eBook

Data Management 101 on Databricks

Learn how Databricks streamlines the data management lifecycle

databricks

Introduction

Given the changing work environment, with more remote workers and new channels, we are seeing greater importance placed on data management.

> **According to Gartner, "The shift from centralized to distributed working requires organizations to make data, and data management capabilities, available more rapidly and in more places than ever before."**

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 managing data as a strategic and valuable resource, which includes collecting data, processing data, governing data, sharing data, analyzing it — and doing this all in a cost-efficient, effective and reliable manner.

Contents

The challenges of data management

Data Management
Data Sharing
Data Ingestion
Data Governance
Data Transformation and Processing
Data Analytics

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. And while we are seeing companies realize the value of their data — through data-driven product decisions, more collaboration or rapid movement into new channels — most businesses struggle to manage and leverage data correctly.

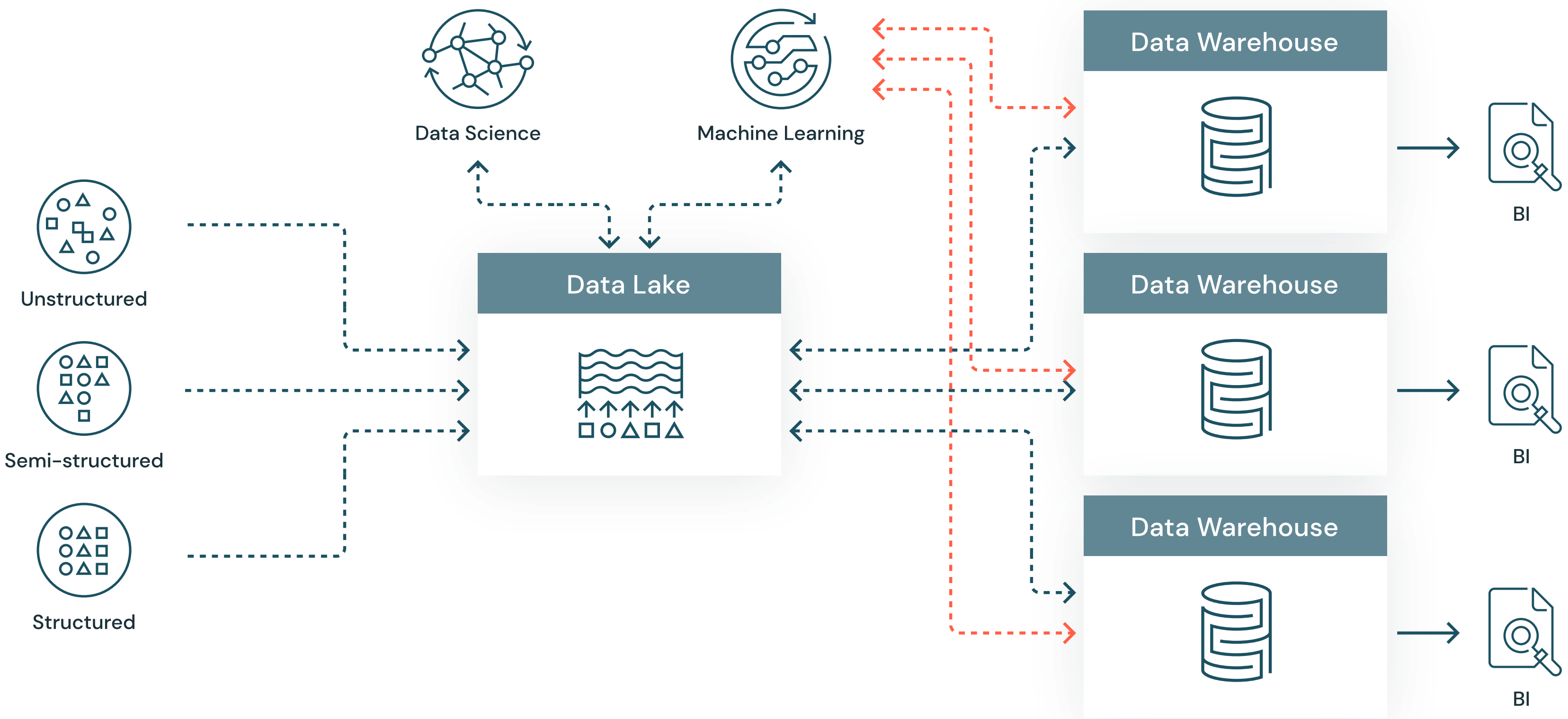
FORRESTER

According to Forrester, up to 73% of company data goes unused for analytics and decision-making, a metric that is costing businesses their success.

The vast majority of company data today flows into a data lake, where teams do data prep and validation in order 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 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.

