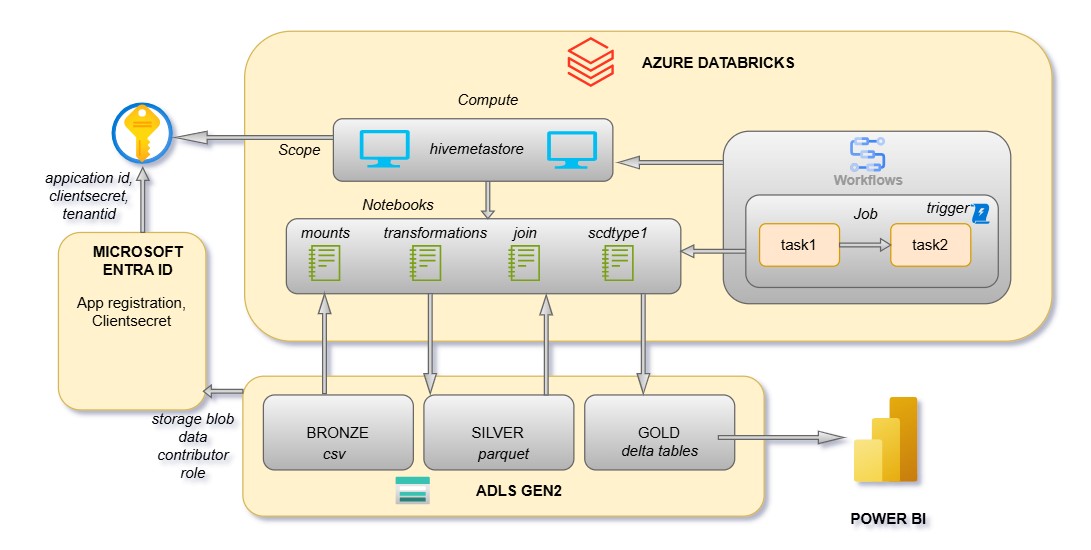
## **Project Overview:**

## This project demonstrates a full data engineering pipeline using **Databricks** to move and transform data through medallion architecture (Bronze → Silver → Gold). The pipeline extracts raw data from a bronze container in ADLS Gen2**,** transform the files with necessary transformations and load into the silver container in ADLS Gen2 in parquet files and then apply **SCDTYPE 1** and store the data in gold container in ADLS Gen2 as delta from where enriched data can be consumed for further analytics downstream.

## **Project Architecture:**



**1.Bronze Layer (Raw Layer)**

Raw data is copied manually as-is to the bronze layer in ADLS Gen2

* **Files Included**:
  + accounts.csv
  + customers.csv
  + loans.csv
  + loan\_payments.csv
  + transactions.csv

**2.Silver Layer (Cleaned/Transformed Layer)**

* **Process**: Data in the Bronze layer is transformed to address data quality issues such as:
  + Removing nulls
  + Casting data types
  + Joining relevant tables (if needed)
* **Target**: Cleaned and structured data is saved in the silver layer of ADLS in parquet.

**3.Gold Layer (Business-Ready Layer)**

* **Source**: Transformed data from the silver layer in parquet.
* **Target**: Transformed data from the silver layer is loaded into gold container in ADLS after applying **scdtype1** in delta for analytics, reporting, or further downstream consumption.

## **Tools & Technologies Used**

**Azure Data Lake Storage Gen2**: Data storage for bronze, silver and gold layers

**Azure Databricks:** Data transformation and orchestration

**Azure Key Vault:** To store secrets and credentials

**Azure Microsoft Entra ID:** For app registration to mount using service principle

## **Databricks Job Overview**

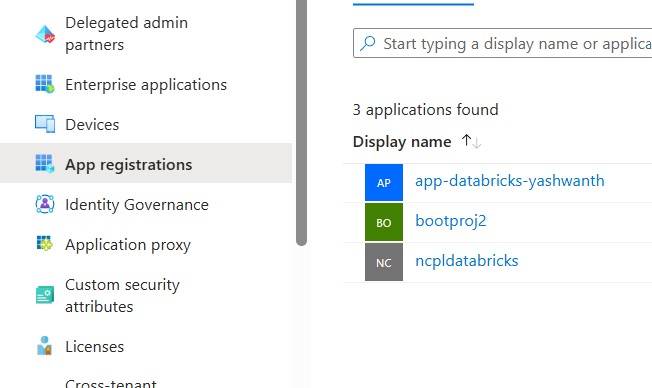
A Databricks job is set up to run two tasks **Bronze to Silver** and **Silver to Gold** which are scheduled to run daily with help of a trigger.

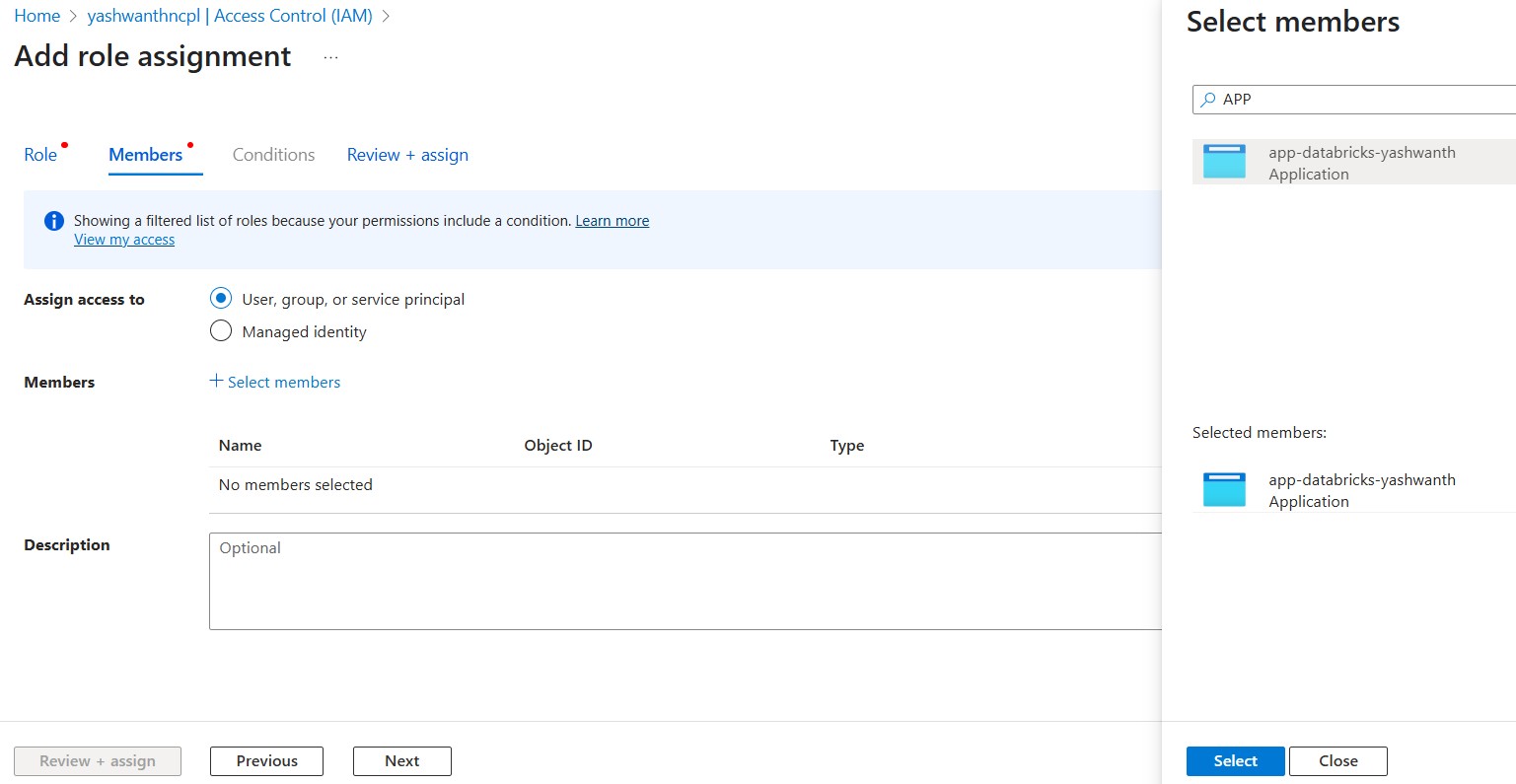
1. **Bronze to Silver** is the task 1 which is to transform data and store into silver container
2. **Silver to Gold** is the task 2 to apply scdtype1 on the files in the silver container and store it to the gold in delta files. This task 2 will run after the task1.

