databricks

Big Book of Data Engineering

3RD EDITION

Contents

01 Introduction to Data Engineering on Databricks

Introduction to Data Engineering on Databricks

A recent MIT Tech Review Report shows that 88% of surveyed organizations are either investing in, adopting or experimenting with generative AI (GenAI) and 71% intend to build their own GenAI models. This increased interest in AI is fueling major investments as AI becomes a differentiating competitive advantage in every industry. As more organizations work to leverage their proprietary data for this purpose, many encounter the same hard truth:

The best GenAI models in the world will not succeed without good data.

This reality emphasizes the importance of building reliable data pipelines that can ingest or stream vast amounts of data efficiently and ensure high data quality. A unified platform and good data engineering are essential components of success in every data and AI initiative, especially in the era of GenAI.

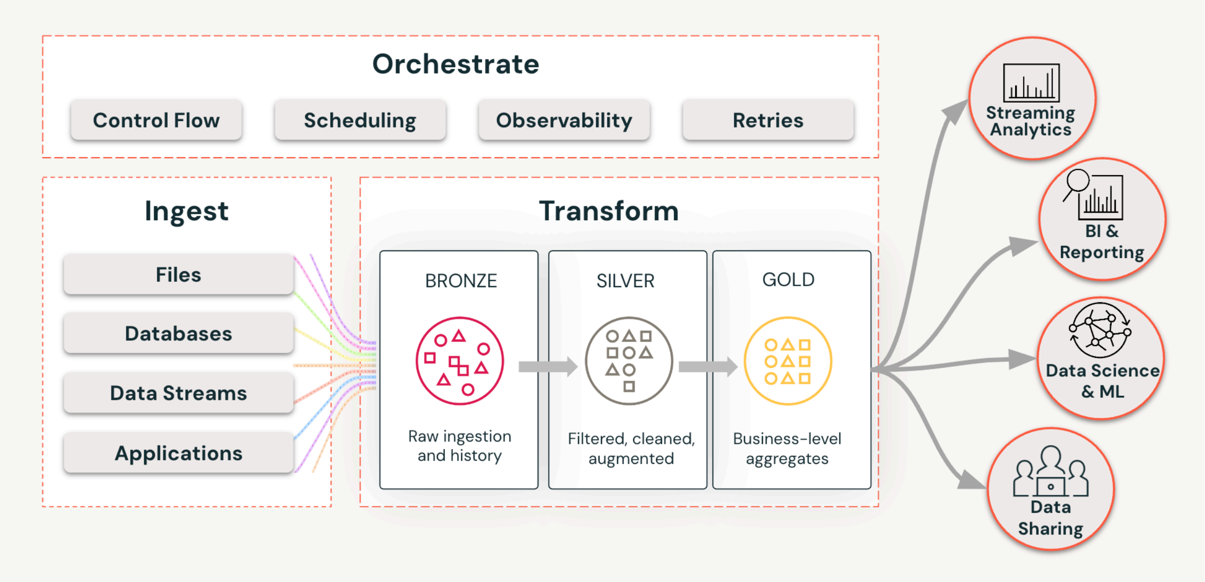
Using practical guidance, useful patterns, best practices and real-world examples, this book will provide you with an understanding of how the Databricks Data Intelligence Platform helps data engineers meet the challenges of this new era.

What is data engineering?

Data engineering is the practice of taking raw data from a data source and processing it so it's stored and organized for a downstream use case such as data analytics, business intelligence (BI) or machine learning (ML) model training. In other words, it's the process of preparing data so value can be extracted from it.

A useful way of thinking about data engineering is by using the following framework, which includes three main parts:

1. **Ingest**
   Data ingestion is the process of bringing data from one or more data sources into a data platform. These data sources can be files stored on-premises or on cloud storage services, databases, applications and, increasingly, data streams that produce real-time events.
2. **Transform**
   Data transformation takes raw ingested data and uses a series of steps (referred to as "transformations") to filter, standardize, clean and finally aggregate it so it's stored in a usable way. A popular pattern is the medallion architecture, which defines three stages in the process — Bronze, Silver and Gold.
3. **Orchestrate**
   Data orchestration refers to the way a data pipeline that performs ingestion and transformation is scheduled and monitored as well as the control of the various pipeline steps and handling failures (e.g., by executing a retry run).



Challenges of data engineering in the AI era

As previously mentioned, data engineering is key to ensuring reliable data for AI initiatives. Data engineers who build and maintain ETL pipelines and the data infrastructure that underpins analytics and AI workloads face specific challenges in this fast-moving landscape.

