The tools that are used in this project are

1. Azure Data Factory

2. Azure Data Lake Storage Gen2

3. Azure Databricks

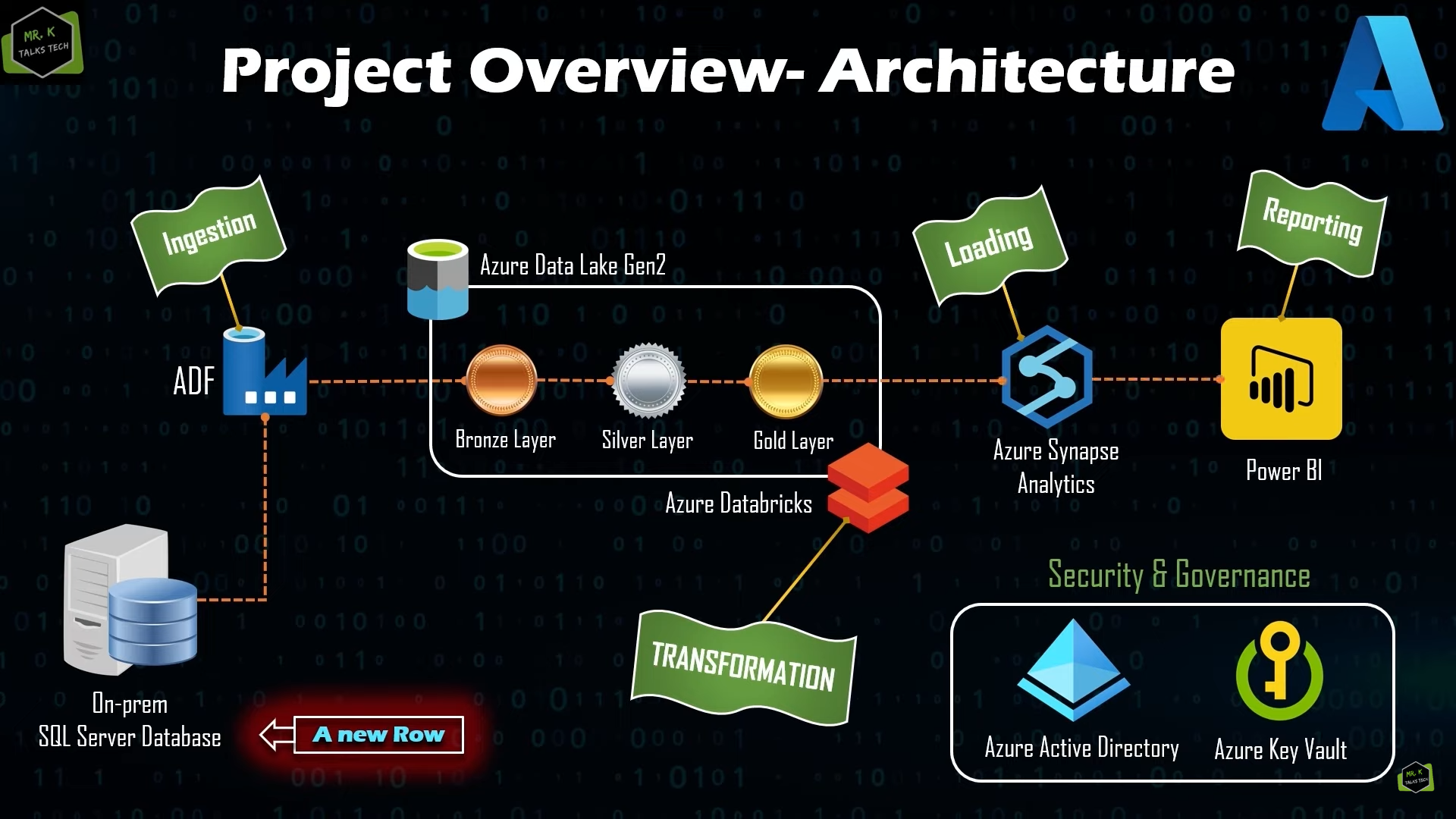
4. Azure Synapse Analytics

5. Azure Key vault

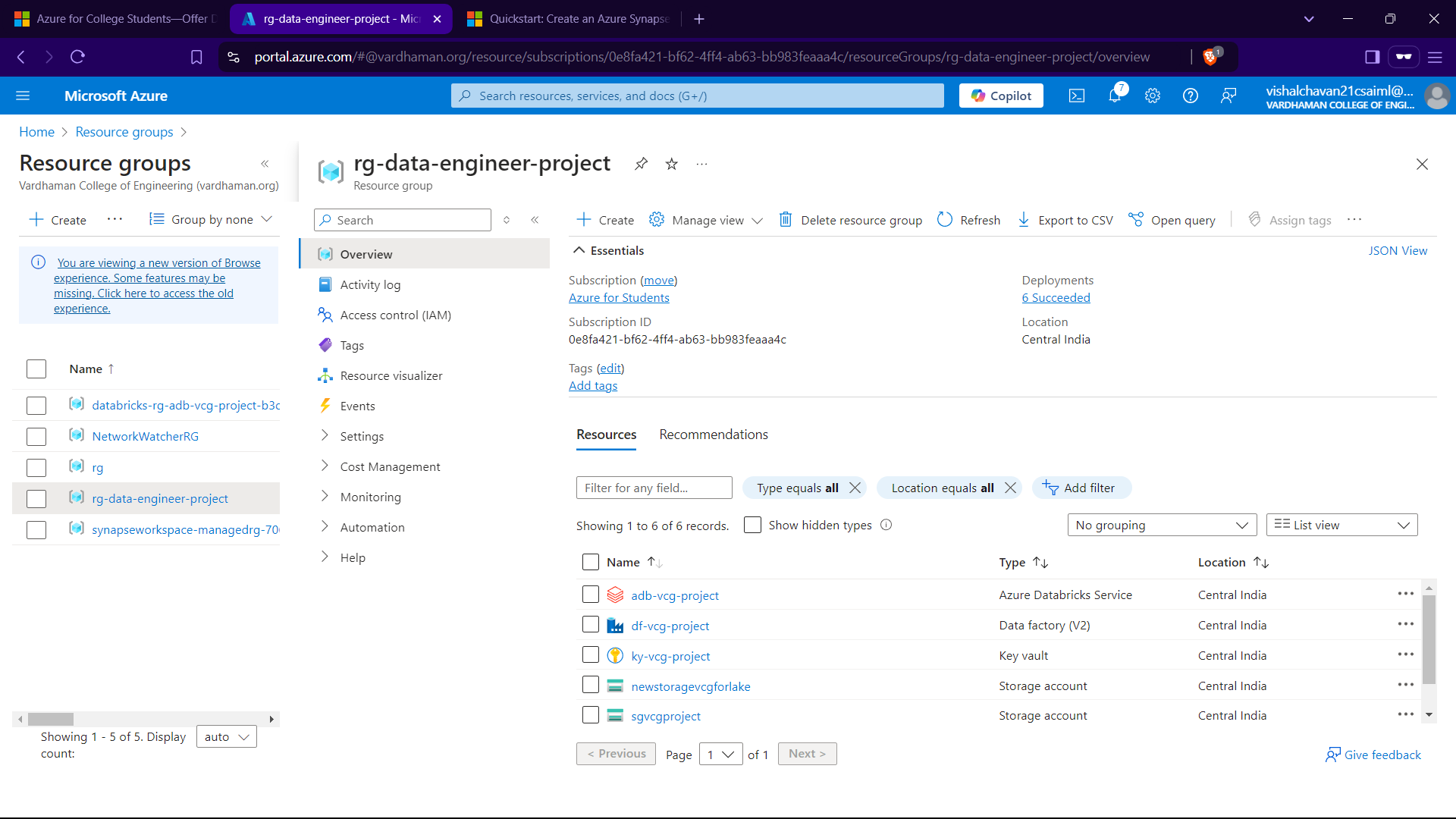
6. Azure Active Directory (AAD) and

7. Microsoft Power BI

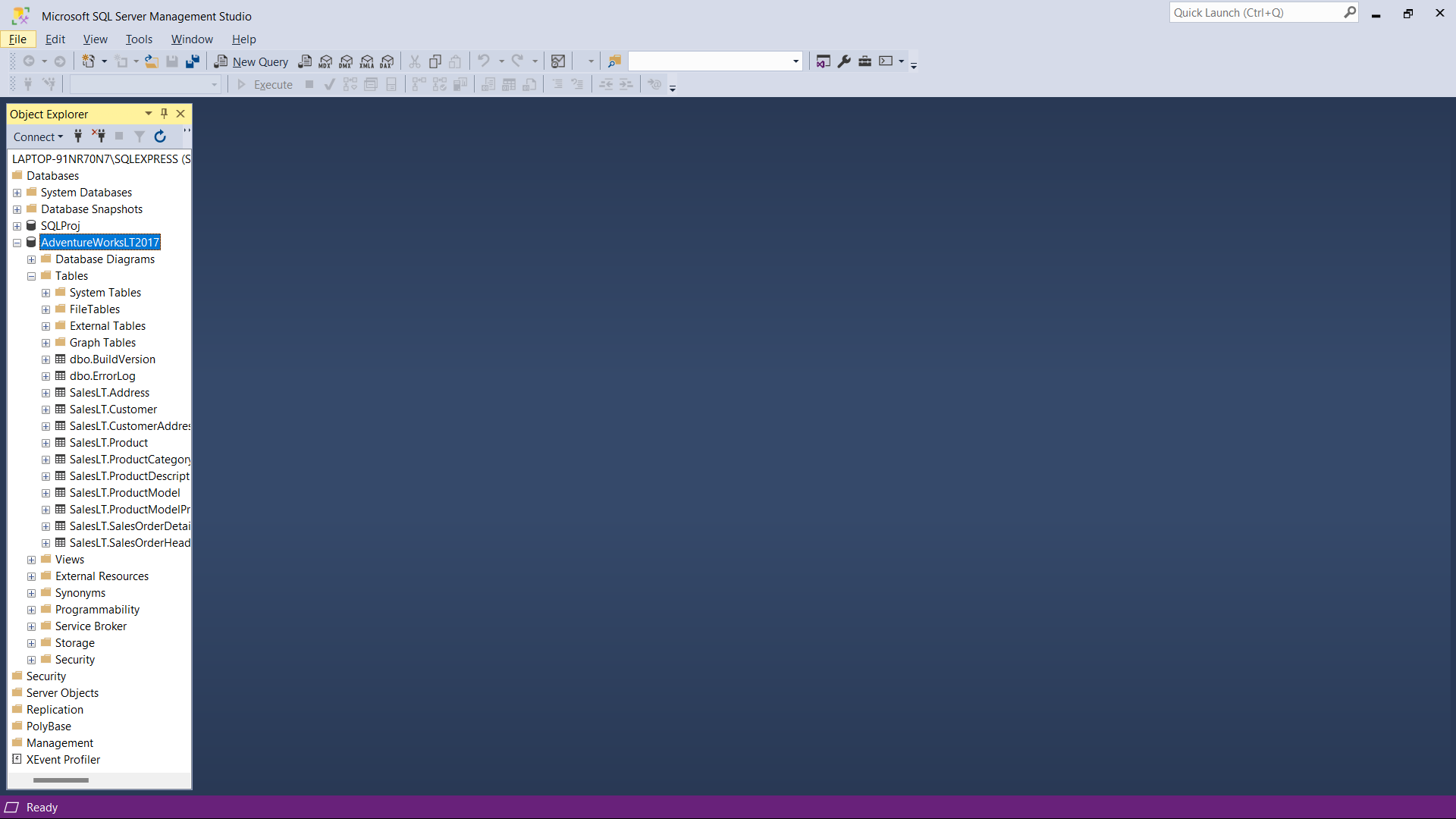
The use case for this project is building an end to end solution by ingesting the tables from on-premise SQL Server database using Azure Data Factory and then store the data in Azure Data Lake. Then Azure databricks is used to transform the RAW data to the most cleanest form of data and then we are using Azure Synapse Analytics to load the clean data and finally using Microsoft Power BI to integrate with Azure synapse analytics to build an interactive dashboard. Also, we are using Azure Active Directory (AAD) and Azure Key Vault for the monitoring and governance purpose.



