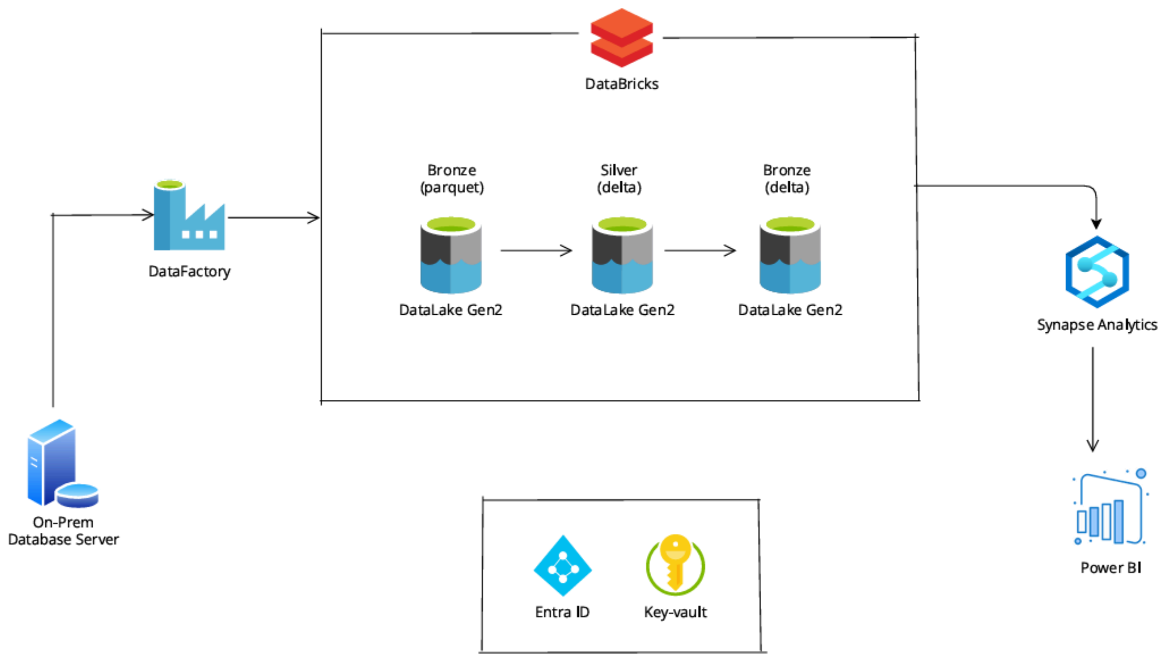
Architecture:



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.

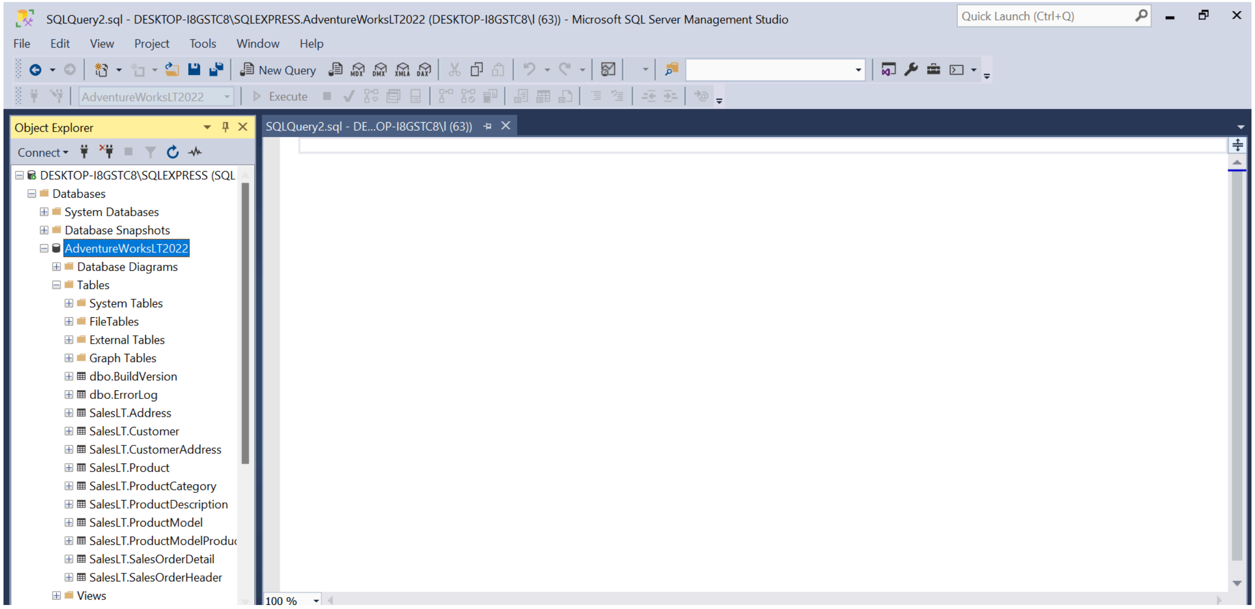
6. Create a Service Principal to connect the ADLS gen2 storage account with Databricks.

7. SQL Server Management Studio.

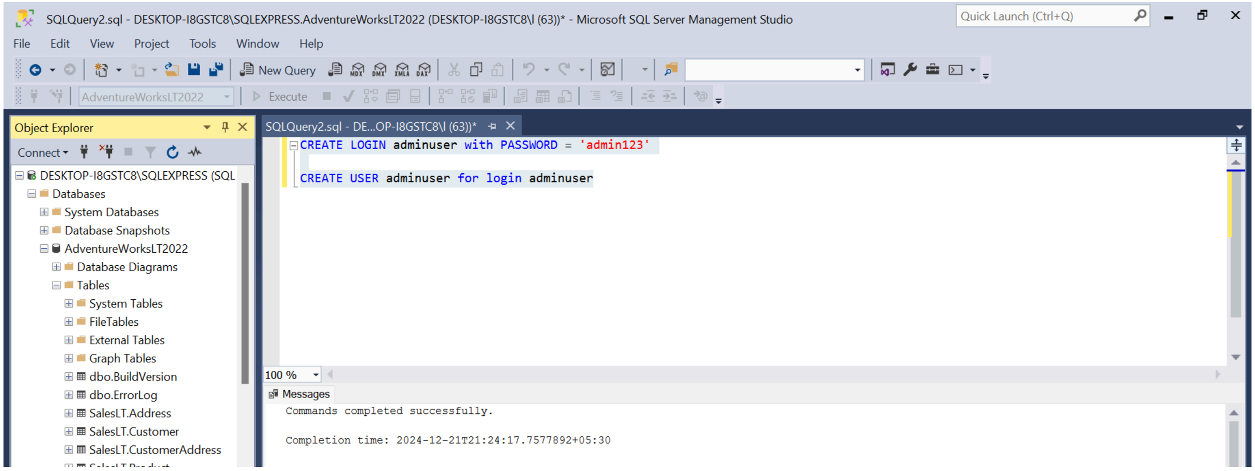
8. Power BI

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



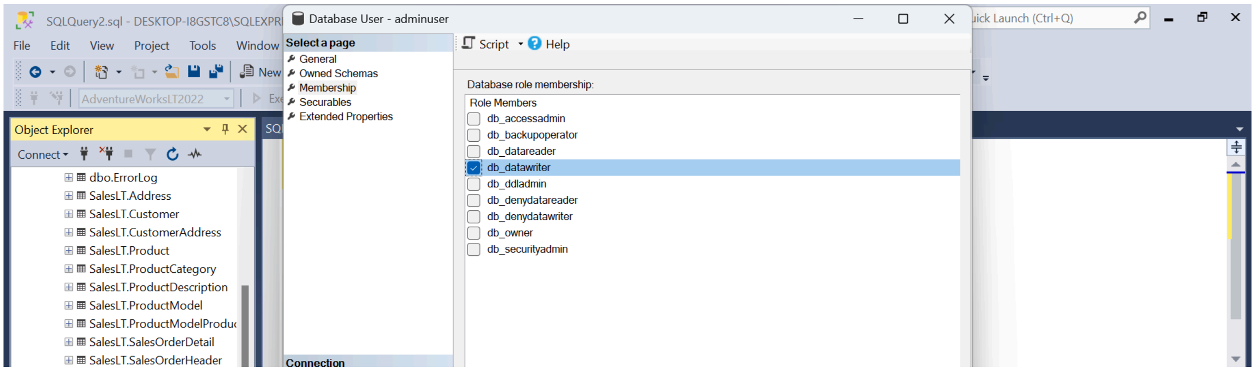
I have created a user login in ssms to connect with Azure.



```
CREATE LOGIN adminuser with PASSWORD = 'admin123'

CREATE USER adminuser for login adminuser
```

```
Commands completed successfully.

Completion time: 2024-12-21T21:24:17.7577892+05:30
```

