the data you want, necessitating extensive curation and planning. The rise of generative (Gen)AI has amplified concerns around the security, privacy, and accuracy of large language models (LLMs). Companies want intelligence on their own company's data estate, with their own business terms and definitions that aren't diluted by general-purpose 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

The Data Intelligence Platform was created to provide AI tuned to your business for more intelligent insights, all with a lakehouse foundation for unified data and governance. 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 all aspects of data management by employing AI 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 use of natural language:** By leveraging GenAI, data intelligence lets you simply converse with your own data — making it easy for anyone to use data for their work.
- **Semantic cataloging and discovery:** Data intelligence understands each organization's data model, metrics, and key performance indicators (KPIs) to offer unparalleled discovery features and automatically identify discrepancies in data use.
- **Automated management and optimization:** Data intelligence learns usage patterns and improves data layout, partitioning, and indexing without manual tuning.
- **Enhanced governance and privacy:** Data intelligence can automatically detect, classify, and prevent misuse of sensitive data while simplifying management using natural language.
- **First-class support for AI applications:** Data intelligence enhances enterprise AI applications by creating and leveraging the business's own terminology and semantics (metrics, KPIs, and so on) to deliver more 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 a lakehouse, which offers a unified governance layer across data and AI, a unified, open, format-agnostic storage layer, and a single unified platform that spans Extract, Transform, Load (ETL), SQL, machine learning (ML), AI, and BI.

In addition, Databricks provides capabilities to build and manage AI agents and models, which fuel all parts of the Data Intelligence Platform, including

- **AI/BI Genie:** Lets you simply talk with your own data in the context of your business using AI/BI dashboards that go beyond traditional BI tools — letting business users self-serve their own AI-driven insights without waiting on the BI teams. All governed with human-based guidance for the most accurate and informed insights.
- **Platform optimization:** Automatically adjusts settings like column indexing and partition layout, strengthening a 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 the coding and debugging experience with languages such as Python and SQL. Either simply chat with the AI Assistant or leverage Agent mode for even more of an intelligent companion to plan and automate entire multistep solutions, and more advanced analytics through the Data Science Agent.
- **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. Databricks AI makes it possible to build, deploy, and manage AI applications and agent systems — such as Retrieval Augmented Generation (RAG) pipelines and vector indexes — without duplicating data. Agent Bricks extends this by providing prebuilt, production-ready AI agents optimized on your data.

Databricks provides a unified platform to build agent systems and supports:

- **Agents that reason over your data:** Databricks provides an efficient and secure way to connect your enterprise data to AI agents, including Databricks' own collection of Agent Bricks. With the AI platform built on a 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:** Databricks 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. Databricks 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:** Databricks 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, Databricks 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, and vector databases. These multiple interacting components offer much higher quality outputs than a single model, allowing organizations to evaluate, deliver more accurate, safe, and governed AI applications efficiently.

Figure 4-1 shows how Databricks leverages AI throughout its unified platform for building and managing compound AI systems.

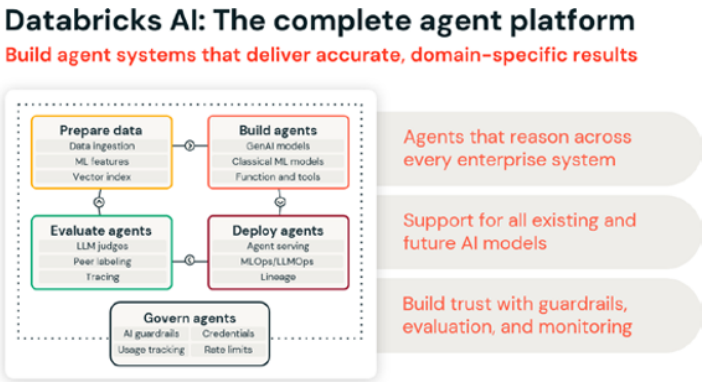


FIGURE 4-1: Building AI agent systems on the Databricks Data Intelligence Platform.

The development of intelligent AI applications is accelerated through Databricks Apps and Agent Bricks.

Databricks Apps enables you to quickly and easily build secure data and AI applications on your own data estate. There is tremendous value in bringing new applications to market more quickly. Databricks Apps is run on serverless compute, powered by data intelligence, secured and governed out of the box, and works with a robust open ecosystem. This includes all popular Python frameworks such as Dash, Streamlit, and Gradio. Databricks Apps represents the architectural shift of moving the app to where the data and AI reside, instead of the old way of moving the data and AI to where the app resides.

Agent Bricks uses research-backed innovations to streamline your process of building, evaluating, and optimizing AI agent systems — grounded on your data. You can create custom evaluations of how models are performing, swap in new models as the market evolves, balance cost versus quality

and more. For example, extract insights from documents, build high-quality Q&A agents, or create custom LLMs for more complex and low-level tasks, and supervise stitching multiple agents together. Success isn't about just selecting the best model, and Agent Bricks lets your team focus on value, not infrastructure.

