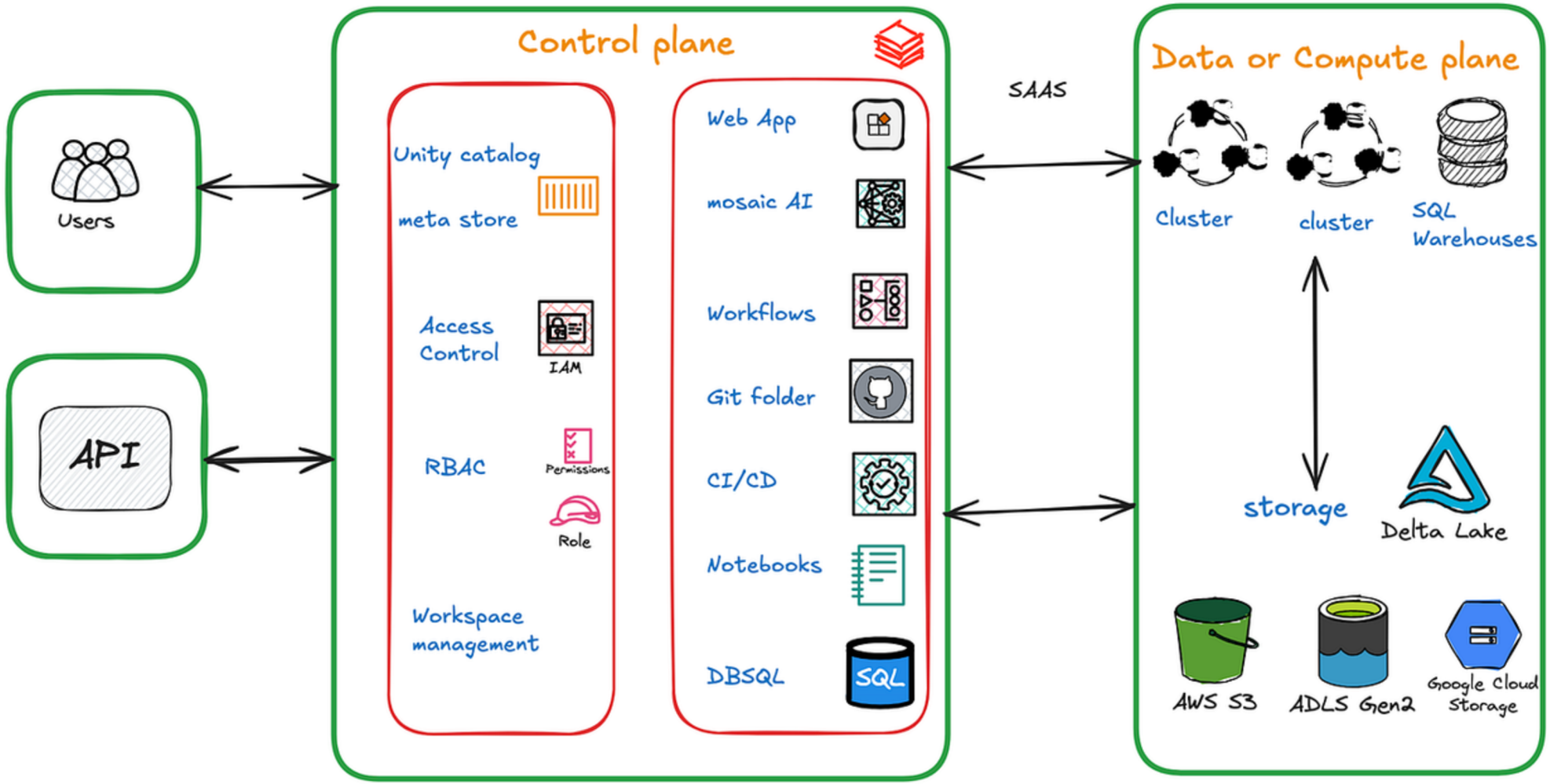
Databricks Architecture

In simple words..

Databricks Architecture



Databricks Architecture Overview

- Databricks is a cloud-based data platform built on Apache Spark that helps you handle big data, machine learning, and analytics in one place.

