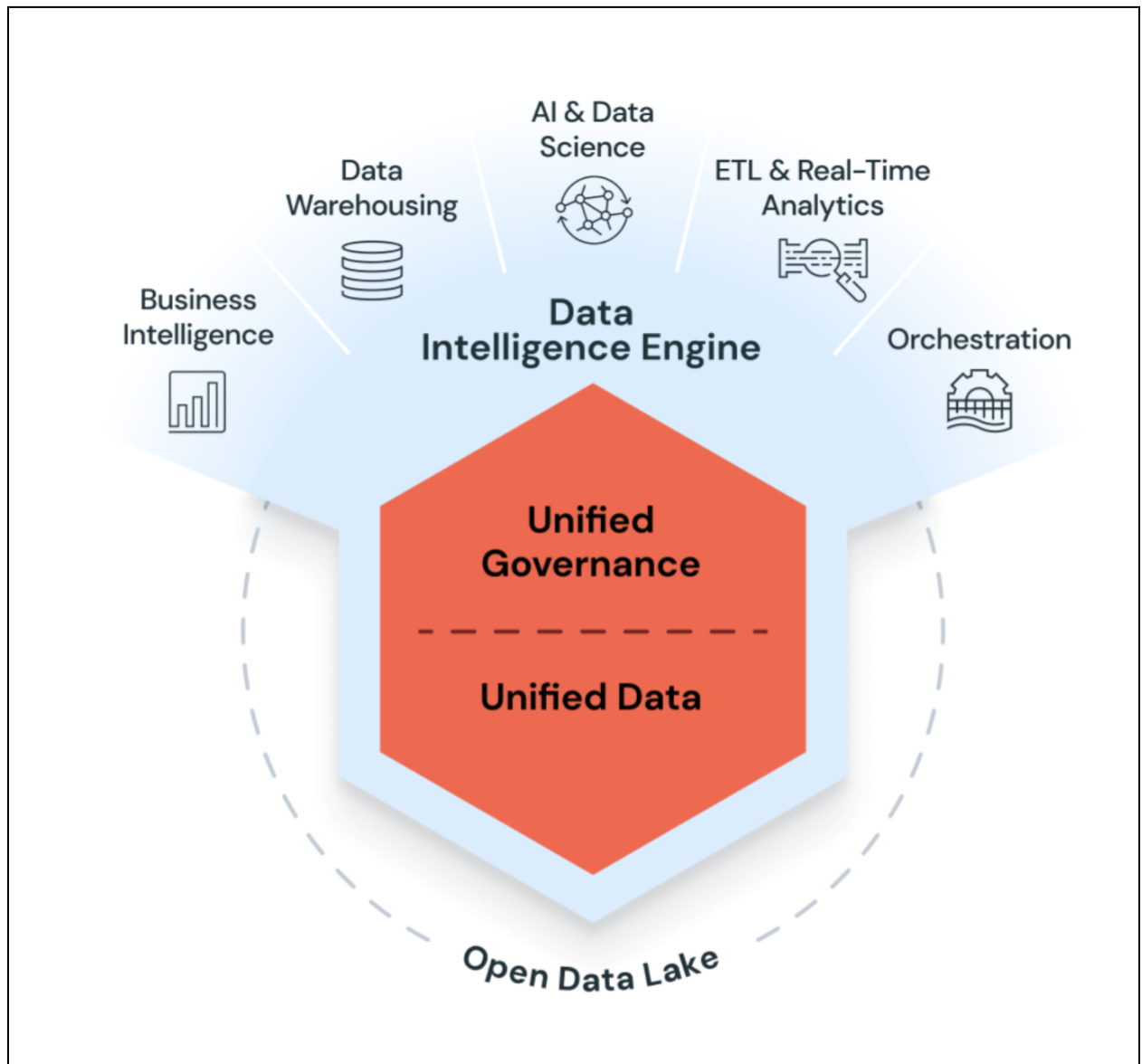


What is Databricks?

Databricks is a unified, open analytics platform for building, deploying, sharing, and maintaining enterprise-grade data, analytics, and AI solutions at scale. The Databricks Data Intelligence Platform integrates with cloud storage and security in your cloud account, and manages and deploys cloud infrastructure for you.



Databricks uses generative AI with the [data lakehouse](#) to understand the unique semantics of your data. Then, it automatically optimizes performance and manages infrastructure to match your business needs.

Natural language processing learns your business's language, so you can search and discover data by asking a question in your own words. Natural language assistance helps you write code, troubleshoot errors, and find answers in documentation.

Managed open source integration

Databricks is committed to the open source community and manages updates of open source integrations with the Databricks Runtime releases. The following technologies are open source projects originally created by Databricks employees:

- [Delta Lake](#) and [Delta Sharing](#)
- [MLflow](#)
- [Apache Spark](#) and [Structured Streaming](#)
- [Redash](#)
- [Unity Catalog](#)

Common use cases

The following use cases highlight some of the ways customers use Databricks to accomplish tasks essential to processing, storing, and analyzing the data that drives critical business functions and decisions.