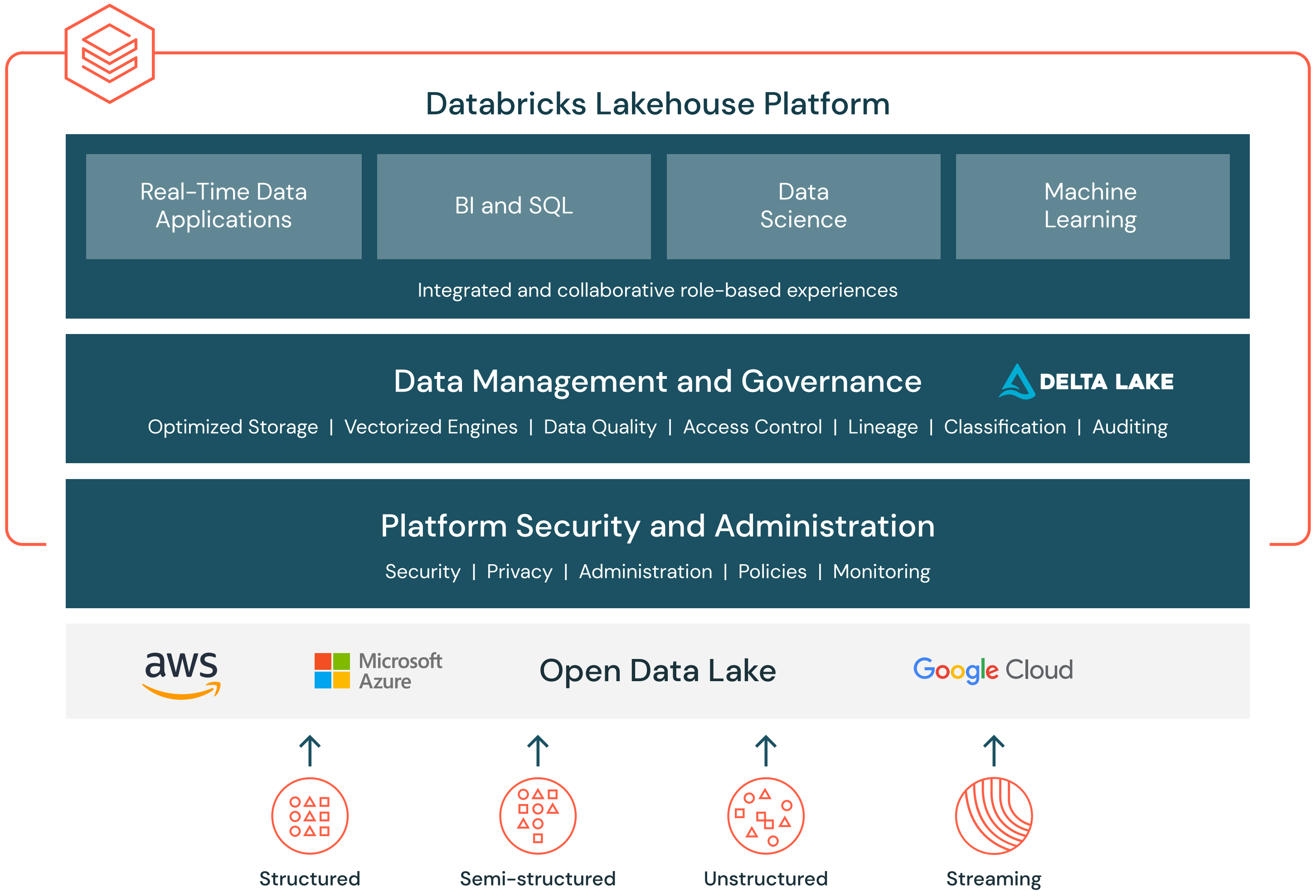


Data management on Databricks

Unifying these systems can be transformational in how we think about data. And the Databricks Lakehouse Platform does just that — unifies all these disparate workloads, teams and data, and provides an end-to-end data management solution for all phases of the data management lifecycle. And with Delta Lake bringing reliability, performance and security to a data lake — and forming the foundation of a lakehouse — data engineers can avoid these architecture challenges. Let's take a look at the phases of data management on Databricks.

