

eBook

# Data Management 101, 2nd Edition

Learn how Databricks streamlines the data management lifecycle



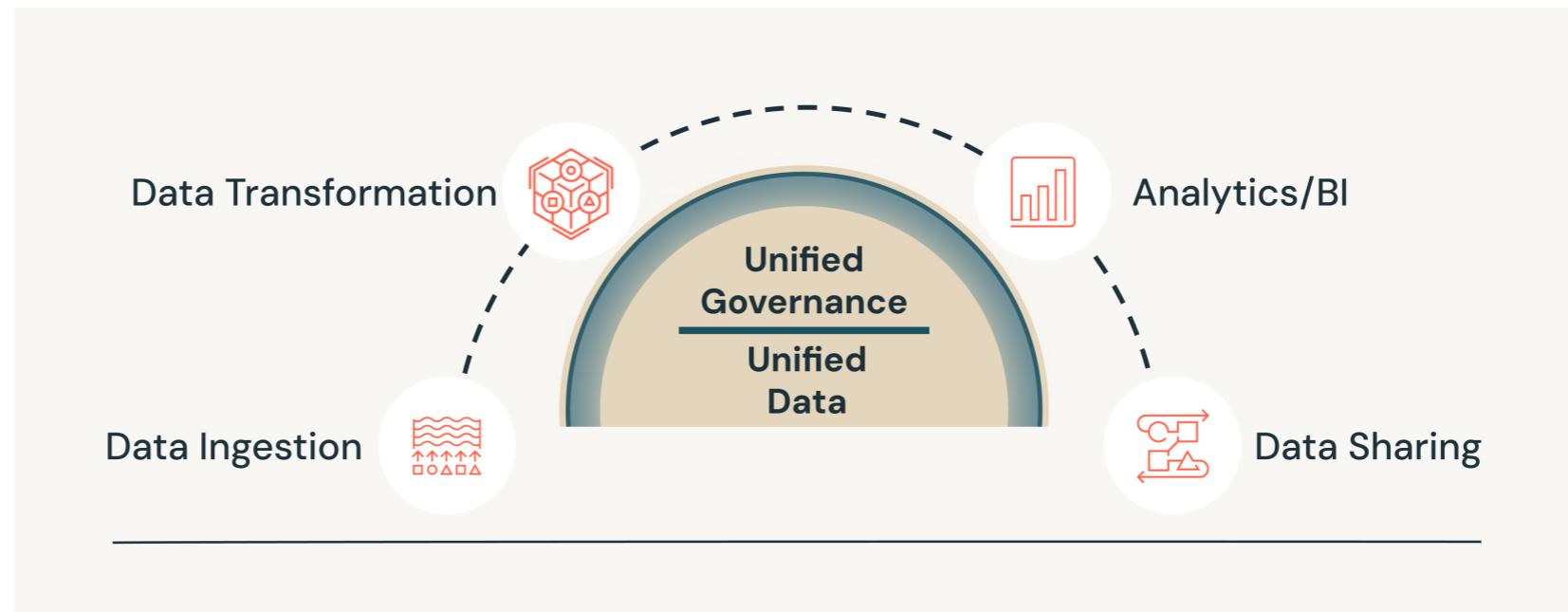
# Contents

<b>Introduction</b> .....	3
The challenges of data management .....	4
<b>Data Management on Databricks</b> .....	6
Data and AI governance .....	8
<b>Data Ingestion</b> .....	12
<b>Data Transformation, Quality and Processing</b> .....	16
<b>Working With Data</b> .....	19
Empowering decisions with data warehousing.....	19
AI-driven business intelligence: Democratizing data across organizations .....	21
<b>Data Sharing and Collaboration</b> .....	22
<b>The Role of AI in Data Management</b> .....	24
Bringing AI to the data lakehouse.....	25
Building enterprise AI on the foundation of data management .....	26
<b>Conclusion</b> .....	27
<b>About Databricks</b> .....	28

## Introduction

Companies are rapidly adopting the data lakehouse architecture to enable their organizations to better use data for analytics and AI use cases. A shift toward the lakehouse means thinking differently about the lifecycle of data.

Data management has been a common practice across industries for many years, although not all organizations have used the term the same way. At Databricks, we view data management as all disciplines related to the lifecycle of data as a strategic and valuable resource, which includes collecting data, processing data, governing data, sharing data, analyzing it and optimizing it — and doing this all in a cost-efficient, effective and reliable manner.



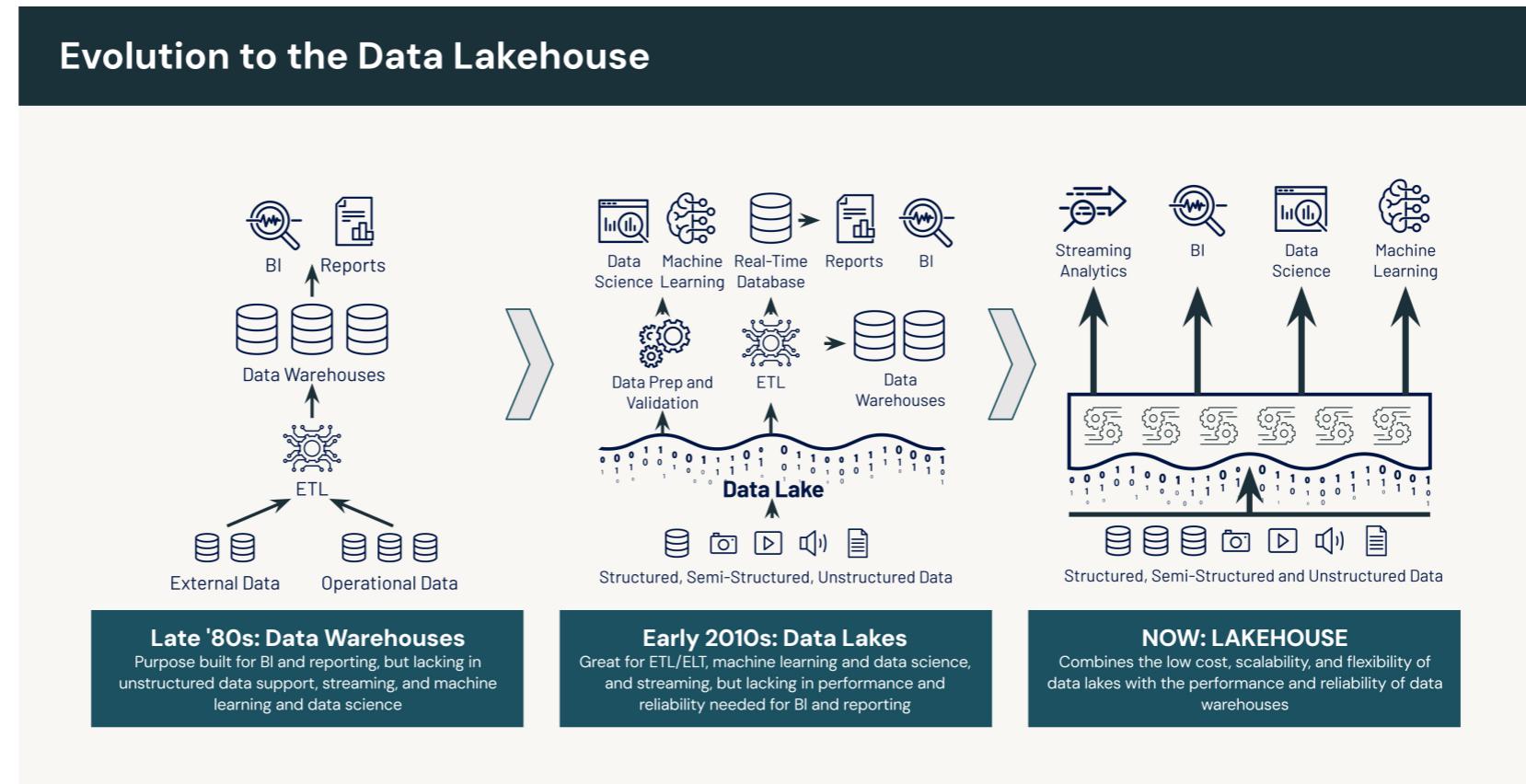
## The challenges of data management

Ultimately, the consistent and reliable flow of data across people, teams and business functions is crucial to an organization's survival and ability to innovate. Organizations are increasingly recognizing the strategic importance of their data through various applications, including generative AI. This includes leveraging data to drive product innovation, facilitating enhanced collaboration among teams and accelerating entry into new market channels. According to [MIT Technology Review Insights](#), 99% of adopters of a data lakehouse architecture achieve their data and AI goals. But this means it's even more important for data to be trustworthy, processed quickly, and governed.

