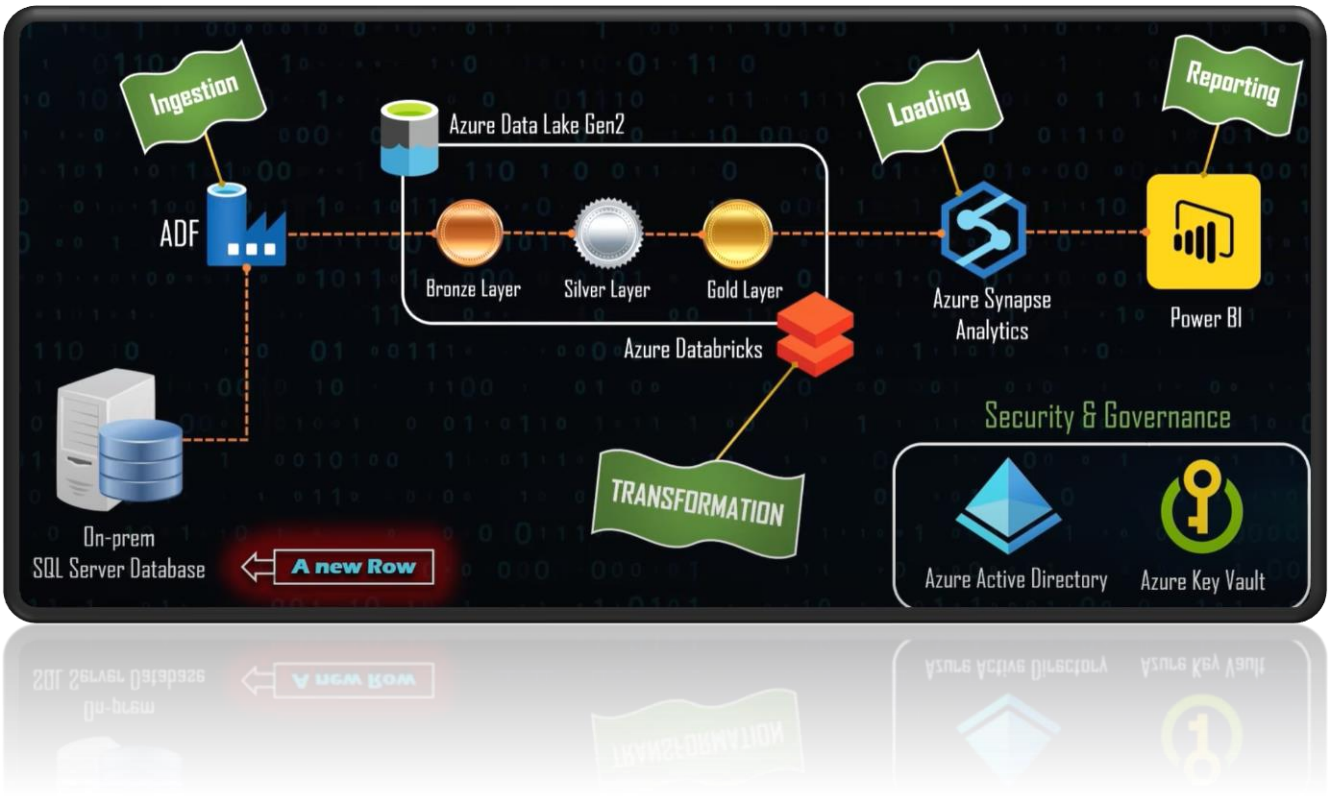
END To END ETL Implementation Project



Siva Rama Krishna Ponnapalli

Contents

END To END ETL Implementation Project 1

1.Set Up Environment 3

2.Data Ingestion using ADF 6

3.Data Transformation With Databricks 15

4.Creating Views Using Synapse 17

5. Publishing Data Analytics Reports – Power BI 22

6. Security and Governance 24

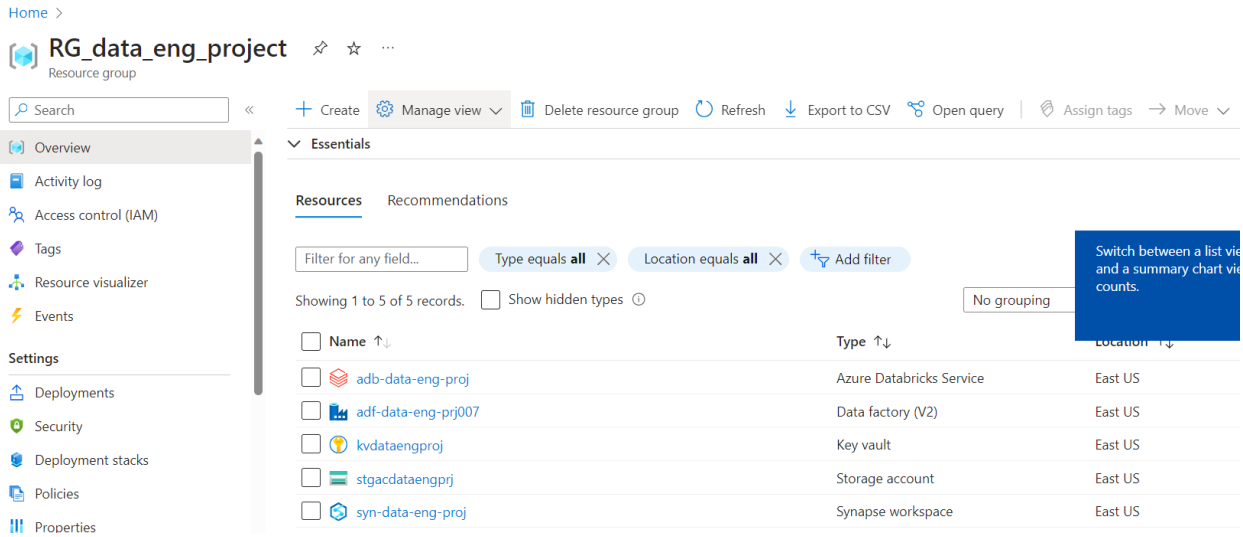
7. End to End Pipeline Testing 25

8. References 27

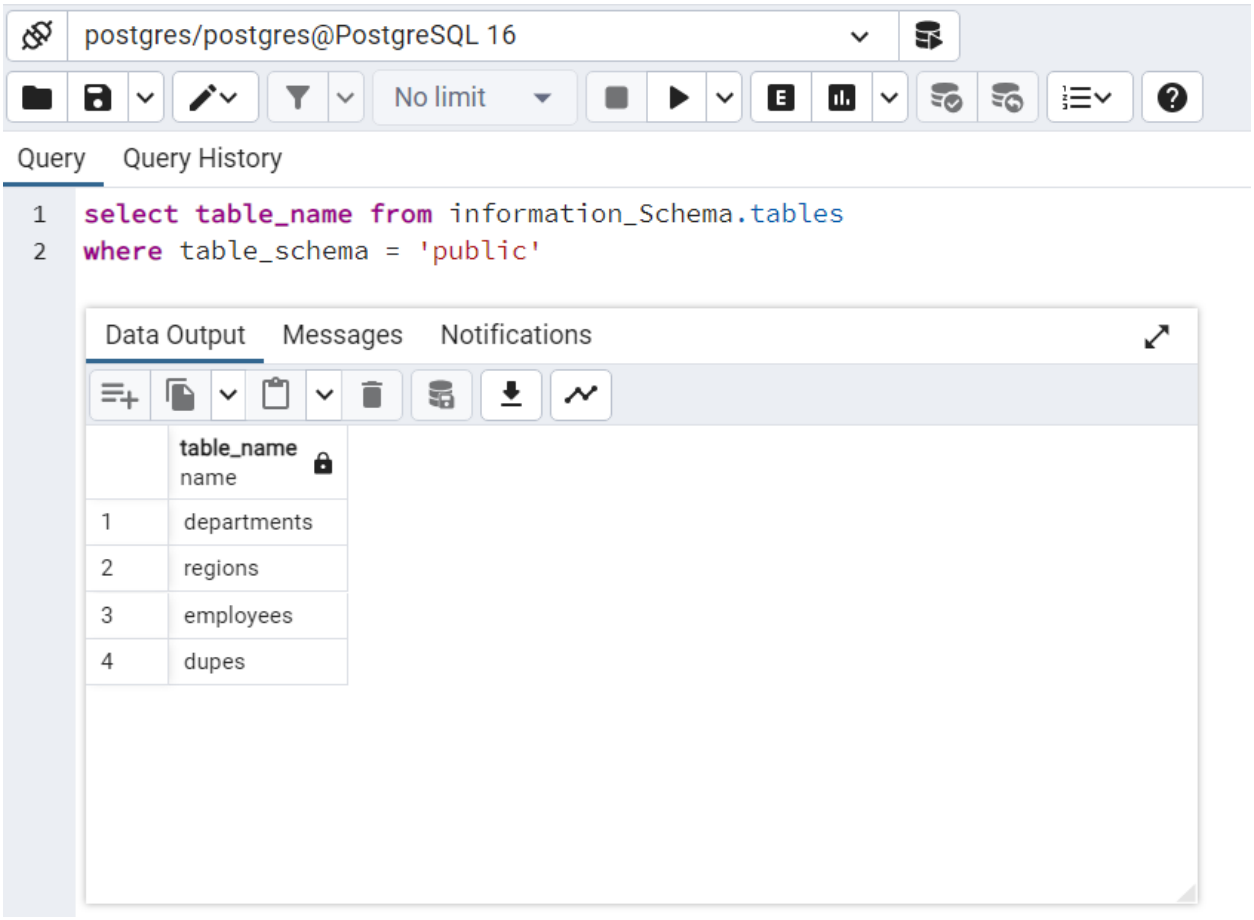
Scenario :
Performing a migration from On-Premise PostgreSQL server to Azure cloud and doing a complete End to end implementation which includes Migration, Extraction, Transformation ,Loading and lastly Analytics .

1.Set Up Environment

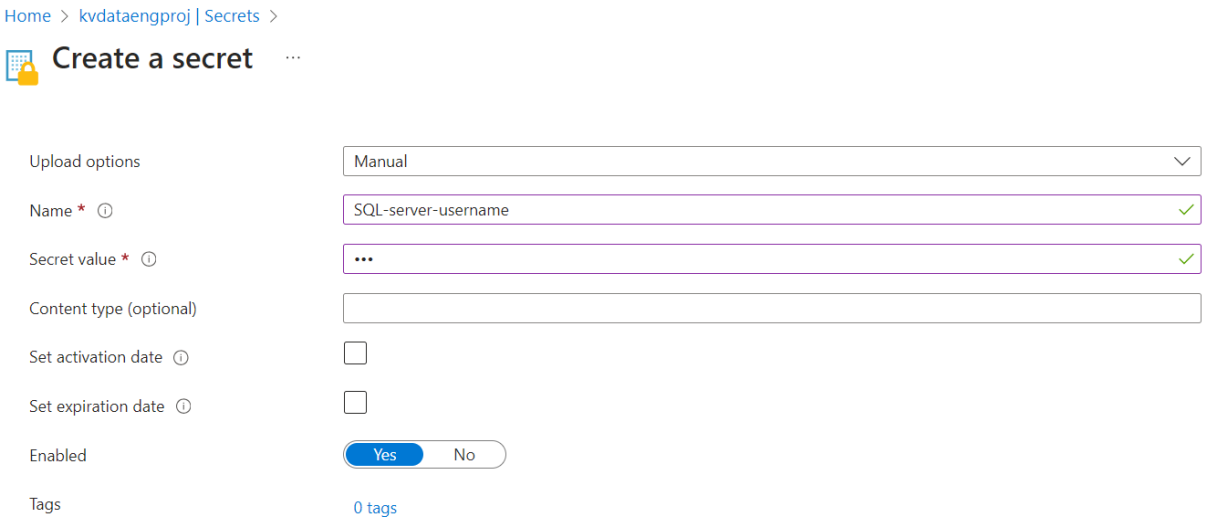
1.1 For the set up environment I created a new resource group for the project and have the following resources up and running.
These Include : **Azure Databricks, Azure Data Factory, Azure Storage Account, Azure Synapse Analytics Workspace and Azure Key Vault**

