**AZURE DATABRICKS**

****Dataricks****

**Azure Databricks** is a **cloud-based analytics platform** optimized for **big data processing, machine learning, and AI**, built on **Apache Spark**. It enables **scalable data engineering, real-time analytics, and collaborative data science** with support for Python, Scala, SQL, and R.

**Why**   
  
It provides a **scalable, high-performance** environment for **big data processing, machine learning, and real-time analytics** with optimized Apache Spark.

**Compare with Synapse & Data Factory**

Data Factory : Use ADF for data movement & orchestration.

Synapse Analytics : Use Synapse for SQL-based analytics & data warehousing.

Databricks : Use Databricks for big data, machine learning, and advanced analytics.

**Components of Databricks**

Workspace : Acts as a central place for managing notebooks, libraries, and datasets.

Clusters : Managed **Apache Spark clusters** that provide computing resources.

Notebooks : Interactive notebooks for writing and executing **Python, Scala, SQL, and R** code.

Jobs : Used to **schedule and automate** workflows (ETL jobs, machine learning training, etc.).

Delta Lake : A **storage layer** that provides ACID transactions, versioning, and schema enforcement.

MLflow : Open-source platform for **tracking experiments, deploying models, and managing machine learning lifecycle**.

Databricks SQL: Optimized for running **SQL queries** on large datasets.Supports **BI dashboards and reporting**.

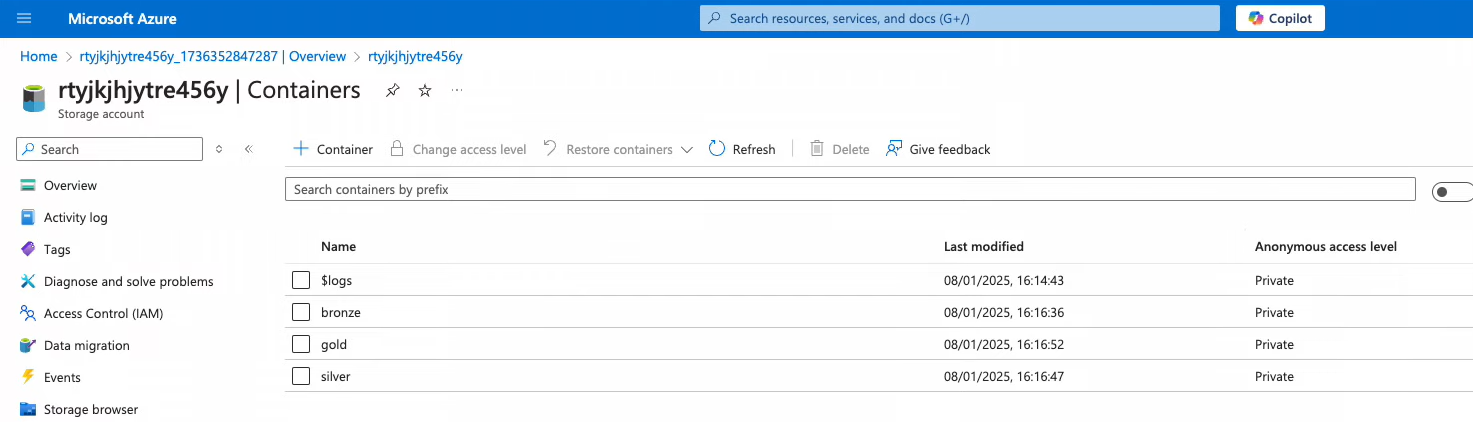
**Workflow**

Before starting workflow we have to create the prerequisites.

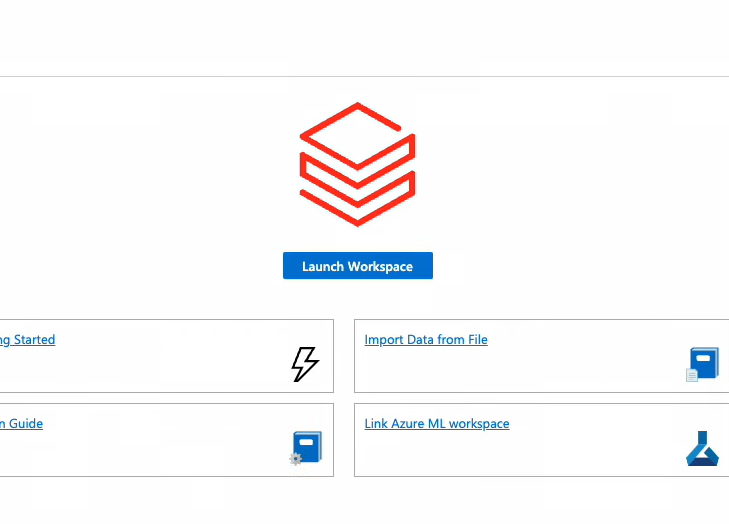
Creating the storage account and containers :

\* Create a separate resource group for Databricks Project in Azure home page.

