eBook

The Data Team's Guide to the Databricks Data Intelligence Platform

databricks

Contents

CHAPTER 1 The data lakehouse 4

CHAPTER 2 The Databricks Data Intelligence Platform 11

CHAPTER 3 Data reliability and performance 17

CHAPTER 4 Unified governance and sharing for data, analytics and AI 27

CHAPTER 5 Security 40

CHAPTER 6 Instant compute and serverless 47

CHAPTER 7 Data warehousing 50

CHAPTER 8 Data engineering 54

CHAPTER 9 Data streaming 66

CHAPTER 10 Data science, AI and machine learning 71

CHAPTER 11 Databricks Technology Partners and the modern data stack 77

CHAPTER 12 Get started with the Databricks Data Intelligence Platform 79

INTRODUCTION

The Data Team's Guide to the Databricks Data Intelligence Platform

The Data Team's Guide to the Databricks Data Intelligence Platform is designed for data practitioners and leaders who are embarking on their journey into the data lakehouse architecture.

In this eBook, you will learn the full capabilities of the data lakehouse architecture and how the Databricks Data Intelligence Platform allows your entire organization to use data and AI.

You will see how it's built on a lakehouse to provide an open, unified foundation for all data and governance, and is powered by a Data Intelligence Engine that understands the uniqueness of your data. From ETL to data warehousing to generative AI, Databricks helps you simplify and accelerate your data and AI goals.

databricks

CHAPTER

01

The data lakehouse

The evolution of data architectures

Data has moved front and center within every organization as data-driven insights have fueled innovation, competitive advantage and better customer experiences.

However, as companies place mandates on becoming more data-driven, their data teams are left in a sprint to deliver the right data for business insights and innovation. With the widespread adoption of cloud, data teams often invest in large-scale complex data systems that have capabilities for streaming, business intelligence, analytics and machine learning to support the overall business objectives.

To support these objectives, data teams have deployed cloud data warehouses and data lakes.

Traditional data systems: The data warehouse and data lake

With the advent of big data, companies began collecting large amounts of data from many different sources, such as weblogs, sensor data and images. Data warehouses — which have a long history as the foundation for decision support and business intelligence applications — cannot handle large volumes of data.

While data warehouses are great for structured data and historical analysis, they weren't designed for unstructured data, semi-structured data, and data with high variety, velocity and volume, making them unsuitable for many types of data.

This led to the introduction of data lakes, providing a single repository of raw data in a variety of formats. While suitable for storing big data, data lakes do not support transactions, nor do they enforce data quality, and their lack of consistency/isolation makes it almost impossible to read, write or process data.

For these reasons, many of the promises of data lakes never materialized and, in many cases, reduced the benefits of data warehouses.

As companies discovered new use cases for data exploration, predictive modeling and prescriptive analytics, the need for a single, flexible, high-performance system only grew. Data teams require systems for diverse data applications including SQL analytics, real-time analytics, data science and machine learning.

To solve for new use cases and new users, a common approach is to use multiple systems — a data lake, several data warehouses and other specialized systems such as streaming, time-series, graph and image databases. But having multiple systems introduces complexity and delay, as data teams invariably need to move or copy data between different systems, effectively losing oversight and governance over data usage.

Challenges with data, analytics and AI

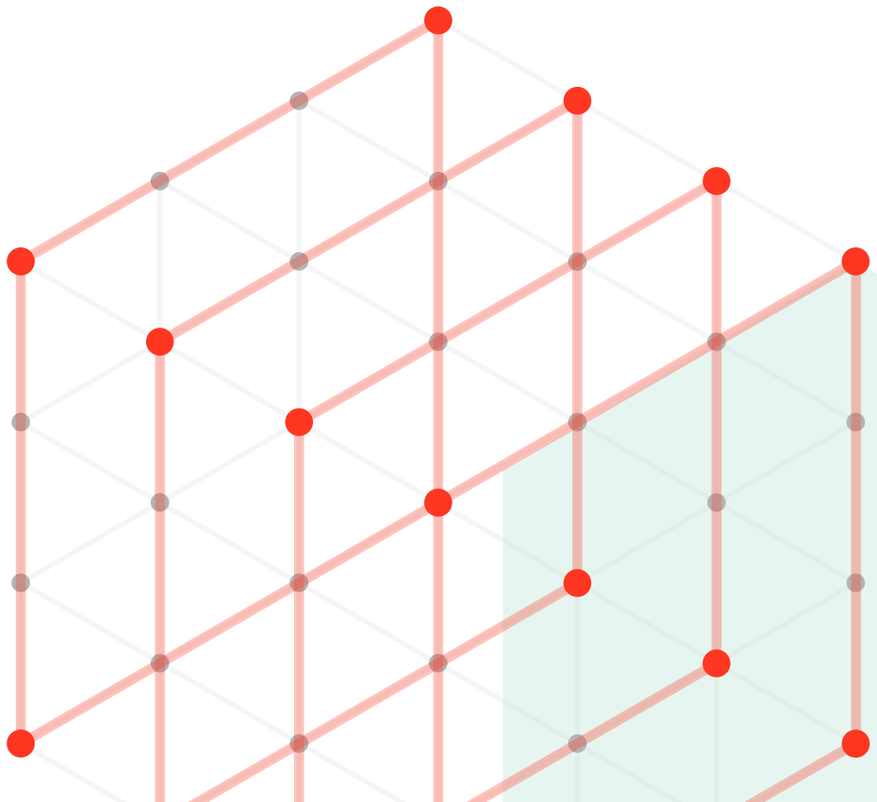
In a recent Accenture study, only 32% of companies reported tangible and measurable value from data. The challenge is that most companies continue to implement two different platforms: data warehouses for BI and data lakes for AI. These platforms are incompatible with each other, but data from both systems is generally needed to deliver game-changing outcomes, which makes success with AI extremely difficult.