- It sits on top of cloud providers like Azure, AWS, or GCP and provides a unified workspace.

Main Components of Databricks Architecture

1. Control Plane
2. DataPlane
3. Workspace
4. Clusters
5. Delta Lake (Storage Layer)

1. Control Plane

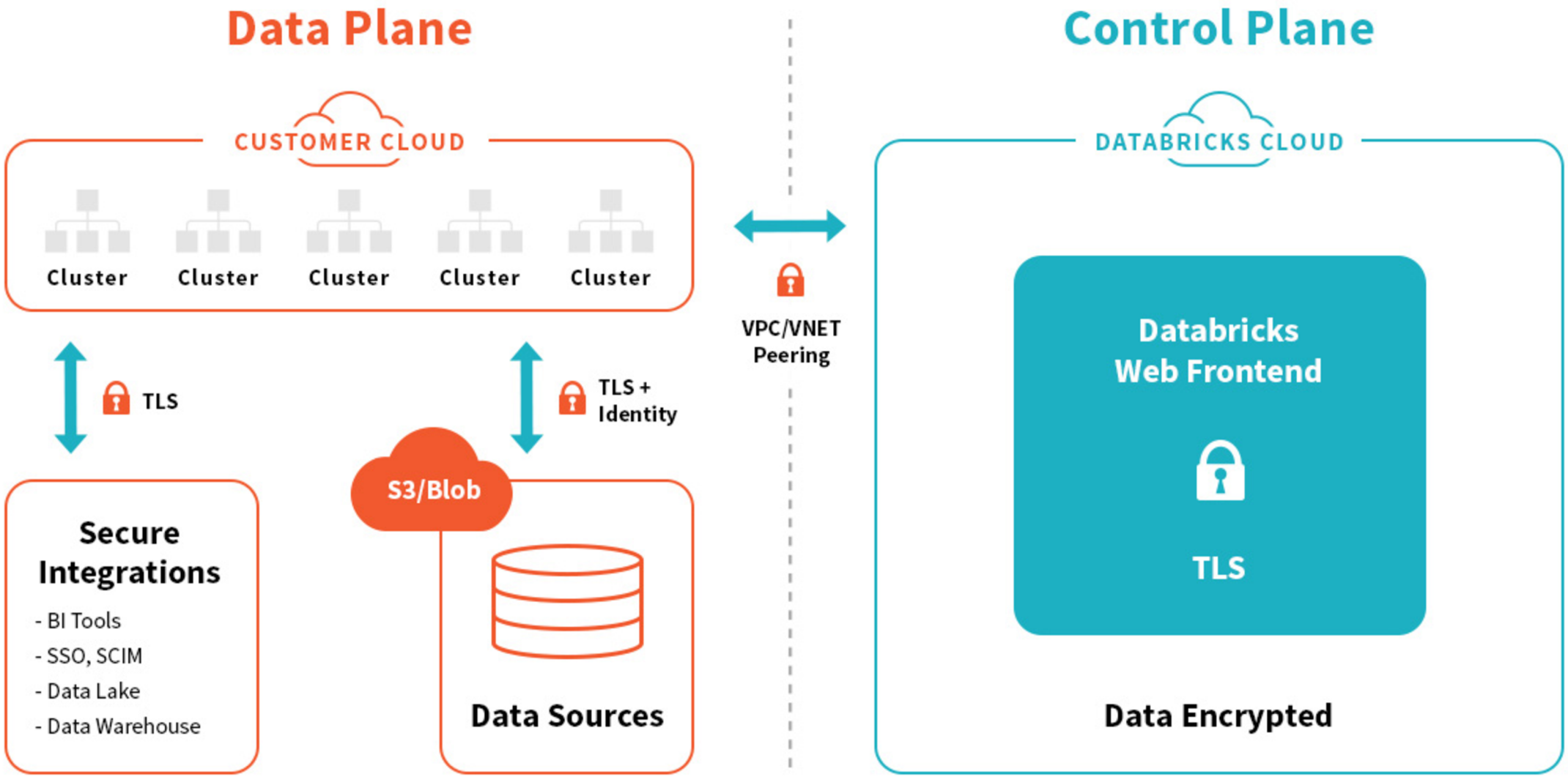
- Managed by Databricks(you don’t manage it).
- Responsible for authentication, jobs scheduling, notebooks,
- REST APIs, UI, cluster management.
- Stores metadata (but not your data).

