What is Databricks Lakehouse?

Learning Objectives

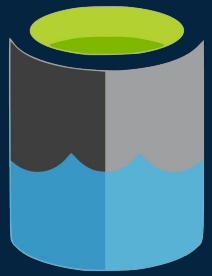
- Explain what a Data Lakehouse is
- Define the Databricks Data Lakehouse architecture
- Become familiar with clusters in Databricks

Data Warehouse

- ETL
- BI and reporting applications
- Strong data governance

Structured Data

Data Lake



- ELT
- BI, Reporting, Data Science applications
- Lack of data governance

Structured, Semi-Structured, Unstructured Data

Data Lakehouse

- ELT
- BI, Reporting, Data Science applications
- Strong data governance

Structured, Semi-Structured, Unstructured Data

Bringing data management and governance to cloud data lakes

Data Warehouse

- Reliable
- Strong governance

Data Lake

- Flexible
- ML and AI workloads

A single unified platform for all data, analytics, and AI workloads

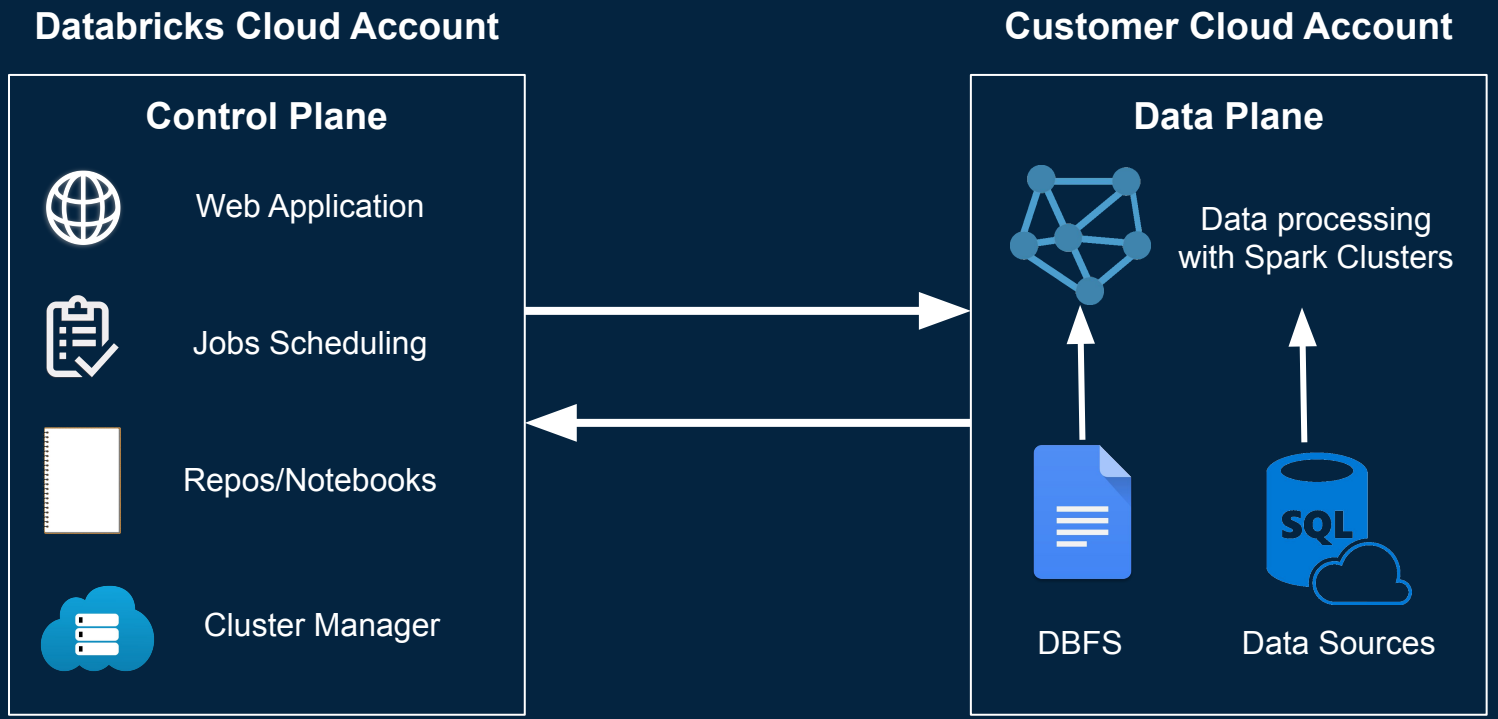
The Databricks Lakehouse platform architecture