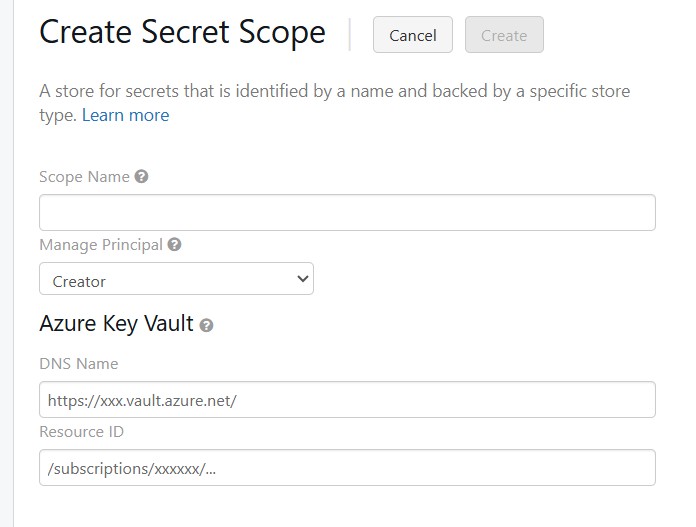
## **Detailed Overview of the Project:**

**Mount ADLS Gen2 to Databricks:**

ADLS Gen2 container are mounted to data bricks using a service principle by creating an **app registration and client secret** in **Microsoft Entra ID** and giving **storage data blob contributor role** to the service principle in the ADLS using IAM and store the client secret,tenantid, application id of the app in Azure Key vault.



To access the secrets from Azure Key vault, we need to create a scope in Azure Databricks by passing key vault uri and DNS name.

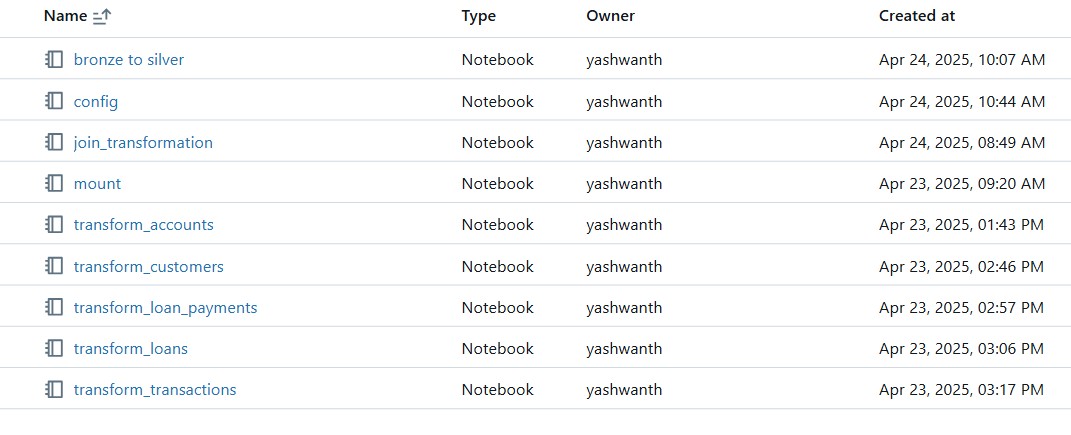


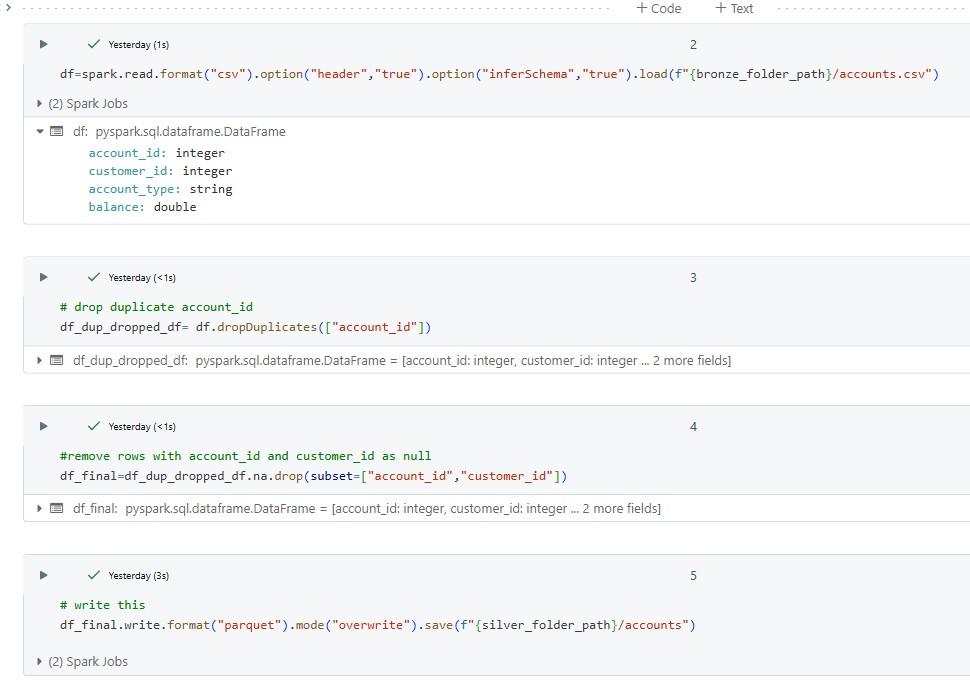
After creating the above details to the scope, we can read the secrets into the notebooks to create mount.



