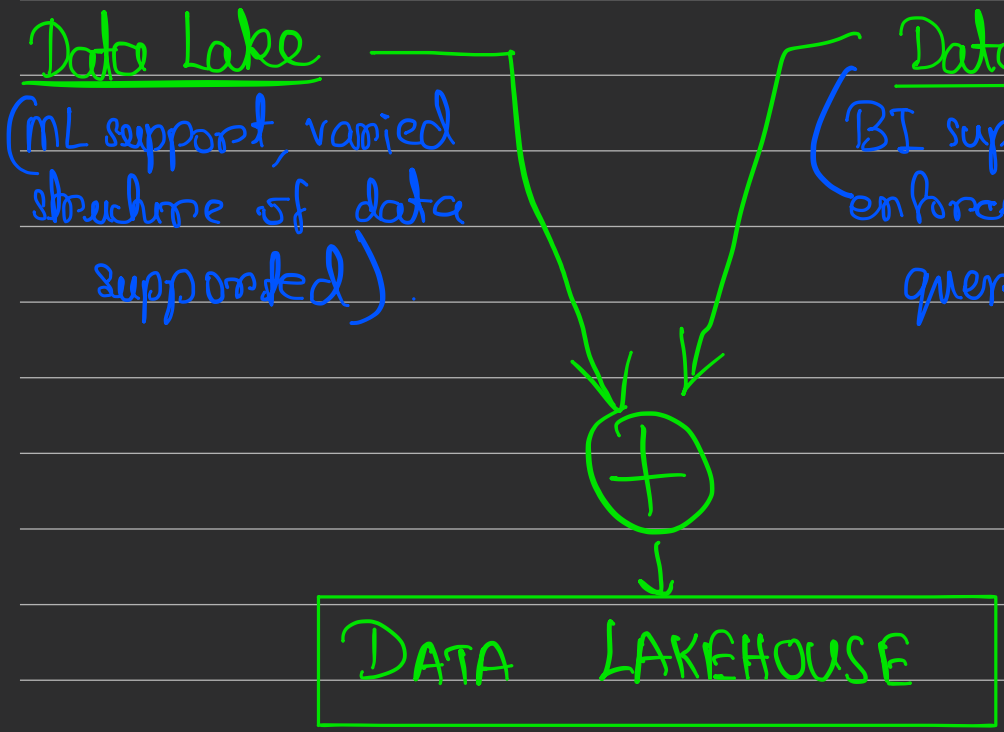
≥ Databricks provided 1st Lakehouse Platform: One simple platform to unify all the data, analytics and AI workloads.

≥ For following different needs :-
Data Warehousing, Data Engg., Streaming, Data Science/ML different technologies are used. Also, data is stored in them in variety of proprietary formats. Transferring data b/w them becomes resource intensive and costly.
Due to multiple formats, there becomes multiple copies of data, which causes inconsistency, and performance slower and management tougher.



≥ At the heart of Data Lakehouse Platform is the DELTA LAKE FORMAT.

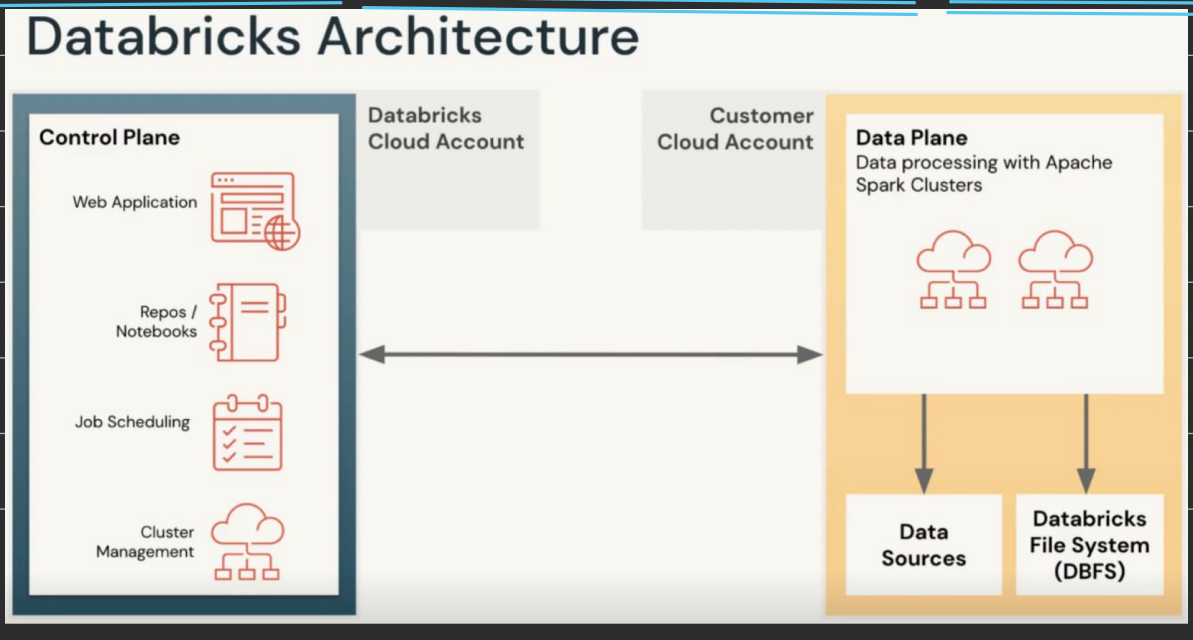
DELTA LAKE provides the ability to build curated Data Lakes that provides reliability, performance and governance like Data warehouse, such as supporting ACID transactions directly on Data Lake.

