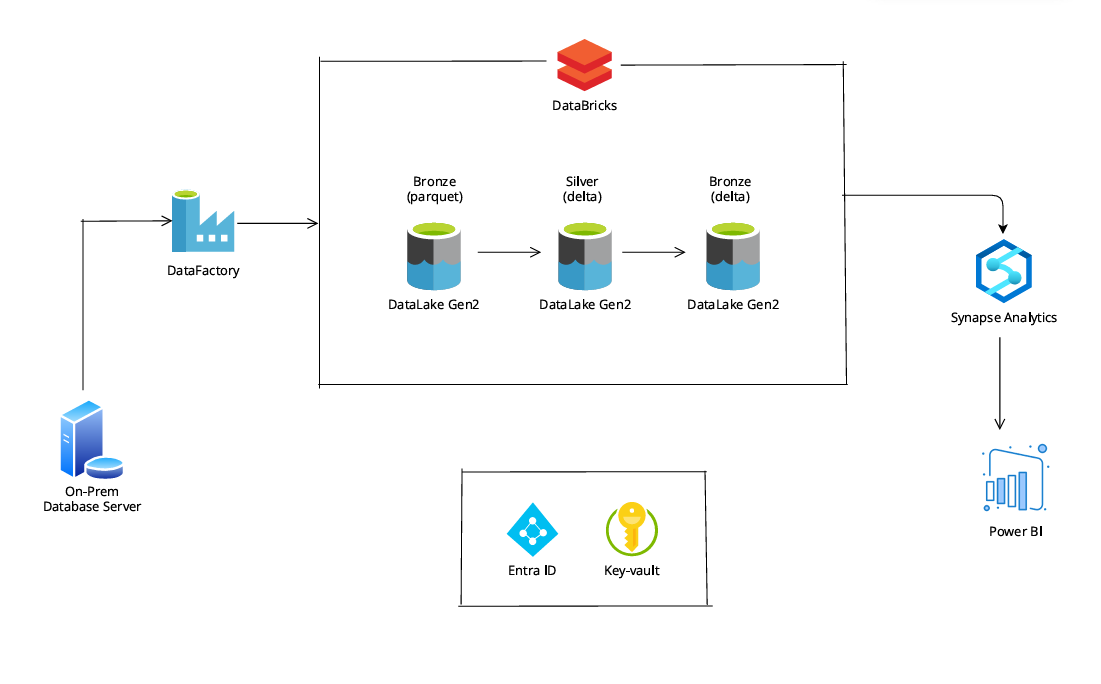
**Architecture:**



**Environment Setup:**

1. Create a Resource Group for the Project.

2. Create Azure Synapse Analytics and Data Lake gen2 Storage Account.

Under storage account created bronze, silver and gold container.

Bronze → to store the raw data

Silver → to store the processed data

Gold → to store the cleaned data that connect to power bi for end users

3. Create a Azure Data Factory.

4. Create Azure Databricks

5. Create a Azure Key Vault.

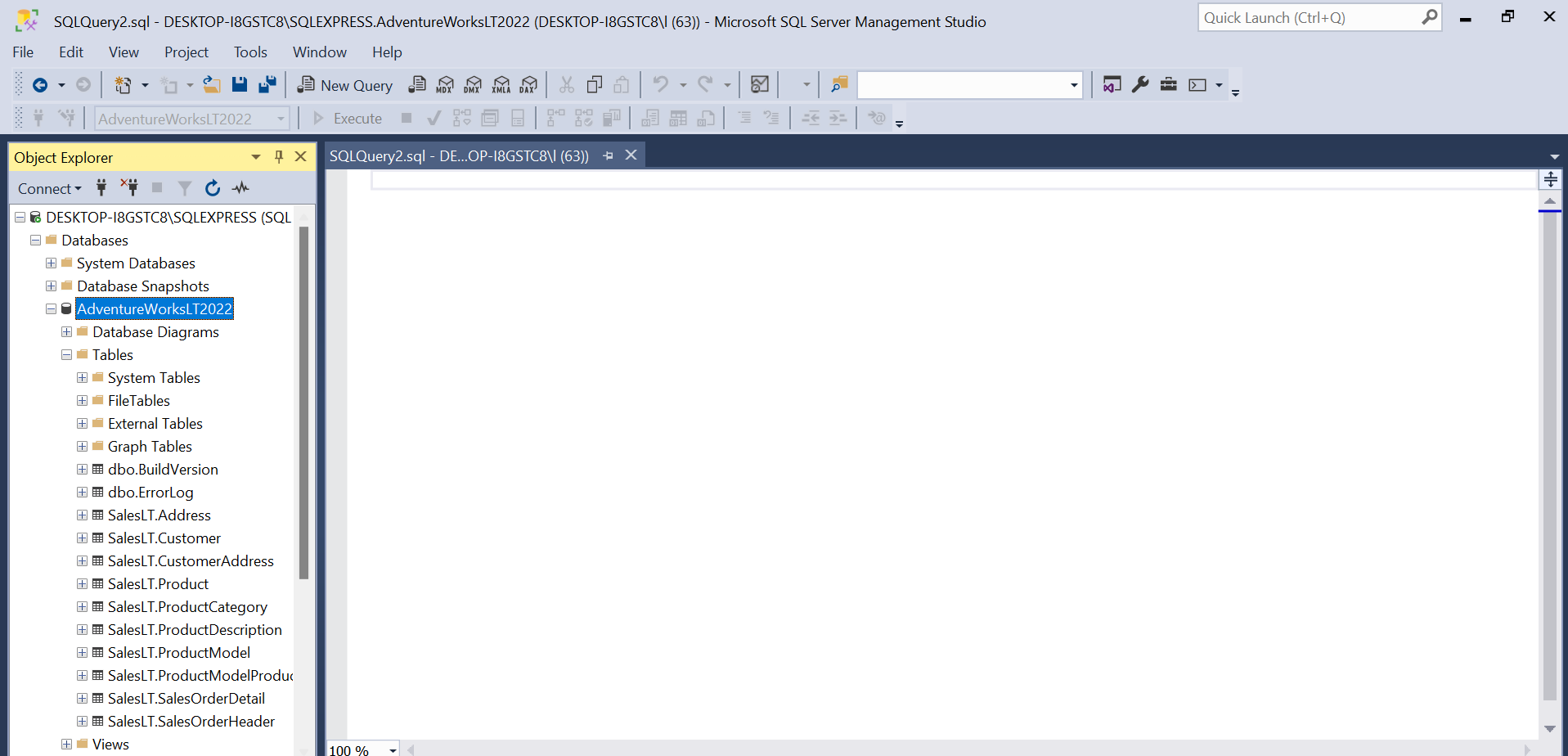
6. Create a Service Principal to connect the ADLS gen2 storage account with Databricks.