**Bronze to Silver:**

In this phase, raw files are transformed and stored into silver container in parquet format after applying the transformations required on the raw data.

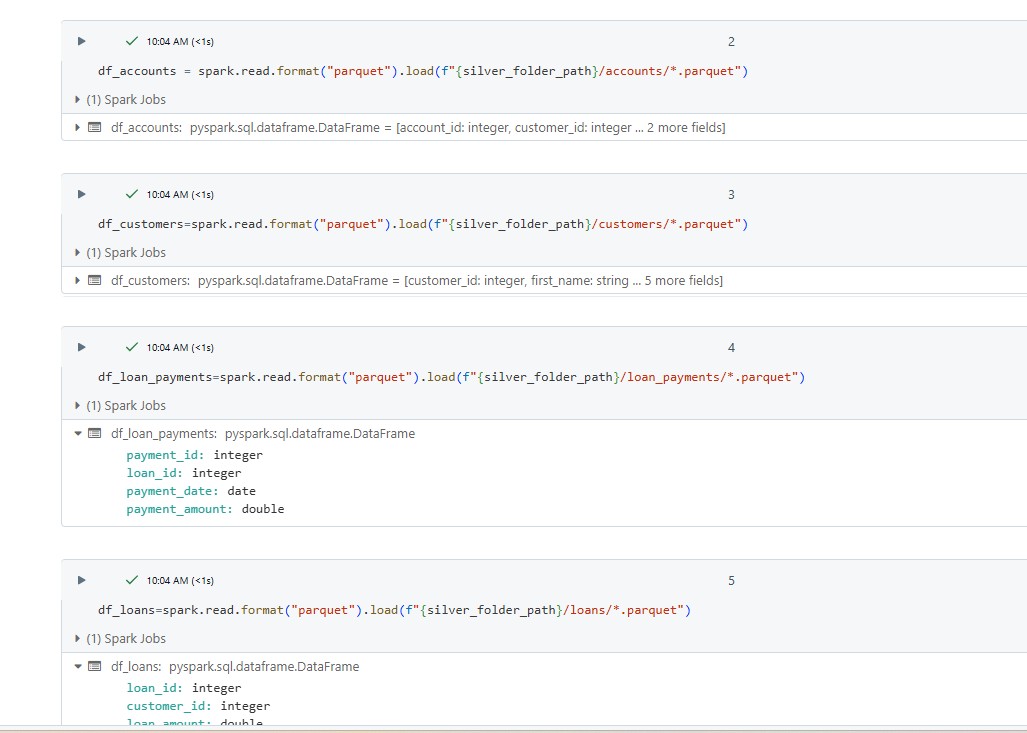


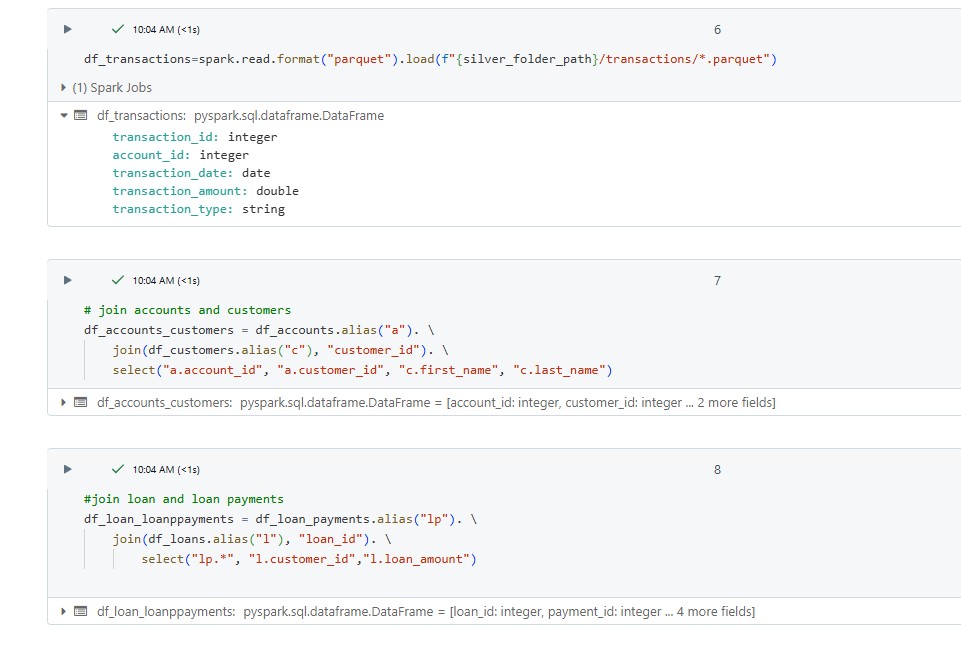


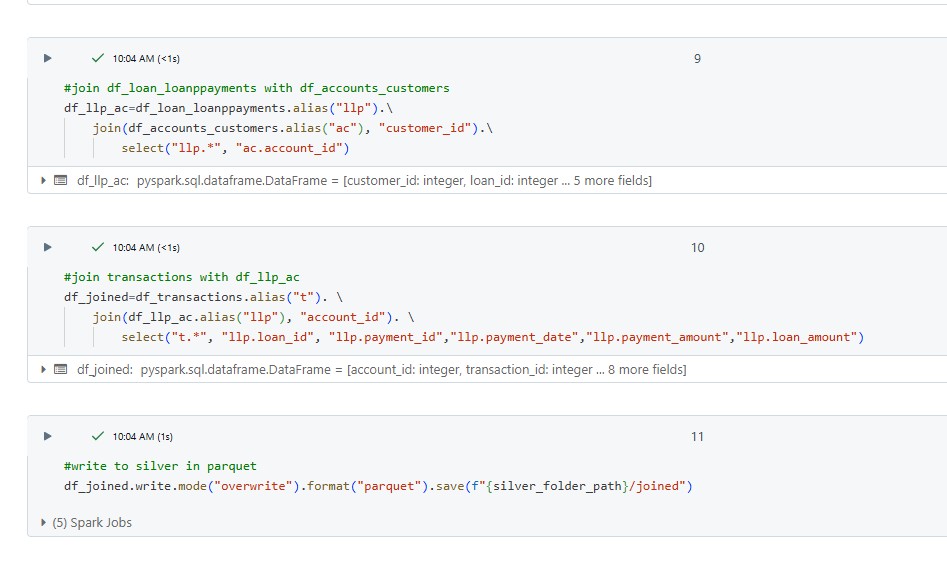
Here, duplicates and null values are dropped using drop\_duplicates() based on the key column and null values are removed using na.drop() based on key column.

