**Sunkari Manohar**

**Assessment 7 – 05/01/2024**

**Azure Databricks**

**Lakehouse:**

Gives unified workspace, that combines the best of Data Warehouse (stores data in structured form) and Data Lake (stores data in unstructured form)

**Key Features of lakehouse:**

* ACID transactions,
* Schema enforcement
* Version control
* Unified metadata management
* Scalability & Performance
* Flexibility

**Architectural components of lake house:**

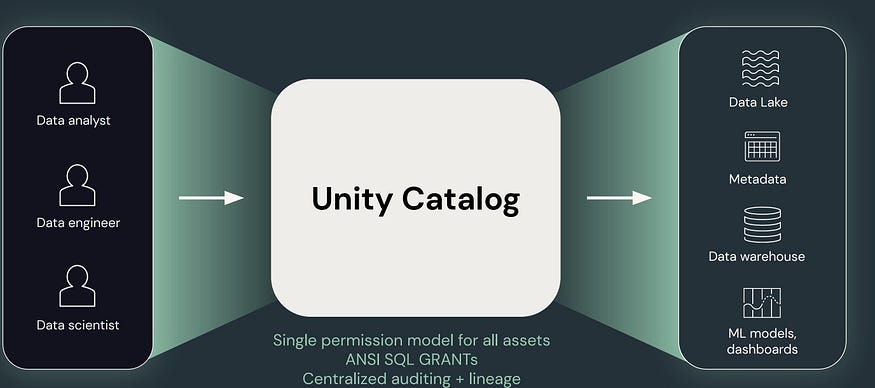
**Delta lake:**

Its is a optimized storage layer which can manage various file formats and provides features like optimizing data retrieval,transactional log which enables acid transactions and allows for data versioining and rollbacks.

**Unity catalog:**

Provides Unified governance for all the things like your files,tables etc present in your lakehouse on any cloud.

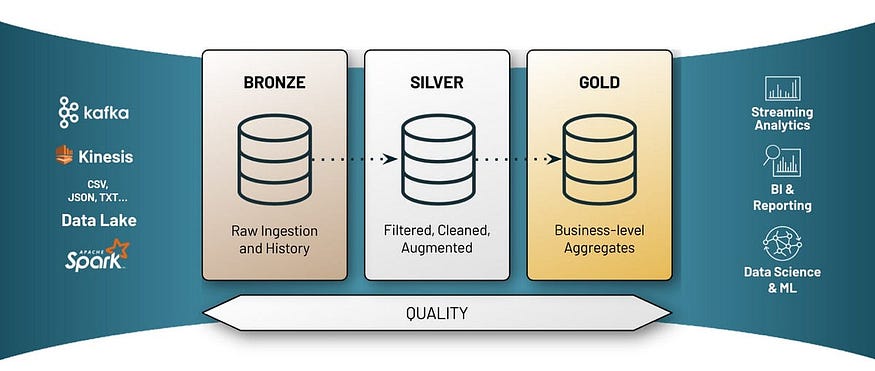
One of the key benefits of using the Unity Catalog in a lakehouse architecture is that it provides a single point of access to all data assets, regardless of where they are stored. This means that users can easily discover and access data assets that are stored in different locations, such as cloud object store.



## **Medallion Lakehouse Architecture:**

In medallion Lakehouse Architecture, the source data is passed through multiple Data ingestion layer before it is processed and made available to the downstream team. The three Data Ingestion Layers are:

* **Bronze Layer:**This layer stores the data in it original format(Raw data) received from multiple sources. It’s used by an organization for audit purpose and to trace back to the data source.
* **Silver Layer:** Silver layer is responsible for cleaning and filtering the Bronze data. It handles the missing data and also does the data type conversation if required.
* **Gold Layer:** This layer contains the transformed, aggregated and modeled data which can be made available for SQL queries.



**Delta live tables(DTL):**

Delta Live Tables or DLT is one of the best way to do ETL process in Lakehouse. It’s responsible for creating, maintaining and testing the data flow pipelines using a declarative approach, which helps data engineer to focus more on the getting the value out of data rather than focusing on tools.

Delta live tables is solely responsible for performing data transformation and managing the task orchestration, cluster management, monitoring, data quality, and error handling.

**Key benefits:**

Accelerates the etl process

Automatically manages your infrastructure

Ensure high data quality

Unify batch and streaming.

**Azure DataFactory:**

Azure Data Factory is a cloud-based data integration service that allows you to create data-driven workflows in the cloud for orchestrating (plan,coordinate)and automating data movement and data transformation.

It allows you to create data pipelines that move and transform data and run them on specified schedule.

ADF pipelines performs 3 steps:

**Step1: Connect and collect-**Connect to all data sources and move them to a location by using a copy activity in a datapipeline to move data from datastores to centralized datasore in cloud .

**Step2: Transform and enrich** - Once it is in centralized datastore transformations can be done

**Ste3: publish**-After transforming deliver the data to another cloud storage or area.

**Key components:**

**Dataset:** Represents datastructures within the datastore,it is the form of the data or files given as input or transformed and sent as output,like blob files,azure sql tables ec.

**Pipeline:** Is a group of activities

**Activities** : Define the actions to be performed .ADF does 2 actions data movement ,transformation

**Linked service:** Like connection strings, define the connection information needed for the service to connect to external resources.

**Copy activity in ADF:**

Copy activity is used to move the data from source to destination/sink using copy tool.

After creating ADF and launching it we get redirected to its workspace

For this activity to perform there needs to be 2 storage accounts or any storages.

We copy files from one to another

Step 1: Start the copy data Tool

Step 2: Complete source configuration

Step 3: Complete destination configuration

Step 4: Review all settings and deployment

Step 5: Monitor the running results