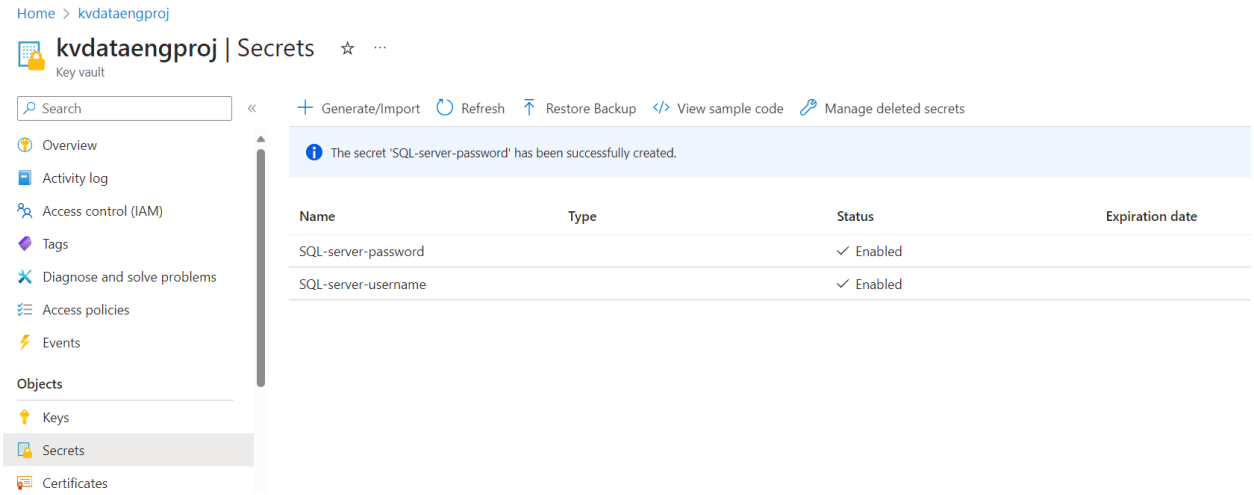


1.2 In my On-Premise Microsoft SQL Server/PostgreSQL, Tables which are being migrated :

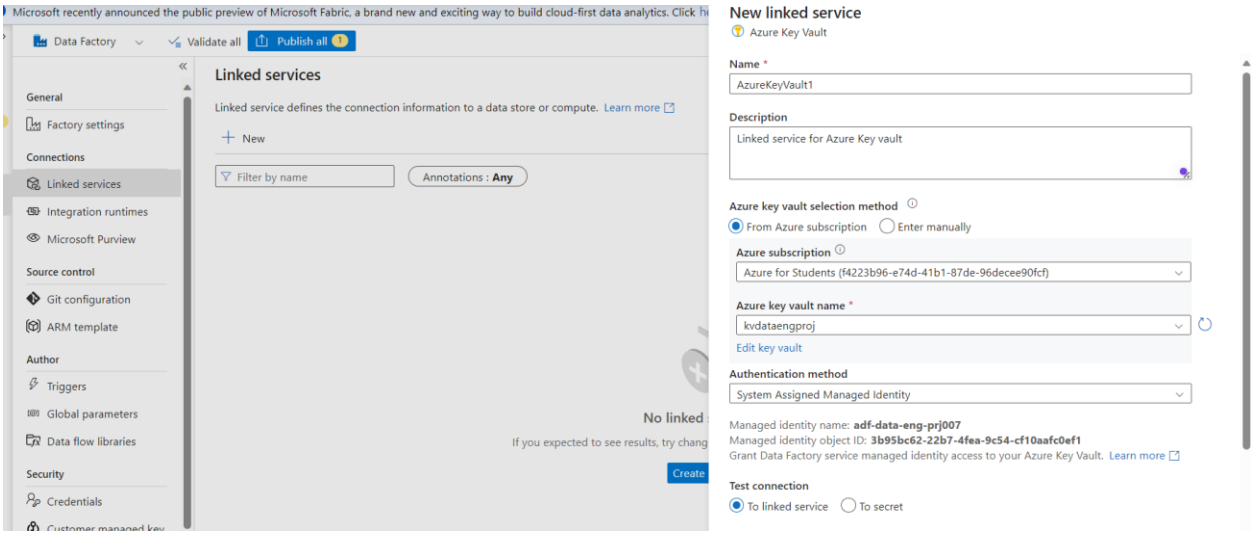


1.3 Created Azure Key Vault secrets Which comprise secrets created for the SQL server user Login Id and Password

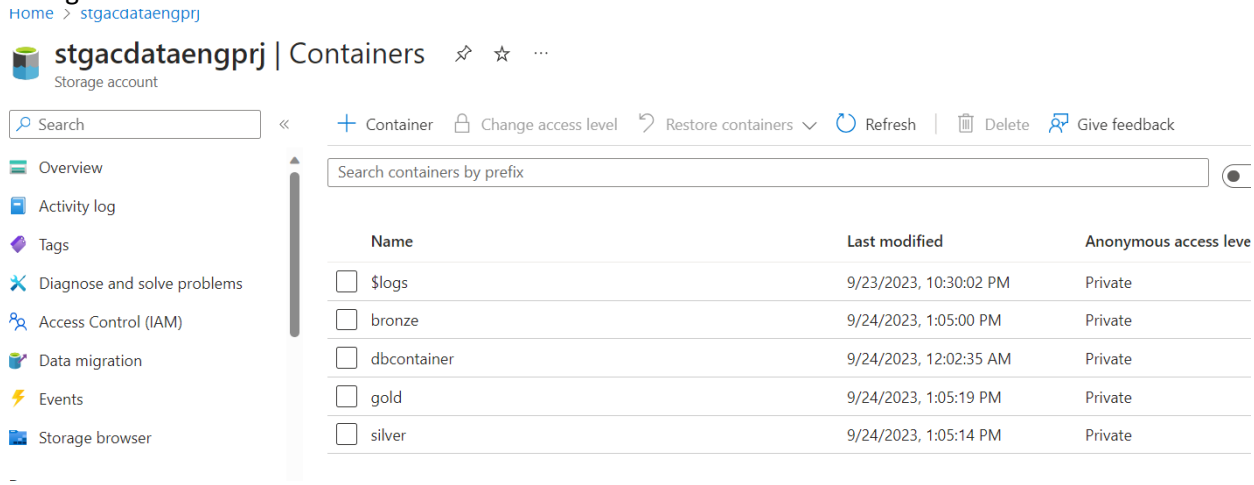




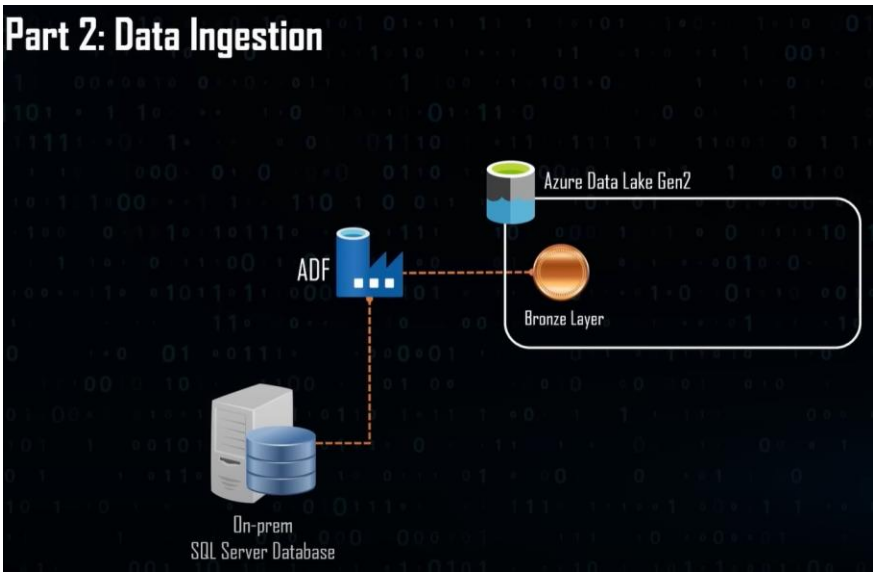
Also, create a linked service for the key vault, as this needs to be linked whenever key vault is used.