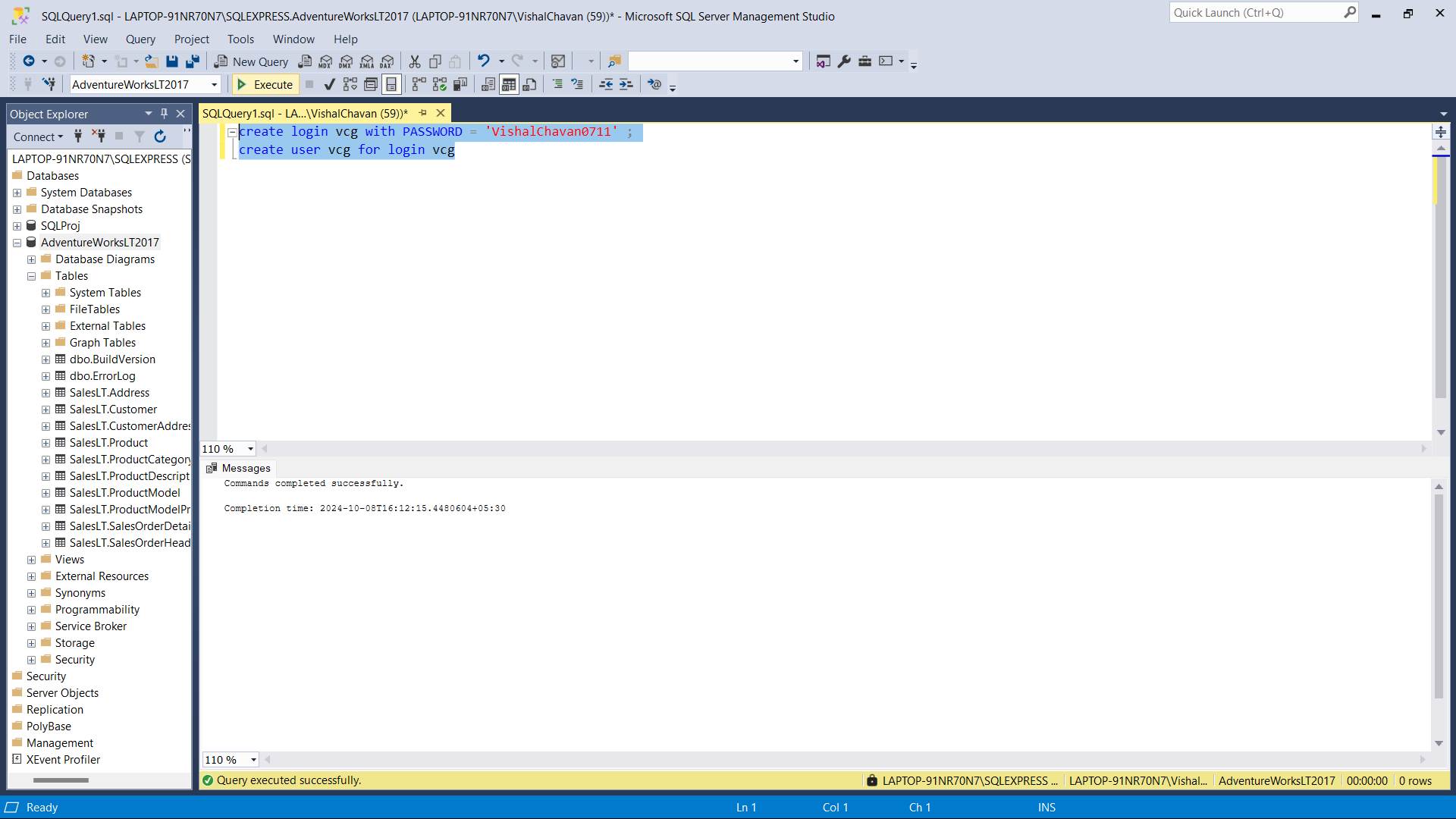
Section I-Environment Setup



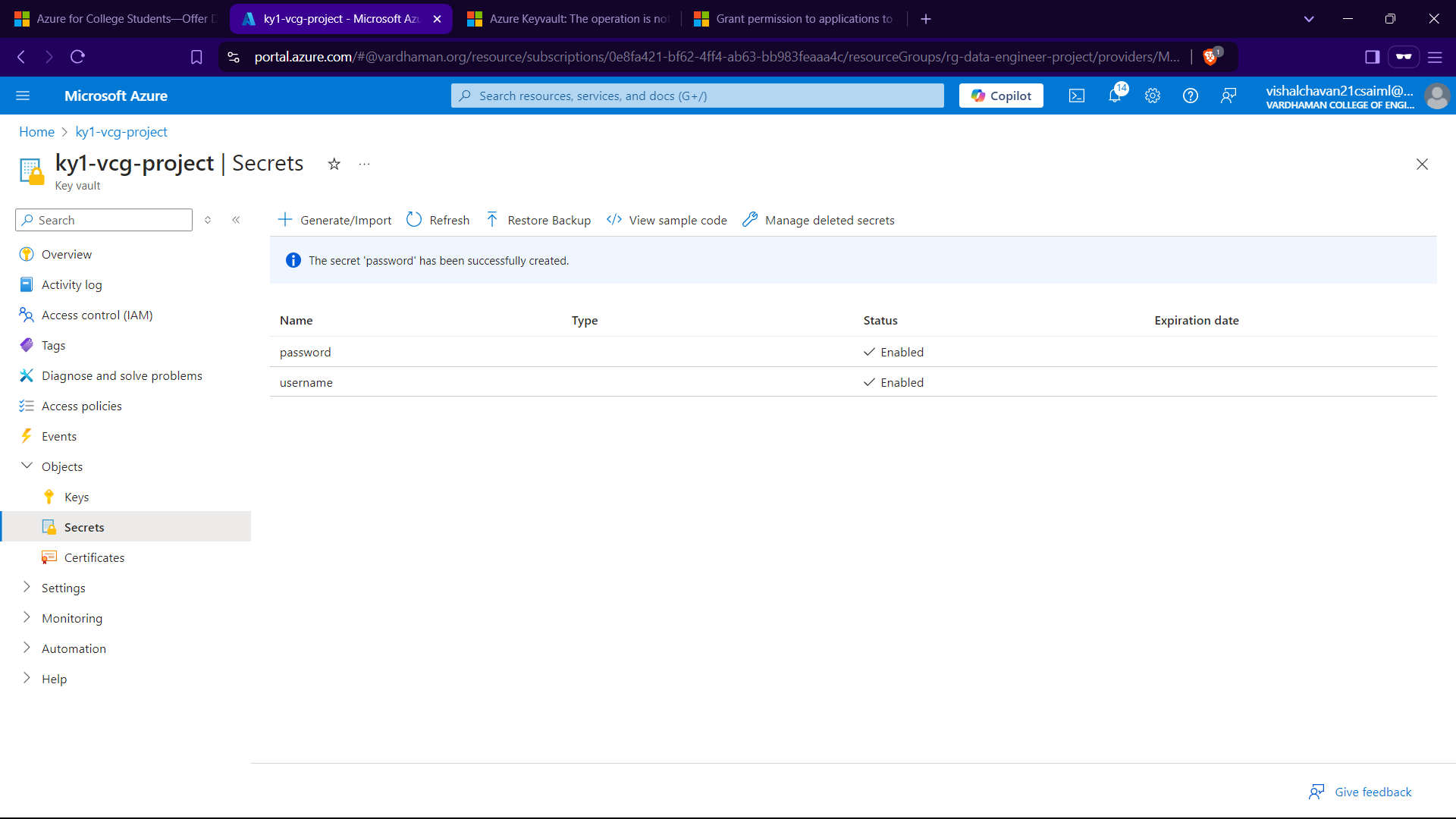
So there is no connection bw azure and this database which is stored in my local machine. And we are interested only in transporting only the SalesLT files not the dbo files.



Making credentials for us to access the DB in further use.and giving the user vcg the db reader aces from security in DB and by clicking o vcg right click properties ->memberships->dbreader->ok

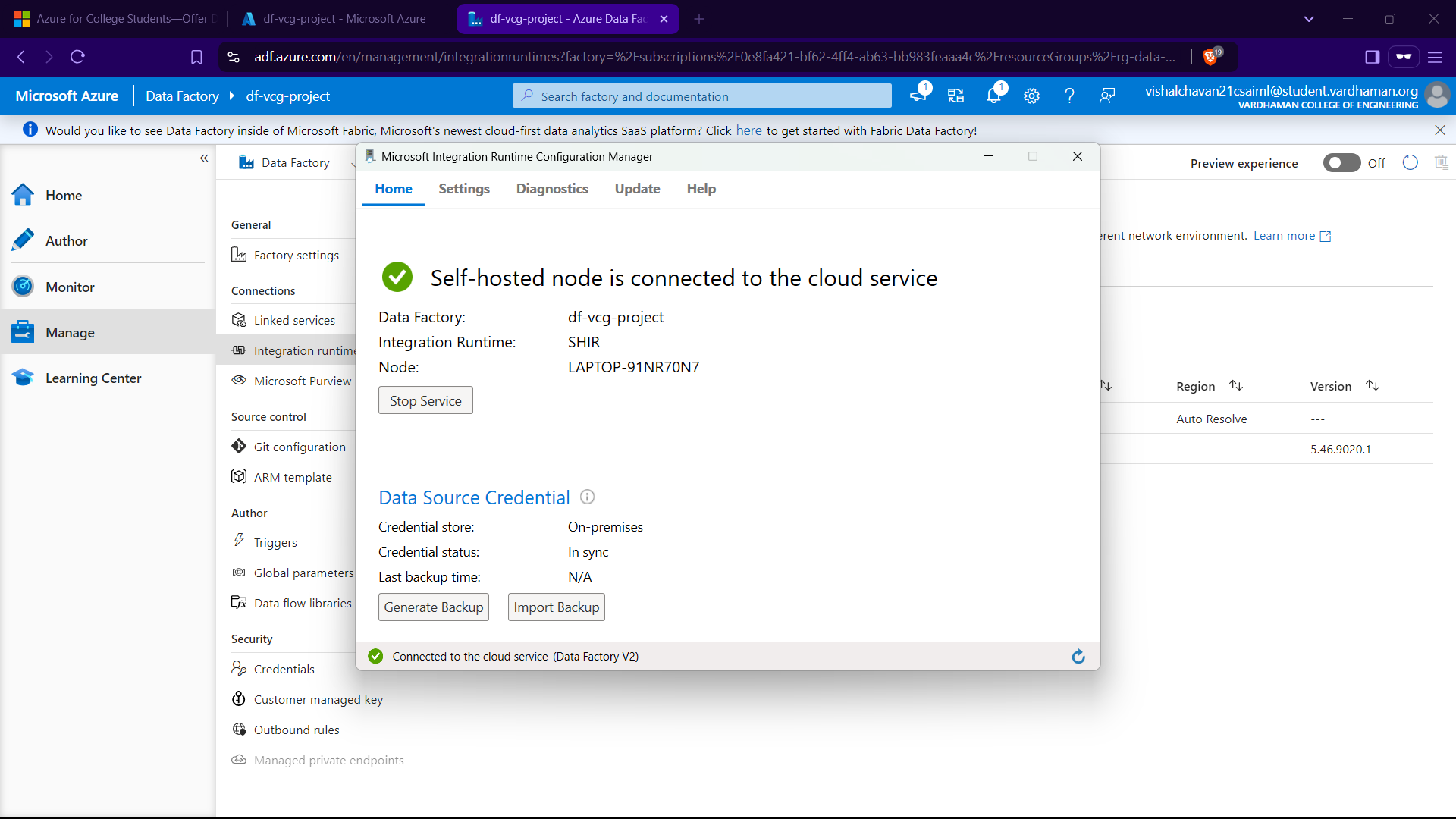


In key vaults assign urself Key Vault Administrator Role/Permission to Genereate Secrets from keyvault.(problem resolved here self google)

