Comprehensive Approach to the Azure Data Pipeline Lifecycle

DOCUMENTATION

The pipeline is designed to ingest, transform, and load data from SQL Server into Azure Synapse Analytics and Power BI.

It contains three levels:

- Microsoft SQL Server and SSMS
- Azure Portal
- Power BI

Level 1: Microsoft SQL Server and SSMS

- Open "SQL Server Management Studio
- Connect to Database engine using appropriate credentials.
- Upload data into your database.
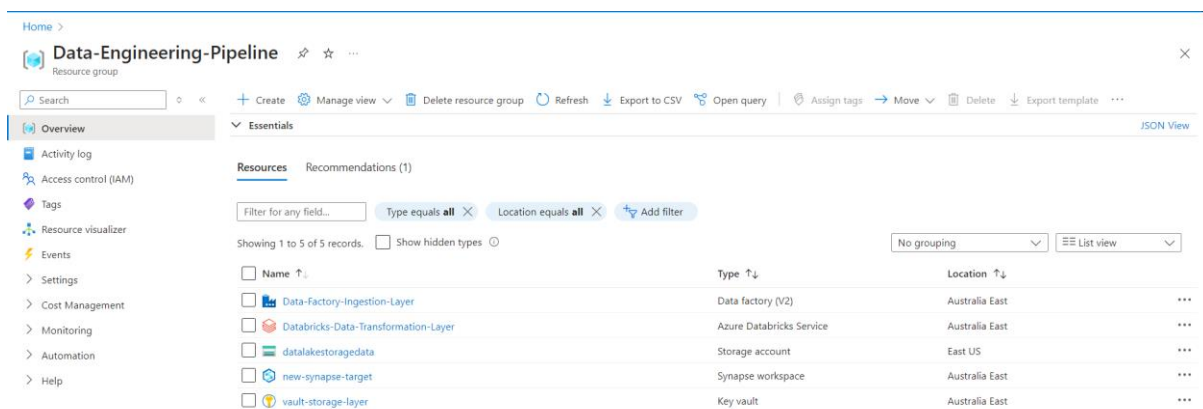- Query the tables for specific requirements (if necessary)
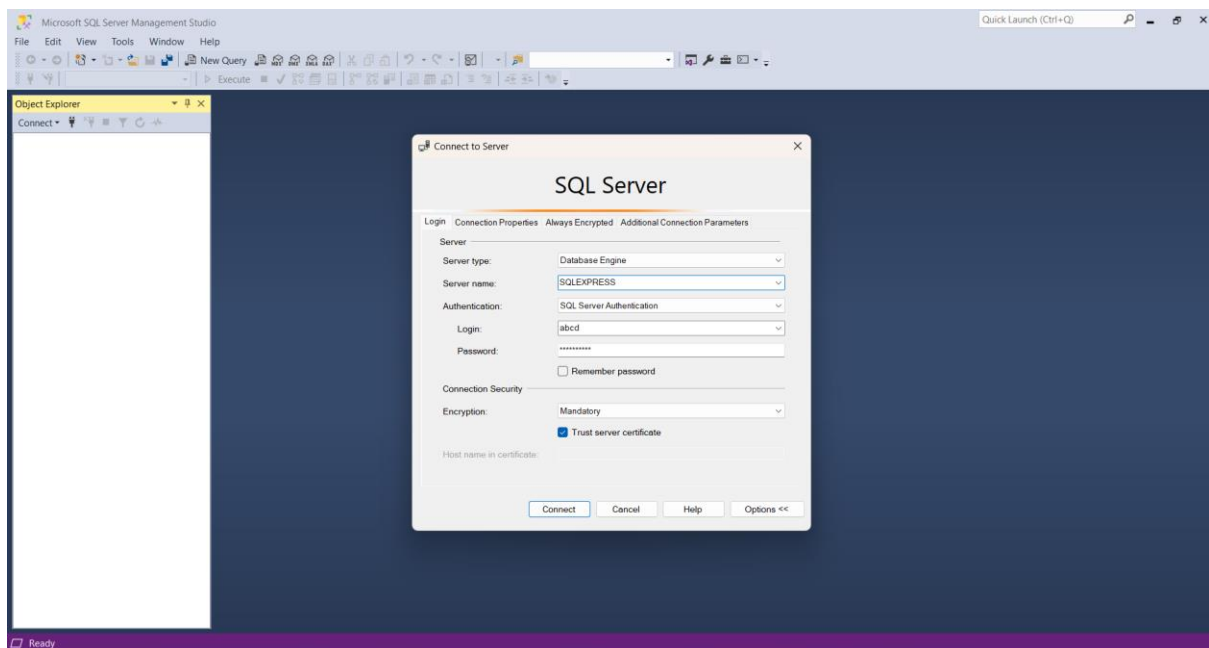
Level 2: Azure portal