The vast majority of company data today flows into a data lake, where teams do data prep and validation to serve downstream data science and machine learning initiatives. At the same time, a huge amount of data is transformed and sent to many different downstream data warehouses for business intelligence (BI) because traditional data lakes are too slow and unreliable for BI workloads.

Depending on the workload, data sometimes also needs to be moved out of the data warehouse and back to the data lake. And increasingly, machine learning workloads are also reading and writing to data warehouses. The underlying reason why this kind of data management is challenging is that there are inherent differences between data lakes and data warehouses.

On one hand, data lakes do a great job supporting machine learning — they have open formats and a big ecosystem — but they have poor support for business intelligence and suffer from complex data quality problems. On the other hand, we have data warehouses that are great for BI applications, but they have limited support for machine learning workloads, and they are proprietary systems with only a SQL interface.



Moreover, data and usage patterns change over time. As data is added to the data lake and is processed into the data warehouse, schemas need to adapt to changing data types and sources. New analytics and AI use cases result in queries that join data in more complex ways. As a result, tables that were optimized for older use cases may not perform well over time. The traditional approach to handling this is to manually repartition and recluster data. It is a time-consuming, complicated and sometimes costly process that often gets deprioritized in favor of new development.

# Data Management on Databricks

Unifying these systems can be transformational in how we think about data. The [Databricks Data Intelligence Platform](#) does just that — unifies all these disparate workloads, teams and data, and provides an end-to-end data management platform for all phases of the data lifecycle.

At the core of the Data Intelligence Platform is an open data lakehouse. Organizations own their data and store it in their preferred cloud data storage in [Parquet-based](#) open source table formats like [Delta Lake](#), [Apache Iceberg™](#), CSV, JSON, AVRO, or other semi- and unstructured data types. Why open source formats? They are portable. With an open data lakehouse, there is no vendor lock-in, either in the format or the storage location.

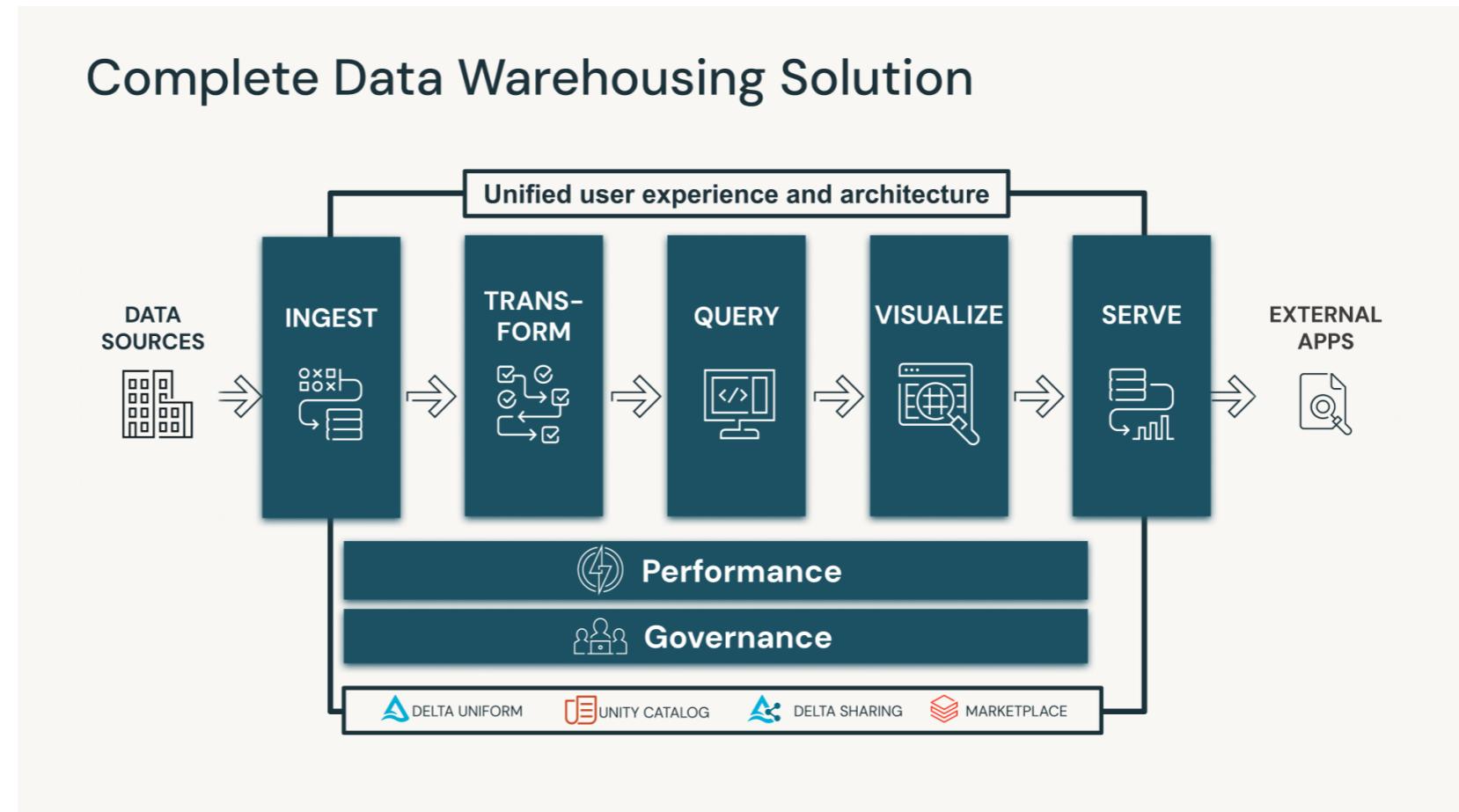
Historically, lakehouses have only been able to support a single open table format, resulting in fragmentation across ecosystems. Organizations have had to choose their platform based on their preferred format, which restricted their choice of compute engine for analysis.

In addition, different lakehouse vendors have historically offered their own catalogs for data discovery and governance. However, each catalog has restrictions on the read or write access for various analytics tools and compute engines. The net result is further fragmentation across the lakehouse ecosystem. No single vendor catalog has had a view of data and AI assets across the entire ecosystem.

At Databricks, [Unity Catalog](#) is the key to solving both of these challenges. Unity Catalog manages reads and writes across engines and formats, including both Delta Lake and Iceberg. Unity Catalog is a full implementation of the [Iceberg REST Catalog API](#), the canonical catalog spec for Iceberg support.

Unity Catalog also offers advanced cataloging capabilities that provide a single view across data assets so it serves as a single entry point to implement governance rules across assets, regardless of format. Teams can access and govern data in foreign catalogs without having to make copies of metadata or data files, because Unity Catalog offers federation and mirroring capabilities.

[Unity Catalog](#) brings unified governance, open connectivity and AI-enabled optimizations to make it easier to implement the data management lifecycle on Databricks.