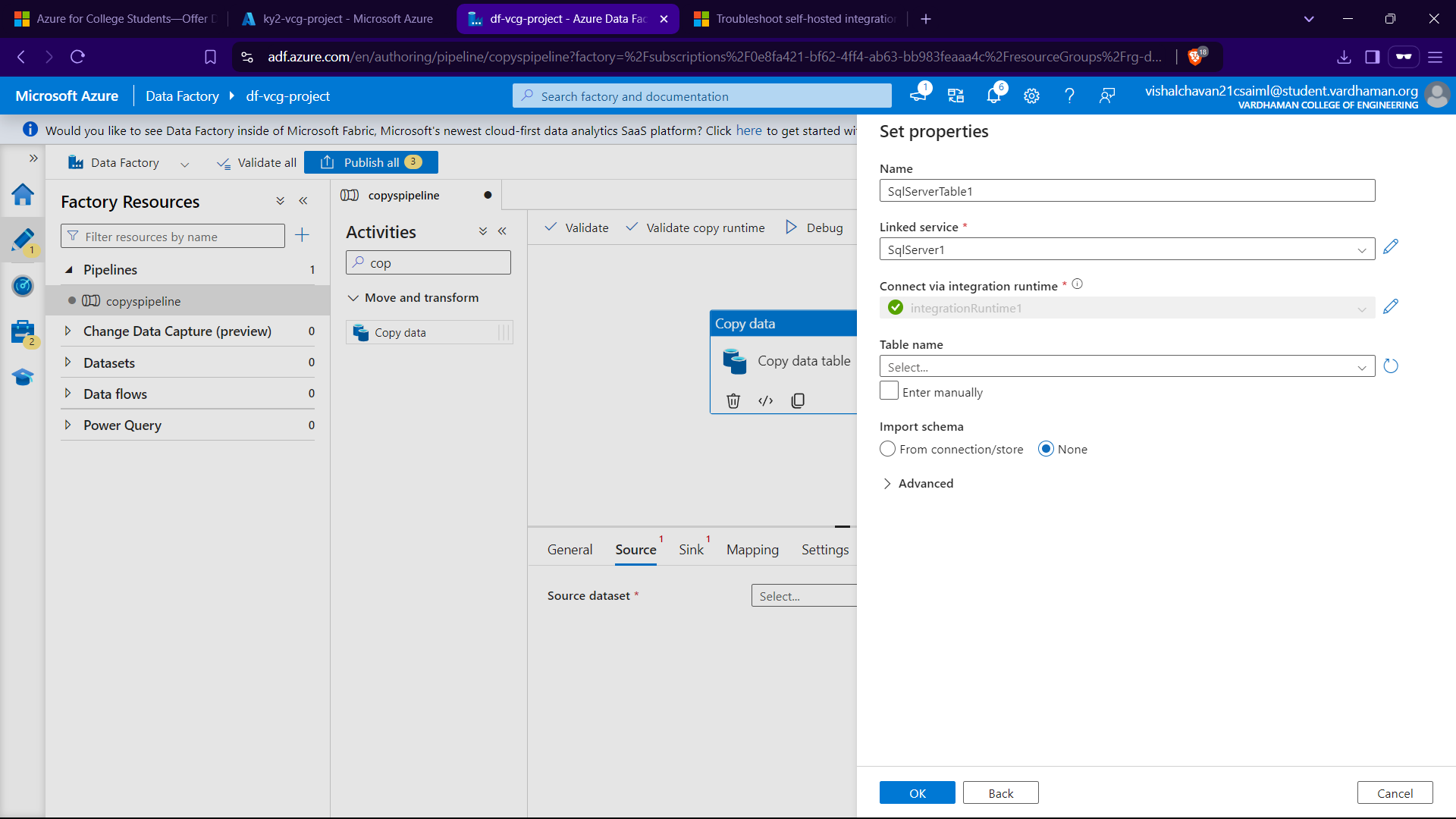


Section II – Data Ingestion

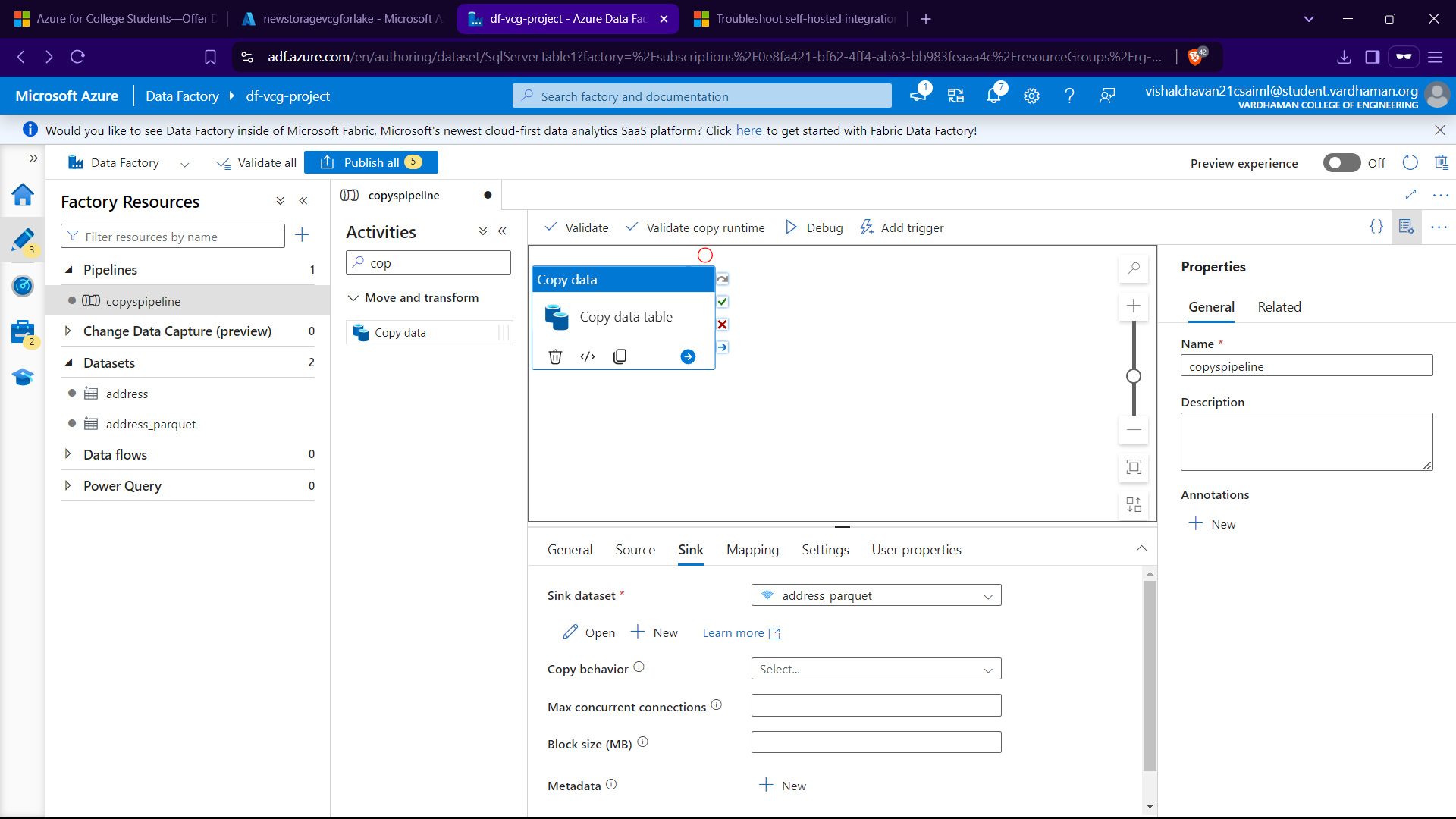
IN ADF go to Manage Integration Runtime Create Self hosted 2 times



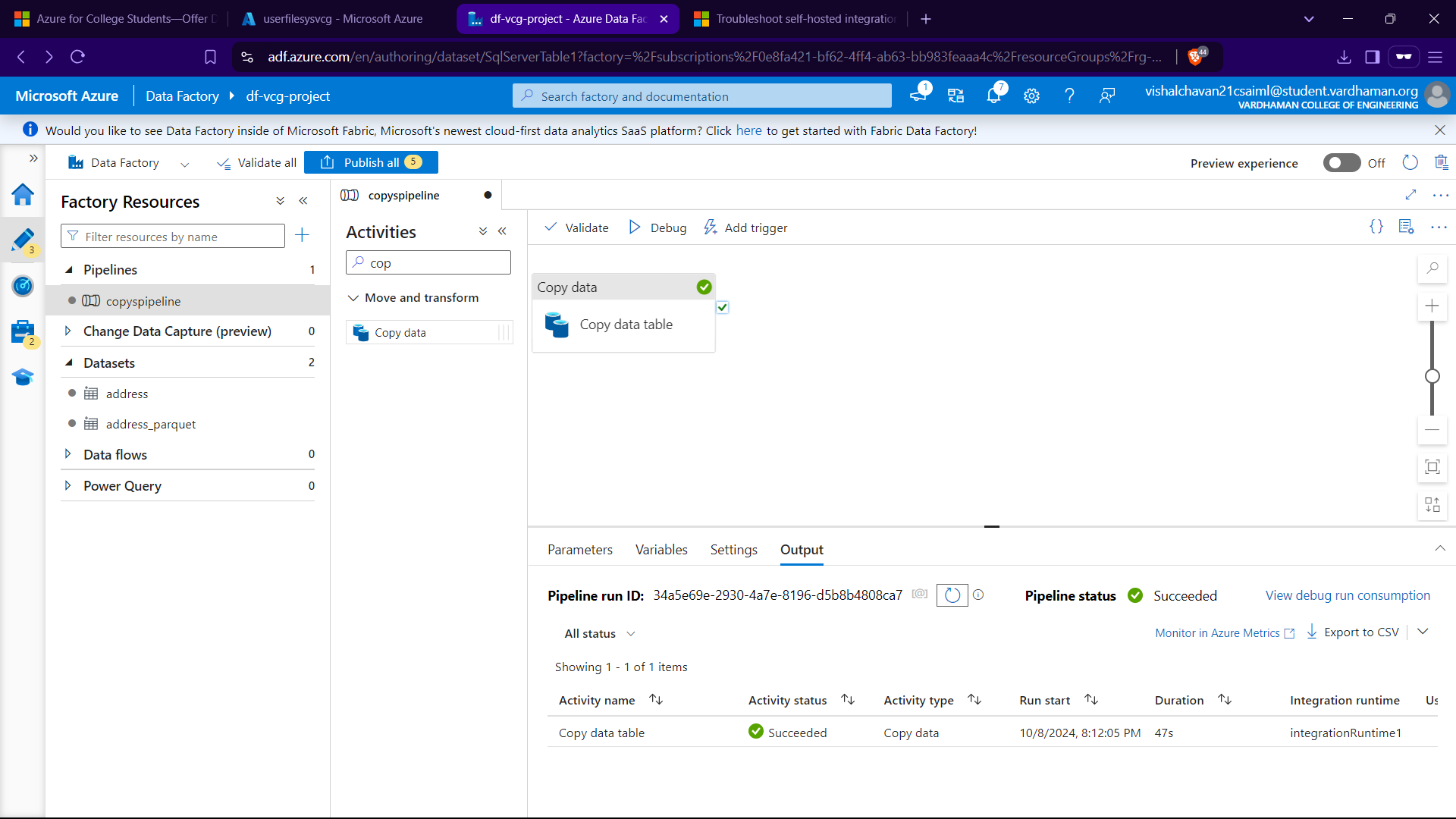
Go to author create a pipe line go to copy data in activites and then sourse and add linked service using keyvault and self hosted integration runtime and select table name