- **Disparate data sources challenge most organizations:** ISG predicts that by 2026, 8 in 10 enterprises will have their data spread across multiple cloud providers and on-premises data centers that span multiple locations. This decentralization creates a dependency on specialized, siloed teams, inefficient pipelines and development with high costs, and slow time to value, which limits the usage of data and blocks innovation.
- **Handling real-time data:** From mobile applications to sensor data on factory floors, more and more data is created and streamed in real time and requires low-latency processing so it can be used in real-time decision-making.
- **Scaling data pipelines reliably:** With data coming in large quantities and often in real time, scaling the compute infrastructure that runs data pipelines is challenging, especially when trying to keep costs low and performance high. Running data pipelines reliably, monitoring data pipelines and troubleshooting when failures occur are some of the most important responsibilities of data engineers.
- **Data quality:** “Garbage in, garbage out.” High data quality is essential to training high-quality models and gaining actionable insights from data. Ensuring data quality is a key challenge for data engineers.
- **Governance and security:** Data governance is becoming a key challenge for organizations that find their data spread across multiple systems, with increasingly larger numbers of internal teams looking to access and utilize it for different purposes. Securing and governing data is also an important regulatory concern many organizations face, especially in highly regulated industries.

These challenges stress the importance of choosing the right data platform for navigating new waters in the age of AI. But a data platform in this new age can also go beyond addressing just the challenges of building AI solutions. The right platform can improve the experience and productivity of data practitioners, including data engineers, by infusing intelligence and using AI to assist with daily engineering tasks.

In other words, the new data platform is a data *intelligence* platform.

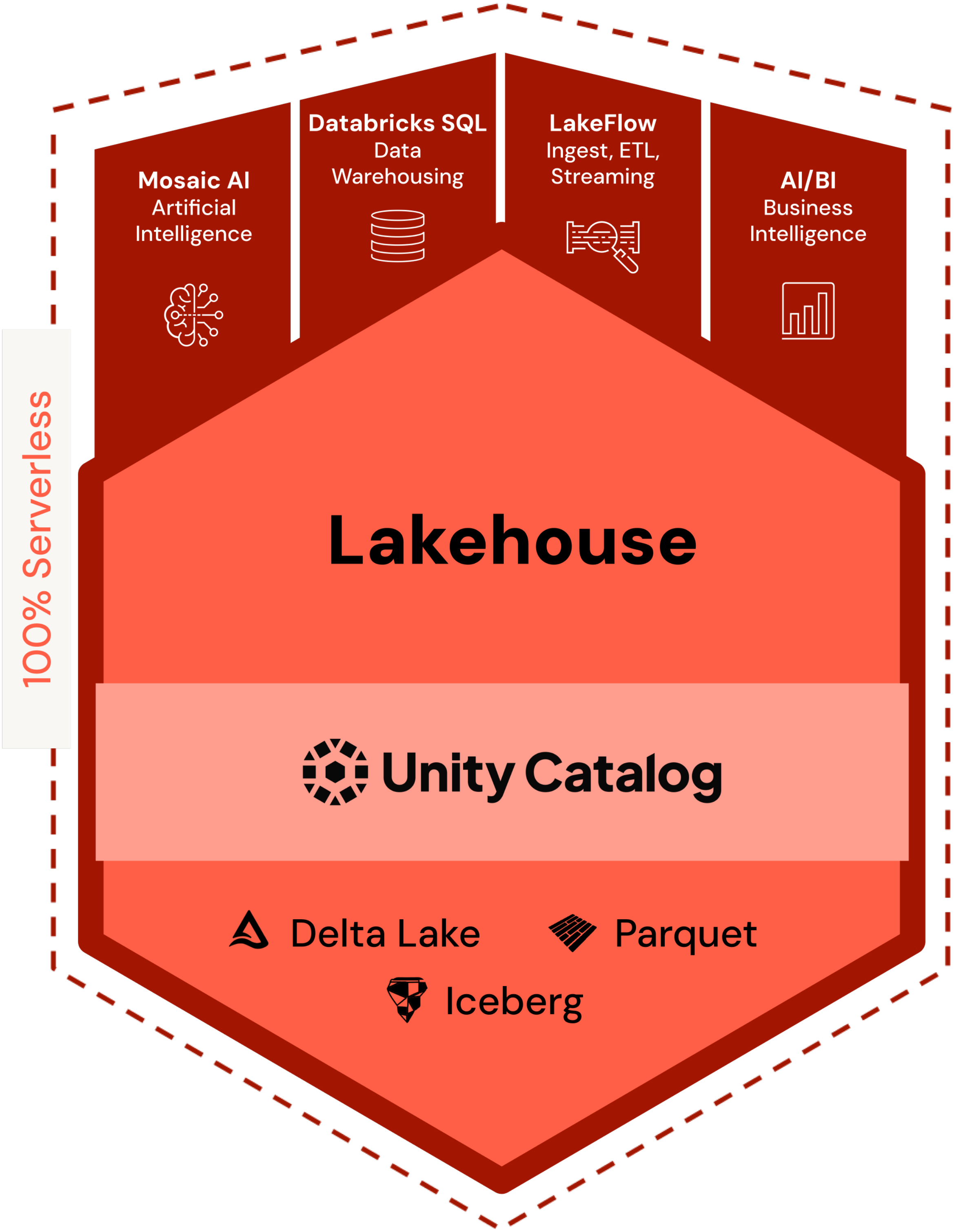
The Databaricks Data Intelligence Platform