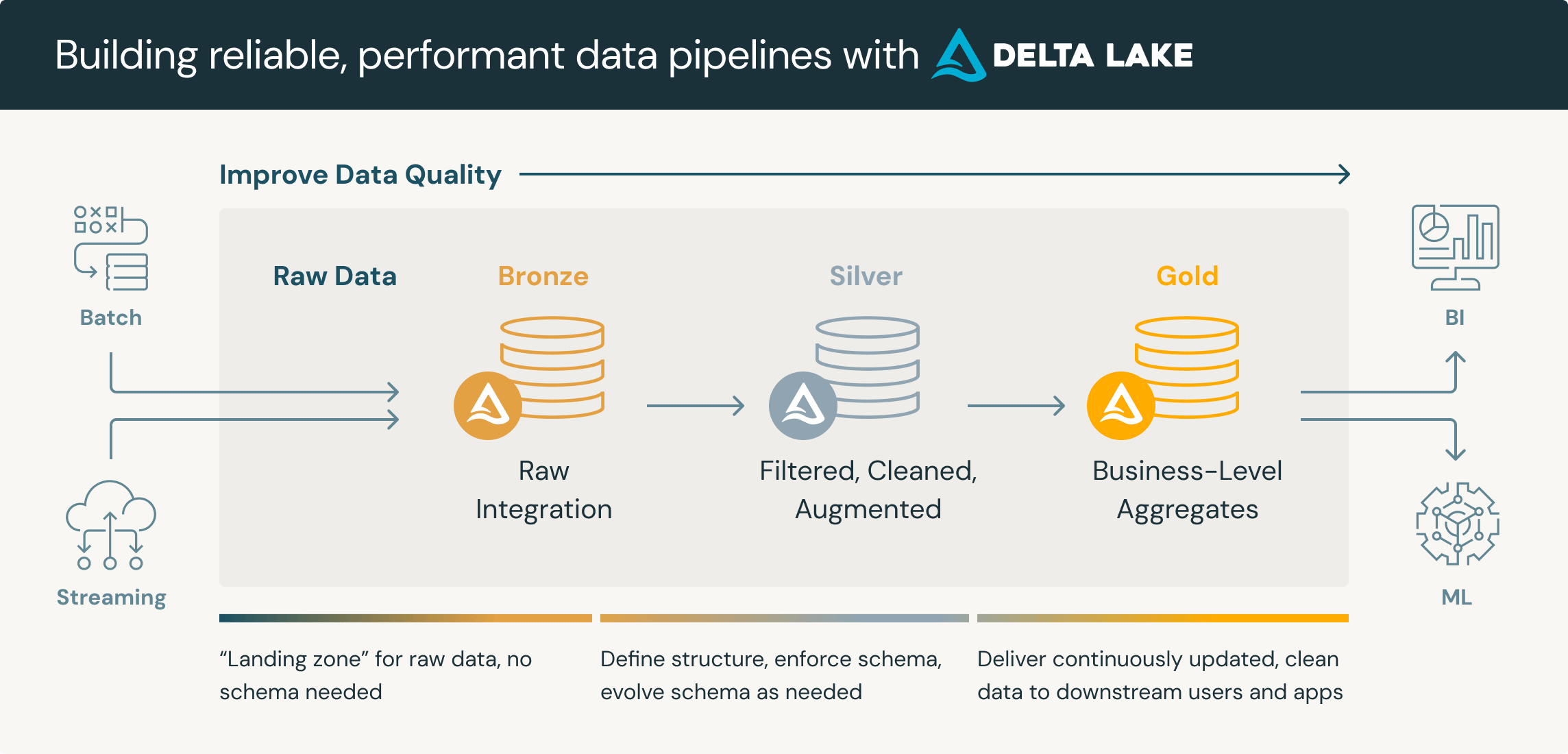
\* In Resource create storage account and create bronze,silver and gold containers.



\* Create a databricks workspace in the same resource group and launch the workspace.

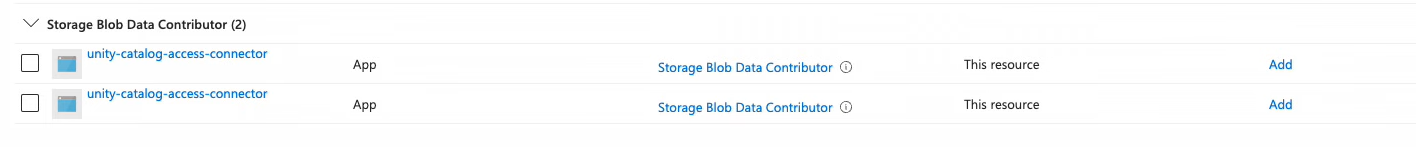


\* Here we are using the medalion architecture for the workflow.