Learn more about the Databricks Lakehouse Platform

Learn more about Delta Lake

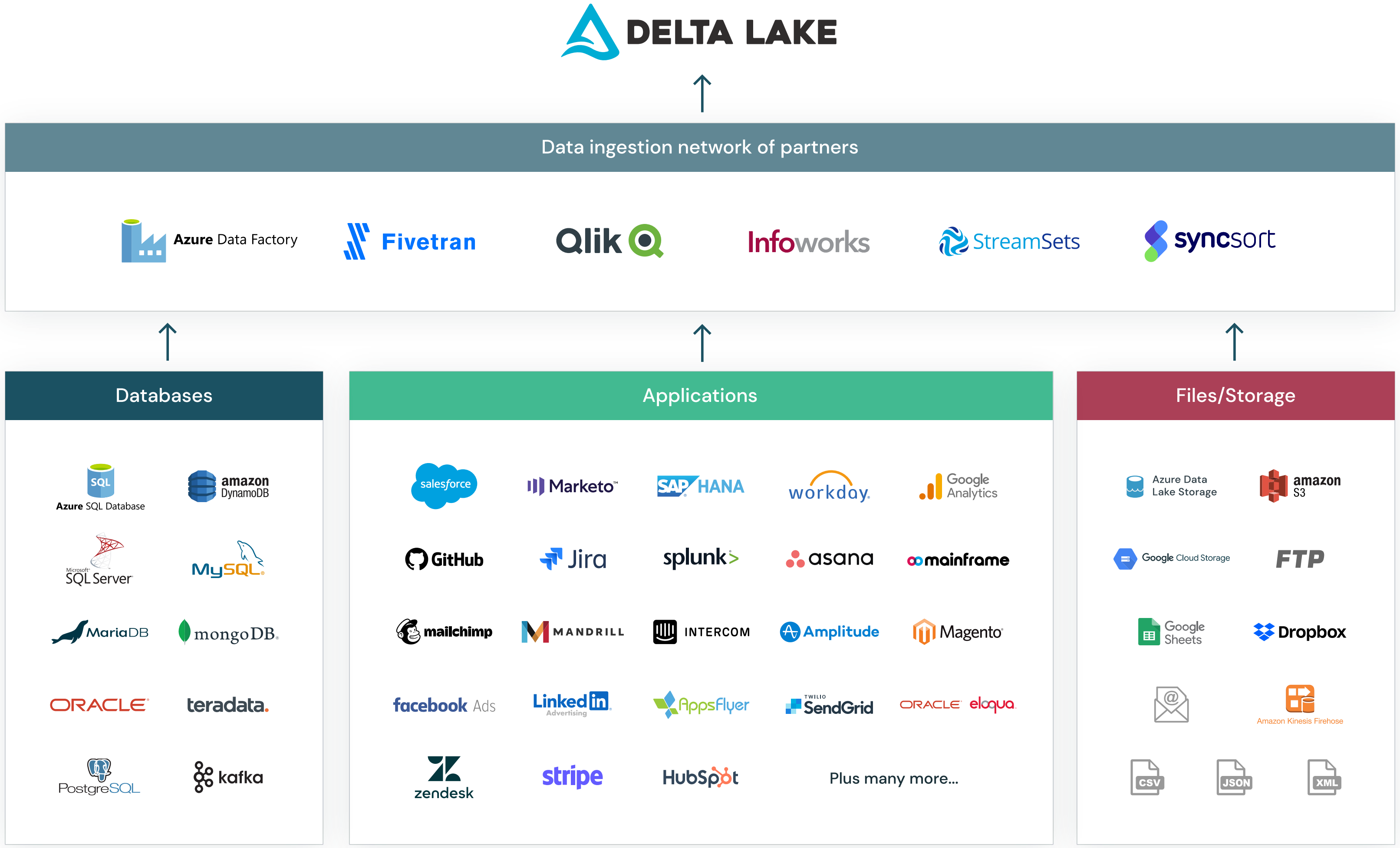

Databricks Lakehouse Platform
Real-Time Data Applications
BI and SQL
Data Science
Machine Learning
Integrated and collaborative role-based experiences
Data Management and Governance
DELTA LAKE
Optimized Storage | Vectorized Engines | Data Quality | Access Control | Lineage | Classification | Auditing
Platform Security and Administration
Security | Privacy | Administration | Policies | Monitoring
aws
Microsoft Azure
Open Data Lake
Google Cloud
Structured
Semi-structured
Unstructured
Streaming

Data ingestion

In today's world, IT organizations are inundated with data siloed across various 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. To support these new use cases and the growing volume and complexity of data, many IT teams are now looking to centralize all their data with a lakehouse architecture built on top of Delta Lake, an open format storage layer.

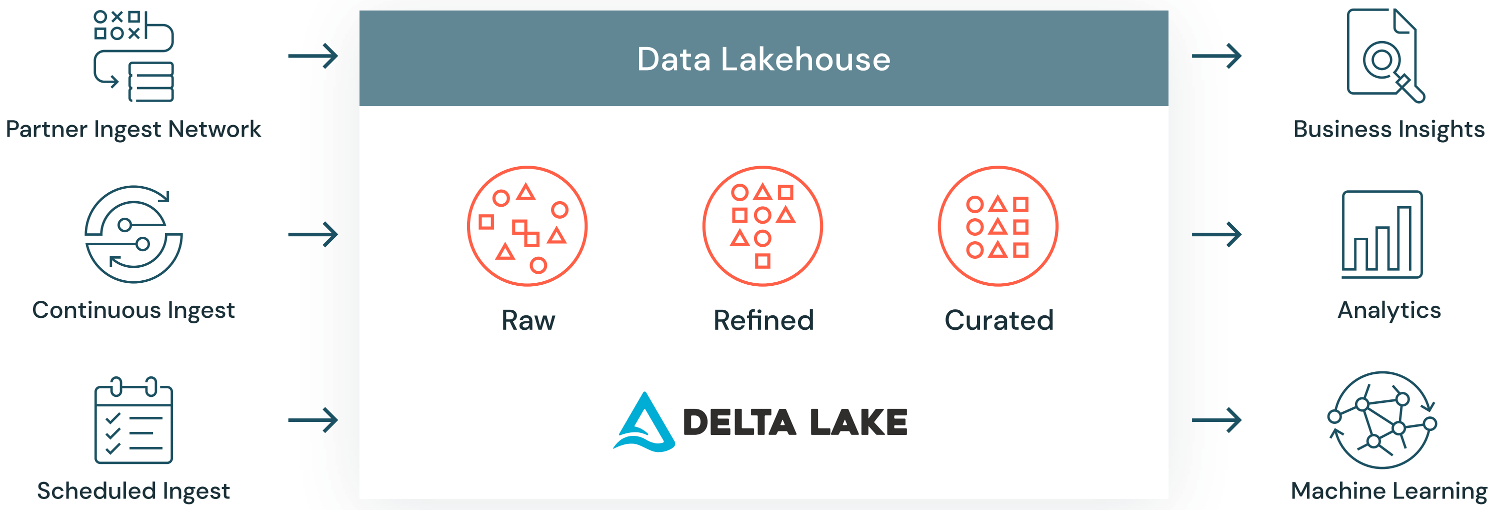
However, the biggest challenge data engineers face in supporting the lakehouse architecture is efficiently moving data from various systems into their lakehouse. Databricks offers two ways to easily ingest data into the lakehouse: through a network of data ingestion partners or by easily ingesting data into Delta Lake with Auto Loader.