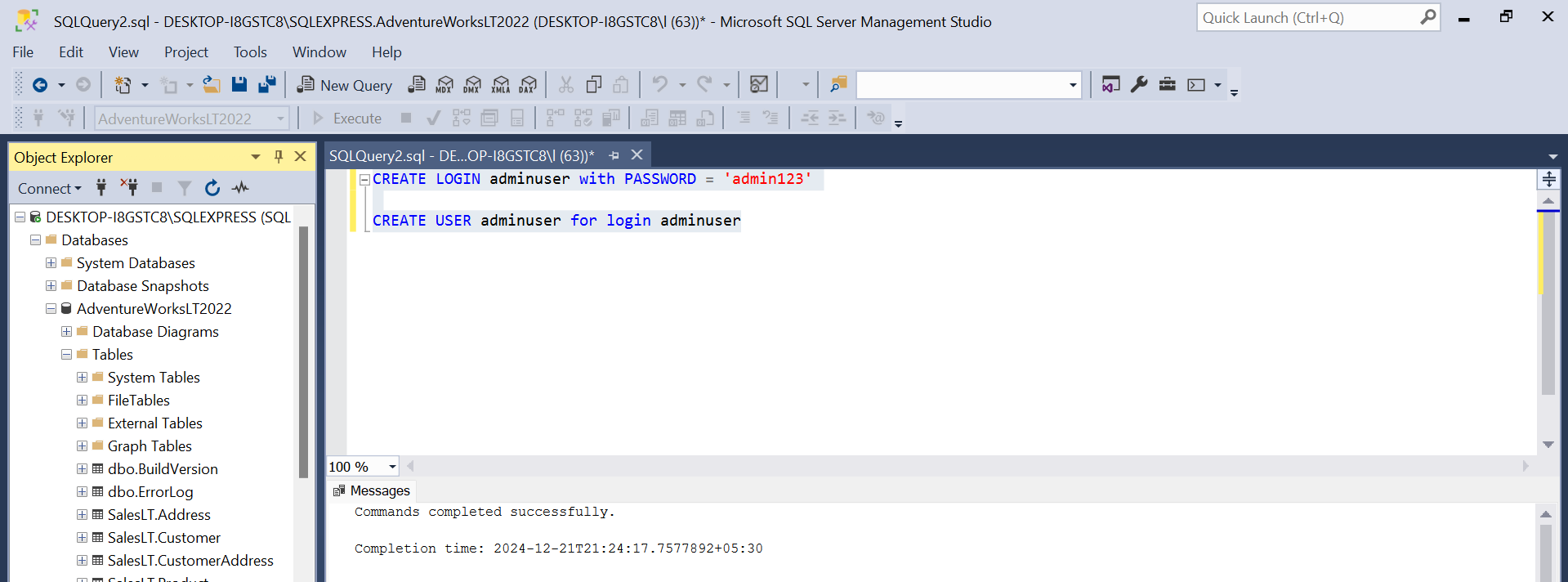
7. SQL Server Management Studio.

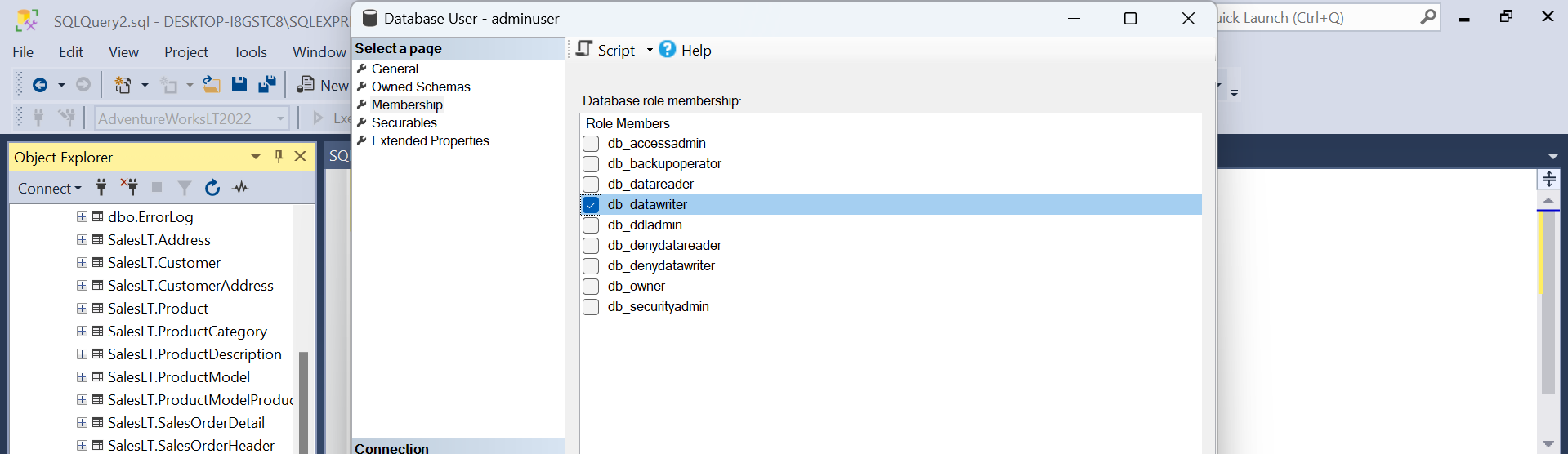
8. Power BI

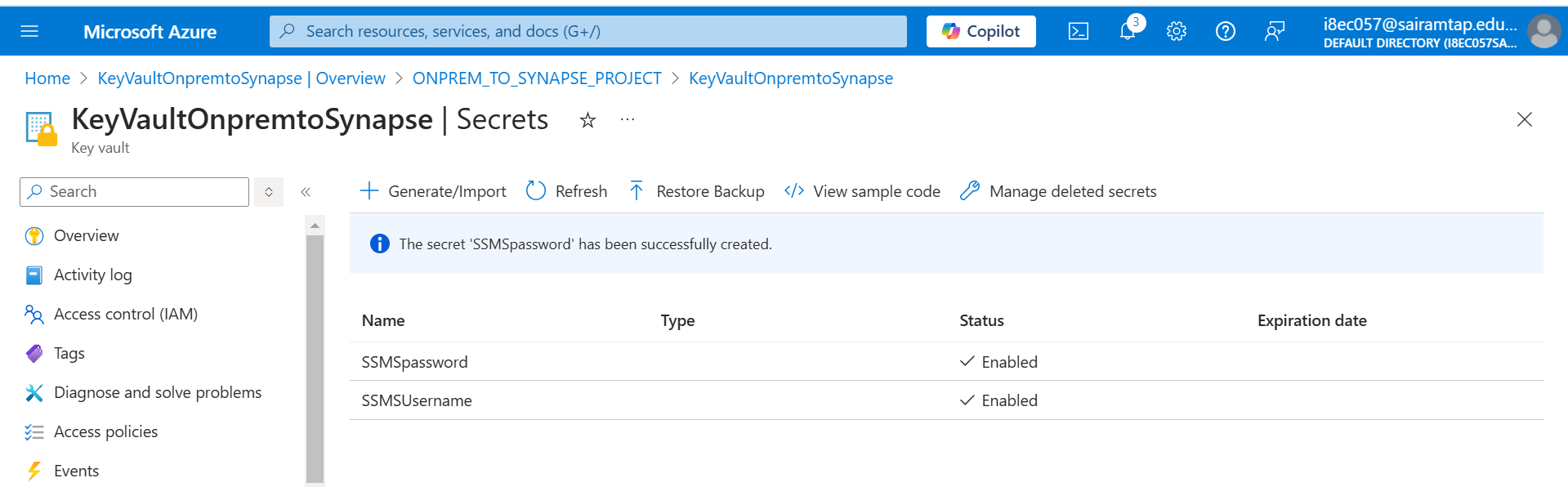
**Dataset required:**

In this project, I have used the microsoft sample database **AdventureWorks.** I have downloaded the sample database from <https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver16&tabs=ssms>



I have created a user login in ssms to connect with Azure.  
  


