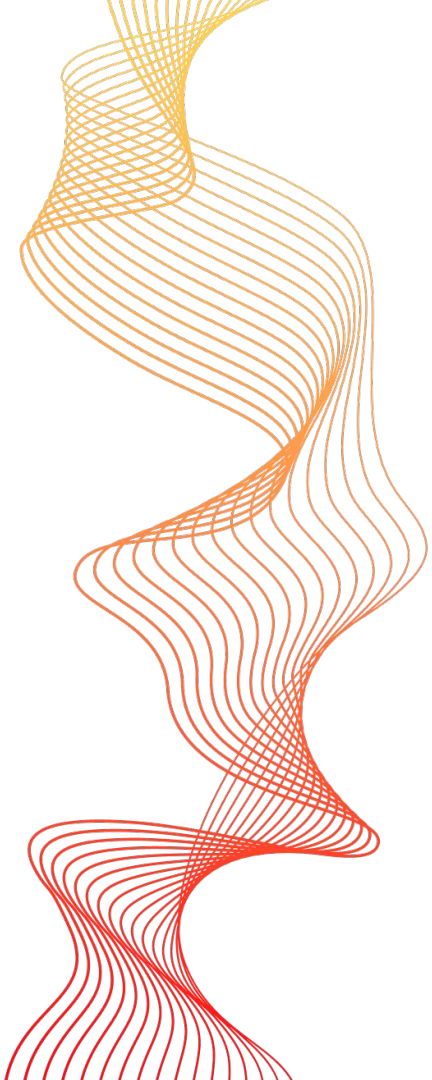
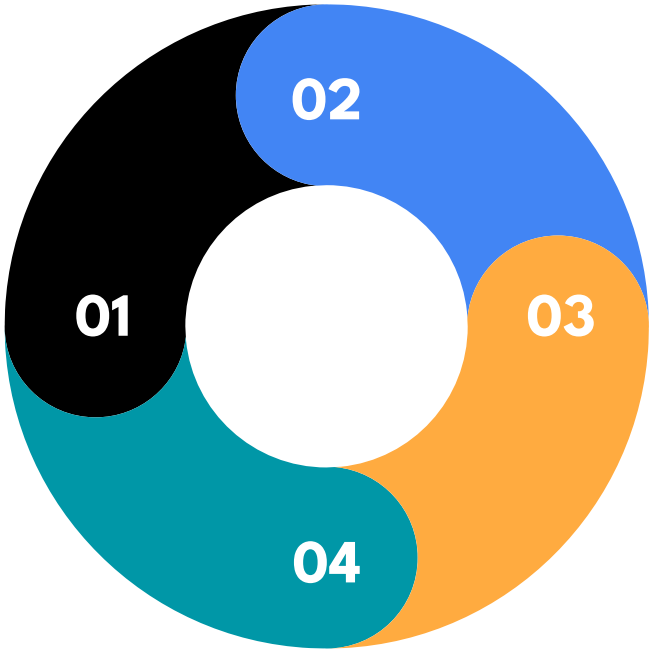
Databricks

An overview of Databricks and its features.



What is Databricks?

Databricks is a unified analytics platform.



Databricks provides a collaborative environment for data scientists, data engineers, and business analysts.

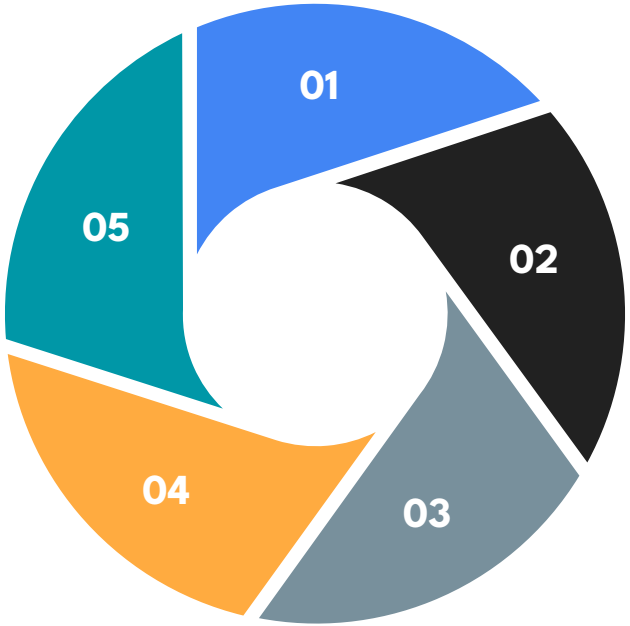
It is built for massive scale and high-performance analytics.