Learn more about the [Databricks Data Intelligence Platform](#)

Learn more about [Unity Catalog](#)

## Data and AI governance

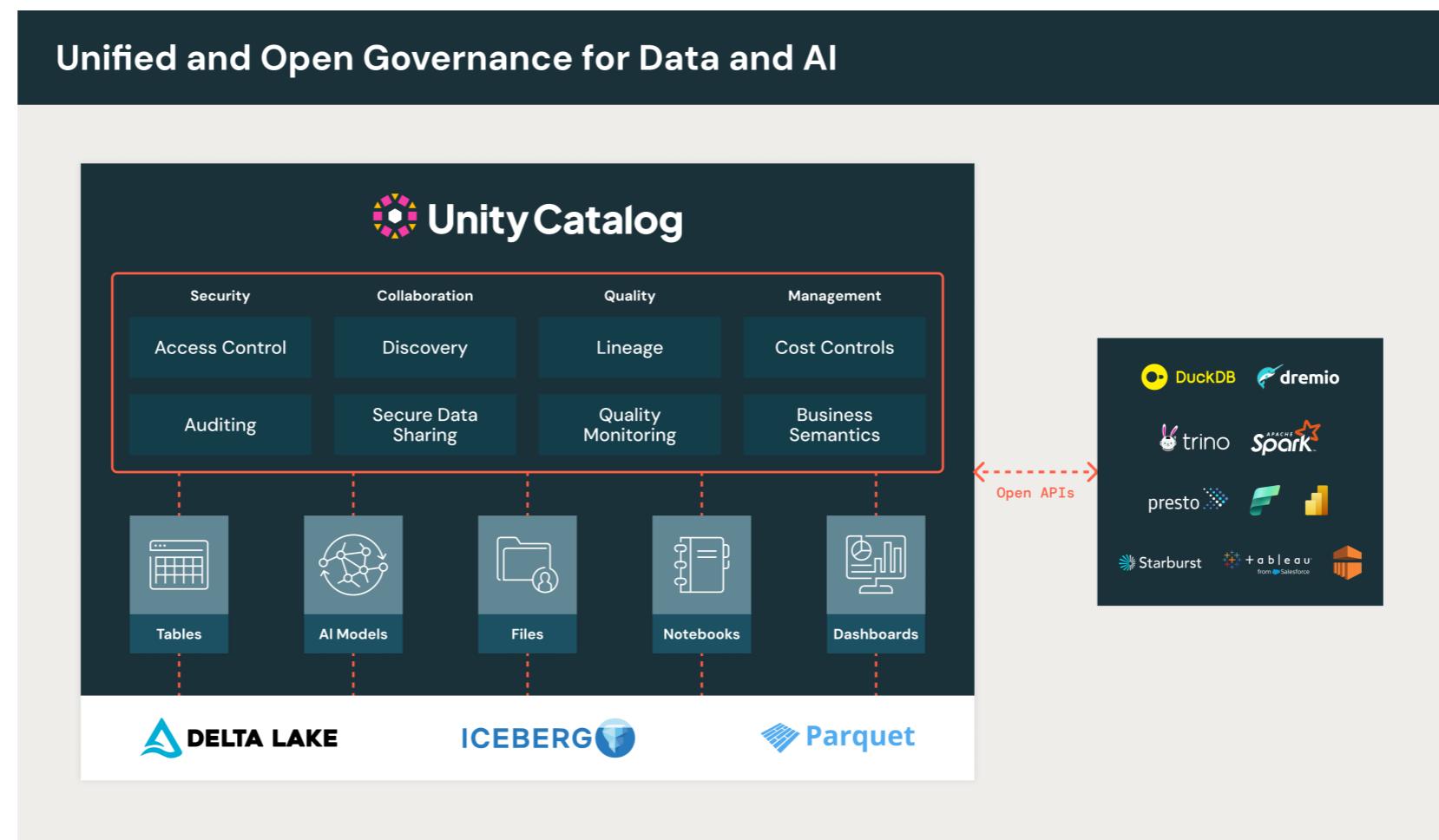
Effective governance is key to making data and AI accessible in the age of generative AI. However, it is also very complex in today's rapidly evolving data and AI landscape. Let's look at why.

Today, organizations are generating enormous amounts of data and AI resources, but they struggle with inconsistent governance across various elements such as structured and unstructured data, files, AI tools, notebooks, dashboards and machine learning models. This complexity is compounded by different data formats like Apache Iceberg, Delta Lake and Parquet, which make it difficult to integrate and standardize data. Additionally, organizations often rely on separate tools for security, cataloging, monitoring and tracking, each with its own limitations and lack of cohesion. This fragmentation of governance leads to operational inefficiencies, elevates compliance risks and hampers innovation. Inconsistent management across formats and tools creates data silos and reduces data quality. It also drives up costs and complicates decision-making as organizations face difficulties maintaining a cohesive view of their data and AI landscapes.

Additionally, organizations are increasingly adopting a wide range of data and AI tools and sourcing data from diverse origins, with teams seeking tailored, best-in-class solutions. However, cross-platform data sharing, connectivity to various data sources and interoperability between tools remain limited. This creates vendor lock-in, limiting flexibility to switch providers or adopt new technologies. Poor interoperability and fragmented data sharing hinder collaboration and scalability, resulting in underutilized data assets, higher costs and missed growth opportunities.

Finally, today's data and AI platforms often lack the built-in intelligence needed to connect business concepts with the underlying data. This gap means organizations depend heavily on technical experts to interpret data into actionable insights, creating a bottleneck. This bottleneck restricts access and use across the organization, especially for nontechnical users, slowing innovation, delaying decisions and limiting the competitive advantage of data and AI.

To address these key governance challenges, the Databricks Data Intelligence Platform provides **Unity Catalog**, the industry's only open and unified governance solution for managing all data and AI assets. As the cornerstone of your data intelligence strategy, Unity Catalog combines the power of lakehouse and AI to deliver contextual, domain-specific insights that boost productivity for both technical and business users across any workload. With an open source foundation, Unity Catalog enables the discovery, access and sharing of trusted data and AI assets across any tool, compute engine or cloud. This unified, open approach drives better collaboration, accelerates data and AI initiatives, and simplifies compliance in a rapidly evolving data landscape.



## Unified governance for data and AI

- Build an enterprise catalog for the curation of all structured and unstructured data, ML models, AI tools, notebooks, metrics
- Leverage any open data formats of your choice, including Delta, Iceberg and Parquet
- Simplify security and compliance through a unified interface for access management and auditing
- Understand data flow and dependencies with automated lineage for data and ML
- Scale and simplify governance with tag-based and attribute-based access controls
- Gain enhanced security with fine-grained access controls on rows and columns
- Monitor and manage usage and cost with out-of-the-box observability dashboards
- Ensure data and AI quality with built-in monitoring and alerting capabilities
- Track sensitive data and AI assets with rich tagging and auto-classification

## Open access and collaboration

- Break down data silos across databases, data warehouses and catalogs with built-in federation capabilities
- Access data and AI assets from any compute engine or tool of your choice with open APIs
- Share data and AI assets across data platforms, clouds and regions without data replication
- Collaborate with your business units and partners on sensitive data across clouds, regions and platforms in a privacy-safe manner

## Built-in data intelligence

- Democratize data and AI with context-aware search and discovery and auto-generated data insights for everyone — from data practitioners to business users
- Accelerate insights with a context-aware assistant that provides domain intelligence for any workload and user
- Drive clarity, better understanding and data discovery with auto-generated comments and tags
- Maximize performance with AI-powered table optimizations that simplify your workflow, reducing complexity and letting the platform handle the fine-tuning

“Databricks Unity Catalog is now an integral part of the PepsiCo Data Foundation, our centralized global system that consolidates over 6 petabytes of data worldwide. It streamlines the onboarding process for more than 1,500 active users and enables unified data discovery for our 30+ digital product teams across the globe, supporting both business intelligence and artificial intelligence applications.”

