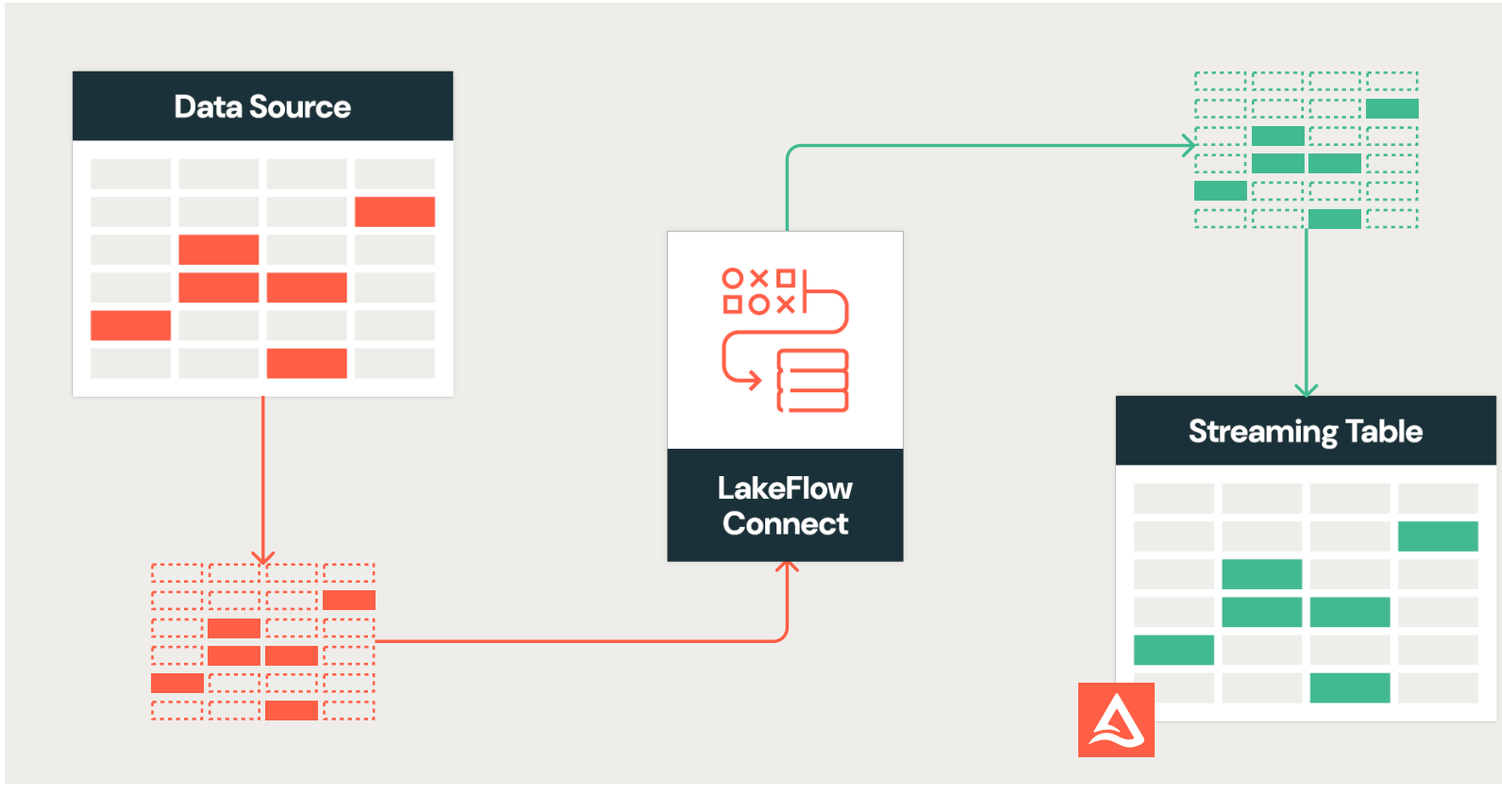
Databricks LakeFlow

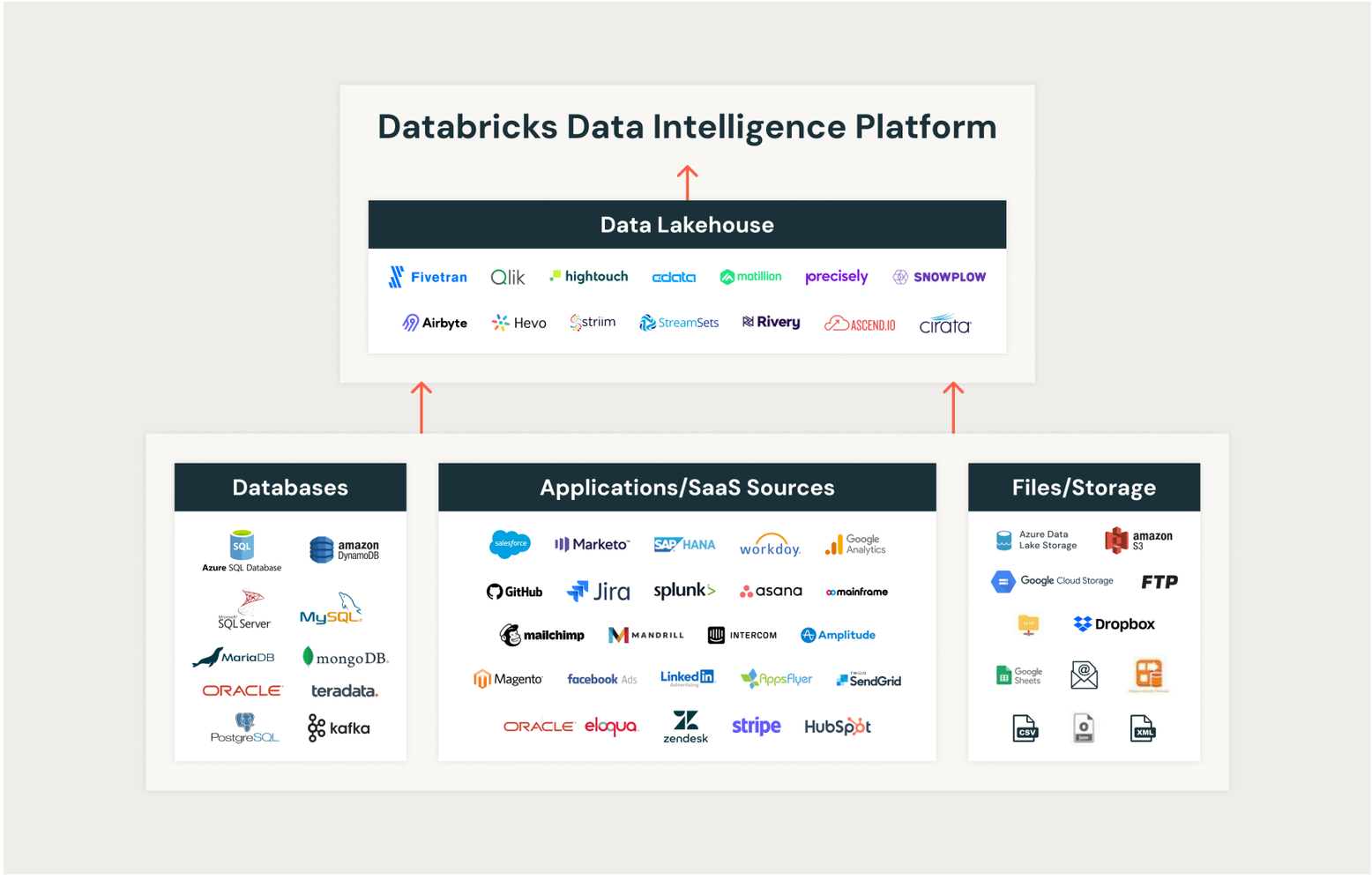
Databricks has announced Databricks LakeFlow — a unified, intelligent solution for data engineering. Native to the Databricks Data Intelligence Platform, LakeFlow empowers users to easily ingest the data they need from external sources (LakeFlow Connect), build and operate data pipelines (LakeFlow Pipelines), 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.