Today, most of the data is landing in the data lake, and a lot of it is unstructured. In fact, according to IDC, about 80% of the data in any organization will be unstructured by 2025. But, this data is where much of the value from AI resides. Subsets of the data are then copied to the data warehouse into structured tables, and back again in some cases.

You also must secure and govern the data in both warehouses and offer fine-grained governance, while lakes tend to be coarser grained at the file level. Then, you stand up different stacks of tools on these platforms to do either BI or AI.

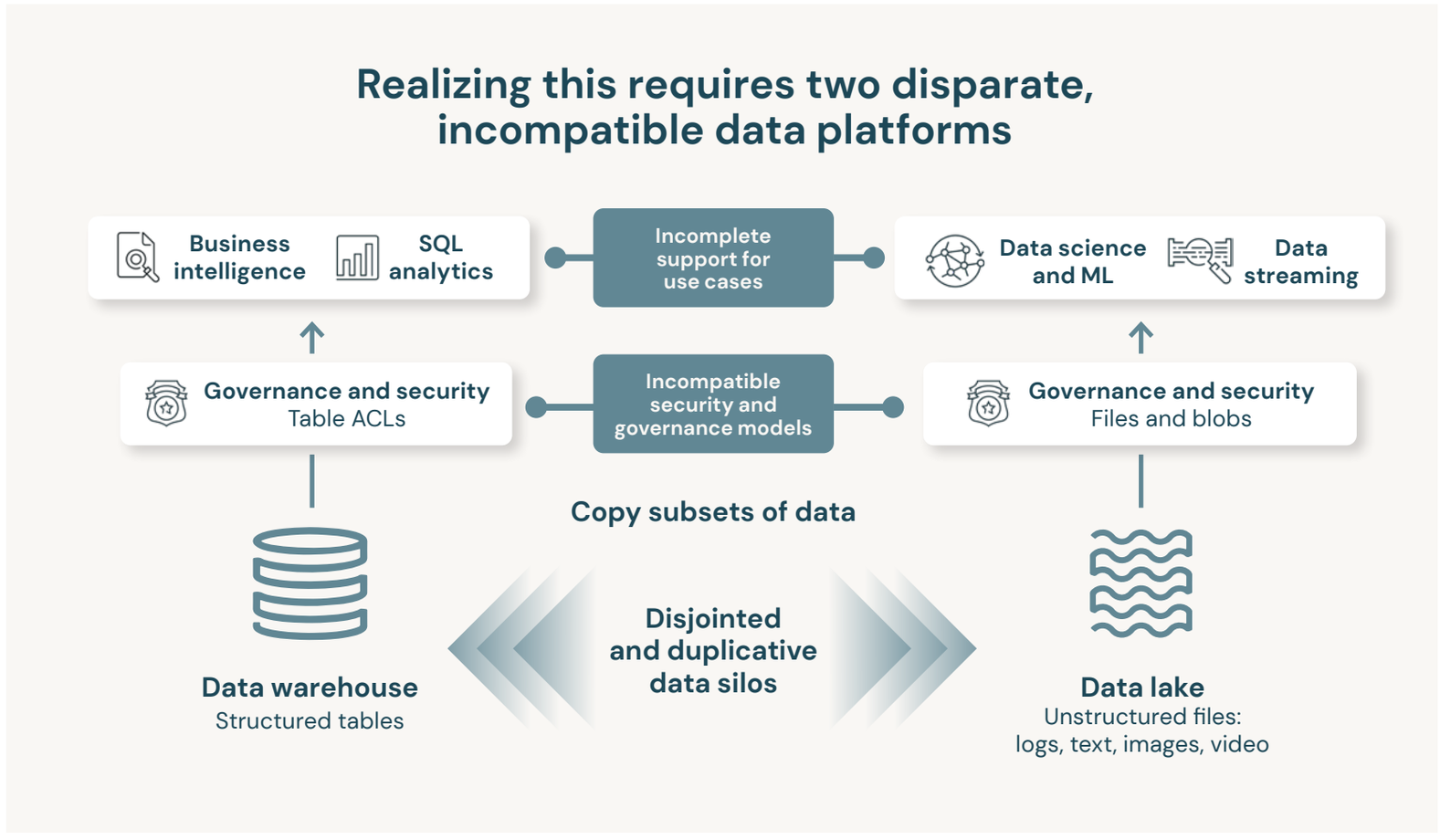
You have now duplicated data in two different systems and the changes you make in one system are unlikely to find their way to the other. So you are going to have data drift almost immediately, not to mention paying to store the same data multiple times.