— Bhaskar Palit, Senior Director, Data and Analytics



Learn more about [Unity Catalog](#)

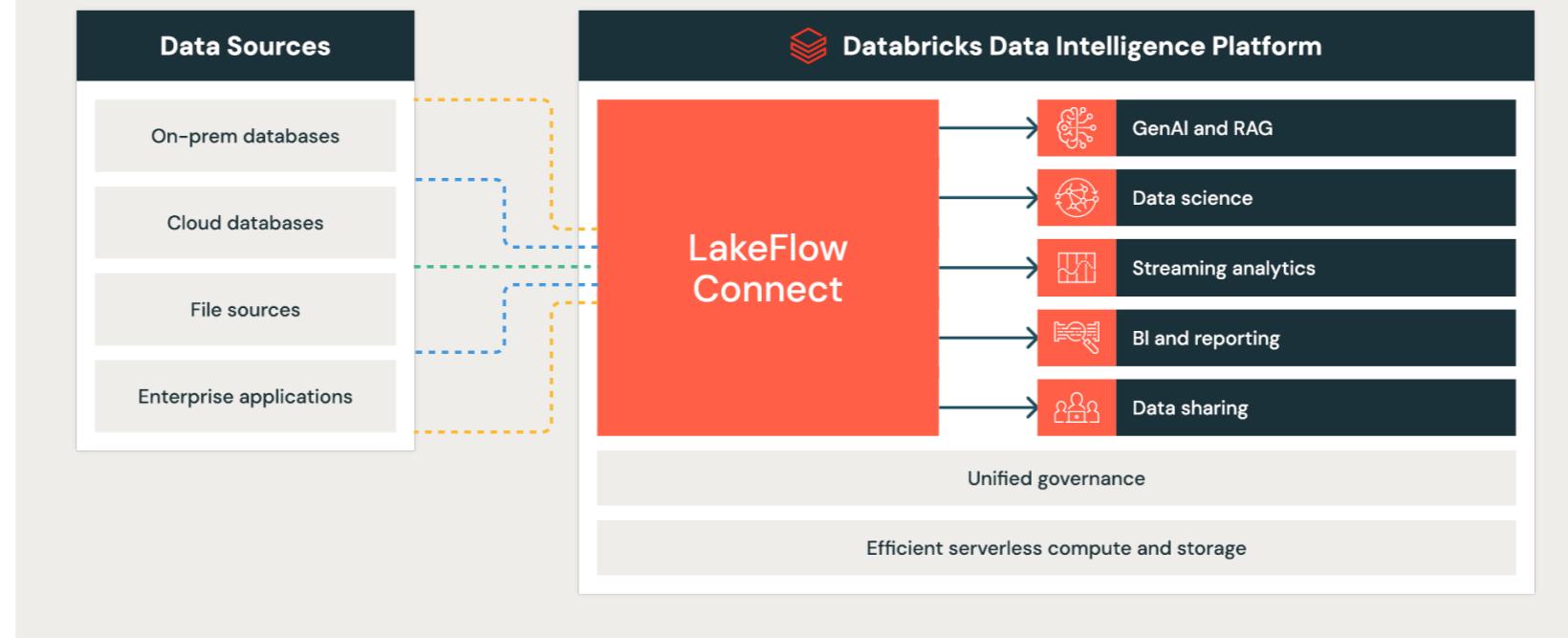
[Download free eBook](#) to learn more about data and AI governance on the Databricks Data Intelligence Platform.

## Data Ingestion

In today's world, IT organizations are inundated with data siloed across various, often proprietary on-premises application systems, databases, data warehouses and SaaS applications. This fragmentation makes it difficult to support new use cases for analytics or machine learning. Data teams often require the creation of complex and unstable connectors to ingest data, with data preparation that involves maintaining intricate logic, which can cause system failures or latency spikes, resulting in a poor customer experience.

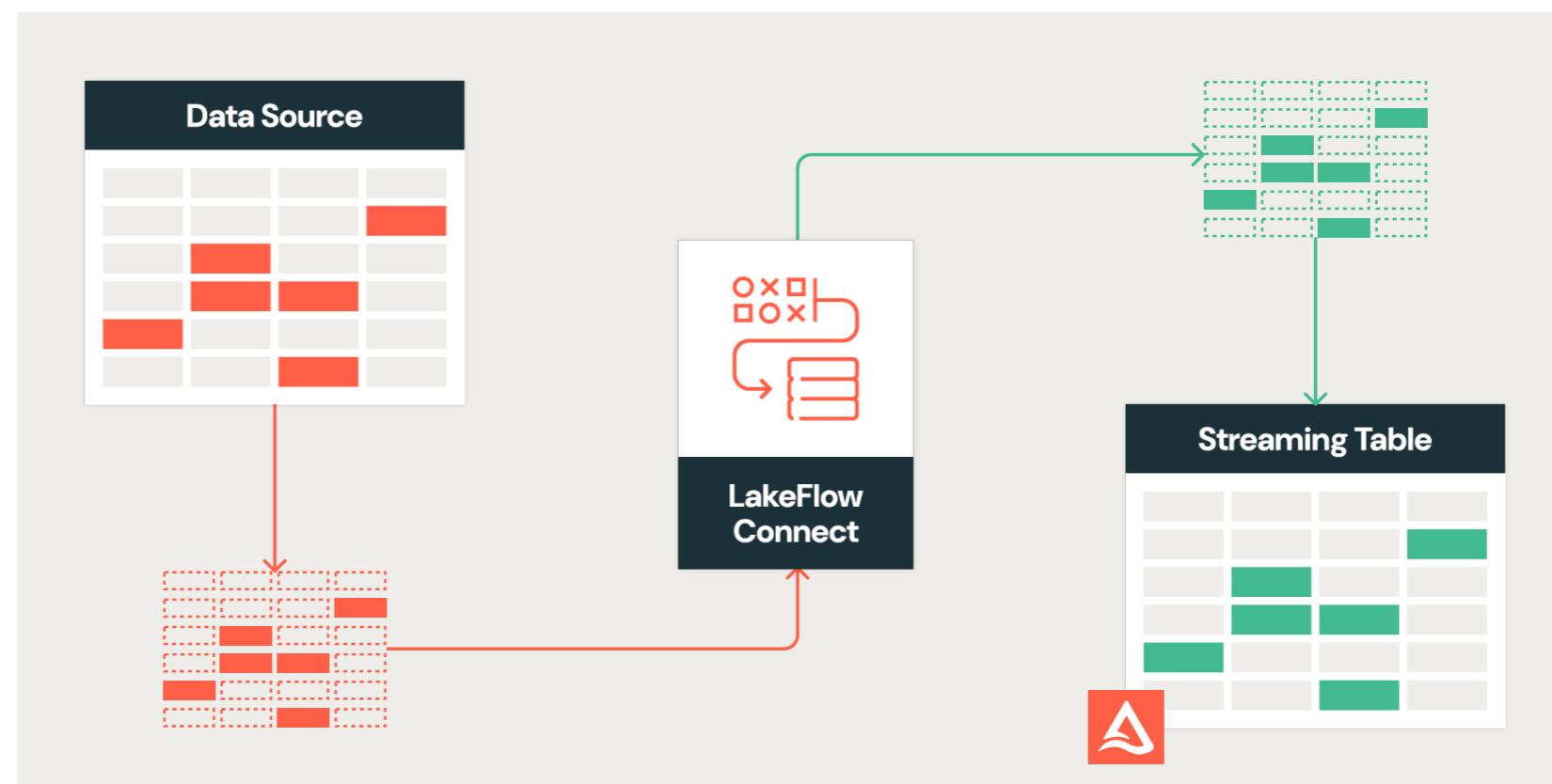
The biggest challenge many data engineers face today is efficiently moving data from various systems into a single, open and unified lakehouse architecture.