It supports various programming languages like Python, Scala, and R.

Key Features

Advanced analytics capabilities.

Real-time data processing.

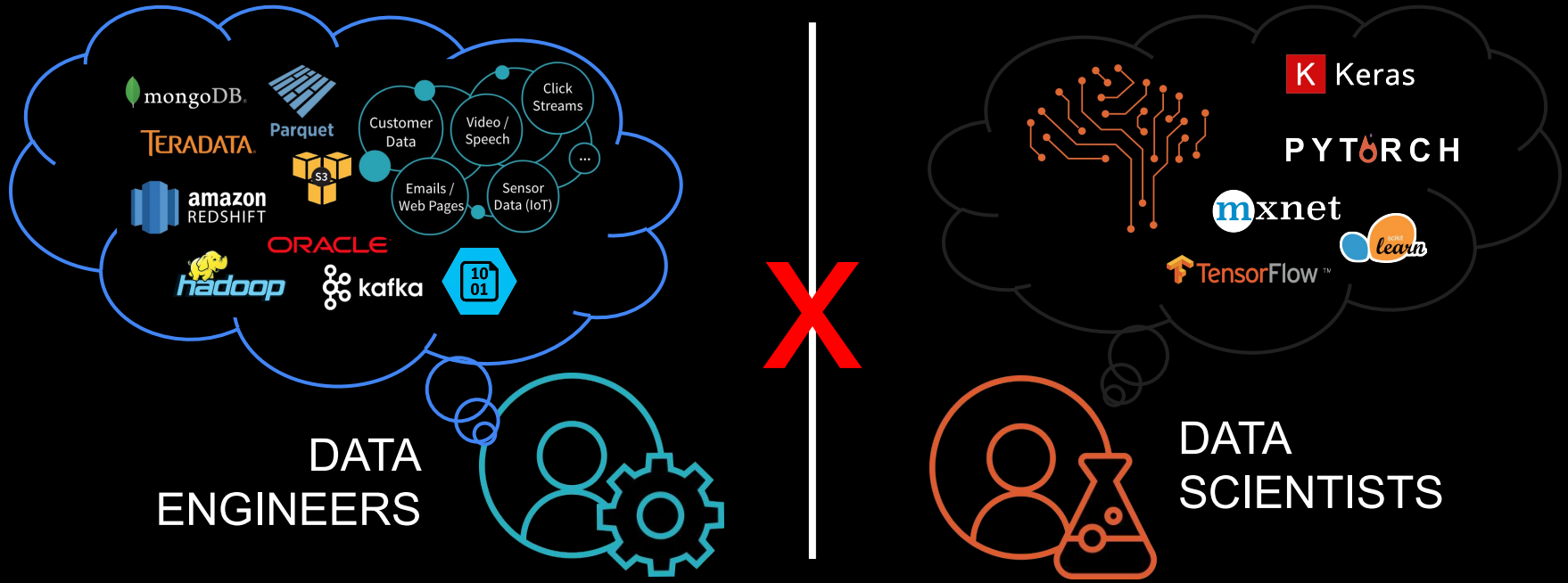


Scalable and elastic cloud infrastructure.

Built-in collaboration tools for teams.

Integration with popular data sources.