The network of data ingestion partners makes it possible to move data from various siloed systems into the lake. The partners have built native integrations with Databricks to ingest and store data in Delta Lake, making data easily accessible for data teams to work with.



On the other hand, many IT organizations have been using cloud storage, such as AWS S3, Microsoft Azure Data Lake Storage or Google Cloud Storage, and have implemented methods to ingest data from various systems. Databricks Auto Loader optimizes file sources, infers schema and incrementally processes new data as it lands in a cloud store with exactly once guarantees, low cost, low latency and minimal DevOps work.

With Auto Loader, data engineers provide a source directory path and start the ingestion job. The new structured streaming source, called "cloudFiles," will automatically set up file notification services that subscribe file events from the input directory and process new files as they arrive, with the option of also processing existing files in that directory.



Getting all the data into the lakehouse is critical to unify machine learning and analytics. With Databricks Auto Loader and our extensive partner integration capabilities, data engineering teams can efficiently move any data type to the data lake.

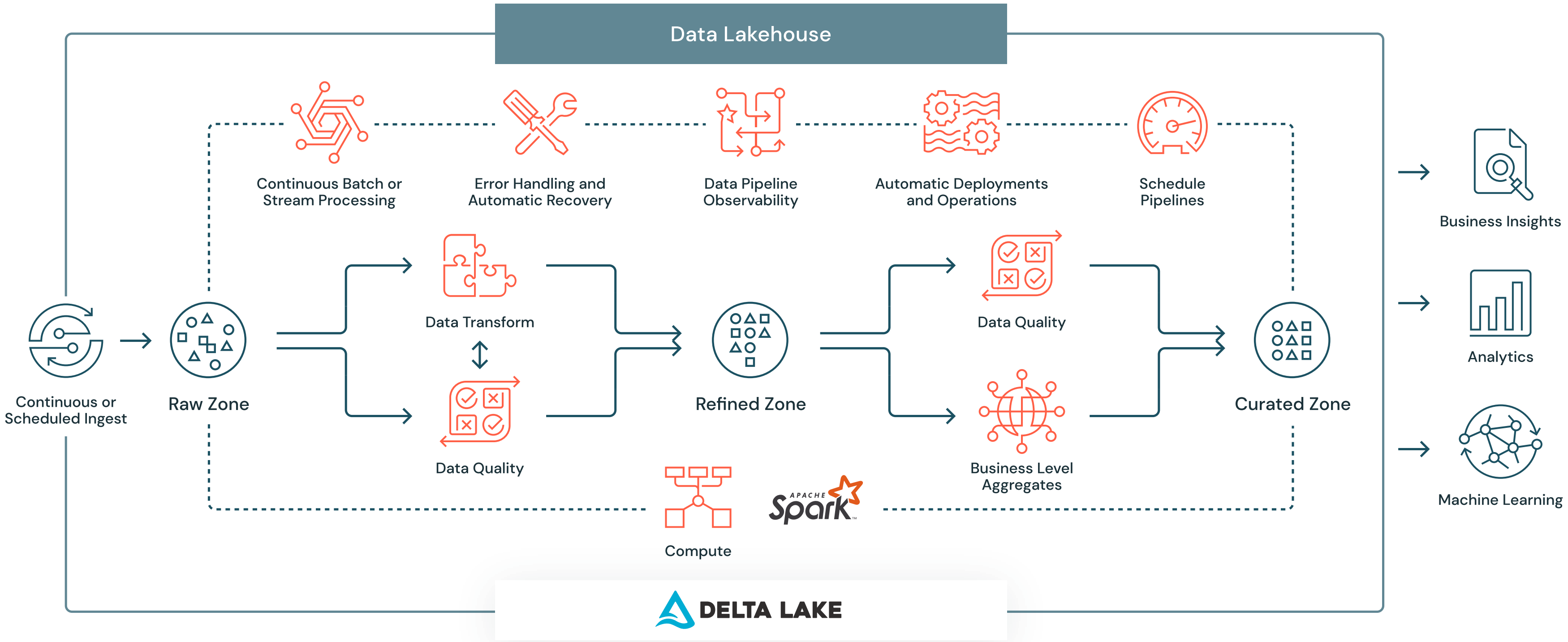
Data
Ingestion

Data ingestion on Databricks

Learn more

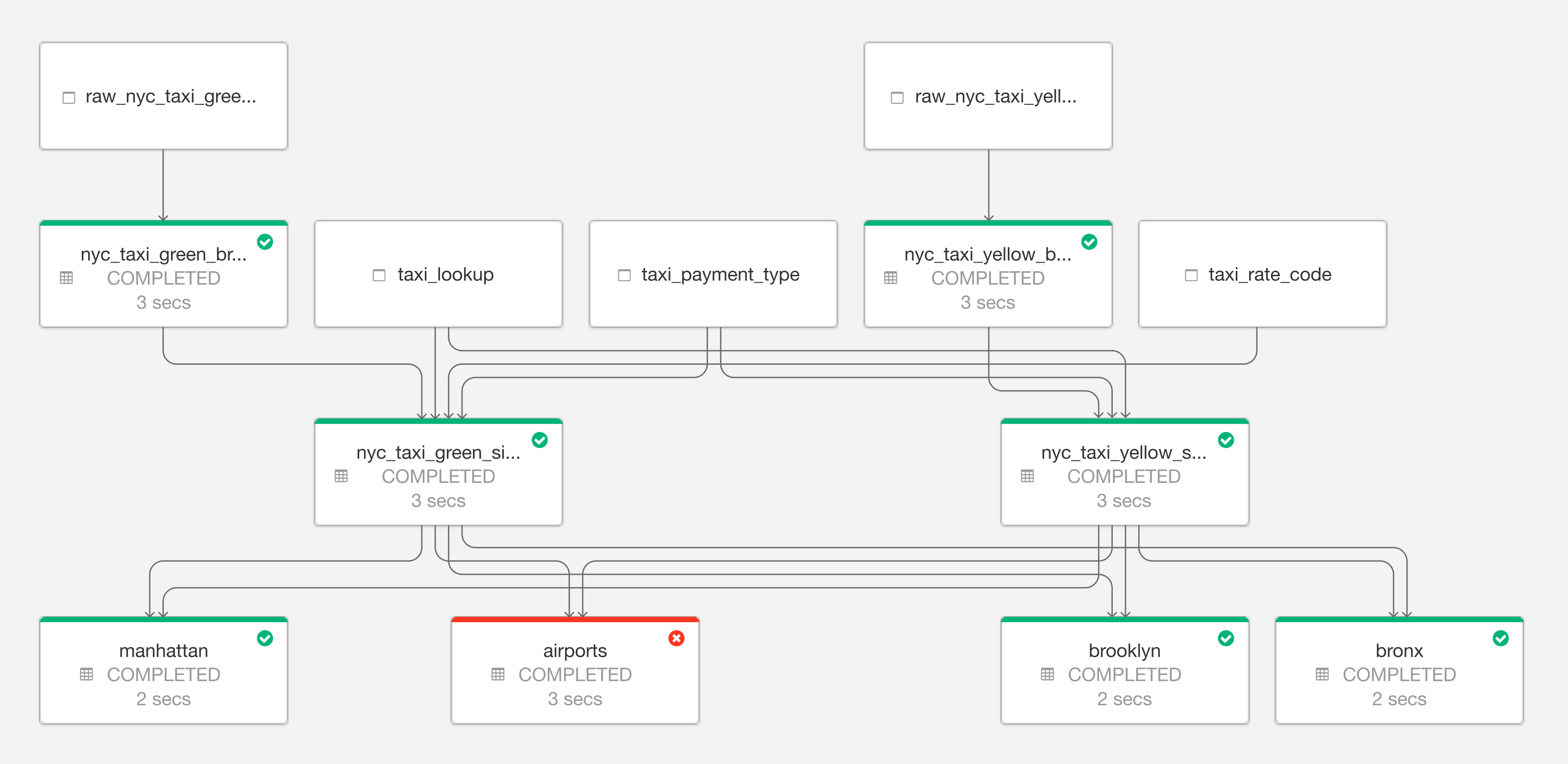
Data transformation, quality and processing

Moving data into the lakehouse solves one of the data management challenges, but in order to make data usable by data analysts or data scientists, data must also be transformed into a clean, reliable source. This is an important step, as outdated or unreliable data can lead to mistakes, inaccuracies or distrust of the insights derived.


Data Lakehouse
Continuous Batch or Stream Processing
Error Handling and Automatic Recovery
Data Pipeline Observability
Automatic Deployments and Operations
Schedule Pipelines
Continuous or Scheduled Ingest
Raw Zone
Data Transform
Data Quality
Refined Zone
Data Quality
Business Level Aggregates
Curated Zone
Compute
Apache Spark
DELTA LAKE
Business Insights
Analytics
Machine Learning