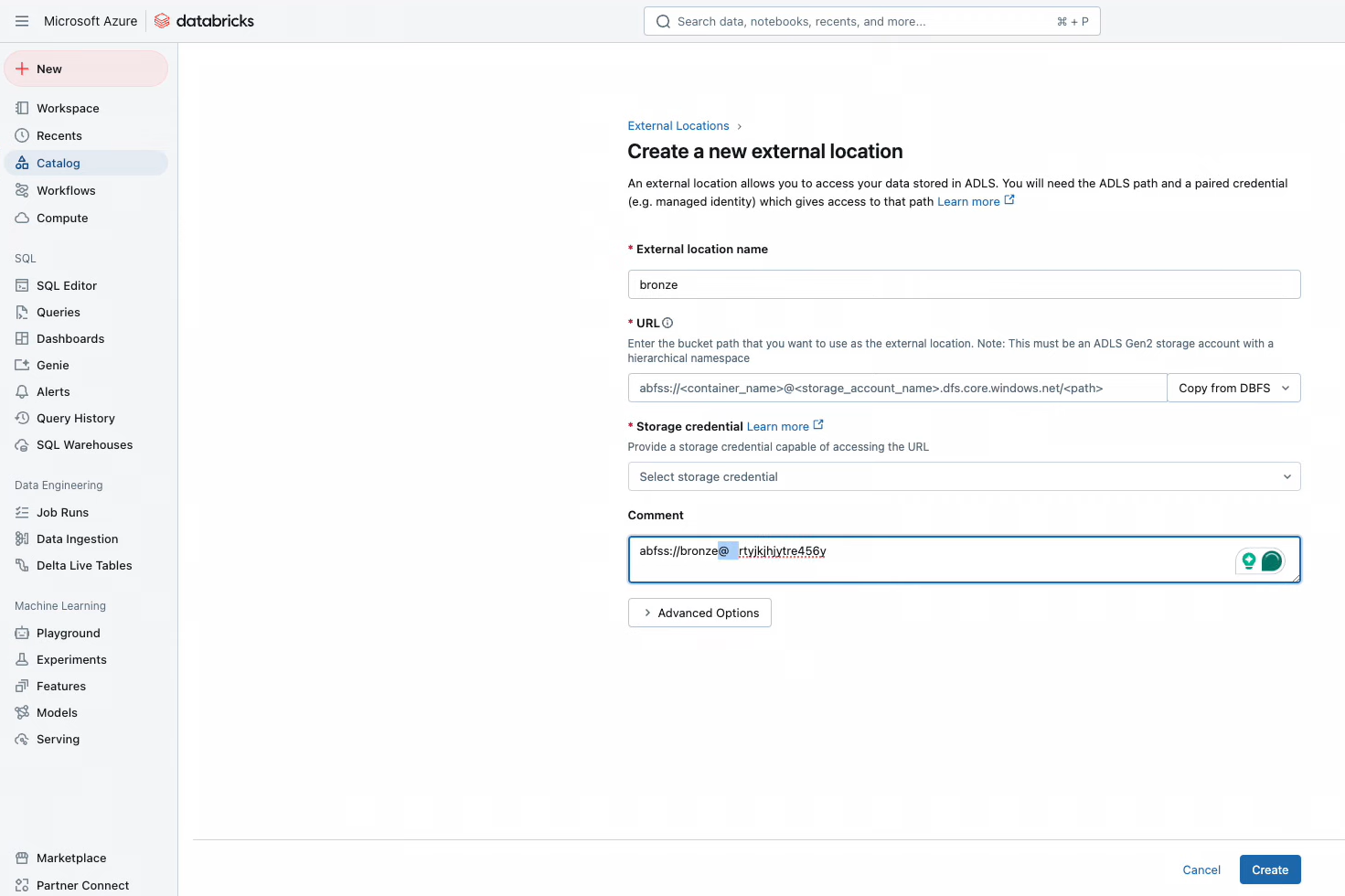
**Step 1:**

Before launching the workspace, we need to manage the identity for Databricks to connect to the storage account. To do this, open the storage account, navigate to **Access Control (IAM)**, click **Add role**, select **Storage Blob Data Contributor**, and then choose your Databricks workspace name.

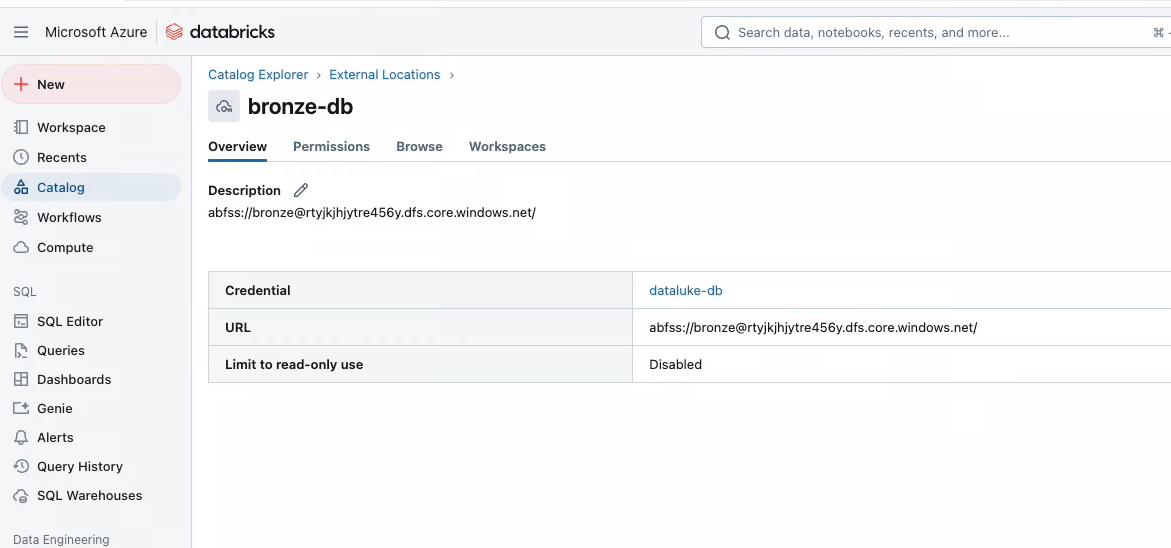


**Step 2 :**

Then open the databricks where we have created in resource group and launch it. Then we have to create the cluster for running the pipelines or notebooks in the databricks. Then in databricks we have create a workspace and created the three notebooks bronze silver and gold where we have to do the transfromations for it.before the we need to create the unity catalog access connector.this will be useful for connected external sources into the databricks. While creating the external source we need to give the credentials of the storage account and containers.