Then, because governance is happening at two distinct levels across these platforms, you are not able to control things consistently.



Finally, the tool stacks on top of these platforms are fundamentally different, which makes it difficult to get any kind of collaboration going between the teams that support them.

This is why AI efforts fail. There is a tremendous amount of complexity and rework being introduced into the system. Time and resources are being wasted trying to get the right data to the right people, and everything is happening too slowly to get in front of the competition.

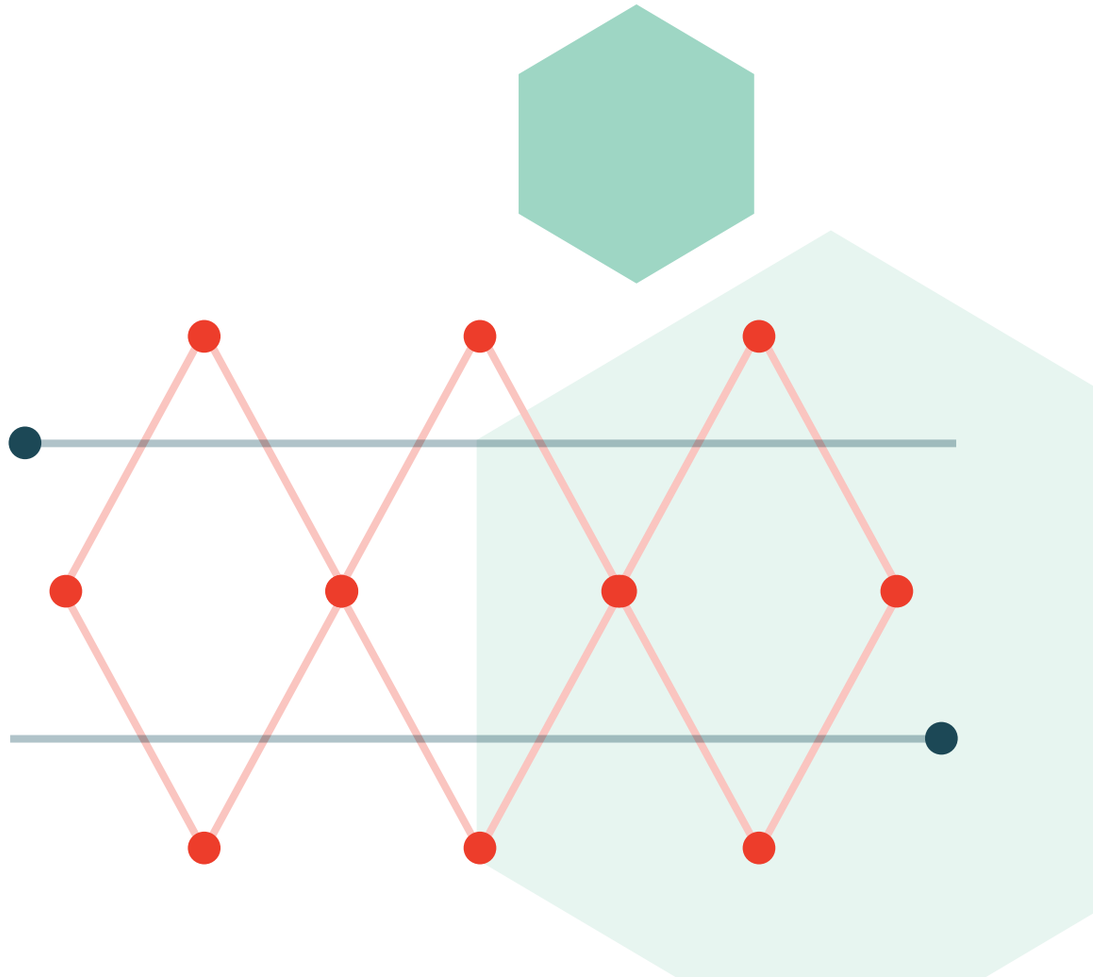


Lakehouse: A new generation of open platforms

Databricks is the inventor and pioneer of the data lakehouse architecture. We are continuously innovating on the lakehouse architecture to help customers deliver on their data, analytics and AI aspirations. The ideal data, analytics and AI platform needs to operate differently. Rather than copying and transforming data in multiple systems, you need one platform that accommodates all data types.