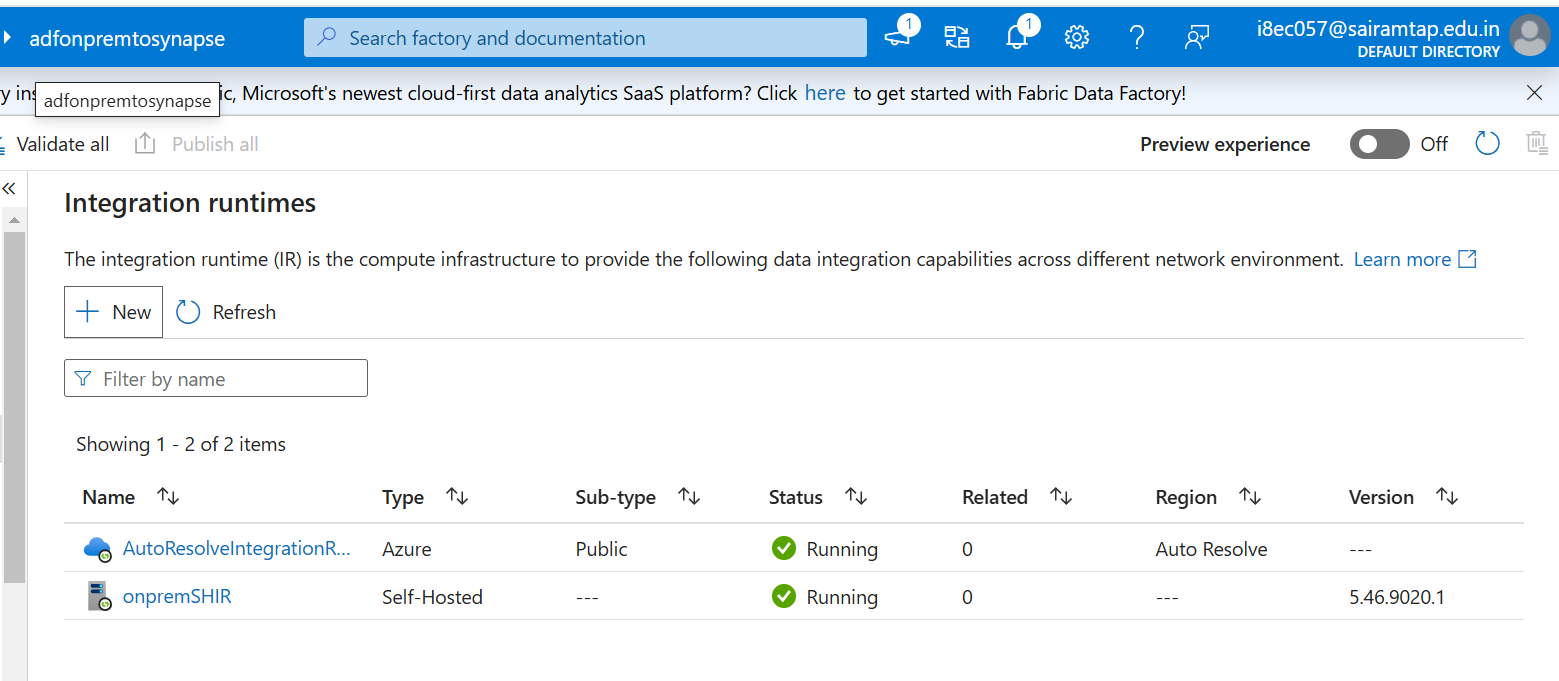
I have provided the datareader access to the created user.  
  


In order to prevent directly using the above login details in Azure resources, I have created a Azure Key Vault to store the above details securely.  
  


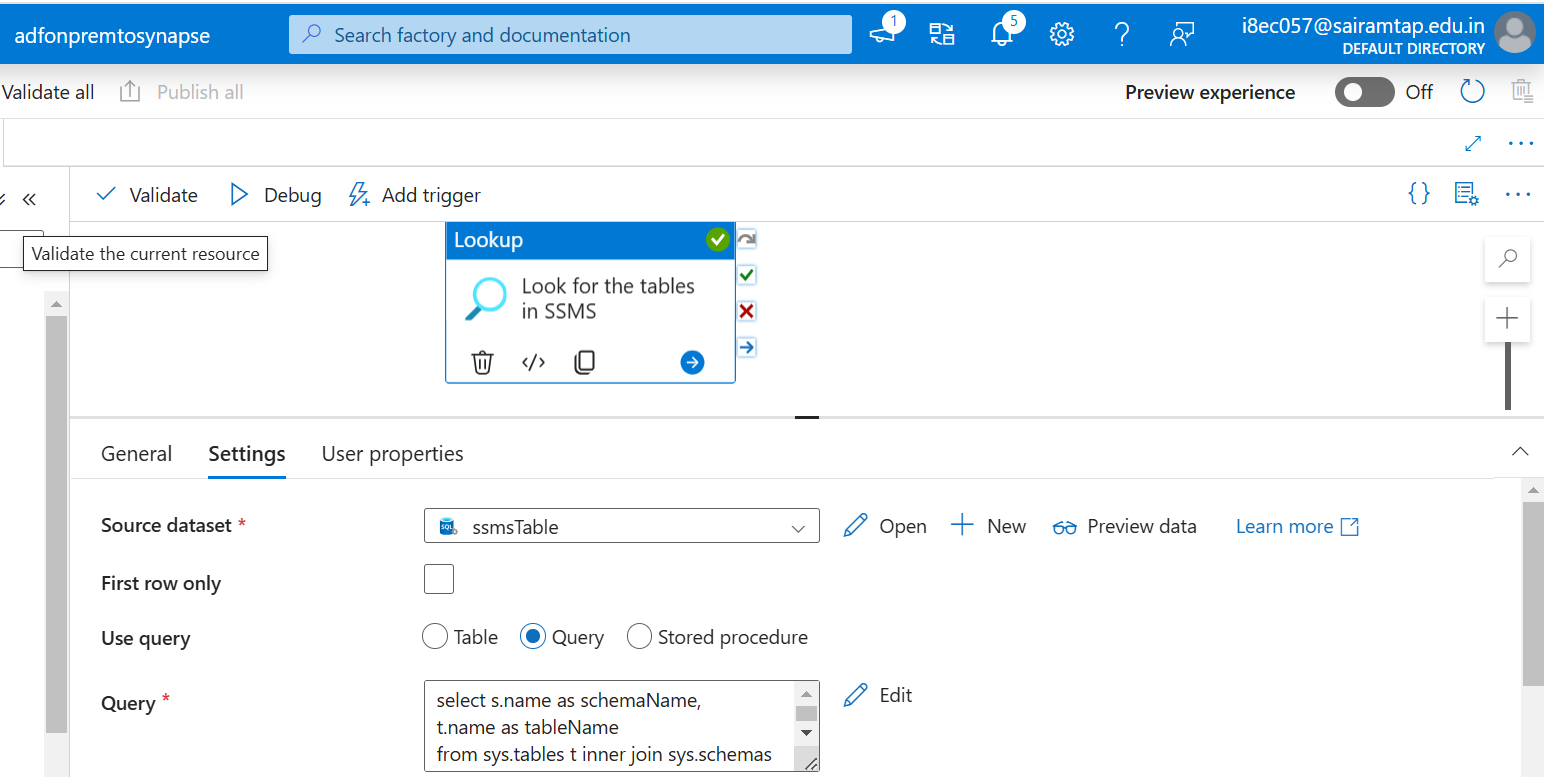
**PHASE 1:**

Data Ingestion

I have created a dynamic pipeline to read all the tables from the SSMS database and copy the tables to the Azure data lake gen2 storage account.

I have created a Self-hosted integration runtime, to connect my on-prem machine with Azure.  
  


Created a Lookup activity to read the tables name available in the SSMS Adventure Database.