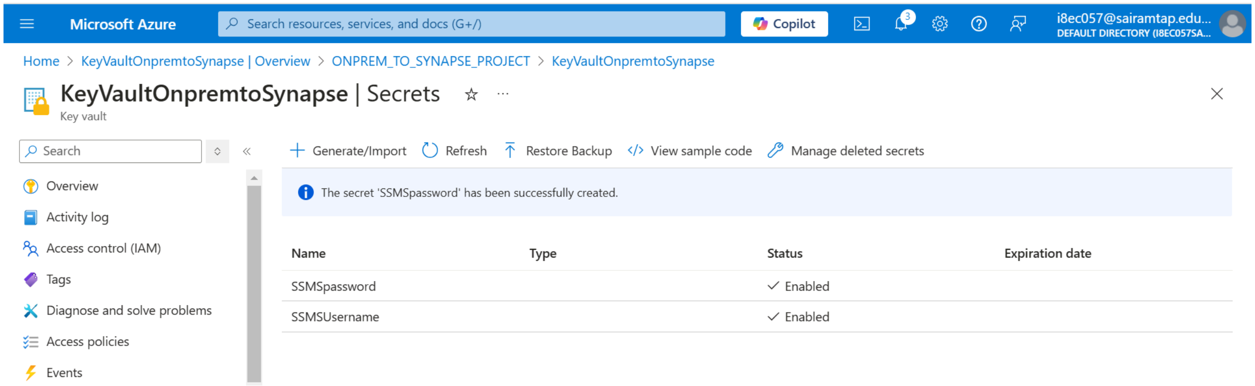
I have provided the datareader access to the created user.



Database role membership:

Role Members
- ☐ db_accessadmin
- ☐ db_backupoperator
- ☐ db_datareader
- ☑ db_datawriter
- ☐ db_ddladmin
- ☐ db_denydatareader
- ☐ db_denydatawriter
- ☐ db_owner
- ☐ db_securityadmin

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.


Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
i8ec057@sairamtap.edu...
DEFAULT DIRECTORY (I8EC057SA...
Home > KeyVaultOnpremtoSynapse | Overview > ONPREM_TO_SYNAPSE_PROJECT > KeyVaultOnpremtoSynapse
KeyVaultOnpremtoSynapse | Secrets
Key vault
Search
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Access policies
Events
Generate/Import
Refresh
Restore Backup
View sample code
Manage deleted secrets

The secret 'SSMSpassword' has been successfully created.

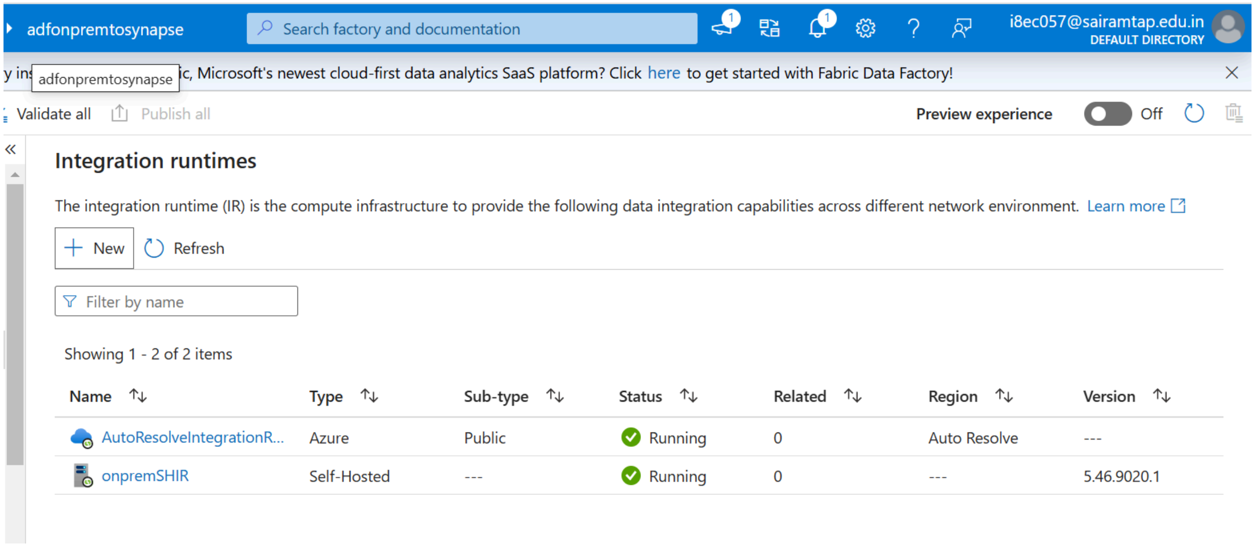
Name	Type	Status	Expiration date
SSMSpassword		✓ Enabled	
SSMSUsername		✓ Enabled	

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.



y in c, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started with Fabric Data Factory!

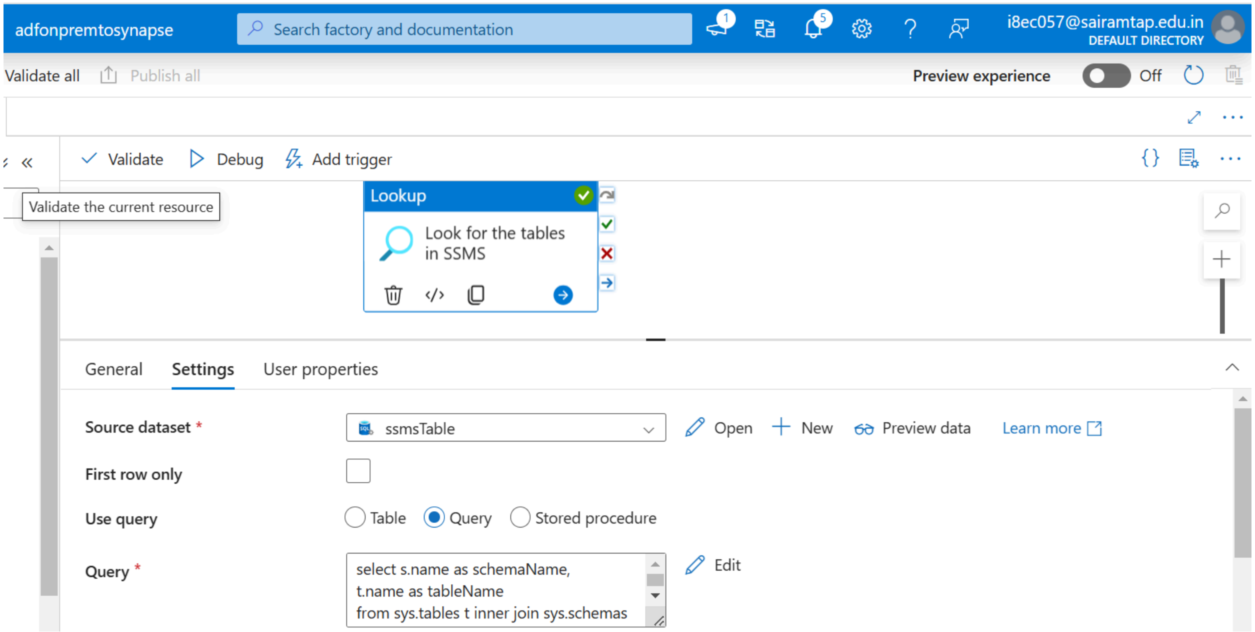
Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment. Learn more