Data engineers have the difficult and laborious task of cleansing complex, diverse data and transforming it into a format fit for analysis, reporting or machine learning. This requires the data engineer to know the ins and outs of the data infrastructure platform, and requires the building of complex queries (transformations) in various languages, stitching together queries for production. For many organizations, this complexity in the data management phase limits their ability for downstream analysis, data science and machine learning.

To help eliminate the complexity, Databricks Delta Live Tables (DLT) gives data engineering teams a massively scalable ETL framework to build declarative data pipelines in SQL or Python. With DLT, data engineers can apply in-line data quality parameters to manage governance and compliance with deep visibility into data pipeline operations on a fully managed and secure lakehouse platform across multiple clouds.



DLT provides a simple way of creating, standardizing and maintaining ETL. DLT data pipelines automatically 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 deliver trusted data, data engineers define rules about the expected quality of data within the data pipeline. DLT enables teams to analyze and monitor data quality continuously to reduce the spread of incorrect and inconsistent data.

> **"Delta Live Tables has helped our teams save time and effort in managing data at scale...With this capability augmenting the existing lakehouse architecture, Databricks is disrupting the ETL and data warehouse markets, which is important for companies like ours."**
>
> — Dan Jeavons, General Manager, Data Science, Shell

A key aspect of successful data engineering implementation is having engineers focus on developing and testing ETL and spending less time on building out infrastructure. Delta Live Tables abstracts the underlying data pipeline definition from the pipeline execution. This means at pipeline execution, DLT optimizes the pipeline, automatically builds the execution graph for the underlying data pipeline queries, manages the infrastructure with dynamic resourcing and provides a visual graph for end-to-end pipeline visibility on overall pipeline health for performance, latency, quality and more.