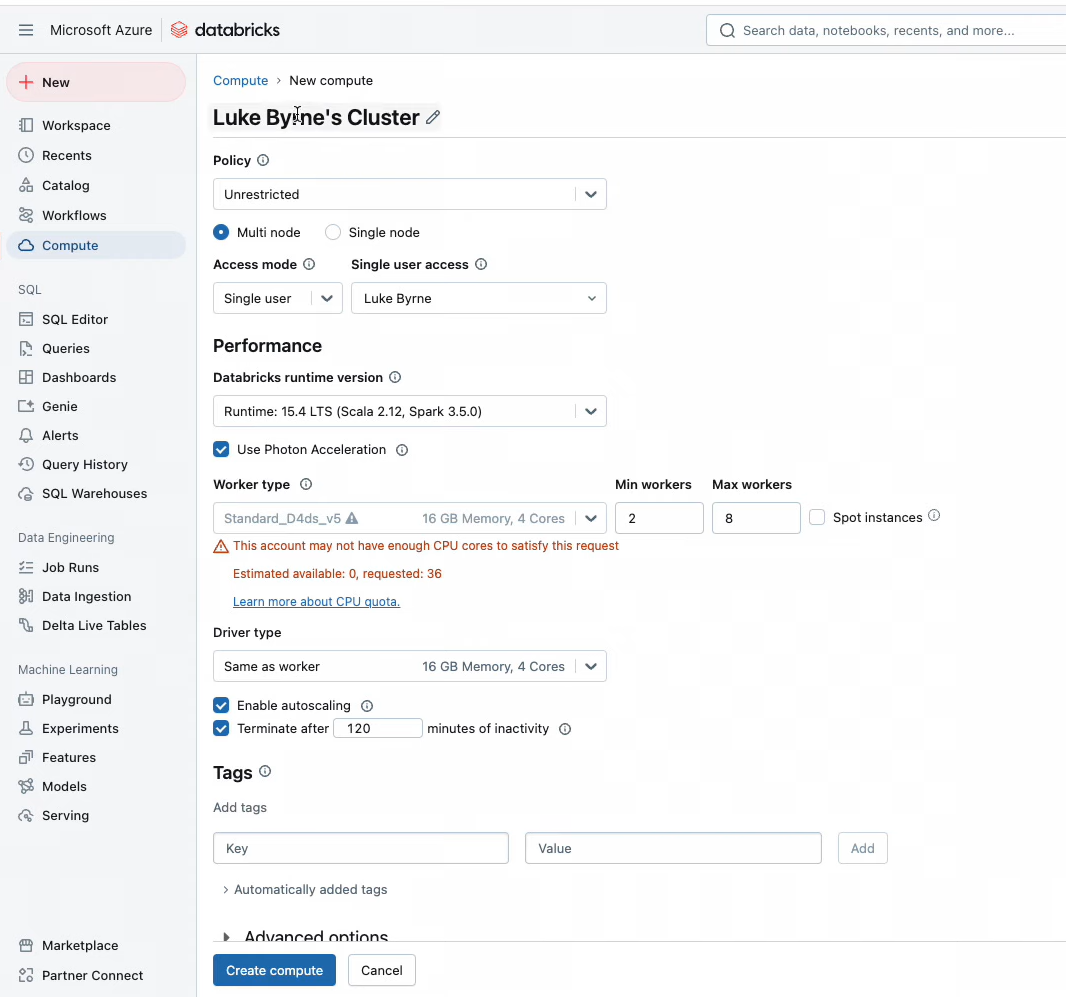


Same process execute for silver and gold creating a new external resource from the storage account credentials.It will show like below after creating the external resource in databricks. The external source is there in **Catalog tab.**



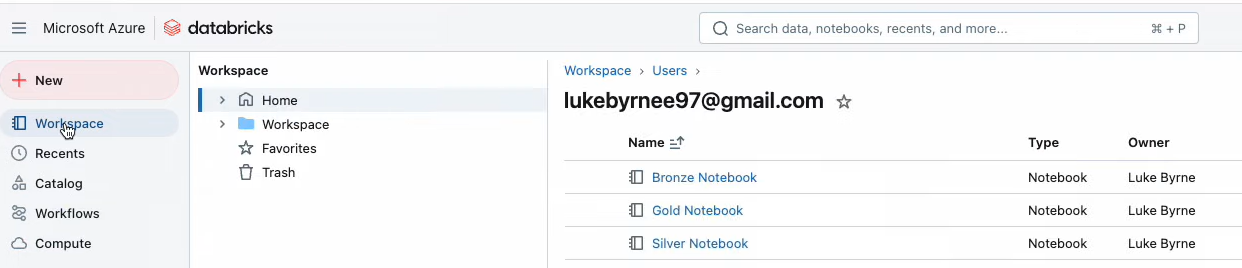
**Step 3 :**

Creating a cluster is essential for running notebooks in Databricks. To do this, navigate to the **Compute** **tab**, where you will find the option to create a cluster. Once created, the tab will display the cluster details. Additionally, there is an option to automatically terminate the cluster when it is not in use. Once it will created we’ll use it in the notebooks to run.

