**AZURE DATABRICKS**

**What are Databricks?**

Azure Databricks is unified, open analytics platform for building and deploying, sharing, and maintaining enterprise- grade data, analytic AI solution scale.

**Databricks uses:**

Our customers use Databricks to process, store, clean, share, analyze, model, and monetize their data sets with solution from BI to machine learning.

**How does Databricks work with azure:**

Databricks is cloud-based data engineering tool that is widely used by companies to process and transform large quantity of data and explore the data.

**Data Lakehouse:**

The Databricks Lakehouse combines the ACID transactions and governance of enterprise data warehouse with flexibility and cost -efficiency of the data lake to enable business intelligence (BI) and machine learning (MI)on all data.

**ACID (Atomicity, Consistency, Isolation, Durability):**

**Atomicity:** atomicity means that all transactions either succeed or fail completely.

**Consistency:** it is guarantees related to how a given state of the data is observed by simultaneous operations.

**Isolation**: it is referring how to simultaneous operations potentially conflict with on another.

**Durability**: it is means that committed changes are permanent.

**Components of Databricks Lakehouse**:

They are 2 primary components.

1.Delta tables

2. Unity catalog

1.**Delta tables:**

Tables created on Azure Databricks use the Delta Lake protocol by default.

* Metadata used to reference the table is added to the meta store in the declared schema or database.
* Data and table metadata are saved to a directory in cloud object storage.

2.**Unity Catalog:**

Unity Catalog unifies data governance and discovery on Azure Databricks. Available in notebooks, jobs, Delta Live Tables, and Databricks SQL, Unity Catalog provides features and UIs that enable workloads and users designed for both data lakes and data warehouses.

* Account-level management of the Unity Catalog meta store means databases, data objects.

**What data objects are Databricks Lakehouse:**

The data bricks Lakehouse architecture combine data stored with the delta lake protocol in cloud object storage with metadata registered to a metastore. they are 5 primary objects in the Databricks Lakehouse.

1. **Catalog**: a group of databases.
2. **Database**: a group of objects in a catalog. databases contain tables, views, and functions.
3. **Table**: a collection of row and columns stored as data files in object storage.
4. **View**: a saved query typically against one or more tables or data sources.
5. **Function**: saved logic that returns a scalar value or set of rows.

**Metastore:**

The metastore contains all the metadata that defines data objects in the Lakehouse.

Azure Databricks metastore options.

1. **Unity catalog metastore**: unity catalog provides centralized access control, auditing, lineage, and data discovery capabilities.
2. **Built-in hive metasore**: each azure Databricks workspace includes a built- in hive metastore as a managed service. It provides a less centralized data governance model than unity catalog.
3. **External hive metastore**: it brings your own metastore to azure Databricks. its cluster can connect to existing external Apache hive metastores.

**Catalog:**

A catalog is a group of databases. And schema contains table and views

**Database:**

A database is a collection of data objects, such as tables or views and functions.

Databases will always be associated with a location on cloud object storage.

**Tables**:

It is collection of structured data.

**Managed table**:

Azure Databricks managed both the metadata and the data for a managed table.

**Functions**:

Functions allow you to associate user- defined logic with a database. Functions can return either scalar values or sets of rows. Functions are used to aggregate data.

**Data lake:**

Is a centralized repository design to store, process and secure large amount of structure, semi structured and unstructured data.

**Data Ingestion:**

the process of data importing, loading and processing data for later use or storage in a database.

**workflows:**

workflows can help streamline and transferring, automate repeatable business tasks, minimizing room for errors and increasing overall efficiency.

**delta live table:**

is a framework that simplifies ETL and streaming data processing

**notebooks:**

A notebook is a web-based interface to documents containing a series of runnable cells(commands) that operate on files and tables, visualizations, and narrative text.

**jobs:**

jobs are one mechanism for running code in azure Databricks.

**Delta Lake:**

Delta lake is designed to work with Apache spark.

**Data Governance:**

It is used to secure purpose, is everything do to ensure data and secure, private, available, and accurate and usable.

**ACL (access control list):**

A list of permissions attached to the workspace cluster, job, table or experiment.