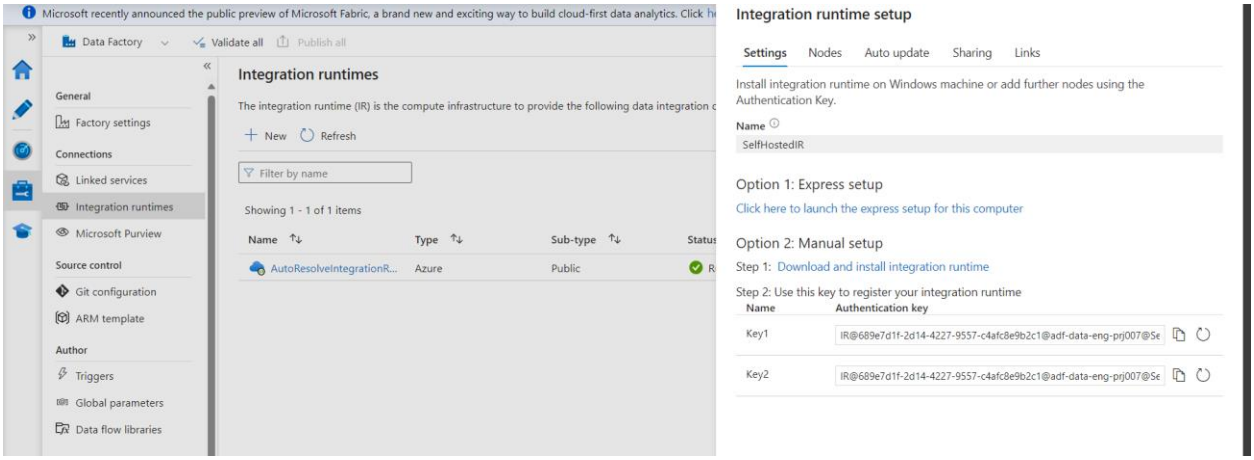
1.4 Storage account and containers



2.Data Ingestion using ADF

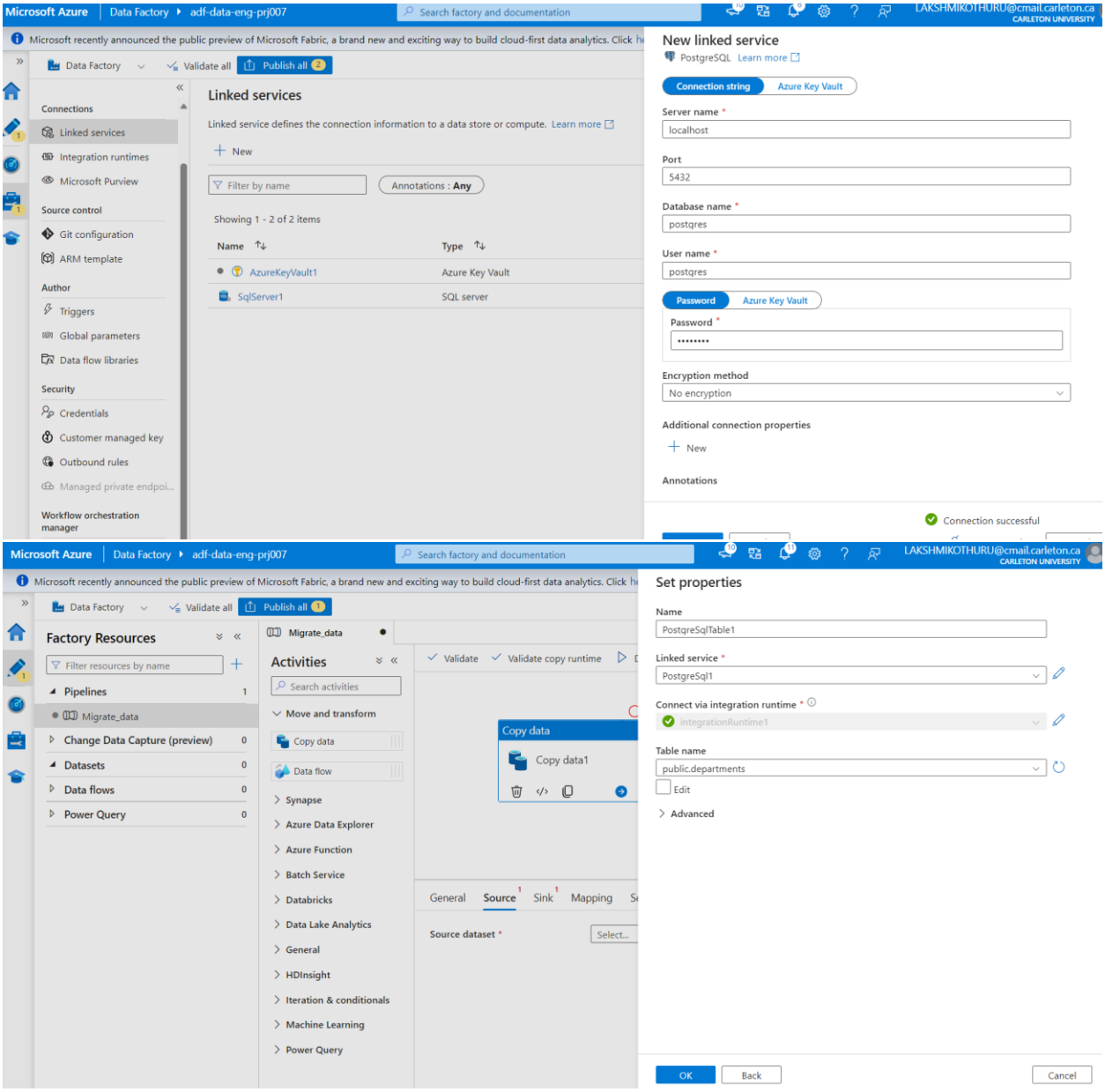


2.1 Creating a Self-Hosted Integration Runtime in Azure Data Factory as we are connecting to the on-premises server.

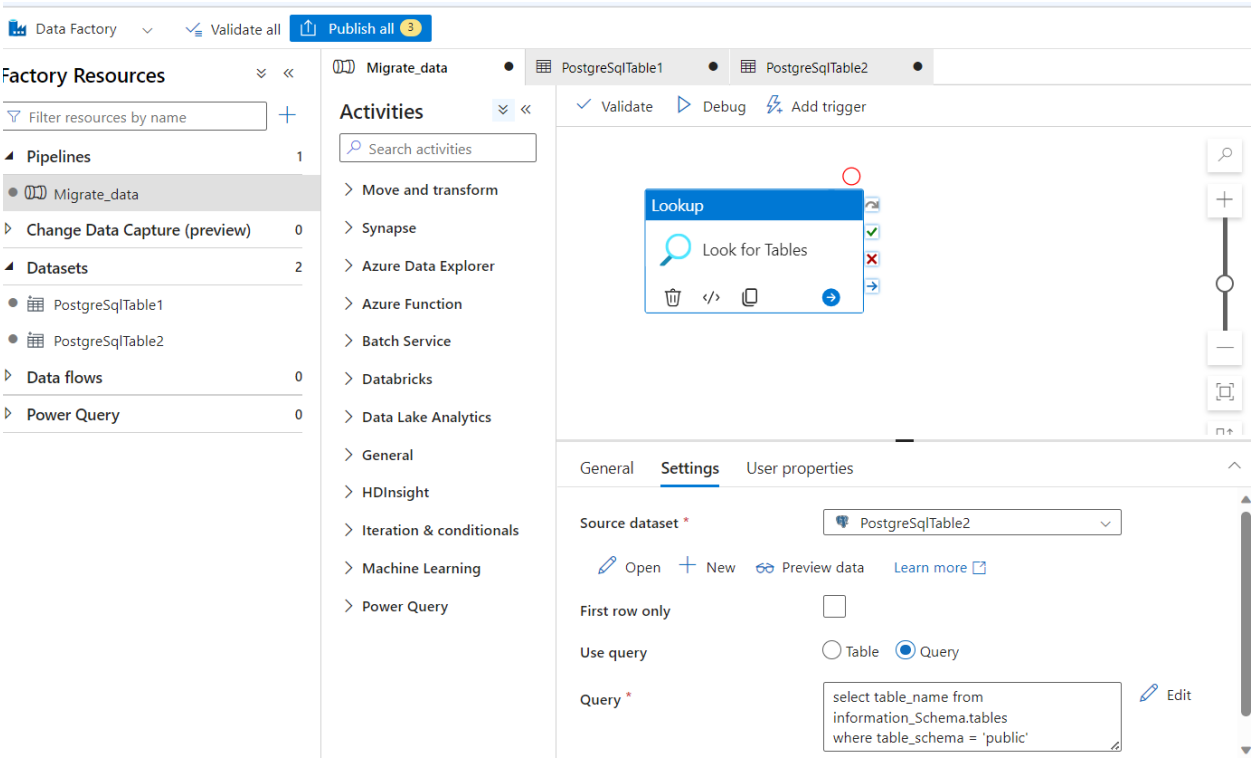


2.2 Create a dataset and a linked for the source of Copy data pipeline

Linked Service for Postgresql

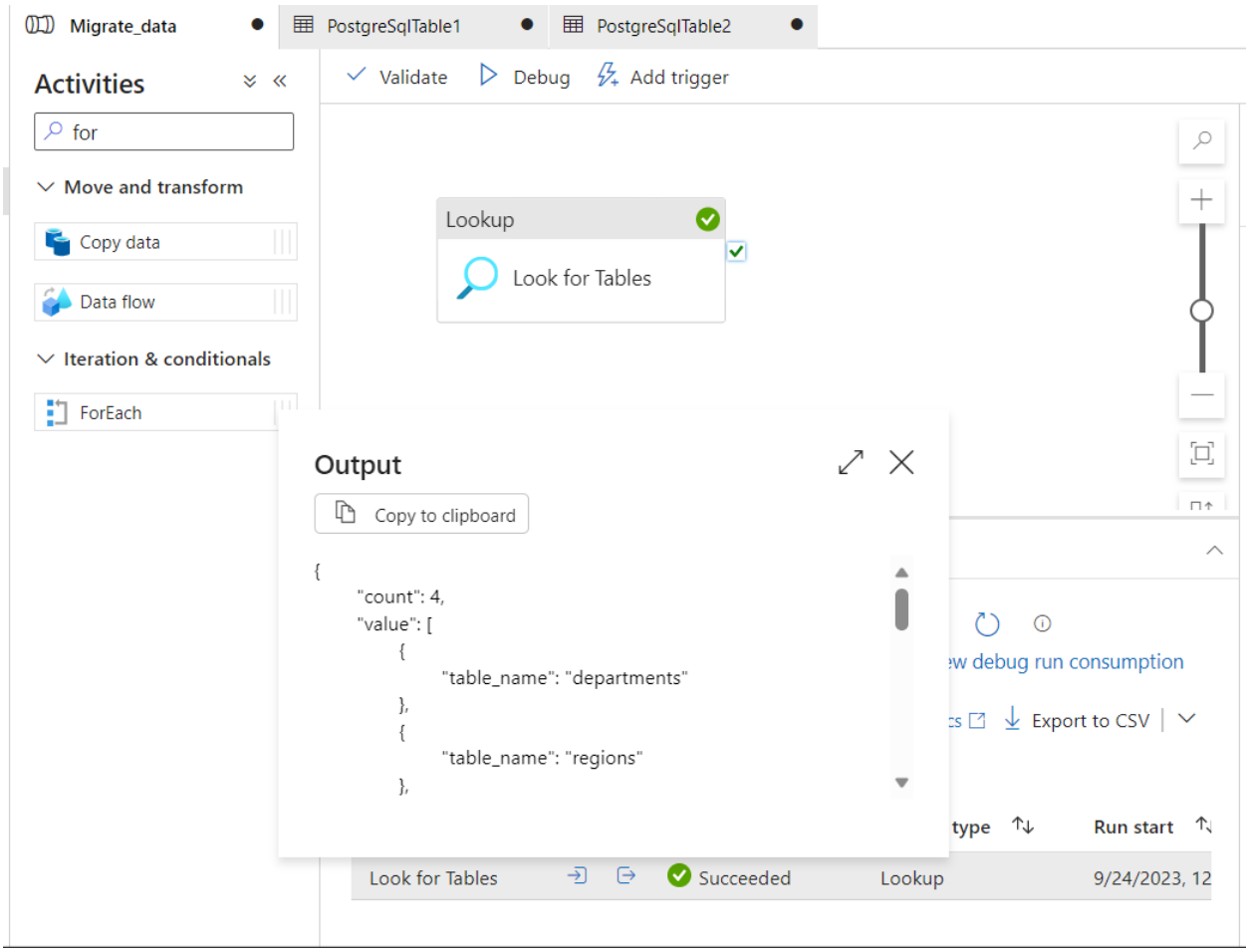


2.3 Creating a pipeline in ADF with Lookup activity .



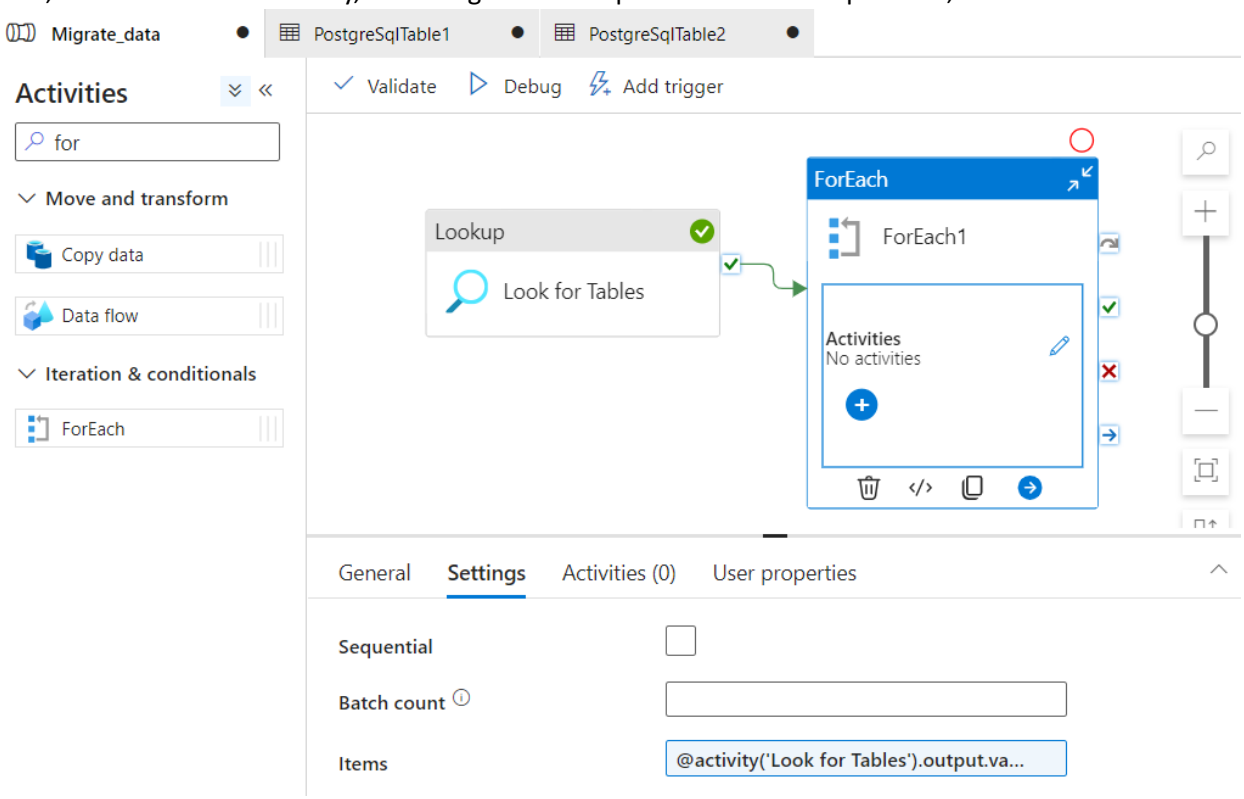
Query : select table_name from information_Schema.tables
where table_schema = 'public'

Validated the activity and debugged the activity to get the output as follows

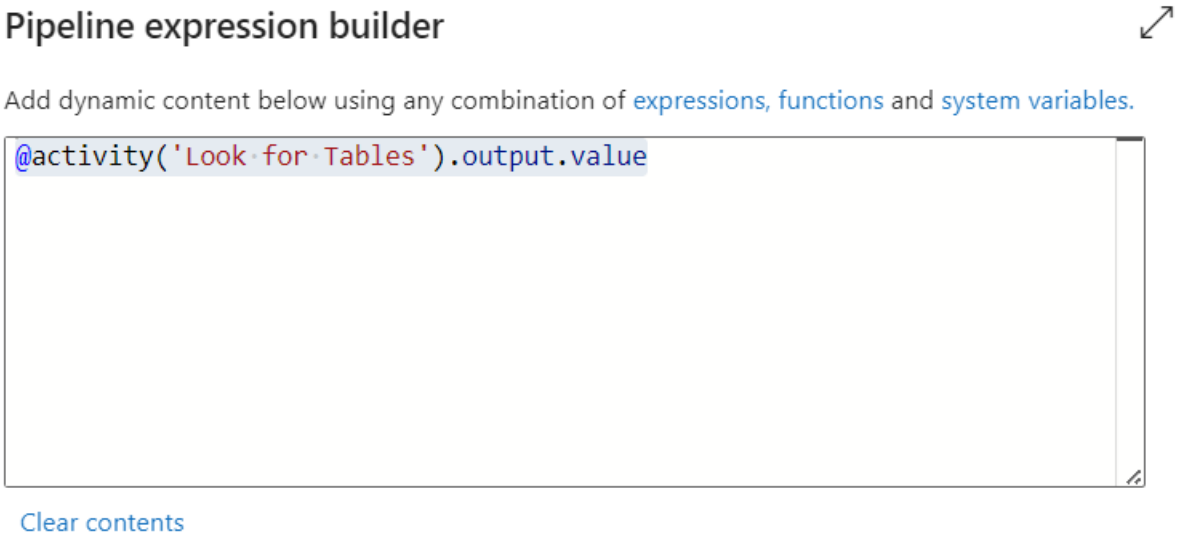


The JSON format output includes the list of tables in my PostgreSQL database.

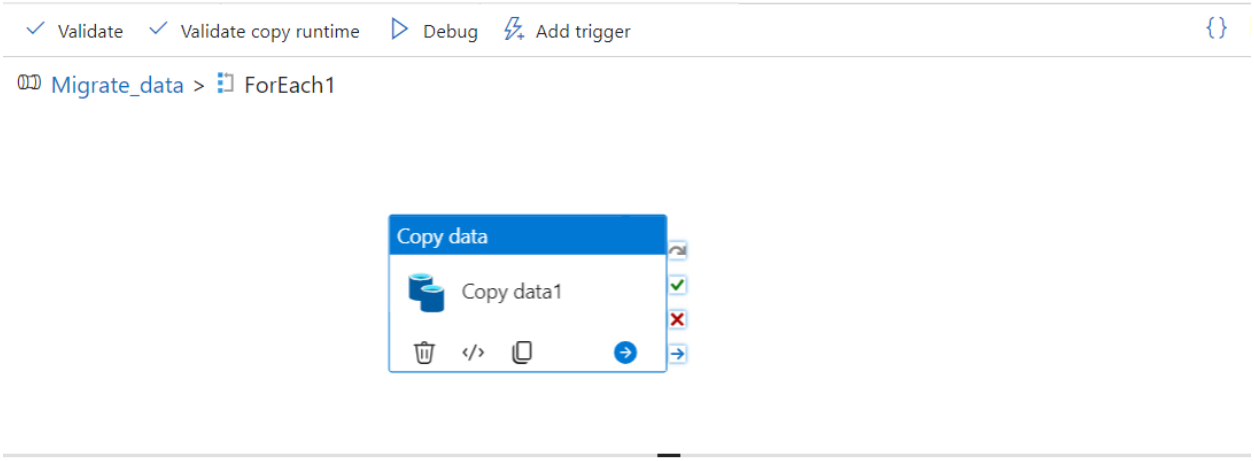
2.4 Now, with the for each activity, we configure the output from the look up tables ;

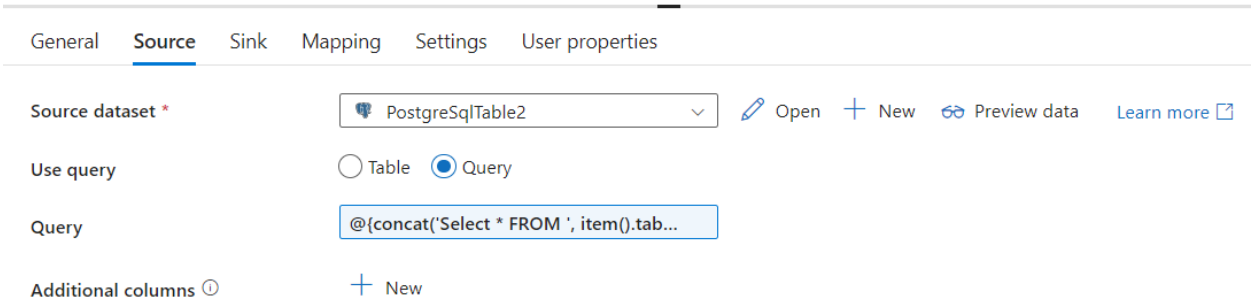


In the items, we add dynamic content and take the output from lookup activity with expression