Similar transformations are applied to all data frames and stored in bronze to silver folder.

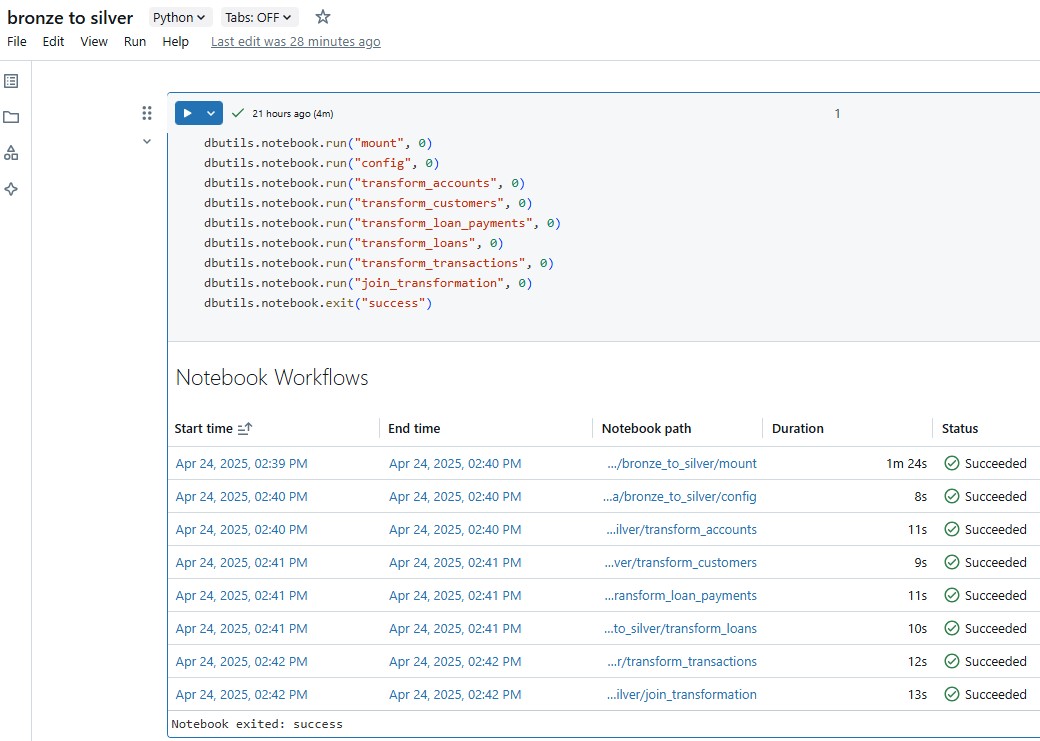
After transforming all the raw files and storing it into silver, then we do a join of the all the silver parquet files as shown below.



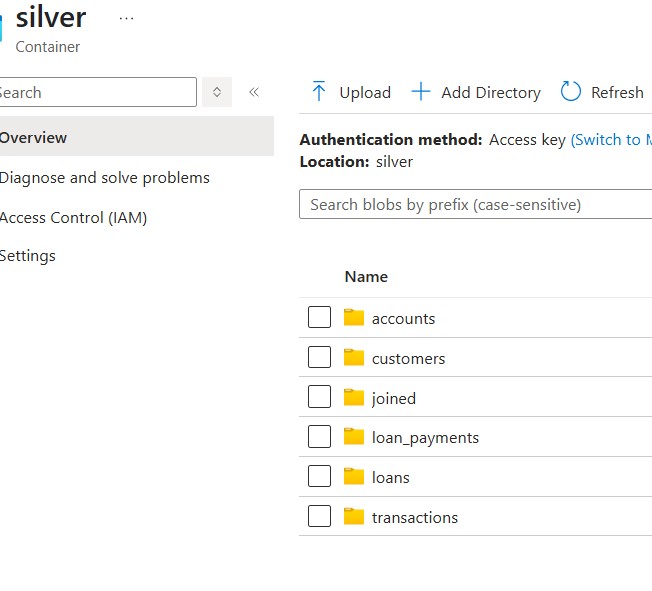




A master notebook will have the code to run all the Bronze to silver files sequentially; this file will be used in the databricks job to run all bronze to silver files



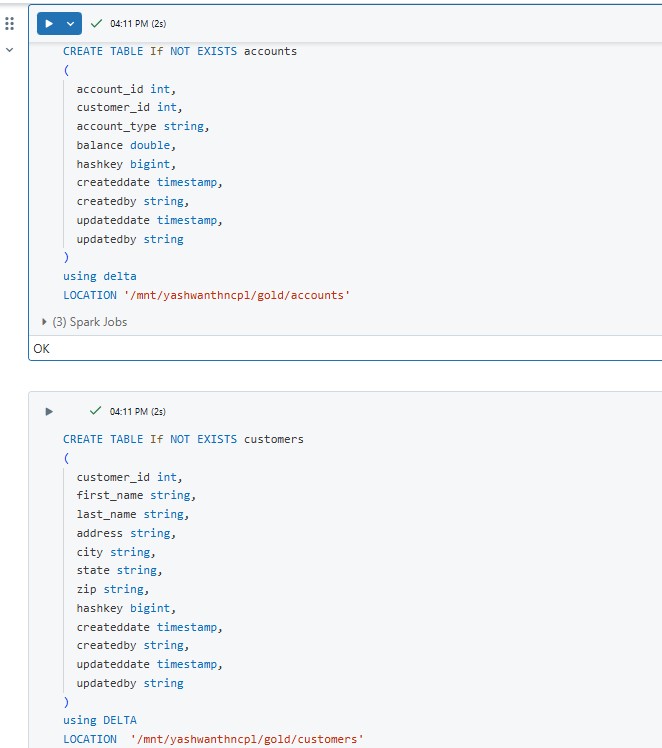
After the successful test run, we will have the data in ADLS silver container as shown below. Folders are created for each raw file in the bronze and for the joined file.