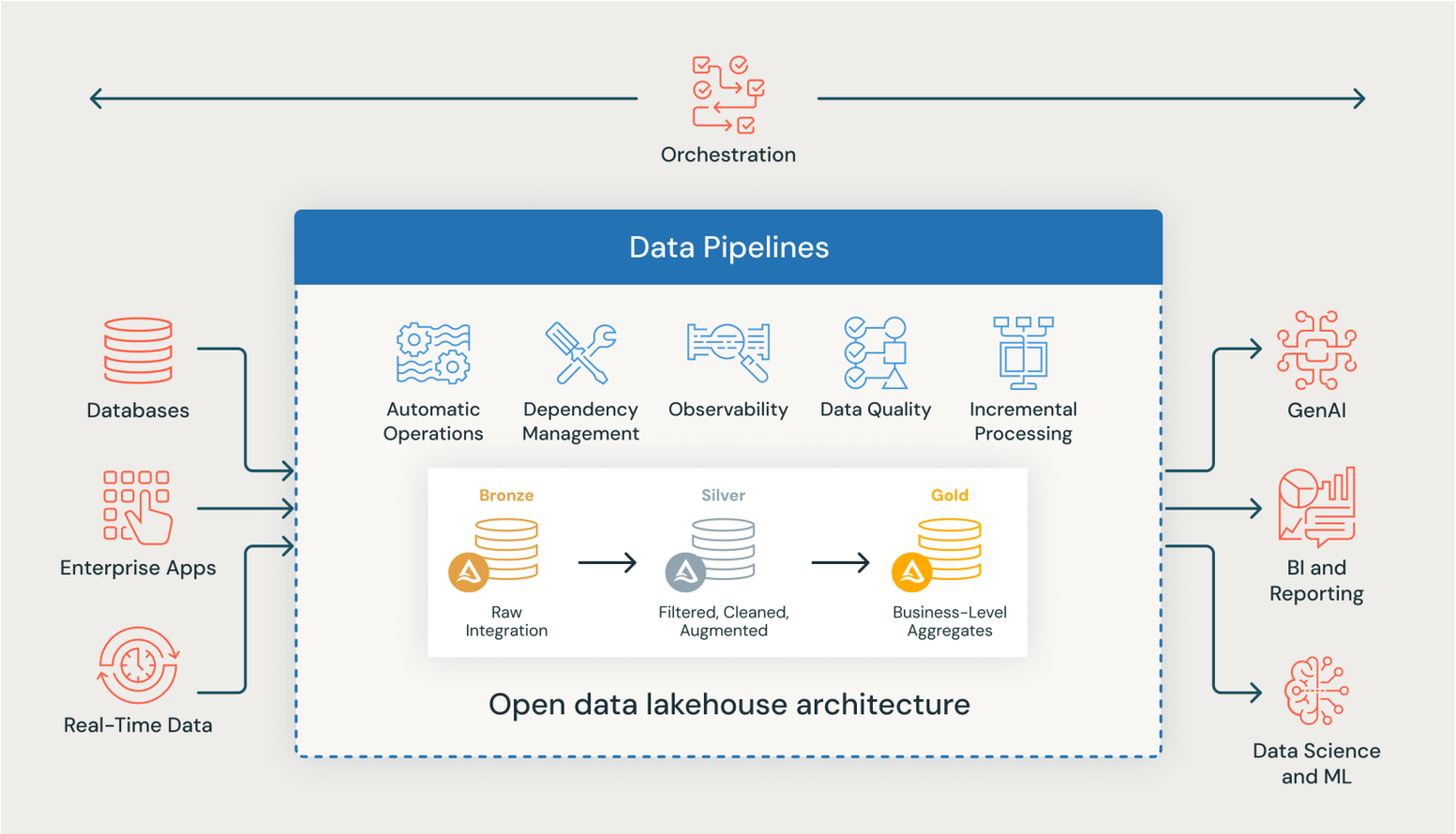
In addition to these native solutions, Databricks has a broad network of data ingestion partners that make it possible to move data from various siloed systems into your data platform. These partners offer a wide range of connectors and native integrations with Databricks to ingest and store data in 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.