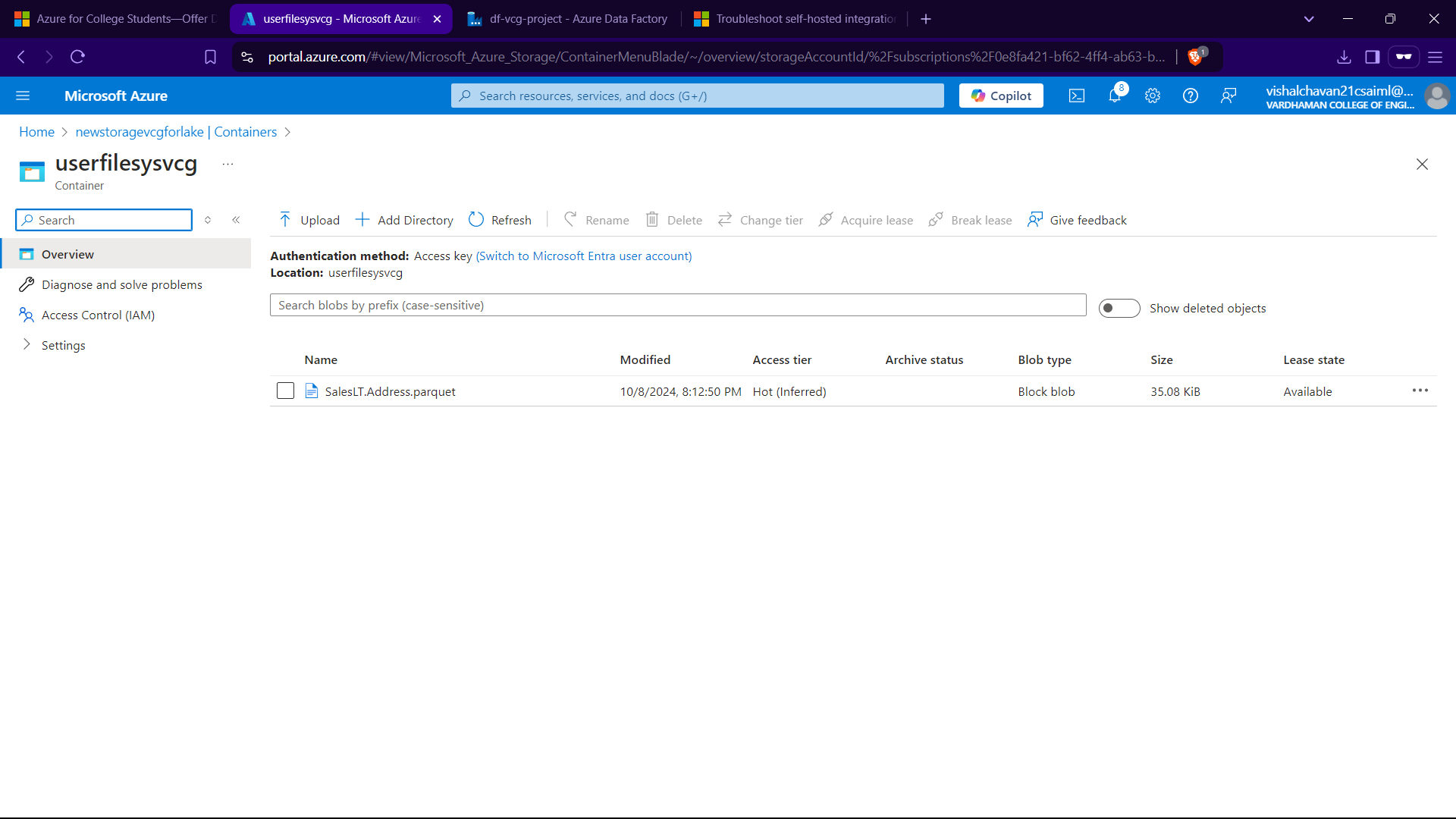
Go to sink for azure datalake and parquet format of data with Auto IR unlike for source ie Self hosted



Now to test the pipe line we can use Debug and trigger option here we use debug



After debug we go to the container specified fot datalake and hit refresh to get the address table in container



Publish all the canges in ADF->pipeline

Till above we saw that one pipeline was used to copy one file from onpremise to the Cloud now we will make the whole dataset to be copied in one pipeline itself.

Drag the loop up activity in new pipeline and drag and drop it.

After managing setting don’t select any tables and create then in settings select query as give this:

select

s.name as SchemaName,

t.name as TableName

from sys.tables t

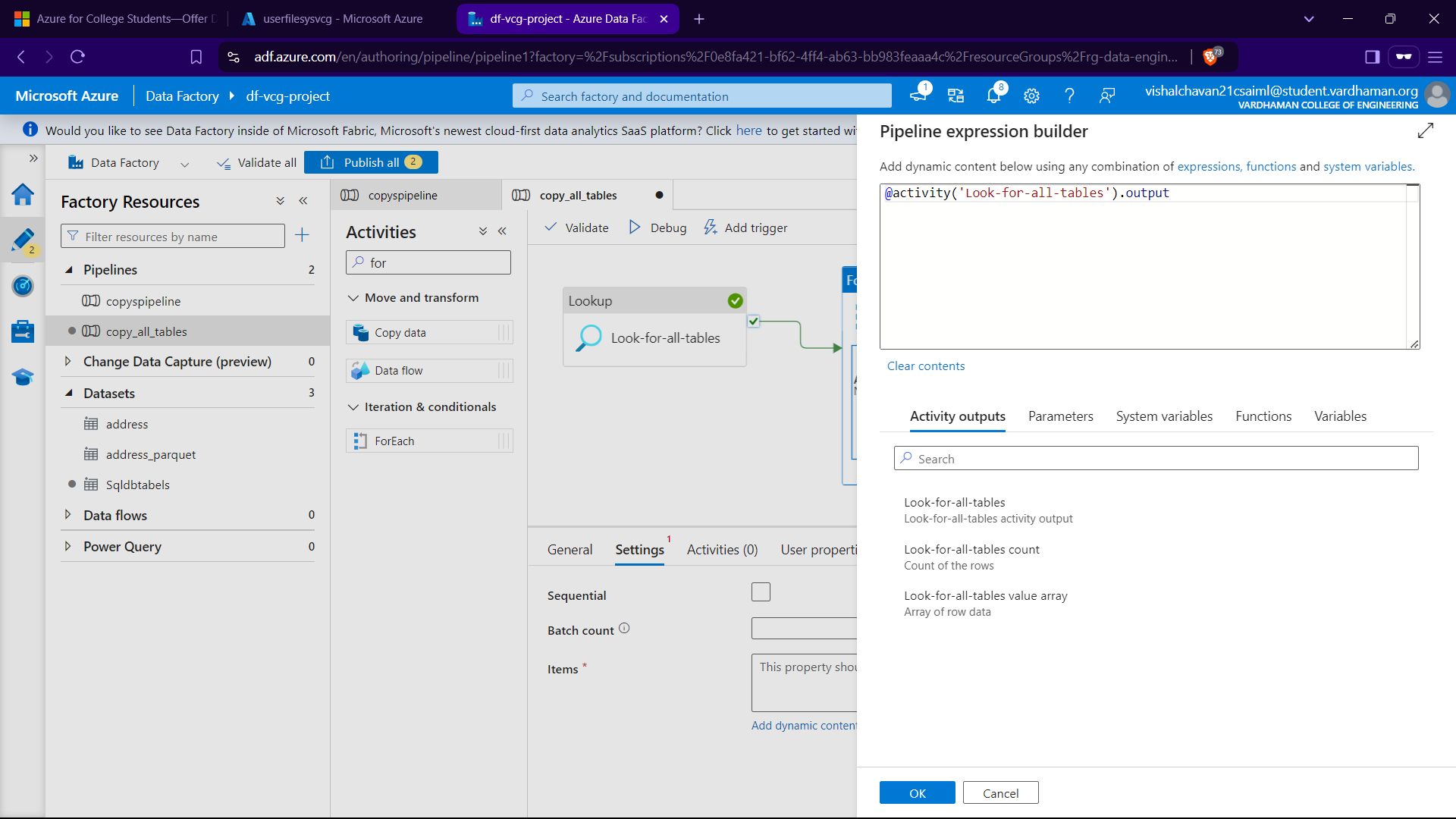
inner join sys.schemas s

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

where s.name='SalesLT'

then click on preview data to make sure if its working fine(listing all tables) and debug it

now D&D from activities :”For each”



Then in foreach acctivity go again d&d copy data and perform the source with query as:

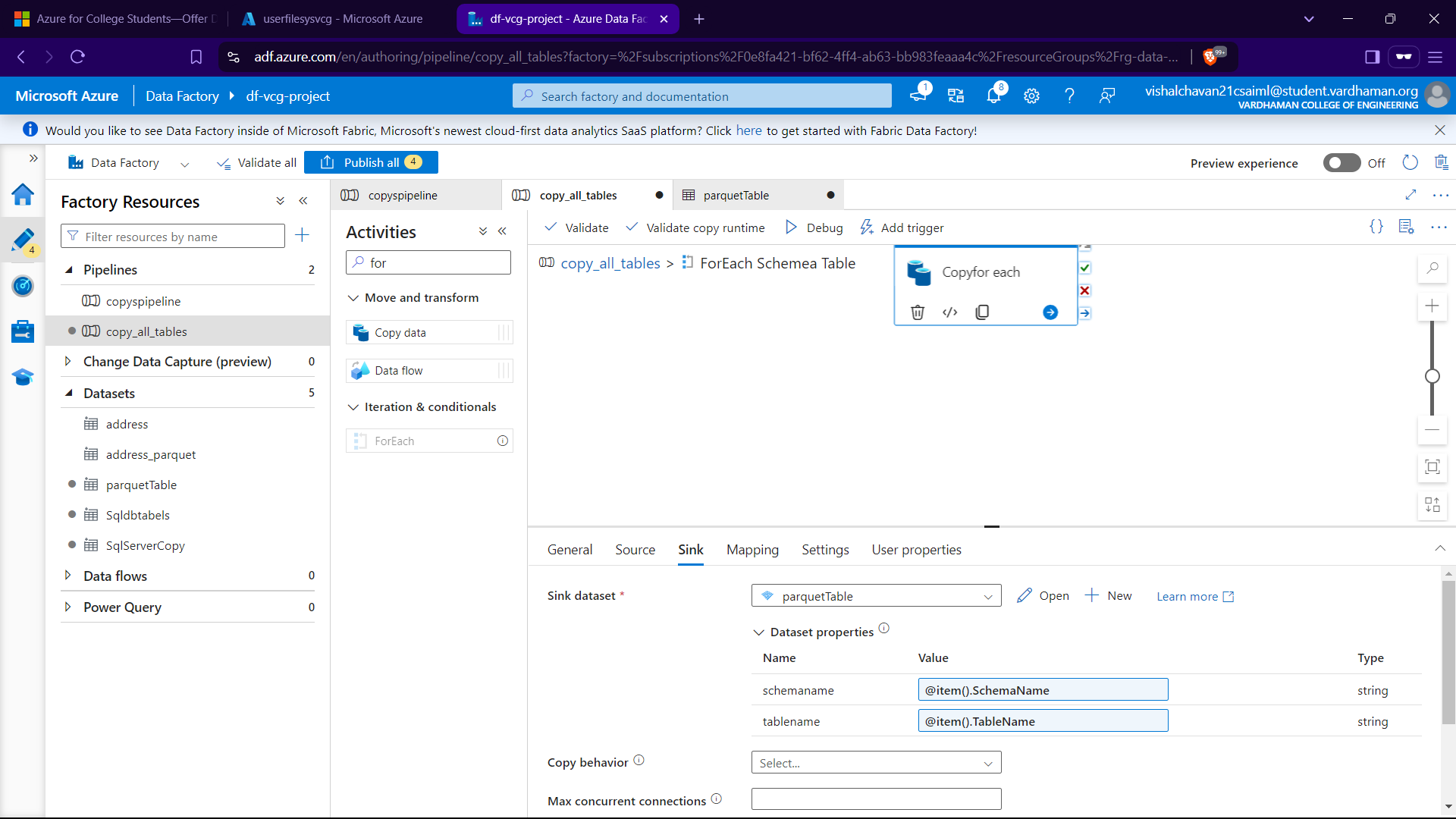
@{concat('SELECT \* FROM ',item().SchemaName,'.',item().TableName)}

And sink as normal by datalake.