With all these DLT components in place, data engineers can focus solely on transforming, cleansing and delivering quality data for machine learning and analytics.

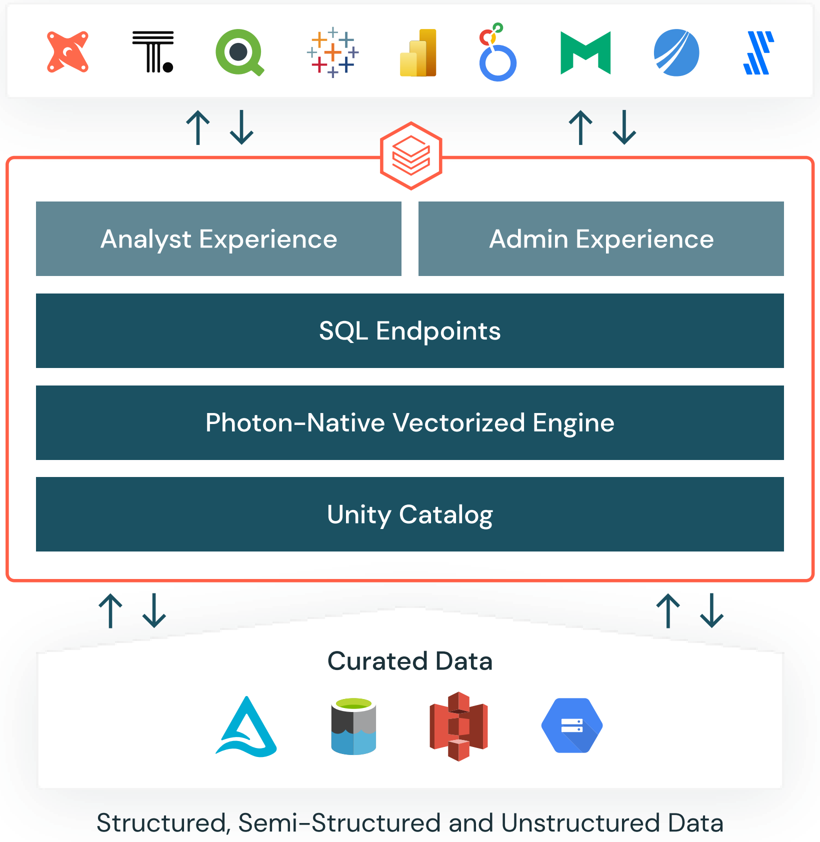
Data Transformation and Processing

Data transformation on Databricks with Delta Live Tables

Learn more

Data analytics

Now that data is available for consumption, data analysts can derive insights to 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. To simplify access and query a lakehouse, Databricks SQL allows 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 data lake data directly within Databricks.


Analyst Experience
Admin Experience
SQL Endpoints
Photon-Native Vectorized Engine
Unity Catalog
Curated Data
Structured, Semi-Structured and Unstructured Data

A dedicated SQL workspace brings familiarity for data analysts to 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. Re-engineered and optimized connectors ensure fast performance, low latency and high user concurrency to your data lake. 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-grain control and visibility over how data is used and accessed across the entire lakehouse for analytics. Administrators have visibility into Databricks SQL usage: the history of all executed queries to understand performance, where each query ran, how long a query ran and which user ran the workload. All this information is captured and made available for administrators to easily triage, troubleshoot and understand performance.

