eBook

Data Management 101, 2nd Edition

Learn how Databricks streamlines the data management lifecycle

databricks

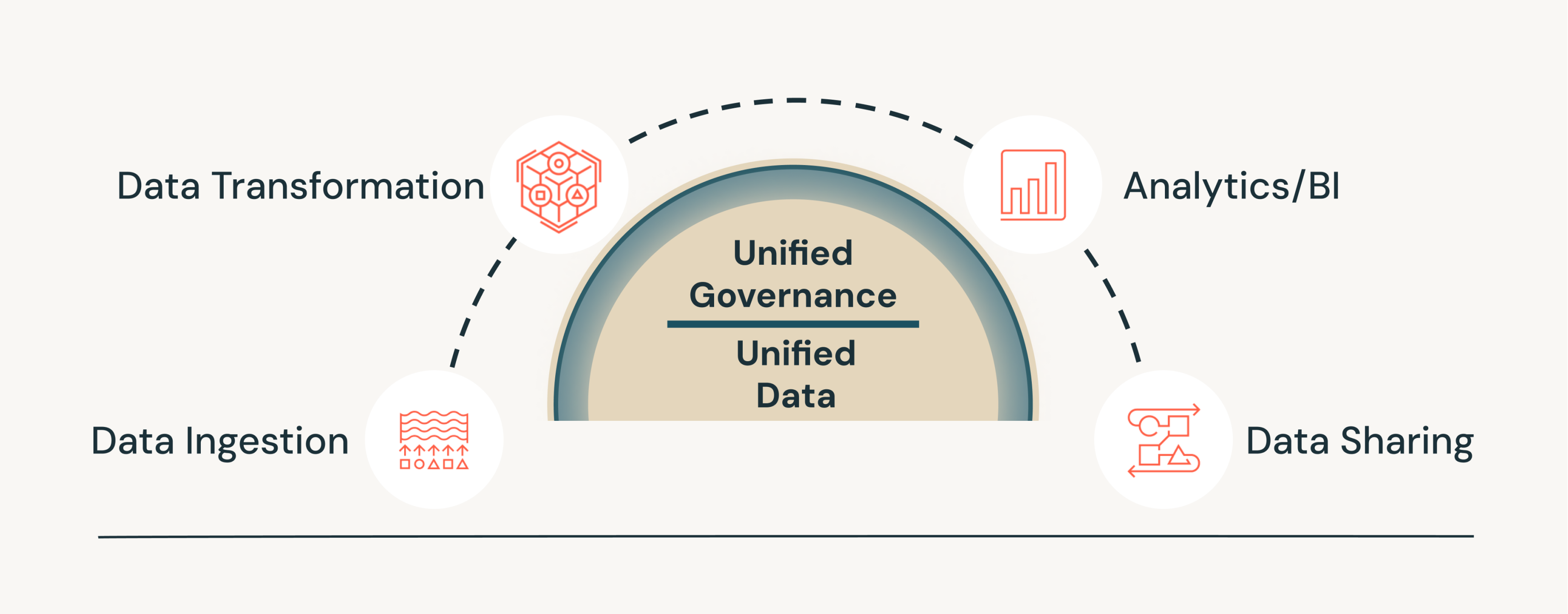
Contents

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

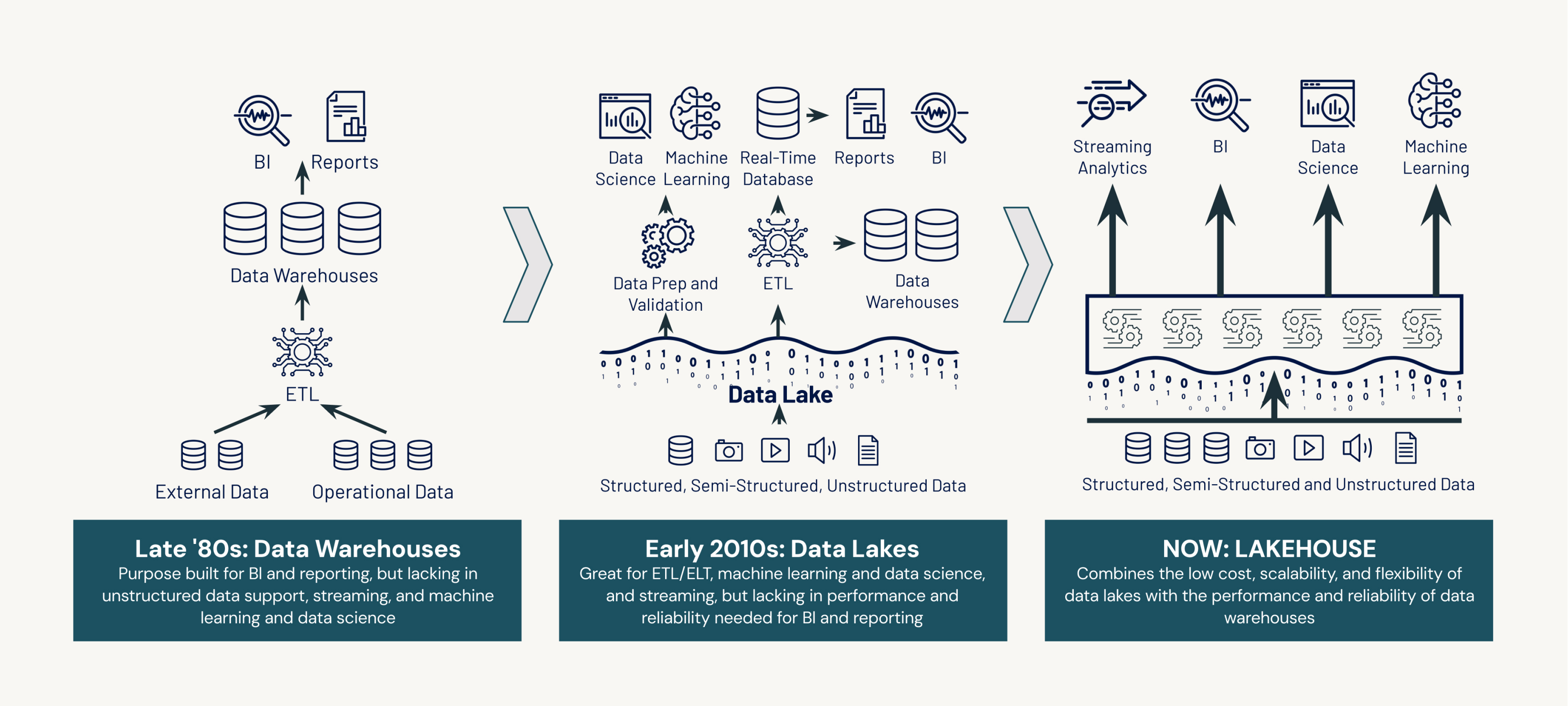
Ultimately, the consistent and reliable flow of data across people, teams and business functions is crucial to an organization’s survival and ability to innovate. Organizations are increasingly recognizing the strategic importance of their data through various applications, including generative AI. This 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.

Evolution to the Data Lakehouse



Late '80s: Data Warehouses
Purpose built for BI and reporting, but lacking in unstructured data support, streaming, and machine learning and data science

Early 2010s: Data Lakes
Great for ETL/ELT, machine learning and data science, and streaming, but lacking in performance and reliability needed for BI and reporting

NOW: LAKEHOUSE
Combines the low cost, scalability, and flexibility of data lakes with the performance and reliability of data warehouses

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.