- Sign up/Login using credentials
- Create a Resource group

    List of resources to be created under this resource group are:

    - Azure Data Factory (V2)
    - Azure Databricks service
    - Azure Key vault
    - Storage Account (Azure Data Lake Storage Gen 2 (ADLS Gen 2))
   - Azure Synapse Analytics Workspace

- Under "Access Control (IAM)" section of your resource group – Add "Owner" role assignment. (Not Managed Identity) (Select user or group option)
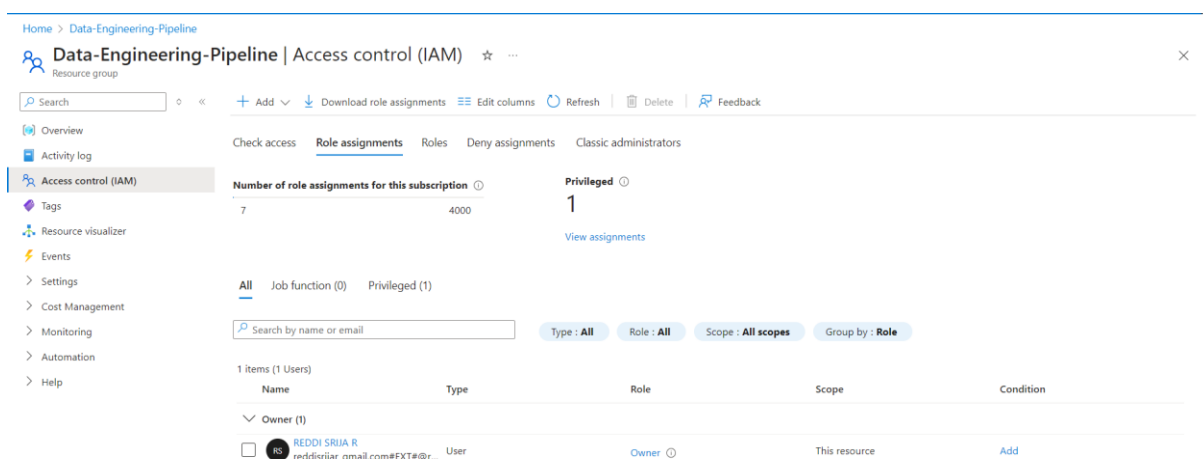
- Your project pipeline will be built in four layers:
    - Data Ingestion Layer
    - Data Transformation Layer
    - Data Loading Layer
    - Data Reporting Layer

1. Data Ingestion Layer

- Launch "Azure Data Factory Studio"
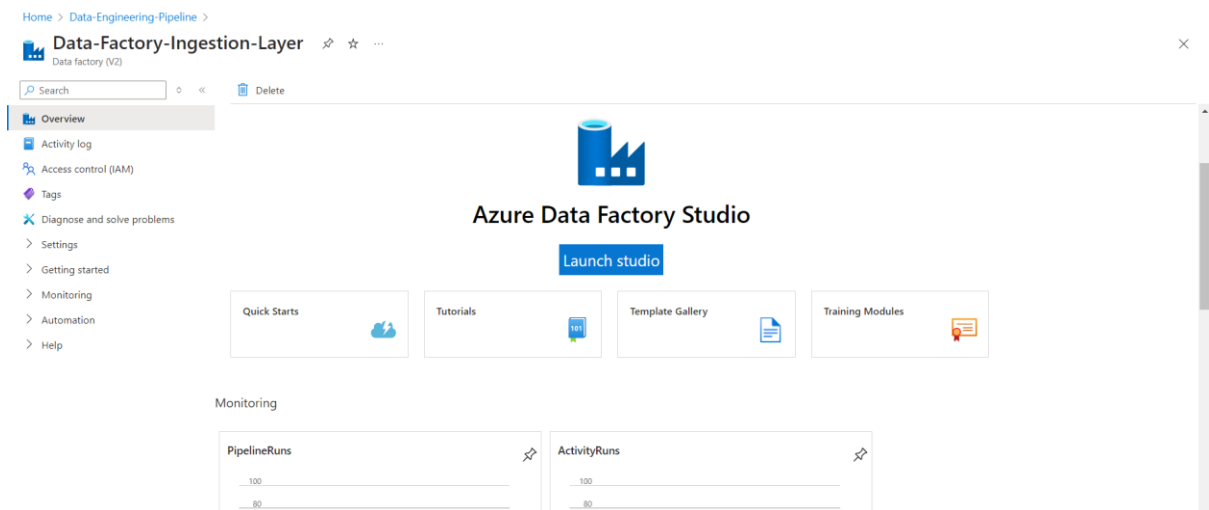
- Navigate to "Manage" section in left pane