👉 Think of this as the “brain” that manages everything.

2. Data Plane

- This runs in side your cloud account(Azure, AWS, GCP).
- Yourdata andcomputeclusterslivehere.
- Databricks spins up Spark clusters on-demand in your cloud environment.

👉Think of this as the “muscle” where actual data processing happens.

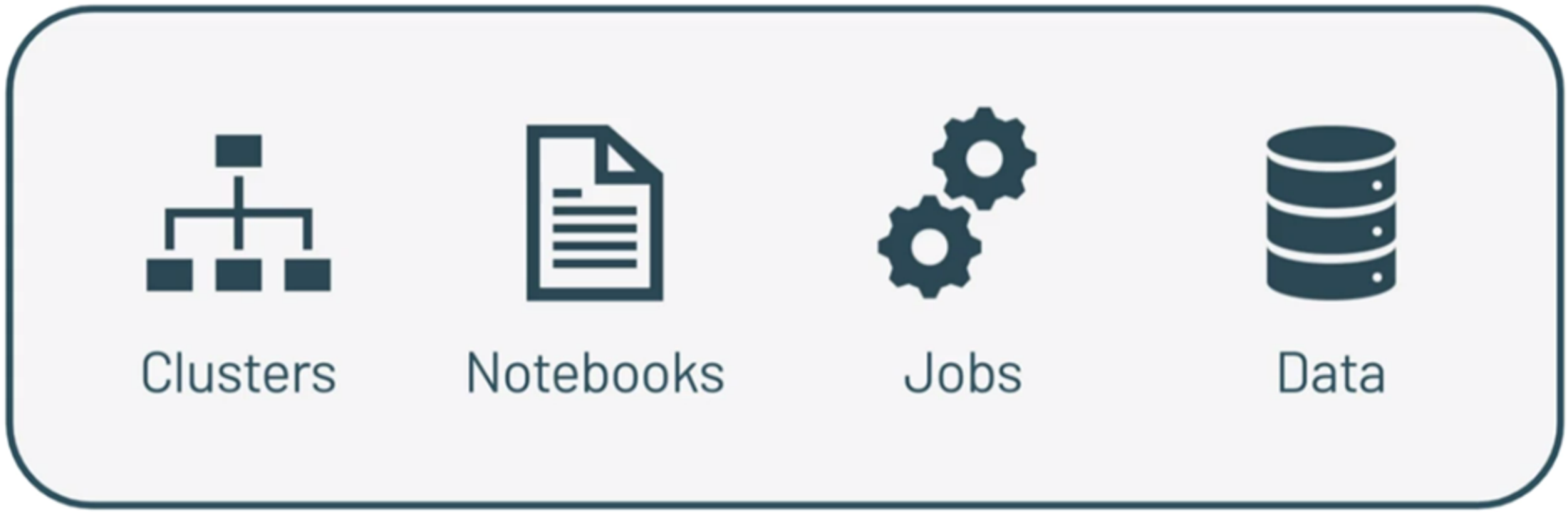

Data Plane
CUSTOMER CLOUD
Cluster
Cluster
Cluster
Cluster
Cluster
TLS
TLS +
Identity
S3/Blob
Secure
Integrations
- BI Tools
- SSO, SCIM
- Data Lake
- Data Warehouse
Data Sources
VPC/VNET
Peering
Control Plane
DATABRICKS CLOUD
Databricks
Web Frontend
TLS
Data Encrypted

3. Workspace

- A collaborative **UI/Notebook environment** for developers, data scientists, and analysts.
- Supports **Python, SQL, R, Scala** in the same notebook.

👉 Example: A Data Engineer writes ETL in PySpark, while a Data Analyst writes SQL in the same workspace.