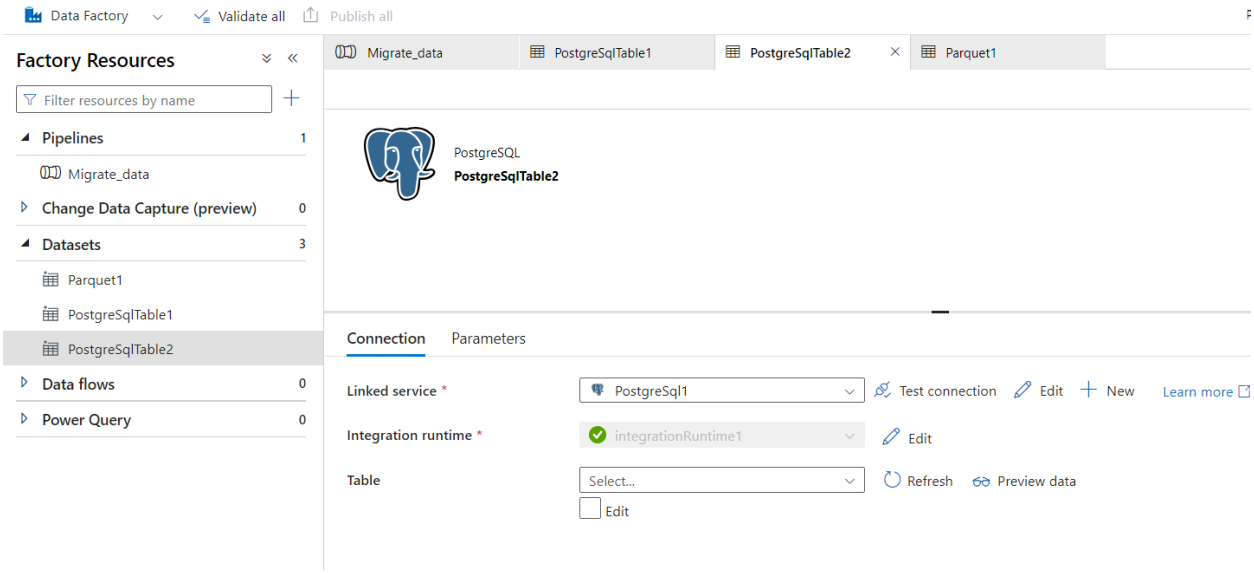


2.5 Configure a **copy activity** as a sub activity in the for each activity





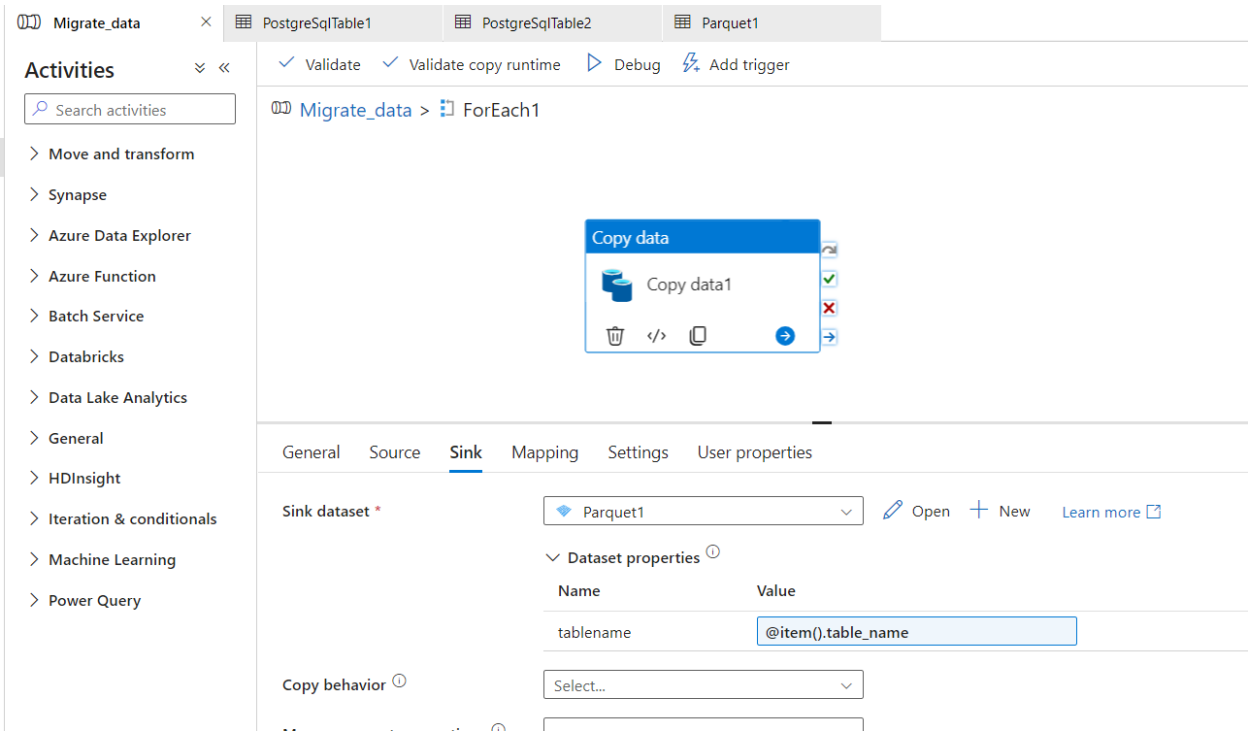
Created a new dataset with the existing self hosted Integration runtime



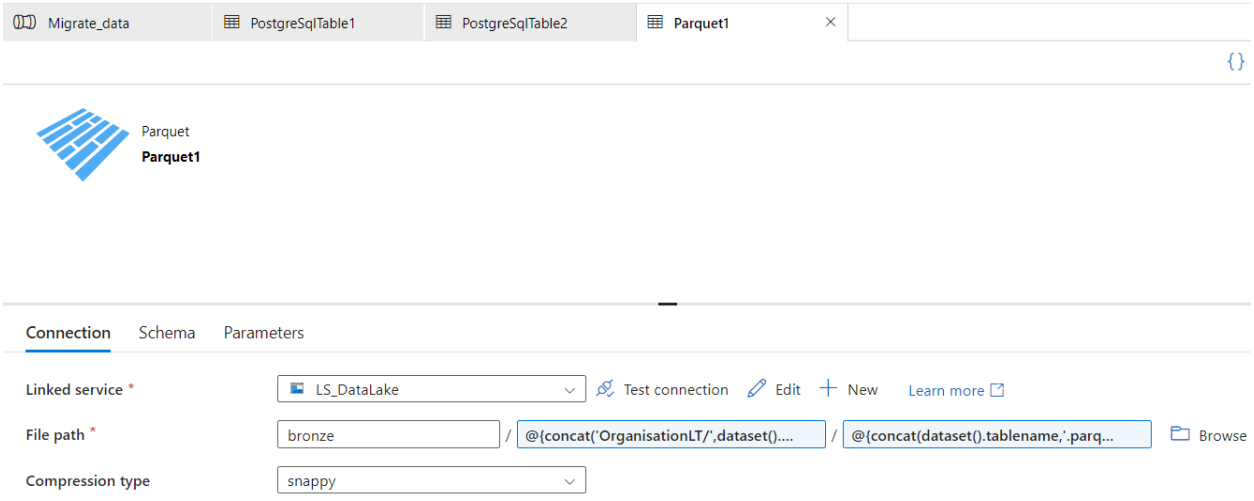
A simple Concat query with the dynamic content from the previous activity

```
@{concat('Select * FROM ', item().table_name)}
```

Sink Dataset with parquet format , with table name being parameterised with dynamic content , as we have multiple tables being migrated.



Configured the dataset to store the dataset in parquet and store it in a hierarchical fashion.

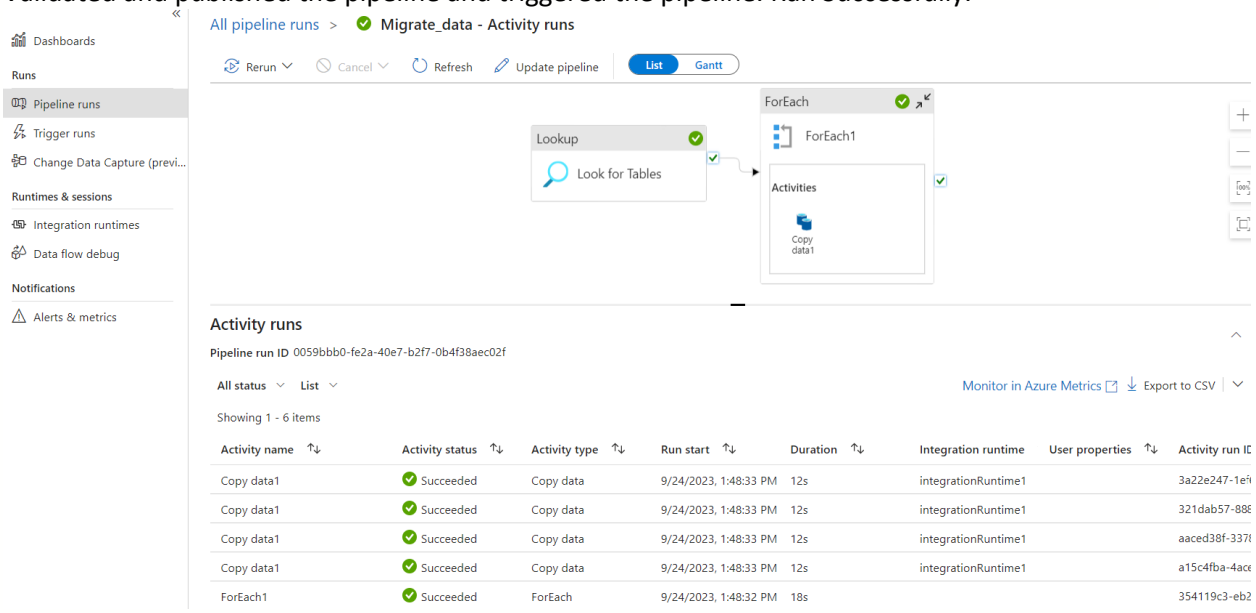


Container_name/OrganisationLT/Table_name/Tablename.parquet

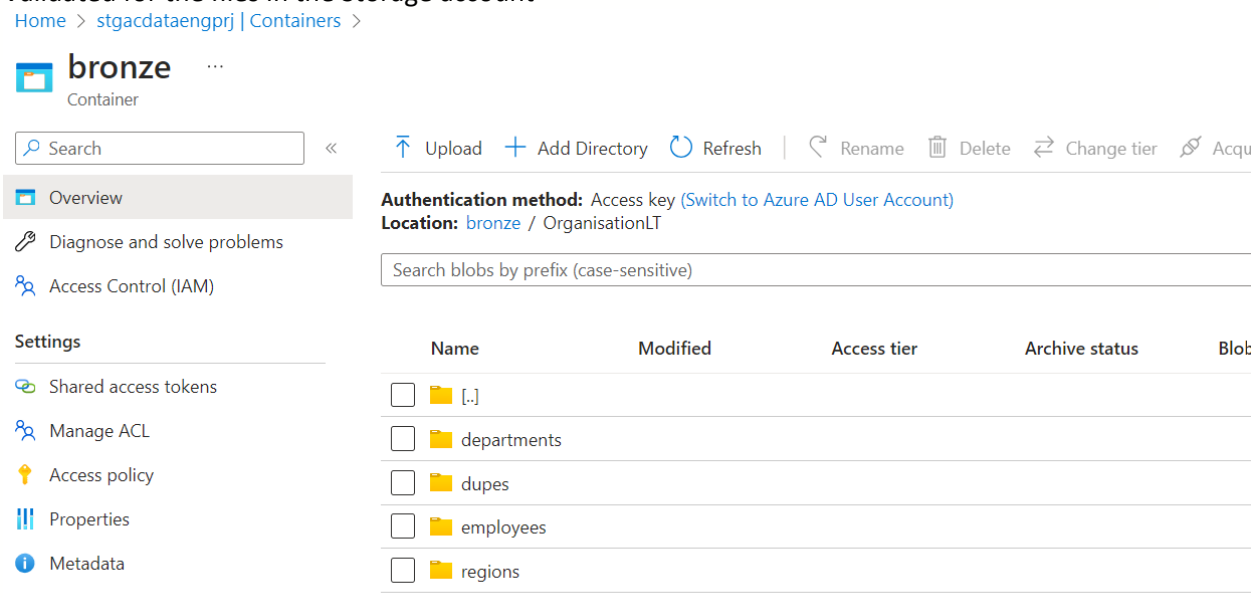
For Direcotry : `@{concat('OrganisationLT/',dataset().tablename)}`

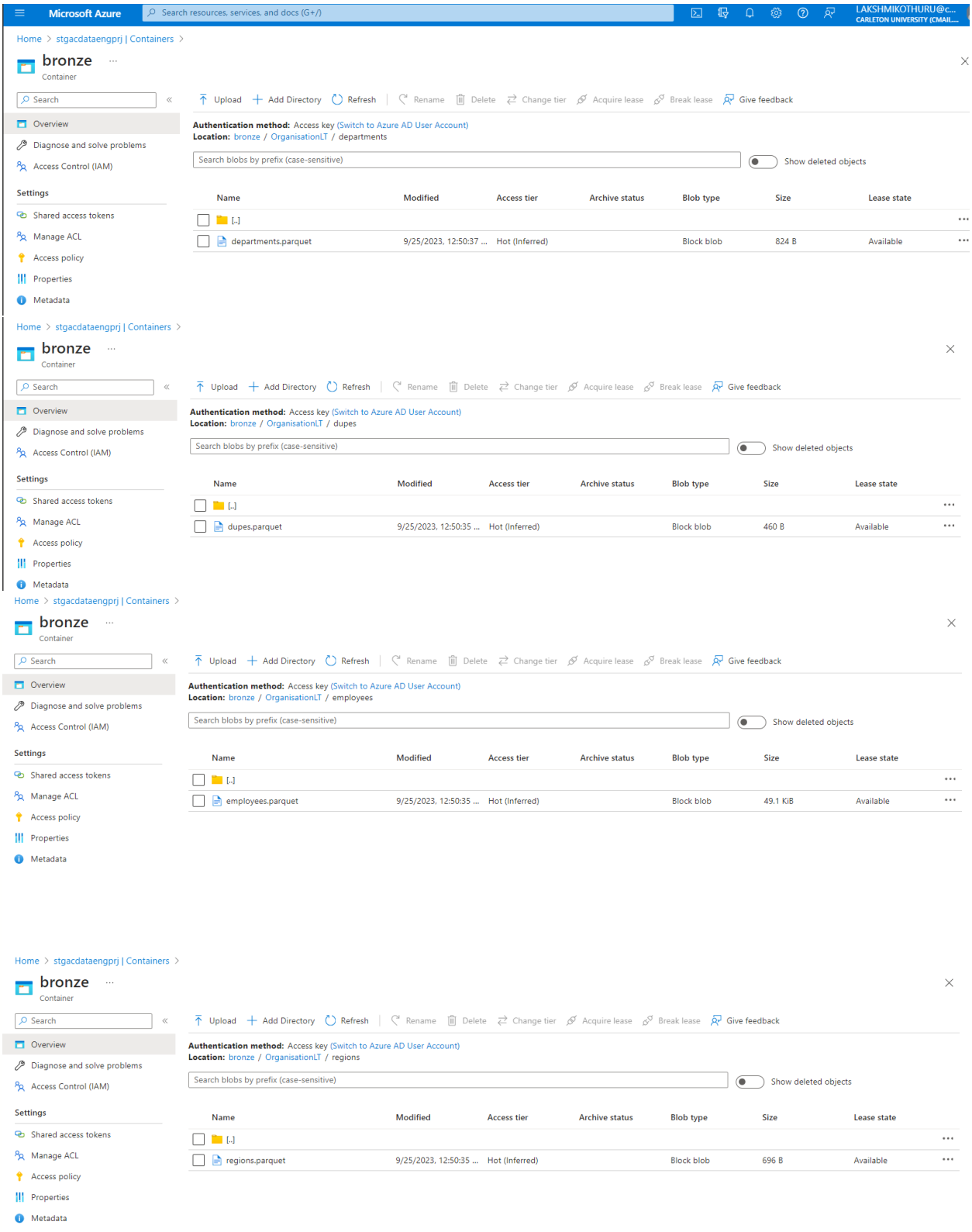
`For File name : @{concat(dataset().tablename,'.parquet')}`