- In "Integration service"
    - Default service named "Auto Resolve Integration Runtime" will be available – used to connect any cloud-based resources.

    - Create a new "Integration Runtime" – used to connect any on-premise resources. (As we are ingesting data from on-prem SQL Server)
    - Select either option and follow the instructions.
        - If Option 1: Click the link to download -> run the .exe file -> you can choose to either start or stop the created integration runtime service.
        - If Option 2: Click the link to download -> use provided either of the key to complete the setup.
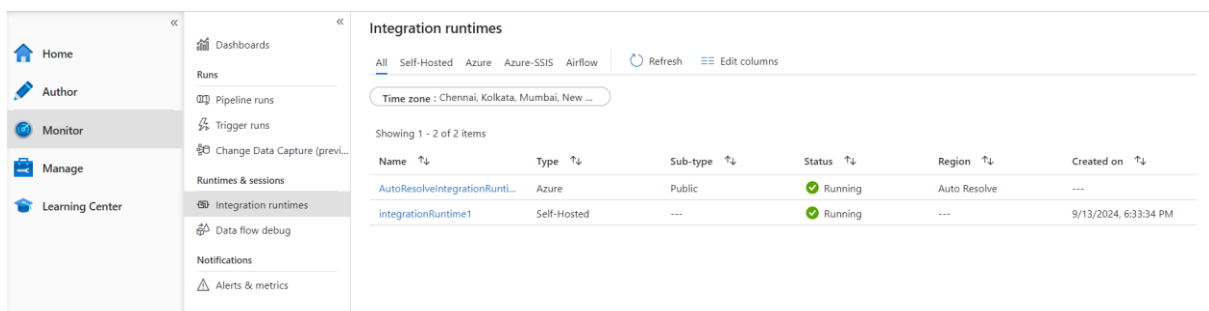
- Create a connection under "Linked Services" section for:
    - SQL server (use integration runtime)