### First-Party Ingestion for the Data Intelligence Platform



## Databricks LakeFlow

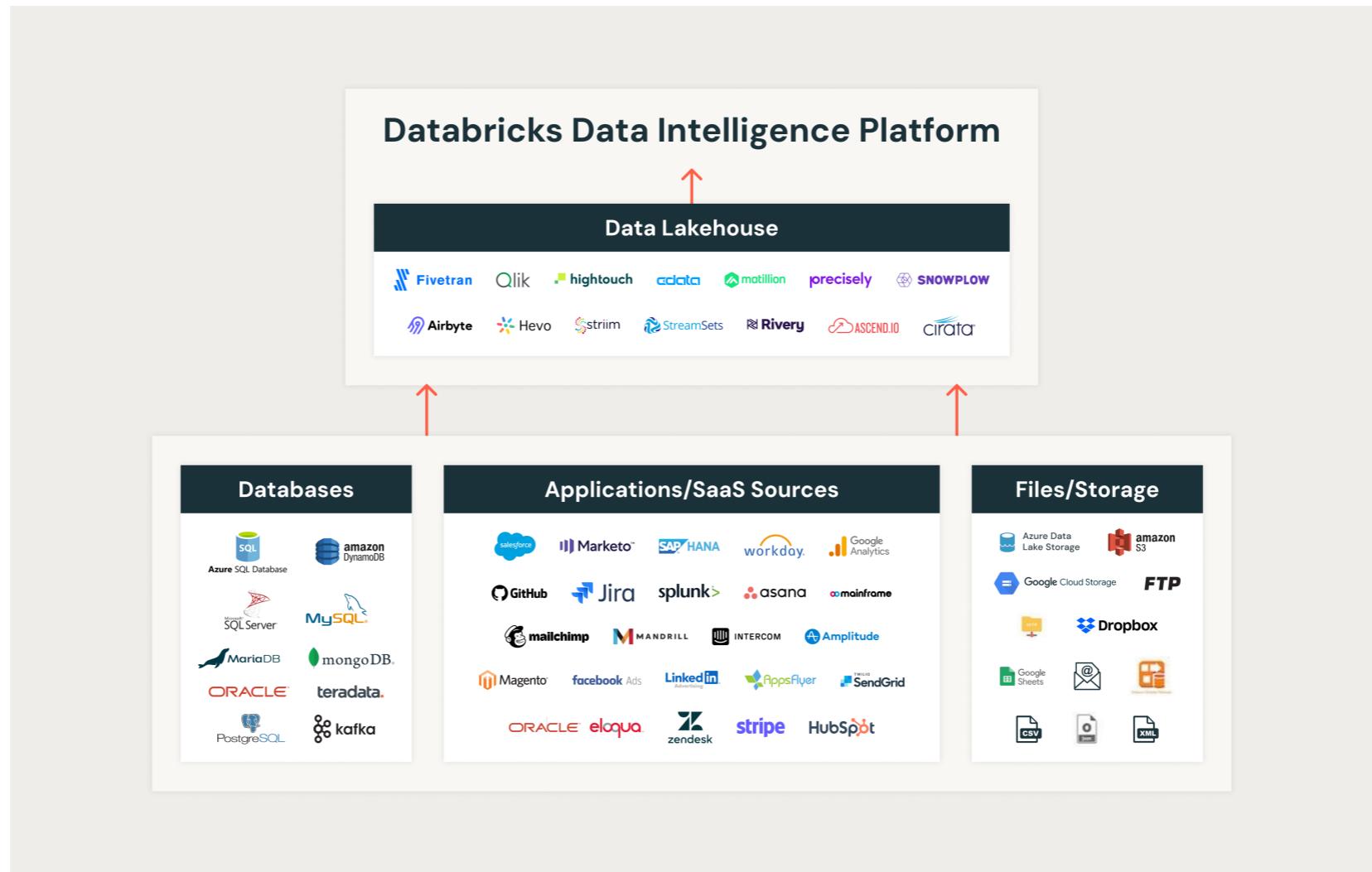
Databricks has announced Databricks LakeFlow — a unified, intelligent solution for data engineering. Native to the Databricks Data Intelligence Platform, LakeFlow empowers users to easily ingest the data they need from external sources (LakeFlow Connect), build and operate data pipelines (LakeFlow Pipelines), and orchestrate anything on the data platform (LakeFlow Jobs). As this unified data engineering experience is built, data professionals can continue taking advantage of all existing tooling, with no manual migrations required. Read the [announcement blog](#) for more information.



LakeFlow Connect offers native connectors for popular data sources. Databricks makes it easier to ingest data directly from popular SaaS applications such as Salesforce, databases such as SQL Server, and file sources such as SFTP, so any practitioner can build incremental data pipelines at scale. These built-in connectors provide efficient end-to-end incremental ingestion, easy setup with a simple UI or API access, and governance via Unity Catalog — all powered by the Databricks Data Intelligence Platform.

In addition to LakeFlow Connect, Databricks continues to offer Databricks Auto Loader, a connector for cloud object storage that is compatible with Structured Streaming and Delta Live Tables. Auto Loader allows you to incrementally ingest files as they arrive in cloud storage, such as Amazon S3, Azure Data Lake Storage and Google Cloud Storage. Using Delta Live Tables and Auto Loader provides incremental data ingestion and allows practitioners to benefit from scalability, performance, schema inference and evolution support — as well as low cost, low latency and minimal DevOps work.

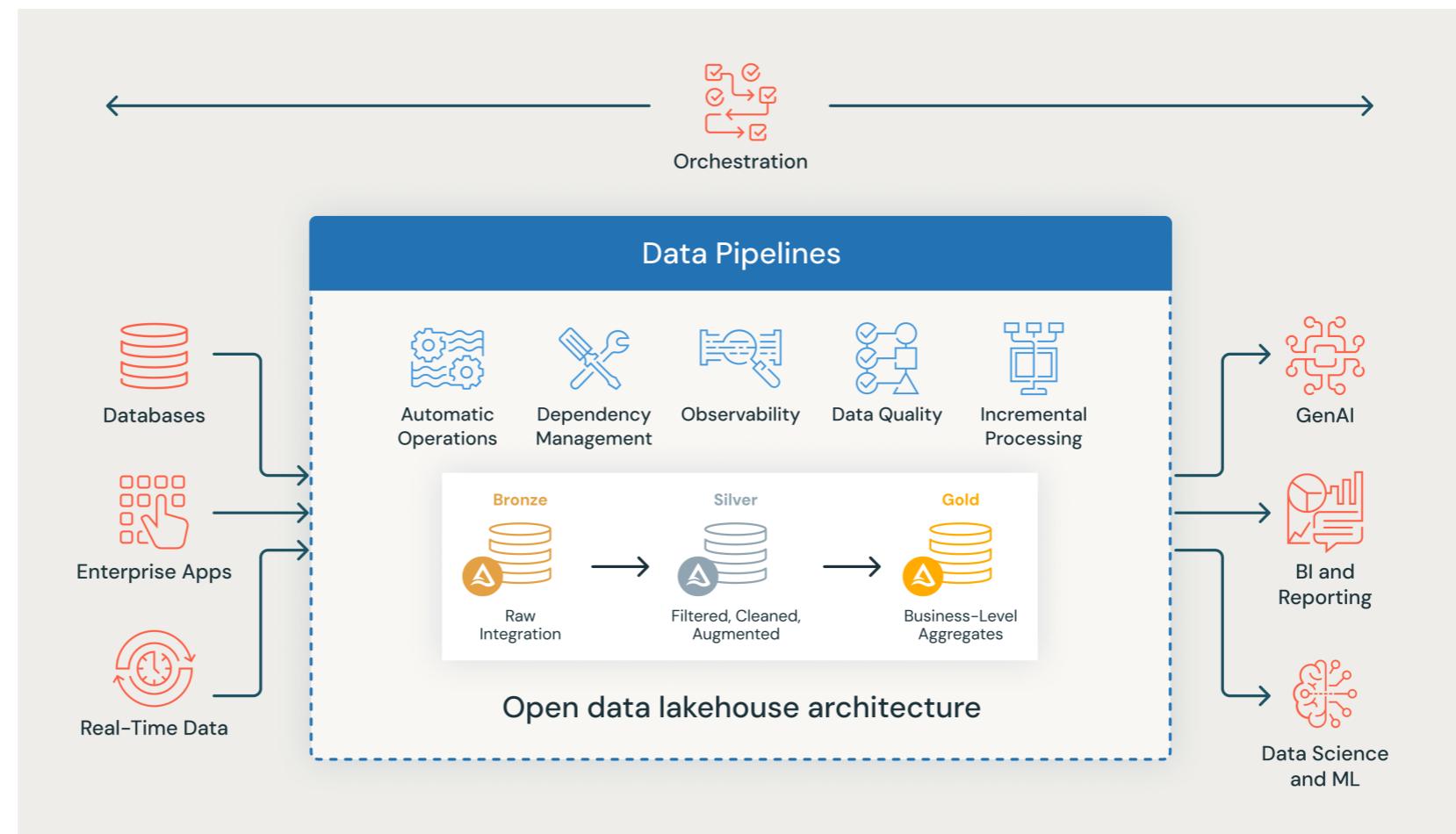
In addition to these native solutions, Databricks has a broad network of data ingestion partners that make it possible to move data from various siloed systems into your data platform. These partners offer a wide range of connectors and native integrations with Databricks to ingest and store data in Delta Lake, making data easily accessible and manageable for data teams. Our partners' solutions enable customers to leverage the reliability and scalability of the Databricks Data Intelligence Platform to innovate faster while deriving valuable data insights. With Databricks Technology Partners, you can choose from 500+ additional pre-built connectors to meet any use case for data engineering.



With the Databricks Data Intelligence Platform, data engineering teams can take that first step of efficiently ingesting any data type into their data lake to extract value.