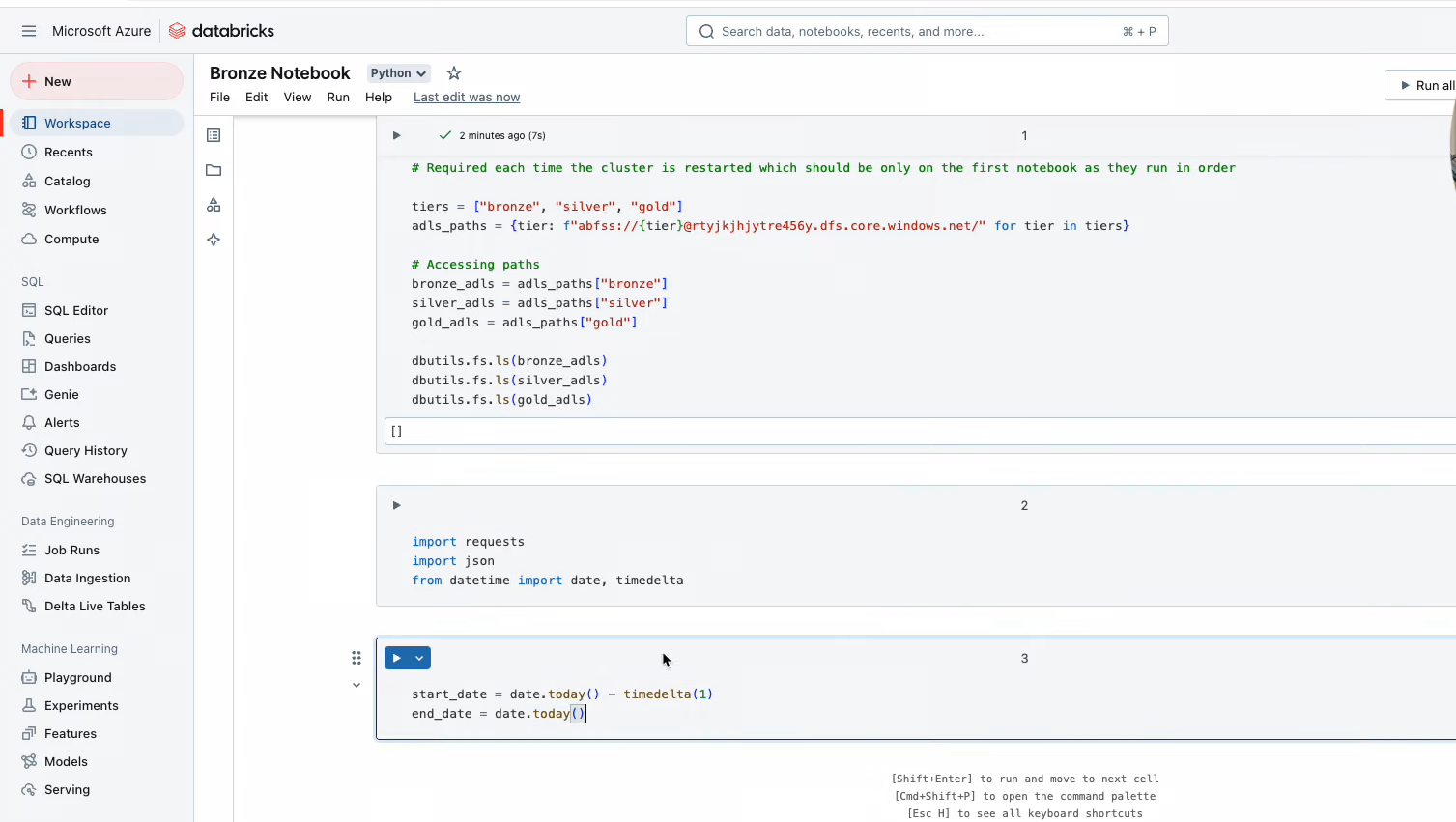


**Step 4 :**

After creating the clusters go to the **Workspace tab** and create the bronze, silver and gold notebooks for the transformation.

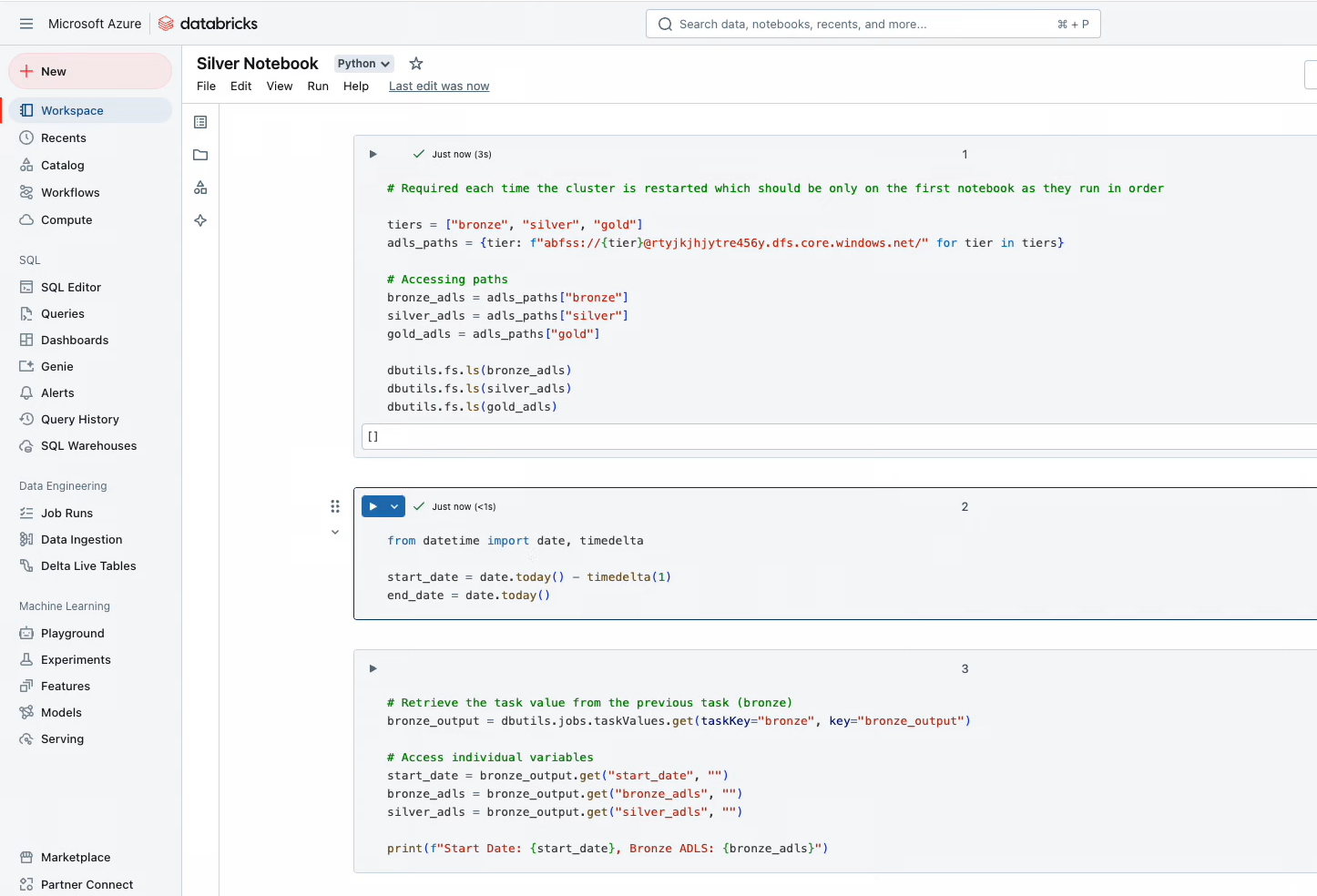


Then open the bronze notebook give the credentials for importing data into bronze container by using API link and Storage account credentials so that it can able to understand the source and destination and fetch the data easily.