Ideally, the platform must be open, so that you are not locked into any walled gardens. You would also have one security and governance model. It would not only manage all data types, but it would also be cloud-agnostic to govern data wherever it is stored.

Last, it would support all major data, analytics and AI workloads, so that your teams can easily collaborate and get access to all the data they need to innovate. The lakehouse architecture is a prerequisite for data intelligence, because it unifies data and AI. Adding an integrated data intelligence engine that understands the semantics of your data and all the metadata to the lakehouse architecture is what creates a data intelligence platform.

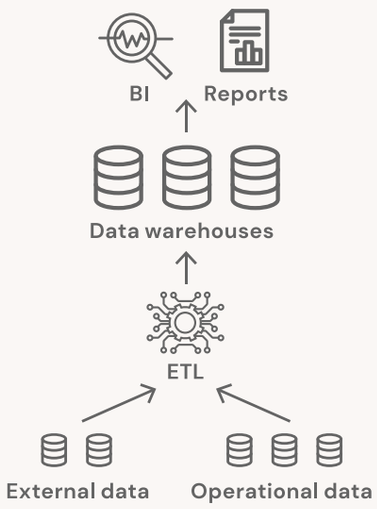


databricks

Data lakehouse

One architecture to unify all your data, analytics and AI workloads

Data warehouse



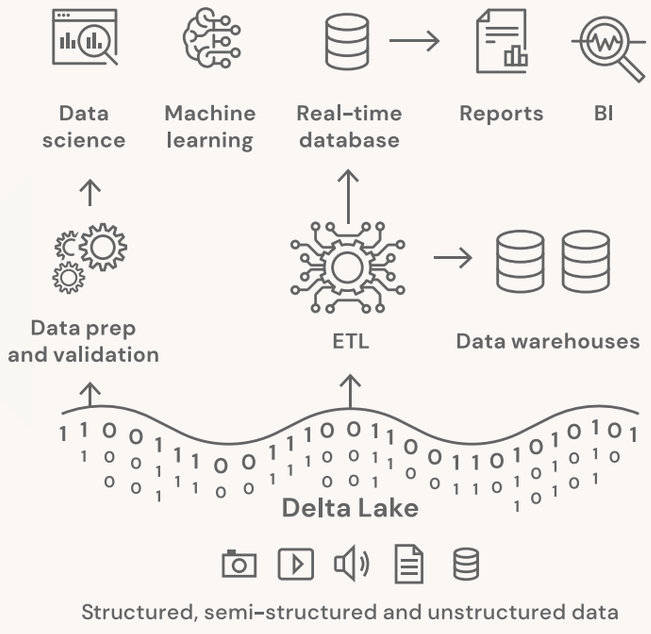
Lakehouse Architecture

All machine learning, SQL, BI, and streaming use cases

One security and governance approach for all data assets on all clouds

An open and reliable data platform to efficiently handle all data types

Delta Lake



databricks

Key features for a lakehouse

Recent innovations with the data lakehouse architecture can help simplify your data and AI workloads, ease collaboration for data teams, and maintain the kind of flexibility and openness that allows your organization to stay agile as you scale. Here are key features to consider when evaluating data lakehouse architectures:

Transaction support: In an enterprise lakehouse, many data pipelines will often be reading and writing data concurrently. Support for ACID (Atomicity, Consistency, Isolation and Durability) transactions ensures consistency as multiple parties concurrently read or write data.

Schema enforcement and governance: The lakehouse should have a way to support schema enforcement and evolution, supporting data warehouse schema paradigms such as star/snowflake. The system should be able to reason about data integrity, and it should have robust governance and auditing mechanisms.

Data governance: Capabilities including auditing, retention and lineage have become essential, particularly considering recent privacy regulations. Tools that allow data discovery have become popular, such as data catalogs and data usage metrics.

BI support: Lakehouses allow the use of BI tools directly on the source data. This reduces staleness and latency, improves recency and lowers cost by not having to operationalize two copies of the data in both a data lake and a warehouse.

Storage decoupled from compute: In practice, this means storage and compute use separate clusters, thus these systems can scale to many more concurrent users and larger data sizes. Some modern data warehouses also have this property.