Query:

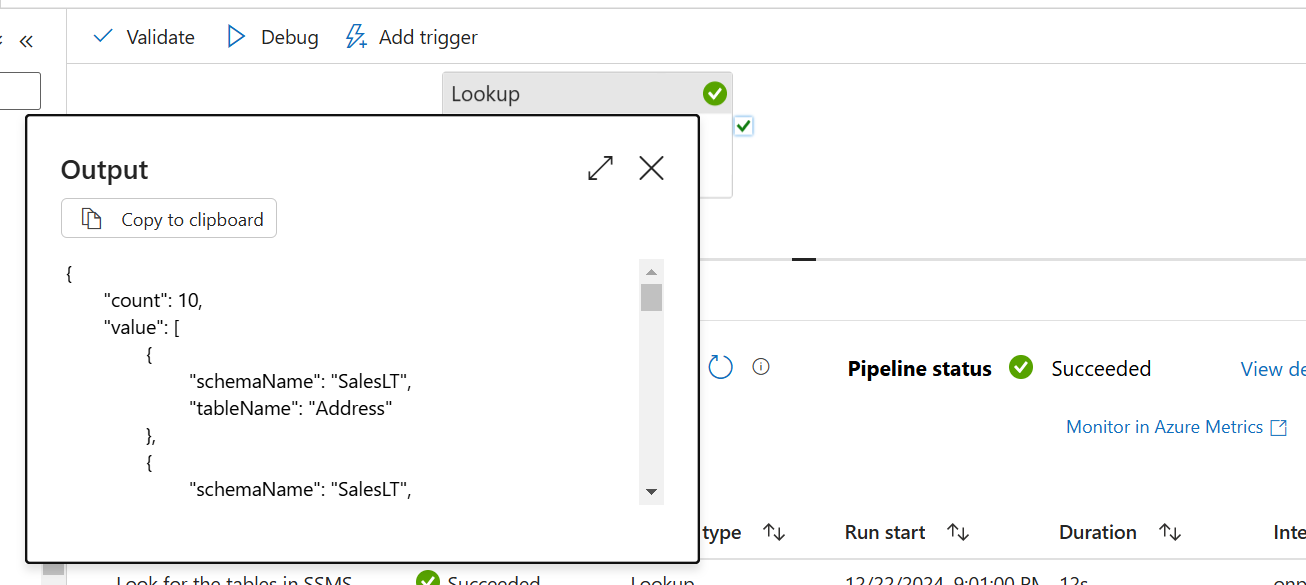
select s.name as schemaName,

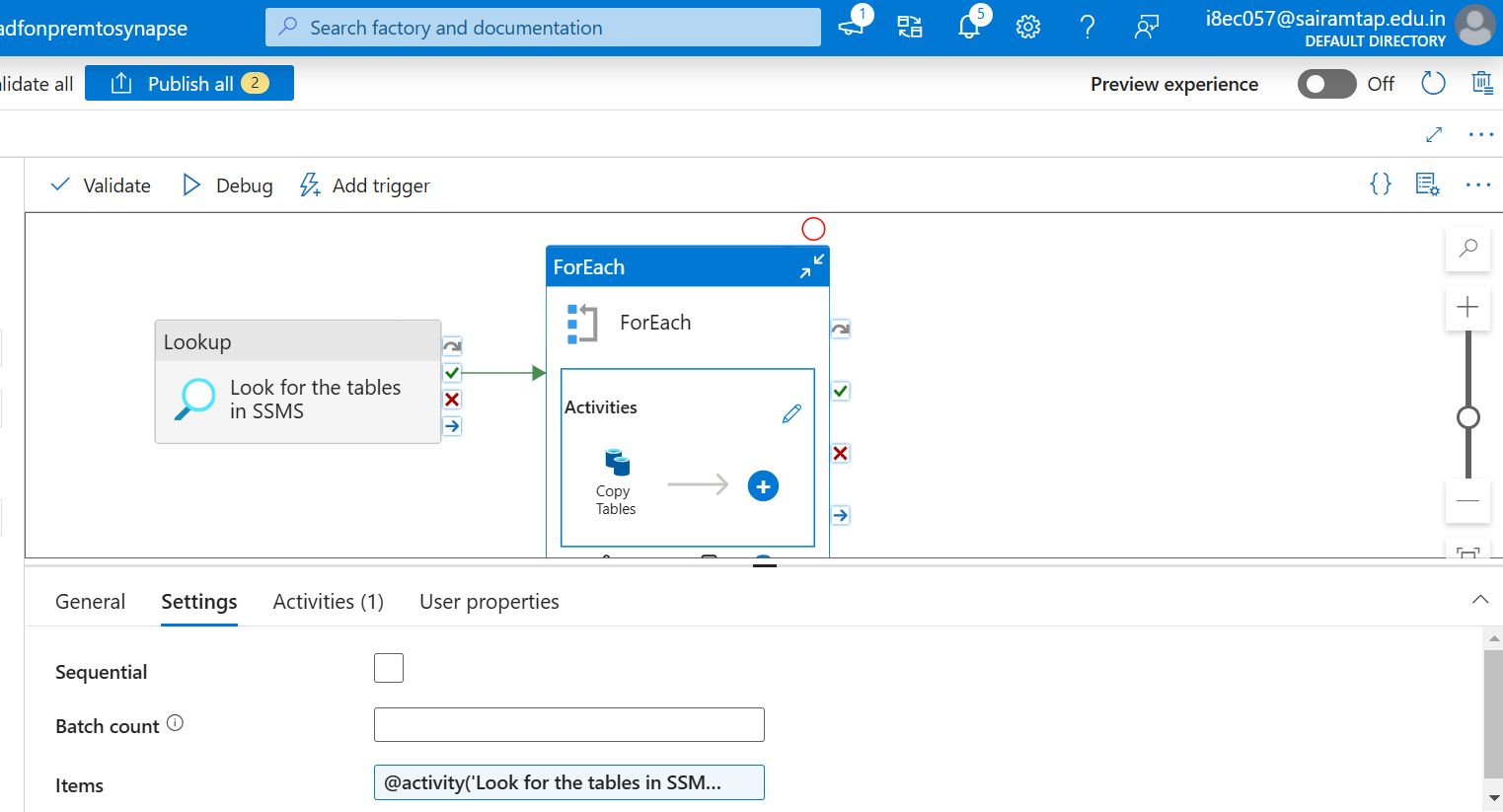
t.name as tableName

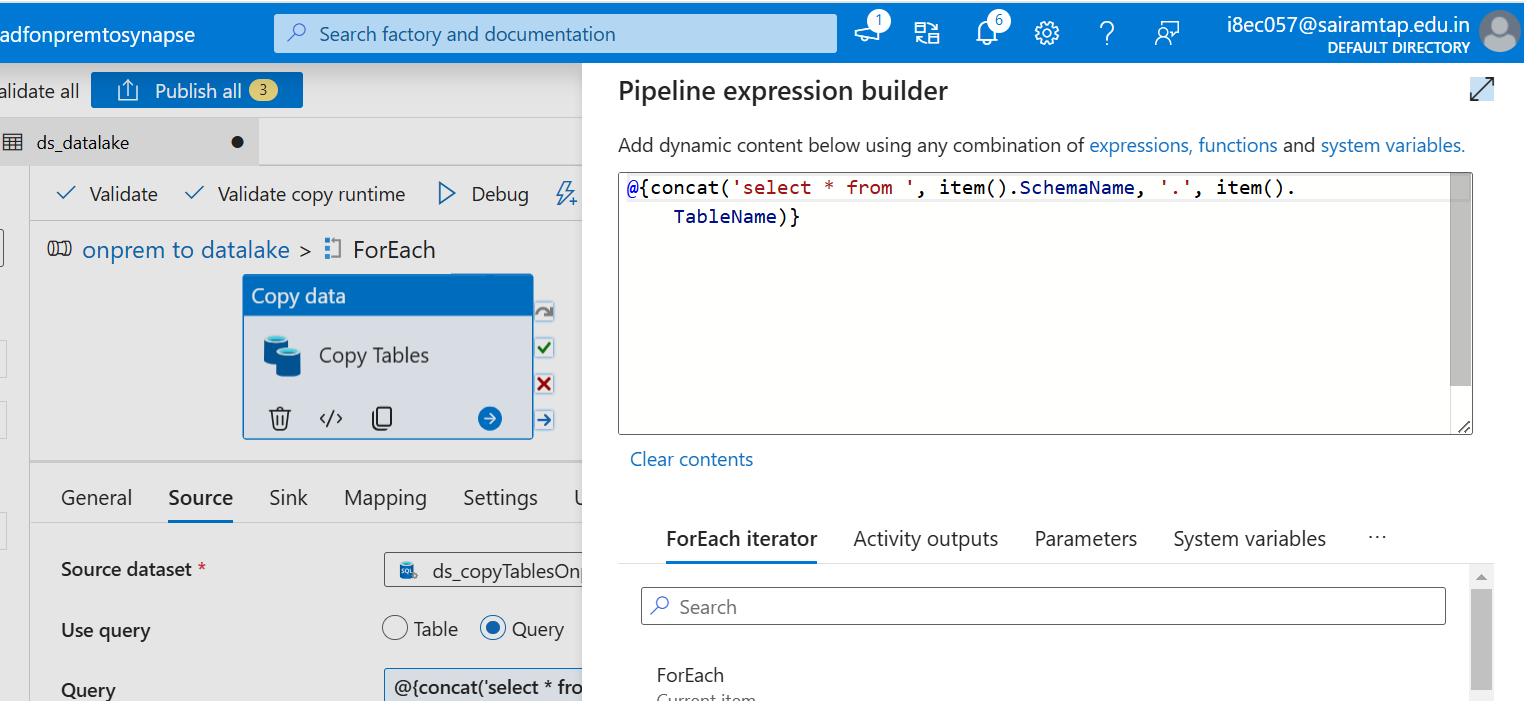
from sys.tables t inner join sys.schemas s