    - Key vault (to automatically detect credentials for connection between azure services)
   - ADLS Gen 2 (use auto resolve runtime)
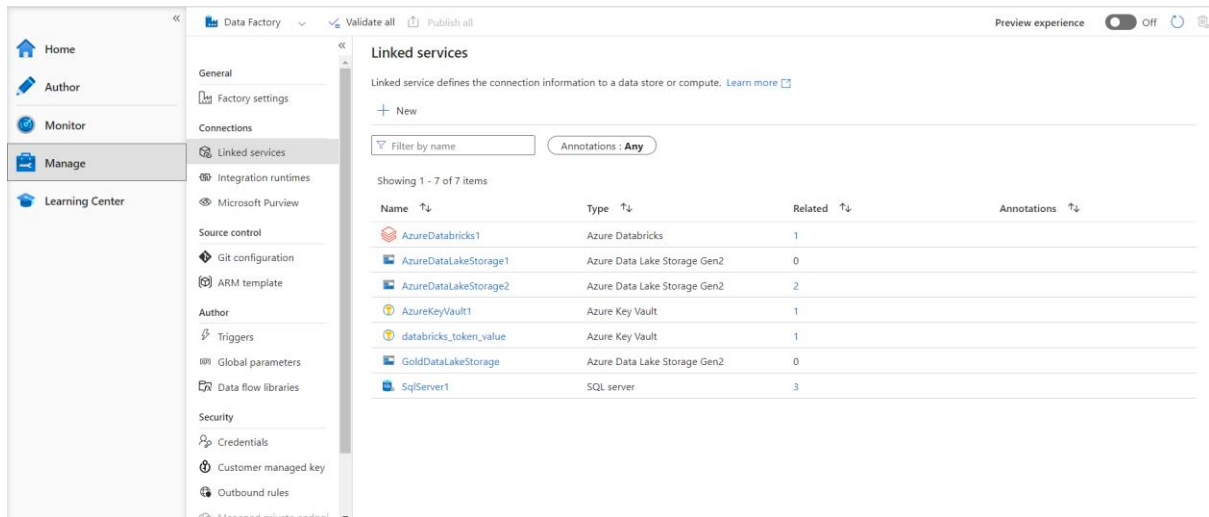
- Navigate to "Author" section in left pane.
- Create a new pipeline to load data automatically
    - Select "Lookup Tables" – to copy tables from our on-prem SQL Server in 'parquet' data format; from "Activities" section.
      Click "Publish" or "Publish All" to save your work in ADF (Azure Data Factory).
    - Create a "Bronze", "Silver" and "Gold" containers in 'storage account' resource (Data storage section) in your resource group.
    - Provide that Bronze container's location to the pipeline for raw data to get ingested.
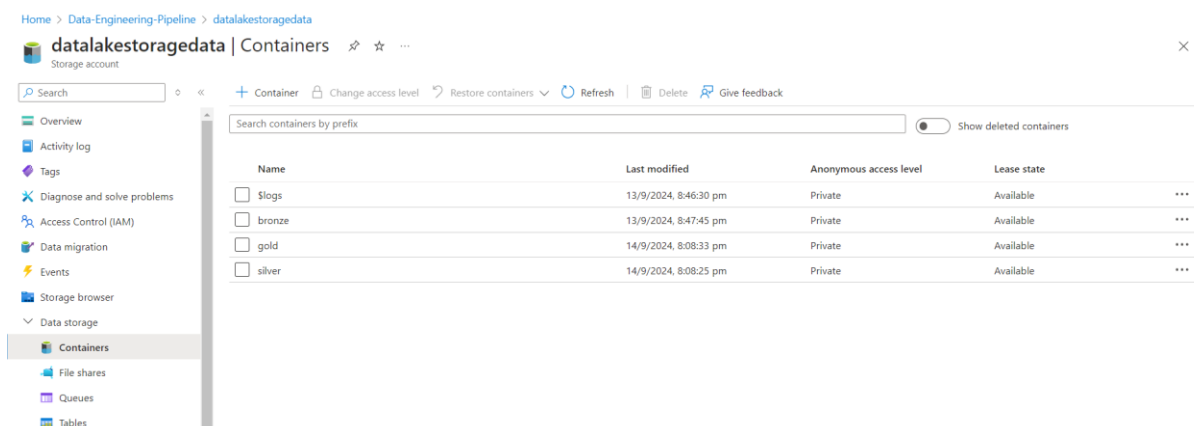