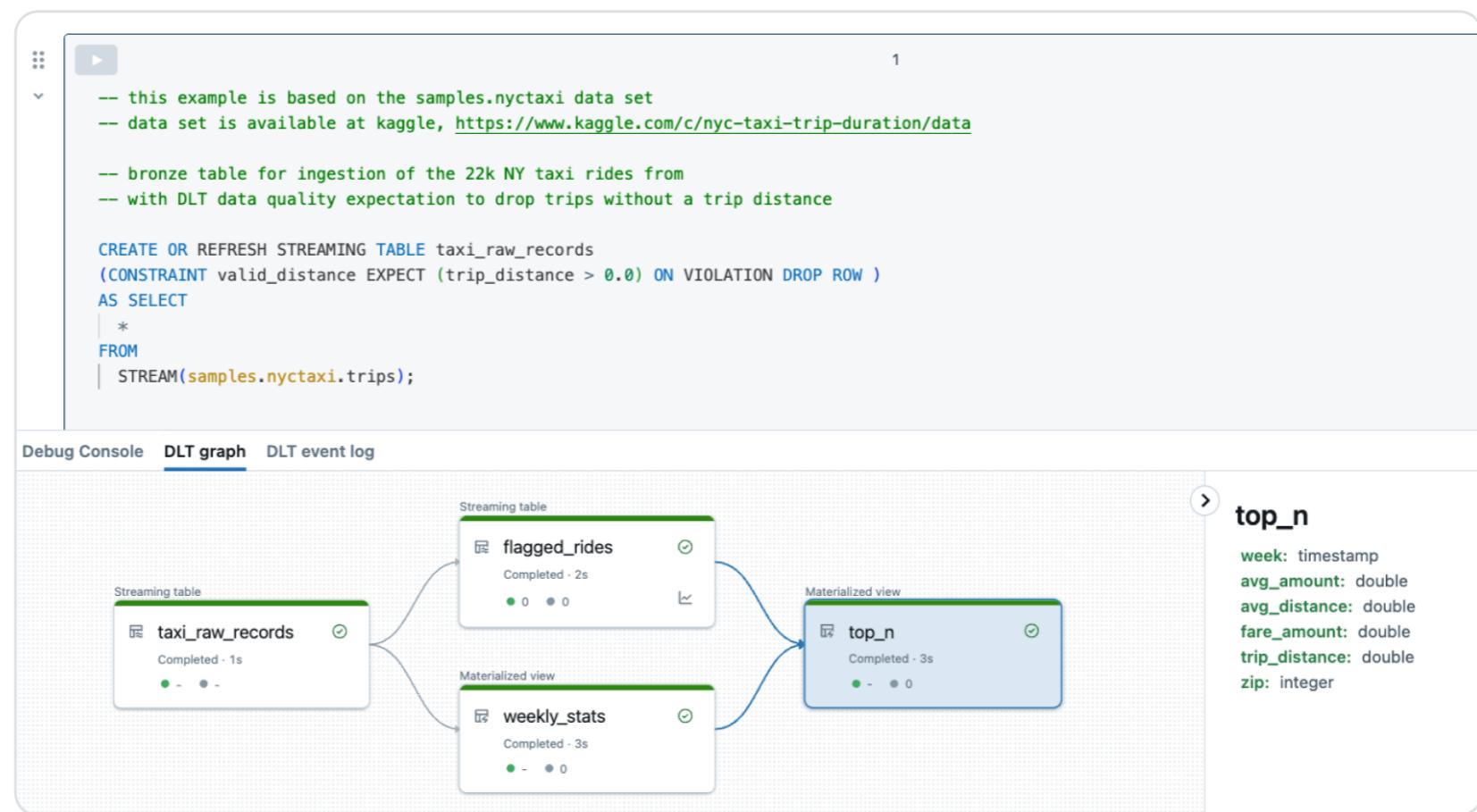
# Data Transformation, Quality and Processing

Moving data into a data lakehouse solves one of the data management challenges, but in order for data analysts or scientists to use it, it must also be transformed into a clean, reliable product for end users. This is an important step, as outdated or unreliable data can lead to mistakes, inaccuracies or distrust.



Data engineers have the difficult and laborious task of cleansing complex, diverse data and transforming it into a format fit for analysis, reporting, data science/machine learning or GenAI use cases. This requires the data engineer to know the ins and outs of the organization's data stack(s), and requires the building of complex queries (transformations) in various languages, stitching together queries for production. For many organizations, the complexity in this phase of the data management lifecycle limits the ability of business groups to extract meaningful value from the source data.

To reduce the complexity of pipeline creation and management, Databricks **Delta Live Tables** (DLT) gives data engineering teams a massively scalable ETL framework to declaratively build data pipelines in SQL or Python. Building pipelines in DLT, data engineers can simply declare the required transformations and let DLT automatically manage task orchestration, cluster management, monitoring, data quality and error handling.



Declarative data pipelines provide a simple way of creating, standardizing and maintaining ETL. These data pipelines autonomously adapt to changes in the data, code or environment, allowing data engineers to focus on developing, validating and testing data that is being transformed. To validate data trustworthiness in real time, data engineers can even define rules about the expected quality of data within the data pipeline. Delta Live Tables enables teams to analyze and monitor data quality continuously to reduce the spread of incorrect and inconsistent data.

“Complex architectures ... were challenging to implement with a classic multicloud data warehouse architecture. Both data scientists and data engineers can now perform such changes using scalable Delta Live Tables with no barriers to entry.”

— Sai Ravuru, Senior Manager of Data Science and Analytics, JetBlue



DLT with serverless compute enables the incremental refresh of complex transformations, allowing for end-to-end incremental processing across the ETL pipeline in both ingestion and transformation. And from just a few lines of code, DLT determines the most efficient way to build and execute your streaming or batch data pipelines, optimizing for price/performance while minimizing complexity.

With all these Delta Live Tables components in place, data engineers can focus solely on transforming, cleansing and delivering quality data for downstream use — analytics or AI.

#### [Learn more about Delta Live Tables](#)

Orchestration is another crucial element of data processing. Data teams must manage the ongoing orchestration of tasks like running ETL or ML pipelines, notebook code, executing scripts, running queries, refreshing dashboards, training models, and so on. To accommodate these needs, [Databricks Workflows](#) lets you easily define, manage and monitor multitask workflows for ETL, analytics and machine learning pipelines.

**Note:** DLT pipelines and Databricks Workflows are evolving to LakeFlow Pipelines and LakeFlow Jobs, respectively, for a more unified authoring and operational experience.