We want the structure as:

Container/Schema/tableName/TableName.parquet

After adding 2 parameters in parquetTable



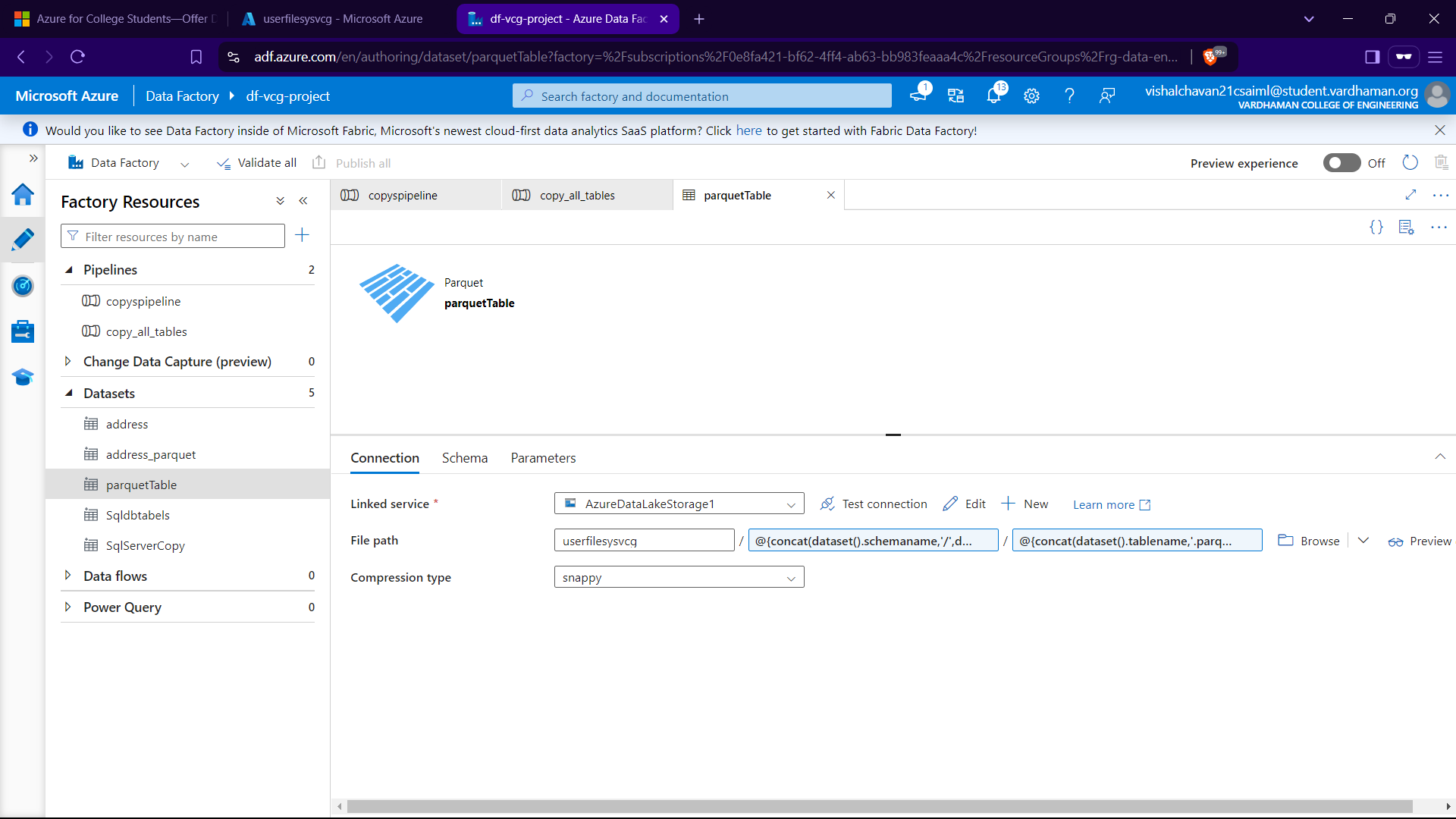
In File name in parquetTable->Connection

For dir name :

@{concat(dataset().schemaname,'/',dataset().tablename)}

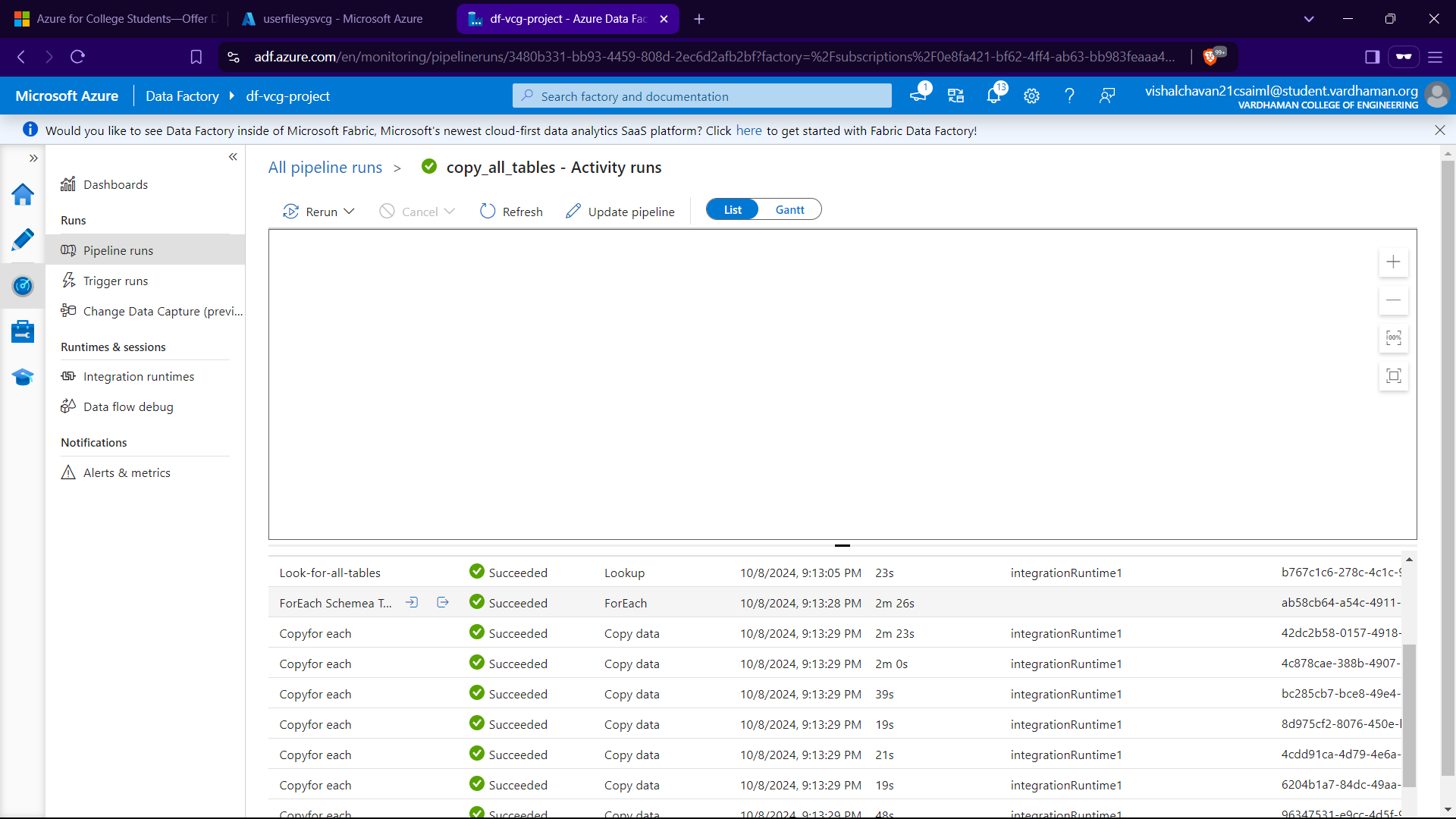
For file name:

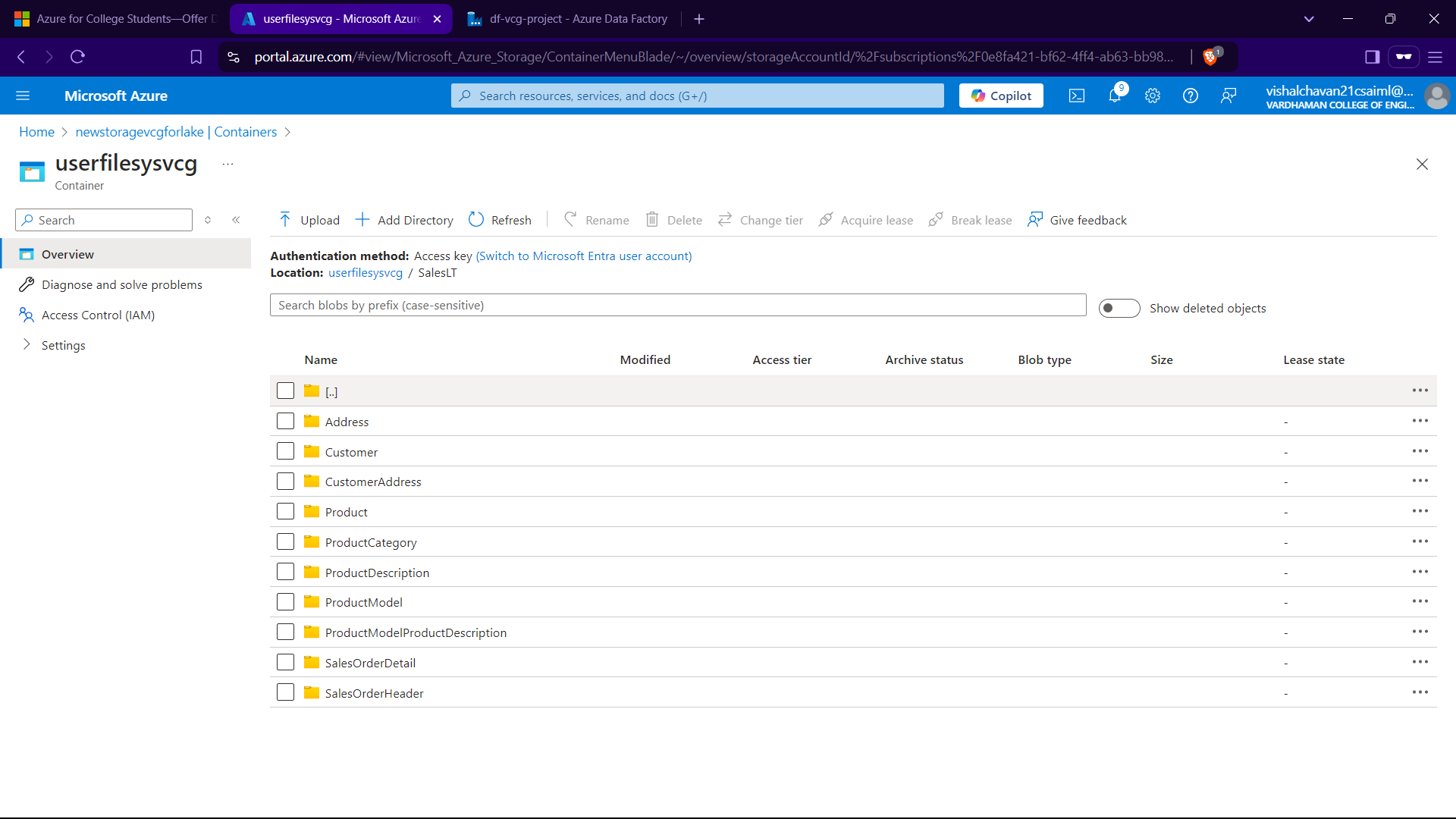
@}concat(dataset().tablename,'.parquet')}