REMEMBER

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

IN THIS CHAPTER

- Eliminating data silos
- Allowing open governance
- Enabling AI and BI 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. These benefits are discussed throughout this book, but here is an overview:

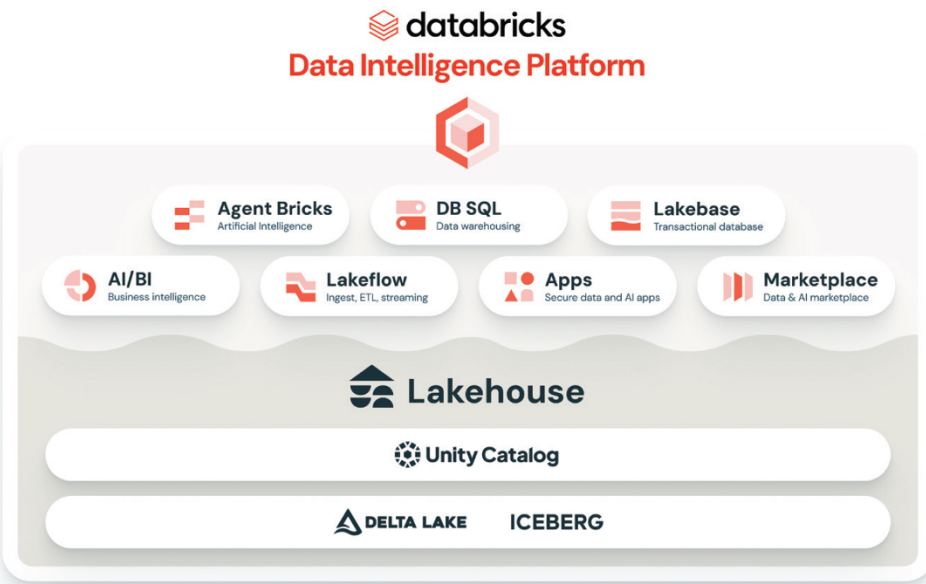
- **Eliminates data silos:** All your data estate is centralized and unified across the architecture, including structured, semistructured, streaming, and unstructured data.
- **Unifies the best of data warehousing (DW), business intelligence (BI), online transaction processing (OLTP), machine learning (ML), and artificial intelligence (AI):** Data lakehouses are the foundation for all types of workloads, combining the best of all 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 enable more personas across the business that can work together to move faster and simpler, have more scalability, and be more cost-effective.
- **Reduces costs and data redundancy:** Data lakehouses eliminate the costly need to create and move redundant copies of data and reduce costly license fees incurred 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 Platform that helps you more intelligently 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.
- **Open source:** Lakehouses leverage open source at every layer to prevent vendor lock-in and provide transparency. Get unified multicloud storage for data reliability (Delta Lake and Iceberg), processing (Apache Spark), managing the ML lifecycle (MLflow), and securely sharing live data (Delta Sharing).
- **AI applications:** Lakehouses connect applications to relevant business data, creating and leveraging the business's own terminology and semantics learned by the data intelligence. They're enhanced with AI agents that can reason on data from across your data estate. Build, evaluate, deploy, and govern agents that reason across every enterprise system, with support for current and future foundational models.
- **Serves as the foundation for data intelligence:** Data intelligence better understands the semantics of your data lakehouse and uses it for more intelligent searches, code assistance, automation of scaling compute up and down as needed, visualizations with Databricks AI/BI, and democratization to allow nontechnical people to query their own data by using natural language.

databricks

The Databricks Data Intelligence Platform

The Databricks Data Intelligence Platform allows your entire organization to use data and AI. It's built on a lakehouse to provide an open, unified foundation for all data and governance, and is powered by data intelligence that understands the uniqueness of your data.



Unify and govern all your data

While many companies are adopting data lakehouses, understanding and using them effectively remains a challenge. Your organization's success with data and artificial intelligence (AI) depends on transforming raw data into actionable insights. This book simplifies the complexities of the data lakehouse, guiding you to harness data intelligence, unlock its potential for data-driven decisions, and drive sustainable business growth.

Inside...

- Simplify with a lakehouse architecture
- Break down silos for unified data
- Unified governance for all data and AI
- Optimize performance and scalability
- Realize actionable insights with AI Apps
- Leverage open source and standards
- Support AI and BI workloads

databricks

Ari Kaplan is Databricks' Head of Technical Evangelism. He created the Chicago Cubs' & Baltimore Orioles' analytics departments. **Amit Kara,** Director of Technical Marketing at Databricks, is a seasoned expert in data management, who helps organizations unlock the full potential of their data.

Go to Dummies.com™
for videos, step-by-step photos, how-to articles, or to shop!

ISBN: 978-1-394-39663-4
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

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