2.6 Validated and published the pipeline and triggered the pipeline. Ran Successfully.



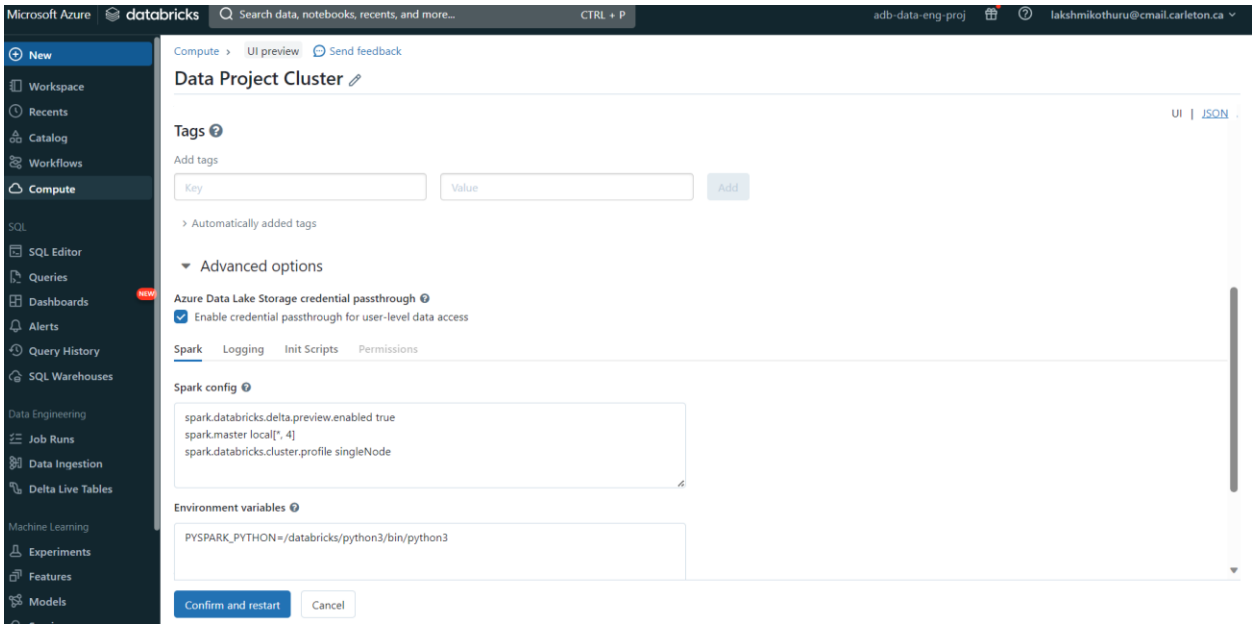
2.7 Validated for the files in the Storage account




Microsoft Azure
Search resources, services, and docs (G+/)
LAKSHMIKOTHURU@c...
CARLETON UNIVERSITY (CMAIL...
Home > stgacdataengprj | Containers >
bronze
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Azure AD User Account)
Location: bronze / OrganisationLT / departments
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
[..]
departments.parquet
9/25/2023, 12:50:37 ...
Hot (Inferred)
Block blob
824 B
Available
Home > stgacdataengprj | Containers >
bronze
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Azure AD User Account)
Location: bronze / OrganisationLT / dupes
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
[..]
dupes.parquet
9/25/2023, 12:50:35 ...
Hot (Inferred)
Block blob
460 B
Available
Home > stgacdataengprj | Containers >
bronze
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Azure AD User Account)
Location: bronze / OrganisationLT / employees
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
[..]
employees.parquet
9/25/2023, 12:50:35 ...
Hot (Inferred)
Block blob
49.1 KiB
Available
Home > stgacdataengprj | Containers >
bronze
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Azure AD User Account)
Location: bronze / OrganisationLT / regions
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
[..]
regions.parquet
9/25/2023, 12:50:35 ...
Hot (Inferred)
Block blob
696 B
Available

3.Data Transformation With Databricks

3.1 Create a databricks single-node cluster while enabling credential passthrough for user-level data access

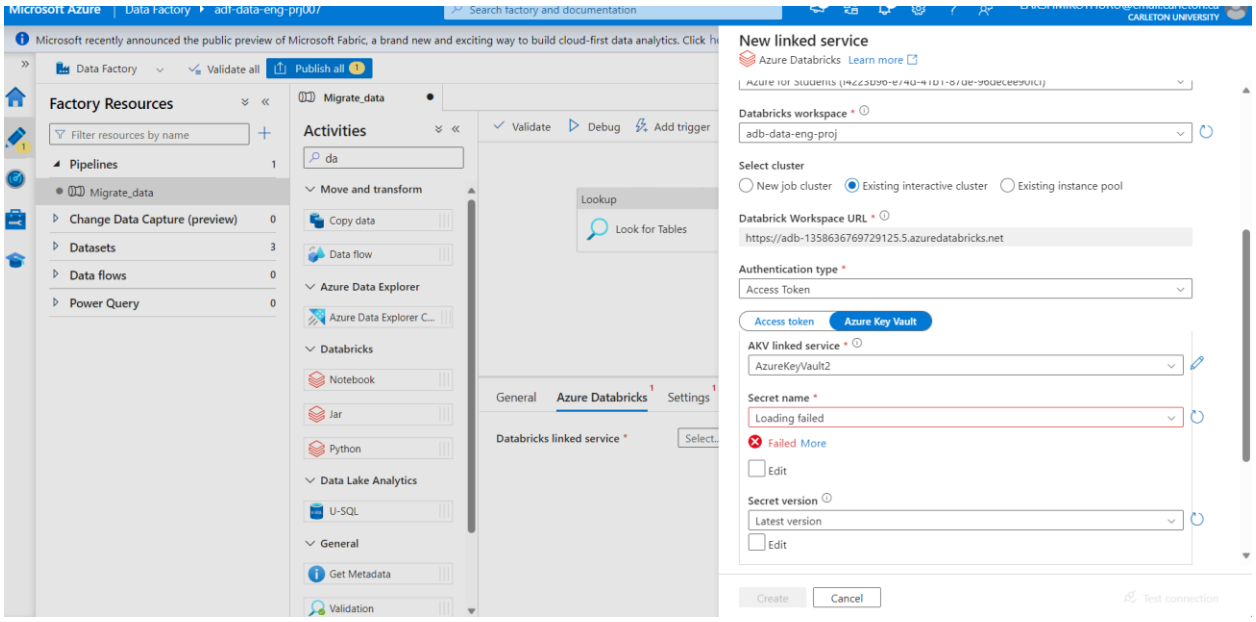


[Access Azure Data Lake Storage using Azure Active Directory credential passthrough (legacy) - Azure Databricks | Microsoft Learn]

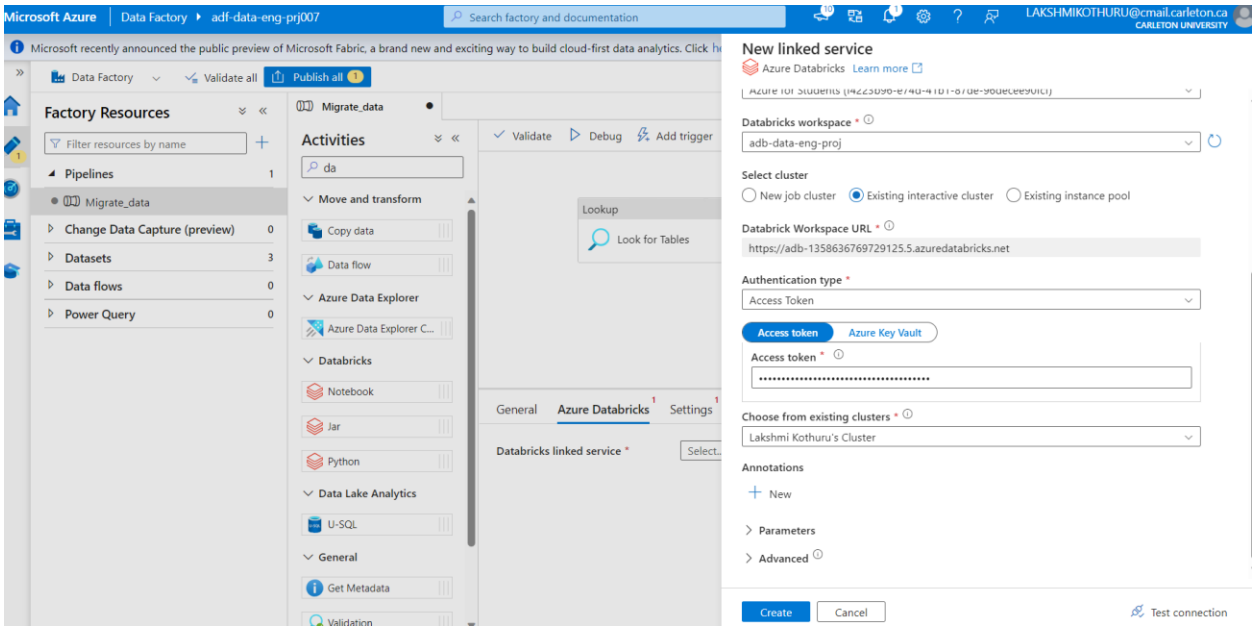
I did mounting through credential passthrough method, which surprisingly worked well, unlike in my other projects where I couldn't make this work.

3.2 Created a databricks notebook activity for bronze to silver notebook, created linked service and tried utilizing Keyvault.

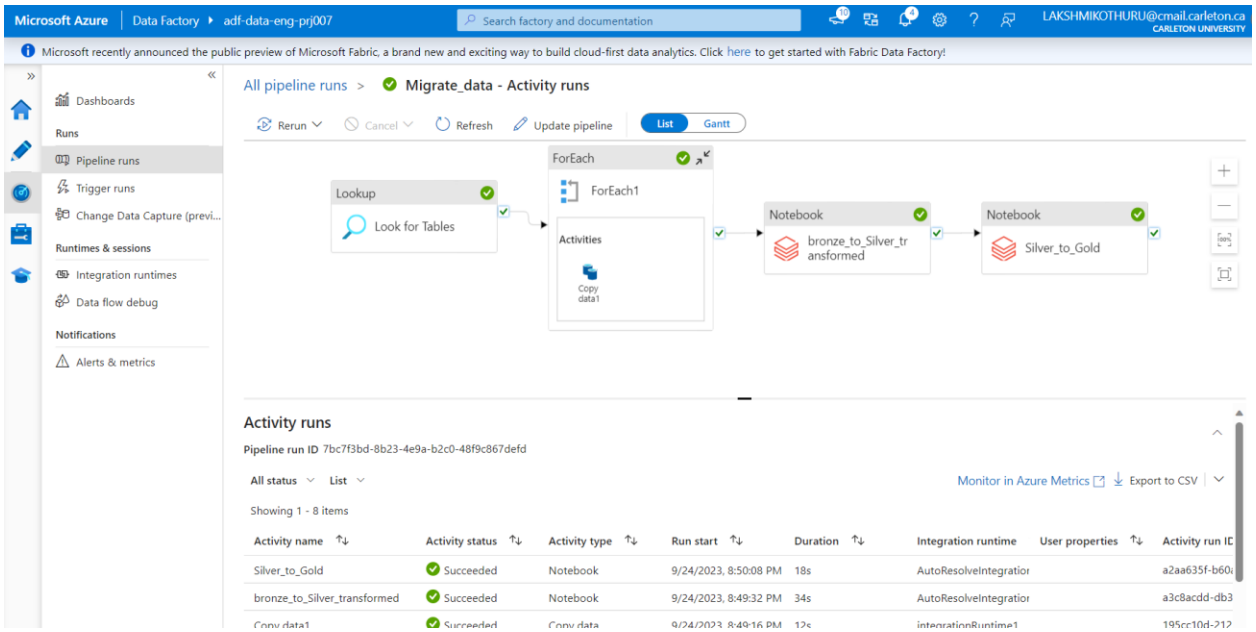
Since My account is not a premium account I'm unable to load the key vault secret name, even after applying role assignments.



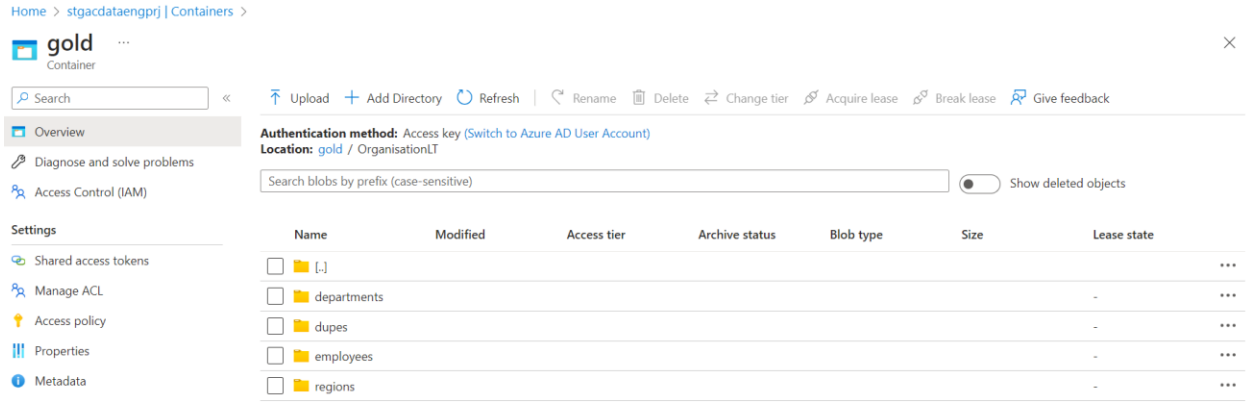
So I directly used the Access token , without Key Vault ,which is not a preferred approach