Openness: The storage formats, such as Apache Parquet, are open and standardized, so a variety of tools and engines, including machine learning and Python/R libraries, can efficiently access the data directly.

Support for diverse data types (unstructured and structured): The lakehouse can be used to store, refine, analyze and access data types needed for many new data applications, including images, video, audio, semi-structured data and text.

Support for diverse workloads: Use the same data repository for a range of workloads including data science, machine learning and SQL analytics. Multiple tools might be needed to support all these workloads.

End-to-end streaming: Real-time reports are the norm in many enterprises. Support for streaming eliminates the need for separate systems dedicated to serving real-time data applications.

Learn more

- Lakehouse: A New Generation of Open Platforms That Unify Data Warehousing and Advanced Analytics
- Building the Data Lakehouse by Bill Inmon, Father of the Data Warehouse
- What Is a Data Lakehouse?

CHAPTER

02

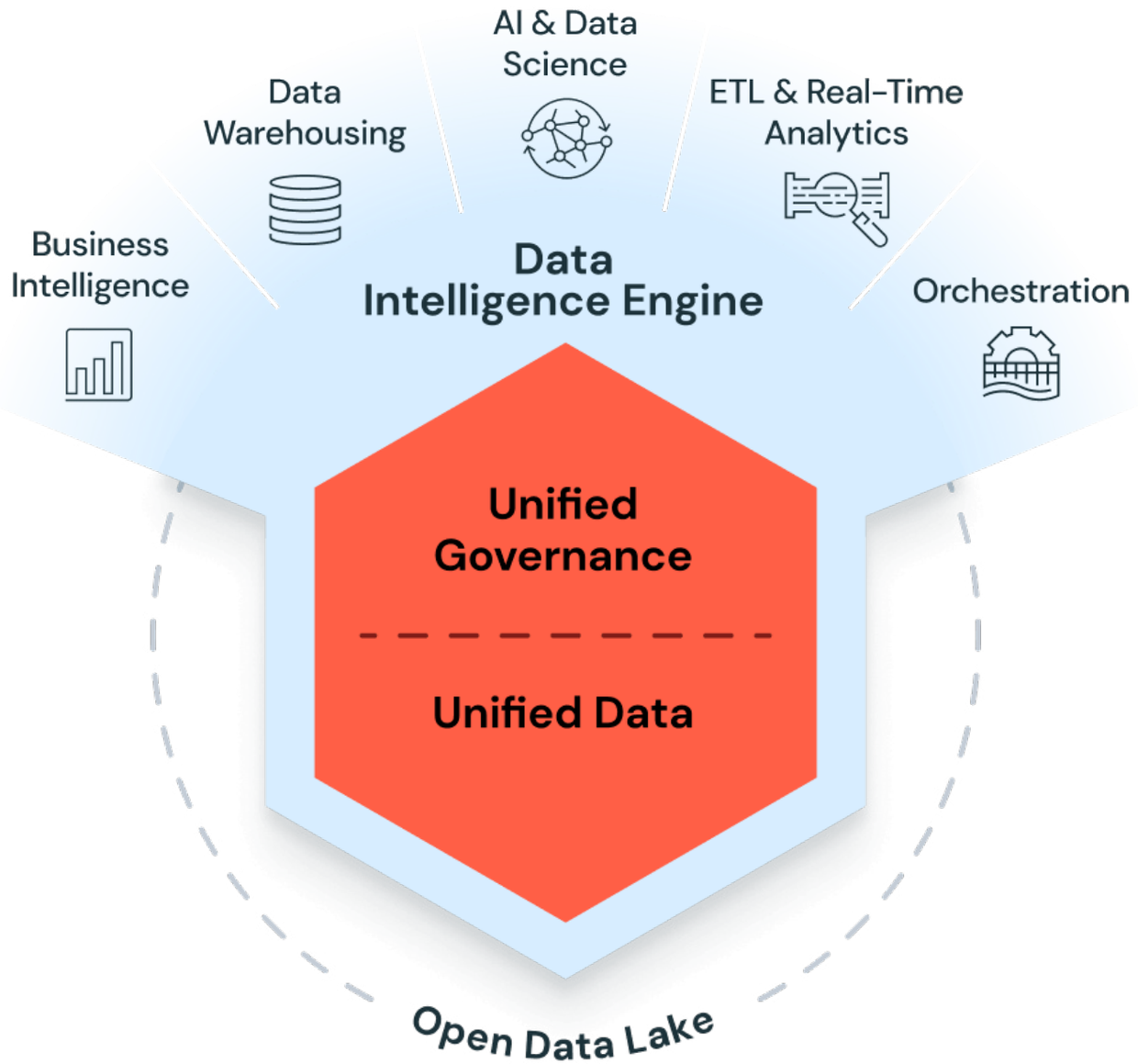
The Databricks Data Intelligence Platform

databricks

What is 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 a Data Intelligence Engine that understands the uniqueness of your data.