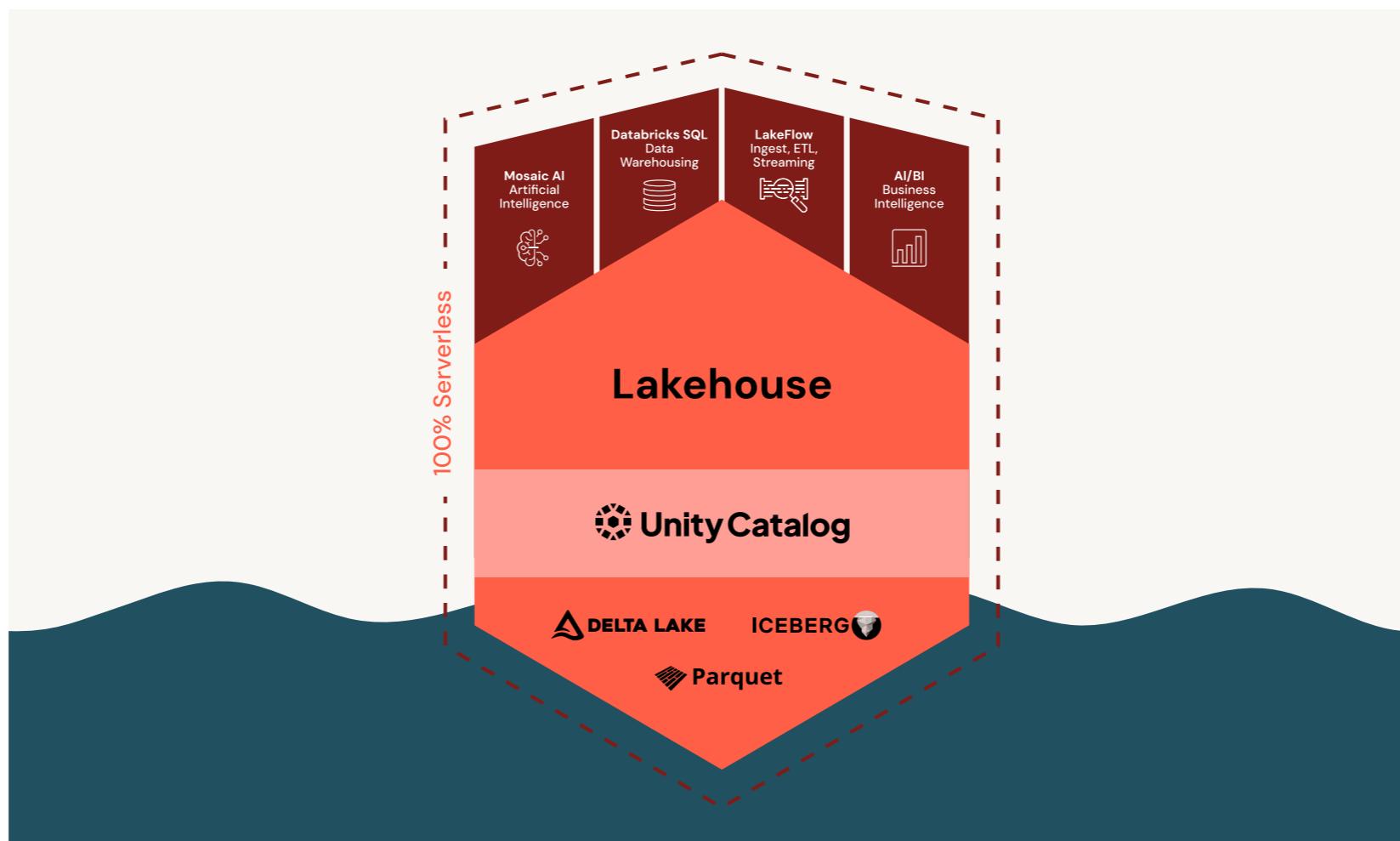
# Working With Data

## Empowering decisions with data warehousing

Now that data is available for consumption, all data workers in an organization can start using that data to derive insights, build models, build apps and drive business decisions.

Typically, to access well-conformed data within a data lake, an analyst would need to leverage Apache Spark™ or use a developer interface to access data. Teams can do this, or they can process and distill data into warehouses for reporting.

To simplify access and query a lakehouse, [Databricks SQL](#) helps data analysts to perform deeper analysis with a SQL-native experience to run BI and SQL workloads on a multicloud lakehouse architecture. Databricks SQL complements existing BI tools with a SQL-native interface that allows data analysts and data scientists to query lakehouse data directly within Databricks.



A dedicated SQL workspace brings familiarity for data analysts to chat with their data in natural language, run ad hoc queries on the lakehouse, create rich visualizations to explore queries from a different perspective and organize those visualizations into drag-and-drop dashboards, which can be shared with stakeholders across the organization. Within the workspace, analysts can explore schema, save queries as snippets for reuse and schedule queries for automatic refresh.

Customers can maximize existing investments by connecting their preferred BI tools to their lakehouse with Databricks SQL endpoints. Reengineered and optimized connectors ensure fast performance, low latency and high user concurrency to your lakehouse. This means that analysts can use the best tool for the job on one single source of truth for your data while minimizing more ETL and data silos.

**“Now more than ever, organizations need a data strategy that enables speed and agility to be adaptable. As organizations are rapidly moving their data to the cloud, we’re seeing growing interest in doing analytics on the data lake. The introduction of Databricks SQL delivers an entirely new experience for customers to tap into insights from massive volumes of data with the performance, reliability and scale they need. We’re proud to partner with Databricks to bring that opportunity to life.”**

— Francois Ajenstat, Chief Product Officer, Tableau

Finally, for governance and administration, administrators can apply SQL data access controls on tables for fine-grained control and visibility over how data is used and accessed across the entire lakehouse for analytics. Additionally, administrators have visibility into Databricks SQL usage, AI usage and costs: the history of all executed queries to understand performance, where each query ran, how long a query ran, which user ran the workload, how many users are engaging with the AI assistant, cost predictions and more. All this information is captured and made available for administrators to easily triage, troubleshoot and understand how people are using their intelligent data warehouse.

[Learn more about Databricks SQL](#)

## AI-driven business intelligence: Democratizing data across organizations

AI is revolutionizing business intelligence by making data more accessible throughout organizations. Databricks AI/BI built for the lakehouse architecture includes Dashboards and Genie, which are redefining data access across organizations, making data-driven insights available to everyone, from business leaders to technical teams. This democratization of data access ensures that each user, regardless of technical skill, can interact with data in meaningful ways, empowering informed decision-making throughout the organization.

**AI/BI Dashboards** offer a user-friendly, low-code interface for creating data visualizations using natural language commands. This intuitive approach enables users to transform complex data into actionable insights without requiring expertise in SQL or other programming languages. The Dashboards allow users to rapidly iterate on questions and visualize answers, fostering an agile approach to data analysis that directly supports strategic decision-making and operational efficiency.

**Genie**, Databricks' conversational tool, takes data interaction a step further by enabling users to ask questions and receive insights through natural language. Tailored to understand each organization's unique terminology and data model, Genie turns data queries into everyday language, helping nontechnical users engage with data in real time. With Genie, employees can ask detailed questions, explore trends and understand metrics on demand, making data insights immediately accessible to those who need them most.

Together, AI/BI Dashboards and AI/BI Genie transform how organizations interact with data, reducing reliance on specialized data teams and making insights accessible across departments. By putting data within reach for all data workers, Databricks AI/BI tools foster a culture of data literacy and inclusivity, empowering each user to contribute to data-driven decisions, innovation and growth.

## Data Sharing and Collaboration

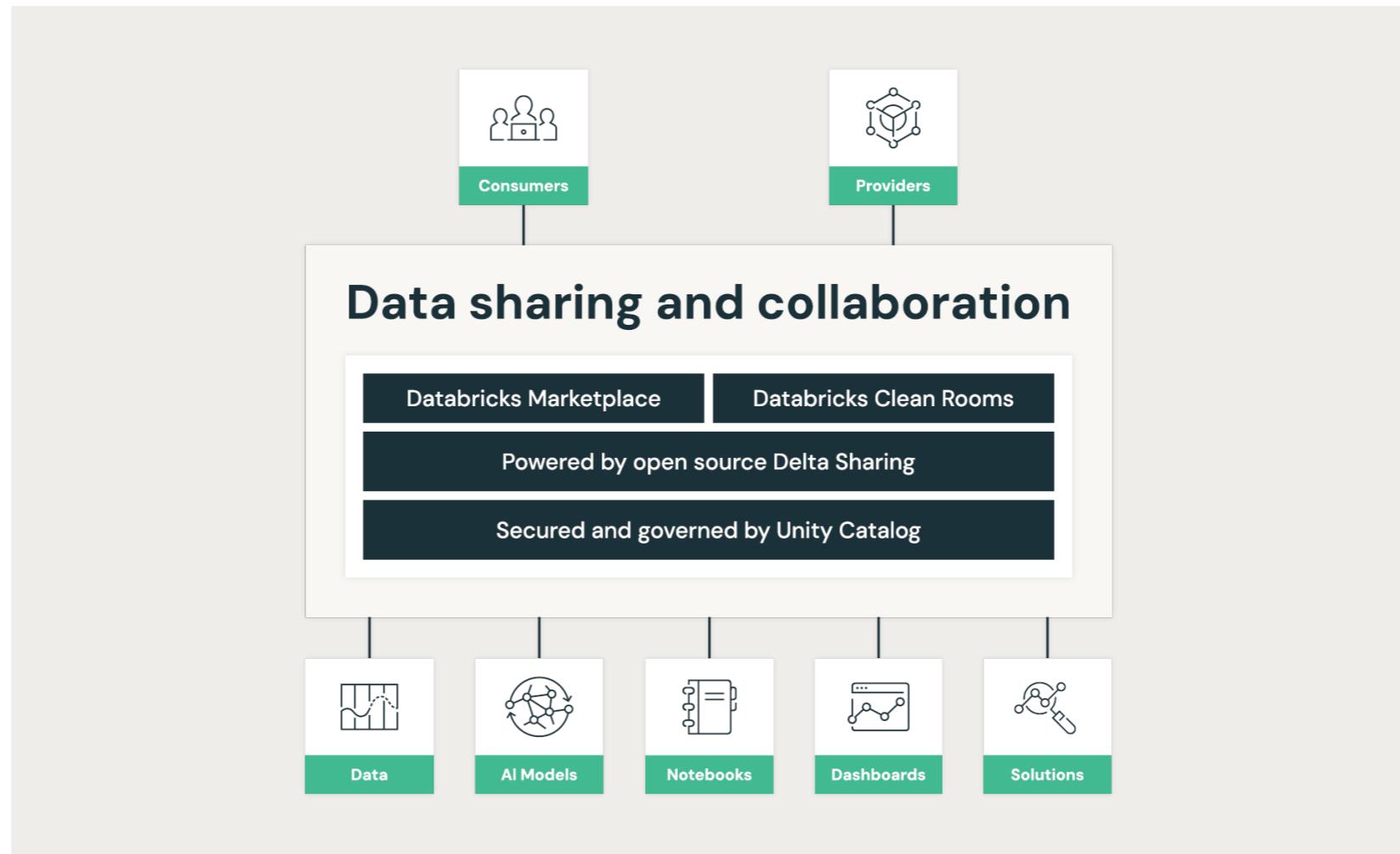
As organizations stand up lakehouse architectures, the supply and demand of cleansed and trusted data doesn't end with analytics and machine learning. Companies need to be able to share and collaborate on data beyond their four walls. Therefore, it is mission-critical that your data strategy aligns with your business strategy by incorporating a secure, flexible and open sharing solution with the broadest ecosystem. However, fragmentation across cloud platforms complicates data sharing, leading to increased costs, storage duplication and privacy risks. Traditional methods hinder AI innovation by limiting the efficient sharing of models and notebooks, and balancing collaboration and privacy continues to be a challenge.