Showing 1 - 2 of 2 items

Name	Type	Sub-type	Status	Related	Region	Version
AutoResolveIntegrationR...	Azure	Public	✓ Running	0	Auto Resolve	---
onpremSHIR	Self-Hosted	---	✓ Running	0	---	5.46.9020.1

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



Source dataset * ssmsTable

First row only □

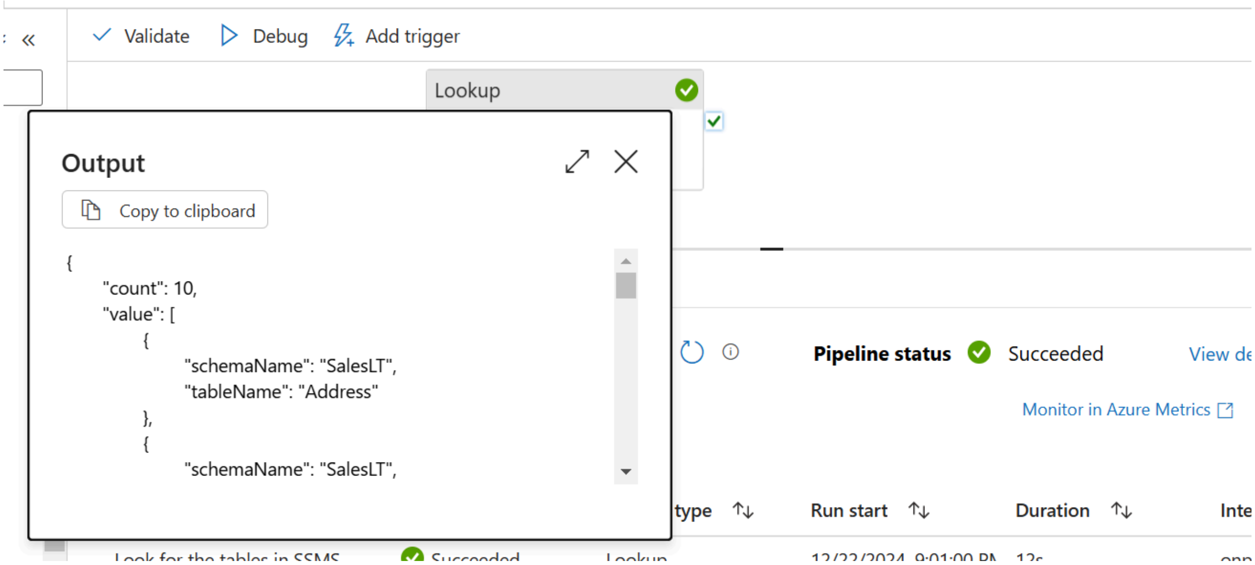
Use query ○ Table ● Query ○ Stored procedure

Query *

```
select s.name as schemaName,
t.name as tableName
from sys.tables t inner join sys.schemas
```

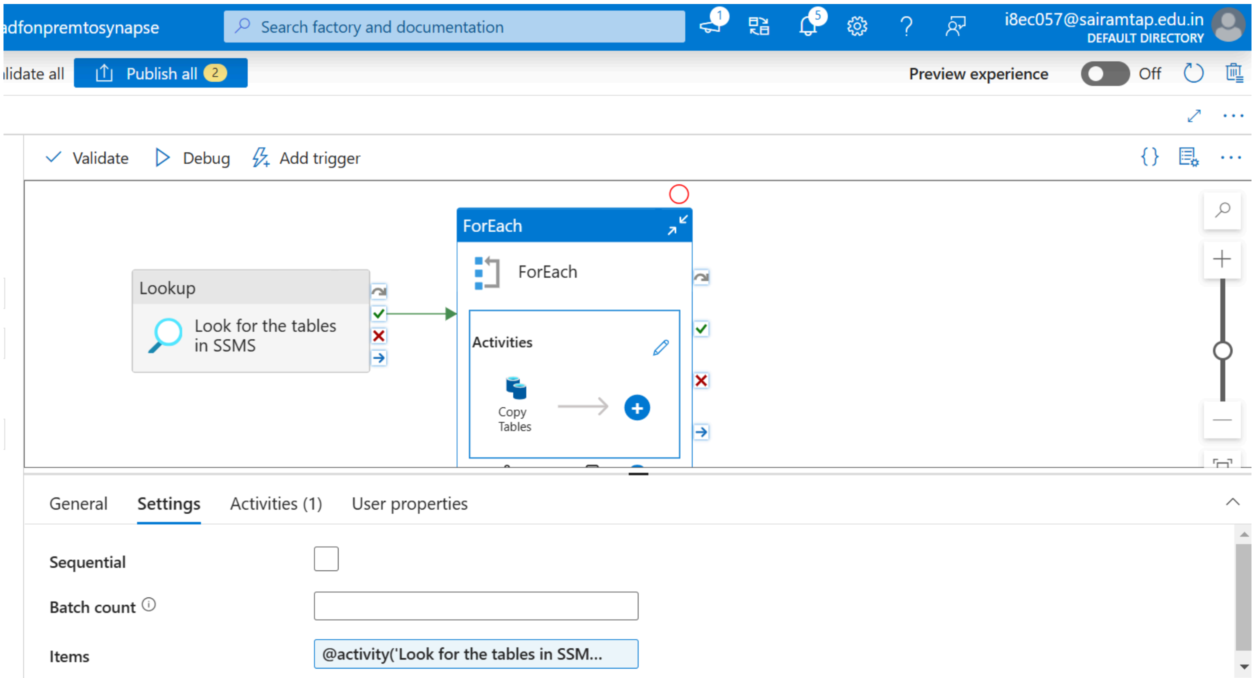
Query:
select s.name as schemaName,
t.name as tableName
from sys.tables t inner join sys.schemas s
on t.schema_id = s.schema_id
where s.name = 'SalesLT';

The output of the lookup activity:



```
{
      "count": 10,
      "value": [
            {
                  "schemaName": "SalesLT",
                  "tableName": "Address"
            },
            {
                  "schemaName": "SalesLT",
```

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

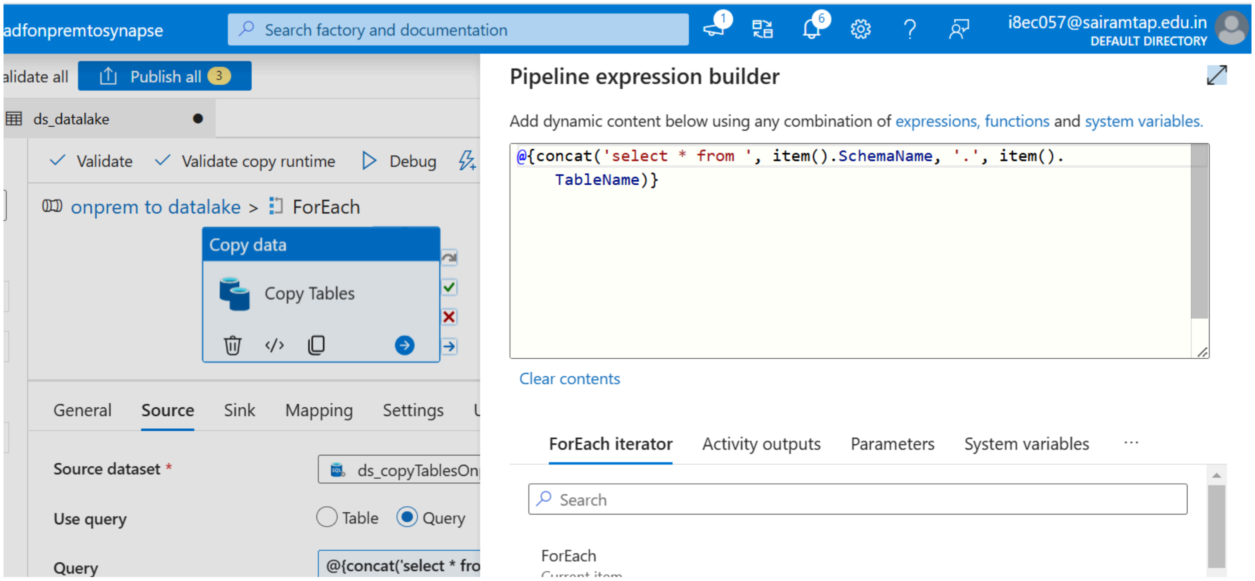


Sequential □

Batch count

Items @activity('Look for the tables in SSM...