Databricks' mission is to democratize data and AI, allowing organizations to use their unique data to build or fine-tune their own machine learning and generative AI models so they can produce new insights that lead to business innovation.

The Databricks Data Intelligence Platform is built on a lakehouse architecture 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. With these capabilities at its foundation, the Data Intelligence Platform lets Databricks customers run a variety of workloads, from business intelligence and data warehousing to AI and data science.

To get a better understanding of the Databricks Platform, here's an overview of the different parts of the architecture as it relates to data engineering.

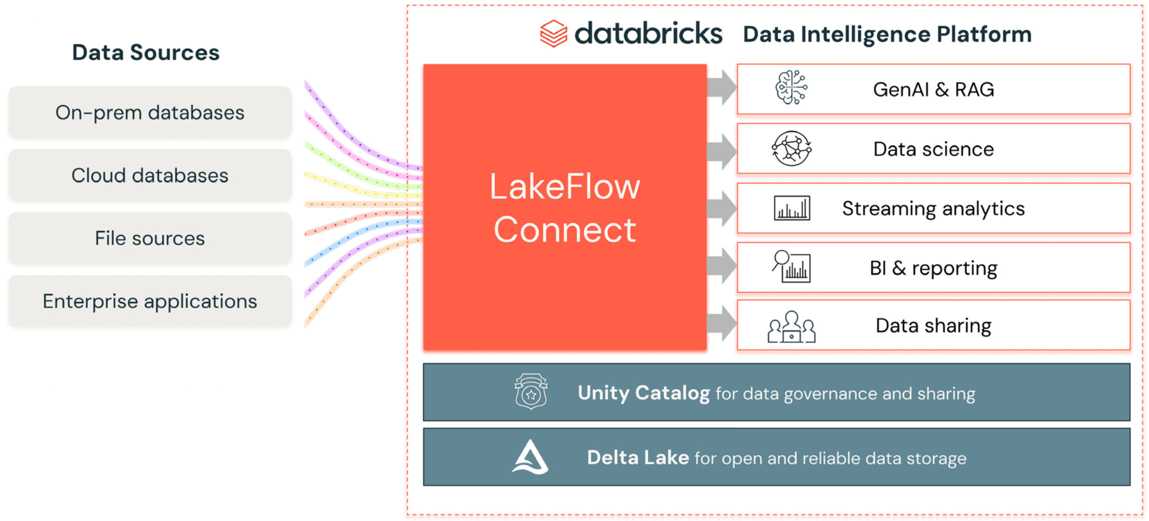
The Databricks Data Intelligence Platform enables you to execute all your data and AI initiatives. As a 100% serverless platform, it provides you with built-in features such as disaster recovery, cost controls and enterprise security. Key components feature Mosaic AI with end-to-end AI for both generative and classical AI; Databricks SQL, the most performant data warehouse in the cloud; efficient data ingestion and reliable transformation tools such as Workflows and Delta Live Tables (DLT) to ensure you can manage all your data for any workload; and AI/BI that integrates deeply with Databricks SQL to easily extend business intelligence across your business.



Data ingestion with Databricks LakeFlow

Databricks enables organizations to efficiently move data from various systems into a single, open and unified lakehouse architecture. Databricks LakeFlow Connect provides native data ingestion connectors for popular SaaS applications, databases and file sources that any practitioner can use to 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 Data Intelligence Platform. LakeFlow Connect is part of LakeFlow — Databricks' new unified data engineering solution spanning ingestion, transformation and orchestration — and is the first of these three components to roll out, compatible with existing tooling. In addition to LakeFlow Connect, Databricks Auto Loader, a connector for cloud object storage, is compatible with Structured Streaming and Delta Live Tables.

