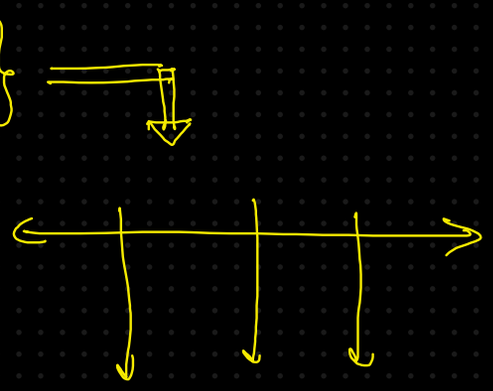
Spark → advance optimization
→ memory mgnt.



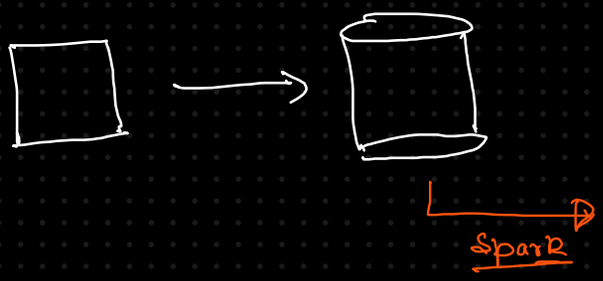
Databricks ⇒ mini project

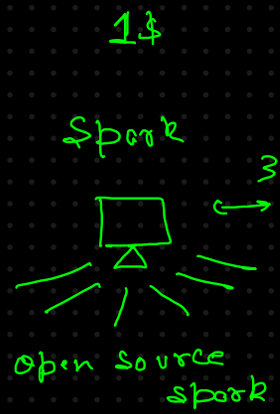
Hive ⇒

Databricks

Databricks is a cloud-based, managed data analytics platform built on Apache Spark

interactive workspace ⟵ data eng. data analyst



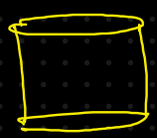


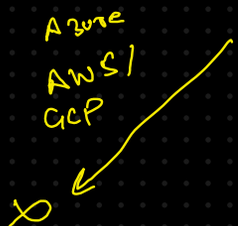
Databricks is created by inventors of Spark.

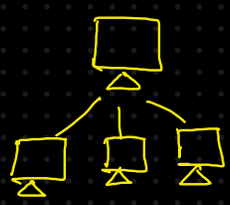
Spark ⟶ Apache foundation

↓

Databricks (paid)

Goldman Sachs ⟶  Compliance





Apache spark ⟹ performance optimization feature

Open Source Spark

Complex Infrastructure setup

Manual software install & updates

Lack of user Interface

Difficult security management

Version compatibility issue

Databricks

Fully managed cluster

Auto configured environment

Web-based workspace

Built in security & governance

Optimized spark runtime

Databricks architecture

Hardware + Software

AWS
GCP
Azure } m/c

databricks

Workspace

Notebooks

Feature	Description
Workspace	Organizes notebooks, libraries, and files.
Clusters	Manages Spark clusters for processing.
Jobs	Runs scheduled workloads and ETL tasks.
Data	Stores datasets and connects external storage.
Notebooks	Interactive coding interface (supports Python, Scala, SQL, R).

Cluster Type	Purpose
All-Purpose Cluster	Interactive work (for development, shared by multiple users).
Job Cluster	Runs specific jobs and shuts down after completion (cost-effective).
Cluster Pool	Manages pre-warmed clusters for faster job execution.

Understanding the databricks architecture

1. How computations are executed
2. How data is stored and accessed
3. How databricks differs from open source spark

High level architecture:

2 layer architecture
- Control plane : managed by databricks
- Data plane : runs inside the cloud environment (aws, gcp or azure)

* However in community edition, a free tier version is provided by databricks & its not integrated with any service

Control Plane

UI
cluster manager
Job scheduler
Notebook execution
Workspace management

Data Plane

Driver node
worker node
DBFS
Compute resources

Feature	Community Edition	Cloud-Based Databricks
Cluster Type	Single-node cluster (only a Driver)	Multi-node cluster with Worker nodes
Data Processing	Limited to small datasets	Scales for big data workloads
Storage	Uses DBFS (local filesystem)	Can integrate with AWS S3, Azure Blob, GCS
Cloud Connectivity	No cloud support	Supports AWS, GCP, Azure
Best For	Learning, small-scale testing	Production workloads, enterprise applications

Data bricks File System (DBFS)

fat32 ↔ ntfs
ext
apfs
hdfs

DBFS is a distributed file system in databricks that allows users to store, manage and interact with files

has to be enabled & then regen.

Why use DBFS ?

→ act as an abstraction layer over the cloud storage (GCS, AWS, Azure)

→ supports structured and unstructured data (csv, parquet, images)

→ integrates very nicely with spark, DB Notebook & delta lake

→ Provides a simple file system interface