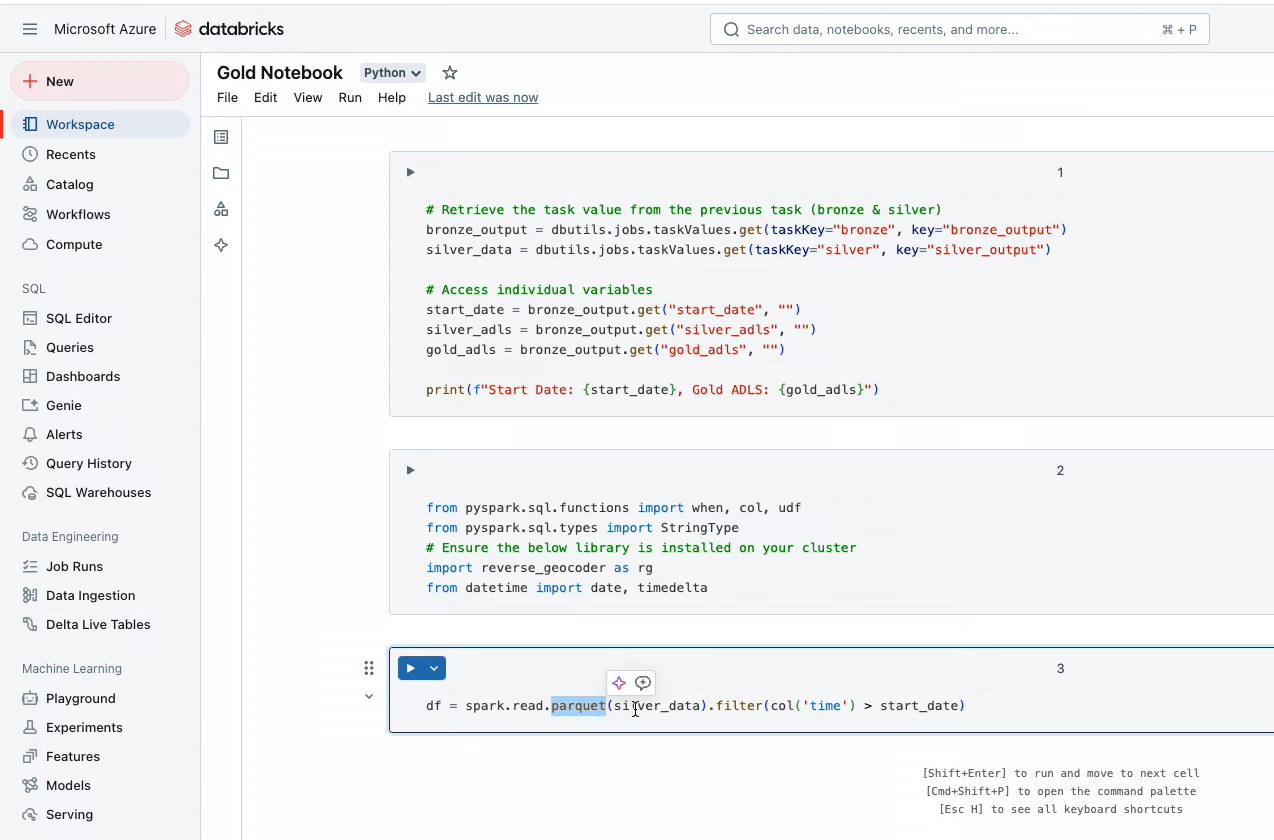


The code we wrote just for importing data by using API and give the destination to storage account bronze container.

Then open Silver container and do the major transformation for the dataset and send the transformed data to the bronze container to silver container.