Databricks Workspace

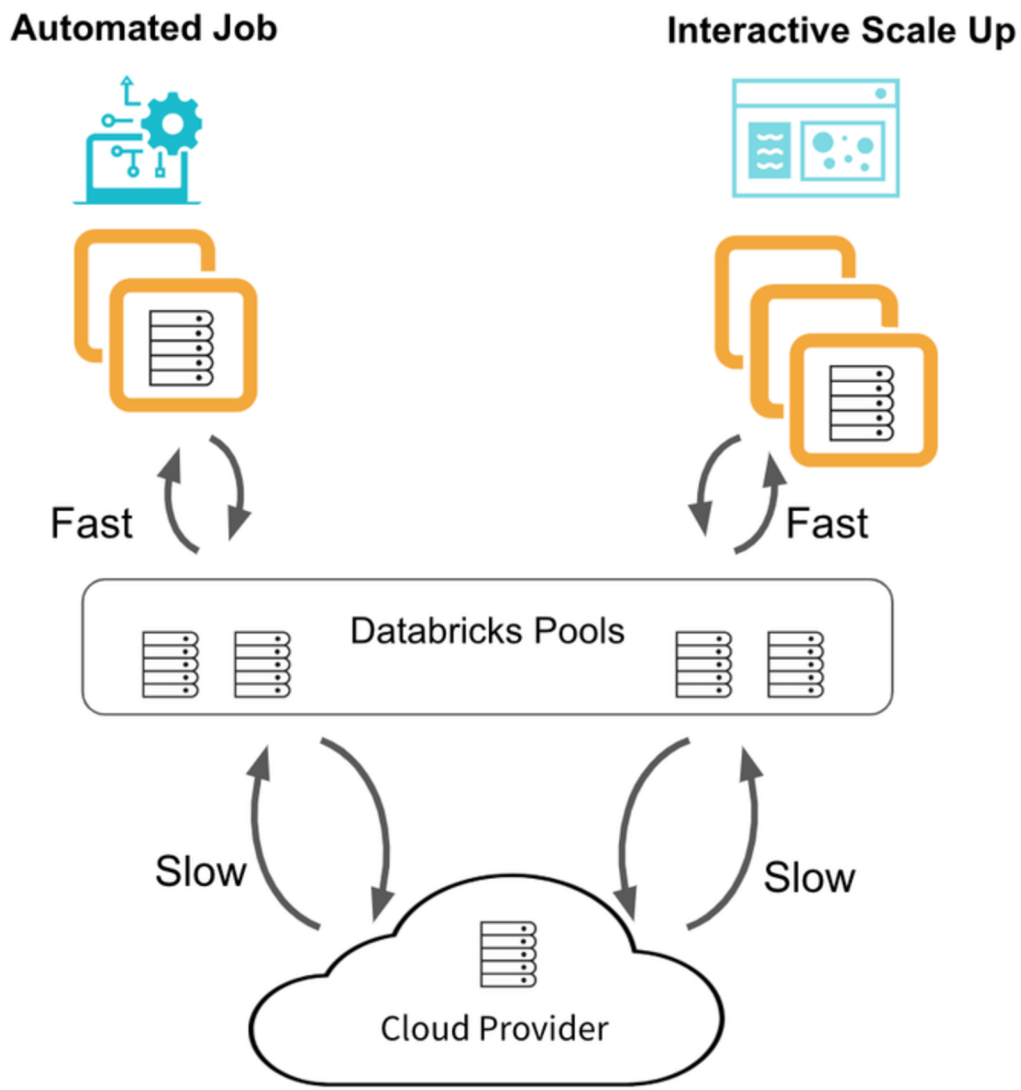


4. Clusters

- Virtual machines(Spark Runtime) that run in your cloud environment.