Data
Analytics

Data analytics on Databricks with Databricks SQL

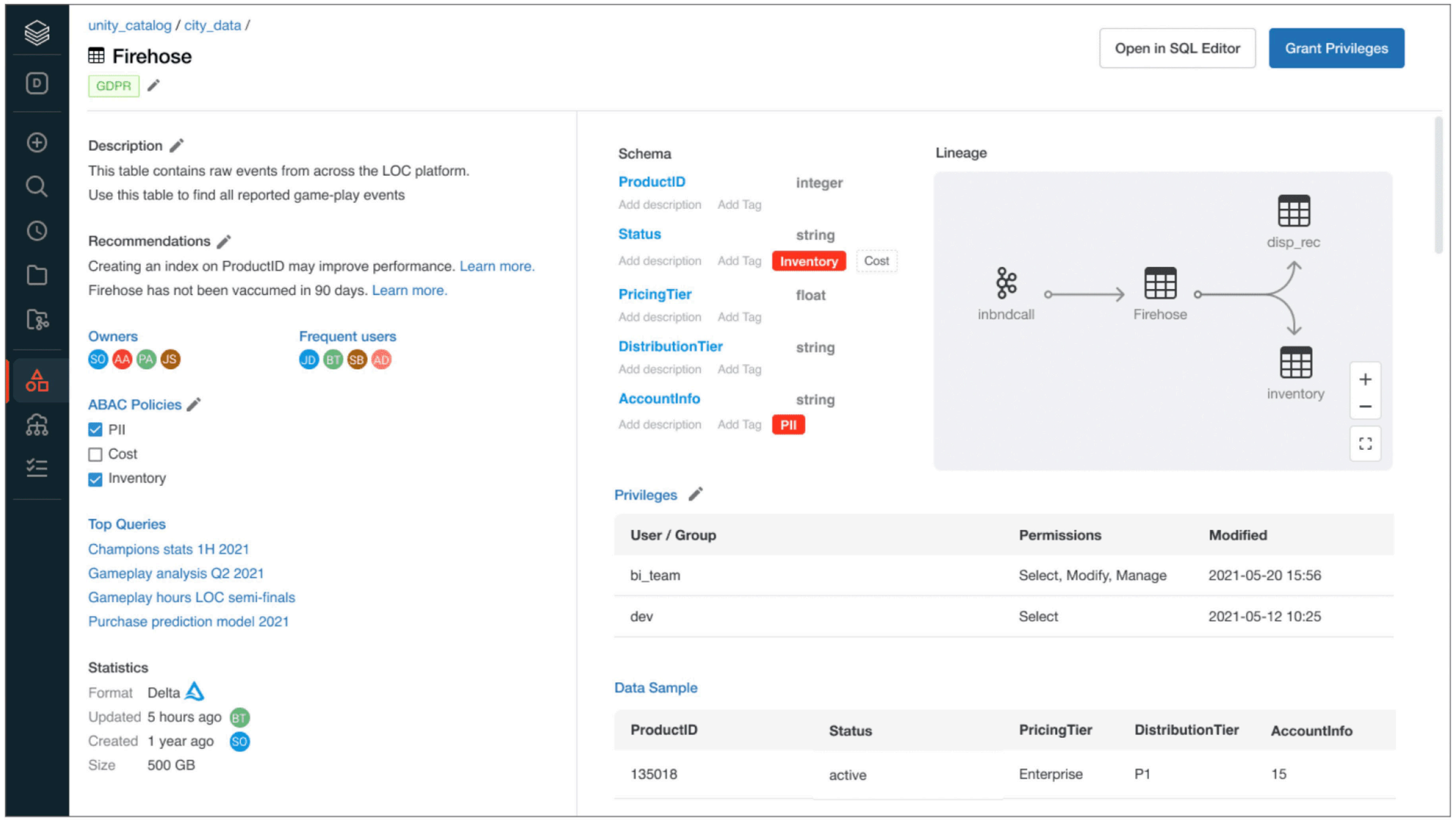
Learn more

Data governance

Many organizations start building out data lakes as a means to solve for analytics and machine learning, making data governance an afterthought. But with the rapid adoption of lakehouse architectures, data is being democratized and accessed throughout the organization. To govern data lakes, administrators have relied on cloud-vendor-specific security controls, such as IAM roles or RBAC and file-oriented access control to manage data. However, this technical security mechanism does not meet the requirements for data governance and of data teams. Data governance defines who within an organization has authority and control over data assets and how those assets may be used.

To more effectively govern data, the Databricks Unity Catalog brings fine-grain governance and security to the lakehouse using standard ANSI SQL or a simple UI, enabling data stewards to safely open their lakehouse for broad internal consumption. With the SQL-based interface, data stewards will be able to apply attribute-based access controls to tag and apply policies to similar data objects with the same attribute. Additionally, data stewards can apply strong governance to other data assets like ML models, dashboards and external data sources all within the same interface.

As organizations modernize their data platforms from on-premises to cloud, many are moving beyond a single-cloud environment for governing data. Instead, they're choosing a multicloud strategy, often working with the three leading cloud providers — AWS, Azure and GCP — across geographic regions. Managing all this data across multiple cloud platforms, storage and other catalogs can be a challenge for democratizing data throughout an organization. The Unity Catalog will enable a secure single point of control to centrally manage, track and audit data trails.



unity_catalog / city_data /

Firehose

GDPR

Description

This table contains raw events from across the LOC platform.
Use this table to find all reported game-play events

Recommendations

Creating an index on ProductID may improve performance. Learn more.
Firehose has not been vaccumed in 90 days. Learn more.

Owners

SO AA PA JS

Frequent users

JD BT SB AD

ABAC Policies

☑ PII
☐ Cost
☑ Inventory

Top Queries

Champions stats 1H 2021
Gameplay analysis Q2 2021
Gameplay hours LOC semi-finals
Purchase prediction model 2021

Statistics

Format	Delta
Updated	5 hours ago BT
Created	1 year ago SO
Size	500 GB

Schema

ProductID	integer	
Add description Add Tag		
Status	string	
Add description Add Tag		Inventory Cost
PricingTier	float	
Add description Add Tag		
DistributionTier	string	
Add description Add Tag		
AccountInfo	string	
Add description Add Tag		PII

Lineage

Privileges

User / Group	Permissions	Modified
bi_team	Select, Modify, Manage	2021-05-20 15:56
dev	Select	2021-05-12 10:25

Data Sample

ProductID	Status	PricingTier	DistributionTier	AccountInfo
135018	active	Enterprise	P1	15

Finally, Unity Catalog will make it easy to discover, describe, audit and govern data assets from one central location. Data stewards can set or review all permissions visually, and the catalog captures audit and lineage information that shows you how each data asset was produced and accessed. Data lineage, role-based security policies, table or column level tags, and central auditing capabilities will make it easy for data stewards to confidently manage and secure data access to meet compliance and privacy needs, directly on the lakehouse. The UI is designed for collaboration so that data users will be able to document each asset and see who uses it.