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



Add dynamic content below using any combination of expressions, functions and system variables.

```
@{concat('select * from ', item().SchemaName, '.', item().
    TableName)}
```

Source dataset * ds_copyTablesOn

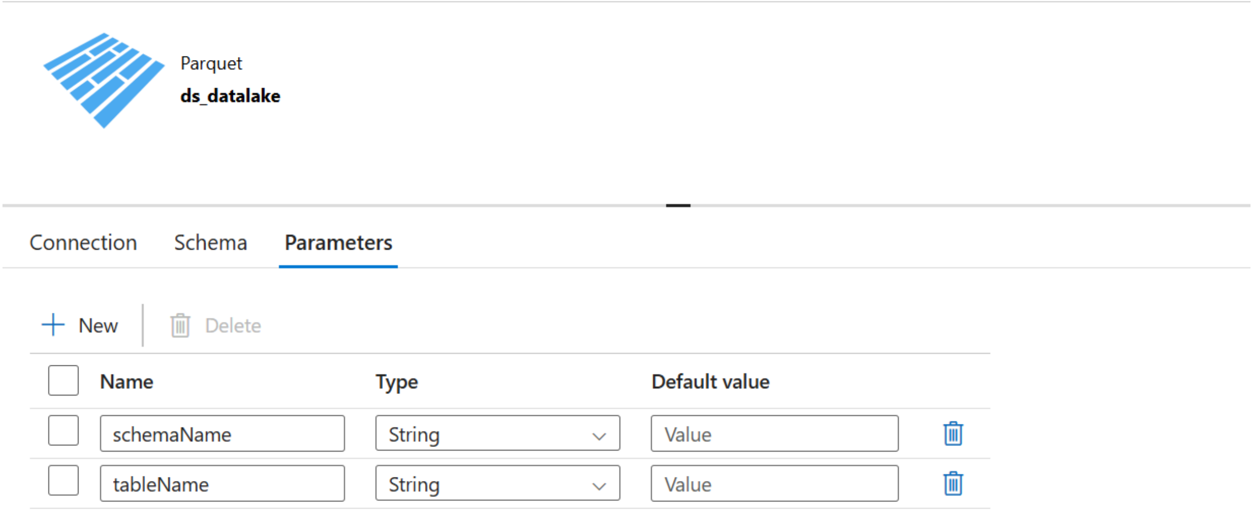
Use query ○ Table ● Query

Query @{concat('select * fro

To create a folder like structure below:

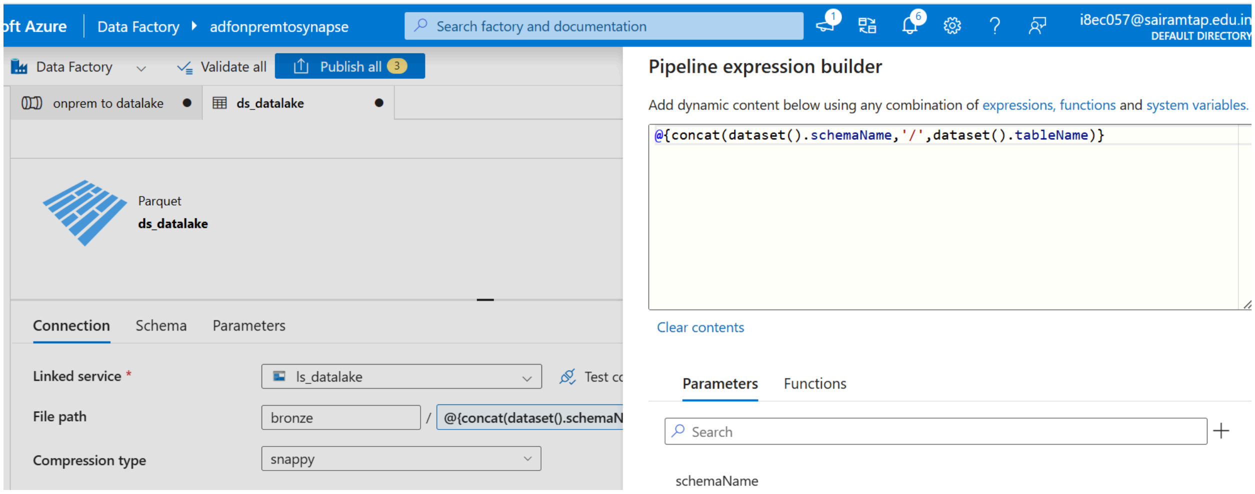
bronze / schemaName / tableName / tableName.parquet

I have created a two parameter in sink dataset.



Parquet
ds_datalake

☐	Name	Type	Default value
☐	schemaName	String	Value
☐	tableName	String	Value



Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

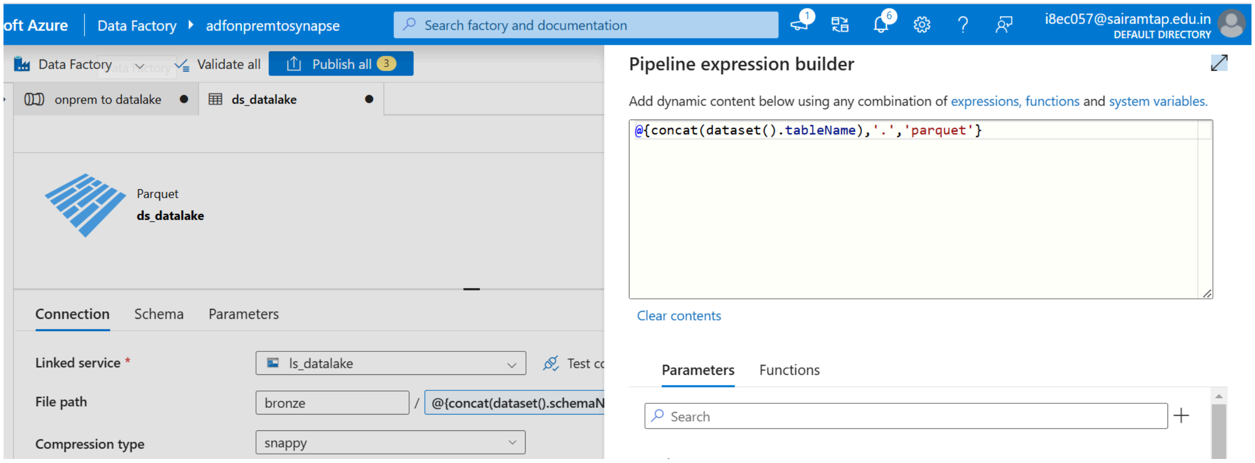
```
@{concat(dataset().schemaName,'/',dataset().tableName)}
```