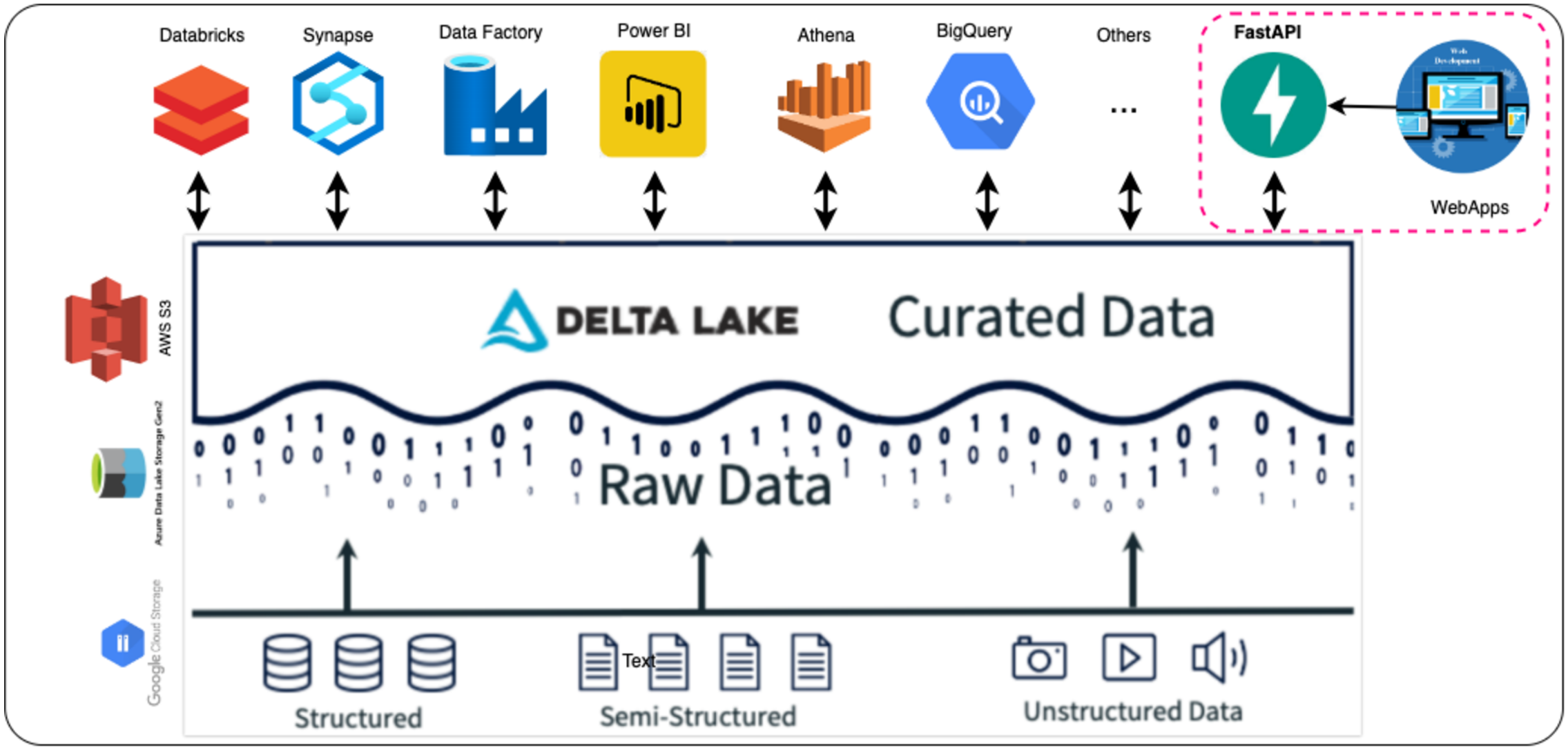
- **Types:**

  -**Interactive clusters:** used for exploration/analysis.
  -**Job clusters:** auto-created for scheduled ETL/ML jobs.


Automated Job
Interactive Scale Up
Fast
Fast
Databricks Pools
Slow
Slow
Cloud Provider