- Under Activities, select "Copy Each" and integrate it with "For Each" Activity – to run the activity for each table inside your database.

## 2. Data Transformation Layer

- Launch Databricks workspace
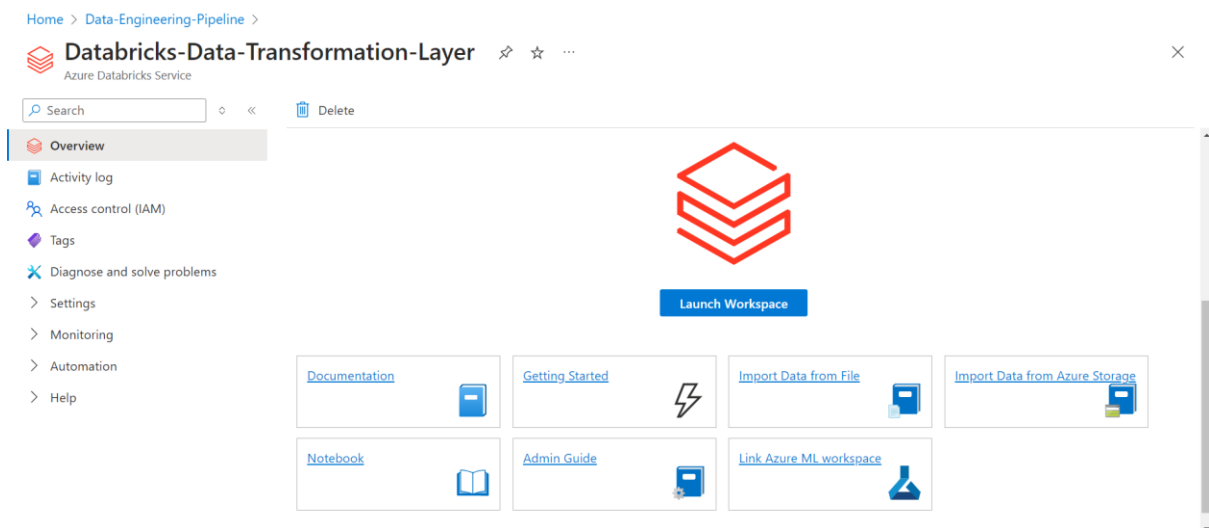
- Under "Compute" section – create new cluster
- Under "Workspace" section – inside 'shared' folder – create three notebooks
  - Storage Mount – to mount our cloud resources
  - Bronze to silver notebook – to store transformed data
  - Silver to gold notebook – to stored fully transformed data
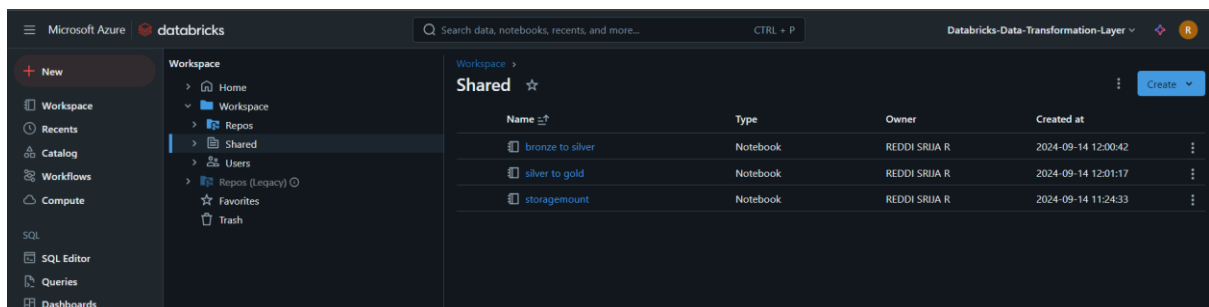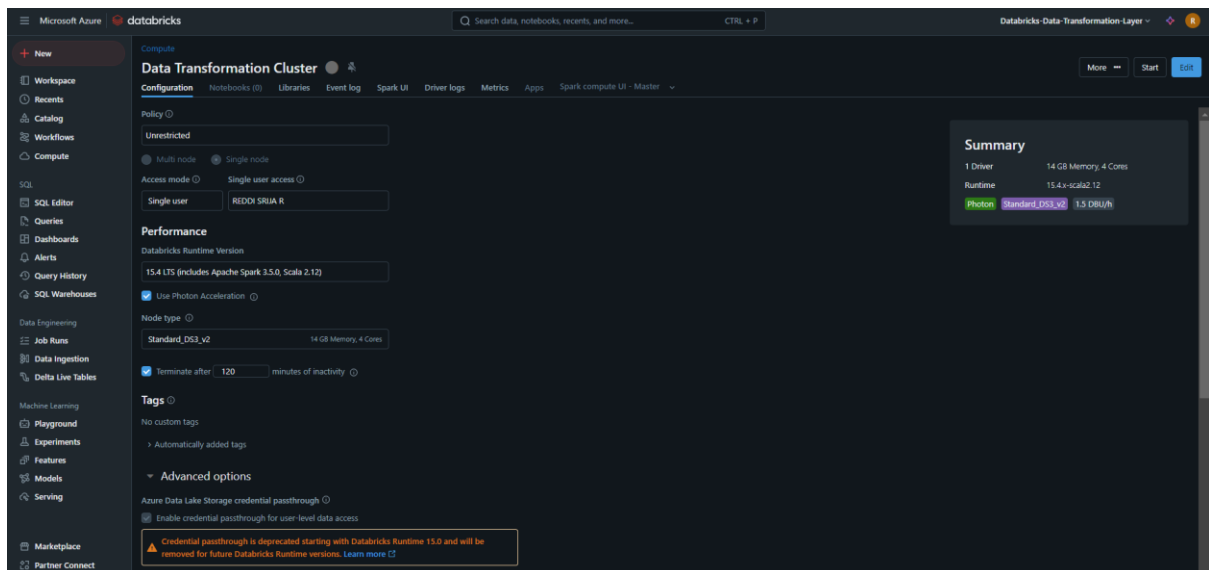
- Navigate to Azure data factory:
    - In Manage -> linked services -> create a connection for "Databricks" with an access token.
    - Access token can be created in "Databricks" resource in your resource group.