Data & AI People are in Silos



Azure Databricks

AZURE DATA SOURCES

Blob Storage

Data Lake Store

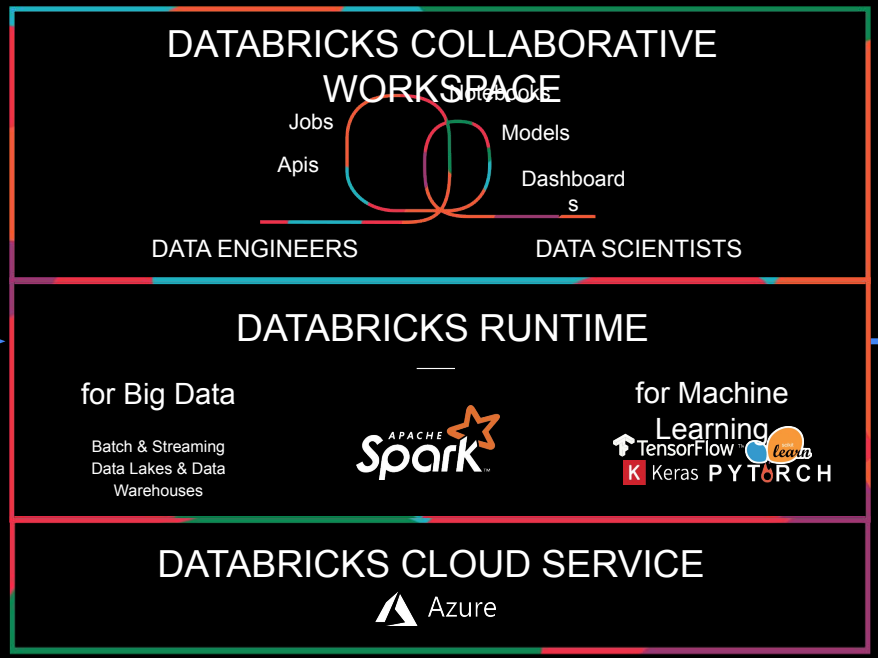
SQL Data Warehouse

Cosmos DB

Event Hub

IoT Hub

Azure Data Factory



Power BI

BI Reporting
Dashboards

Active Directory

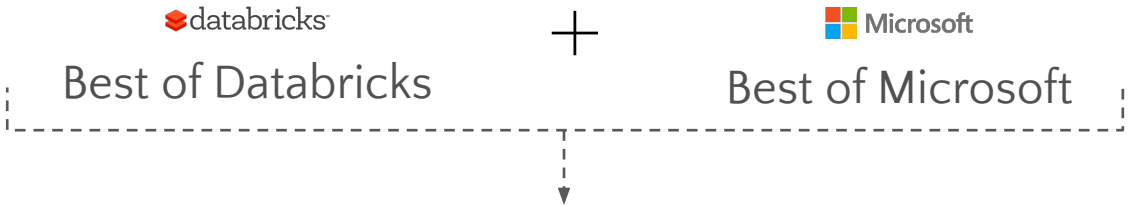
Security Integration

Azure Portal
One-Click setup
Unified Billing

databricks

What is Azure Databricks?

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure



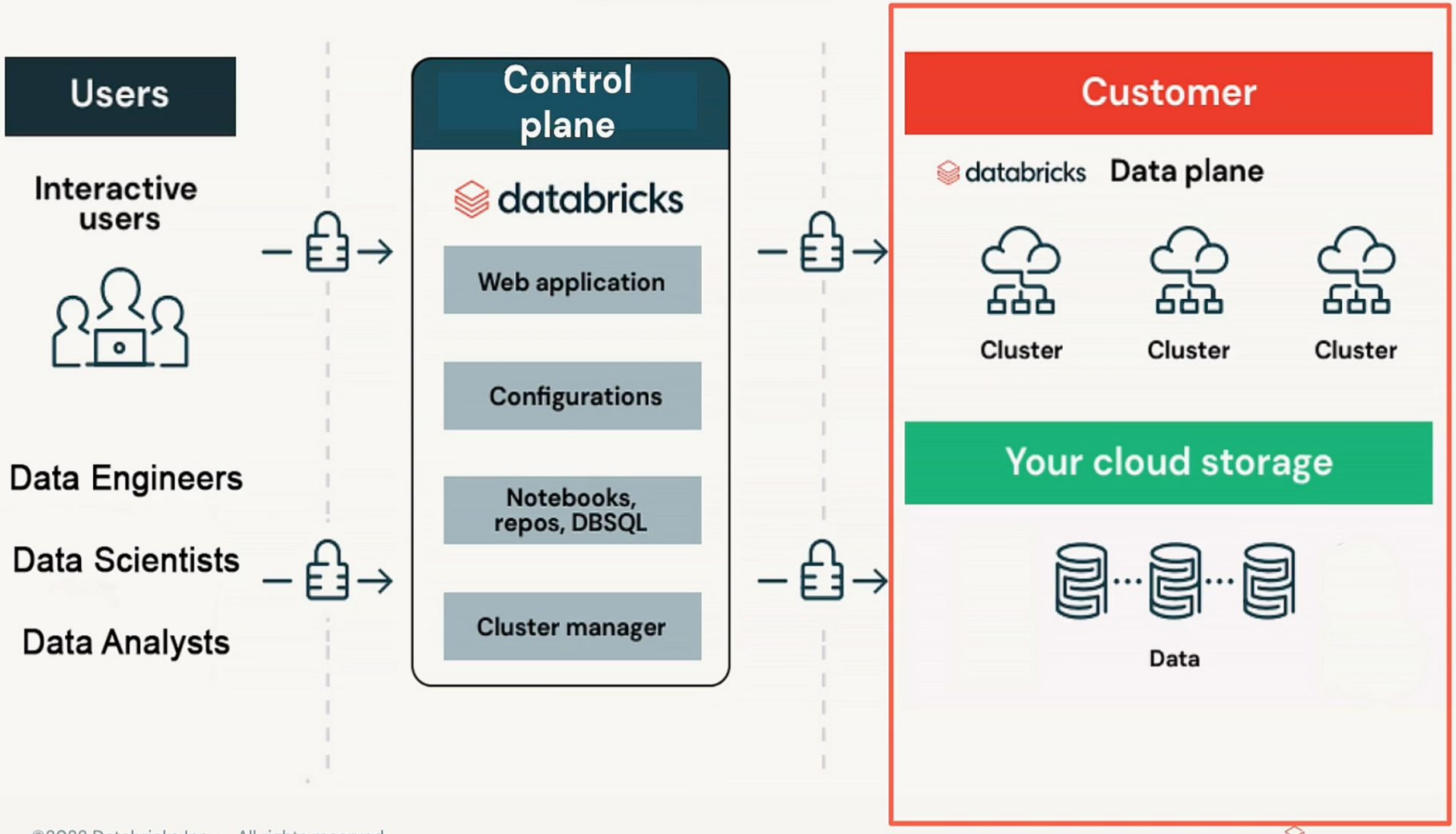
- Designed in collaboration with the founders of Apache Spark
- One-click set up; streamlined workflows
- Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.
- Native integration with Azure services (Power BI, SQL DW, Cosmos DB, Blob Storage)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