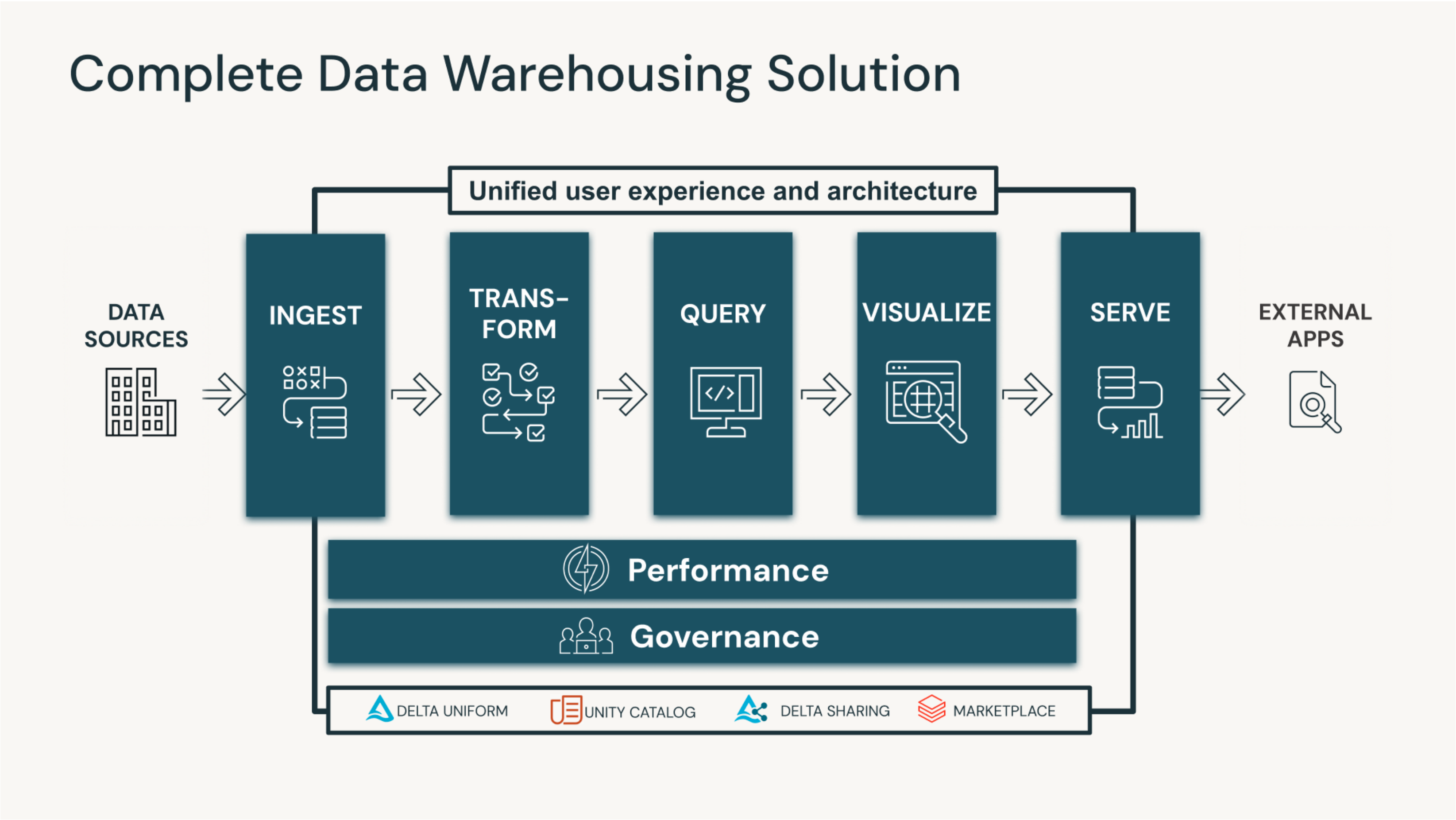
Historically, lakehouses have only been able to support a single open table format, resulting in fragmentation across ecosystems. Organizations have had to choose their platform based on their preferred format, which restricted their choice of compute engine for analysis.

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.