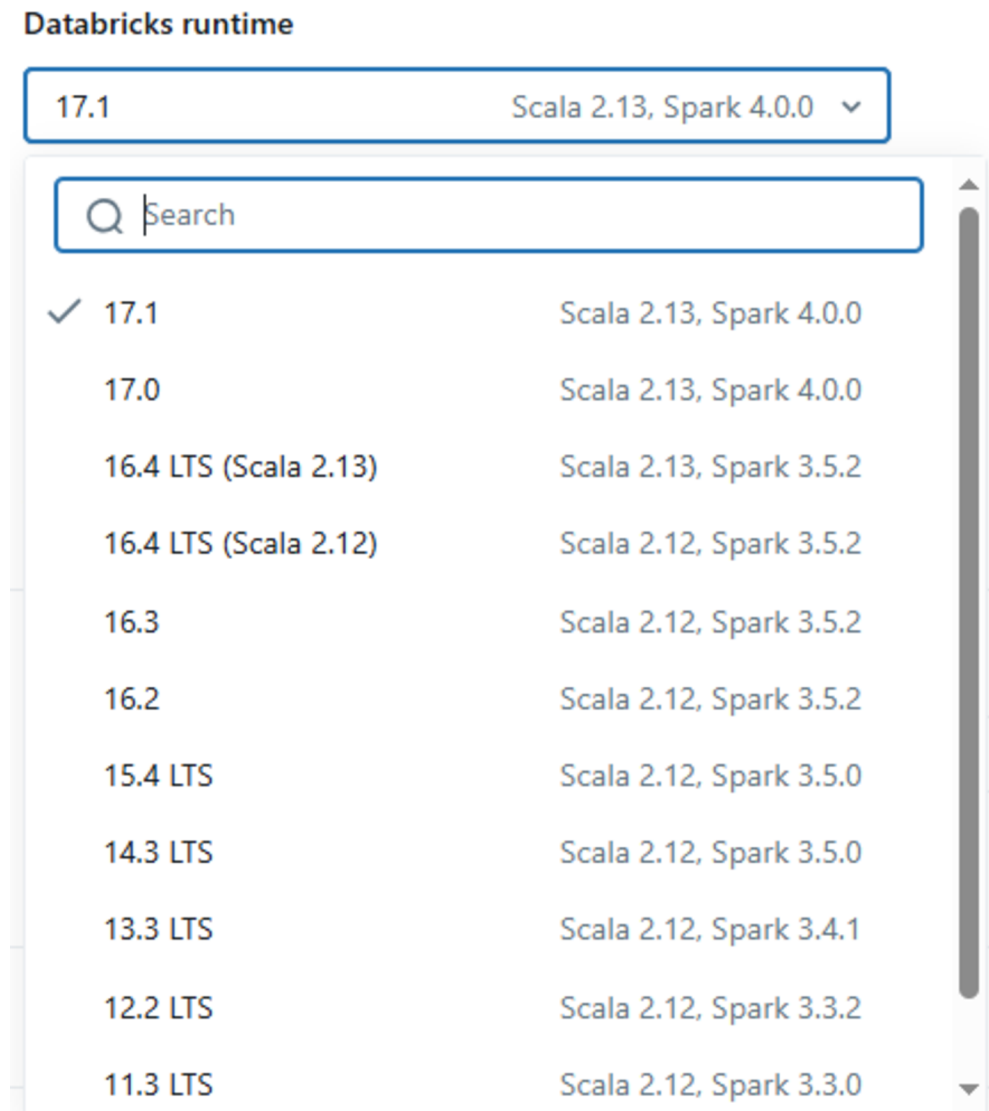
5. Delta Lake(Storage Layer)

- An open-source storage format on top of cloud storage (S3, ADLS, GCS).
- Provides ACID transactions, schema enforcement, time travel.
- Makes data lakes reliable→ calleda Lakehouse.


Databricks
Synapse
Data Factory
Power BI
Athena
BigQuery
Others
FastAPI
WebApps
AWS S3
Azure Data Lake Storage Gen2
Google Cloud Storage
DELTA LAKE
Curated Data
Raw Data
Text
Structured
Semi-Structured
Unstructured Data