Build an enterprise data lakehouse

The data lakehouse combines enterprise data warehouses and data lakes to accelerate, simplify, and unify enterprise data solutions. Data engineers, data scientists, analysts, and production systems can all use the data lakehouse as their

single source of truth, providing access to consistent data and reducing the complexities of building, maintaining, and syncing many distributed data systems. See [What is a data lakehouse?](#).

ETL and data engineering

Whether you're generating dashboards or powering artificial intelligence applications, data engineering provides the backbone for data-centric companies by making sure data is available, clean, and stored in data models for efficient discovery and use. Databricks combines the power of [Apache Spark](#) with [Delta](#) and custom tools to provide an unrivaled ETL experience. Use SQL, Python, and Scala to compose ETL logic and orchestrate scheduled job deployment with a few clicks.

[Lakeflow Spark Declarative Pipelines](#) further simplifies ETL by intelligently managing dependencies between datasets and automatically deploying and scaling production infrastructure to ensure timely and accurate data delivery to your specifications.

Databricks provides tools for [data ingestion](#), including [Auto Loader](#), an efficient and scalable tool for incrementally and idempotently loading data from cloud object storage and data lakes into the data lakehouse.

Machine learning, AI, and data science

Databricks machine learning expands the core functionality of the platform with a suite of tools tailored to the needs of data scientists and ML engineers, including [MLflow](#) and [Databricks Runtime for Machine Learning](#).

Large language models and generative AI

Databricks Runtime for Machine Learning includes libraries like [Hugging Face Transformers](#) that allow you to integrate existing pre-trained models or other open source libraries into your workflow. The Databricks MLflow integration makes it easy to use the MLflow tracking service with transformer pipelines, models, and processing components. Integrate [OpenAI](#) models or solutions from partners like [John Snow Labs](#) in your Databricks workflows.

With Databricks, customize a LLM on your data for your specific task. With the support of open source tooling, such as Hugging Face and DeepSpeed, you can efficiently take a foundation LLM and start training with your own data for more accuracy for your domain and workload.

In addition, Databricks provides AI functions that SQL data analysts can use to access LLM models, including from OpenAI, directly within their data pipelines and workflows. See [Apply AI on data using Databricks AI Functions](#).

Data warehousing, analytics, and BI

Databricks combines user-friendly UIs with cost-effective compute resources and infinitely scalable, affordable storage to provide a powerful platform for running analytic queries. Administrators configure scalable compute clusters as [SQL warehouses](#), allowing end users to execute queries without worrying about any of the complexities of working in the cloud. SQL users can run queries against data in the lakehouse using the [SQL query editor](#) or in notebooks. [Notebooks](#) support Python, R, and Scala in addition to SQL, and allow users to embed the same [visualizations](#) available in [legacy dashboards](#) alongside links, images, and commentary written in markdown.

Data governance and secure data sharing

Unity Catalog provides a unified data governance model for the data lakehouse. Cloud administrators configure and integrate coarse access control permissions for Unity Catalog, and then Databricks administrators can manage permissions for teams and individuals. Privileges are managed with access control lists (ACLs) through either user-friendly UIs or SQL syntax, making it easier for database administrators to secure access to data without needing to scale on cloud-native identity access management (IAM) and networking.

Unity Catalog makes running secure analytics in the cloud simple, and provides a division of responsibility that helps limit the reskilling or upskilling necessary for both administrators and end users of the platform. See [What is Unity Catalog?](#).

The lakehouse makes data sharing within your organization as simple as granting query access to a table or view. For sharing outside of your secure environment, Unity Catalog features a managed version of [Delta Sharing](#).

DevOps, CI/CD, and task orchestration

The development lifecycles for ETL pipelines, ML models, and analytics dashboards each present their own unique challenges. Databricks allows all of your users to leverage a single data source, which reduces duplicate efforts and out-of-sync reporting. By additionally providing a suite of common tools for versioning, automating, scheduling, deploying code and production resources, you can simplify your overhead for monitoring, orchestration, and operations.

[Jobs](#) schedule Databricks notebooks, SQL queries, and other arbitrary code. [Databricks Asset Bundles](#) allow you to define, deploy, and run Databricks resources such as jobs and pipelines programmatically. [Git folders](#) let you sync Databricks projects with a number of popular git providers.

For CI/CD best practices and recommendations, see [Best practices and recommended CI/CD workflows on Databricks](#). For a complete overview of tools for developers, see [Develop on Databricks](#).

Real-time and streaming analytics

Databricks leverages Apache Spark Structured Streaming to work with streaming data and incremental data changes. Structured Streaming integrates tightly with Delta Lake, and these technologies provide the foundations for both Lakeflow Spark Declarative Pipelines and Auto Loader. See [Structured Streaming concepts](#).

Online transactional processing

Lakebase is an online transactional processing (OLTP) database that is fully integrated with the Databricks Data Intelligence Platform. This fully managed Postgres database allows you to create and manage OLTP databases stored in Databricks-managed storage. See [What is Lakebase?](#).