Complete Data Warehousing Solution

Learn more about the Databricks Data Intelligence Platform

Learn more about Unity Catalog

Data and AI governance

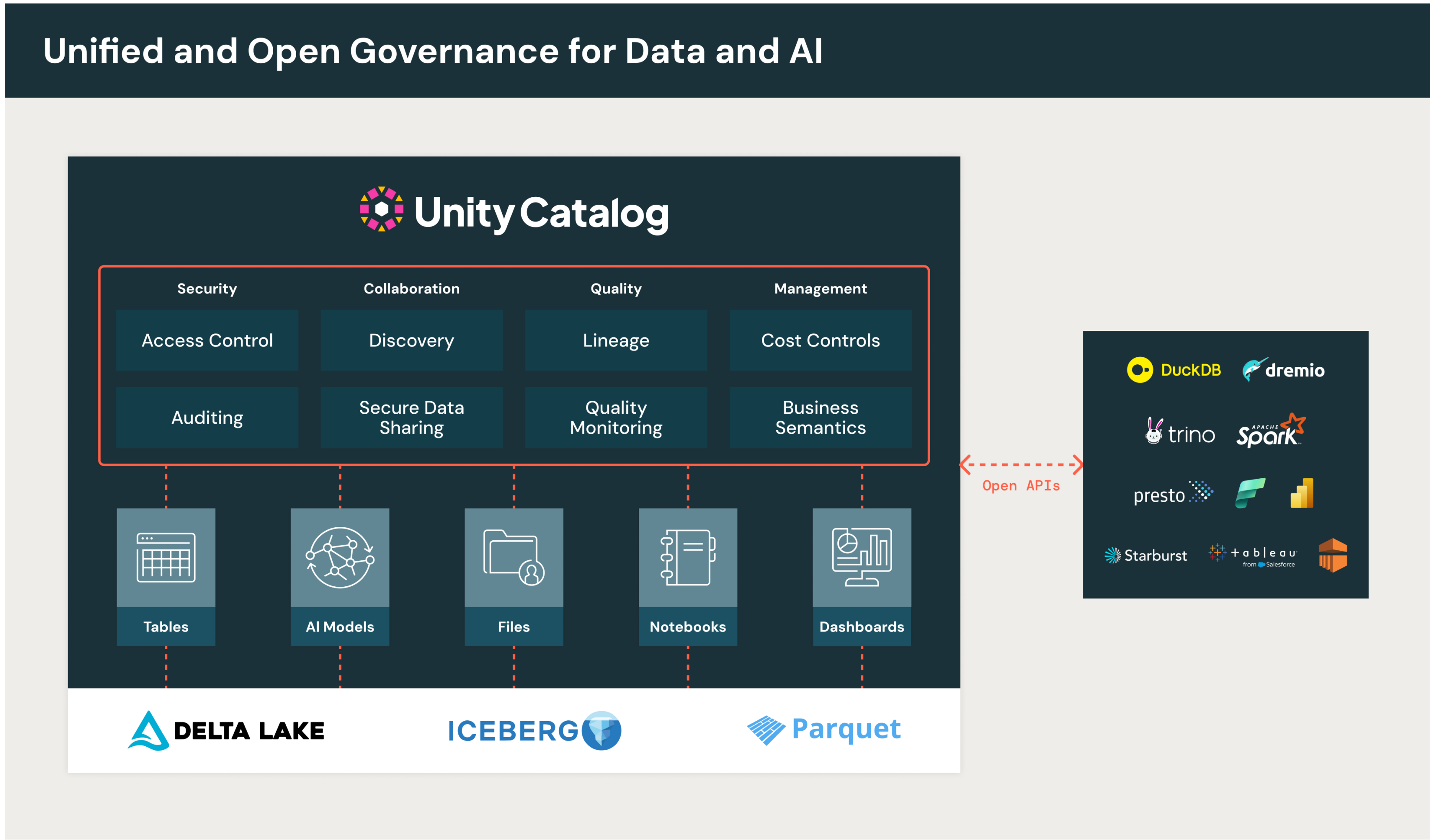
Effective governance is key to making data and AI accessible in the age of generative AI. However, it is also very complex in today's rapidly evolving data and AI landscape. Let's look at why.