3.3 Pipeline Run Successfully

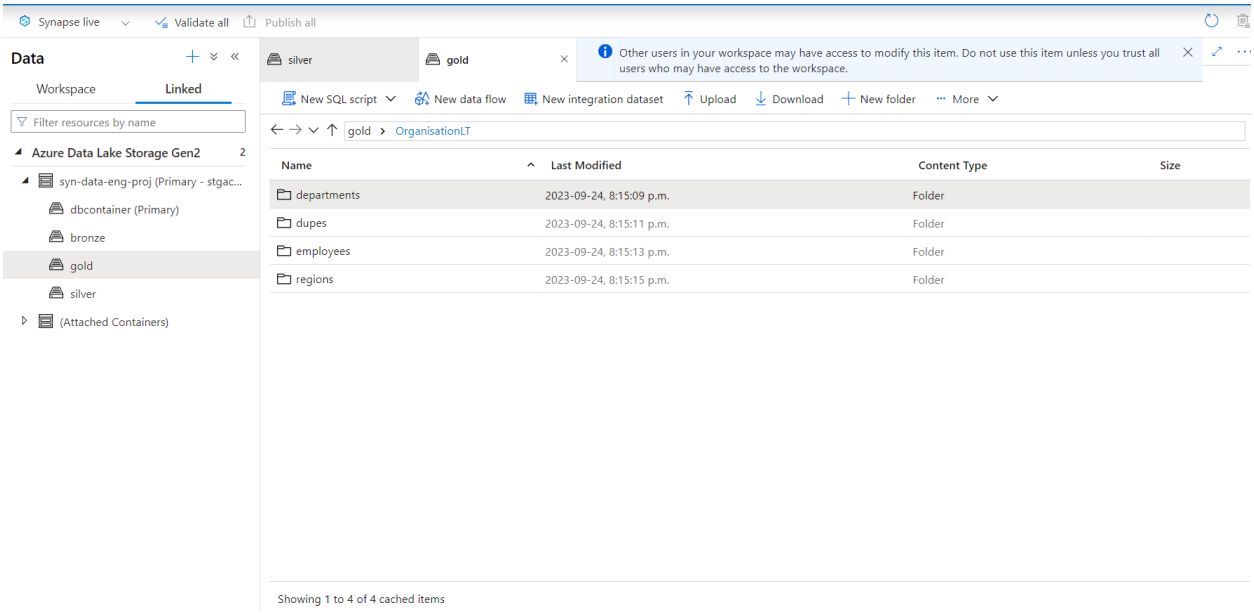


3.4 Verifying for data in the gold container

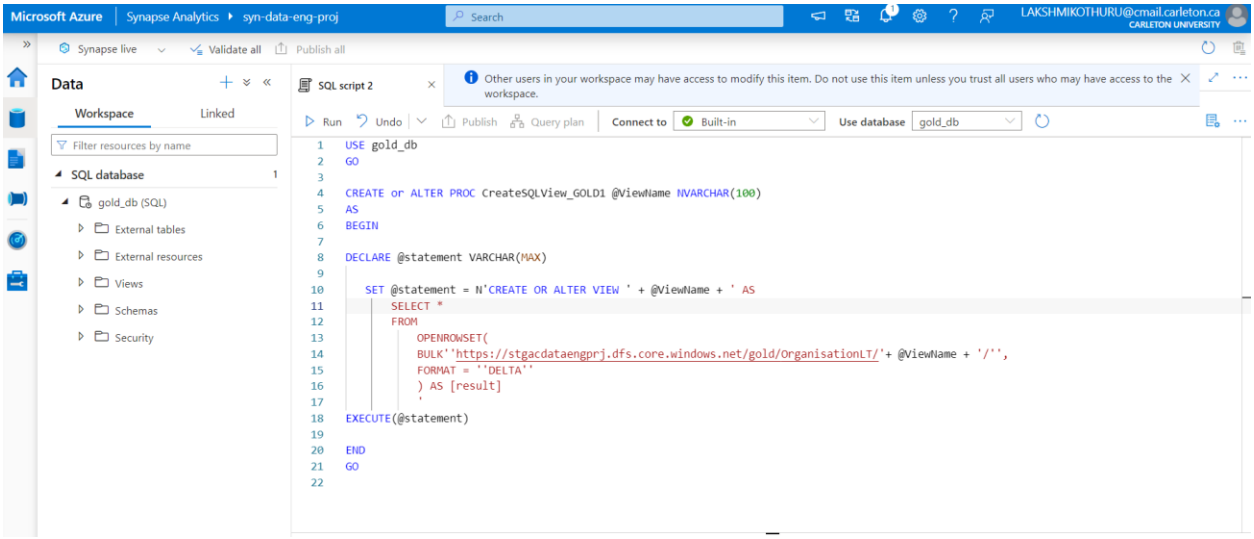


4.Creating Views Using Synapse

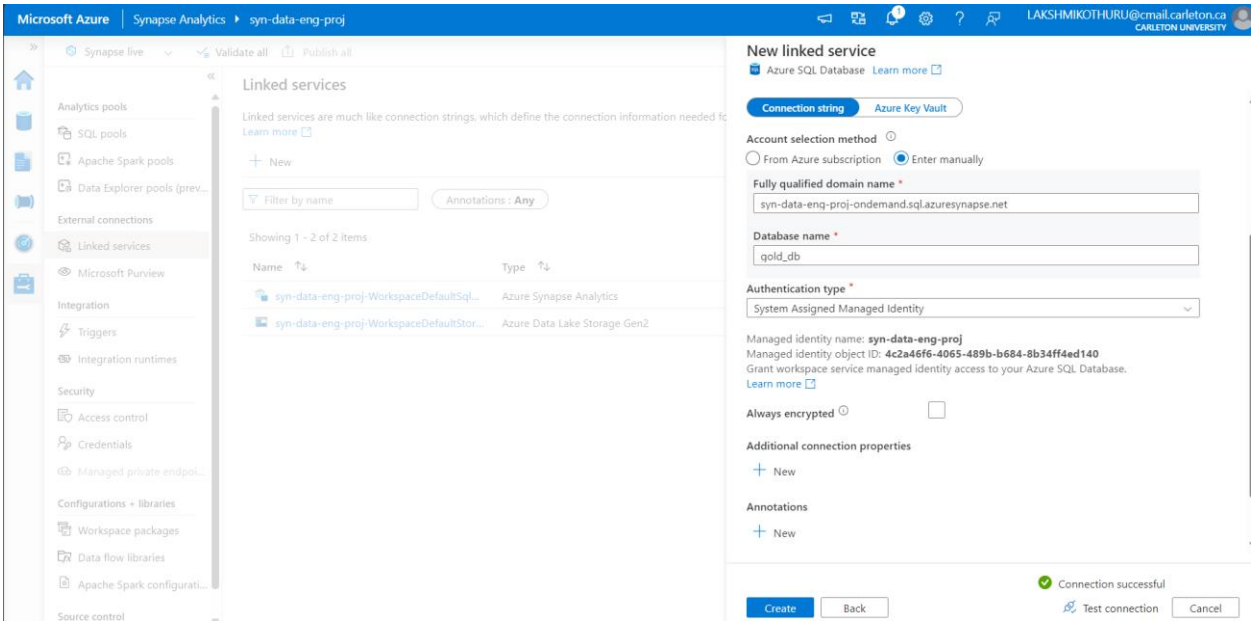
Created a Synapse Workspace,a database **gold_db** and used the linked feature to look at the storage containers which also link the tables



Created a procedure for parameterising and creating the views for all the tables in the gold container



Now I'm creating a pipeline for automating this task , creating a new linked service for the SQL database(serverless) which has the stored procedure.

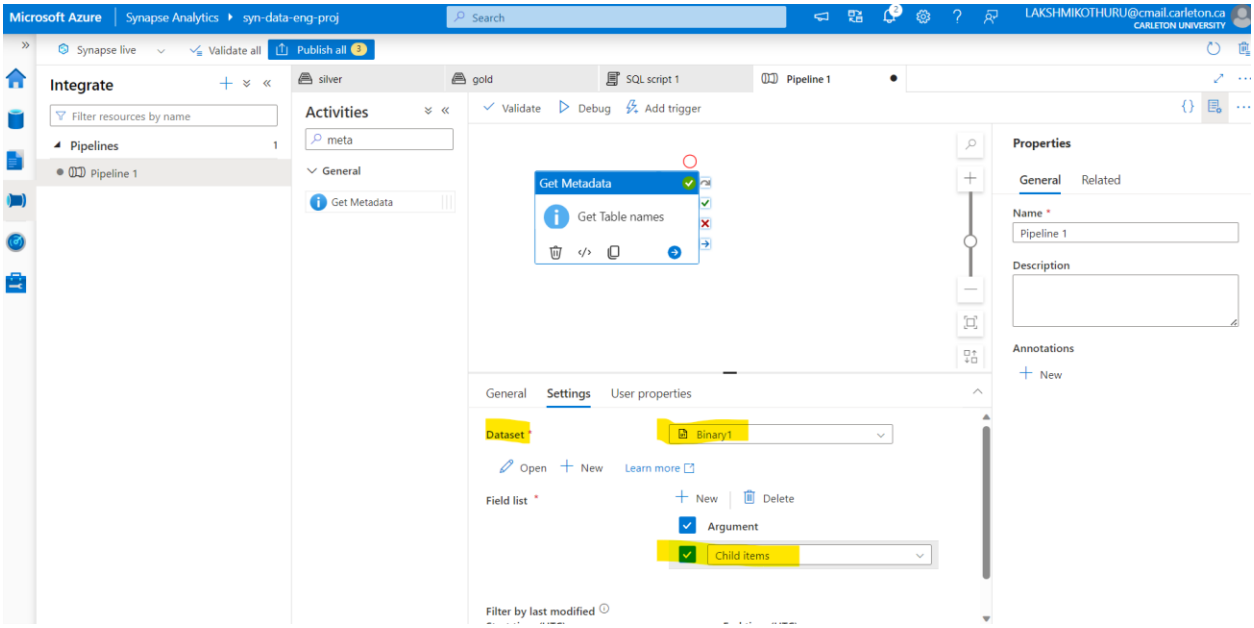


Entered Connection string details manually as we can't populate the database name .

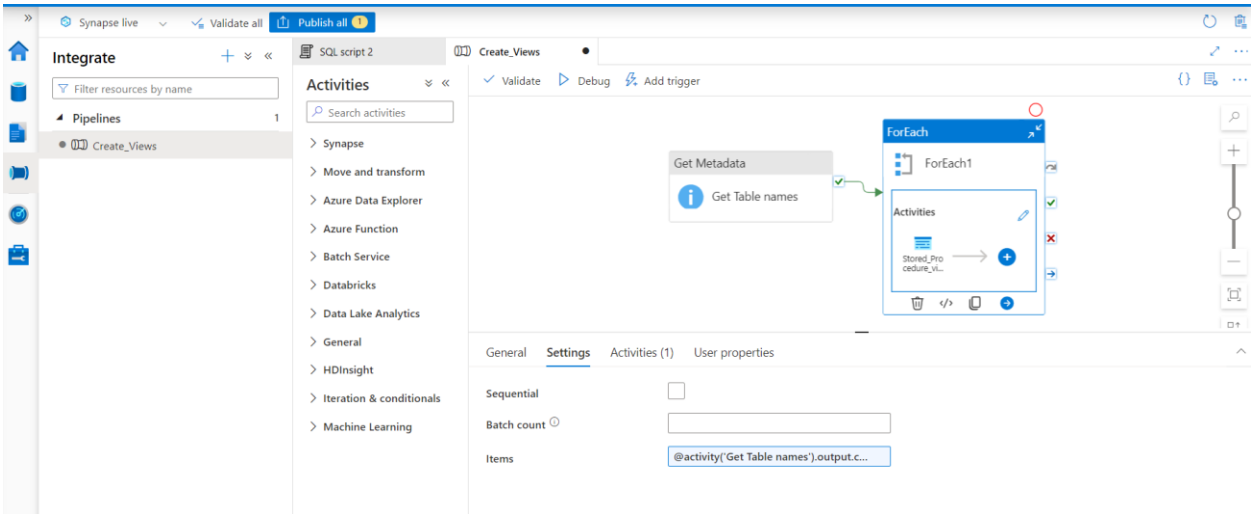
Used the domain name which can be found under **syn-data-eng-proj | Properties| Serverless SQL endpoint**

Create a pipeline in the synapse studio and use the activity GET METADATA and create a dataset in binary format as we need only table names.

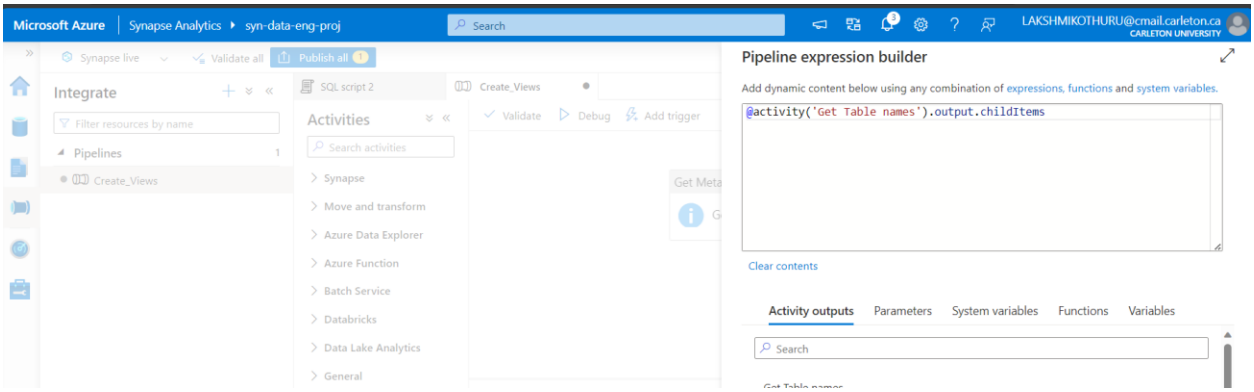
Pass the Field list argument as Child Items as we need table name .



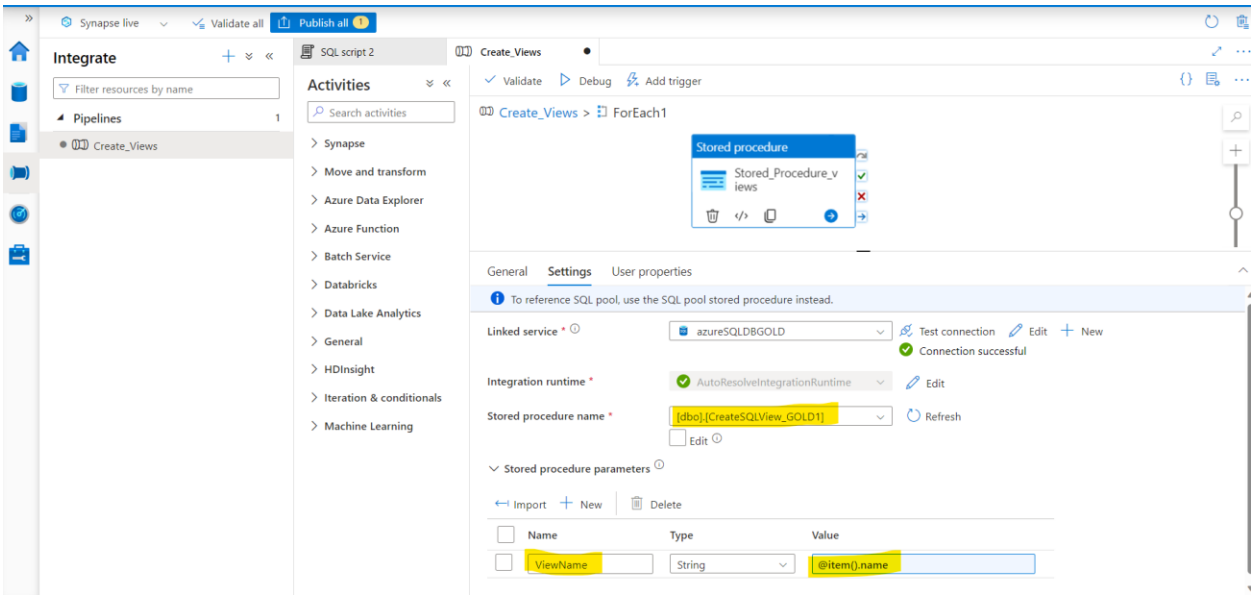
Next, I'm using ForEach, passing the childnames to the Stored procedure I created earlier.