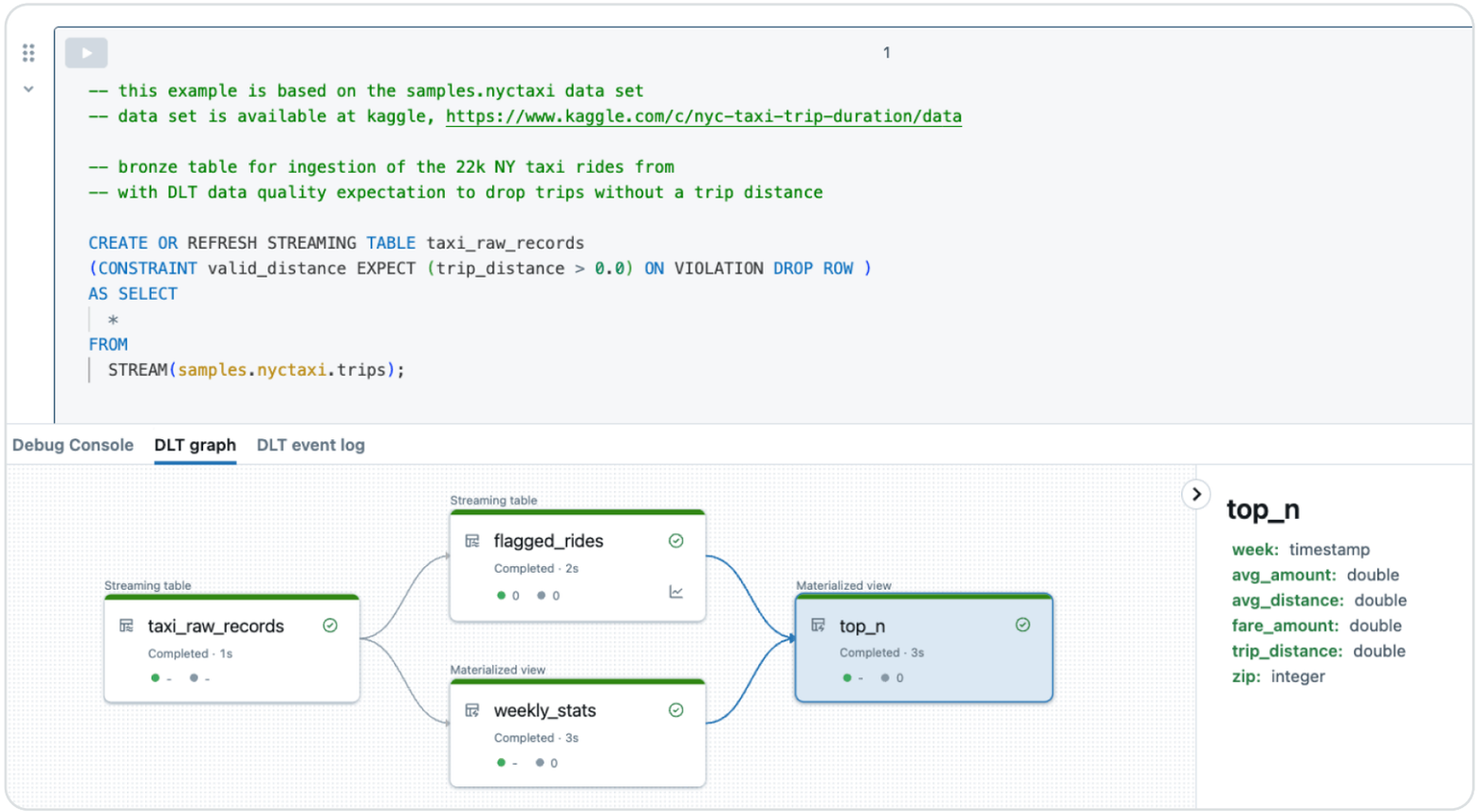
Data Transformation, Quality and Processing

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

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.