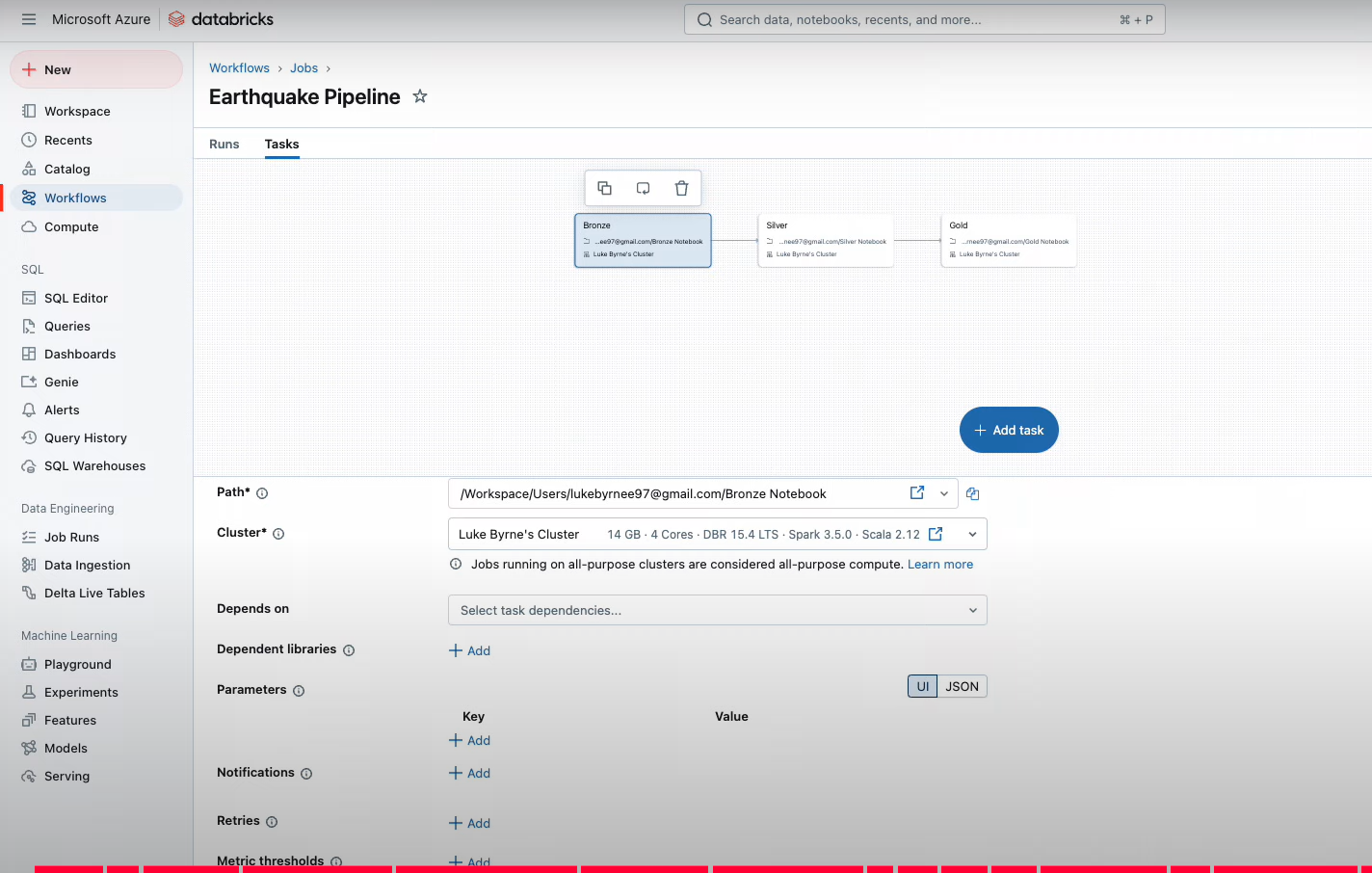


Then open the gold container and do the scheduled transformation in it and send data from silver container to gold container.



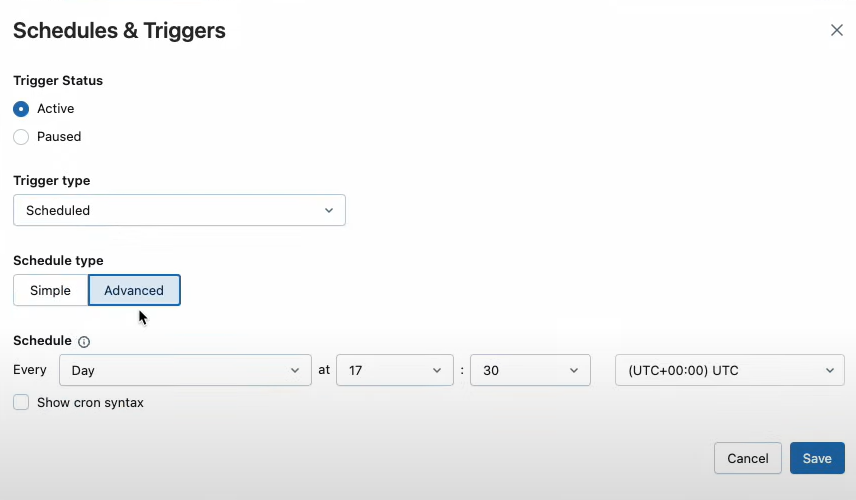
**Step 5 :**

After that we have to create a pipeline to automatically run the workflow we can create this pipelines in **Workflows tab**. We have to create three pipelines and give the notebook credentials to it as shown in below .