Now Publish

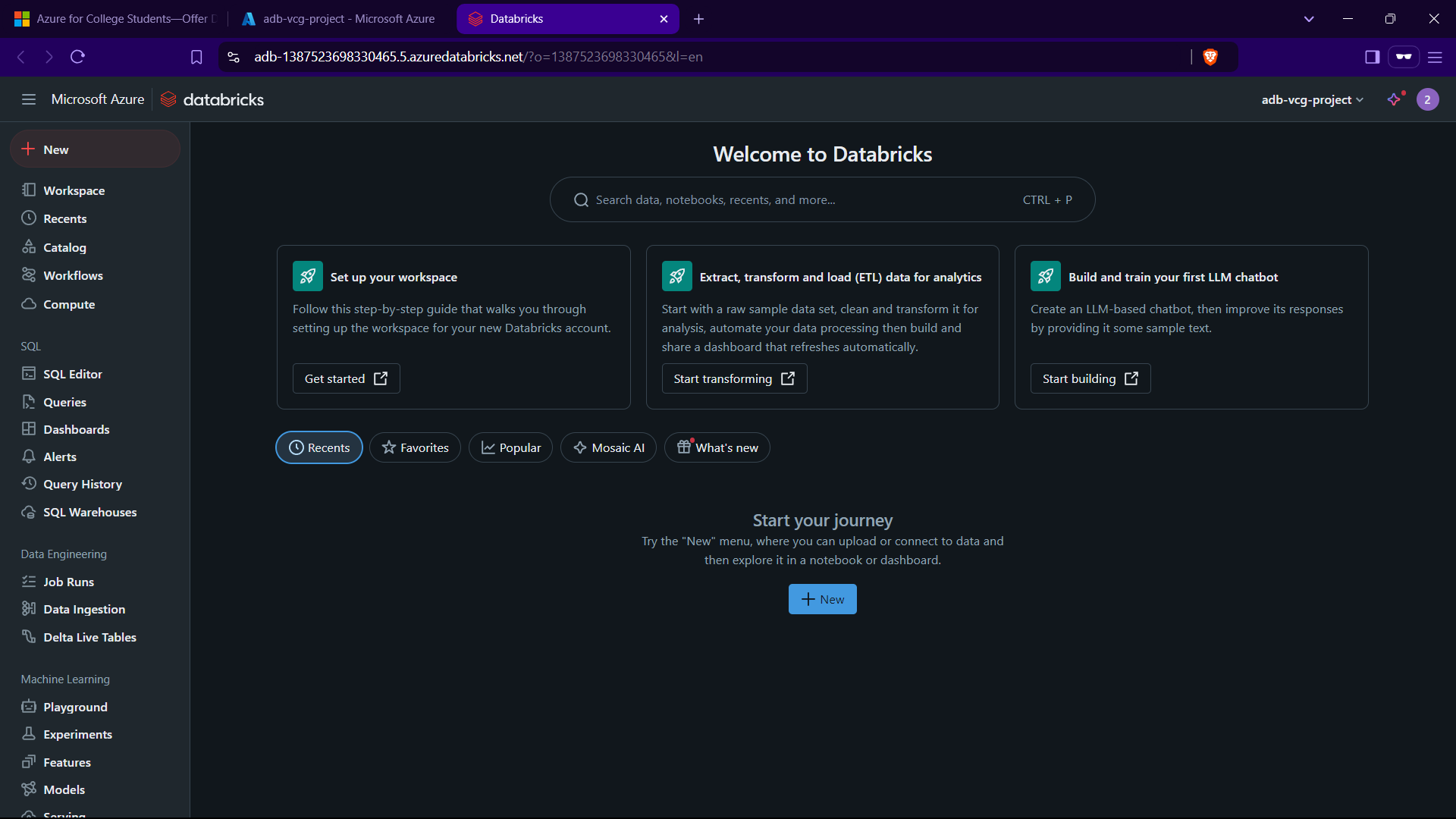
For this lets explore Add Trigger Option->TriggerNow->ok



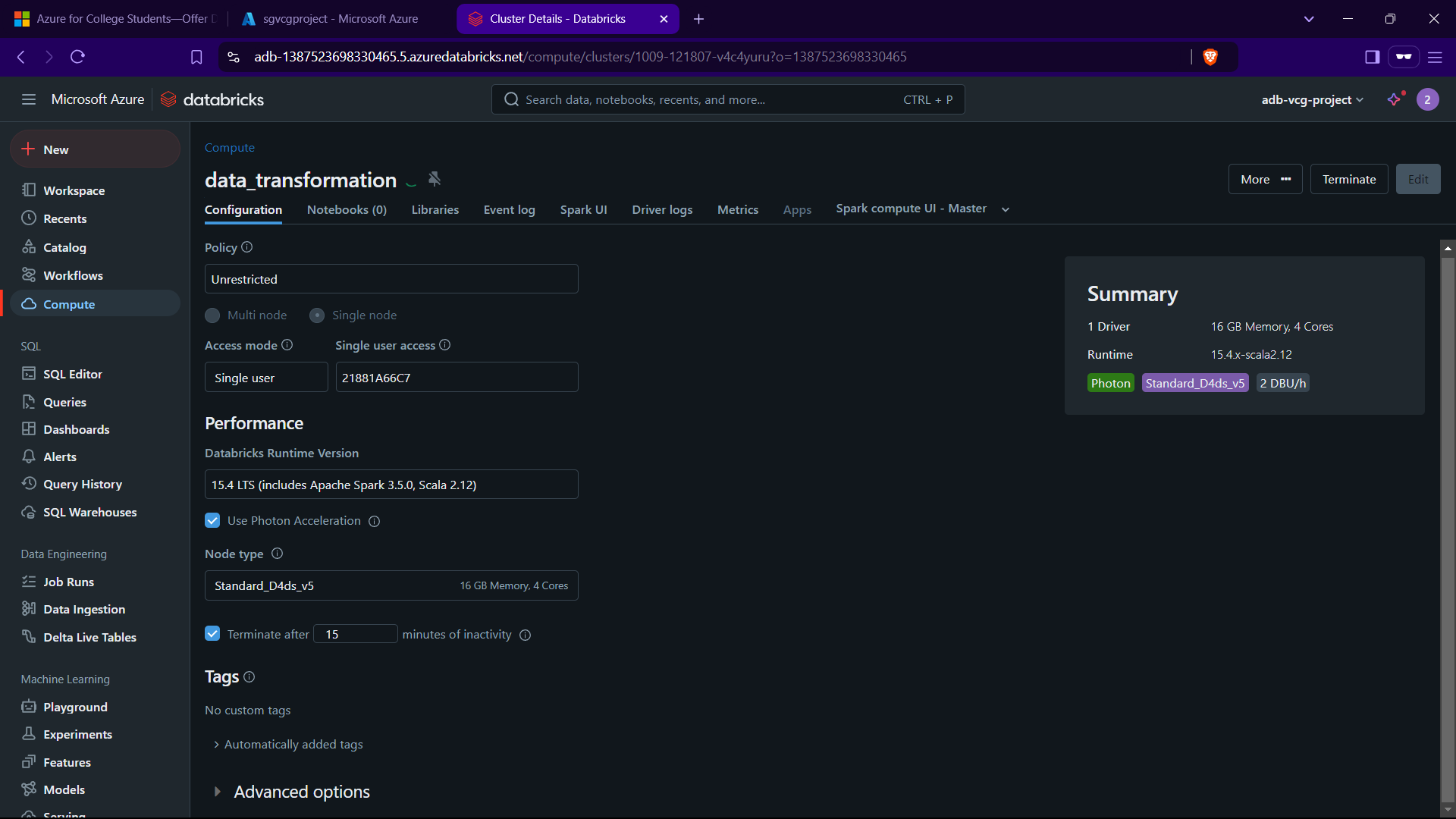


Till here we did the Data Ingestion Process and have Completed till Bronze Layer.

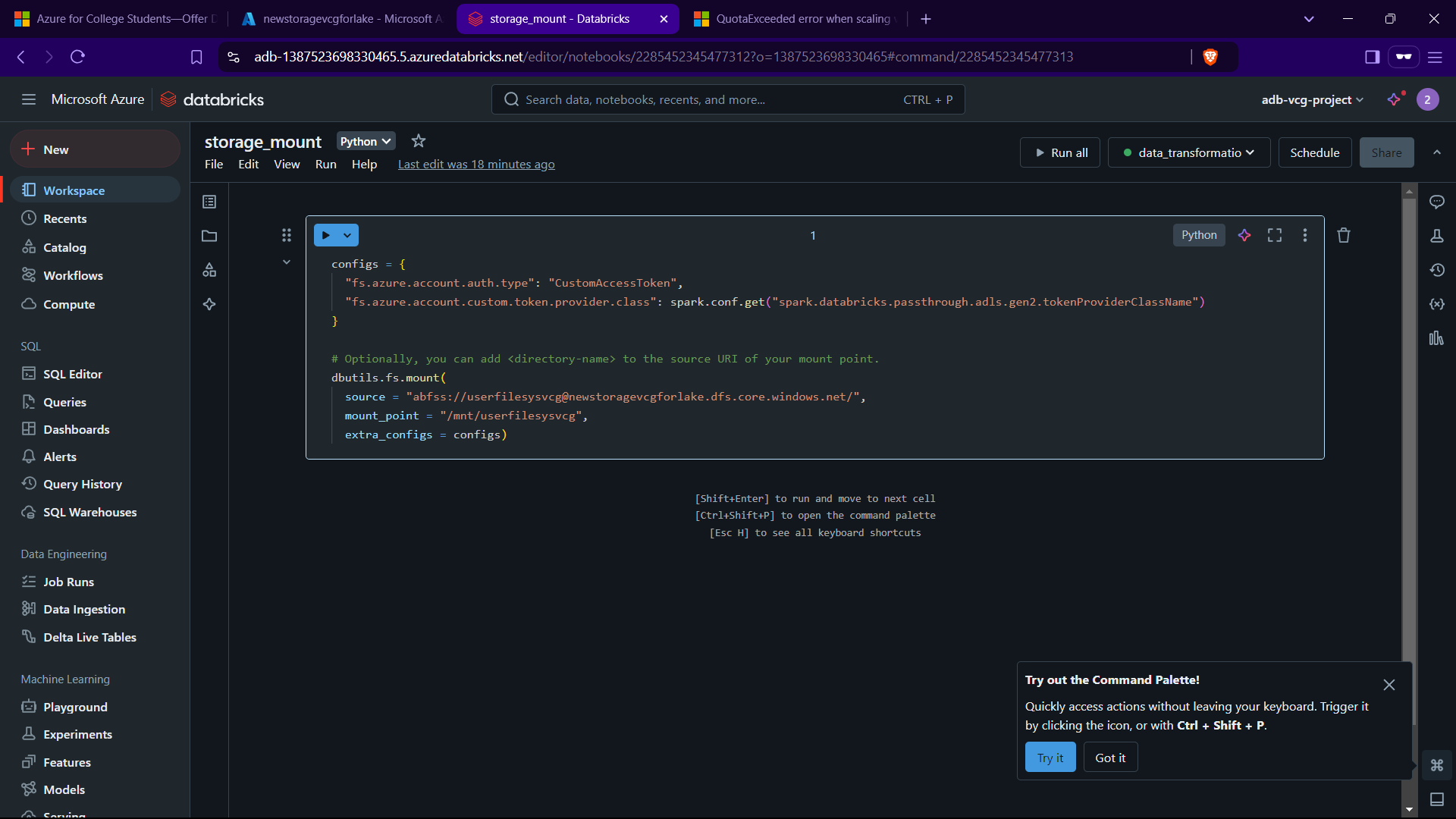
Section III: Data Transformation we use DataBricks