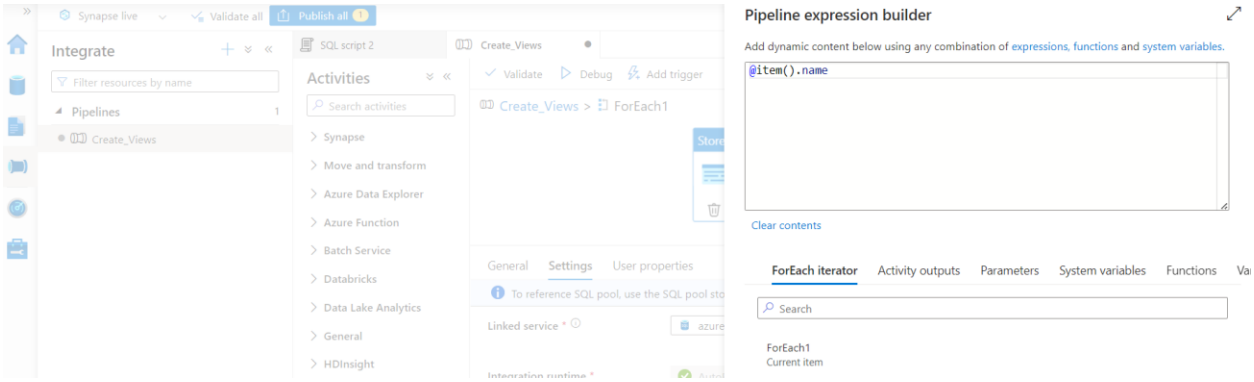


Expression to get table name from the GET METADATA activity

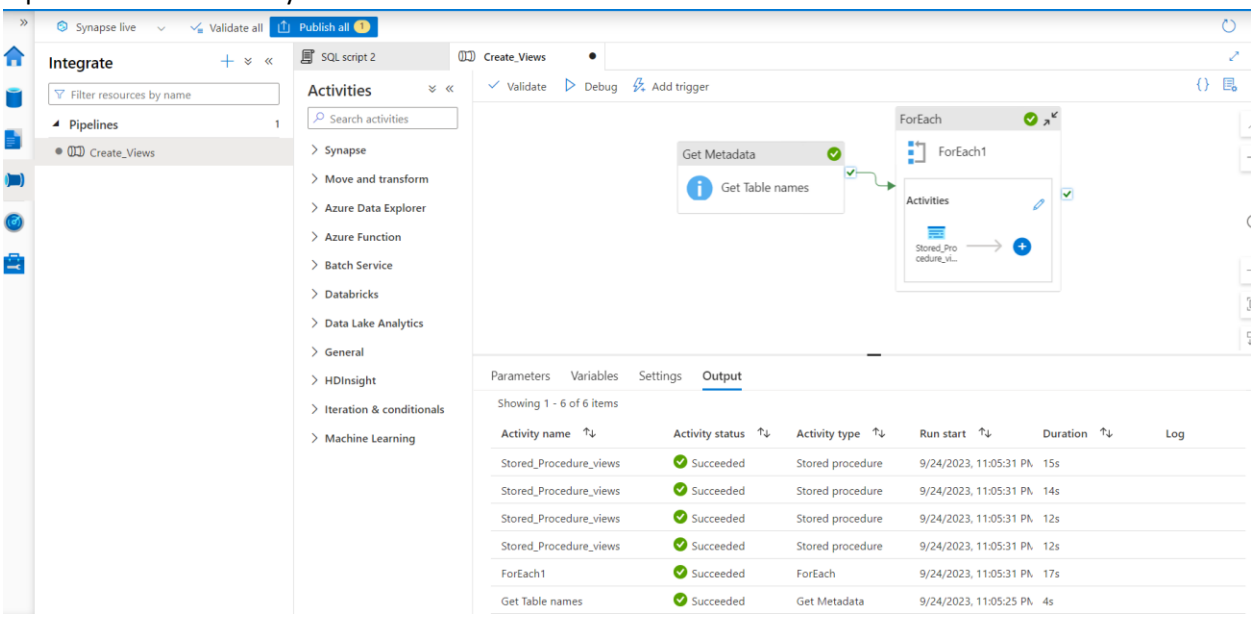


In the For each activity, I'm now calling a stored procedure activity and linking the stored procedure created in SQL Pool.

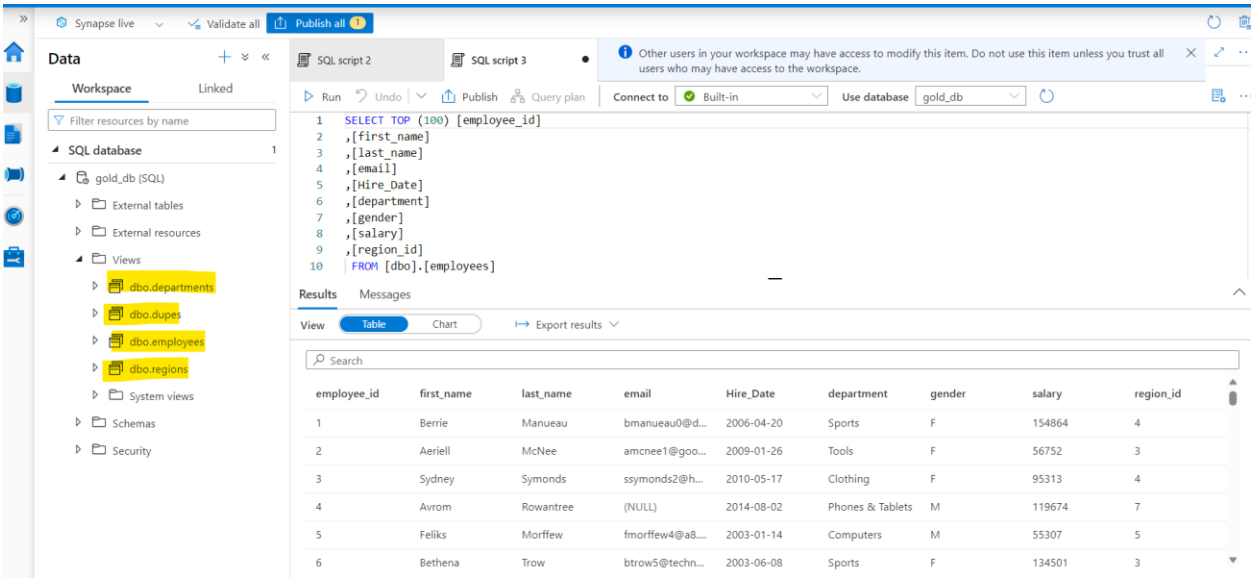

Synapse live
Validate all
Publish all 1
Integrate
Filter resources by name
Pipelines 1
Create_Views
SQL script 2
Create_Views
Activities
Search activities
Synapse
Move and transform
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Validate
Debug
Add trigger
Create_Views > ForEach1
Stored procedure
Stored_Procedure_views
General
Settings
User properties
To reference SQL pool, use the SQL pool stored procedure instead.
Linked service *
azureSQLDBGOLD
Test connection
Edit
New
Connection successful
Integration runtime *
AutoResolveIntegrationRuntime
Edit
Stored procedure name *
[dbo].[CreateSQLView_GOLD1]
Refresh
Edit
Stored procedure parameters
Import
New
Delete
Name
Type
Value
ViewName
String
@item().name


Pipeline expression builder
Add dynamic content below using any combination of expressions, functions and system variables.
@item().name
Clear contents
ForEach iterator
Activity outputs
Parameters
System variables
Functions
Search
ForEach1
Current item

Pipeline Ran successfully

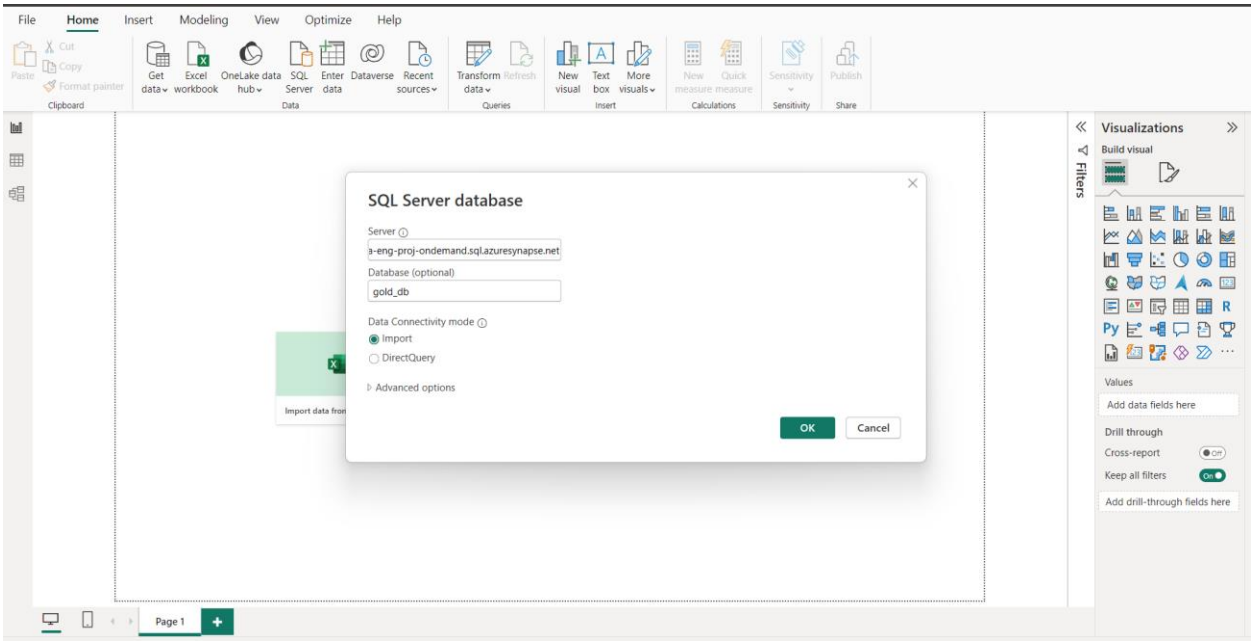


Vailidating the views in Serverless SQL POOL

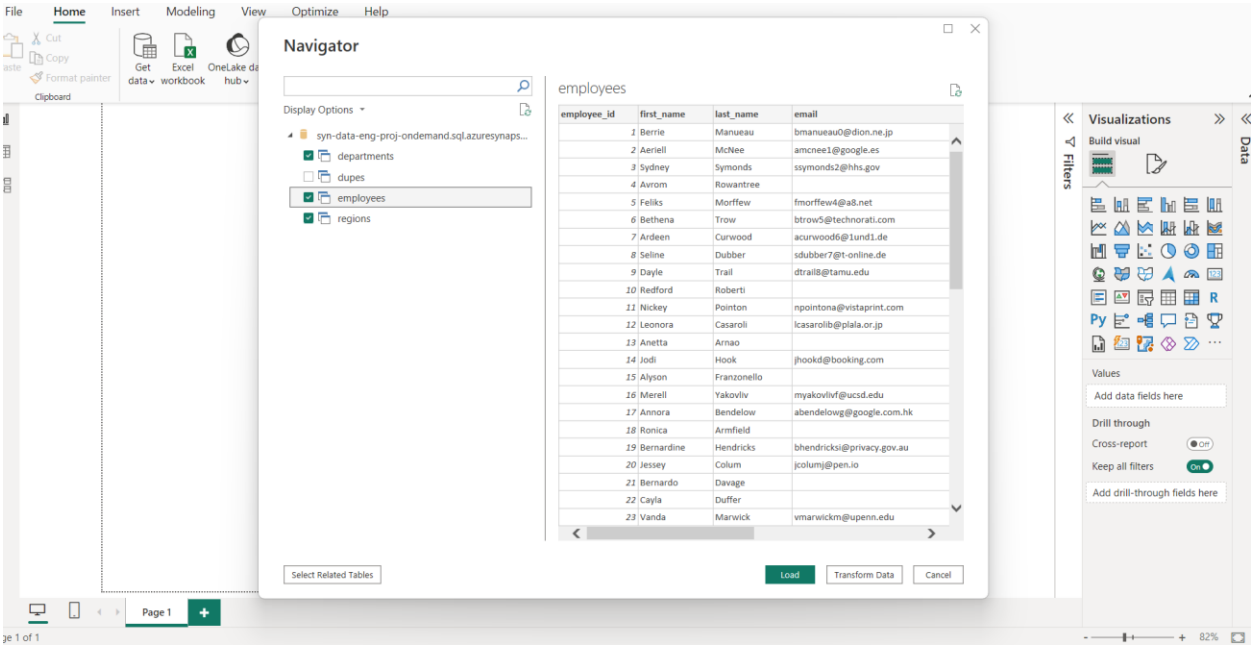


5. Publishing Data Analytics Reports – Power BI

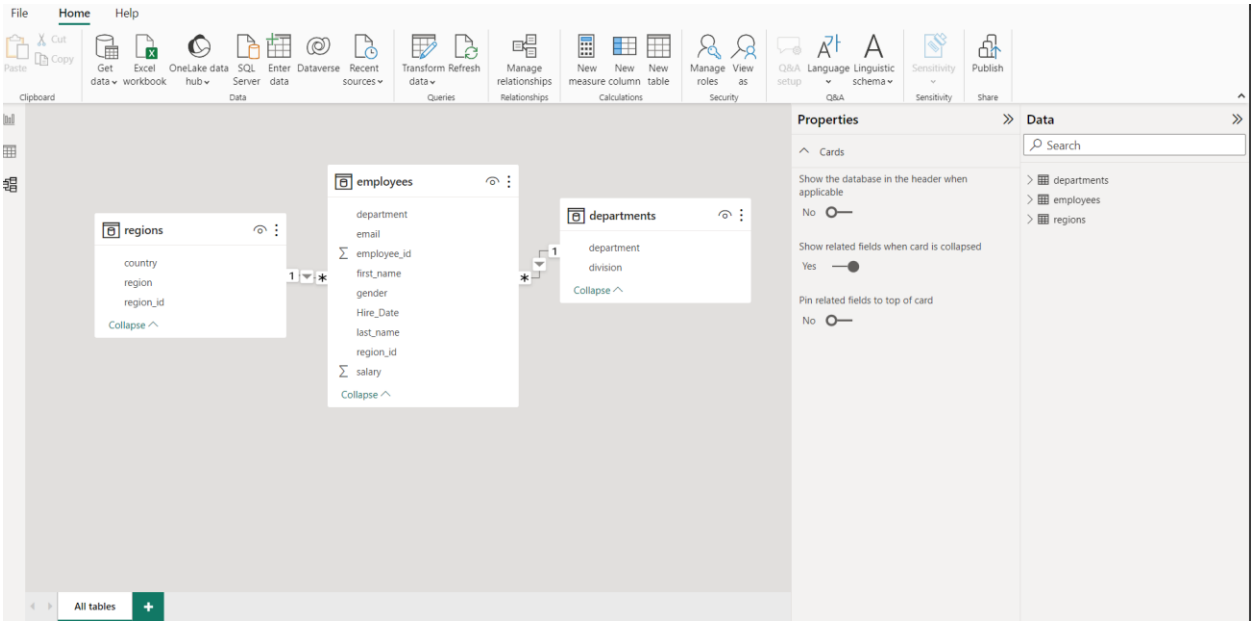
Here I'm Importing data from Azure SQL Database. I've given the Serverless SQL server endpoint , which can be found **synapse-workspace** | Properties| Serverless SQL endpoint



