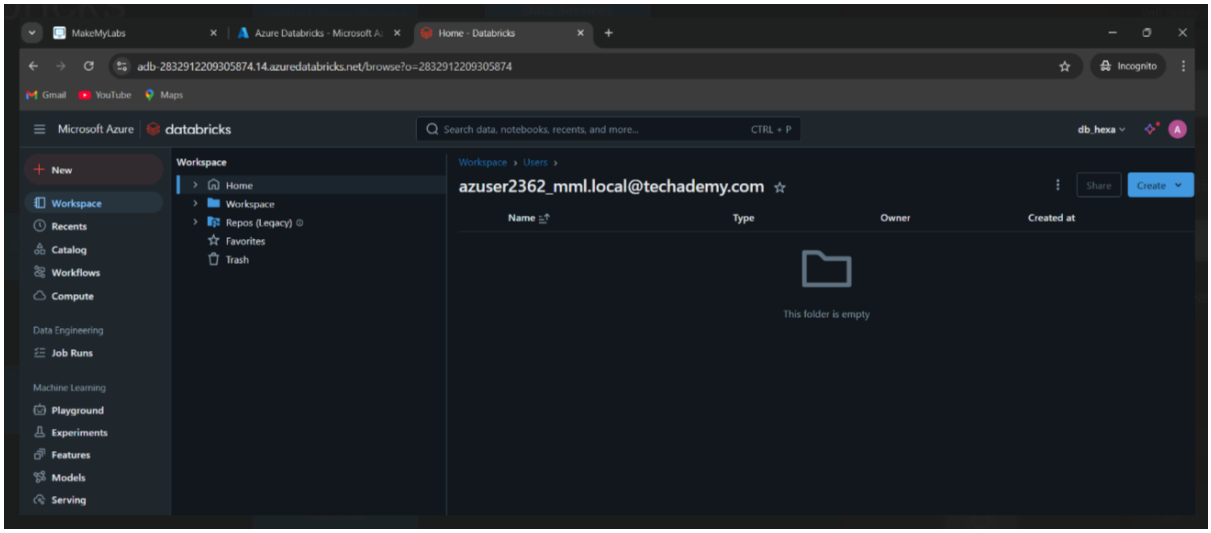
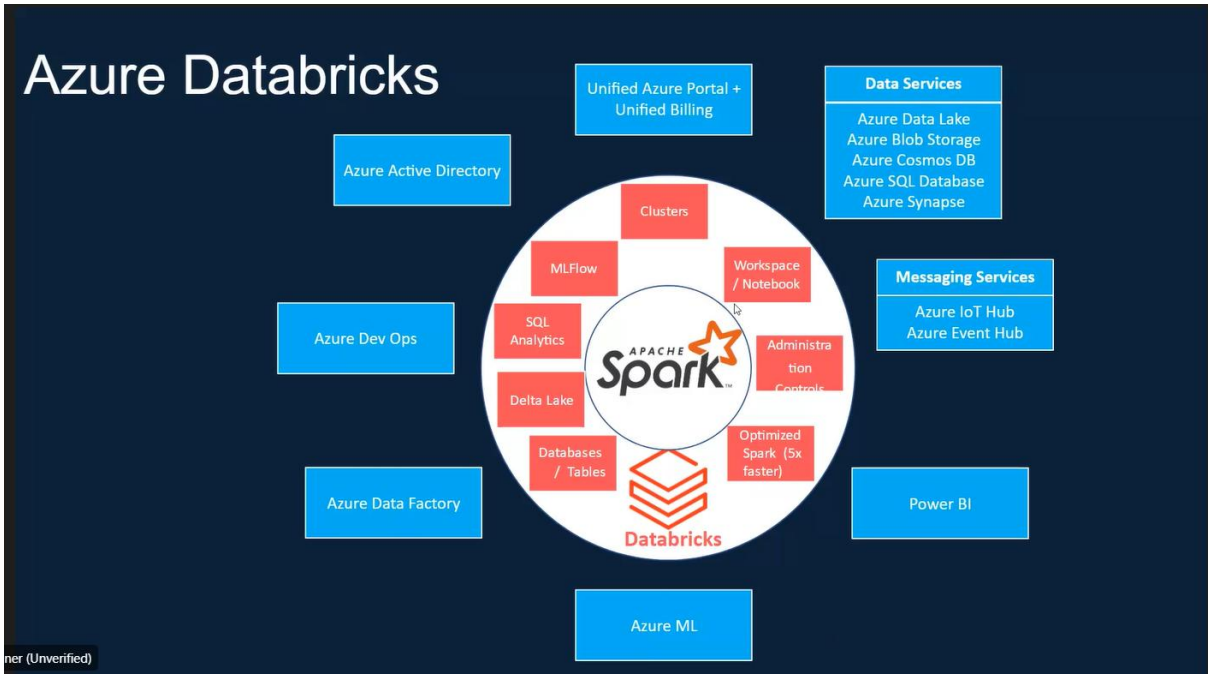
Azure Databricks Day1