Native ingestion for the Data Intelligence Platform



Data reliability and performance with Delta Lake

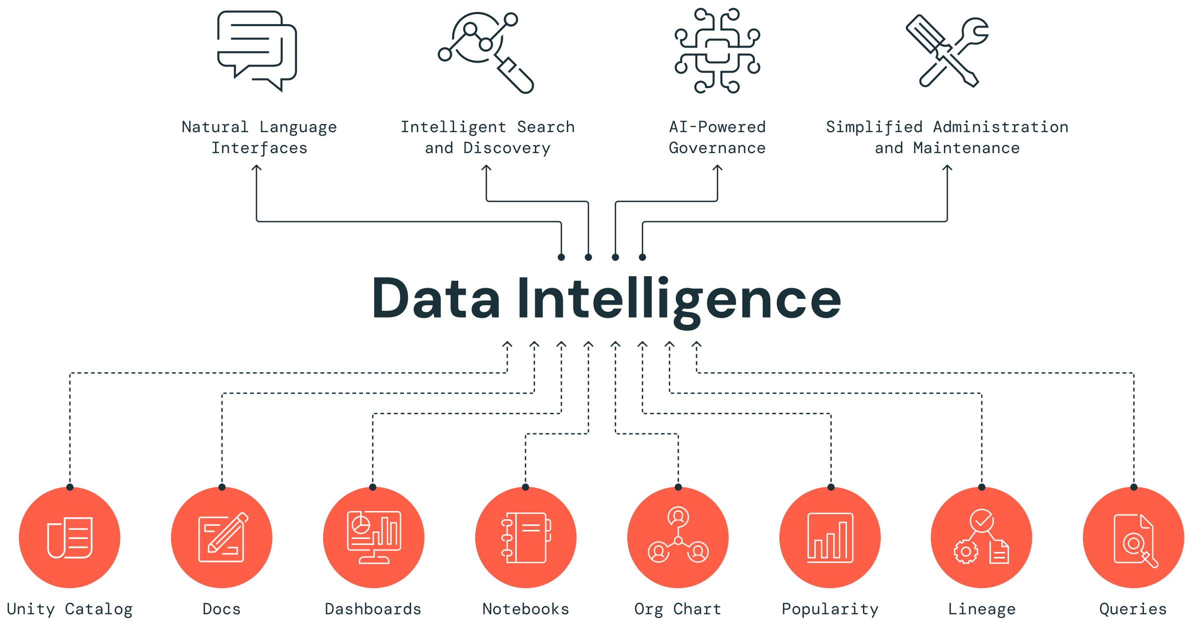
To bring openness, reliability and lifecycle management to data lakes, the Databricks lakehouse architecture is built on the foundation of Delta Lake, an open source, highly performant storage format that solves challenges around unstructured/structured data ingestion, the application of data quality, difficulties with deleting data for compliance or issues with modifying data for data capture. Delta Lake UniForm users can now read Delta tables with Hudi and Apache Iceberg™ clients, keeping them in control of their data. In addition, Delta Sharing enables easy and secure sharing of datasets inside and outside the organization.

Unified governance with Unity Catalog

With Unity Catalog, data engineering and governance teams benefit from an enterprisewide data catalog with a single interface to manage permissions, centralize auditing, automatically track data lineage down to the column level and share data across platforms, clouds and regions.

Accelerate productivity through Data Intelligence

Databricks brings AI to your data to help you bring AI to the world, and at the heart of the Databricks Platform is Data Intelligence. Databricks helps you succeed with AI with your own data to democratize insights and drive down costs. Databricks leverages generative AI with Data Intelligence to power all parts of the platform. Using signals across your entire Databricks environment, including Unity Catalog, dashboards, notebooks, data pipelines and documentation, the Data Intelligence Engine creates highly specialized and accurate generative AI models that understand your data, your usage patterns and your business terminology.



Reliable data pipelines and real-time stream processing with Delta Live Tables

Delta Live Tables (DLT) is a declarative ETL framework that helps data teams simplify and make ETL cost-effective in streaming and batch. Simply define the transformations you want to perform on your data and let DLT pipelines automatically handle task orchestration, cluster management, monitoring, data quality and error management. Engineers can treat their data as code and apply modern software engineering best practices like testing, error handling, monitoring and documentation to deploy reliable pipelines at scale. DLT fully supports both Python and SQL and is tailored to work with both streaming and batch workloads.