Linked service * ls_datalake

File path bronze / @{concat(dataset().schemaN

Compression type snappy

schemaName



Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

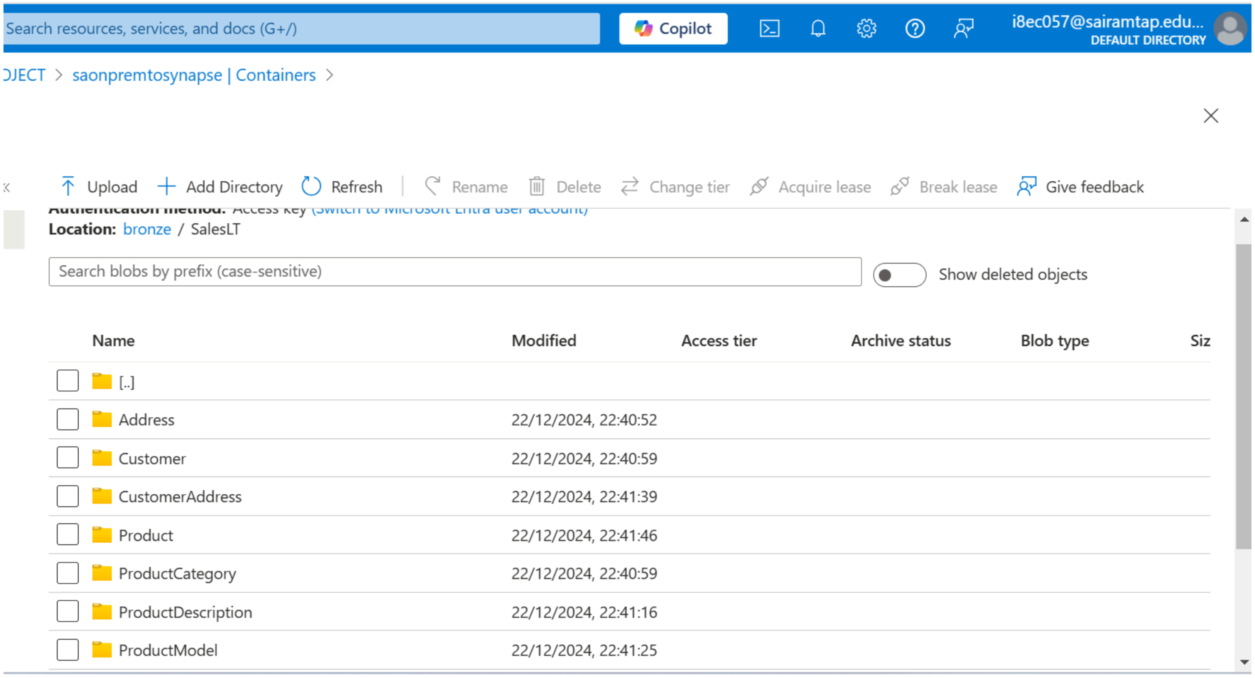
```
@{concat(dataset().tableName),'.','parquet'}
```

Linked service * ls_datalake

File path bronze / @{concat(dataset().schemaN

Compression type snappy

Now successfully copied the data to the bronze container:



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: bronze / SalesLT

	Name	Modified	Access tier	Archive status	Blob type	Siz
□	[..]					
□	Address	22/12/2024, 22:40:52				
□	Customer	22/12/2024, 22:40:59				
□	CustomerAddress	22/12/2024, 22:41:39				
□	Product	22/12/2024, 22:41:46				
□	ProductCategory	22/12/2024, 22:40:59				
□	ProductDescription	22/12/2024, 22:41:16				
□	ProductModel	22/12/2024, 22:41:25				

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.

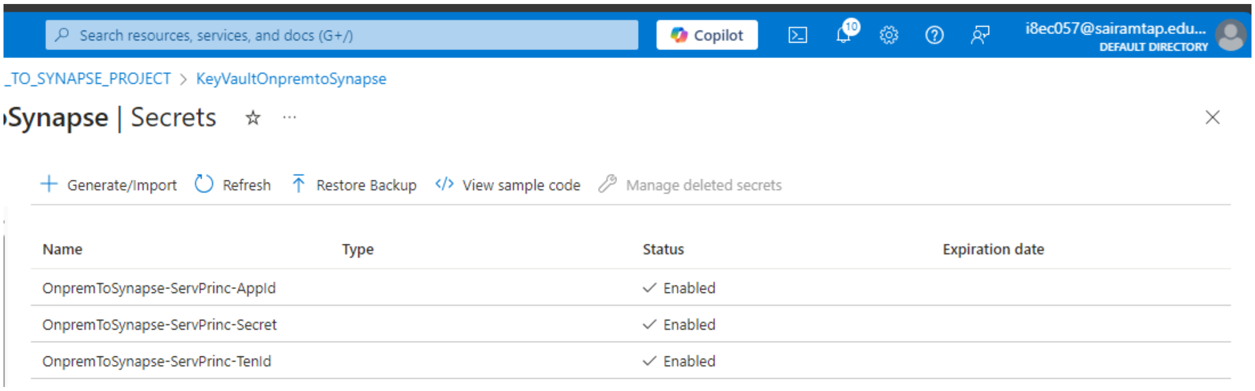
Databricks should have the access to read the files from bronze container (adls gen2 storage account), so I have created a Service principal in Microsoft entra ID to provide the access for the ADLS gen2 storage account to Databricks.

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.



Name	Type	Status	Expiration date
OnpremToSynapse-ServPrinc-AppId		✓ Enabled	
OnpremToSynapse-ServPrinc-Secret		✓ Enabled	
OnpremToSynapse-ServPrinc-TenId		✓ Enabled	

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:

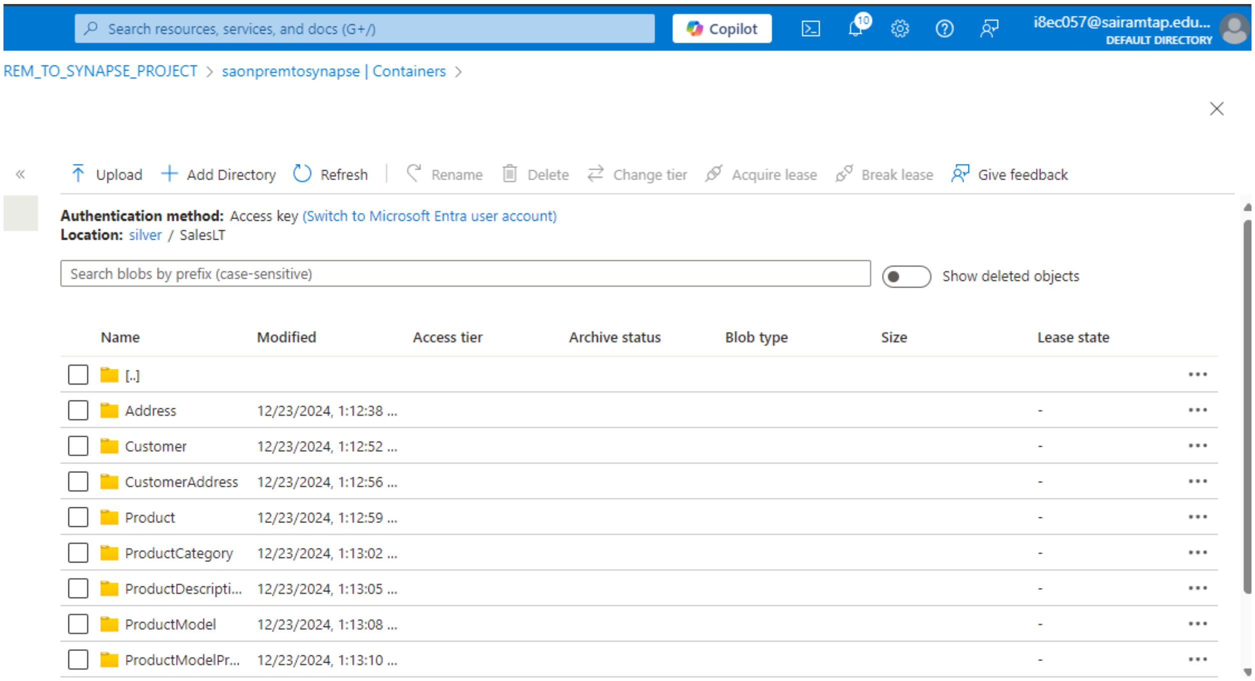
Add the **secrets/createScope** at the end of the databricks instance url.