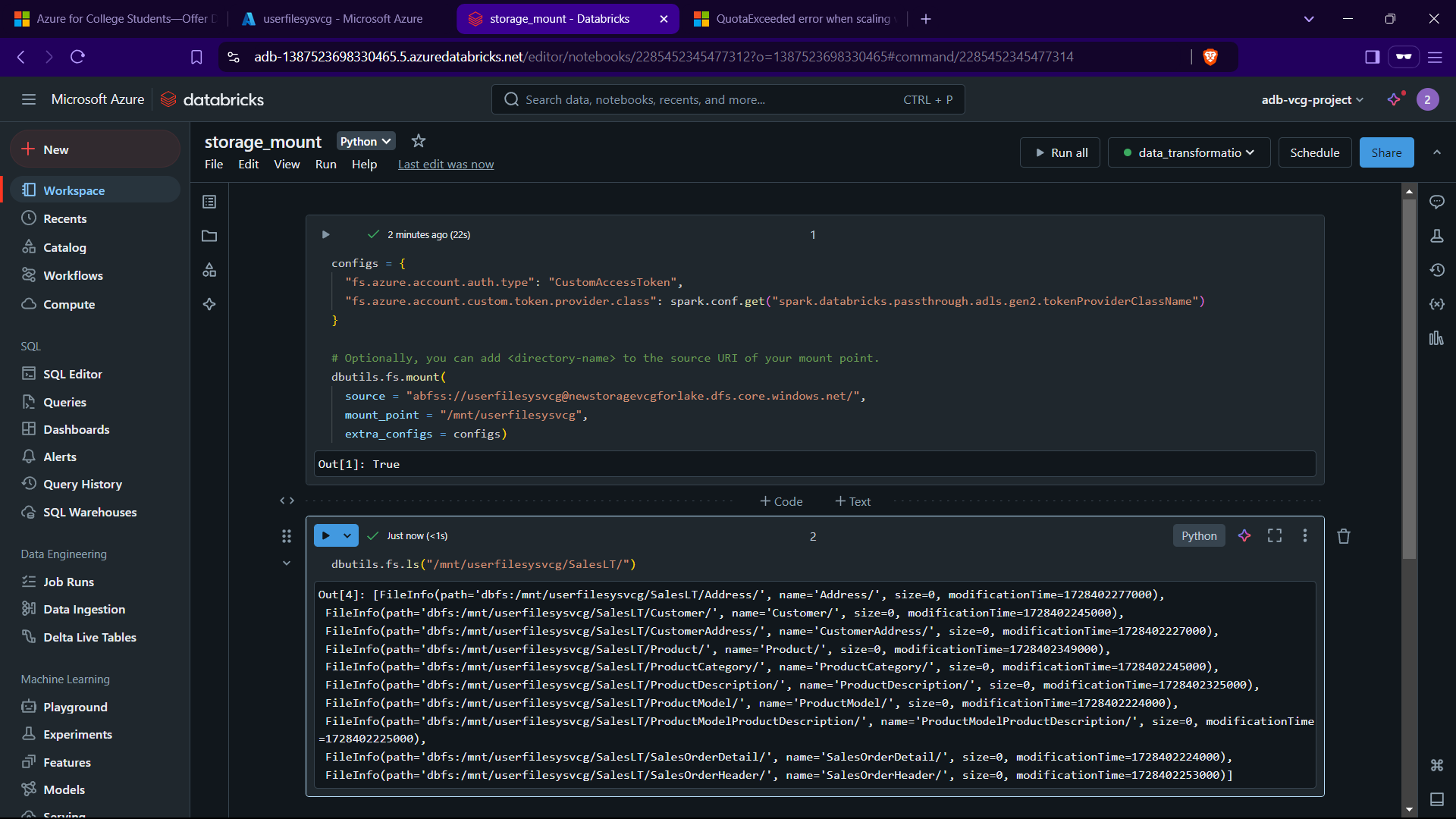
We create a cluster in databricks with enabling credentals in advance option



Created a notebook In workspace and update souce and mount\_point

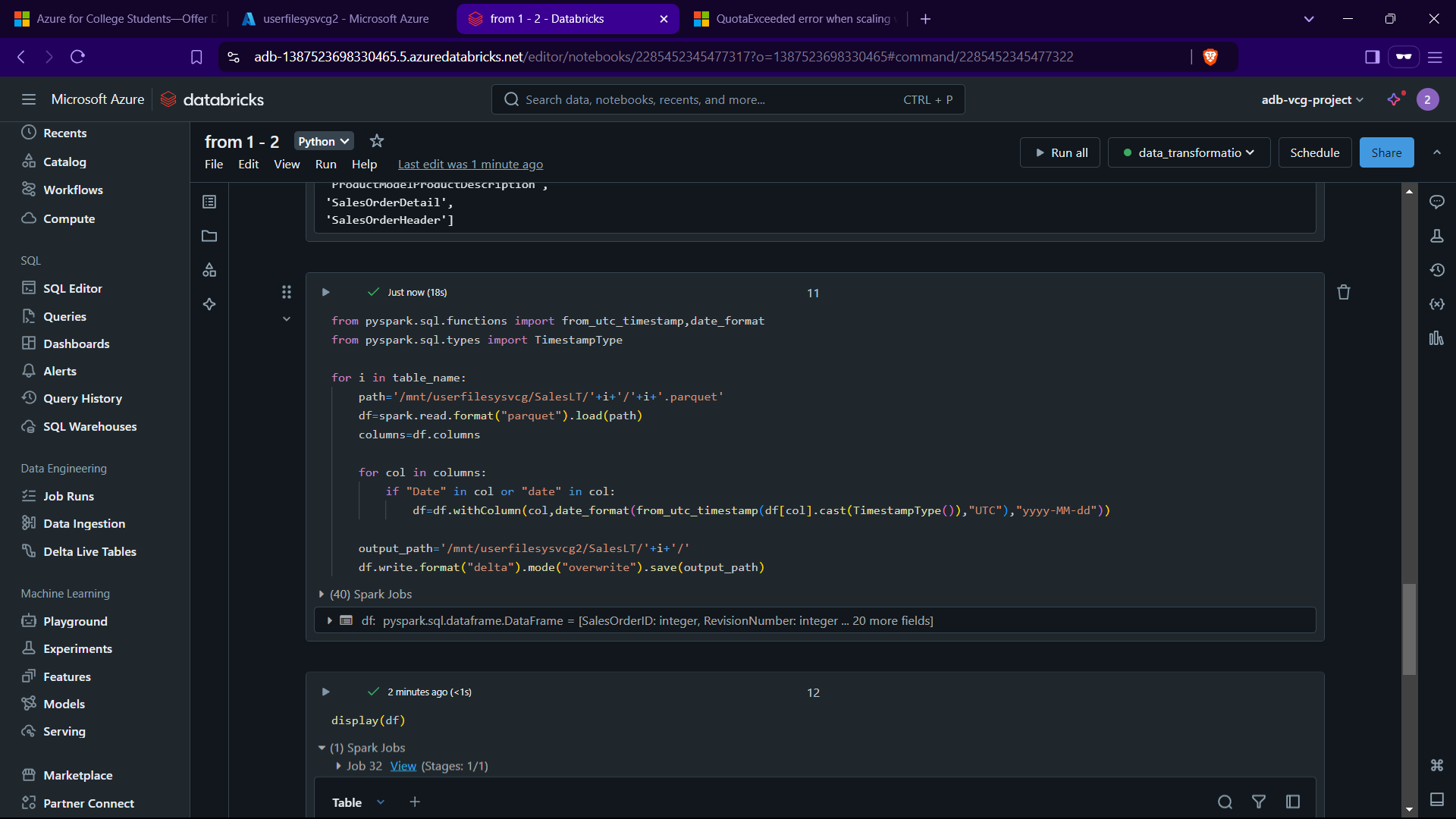


To see the files in the given path

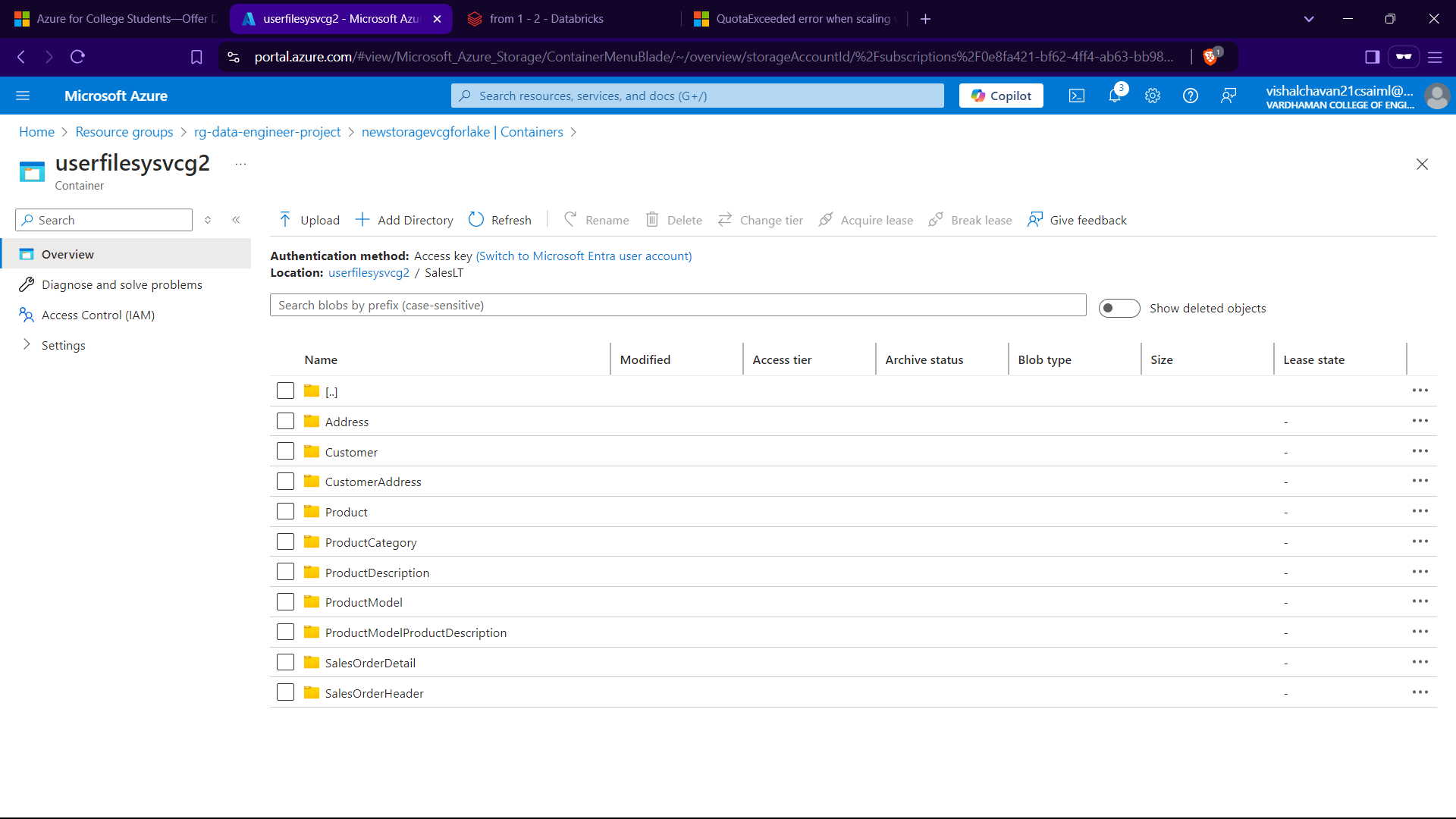


As for Data Transformation we can change try changing the colums data type in the data base in SSMS from datetime to data.