on t.schema\_id = s.schema\_id

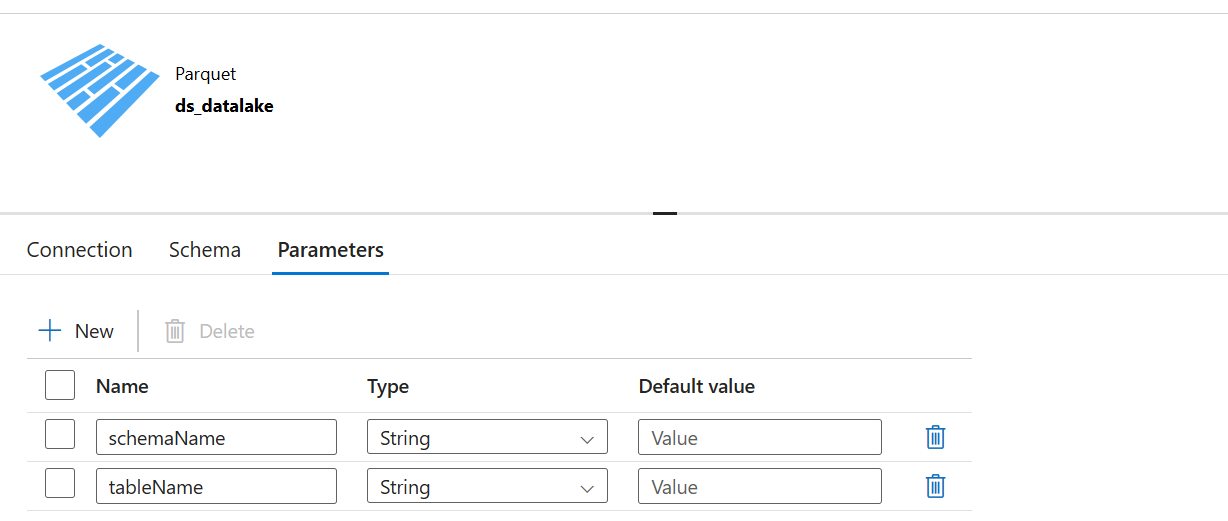
where s.name = 'SalesLT';

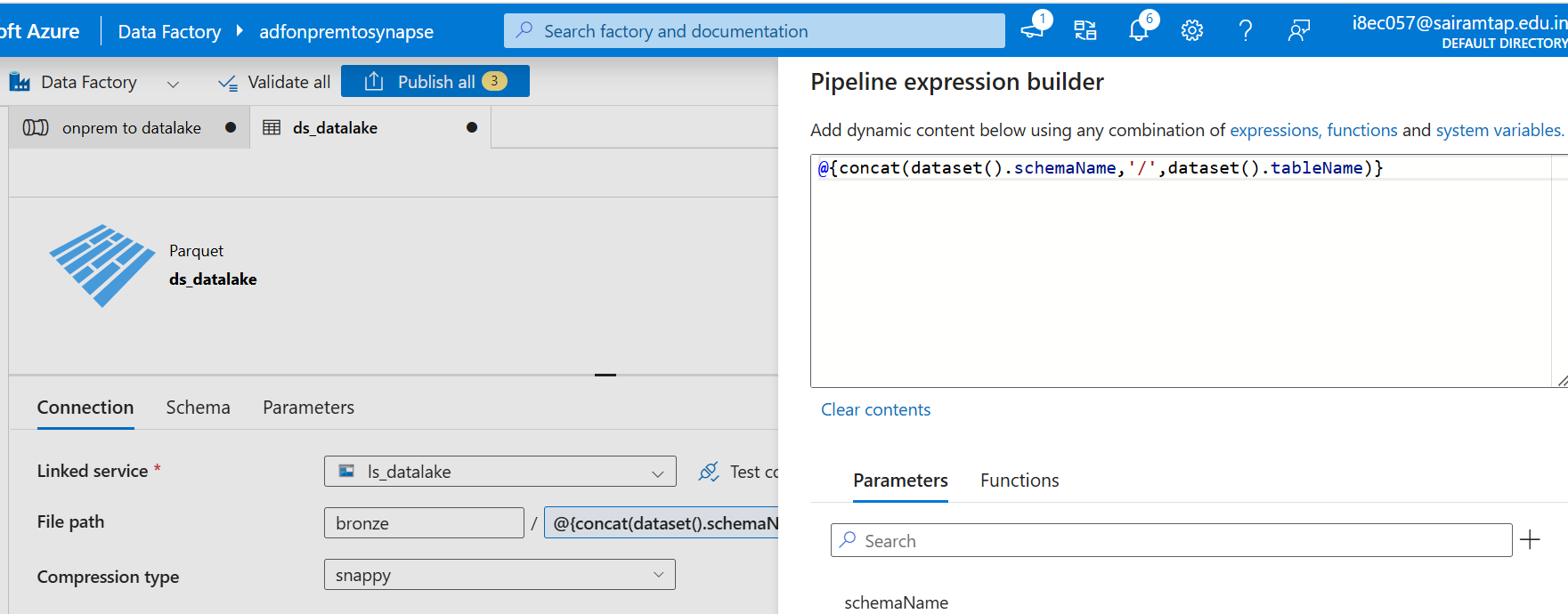
The output of the lookup activity:  
  


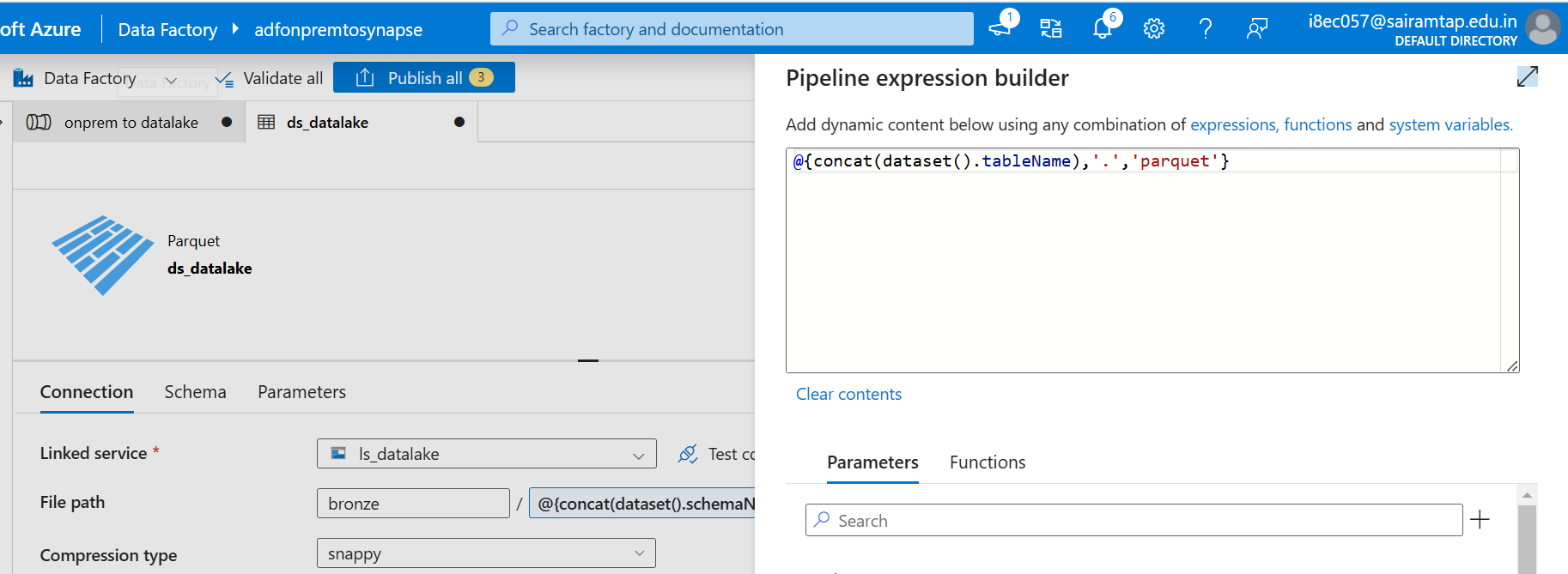
I have created a ForEach activity, to read the output of the lookup activity one by one.  
  