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.



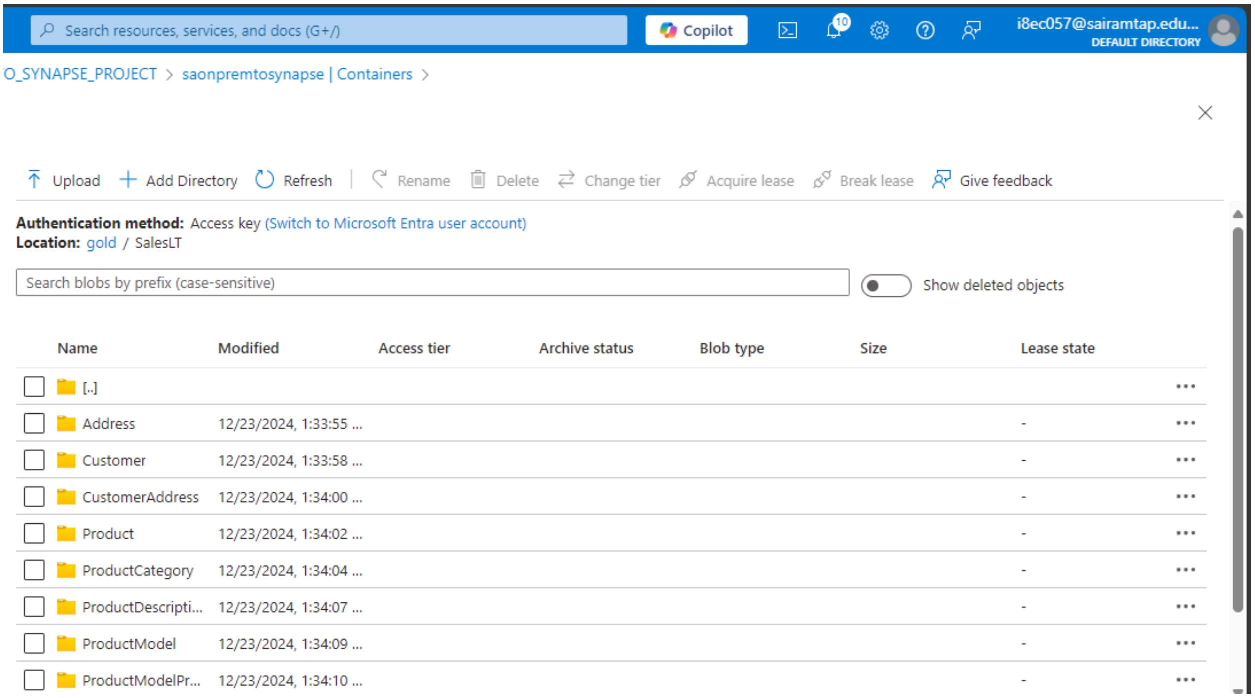
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: silver / SalesLT

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
□	[..]							•••
□	Address	12/23/2024, 1:12:38 ...					-	•••
□	Customer	12/23/2024, 1:12:52 ...					-	•••
□	CustomerAddress	12/23/2024, 1:12:56 ...					-	•••
□	Product	12/23/2024, 1:12:59 ...					-	•••
□	ProductCategory	12/23/2024, 1:13:02 ...					-	•••
□	ProductDescripti...	12/23/2024, 1:13:05 ...					-	•••
□	ProductModel	12/23/2024, 1:13:08 ...					-	•••
□	ProductModelPr...	12/23/2024, 1:13:10 ...					-	•••

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.

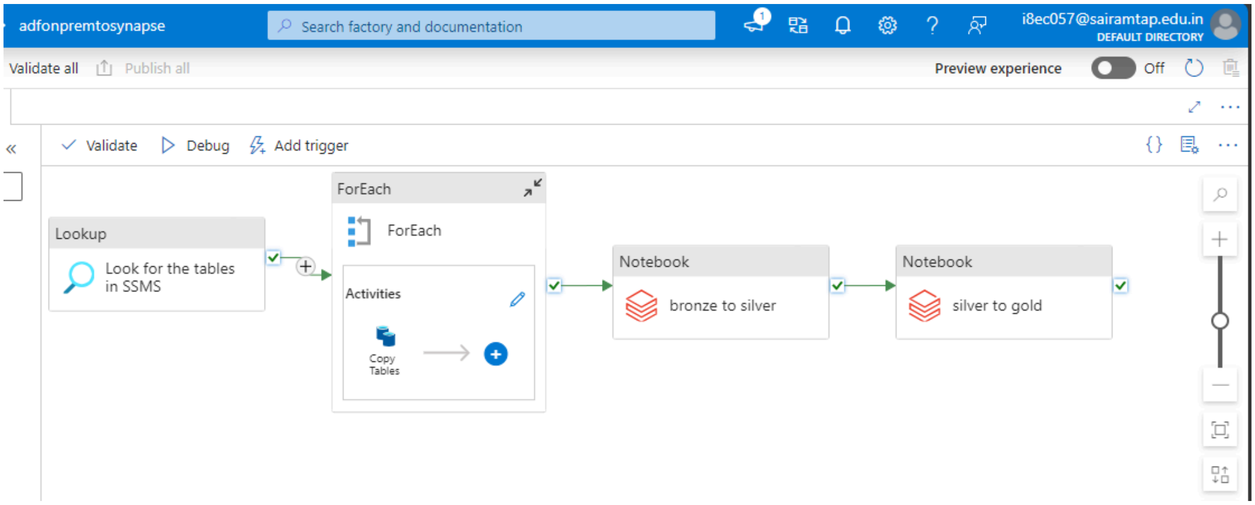


O_SYNAPSE_PROJECT > saonpremtosynapse | Containers >

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: gold / SalesLT

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]						
Address	12/23/2024, 1:33:55 ...					-
Customer	12/23/2024, 1:33:58 ...					-
CustomerAddress	12/23/2024, 1:34:00 ...					-
Product	12/23/2024, 1:34:02 ...					-
ProductCategory	12/23/2024, 1:34:04 ...					-
ProductDescripti...	12/23/2024, 1:34:07 ...					-
ProductModel	12/23/2024, 1:34:09 ...					-
ProductModelPr...	12/23/2024, 1:34:10 ...					-

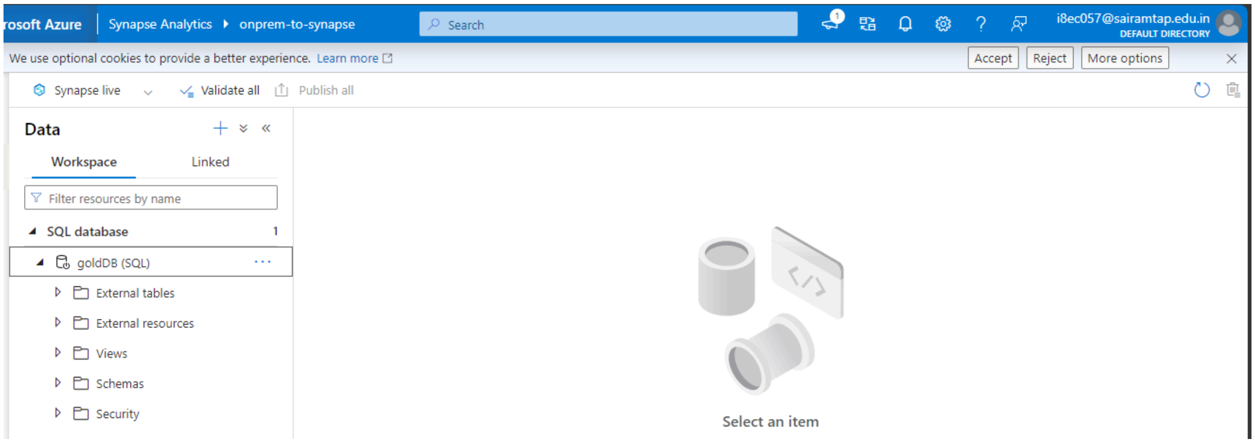
I have attached the databricks notebooks in the Azure Data factory Pipeline.