What is Azure Databricks?

Azure 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 on your behalf.

The Azure Databricks workspace provides a unified interface and tools for most data tasks, including:

- Data processing workflows scheduling and management
- Generating dashboards and visualizations
- Managing security, governance, high availability, and disaster recovery
- Data discovery, annotation, and exploration
- Machine learning (ML) modeling, tracking, and model serving
- Generative AI solutions



What is Delta lake?

Delta Lake brings ACID to object storage

Atomicity means all transactions either succeed or fail completely

Consistency guarantees relate to how a given state of the data is observed by simultaneous operations

Isolation refers to how simultaneous operations conflict with one another. The isolation guarantees that Delta Lake provides do differ from other systems

Durability means that committed changes are permanent

Problems solved by ACID

- Hard to append data
- Modification of existing data difficult
- Jobs failing mid way
- Real-time operations hard
- Costly to keep historical data versions

What is Unity Catalog?

Unity Catalog
Overview

Unified governance across clouds

Fine-grained governance for data lakes across clouds - based on open standard ANSI SQL.

Unified data and AI assets

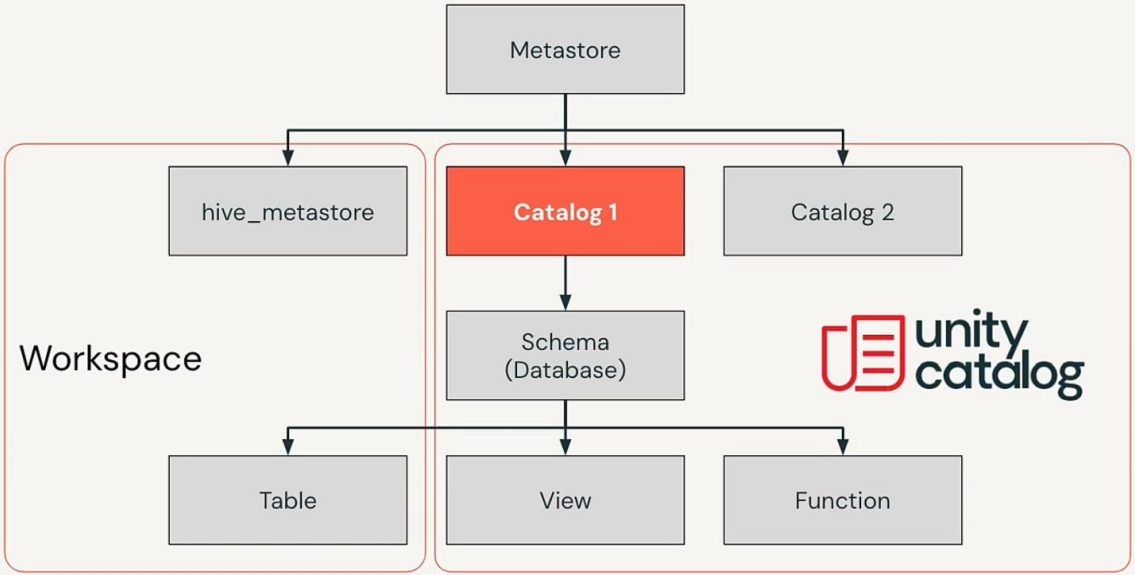
Centrally share, audit, secure and manage all data types with one simple interface.