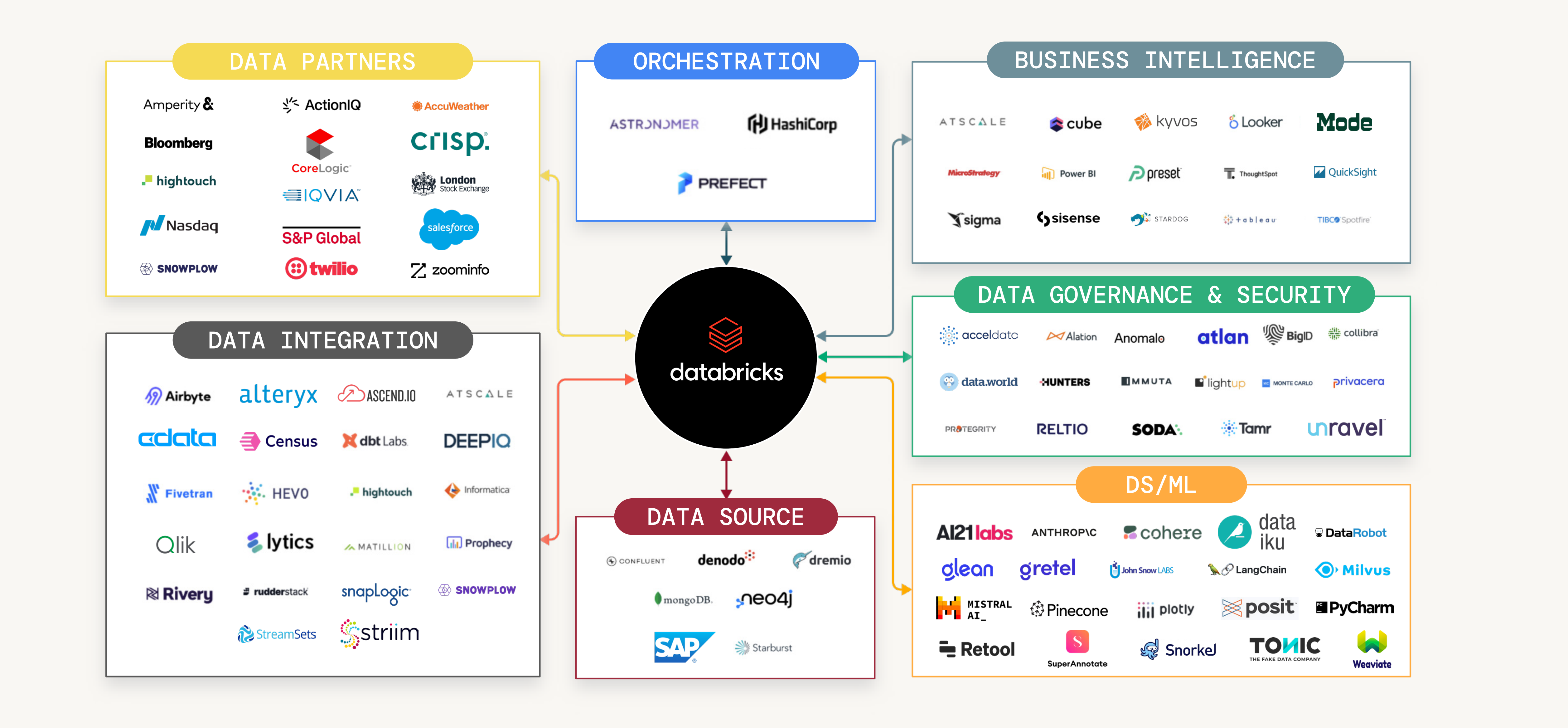
Unified data orchestration with Databricks Workflows

Databricks Workflows offers a simple, reliable orchestration solution for data and AI on the Data Intelligence Platform. Databricks Workflows lets you define multistep workflows to implement ETL pipelines, ML training workflows and more. It offers enhanced control flow capabilities and supports different task types and workflow triggering options. As the platform native orchestrator, Databricks Workflows also provides advanced observability to monitor and visualize workflow execution along with alerting capabilities for when issues arise. Databricks Worklfows offers serverless compute options so you can leverage smart scaling and efficient task execution.

A rich ecosystem of data solutions

The Data Intelligence Platform is built on open source technologies and uses open standards so leading data solutions can be leveraged with anything you build on the lakehouse.

A large collection of technology partners makes it easy and simple to integrate the technologies you rely on when migrating to Databricks — and you are not locked into a closed data technology stack.



The Data Intelligence Platform integrates with a large collection of technologies