PHASE 3:

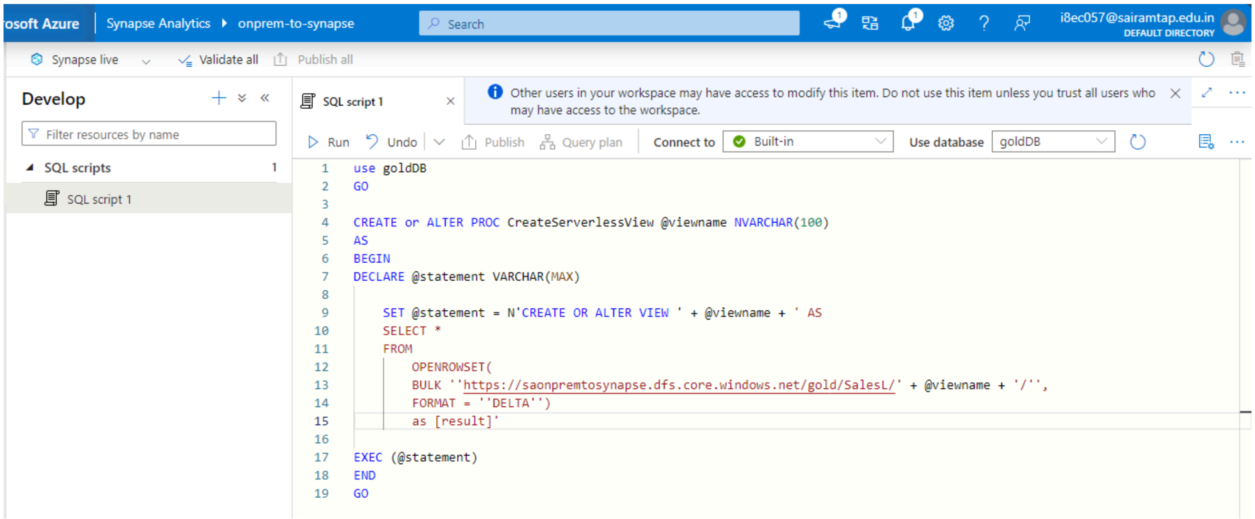
Data Loading

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



We use optional cookies to provide a better experience. Learn more

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



Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

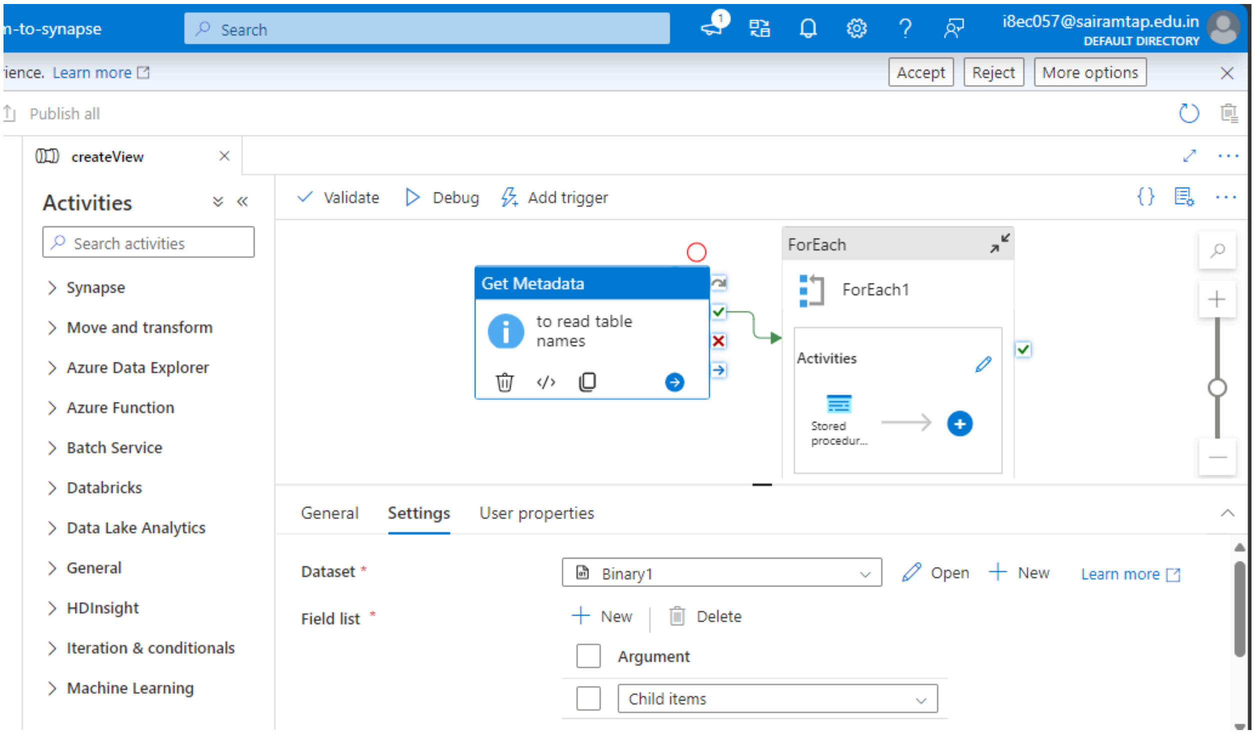
Connect to: Built-in

Use database: goldDB

```sql
use goldDB
GO

CREATE or ALTER PROC CreateServerlessView @viewname NVARCHAR(100)
AS
BEGIN
DECLARE @statement VARCHAR(MAX)

    SET @statement = N'CREATE OR ALTER VIEW ' + @viewname + ' AS
    SELECT *
    FROM
        OPENROWSET(
        BULK ''https://saonpremtosynapse.dfs.core.windows.net/gold/SalesL/' + @viewname + '/'',
        FORMAT = ''DELTA'')
        as [result]'

EXEC (@statement)
END
GO
```

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.



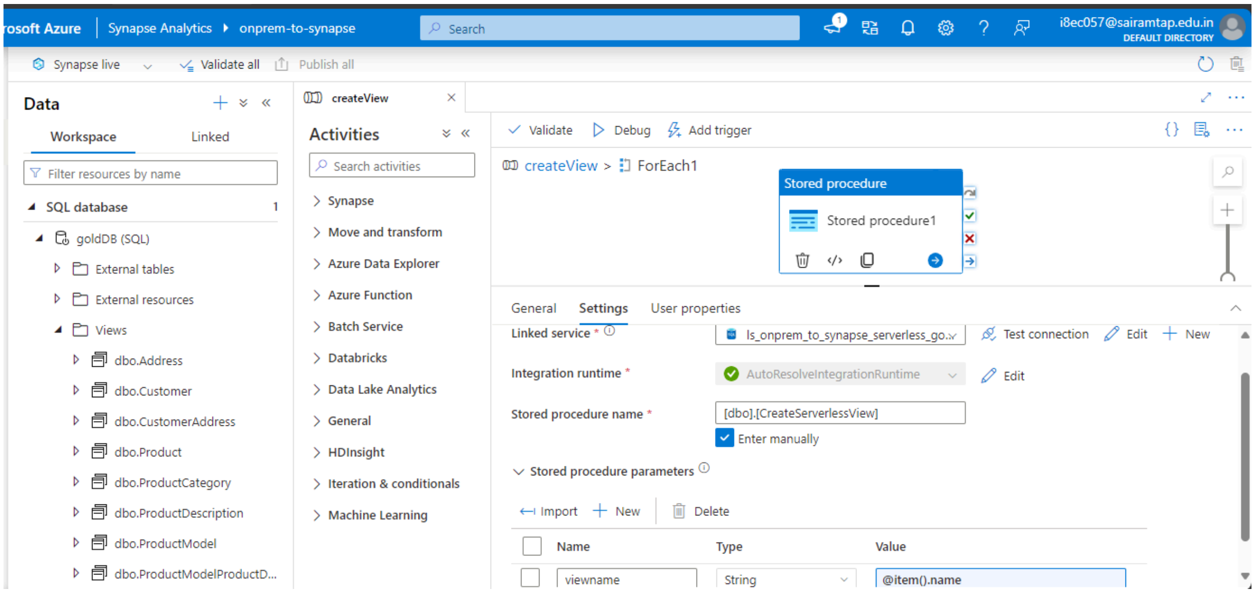
Dataset * Binary1

Field list *

□	Argument
□	Child items

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.