    - Go to your previously created pipeline and select 'Databricks' from Activities section.
    - Connect your "Bronze to silver" and "Silver to gold" notebooks.

    - Notebook's can contain code that performs any type of data transformation necessary for your use case.

    > Code for the notebooks is uploaded in the github with same filename.
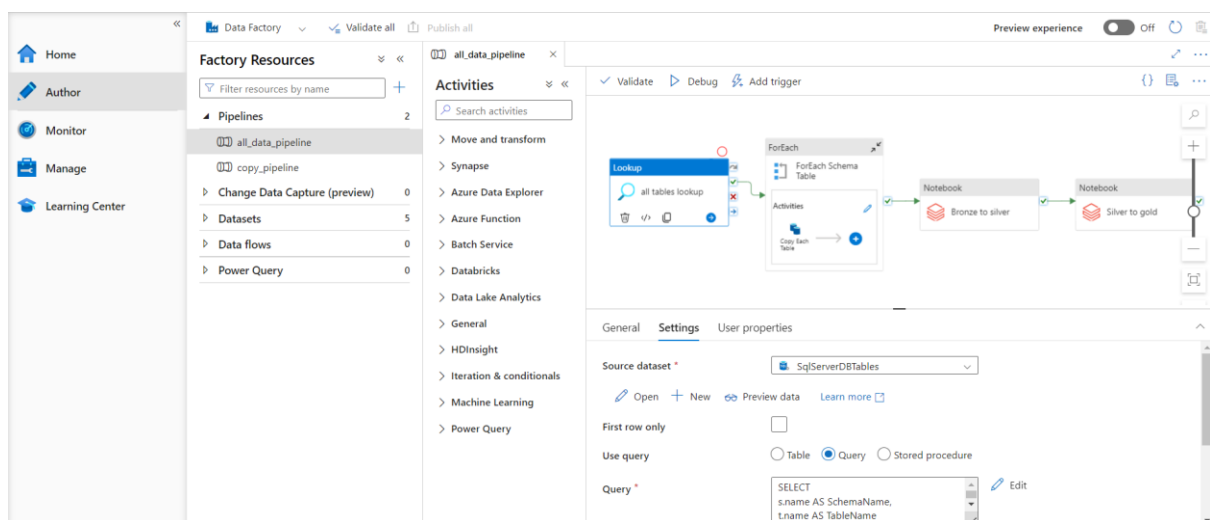
- You can "Debug" and check the working of pipeline

- Click on "Add Trigger" to run the entire pipeline and monitor the progress of your pipeline under "Monitor" section in left pane.

- Trigger has two options: If "Trigger now" is opted, then you have to manually run the process; whereas if opt for "New/Edit", then it will create a trigger which will automatically run our pipeline upon any modifications of your data in your database.

3. Data Loading Layer

- Download OpenJDK for synapse workspace.
- Launch "Open Synapse Studio"

- Create a connection from "Linked services" from Manage section in left pane for "Azure SQL Database".


- Create a "Serverless SQL Database" from "Data" section in left pane. It will be reflected in 'workspace' block.

- Under 'Linked' block, you can expand the sections to find you data containers. Right click and select "Top 1000 rows" to have a preview of your data.
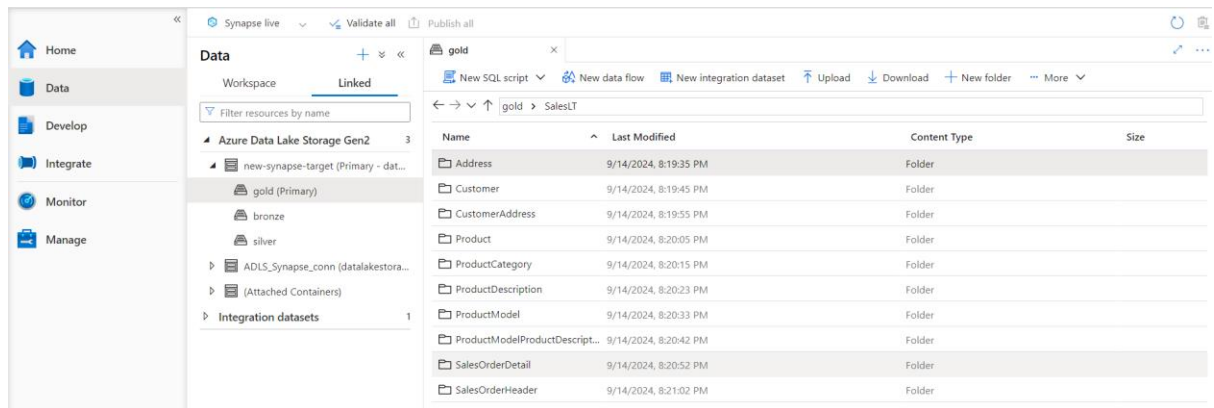
- From "Develop" section in left pane, create a .sql script file or import existing .sql file which should contain SQL code to that creates views of your data into the database you created in synapse workspace. Run and select your previously created database.