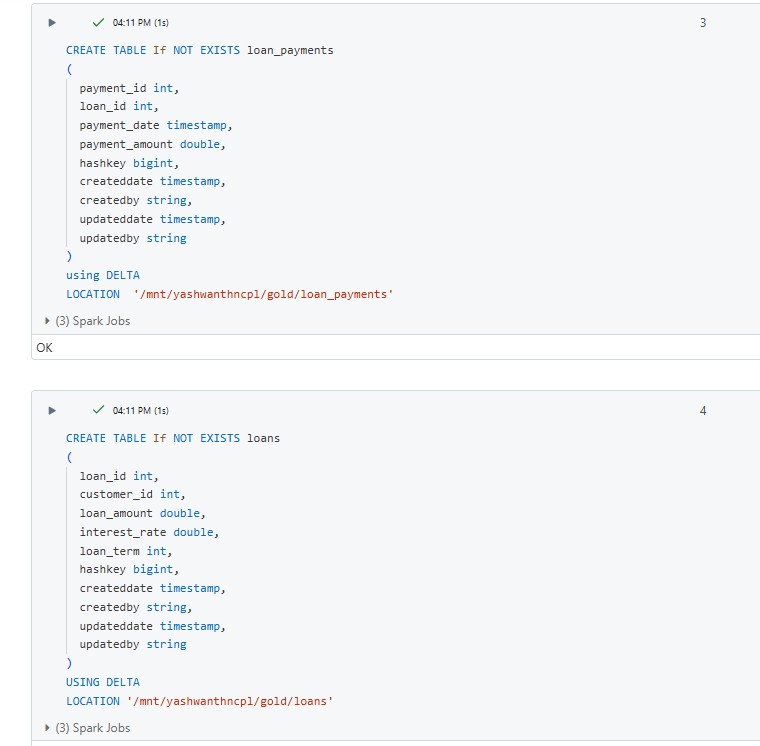


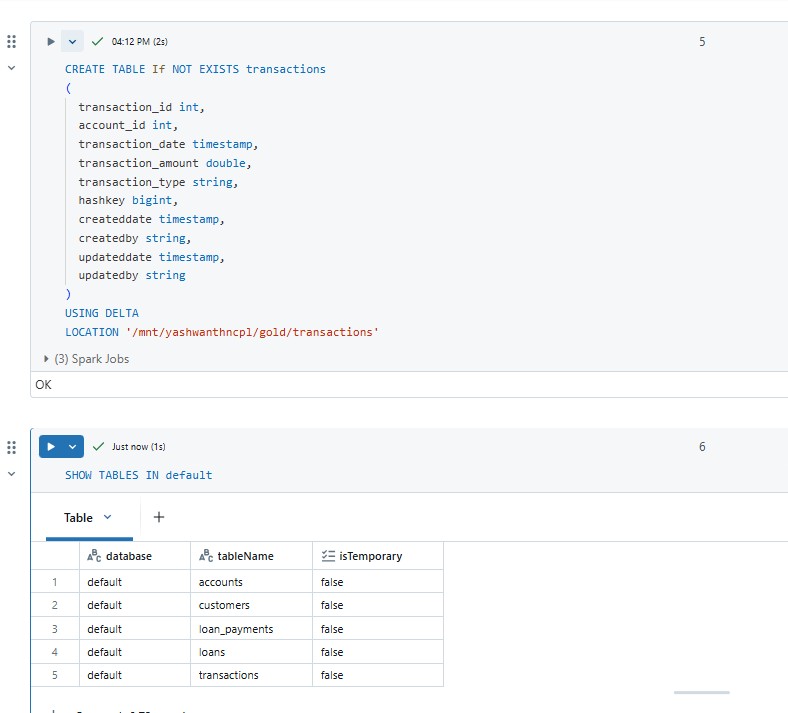
**Silver to Gold:**

In this phase, scdtype1 is applied to all the files in the silver container and then stored into gold container in delta.

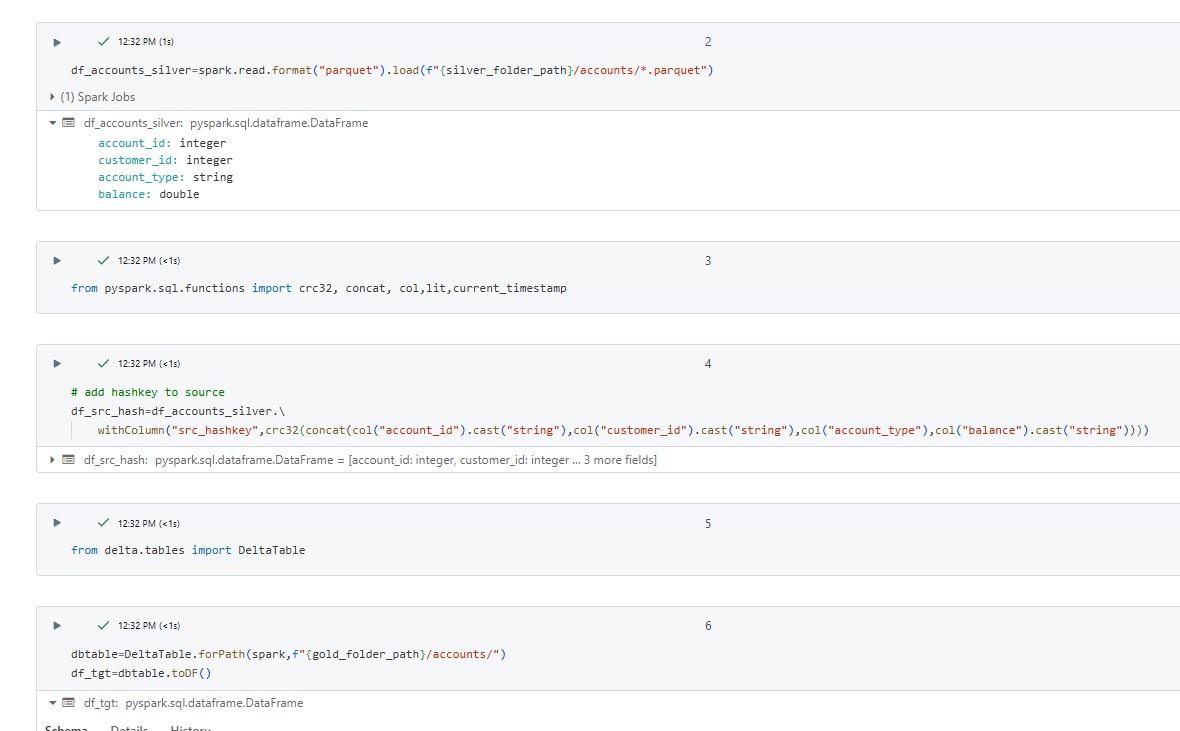
First, we need to create external tables in databricks with location as gold container using delta.