Linked service * ls_onprem_to_synapse_serverless_go...

Integration runtime * AutoResolveIntegrationRuntime

Stored procedure name * [dbo].[CreateServerlessView]

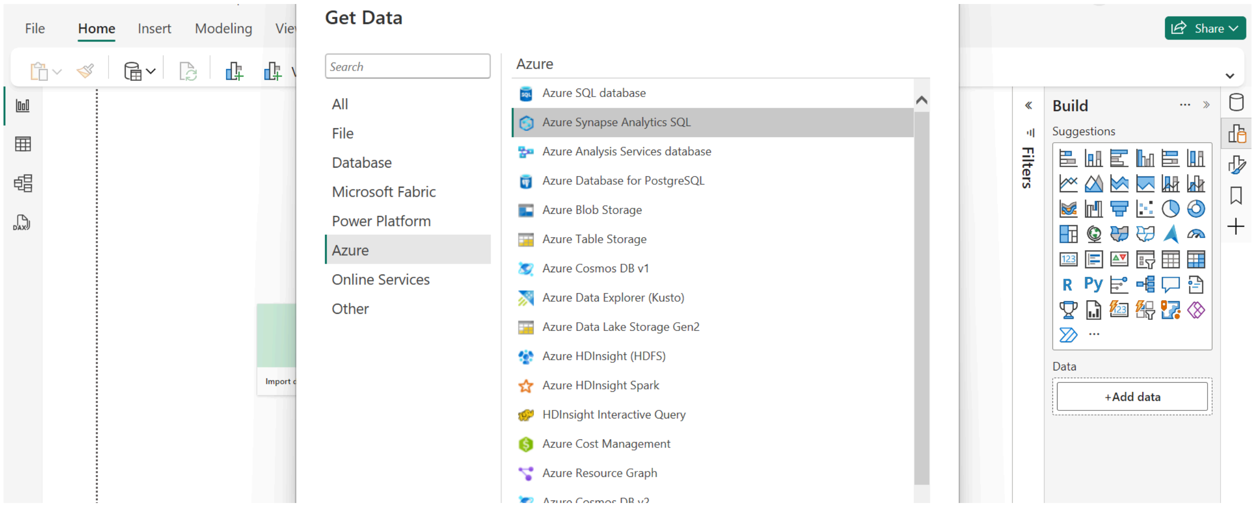
☑ Enter manually

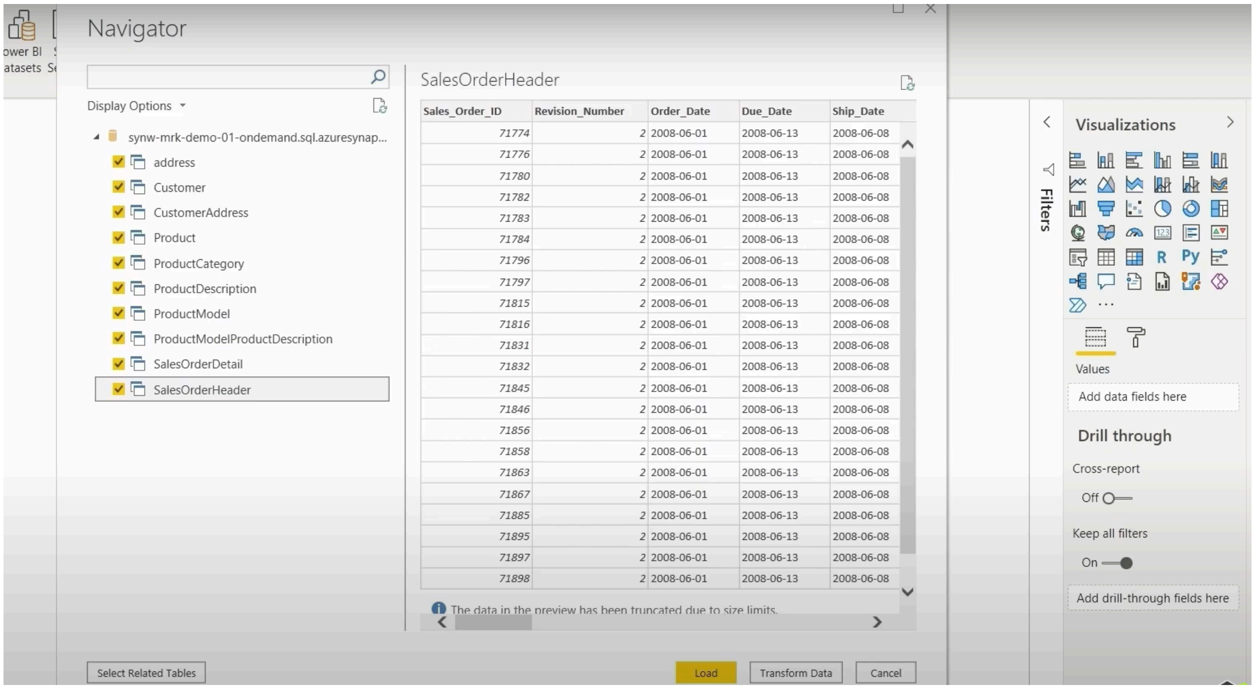
Stored procedure parameters

□	Name	Type	Value
□	viewname	String	@item().name

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:





synw-mrk-demo-01-ondemand.sql.azuresynap...

- ☑ address
- ☑ Customer
- ☑ CustomerAddress
- ☑ Product
- ☑ ProductCategory
- ☑ ProductDescription
- ☑ ProductModel
- ☑ ProductModelProductDescription
- ☑ SalesOrderDetail
- ☑ SalesOrderHeader

SalesOrderHeader

Sales_Order_ID	Revision_Number	Order_Date	Due_Date	Ship_Date
71774	2	2008-06-01	2008-06-13	2008-06-08
71776	2	2008-06-01	2008-06-13	2008-06-08
71780	2	2008-06-01	2008-06-13	2008-06-08
71782	2	2008-06-01	2008-06-13	2008-06-08
71783	2	2008-06-01	2008-06-13	2008-06-08
71784	2	2008-06-01	2008-06-13	2008-06-08
71796	2	2008-06-01	2008-06-13	2008-06-08
71797	2	2008-06-01	2008-06-13	2008-06-08
71815	2	2008-06-01	2008-06-13	2008-06-08
71816	2	2008-06-01	2008-06-13	2008-06-08
71831	2	2008-06-01	2008-06-13	2008-06-08
71832	2	2008-06-01	2008-06-13	2008-06-08
71845	2	2008-06-01	2008-06-13	2008-06-08
71846	2	2008-06-01	2008-06-13	2008-06-08
71856	2	2008-06-01	2008-06-13	2008-06-08
71858	2	2008-06-01	2008-06-13	2008-06-08
71863	2	2008-06-01	2008-06-13	2008-06-08
71867	2	2008-06-01	2008-06-13	2008-06-08
71885	2	2008-06-01	2008-06-13	2008-06-08
71895	2	2008-06-01	2008-06-13	2008-06-08
71897	2	2008-06-01	2008-06-13	2008-06-08
71898	2	2008-06-01	2008-06-13	2008-06-08

The data in the preview has been truncated due to size limits.

Cross-report

Off

Keep all filters

On