The winners in every industry will be data and AI companies. From ETL to data warehousing to generative AI, Databricks helps you simplify and accelerate your data and AI goals.



Benefits of the Databricks Data Intelligence Platform

Intelligent

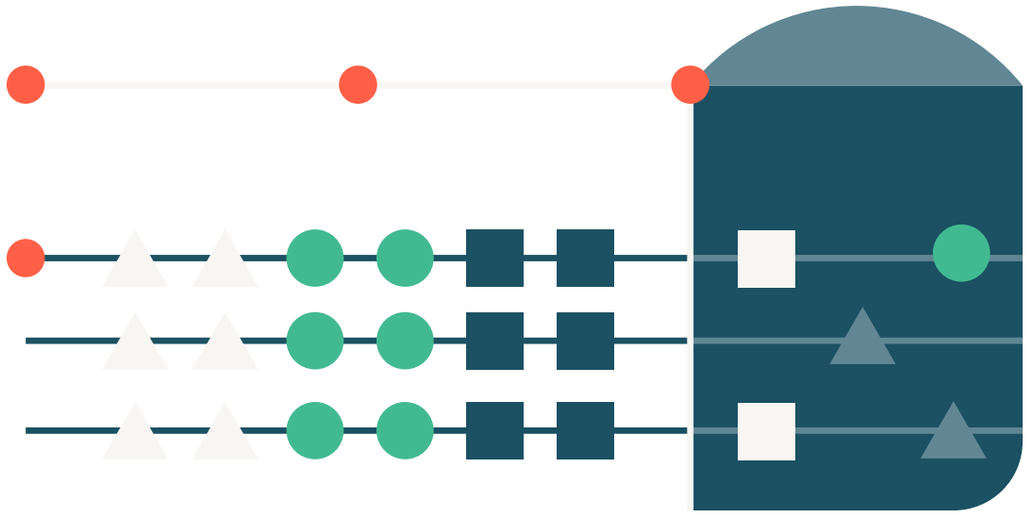
Databricks combines generative AI with the unification benefits of a lakehouse to power a Data Intelligence Engine that understands the unique semantics of your data. This allows the Databricks Platform to automatically optimize performance and manage infrastructure in ways unique to your business.

Simple

Natural language substantially simplifies the user experience on Databricks. The Data Intelligence Engine understands your organization's language, so search and discovery of new data is as easy as asking a question like you would to a coworker. Additionally, developing new data and applications is accelerated through natural language assistance to write code, remediate errors and find answers.

Private

Data and AI applications require strong governance and security, especially with the advent of generative AI. Databricks provides an end-to-end MLOps and AI development solution that's built upon our unified approach to governance and security. You're able to pursue all your AI initiatives — from using APIs like OpenAI to custom-built models — without compromising data privacy and IP control.



The Databricks Data Intelligence Platform architecture

Data reliability and performance

Delta Lake is an open format storage layer built for the data intelligence platform that integrates with all major analytics tools and works with the widest variety of formats to store and process data.

Photon is the next-generation query engine built for the lakehouse that leverages a state-of-the-art vectorized engine for fast querying and provides the best performance for all workloads in the lakehouse.

In Chapter 3, we explore the details of data reliability and performance for the lakehouse.

Unified governance and security

The Databricks Data Intelligence Platform provides unified governance with enterprise scale, security and compliance. The Databricks Unity Catalog (UC) provides governance for your data and AI assets in the lakehouse — files, tables, dashboards and machine learning models — giving you much better control, management and security across clouds.