6. Databricks Runtime

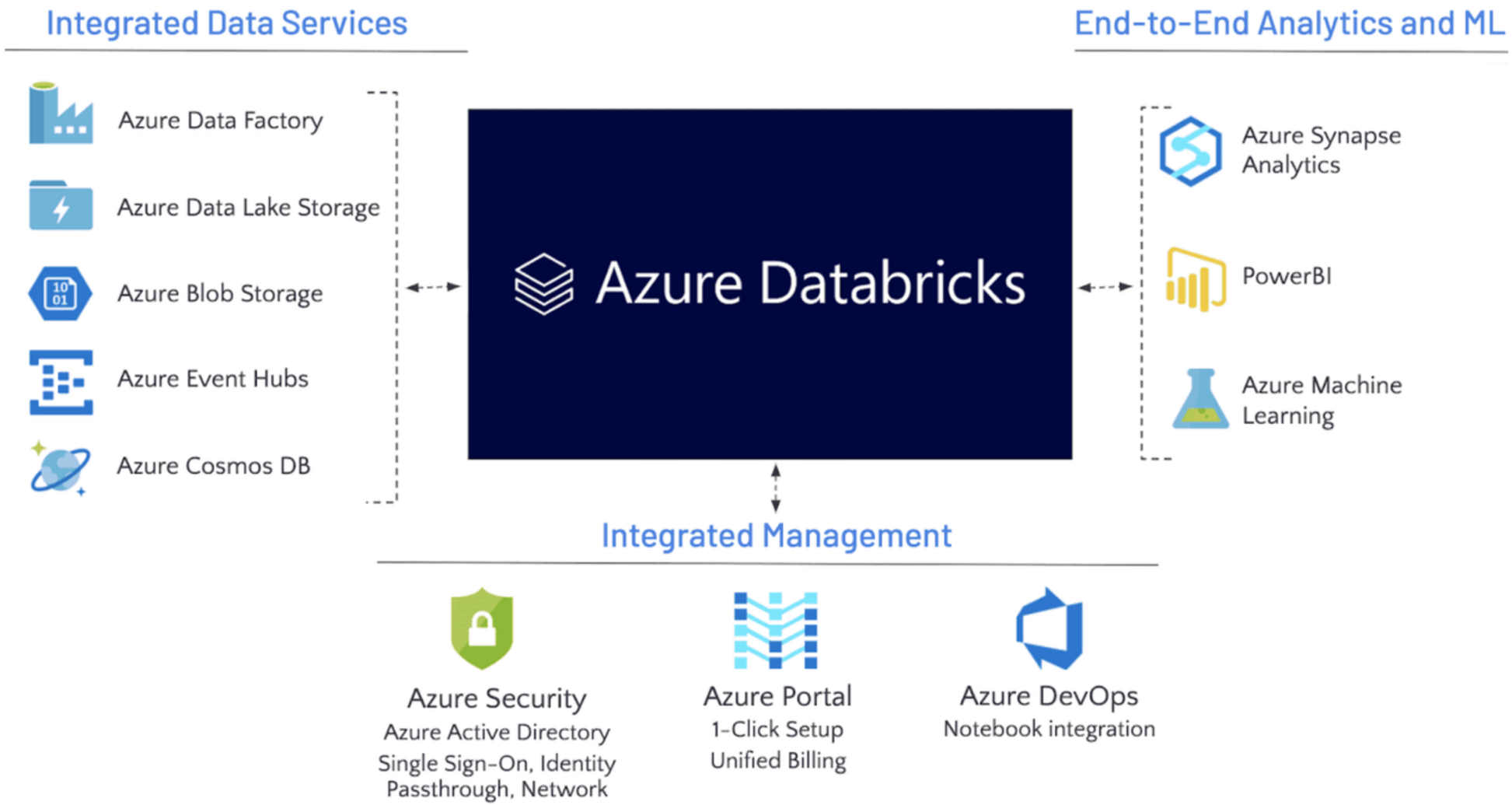
- Pre-packaged Spark environment optimized by Databricks.
- Includes ML flow, Koalas(pandas API),Delta Lake, libraries for ML/AI.



Picking a Databricks Runtime Version

✅ Key Benefit of Databricks Architecture

- Unified platform (ETL +Analytics +ML)
- Secure separation(control plane vs data plane)
- Scalable(auto-scaling Spark clusters)
- Reliable storage (Delta Lake)


Integrated Data Services
Azure Data Factory
Azure Data Lake Storage
Azure Blob Storage
Azure Event Hubs
Azure Cosmos DB
Azure Databricks
End-to-End Analytics and ML
Azure Synapse Analytics
PowerBI
Azure Machine Learning
Integrated Management
Azure Security
Azure Active Directory
Single Sign-On, Identity Passthrough, Network
Azure Portal
1-Click Setup
Unified Billing
Azure DevOps
Notebook integration