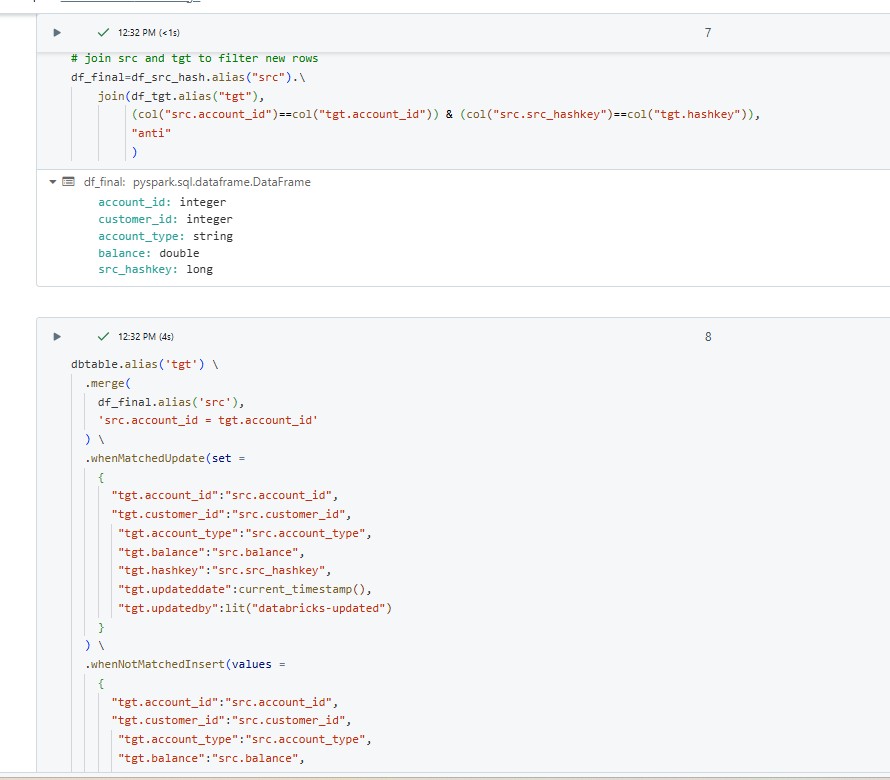






For SCDTYPE 1, source data will be compared with target tables using both unique key and hash key with the help of “anti” join, thus new rows will be filtered and merged into the target table. Below code shows how scdtype 1 is implemented for accounts