Unified existing catalogs

Works in concert with existing data, storage, and catalogs - no hard migration required.

Metastore

Accessing legacy Hive metastore



Catalog

Three-level namespace

Traditional SQL two-level namespace

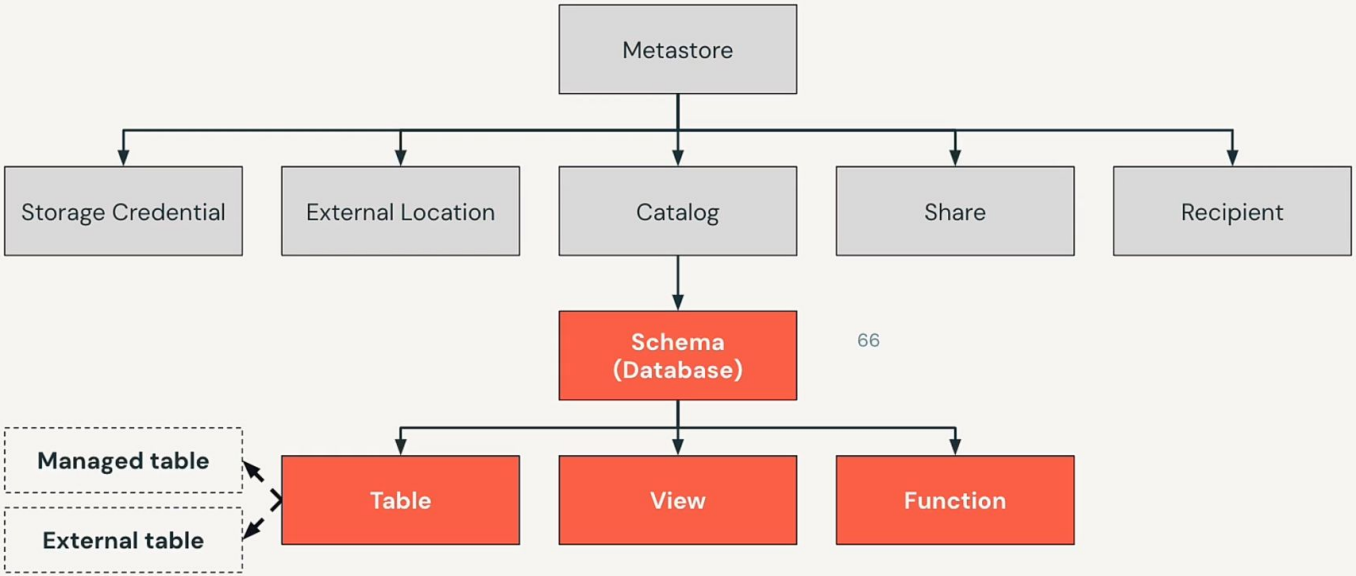
SELECT * FROM schema.table

Unity Catalog three-level namespace

SELECT * FROM catalog.schema.table

Data Objects

Schema (database), tables, views, functions

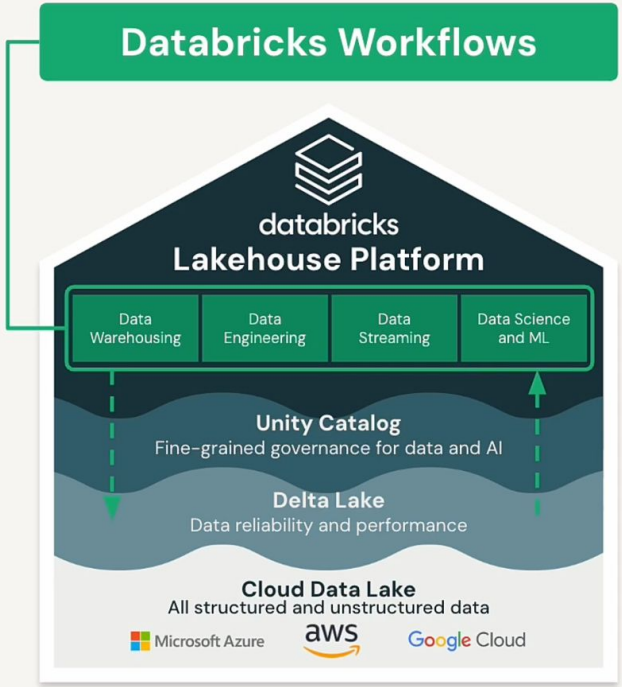


What is Workflow?