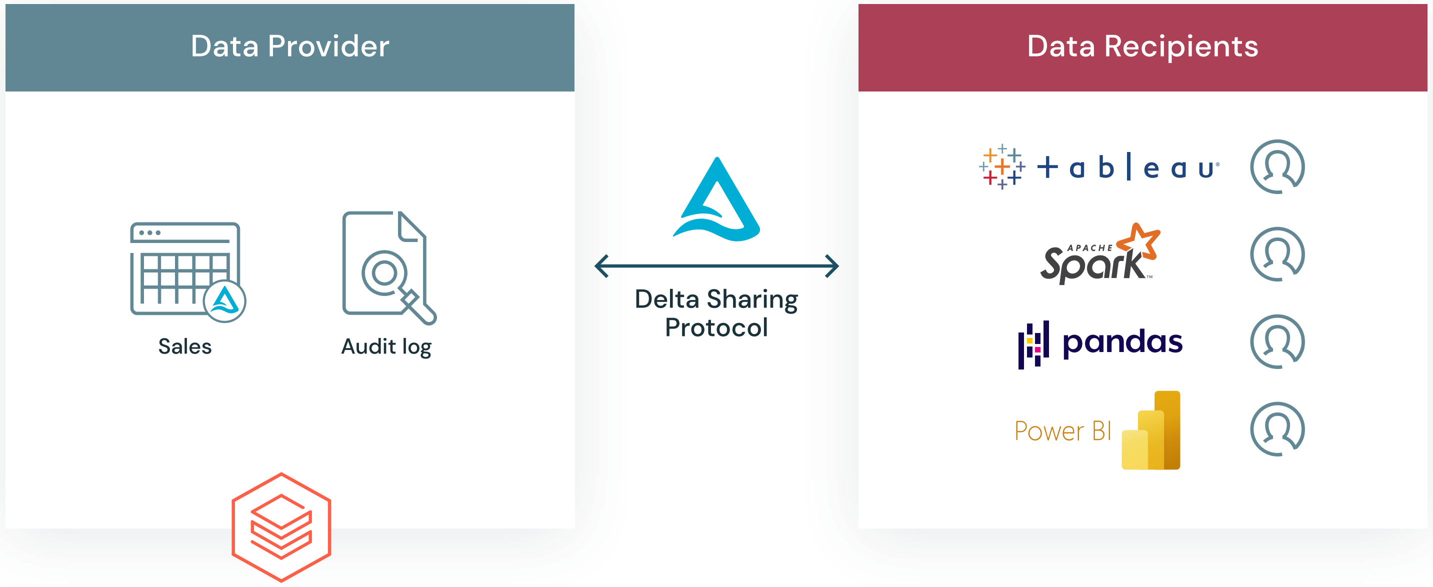
Data
Governance

Data governance on Databricks with Unity Catalog

Learn more

Data sharing

As organizations stand up lakehouse architectures, the supply and demand of cleansed and trusted data doesn't end with analytics and machine learning. As many IT leaders realize in today's data-driven economy, sharing data across organizations — with customers, partners and suppliers — is a key determinant of success in gaining more meaningful insights. However, many organizations fail at data sharing due to a lack of standards, collaboration difficulties when working with large data sets across a large ecosystem of systems or tools, and mitigating risk while sharing data. To address these challenges, Delta Sharing, an open protocol for secure real-time data sharing, simplifies cross-organizational data sharing.



Integrated with the Databricks Lakehouse Platform, Delta Sharing will allow providers to easily use their existing data or workflows to securely share live data in Delta Lake or Apache Parquet format — without copying it to any other servers or cloud object stores. With Delta Sharing's open protocol, data consumers will be able to easily access shared data directly by using open source clients (such as pandas) or commercial BI, analytics or governance clients — data consumers don't need to be on the same platform as providers. The protocol is designed with privacy and compliance requirements in mind. Delta Sharing will give administrators security and privacy controls for granting access to and for tracking and auditing shared data from a single point of enforcement.

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. Delta Sharing will be able to seamlessly share existing large-scale data sets based on the Apache Parquet and Delta Lake formats, and will be supported in the Delta Lake open source project so that existing engines that support Delta Lake can easily implement it.

Data
Sharing

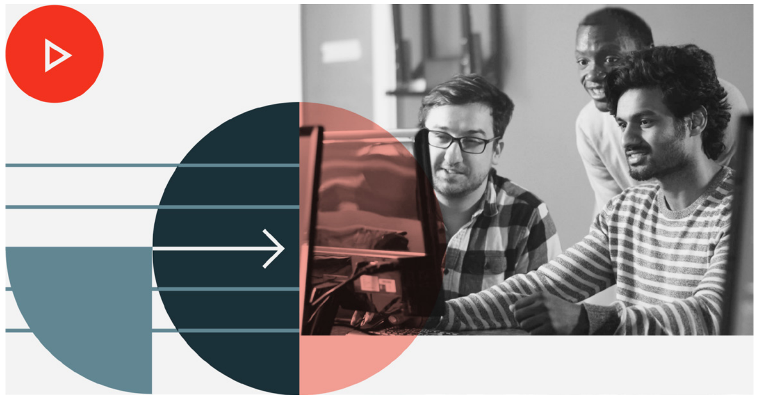
Sharing data on Databricks
with Delta Sharing

Learn more

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 Lakehouse 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 5,000 organizations worldwide — including Comcast, Condé Nast, H&M and over 40% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark™, Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.

databricks