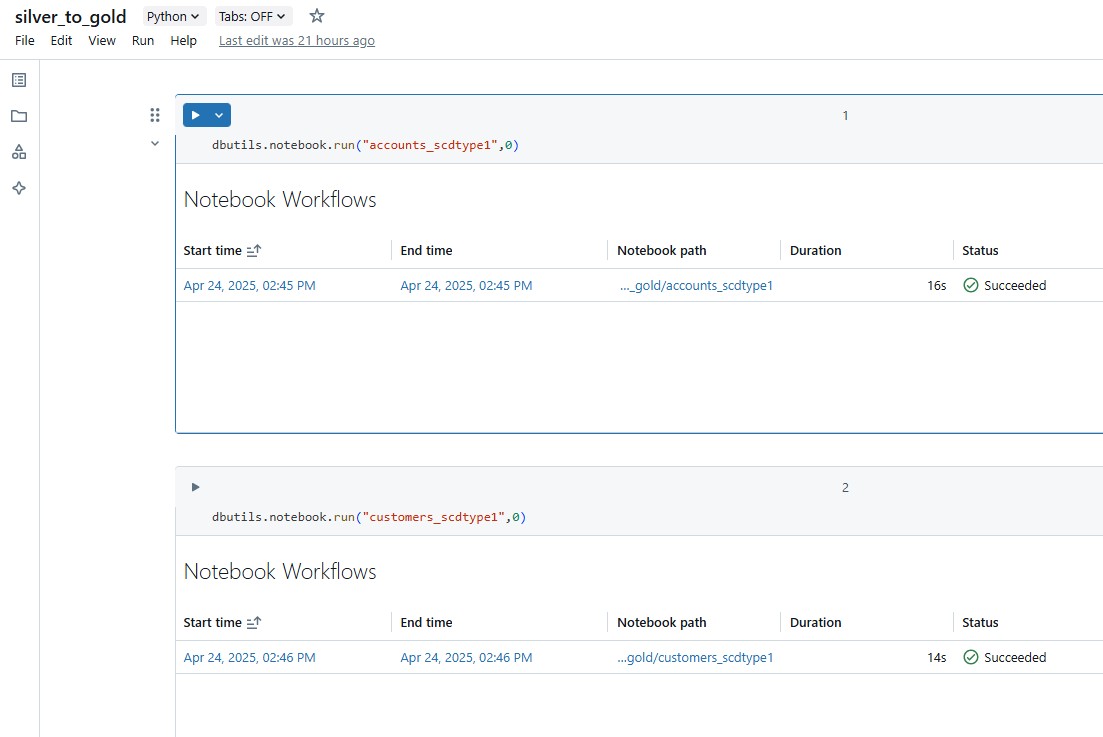


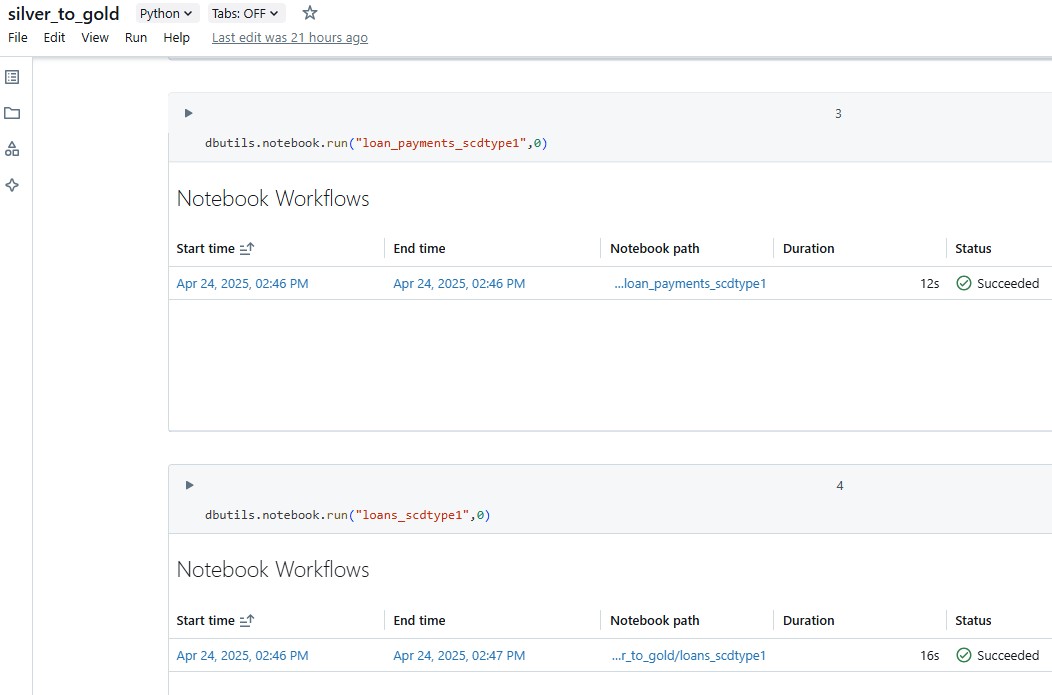


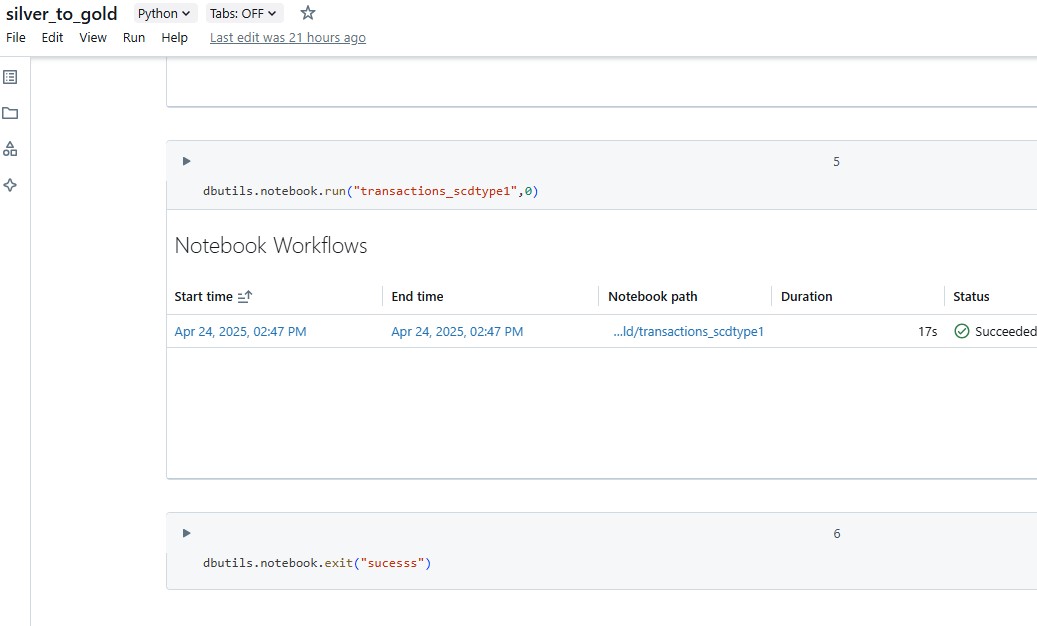


Similar SCDTYPE 1 logic is applied to all the files in silver containers and stored in Silver to Gold folder.

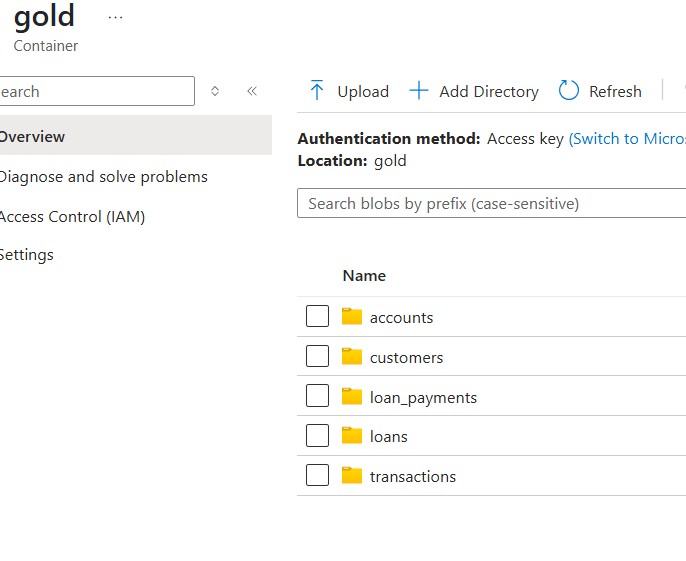
A master notebook is created to run all the notebooks in the silver to gold folder parallely.