Databricks Workflows

Workflows is a **fully-managed cloud-based general-purpose task orchestration service** for the entire Lakehouse.

Workflows is a service for data engineers, data scientists and analysts to build reliable data, analytics and AI workflows on any cloud.



Databricks Workflows

Databricks has two main task orchestration services:

- **Workflow Jobs (Workflows)**
  - Workflows for every job
- **Delta Live Tables (DLT)**
  - Automated data pipelines for Delta Lake

Note: DLT pipeline can be a task in a workflow

DLT versus Workflow Jobs

Considerations

	Delta Live Tables	Workflow Jobs
Source	Notebooks only	JARs, notebooks, DLT, application written in Scala, Java, Python
Dependencies	Automatically determined	Manually set
Cluster	Self-provisioned	Self-provisioned or existing
Timeouts and Retries	Timeouts not supported Retries handled automatically (in production mode)	80 Supported
Import Libraries	Not supported	Supported

Workflow Jobs
Use Cases

Orchestration of Dependent Jobs

Jobs running on schedule, containing dependent tasks/steps

Jobs Workflows

Machine Learning Tasks

Run MLflow notebook task in a job

Jobs Workflows

Arbitrary Code, External API Calls, Custom Tasks

Run tasks in a job which can contain Jar file, Spark Submit, Python Script, SQL task, dbt

Jobs Workflows

How to Leverage Workflows

- Allows you to build simple ETL/ML task orchestration
- Reduces infrastructure overhead
- Easily integrate with external tools
- Enables non-engineers to build their own workflows using simple UI
- Cloud-provider independent
- Enables re-using clusters to reduce cost and startup time

Databricks Repos

Git Versioning

Native integration with Github, Gitlab, Bitbucket and Azure Devops

UI-based workflows

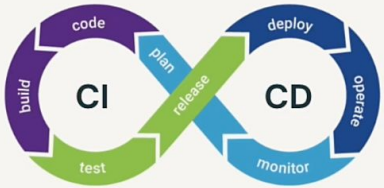
Azure DevOps GitHub

GitLab Bitbucket

CI/CD Integration

API surface to integrate with automation

Simplifies the dev/staging/prod multi-workspace story



Enterprise ready

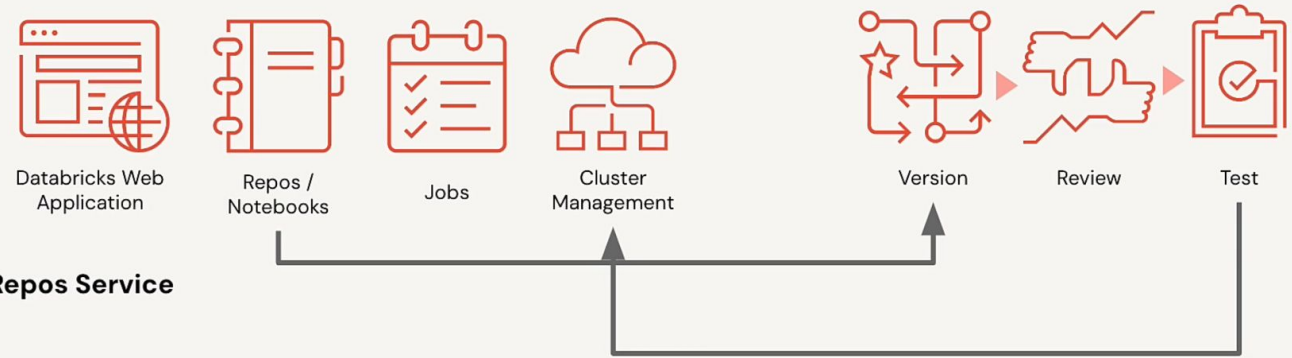
Allow lists to avoid exfiltration

Secret detection to avoid leaking keys

Databricks Repos
CI/CD Integration

Control Plane in Databricks
Manage customer accounts, datasets, and clusters

Git and CI/CD Systems



Demo

What is Cluster?