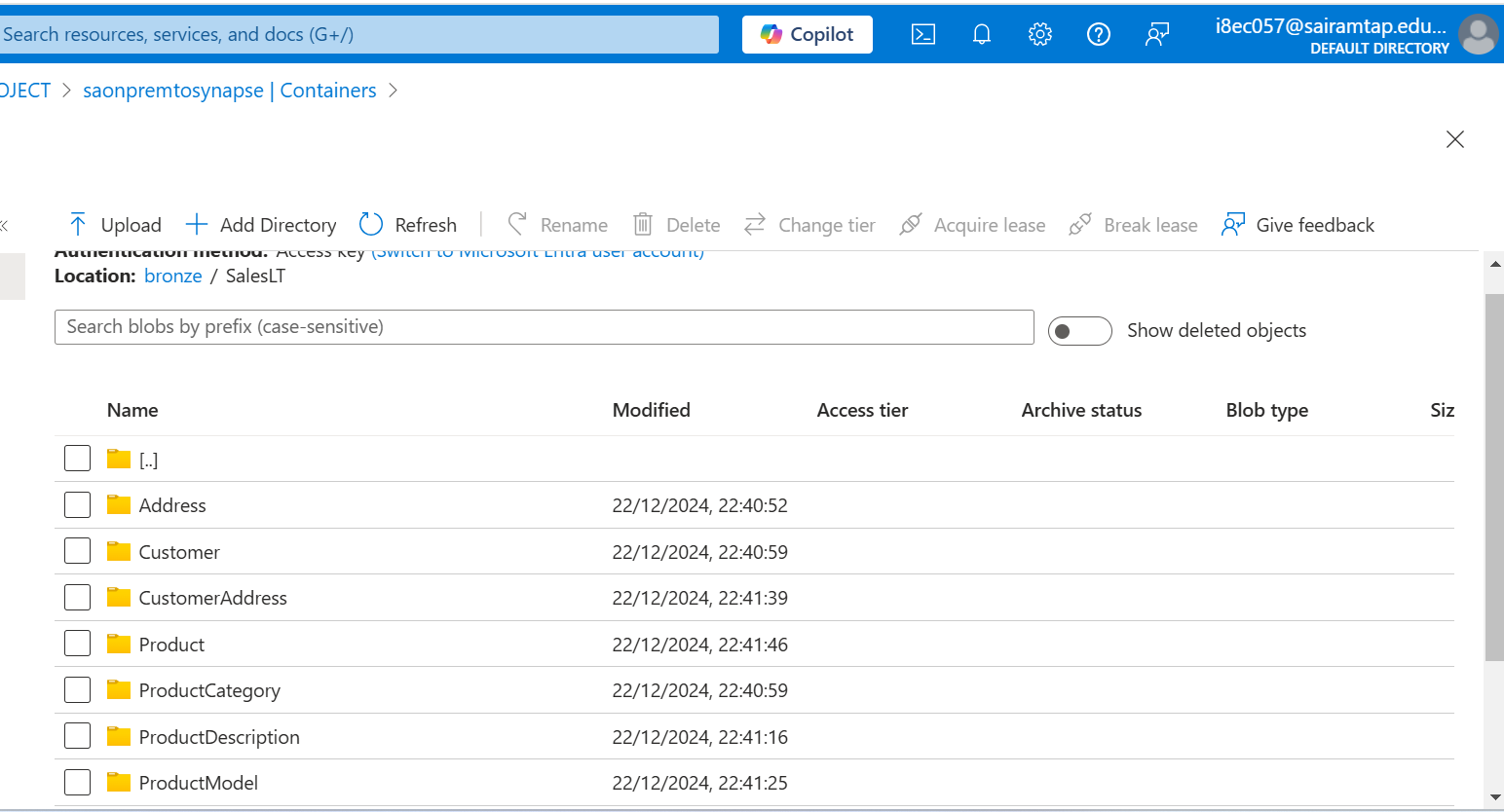
I have created a copy activity inside the ForEach activity, to copy all the tables to datalake.  
  


To create a folder like structure below:  
  
bronze / schemaName / tableName / tableName.parquet  
  
I have created a two parameter in sink dataset.







Now successfully copied the data to the bronze container:  
  


**PHASE 2:**

Data Transformation

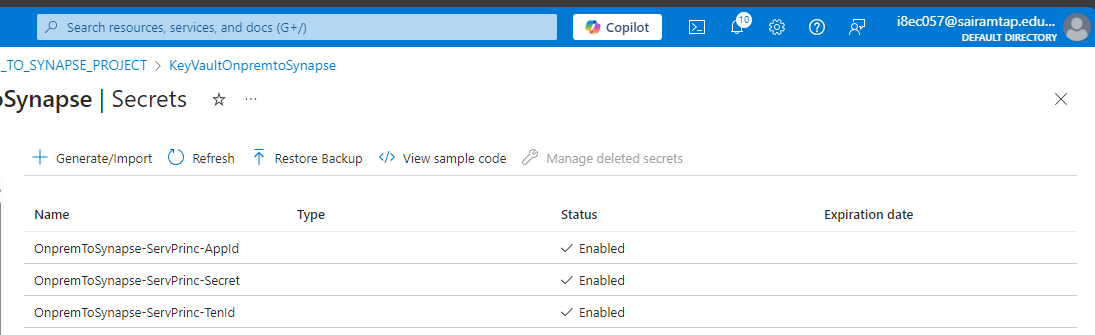
Here in phase 2, I have transformed the data in the bronze container using Databricks and moved the transformed data to silver container.  
  
Databricks should have the access to read the files from bronze container (adls gen2 storage account), so I have created a Service principal in Microsoft entra ID to provide the access for the ADLS gen2 storage account to Databricks.  
  
I have assigned the blob contributor role to the Service principal in the ADLS gen2 storage account.