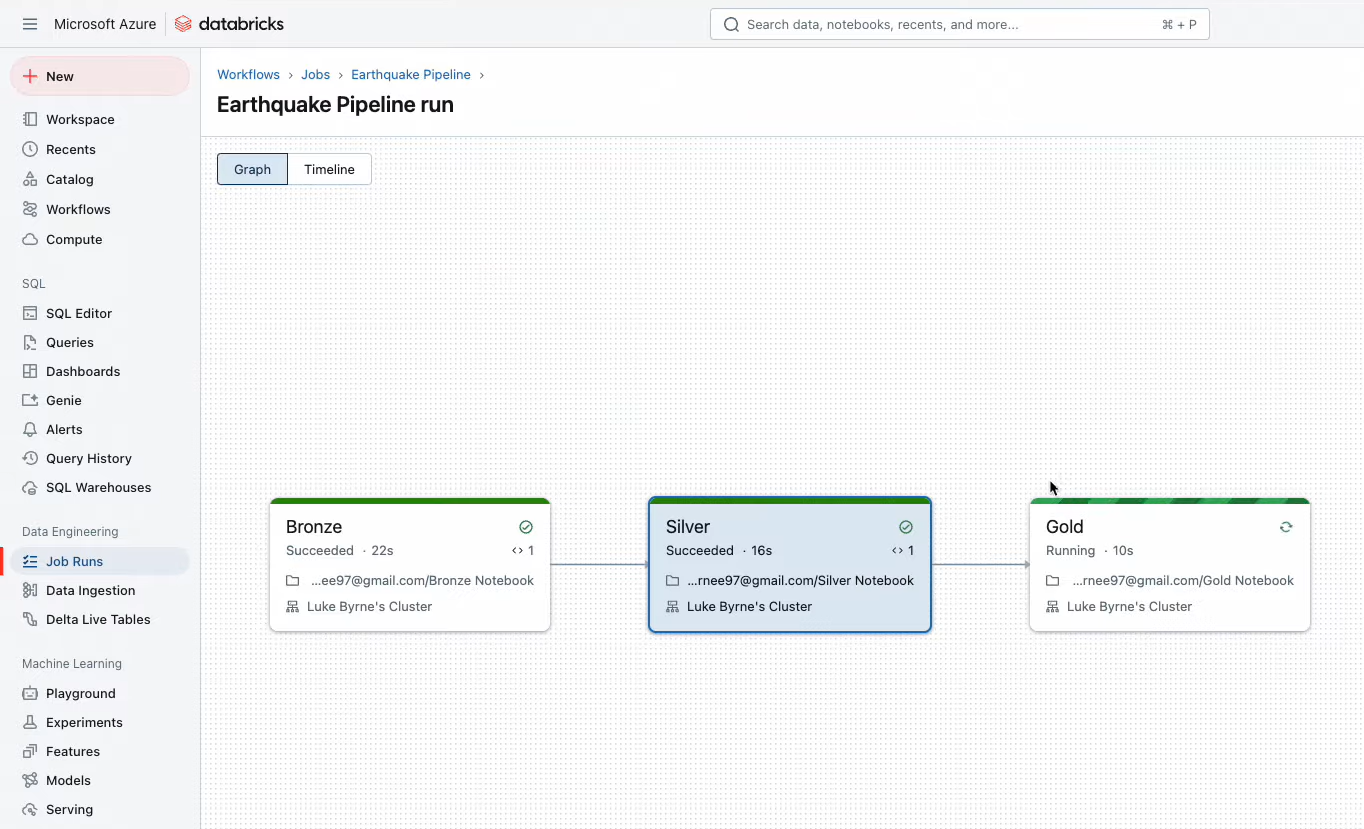
**Step 6 :**

After successfully creating the pipelines we need to schedule the pipelines according to our requirements. Here we have different types of schedules where we can keep the status active or we can pause etc..,. After scheduling this we have to run the pipeline. Click save once schedule is finished.

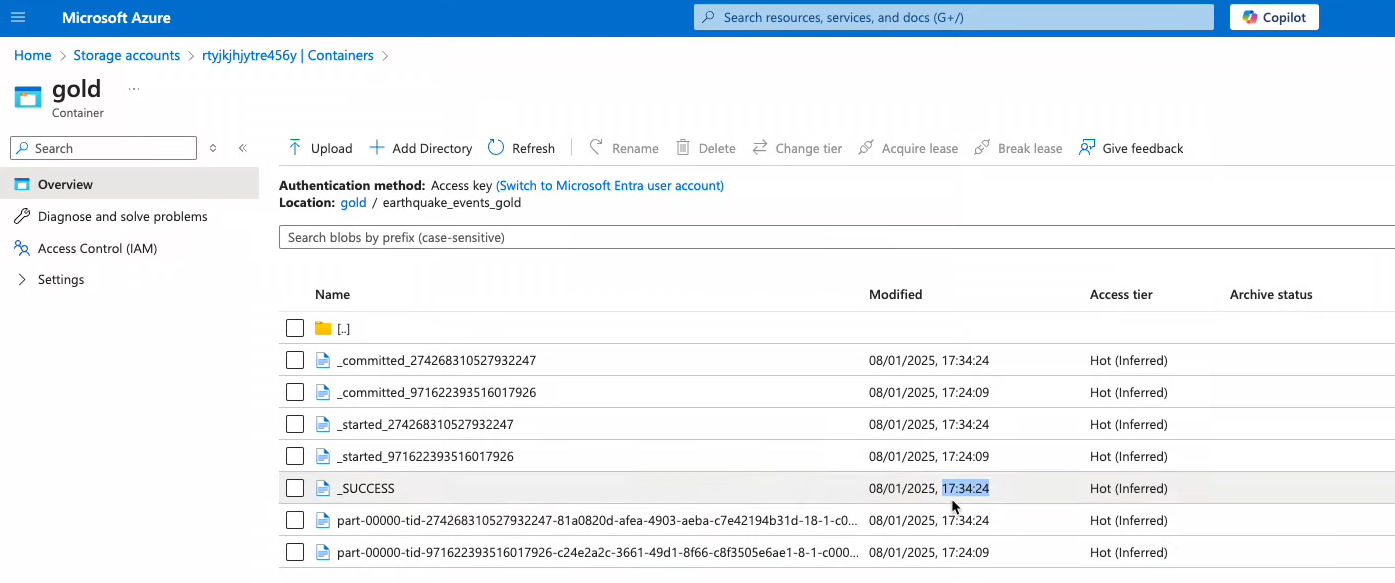


**Step 7 :**

After running the pipeline we need to check it . This is also called as a Job runs this we can see in the databricks platform under the **Data Engineering tab** we can see three options in that open **Job Runs** option and we can see the pipeline runs.



Once Job runs successfully go to the storage account and check the bronze, silver and gold containers whether the data is imported correctly or not.



This process involves completing the entire Azure Data Engineering Project workflow within Databricks without using Data Factory. Databricks supports various methods for handling different data sources, allowing data import using any of its built-in languages.Additionally, it provides flexible and scalable solutions for data transformation, processing, and analysis.