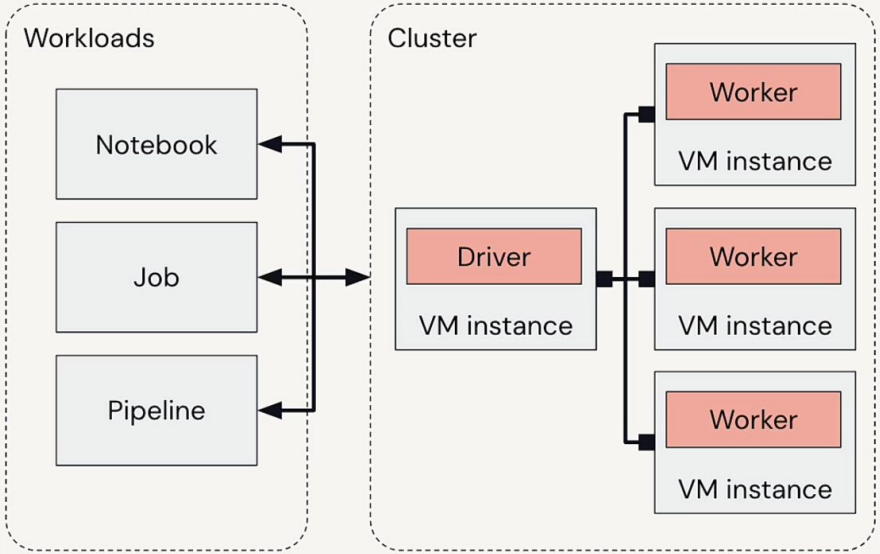
Clusters
Overview

Collection of VM instances

Distributes workloads across workers

Two main types:

1. **All-purpose** clusters for interactive development
2. **Job** clusters for automating workloads



Cluster Types

All-purpose Clusters

Analyze data collaboratively using **interactive** notebooks

Job Clusters

Run **automated** jobs

The Databricks job scheduler creates job clusters when running jobs

Cluster Mode

Single node

Low-cost single-instance cluster catering to single-node machine learning workloads and lightweight exploratory analysis

Standard (Multi Node)

Default mode for workloads developed in any supported language (requires at least two VM instances)

Databricks Runtime Version

Standard

Apache Spark and many other components and updates to provide an optimized big data analytics experiences

Photon

An optional add-on to optimize Spark queries (e.g. SQL, DataFrame)

Machine learning

Adds popular machine learning libraries like TensorFlow, Keras, PyTorch, and XGBoost.

Access Mode

Access mode dropdown	Visible to user	Unity Catalog support	Supported languages
Single user	Always	Yes	Python, SQL, Scala, R
Shared	Always (Premium plan required)	Yes	Python (DBR 11.1+), SQL
No isolation shared	Can be hidden by enforcing user isolation in the admin console or configuring account-level settings	No	Python, SQL, Scala, R
Custom	Only shown for existing clusters *without* access modes (i.e. legacy cluster modes, Standard or High Concurrency); not an option for creating new clusters.	No	Python, SQL, Scala, R

Cluster Policies

Cluster policies can help to achieve the following:

- Standardize cluster configurations
- Provide predefined configurations targeting specific use cases
- Simplify the user experience
- Prevent excessive use and control cost
- Enforce correct tagging

What is DB Notebooks?

Databricks Notebooks

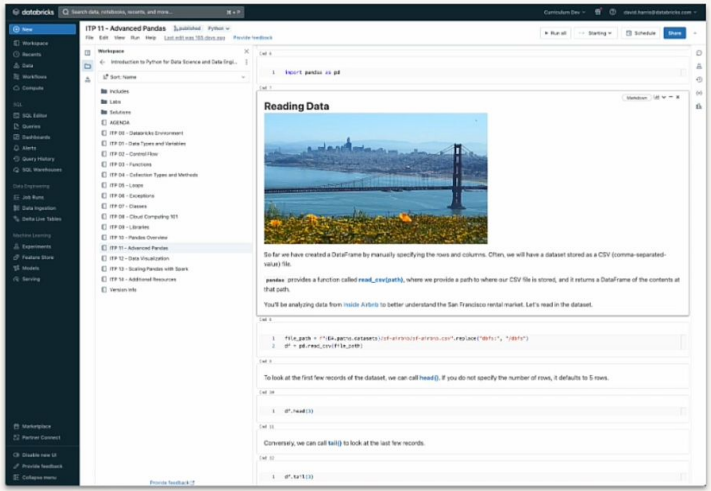
Collaborative, reproducible, and enterprise ready

Multi-language
Use Python, SQL, Scala, and R, all in one Notebook

Collaborative
Real-time co-presence, co-editing, and commenting

Ideal for exploration
Explore, visualize, and summarize data with built-in charts and data profiles

Adaptable
Install standard libraries and use local modules



Reproducible
Automatically track version history, and use git version control with Repos