To connect the ADLS gen2 to Databricks, we need the below details:  
  
Application\_id of ServicePrincipal

Directory\_id of ServicePrincipal

Secret of ServicePrincipal

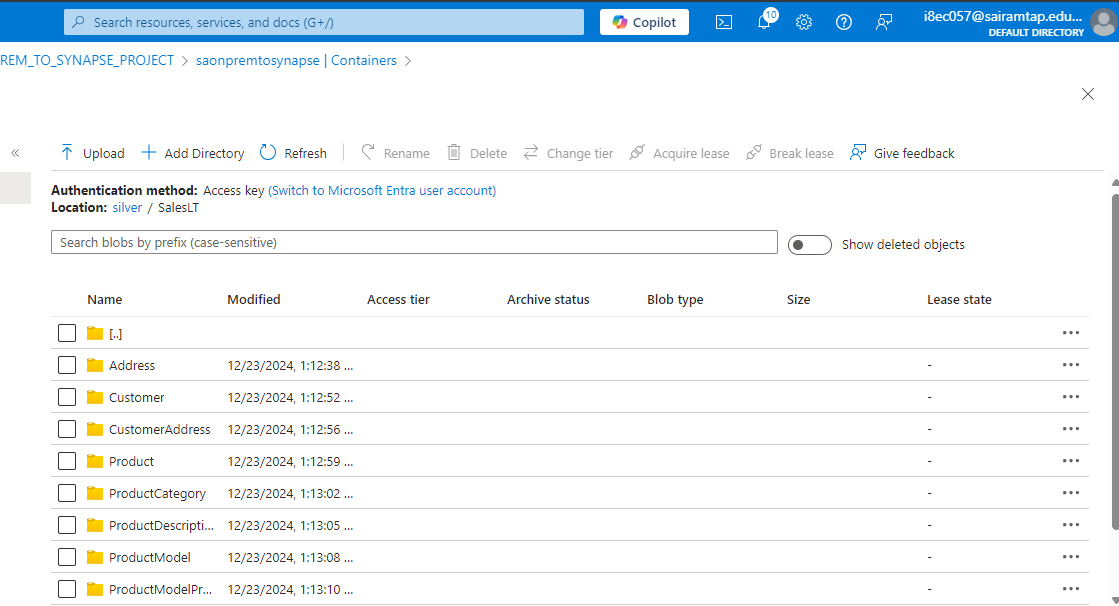
In order to prevent directly using the above details in databricks notebook, I have created a Azure Key Vault to store the above details securely.



In order to use these secrets in the databricks notebook. First we need to create a azure key vault backed secret scope in databricks.  
  
To create a azure key vault backed secret scope in databricks:

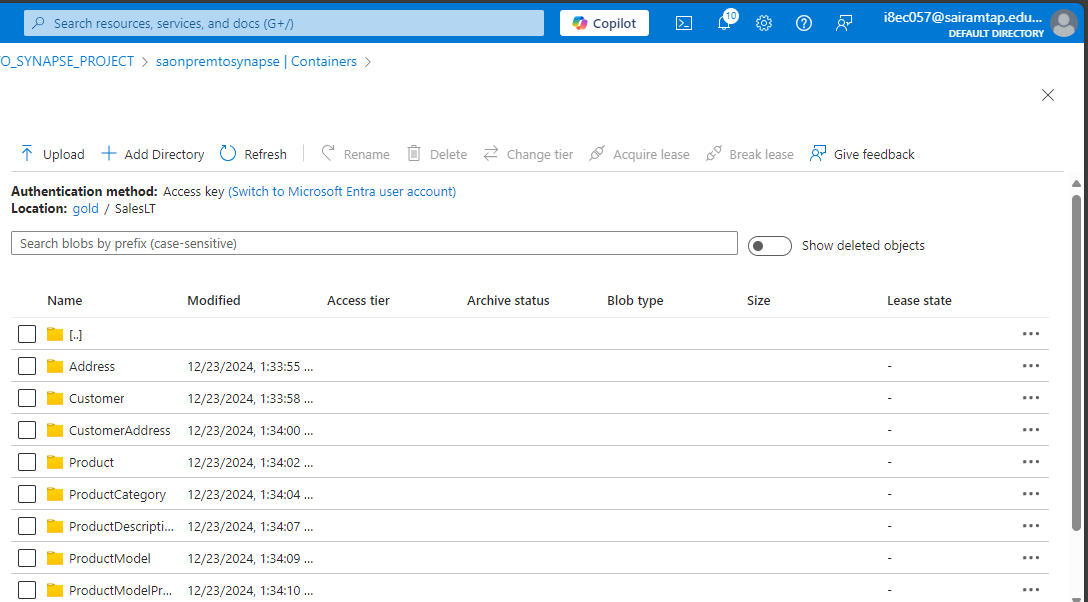
Add the **secrets/createScope** at the end of the databricks instance url.  
  
Go to https://<databricks-instance>#**secrets/createScope**

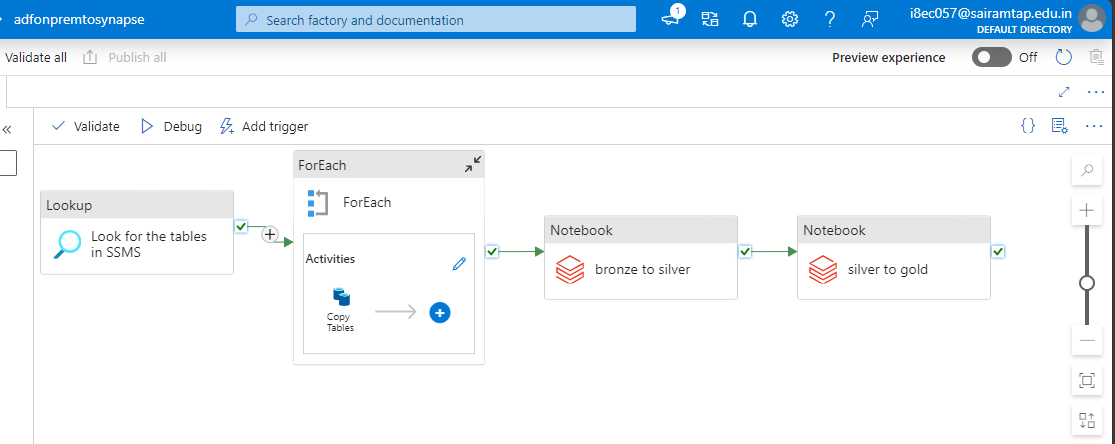
REFER THE **setup** notebook:

REFER THE **bronze\_to\_silver** notebook:  
  
Now the data are transformed and stored in the Silver container in delta format.  
  


Further transformed the data and moved the data to gold container.

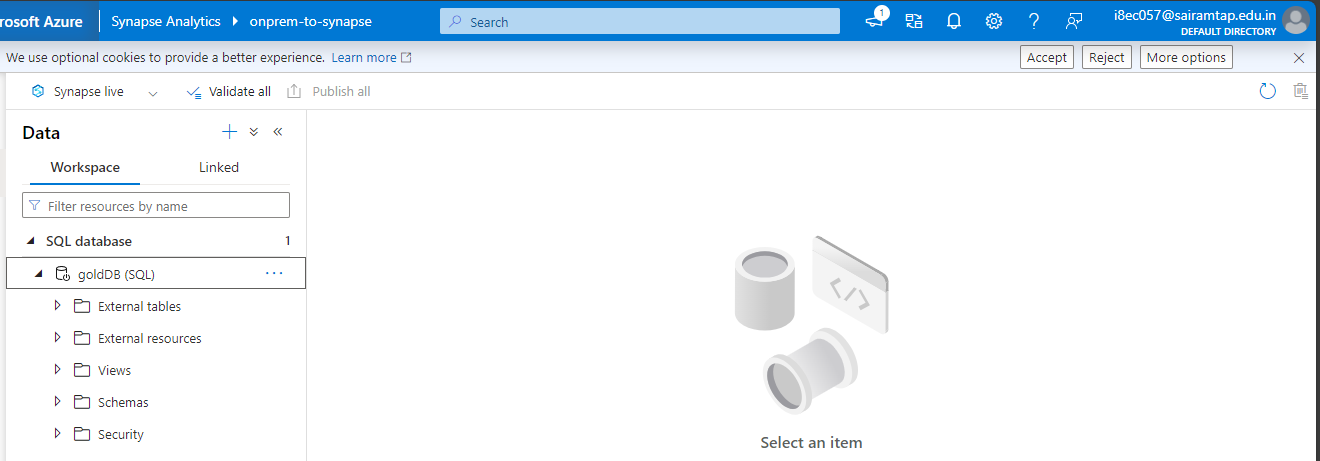
REFER THE **silver\_to\_gold** notebook:  
  
Now the data are moved to the Gold container in Delta Format.