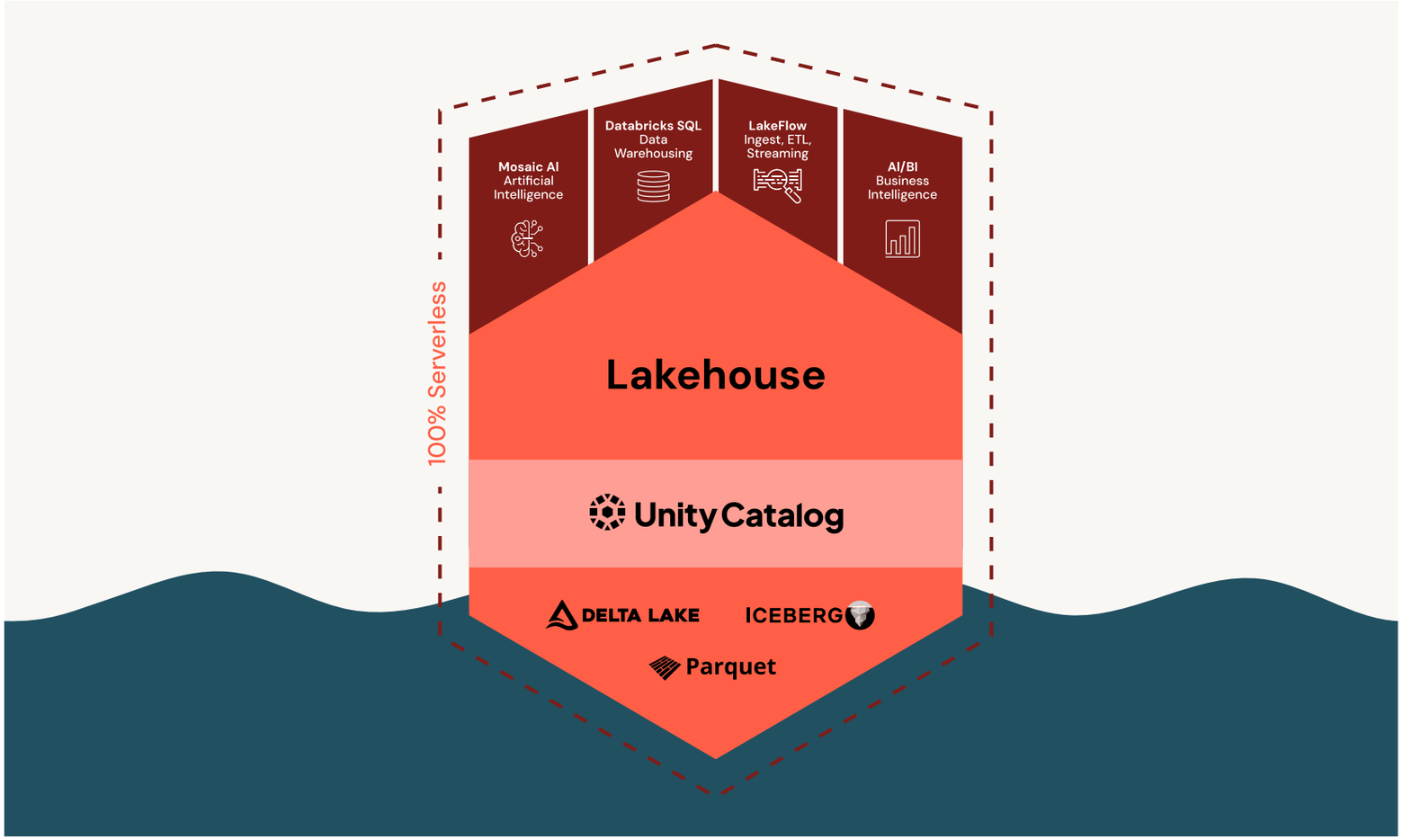
Working With Data

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

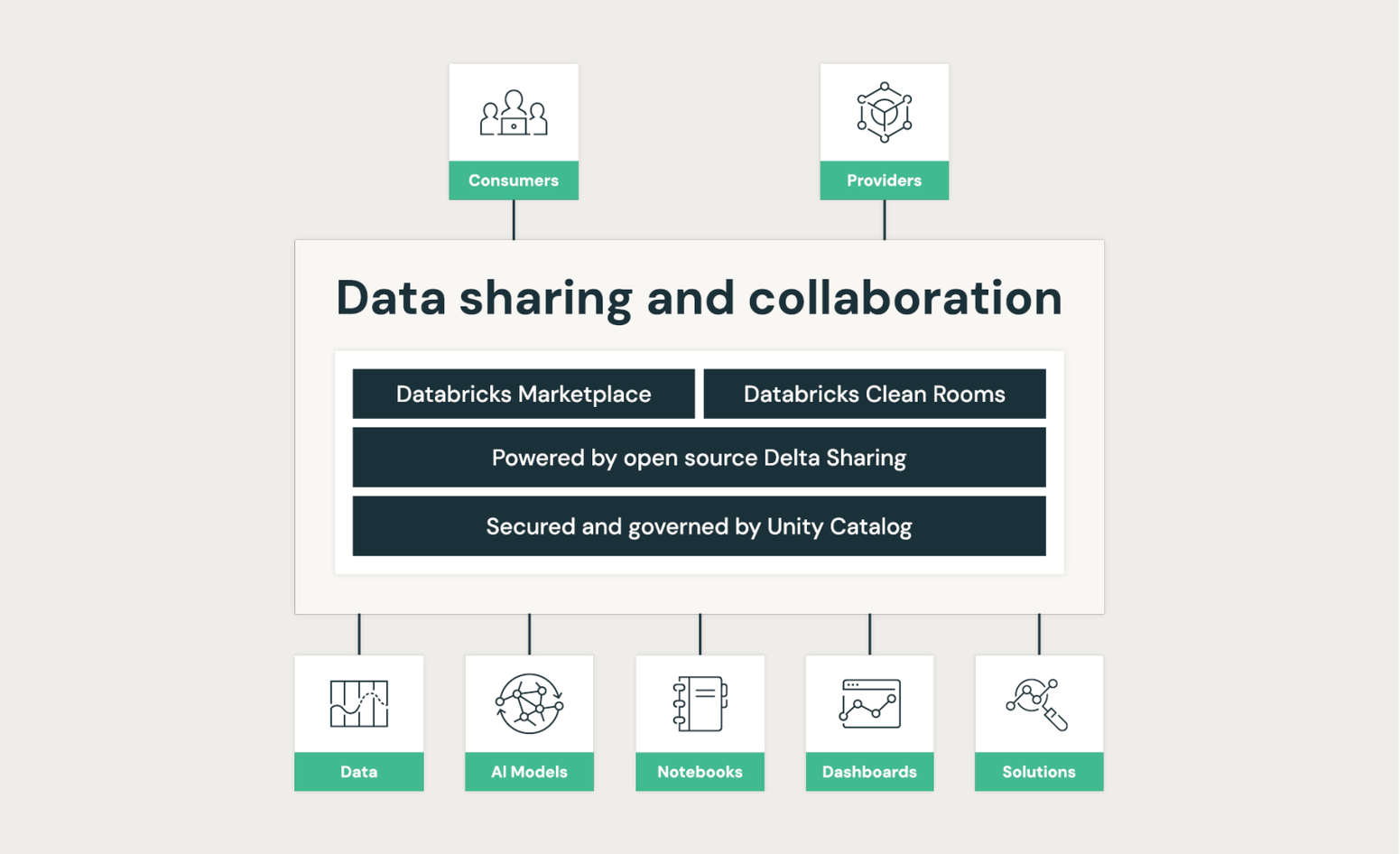
As organizations stand up lakehouse architectures, the supply and demand of cleansed and trusted data doesn't end with analytics and machine learning. Companies need to be able to share and collaborate on data beyond their four walls. Therefore, it is mission-critical that your data strategy aligns with 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.