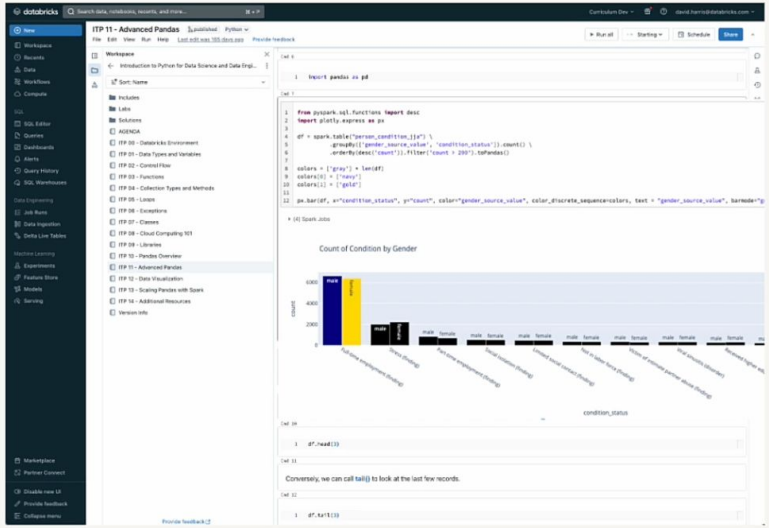
Get to production faster
Quickly schedule notebooks as jobs or create dashboards from their results, all in the Notebook

Enterprise-ready
Enterprise-grade access controls, identity management, and auditability

Databricks Notebooks

Easily develop standard or custom visualizations

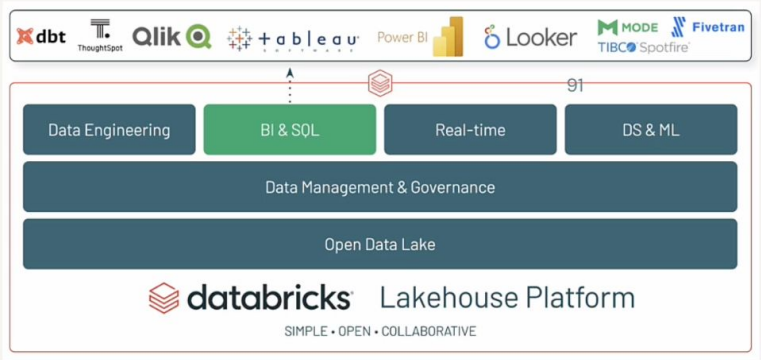
- Create visualizations based on query results or dataframes
- Can use SQL, Python, or Scala
  - Use SQL for out-of-the-box, standard visualizations
  - Use Python and Scala for custom visualizations
- Stitch together custom and standard visuals
- Can be used on existing tables or can write model results to a table for model monitoring

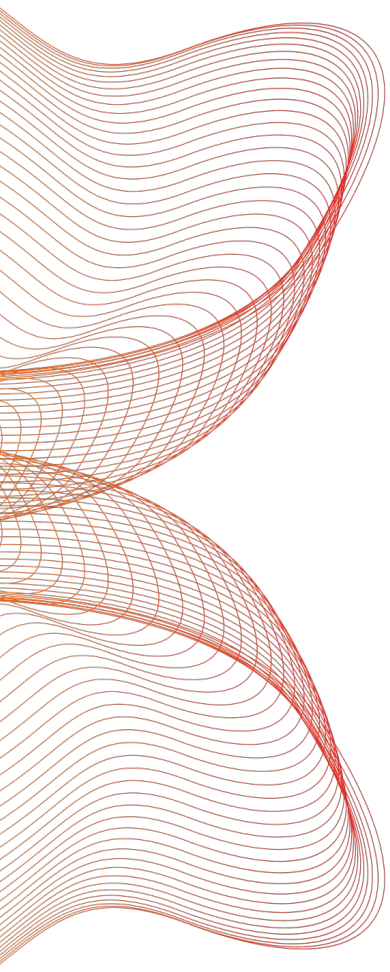


Databricks SQL

Delivering analytics on the freshest data with data warehouse performance and data lake economics

- Better price/performance than other cloud data warehouses
- Simplify discovery and sharing of new insights
- Connect to familiar BI tools, like Tableau or Power BI
- Simplified administration and governance





Use Cases

- Data exploration and visualization.
- Machine learning and AI.
- Predictive analytics and forecasting.
- Realtime monitoring and alerting.

Thank you for your time 😊