- A 'stored procedure' sql file will be developed to automatically detect data from your containers present in ADLS Gen 2 storage account and load into your database as Views.
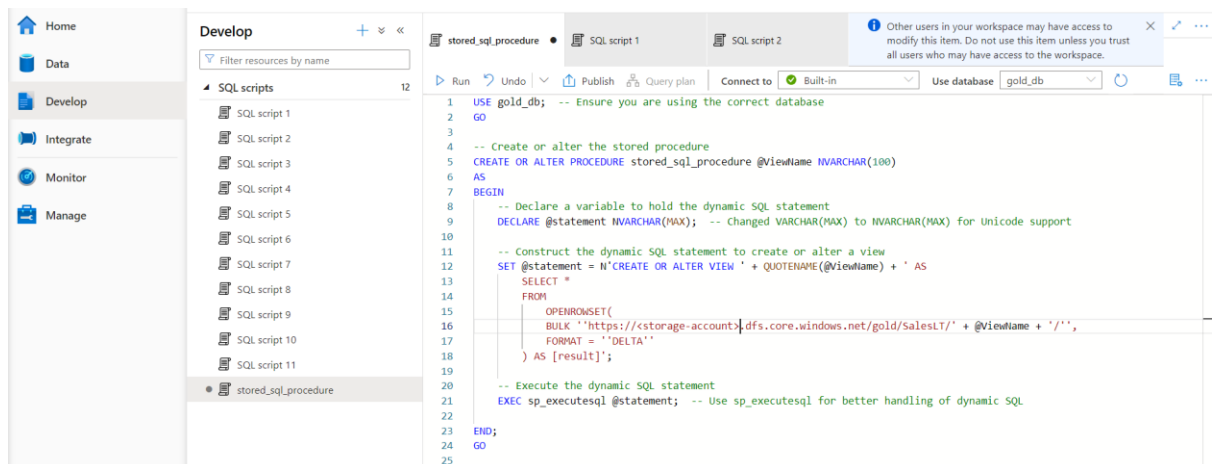
- From synapse workspace – properties – collect "Serverless SQL Endpoint" to use it whenever necessary as "Server name" value.

- From "Integrate" section in left pane, create a pipeline for views creation.
- From activities section – select "Get metadata" option – opt for binary data format while connection and provide "Gold" container's location.
- Select "For Each" from activities section.
- As for the activity of this "For Each" block, select "Stored procedure" from activities section and connect your stored procedure .sql file developed.
- Use Debug or Trigger accordingly.
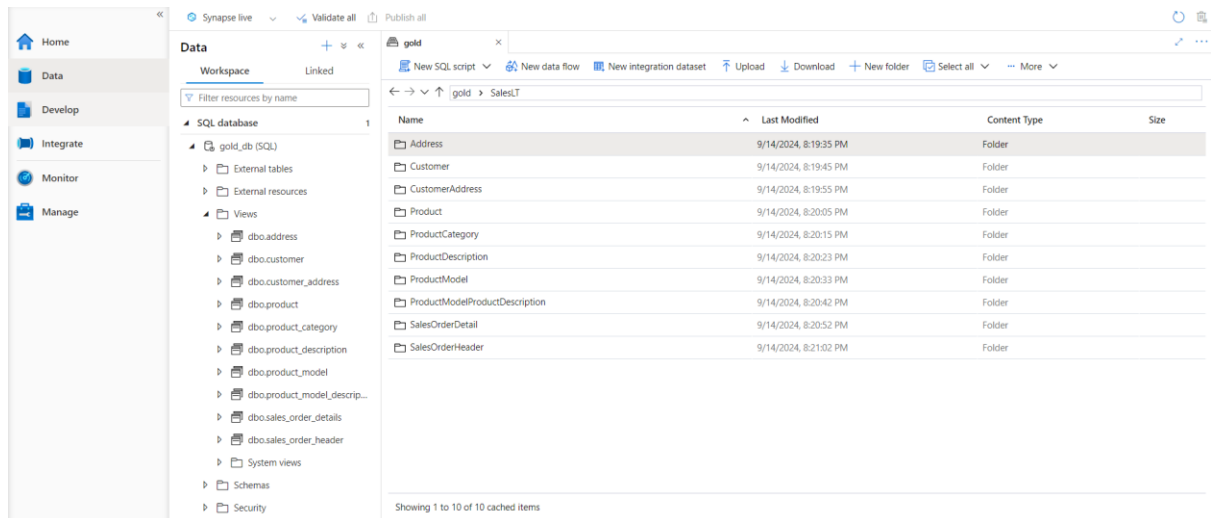
4. Data Reporting