To address these challenges, Databricks offers an open approach to data sharing and collaboration, maximizing reach and impact.

The Databricks Platform is highly interoperable and offers the lowest total cost of ownership (TCO). With zero-copy sharing, you can share a single copy of data across clouds, regions and platforms, eliminating the need for data replication and reducing costs. This approach allows you to use your preferred tools while maintaining full control over storage and compute expenses.

Secondly, Databricks is AI-ready. By enabling the sharing of AI models and notebooks, we unlock a wide range of AI use cases. This seamless sharing accelerates innovation and allows data teams to collaborate effectively across various platforms.

Lastly, Databricks ensures privacy-safe collaboration. Our platform allows you to collaborate with partners privately across clouds, protecting sensitive data without exposing raw information. This privacy-safe environment supports a wide array of use cases, from simple analytics to complex modeling, ensuring your data remains secure.



All this is made possible with the Databricks Data Intelligence Platform, which is built for sharing and collaboration. Databricks Marketplace is the open marketplace for all your data, analytics and AI. Databricks Clean Rooms allow businesses to easily collaborate in a secure environment with their customers and partners on any cloud in a privacy-safe way. And Delta Sharing powers them both. Delta Sharing is the industry's first open protocol for secure data sharing, making it simple to share data with other organizations regardless of which computing platforms they use. And all this is secured and governed by Unity Catalog.

Learn more about [data sharing and collaboration](#)

## The Role of AI in Data Management

Generative AI is pushing companies to become data and AI-driven at their core. To get the most value, they're democratizing data and AI, aiming to integrate intelligence across all operations.

Data intelligence changes data management by using AI to understand enterprise data semantics. Built on the lakehouse — a unified system for querying and managing all data — it analyzes data content, metadata and usage (queries, reports, lineage) to unlock new capabilities.

GenAI takes data intelligence further, enabling deeper understanding and easy interaction with data for all users. With data intelligence, organizations get:

- **Natural language access:** Users interact with data using natural language, customized to organizational jargon.
- **Semantic cataloging and discovery:** AI understands data models and KPIs, enabling better discovery and detecting inconsistencies.
- **Automated management:** Optimizes data layout, partitioning and indexing based on usage.
- **Enhanced governance:** Classifies, detects and prevents misuse of sensitive data while simplifying management.
- **AI workload support:** Connects AI applications to relevant data, leveraging learned semantics for accurate results.

## Bringing AI to the data lakehouse

At Databricks, we're building a **data intelligence platform** on top of our lakehouse. We're excited about the potential of AI in data platforms and continue to enhance our features. Our open data lakehouse is unique, offering (1) unified governance across data and AI, (2) an open, unified storage layer and (3) a unified query engine for ETL, SQL, ML, AI and BI. Mosaic AI powers our Data Intelligence Engine, driving intelligence throughout our platform.

Data intelligence integrates across Databricks, enabling:

- **Data democratization with AI/BI:** Easily creates spaces for business teams to self-serve insights not already answered by their dashboards. AI/BI Genie lets people just converse with their own data, without having to go through others to build dashboards. Genie leverages Unity Catalog so the insights continuously learn your specific business context and semantics while including controls for guidance and security oversight.
- **Enhanced governance:** Data intelligence improves Unity Catalog 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.
- **Platform optimization:** Automatically adjusts settings like column indexing and partition layout, strengthening the lakehouse foundation for better performance and lower TCO.
- **AI assistant:** Enhances Python and SQL code generation 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.

## Building enterprise AI on the foundation of data management

Data intelligence makes building AI solutions easier. Mosaic AI integrates seamlessly to help enterprises create and deploy production-grade ML and AI applications.

- **Production-quality AI:** Delivers accurate outputs tailored to enterprise data, with reinforcement learning by business users. Models can be easily swapped for better accuracy as needed.
- **Unified governance:** Ensures oversight across data and AI assets, managing risk, privacy and accountability from data to models and applications.
- **Cost efficiency at scale:** Mosaic AI optimizes AI deployments, making enterprise-level AI affordable.

Companies are now building adaptable, high-performing and trustworthy AI by leveraging AI systems — integrating multiple models to boost adaptability, customization and transparency.

Databricks, powered by Mosaic AI, is the unified platform for developing and managing AI systems:

1. **Data prep:** Use tools like LakeFlow to ingest and prepare data. With the AI built on the lakehouse, there is no need to duplicate data — instead, you can automatically generate vector indexes and ML features from your production data.
2. **Build agents:** Choose from existing models, train new ones or serve models using Mosaic AI's tools for model training and serving.
3. **Deploy agents:** Deploy models securely at scale using MLflow and Mosaic AI Agent Framework.
4. **Evaluate agents:** Use human and machine evaluations to ensure quality, leveraging Mosaic AI Agent Evaluation and Lakehouse Monitoring.

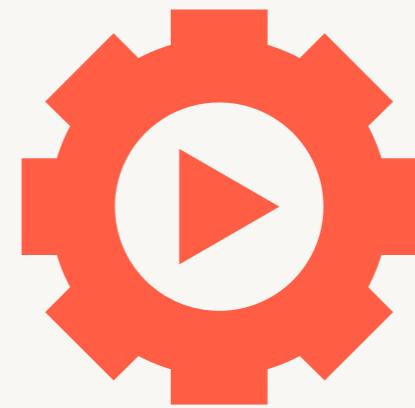
Throughout, Unified Governance keeps data and AI assets secure and compliant, using tools like Unity Catalog and Mosaic AI Gateway for centralized control.

## Conclusion

As we move forward and transition to new ways of working, adopt new technologies and scale operations, investing in effective data management is critical to removing the bottleneck in modernization. With the Databricks Data Intelligence Platform, you can manage your data from ingestion to analytics and truly unify data, analytics and AI.

Learn more about data management on Databricks: [Watch now](#)

Visit our Demo Hub: [Watch demos](#)



## About Databricks

Databricks is the data and AI company. More than 10,000 organizations worldwide — including Block, Comcast, Condé Nast, Rivian, Shell and over 60% 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 [LinkedIn](#), [X](#) and [Facebook](#).