- Delta Lake uses advanced caching & indexing, to make querying of data on data lakes faster.
- Delta Lake supports fine-grained Access Control Lists.

Features of Data Lakehouse :-

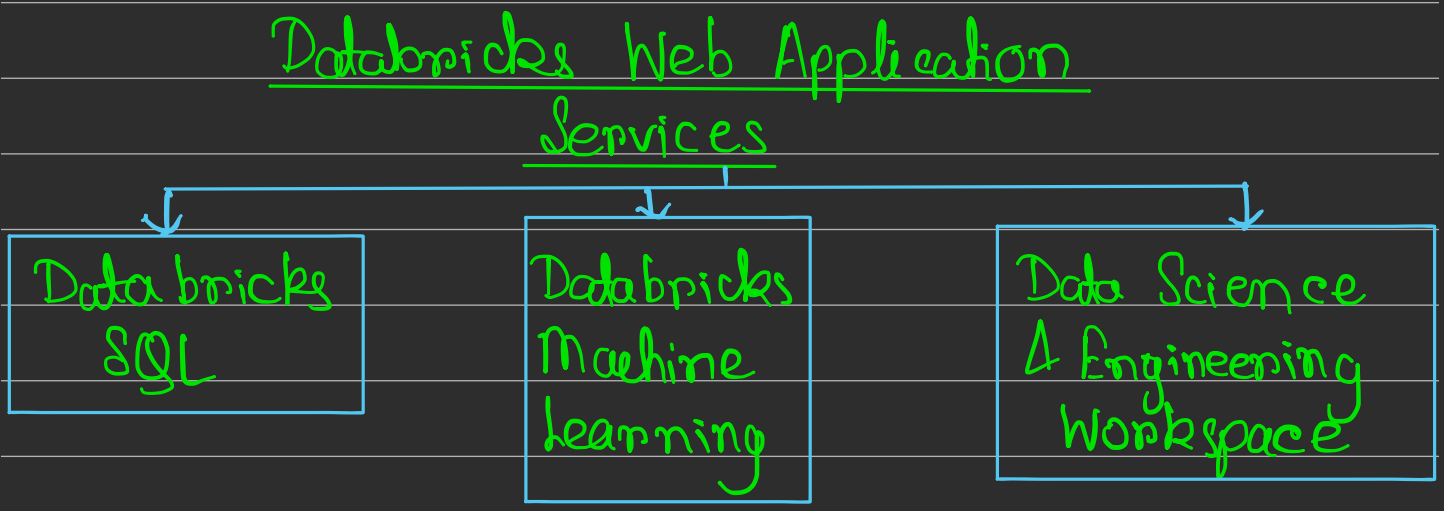
- Simple :- Data exists at one place in one format, to support all workloads.
- Open-Source :- Can support all-proprietary formats.
- Collaborative.

Databricks Architecture / Services :-



⇒ In Control Plane, there is encryption at rest.

⇒ All our data is processed in the Data Plane. All compute resources in the Data Plane resides in the customers' own cloud account (ex:- Clusters)



Repo: Collection of hosted notebooks.

Jobs Orchestration Components → Delta Live Tables
→ Jobs.

Clusters:- Made of one or more VMs.

- Drivers: coordinates activity of executors.
- Executors: run tasks composed of Spark Jobs.

⇒ Types of Clusters :-

1) All-purpose clusters: Analyze data collaboratively using

interactive notebooks.
Created from the workspace or API.

2) Job Clusters : Runs automated jobs.
Created by Databricks Job Scheduler when running jobs.

- Retains upto 30 clusters
- Ensures running of jobs in isolated env.

DBFS (Databricks File System) :-

A distributed, persistent file system available to all of the clusters.

Jobs :-

Allows us to execute tasks on-demand, or based on schedule.

Cluster Management Options :-

1) Terminate :- Stops the cloud resources in use, and cleans the cache, but the config. of cluster remains.

2) Restart :- Clears all the cache and boots the resources afresh. Any output needed, must be stored at a persistent location (ex- DBFS), so that it is available later.

3) Delete :- Stops cluster, frees up the resources, and deletes the cluster configuration.