Workspaces

Runtime Engines

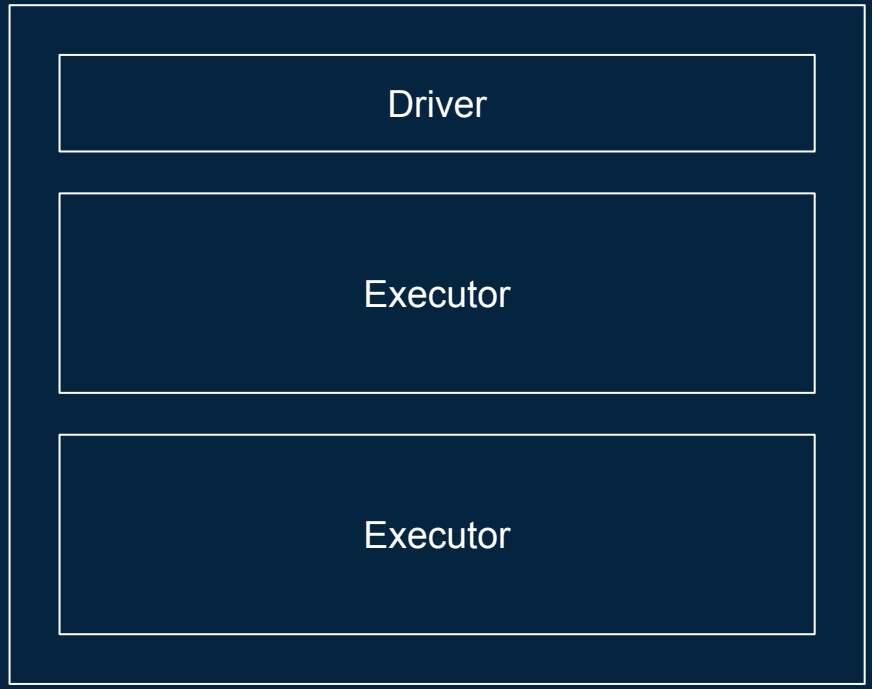
Cloud Service

The Databricks Lakehouse platform architecture



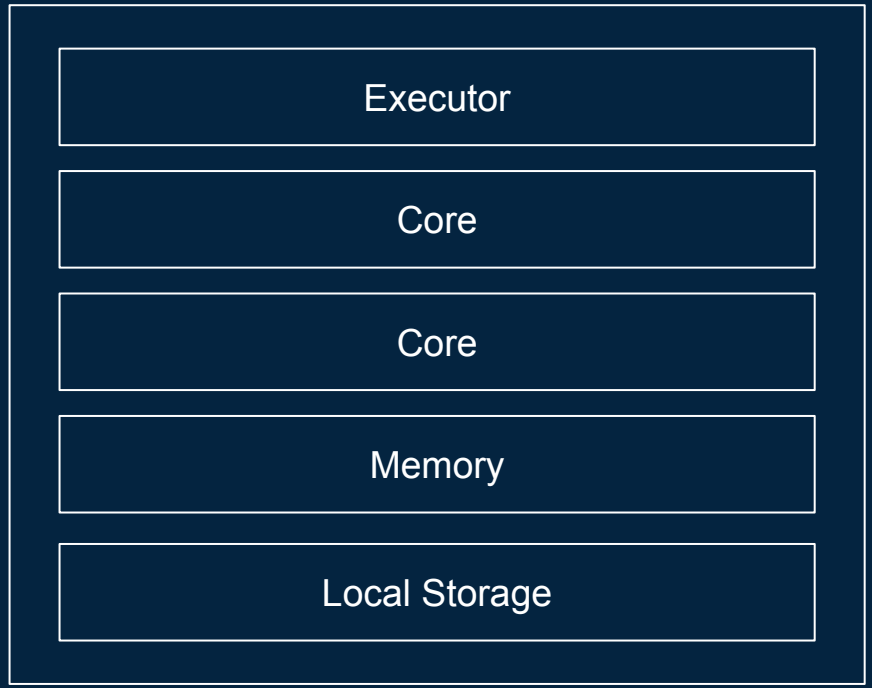
Databricks Clusters

- Databricks Spark clusters consist of one or more virtual machines
- The driver assigns tasks to executors
- The executors run tasks
- A spark job consists of one or more tasks



Databricks Clusters

- Databricks Spark clusters consist of one or more virtual machines
- The driver assigns tasks to executors
- The executors run tasks
- A spark job consists of one or more tasks



Types of clusters in Databricks

All-purpose Clusters

- Support interactive notebooks
- Can be created via the workspace or API
- Not efficient for running automated jobs

Jobs Clusters

- Support automated jobs
- Are created by the Databricks Job scheduler when running jobs