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 and Open Governance for Data and AI

Unity Catalog

Security	Collaboration	Quality	Management
Access Control	Discovery	Lineage	Cost Controls
Auditing	Secure Data Sharing	Quality Monitoring	Business Semantics

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

PEPSICO

Learn more about Unity Catalog

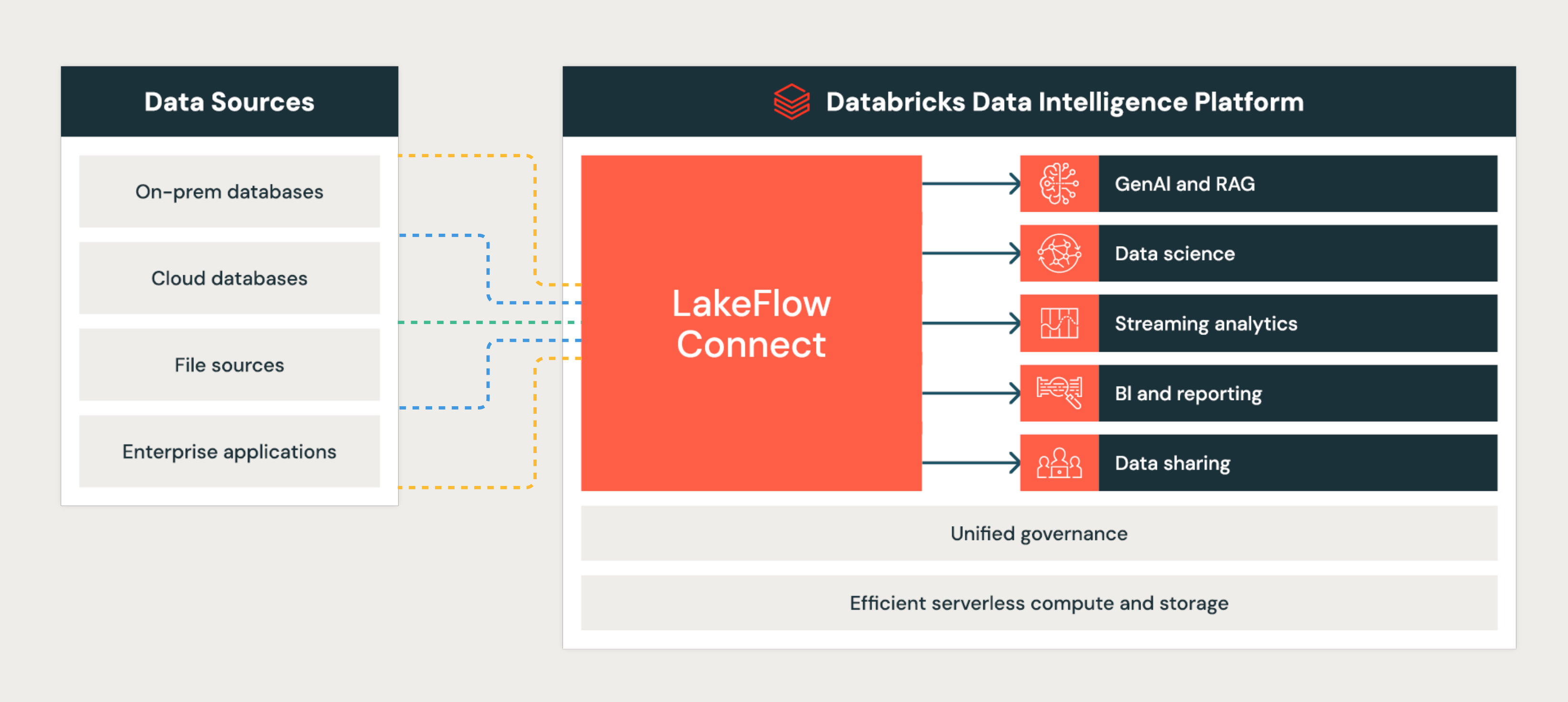
Download free eBook to learn more about data and AI governance on the Databricks Data Intelligence Platform.

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


Data Sources
On-prem databases
Cloud databases
File sources
Enterprise applications
Databricks Data Intelligence Platform
LakeFlow Connect
GenAI and RAG
Data science
Streaming analytics
BI and reporting
Data sharing
Unified governance
Efficient serverless compute and storage