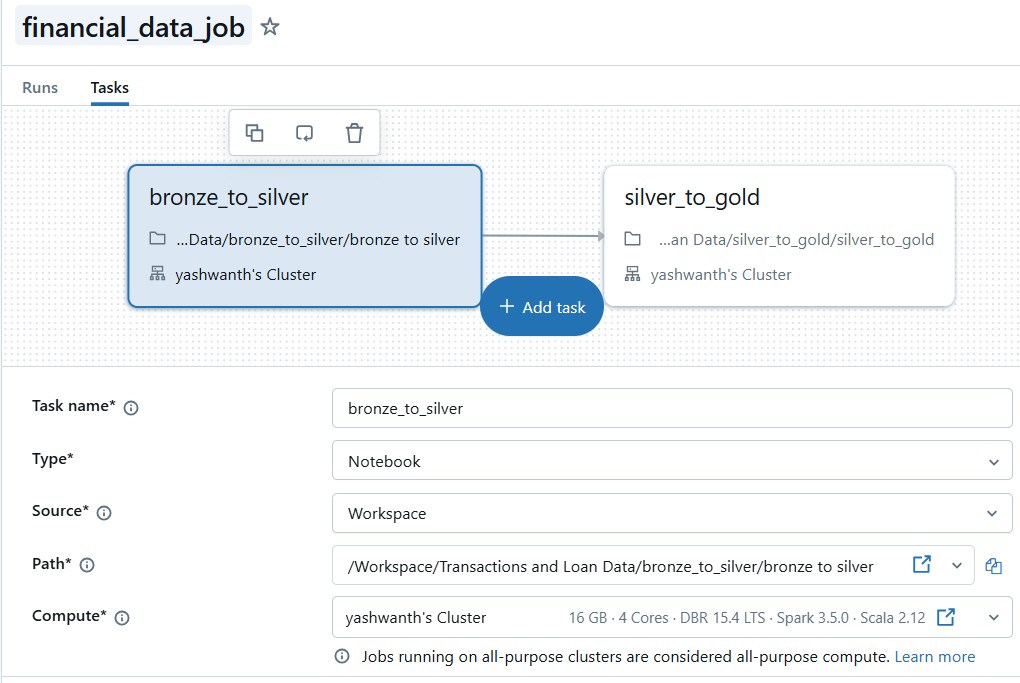


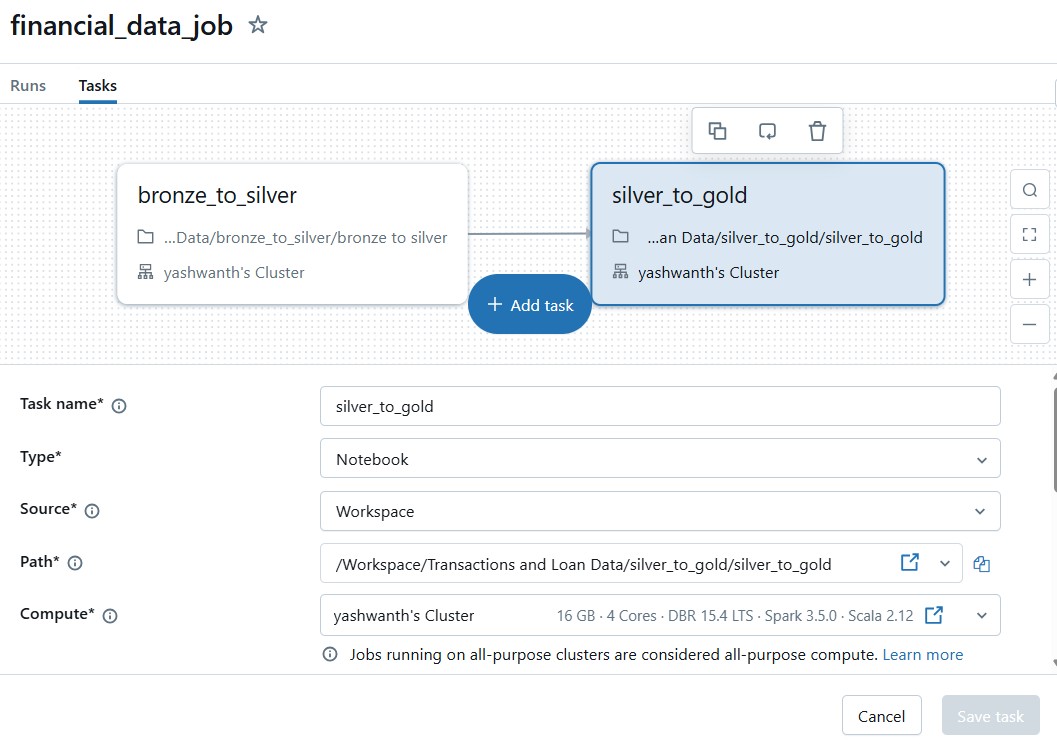
Gold container after running master pipeline

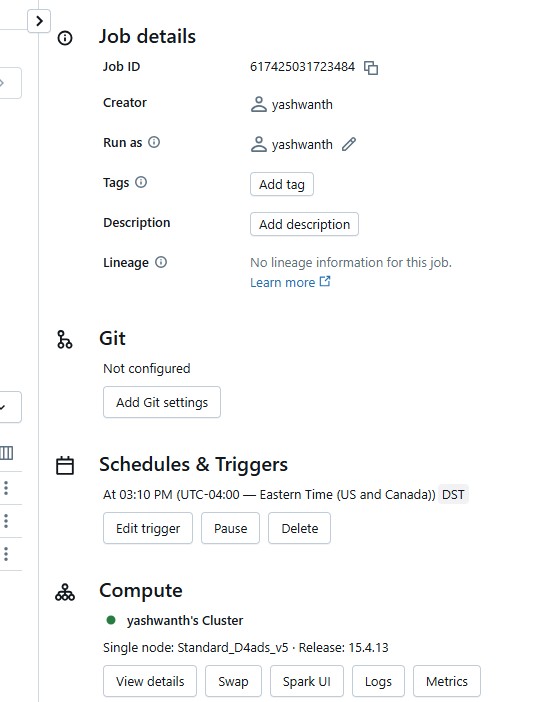


**Databricks Jobs:**

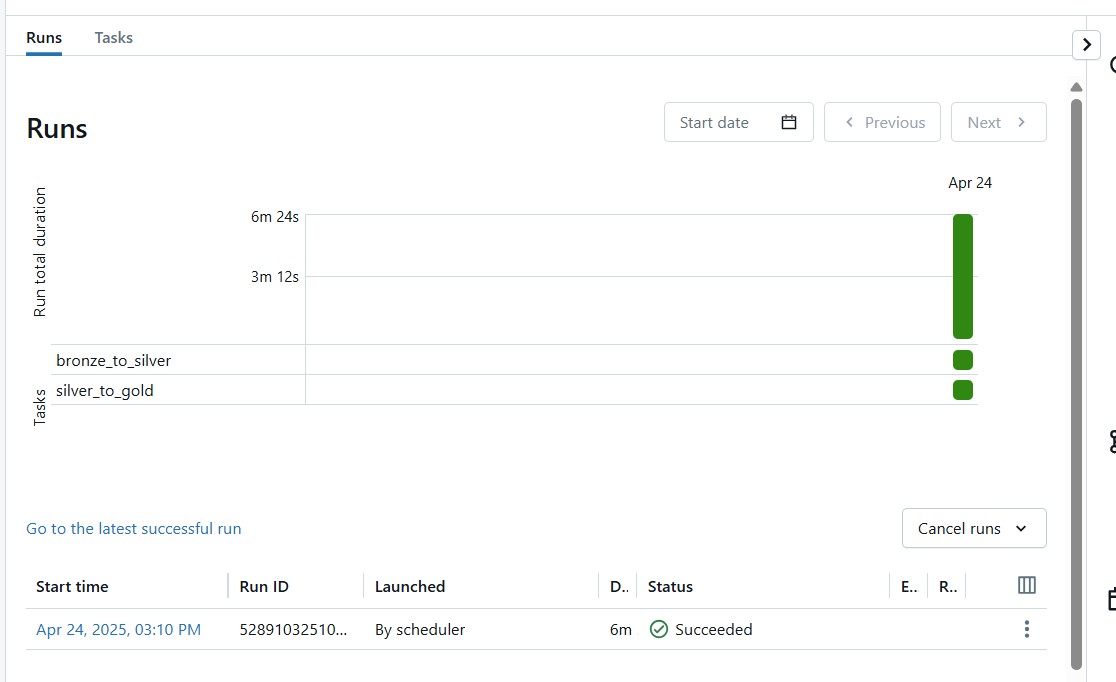
To run the notebooks on schedule, we need to set up a job in databricks workflows as shown below.







Set up a trigger to the job as shown above by selecting schedules and triggers. This job is scheduled to run daily at 03:10 pm and see the scheduled job runs



**Power BI:**

Connect the joined parquet files in ADLS Gen2 to Power BI to create visualizations and publish to fabric workspace