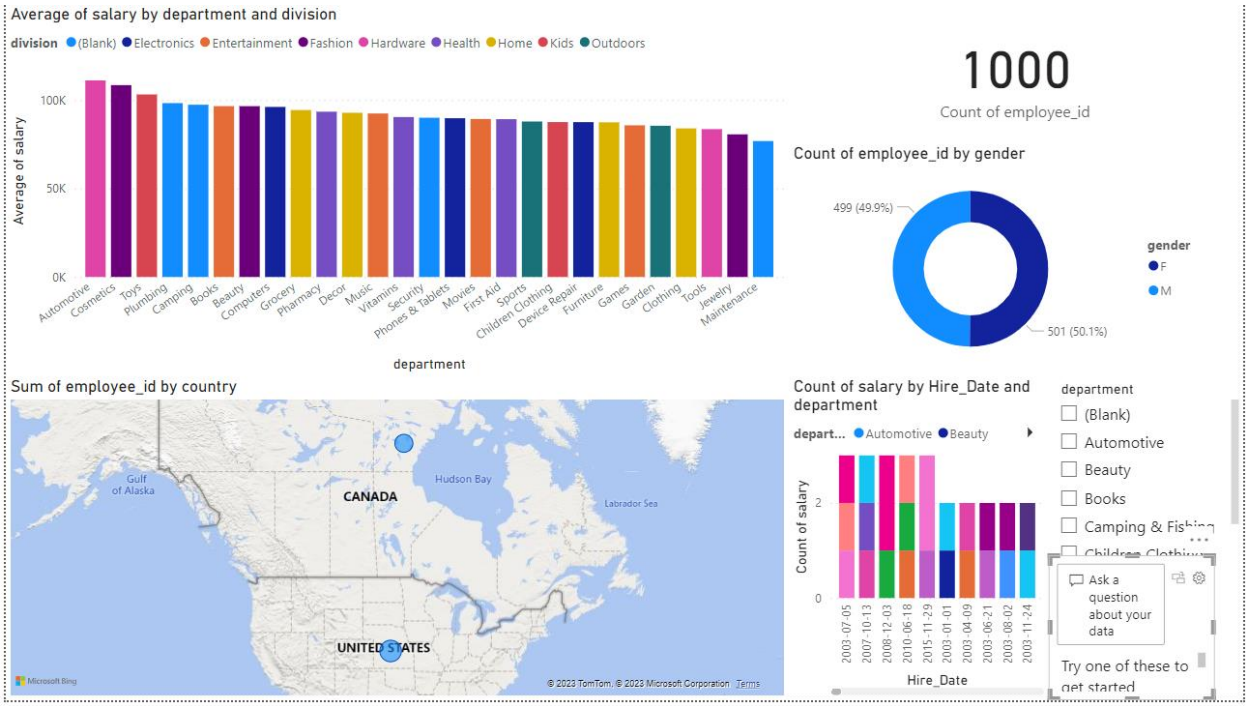
Signed-in with Microsoft account, now able to connect to synapse workspace and import data



Relationship established automatically.



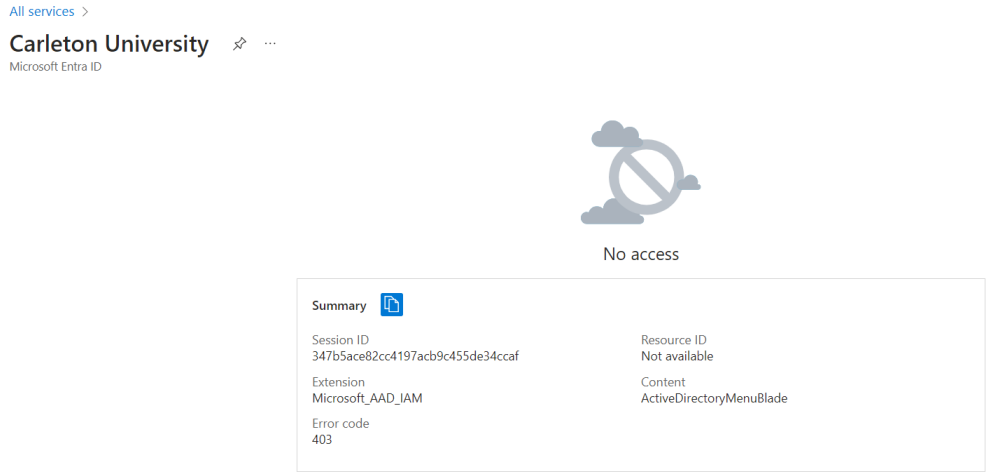
Data Analysed and plotted as a report in Power BI



6. Security and Governance

Next I'll have to create a security group in Azure active directory/Microsoft Entra ID for easy access control in a real-time scenario.

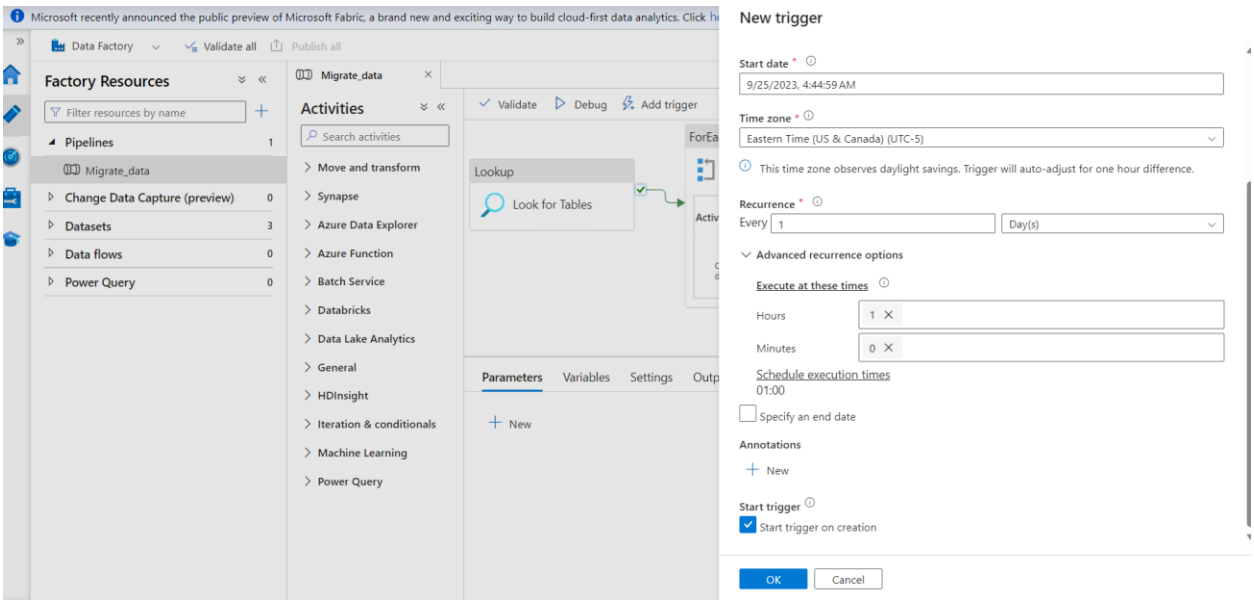
Unfortunately, as mine is a student account I'm unable to access this page.



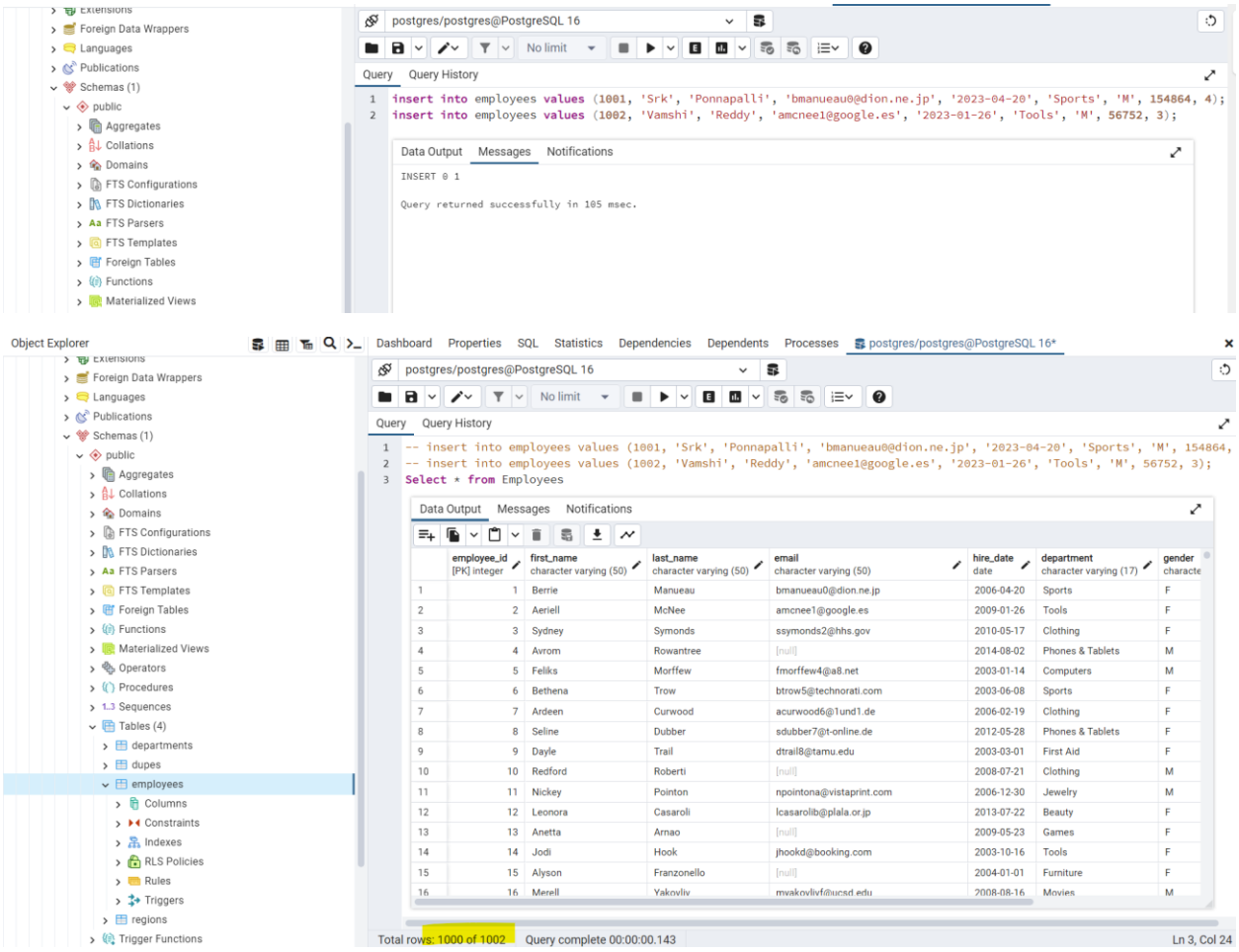
7. End to End Pipeline Testing

This is a end to End Pipeline, meaning any update in the data in the database tables, should automatically be updated everywhere .

To do this we initiate a schedule trigger to run everyday at a specific time.

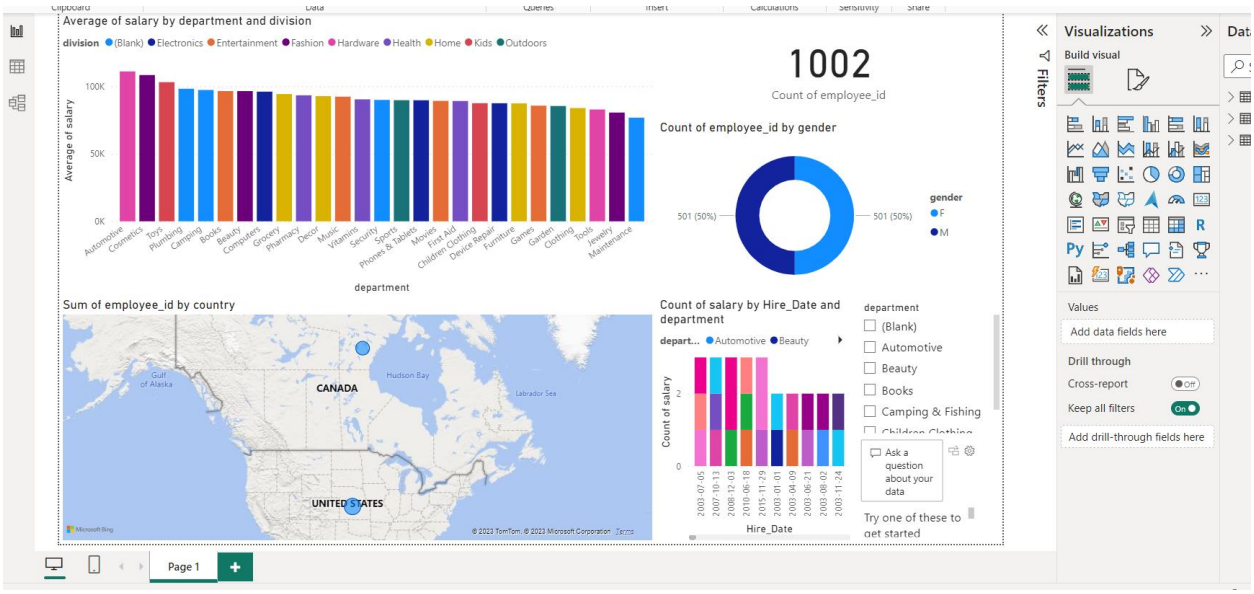


Before this to see the updated the database with 2 records for testing end to end



The Trigger ran successfully, leading to both the pipelines in ADF in Synpase running successfully .

When the Power BI report is refreshed we get the following result.



8. References

1. Azure Data Engineering Project – Mr K Talks
2. Access Azure Data Lake Storage using Azure Active Directory credential passthrough (legacy) - Azure Databricks | Microsoft Learn
3. Access Azure Data Lake Storage using Azure Active Directory credential passthrough (legacy) - Azure Databricks | Microsoft Learn
4. pyspark.sql.DataFrame.join — PySpark master documentation (databricks.com)