



# DATABRICKS

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

It provides an interactive workspace for data engineers, data scientist and analysts to process large scale data easily.

Databricks is created by  
the inventors of spark

Why databricks?

## Open Source Spark

Complex infrastructure setup

Manual software installation and update

Lack of user Interface

Difficult security management

Version compatibility issues.

Manual COX

## Databricks

Fully managed clusters

Auto-configured environment

Web-based interface

Built-in security and governance

Optimized Spark runtime.

Self-driving COX

# Creating an account on Databricks

Databricks Community Edition

≈ 2 times verif. code  
for login

Hardware + Software  
↓  
Intel / Nvidia / AMD      ↓  
                                  chrome

AWS / GCP / Azure      Databricks

databricks  
(community  
edition)

# Understanding the databricks architecture

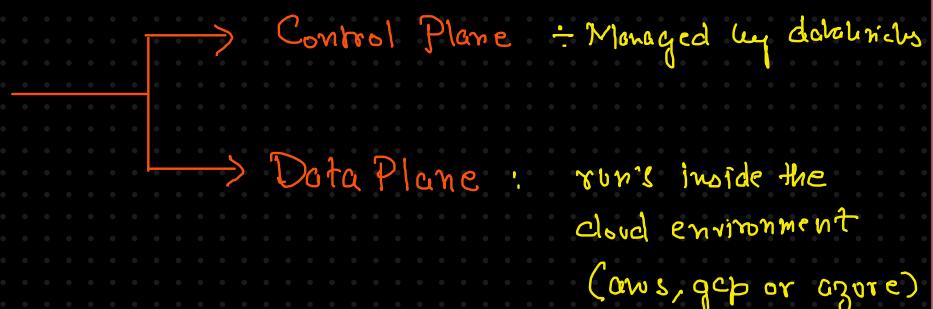
We must make sure to understand the databricks architecture as it clearly :

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

Community  
vs  
cloud

## 1. High level architecture

two layer architecture



\* However in the Community edition, a free tier version is provided by databricks & not having any integration with Cloud (gcp, aws or azure)

## Community edition →

Component	Role	Where It Runs?
Control Plane	Manages UI, job scheduling, and metadata	Hosted by Databricks
Data Plane	Runs Spark computations and data processing	Runs on a single virtual machine managed by Databricks
Databricks File System (DBFS)	Handles data storage inside Databricks	Uses a local filesystem, not cloud storage

cloud

→ run on a cloud  
→ connected to some storage

## Control Plane

UI  
Cluster manager  
Job scheduler  
Notebook execution  
Workspace management

## Data Plane

Driver node  
Worker node  
DBFS  
Computation 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

Read and Process data on Databricks cluster

Refer to the codes file of databricks

## Databricks File System (DBFS)

hdf  
ntfs  
ext  
abfs → abfile

Databricks file system (DBFS) is a distributed file system in databricks that allow users to store, manage and interact with files.

has to be enabled  
& then refresh

Why use DBFS?

- act as a abstraction layer over cloud storage (GCS, S3, Azure Blob)
- supports Structured and unstructured data (CSV, Parquet, Images)
- integrates very nicely with spark, DB notebook & delta lake
- Provides a simple file system interface

DBFS /dbfs

/Filestore

Mount points → Connect dbfs to external cloud storage