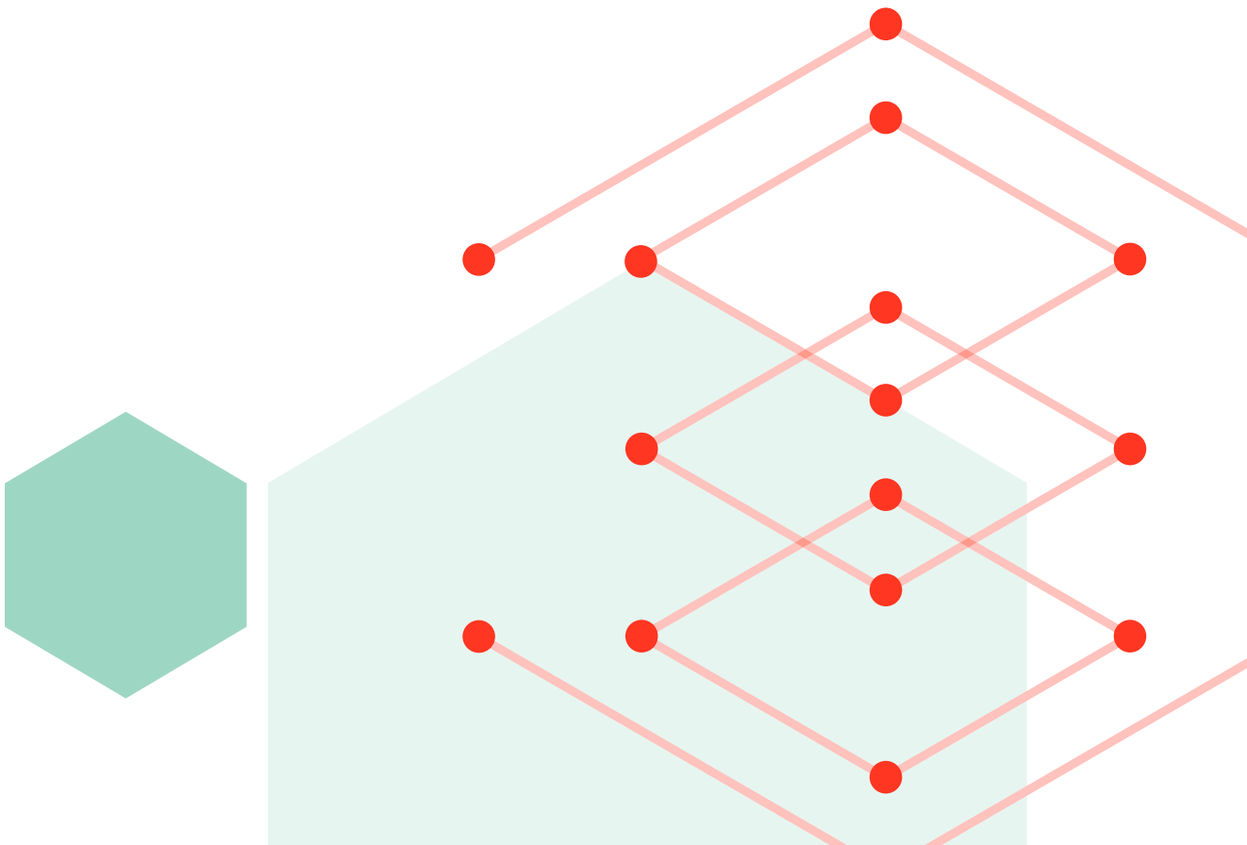
Delta Sharing is an open protocol that allows companies to securely share data across the organization in real time, independent of the platform on which the data resides.

In Chapter 4, we go into the details of unified governance for lakehouse and, in Chapter 5, we dive into the details of security for lakehouse.

Instant compute and serverless

Serverless compute is a fully managed service where Databricks provisions and manages the compute layer on behalf of the customer in the Databricks cloud account instead of the customer account. As of the current release, serverless compute is supported for use with Databricks SQL, Delta Live Tables, Workflows, and Notebooks.

In Chapter 6, we explore the details of instant compute and serverless for the platform.



databricks

The Databricks Data Intelligence Platform workloads

The Databricks Data Intelligence Platform architecture supports different workloads such as data warehousing, data engineering, data streaming, data science and machine learning on one simple, open and multicloud data platform.

Data warehousing

Databricks SQL is a serverless data warehouse built on lakehouse architecture that lets you run all your BI and ETL workloads at scale with up to 12x better price/performance, a unified governance model, open formats and APIs, and your tools of choice — no lock-in.

In Chapter 7, we go into the details of data warehousing on the lakehouse.

Data engineering

Easily ingest and transform batch and streaming data on the Databricks Data Intelligence Platform. Orchestrate reliable production workflows while Databricks automatically manages your infrastructure at scale. Increase the productivity of your teams with built-in data quality testing and support for software development best practices.

In Chapter 8, we go into the details of data engineering on the lakehouse.

Data streaming

The Databricks Data Intelligence Platform dramatically simplifies data streaming to deliver real-time analytics, machine learning and applications on one platform.

Enable your data teams to build streaming data workloads with the languages and tools they already know. Simplify development and operations by automating the production aspects associated with building and maintaining real-time data workloads. Eliminate data silos with a single platform for streaming and batch data.

In Chapter 9, we go into the details of data streaming on the lakehouse.

Data science, AI and machine learning

Streamline the end-to-end data science workflow — from data prep to modeling to sharing insights — with a collaborative and unified data science environment built on an open lakehouse foundation. Get quick access to clean and reliable data, preconfigured compute resources, IDE integration, multi-language support, and built-in advanced visualization tools for maximum flexibility for data analytics teams.

In Chapter 10, we explore the details of data science, AI and machine learning on the data intelligence platform.