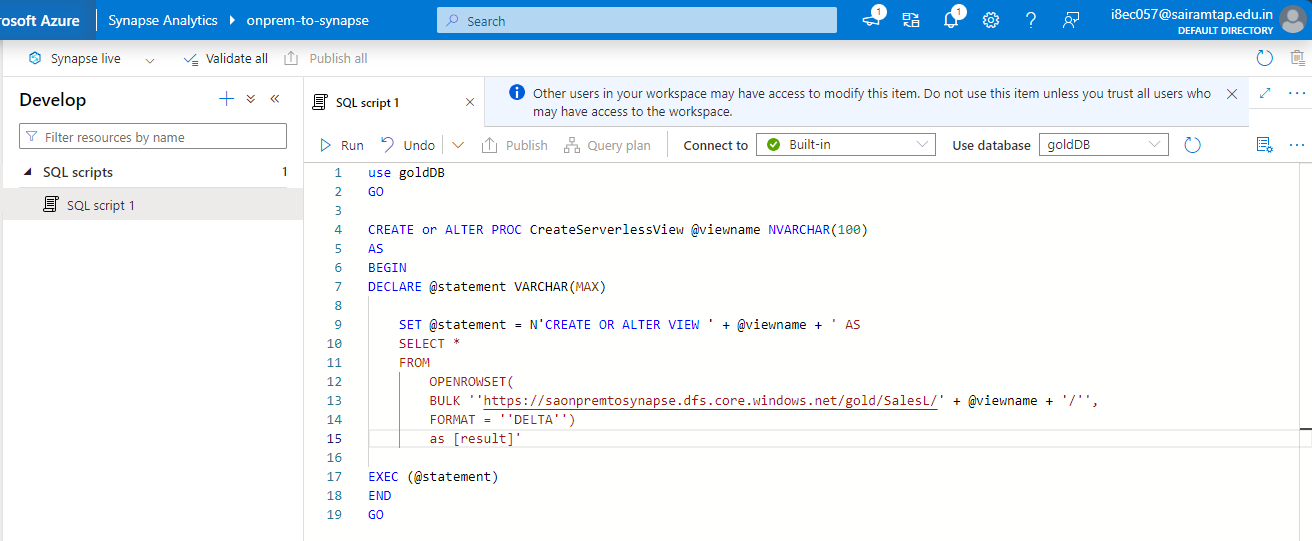


I have attached the databricks notebooks in the Azure Data factory Pipeline.  
  


**PHASE 3:**

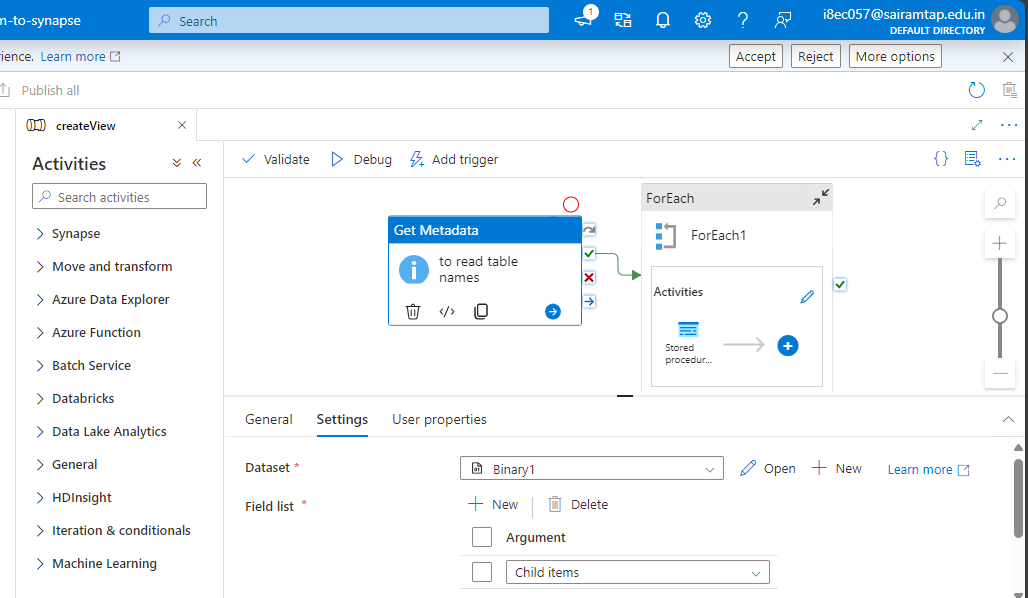
Data Loading

Here in Phase 3, I have created a serverless database in the Azure Synapse Analytics.  
  


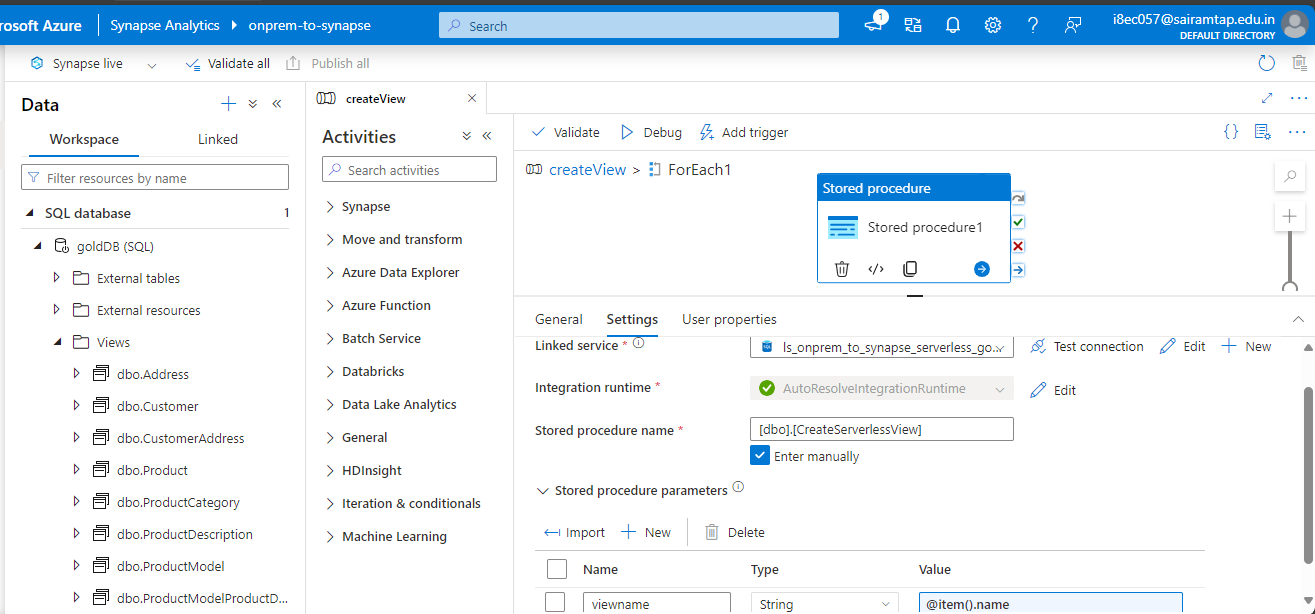
I have created a stored procedure to create views for all the tables in the gold container.  
  


I have created a dynamic pipeline in Synapse Analytics, to create a view for all the tables with the above created stored procedure.

By using GetMeta data activity, I got the list of the filenames in the gold container. With the help of ForEach activity, reading the filenames from the output of the GetMetadata activity one by one and created a StoredProcedure activity inside the ForEach activity.



Finally, created the views in Synapse Analytics for all the tables present in the gold container with this dynamic pipeline.  
  
If there is an change in the data in the gold container, it will automatically reflect in the view also. We need to run the view pipeline only if there is an change in the schema only.



To connect with Power BI, I have installed power BI desktop on my machine.

I have connected my power BI with Azure Synapse as below:  
