# 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 | Databricks |
| --- | --- |
| Complex infrastructure setup | Fully managed clusters |
| Manual software installation and update | Auto-configured environment |
| Lack of user Interface | Web-based interface |
| Difficult security management | Built in security and governance |
| version compatibility issues. | optimized Spark runtime. |
| manual car | self-driving car |

# Creating an account on Databricks

Databricks Community Edition

Hardware + Software

Intel / Nvidia / AMD

chrome

≈ 2 times verif. code for login

AWS / GCP / Azure

Databricks

databricks (community edition)

Databricks

# Understanding the databricks architecture

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

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

- → Control Plane :- Managed by databricks
- → Data Plane : runs inside the cloud environment (aws, gcp or azure)

* 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
- 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 |

# Read and Process data on Databricks cluster

Refer to the codes file of databricks

# Databricks File System (DBFS)

hdfs
ntfs
ext
apfs → apple

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, parque, 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