Workspace creation





Cluster -→Group of virtual machines

- 1.Driver(VM)
- 2.Workers(VM)

When a code is written in notebook,it is given to driver.

Driver has the info of all worker machines and it allocates the work accordingly.

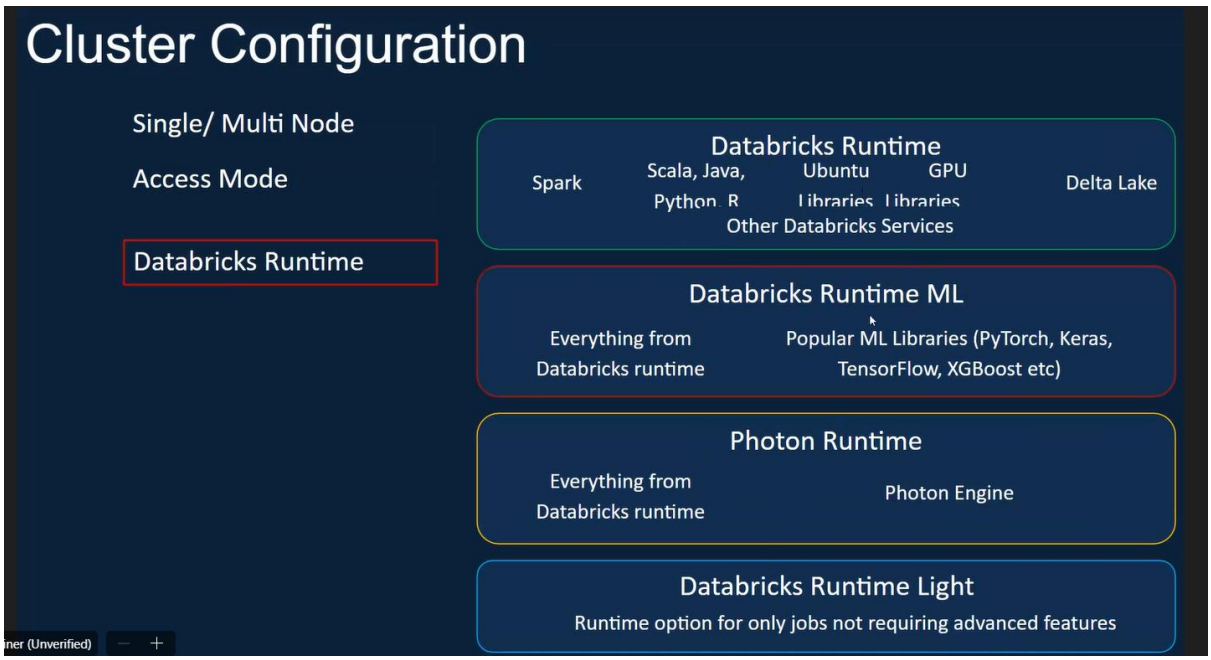
Cluster Types

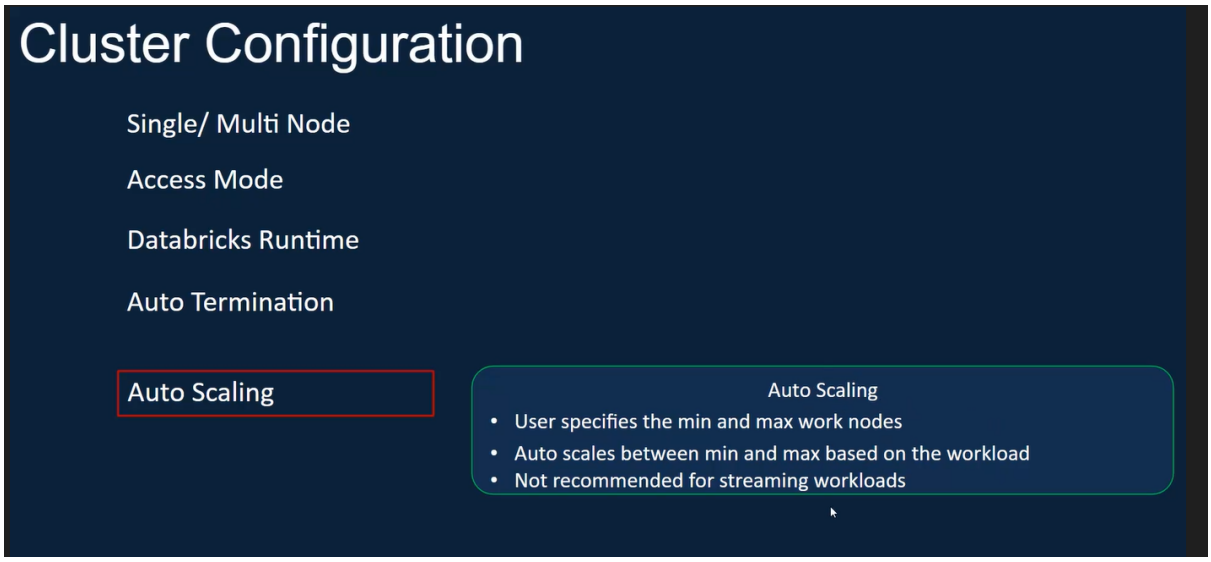
1.All purposes

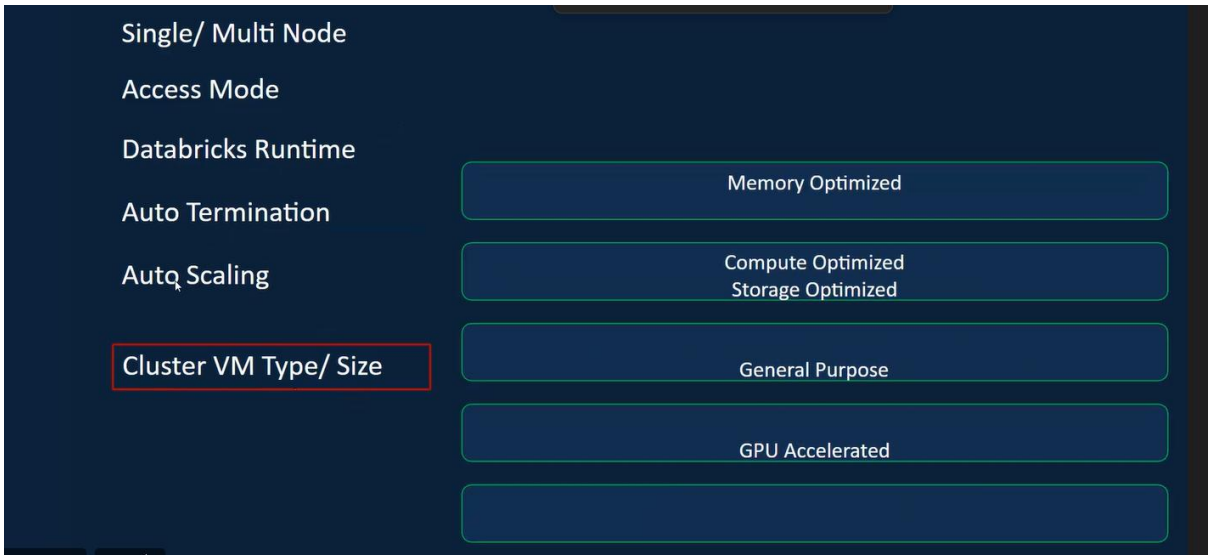
- Created manually

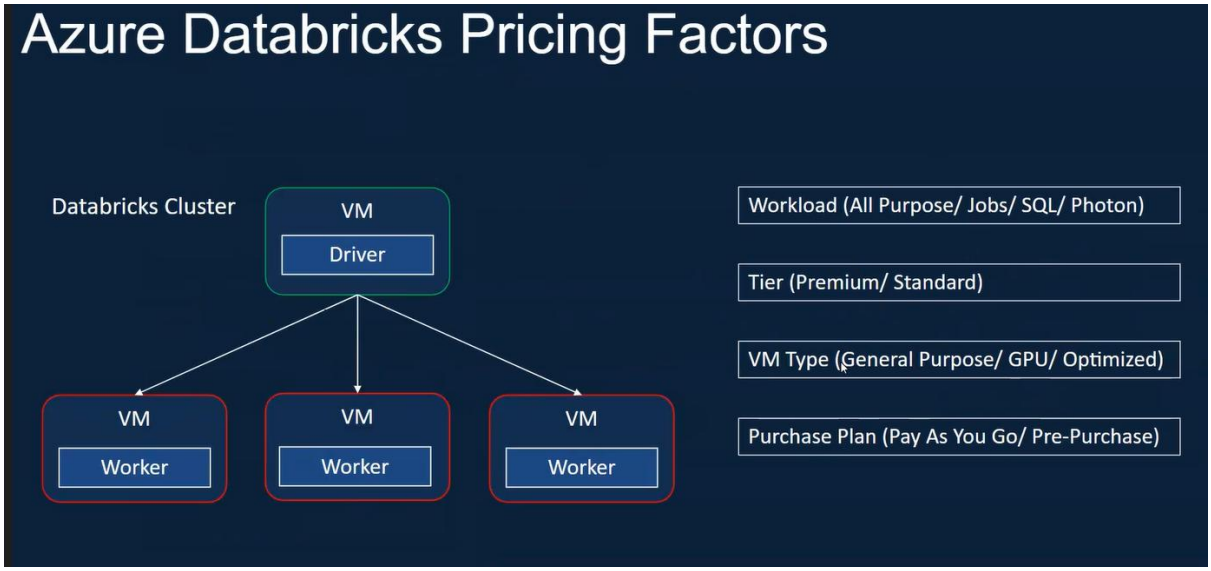
2.Job Cluster

- Created by jobs









DBU(Databricks Unit)

It is a normalized unit of processing power on the Databricks lakehouse platform used for measurement and pricing purposes

- A database is a collection of data or information. Databases are typically accessed electronically and are used to support Online Transaction Processing (OLTP).
    - ACID (Atomicity, Consistency, Isolation, Durability) transactions to ensure data integrity.
- Data Base examples:
    - Relational databases
    - Document databases
    - Key-value databases
    - Wide-column stores
    - Graph databases
- A data warehouse is a system that stores highly structured information from various sources.
- A data lake is a repository of data from disparate sources that is stored in its original, raw format.
- Data lake characteristics
    - Data lakes store large amounts of structured, semi-structured, and unstructured data.
    - They can contain everything from relational data to JSON documents to PDFs to audio files.
    - Data does not need to be transformed in order to be added to the data lake, which means data can be added (or "ingested") incredibly efficiently without upfront planning.

	Database	Data Lake	Data Warehouse
Workloads	Operational and transactional	Analytical	Analytical
Data Type	Structured or semi-structured	Structured, semi-structured, and/or unstructured	Structured and/or semi-structur
Schema Flexibility	Rigid or flexible schema depending on database type	No schema definition required for ingest (schema on read)	Pre-defined and fixed schema definition for ingest (schema on and read)
Data Freshness	Real time	May not be up-to-date based on frequency of ETL processes	May not be up-to-date based on frequency of ETL processes
Users	Application developers	Business analysts, application developers, and data scientists	Business analysts and data scien