- Power BI
  - Get Data - Azure - Synapse Analytics (SQL DW)
  - Provide Credentials and Sign in your Microsoft account
  - Select and Load Data Tables
  - Create visuals and Develop relevant Dashboard

Access Control (IAM)

- Add role assignments
  - Owner
  - Storage Blob Data Coordinator (User and Managed Identity)
  - Storage Blob Data Reader (User and Manager Identity)
  - While creation of Managed Identity – select synapse workspace.
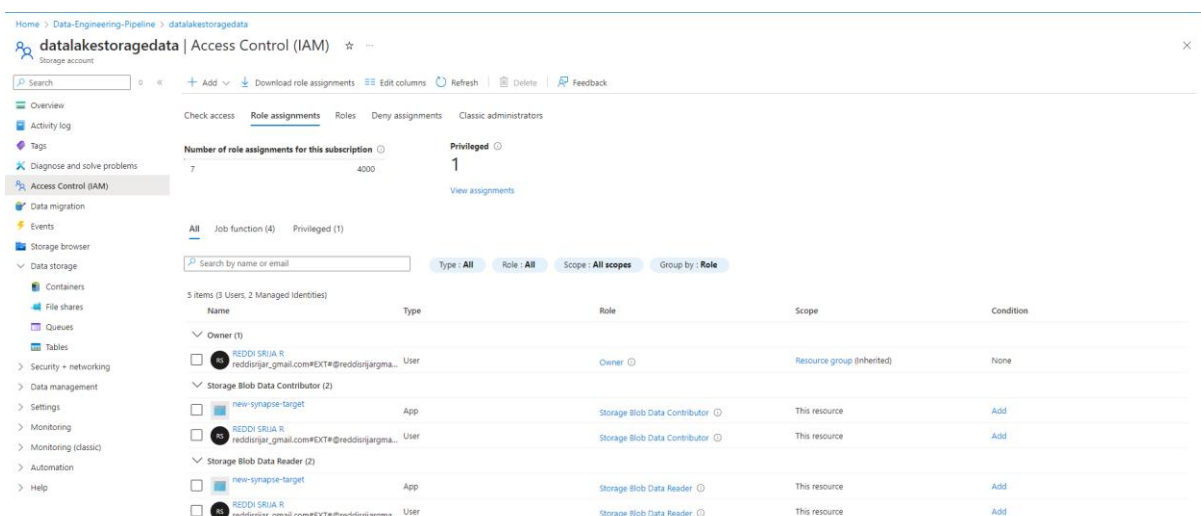
Key Vault

- Generate "Username", "Password" and "Databricks token" under "SECRETS" from "Objects" section.
- To give access of these secrets to other services/resources while connection creation, change "Access Configuration" to "Value access policy".
- Except for "Azure Key Vault" resource, other resource's "Access Configuration" should be "Azure role-based access control" (ADBC control)

Additional Resources:

- https://docs.microsoft.com/en-us/azure/data-factory/

- https://docs.microsoft.com/en-us/azure/databricks/

- https://docs.microsoft.com/en-us/azure/synapse-analytics/

- Mount ADLS Gen 2 storage Documentation: https://learn.microsoft.com/en-us/azure/databricks/archive/credential-passthrough/adls-passthrough