Databricks Data Intelligence Platform and your modern data stack

Partner Connect makes it easy for you to discover data, analytics and AI tools directly within the Databricks platform — and quickly integrate the tools you already use today. With Partner Connect, you can simplify tool integration to just a few clicks and rapidly expand the capabilities of your lakehouse.

In Chapter 11, we go into the details of our technology partners and the modern data stack.

Global adoption of the Databricks Data Intelligence Platform

Today, Databricks has over 10,000+ customers, from Fortune 500 to unicorns across industries doing transformational work. Organizations around the globe are driving change and delivering a new generation of data, analytics and AI applications. We believe that the unfulfilled promise of data and AI can finally be fulfilled with one platform for data analytics, data science and machine learning with the Databricks Data Intelligence Platform.

Learn more

Databricks Data Intelligence Platform

Databricks Data Intelligence Platform Demo Center

Databricks Data Intelligence Platform Customer Stories

Databricks Data Intelligence Platform Documentation

Databricks Data Intelligence Platform Training and Certification

Databricks Data Intelligence Platform Resources

CHAPTER

03

Data reliability and performance

To bring openness, reliability and lifecycle management to data lakes, the Databricks lakehouse architecture is built on the foundation of Delta Lake. Delta Lake solves challenges around unstructured/structured data ingestion, the application of data quality, difficulties with deleting data for compliance or issues with modifying data for data capture.

Although data lakes are great solutions for holding large quantities of raw data, they lack important attributes for data reliability and quality and often don't offer good performance when compared to data warehouses.

Problems with today's data lakes

When it comes to data reliability and quality, examples of these missing attributes include:

- **Lack of ACID transactions:** Makes it impossible to mix updates, appends and reads
- **Lack of schema enforcement:** Creates inconsistent and low-quality data. For example, rejecting writes that don't match a table's schema.
- **Lack of integration with data catalog:** Results in dark data and no single source of truth

Even just the absence of these three attributes can cause a lot of extra work for data engineers as they strive to ensure consistent high-quality data in the pipelines they create.

As for performance, data lakes use object storage, so data is mostly kept in immutable files leading to the following problems:

- **Ineffective partitioning:** In many cases, data engineers resort to "poor man's" indexing practices in the form of partitioning that leads to hundreds of dev hours spent tuning file sizes to improve read/write performance. Often, partitioning proves to be ineffective over time if the wrong field was selected for partitioning or due to high cardinality columns.
- **Too many small files:** With no support for transactions, appending new data takes the form of adding more and more files, leading to "small file problems," a known root cause of query performance degradation.

These challenges are solved with two key technologies that are at the foundation of the lakehouse: Delta Lake and Photon.

What is Delta Lake?

Delta Lake is the only open format storage layer that can automatically and instantly translate across open formats. Delta Lake unifies all data types for transactional, analytical and AI use cases out of the box, with support for streaming and batch operations. Delta Lake offers industry-leading performance and is the foundation of a cost-effective, highly scalable lakehouse.

With Delta Lake Universal Format (UniForm), you will be able to use your favorite Iceberg or Hudi client to read your Delta tables through the Unity Catalog endpoint. Delta Lake 3.0 simplifies the connector ecosystem. Delta Kernel offers a stable library API, making it easier for connectors to incorporate new Delta features without code changes.

Delta Lake features

ACID guarantees

Delta Lake ensures that all data changes written to storage are committed for durability and made visible to readers atomically. In other words, no more partial or corrupted files.

Scalable data and metadata handling

Since Delta Lake is built on data lakes, all reads and writes using Spark or other distributed processing engines are inherently scalable to petabyte-scale. However, unlike most other storage formats and query engines, Delta Lake leverages Spark to scale out all the metadata processing, thus efficiently handling metadata of billions of files for petabyte-scale tables.

Audit history and time travel

The Delta Lake transaction log records details about every change made to data, providing a full audit trail of the changes. These data snapshots allow developers to access and revert to earlier versions of data for audits, rollbacks or to reproduce experiments.

Schema enforcement and schema evolution

Delta Lake automatically prevents the insertion of data with an incorrect schema, i.e., not matching the table schema. And when needed, it allows the table schema to be explicitly and safely evolved to accommodate ever-changing data.

Support for deletes, updates and merges

Most distributed processing frameworks do not support atomic data modification operations on data lakes. Delta Lake supports merge, update and delete operations to enable complex use cases including but not limited to change data capture (CDC), slowly changing dimension (SCD) operations and streaming upserts.

Streaming and batch unification

A Delta Lake table can work both in batch and as a streaming source and sink. The ability to work across a wide variety of latencies, ranging from streaming data ingestion to batch historic backfill, to interactive queries all work out of the box.