In these notebooks we can aslo use magic comands by % sql that notebook will be of py but with this it will also be able to run sql commands in the notebook as well

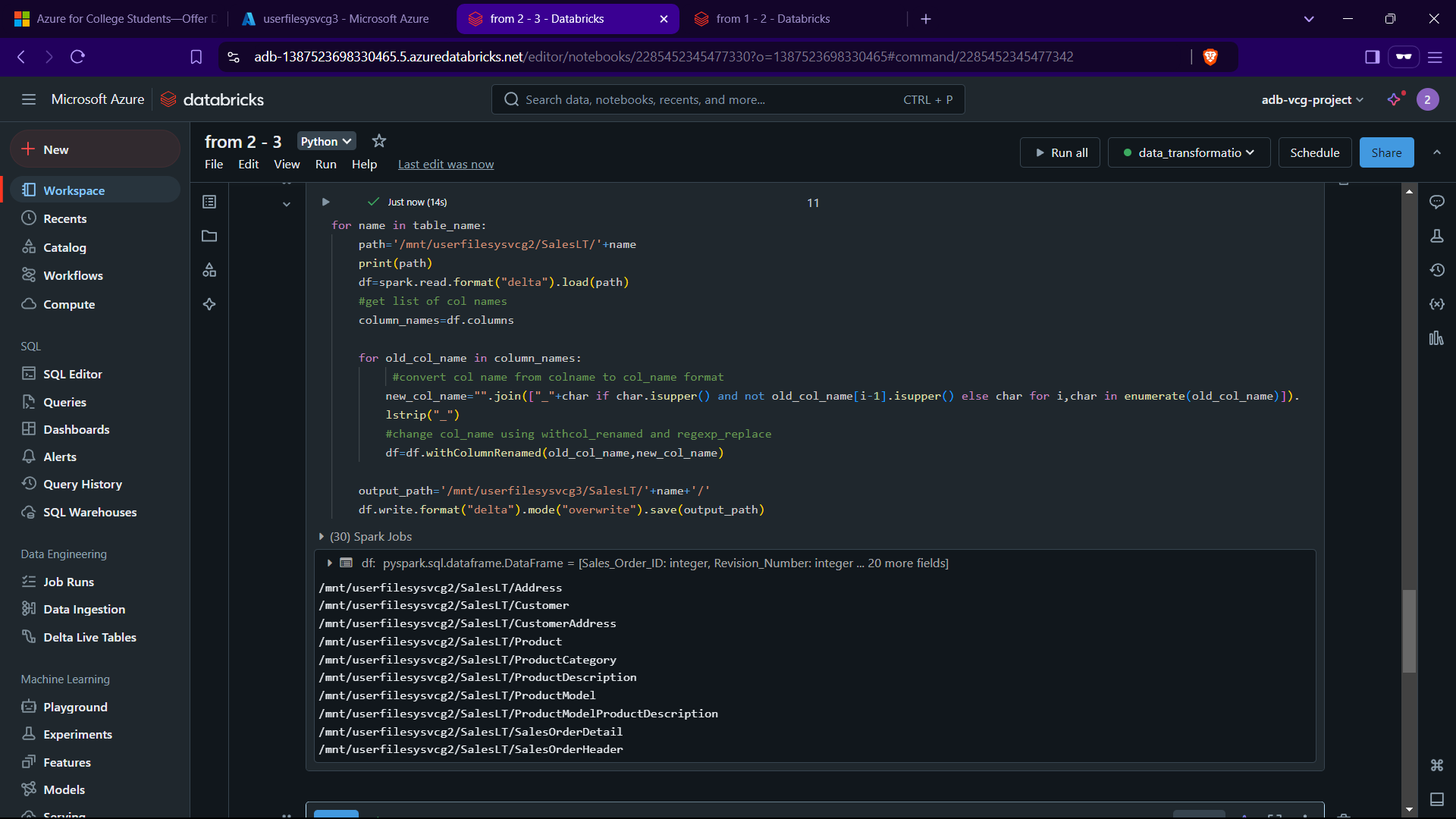


With this code we have successfully copied the entire files from ..vcg to vcg2 container/datalake using databrick

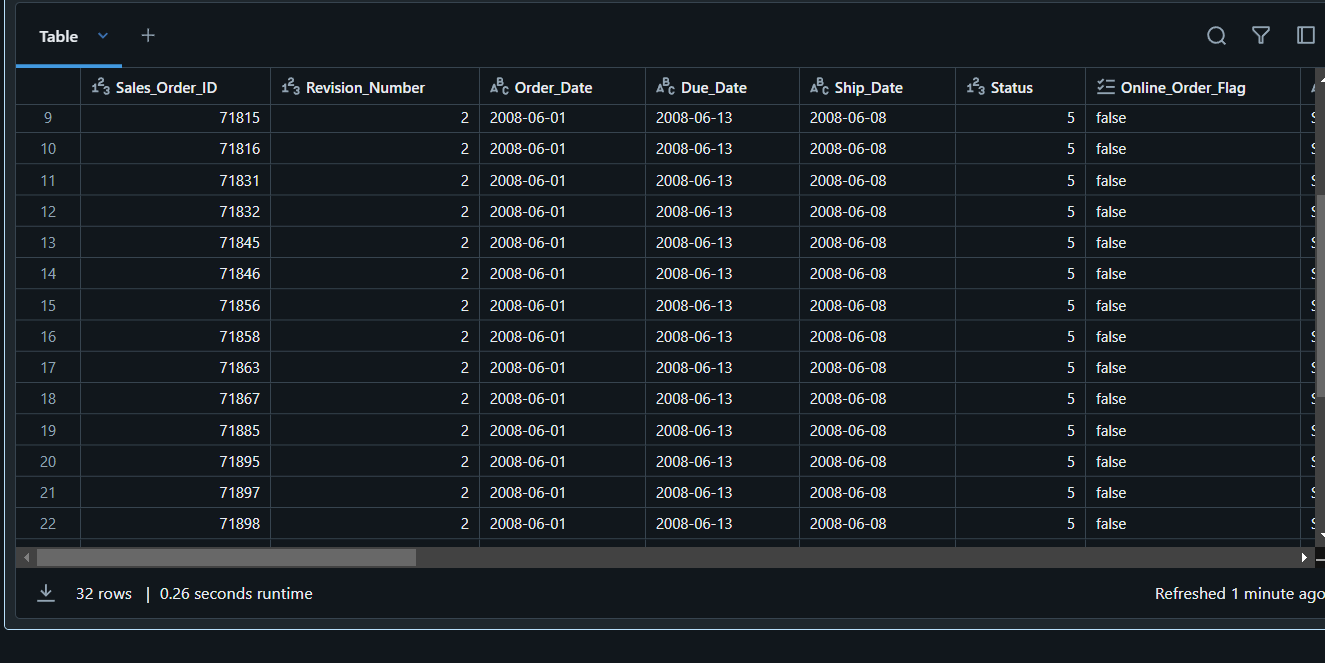


This all was Level 1 Tranformation ie Bronze -> Silver

Now we do from Silver->Gold here we join tables changes naming conf etc..

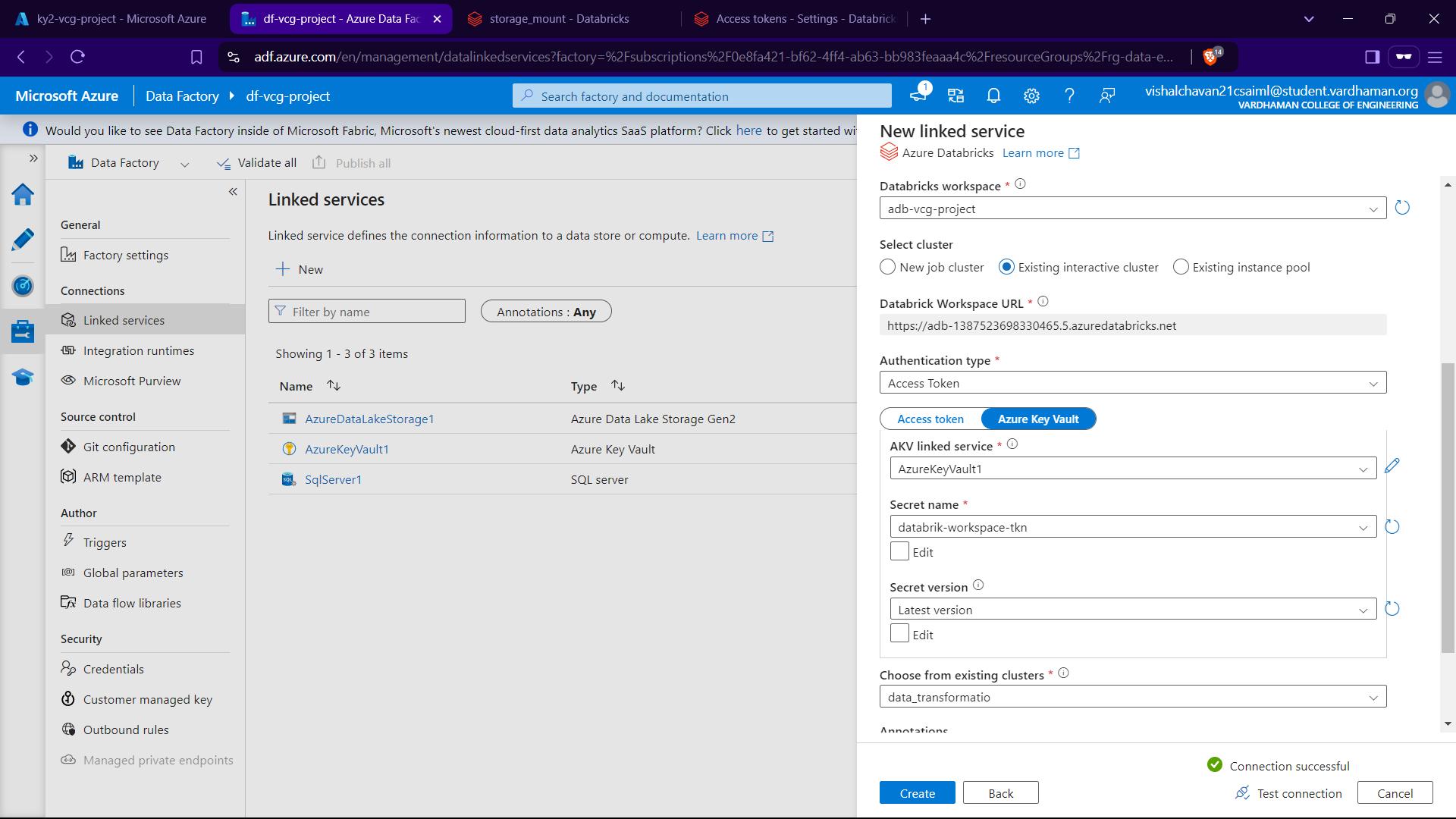


With this we have chaged the col names from vcg2 to vcg3 all names having ”\_” in b/w them.

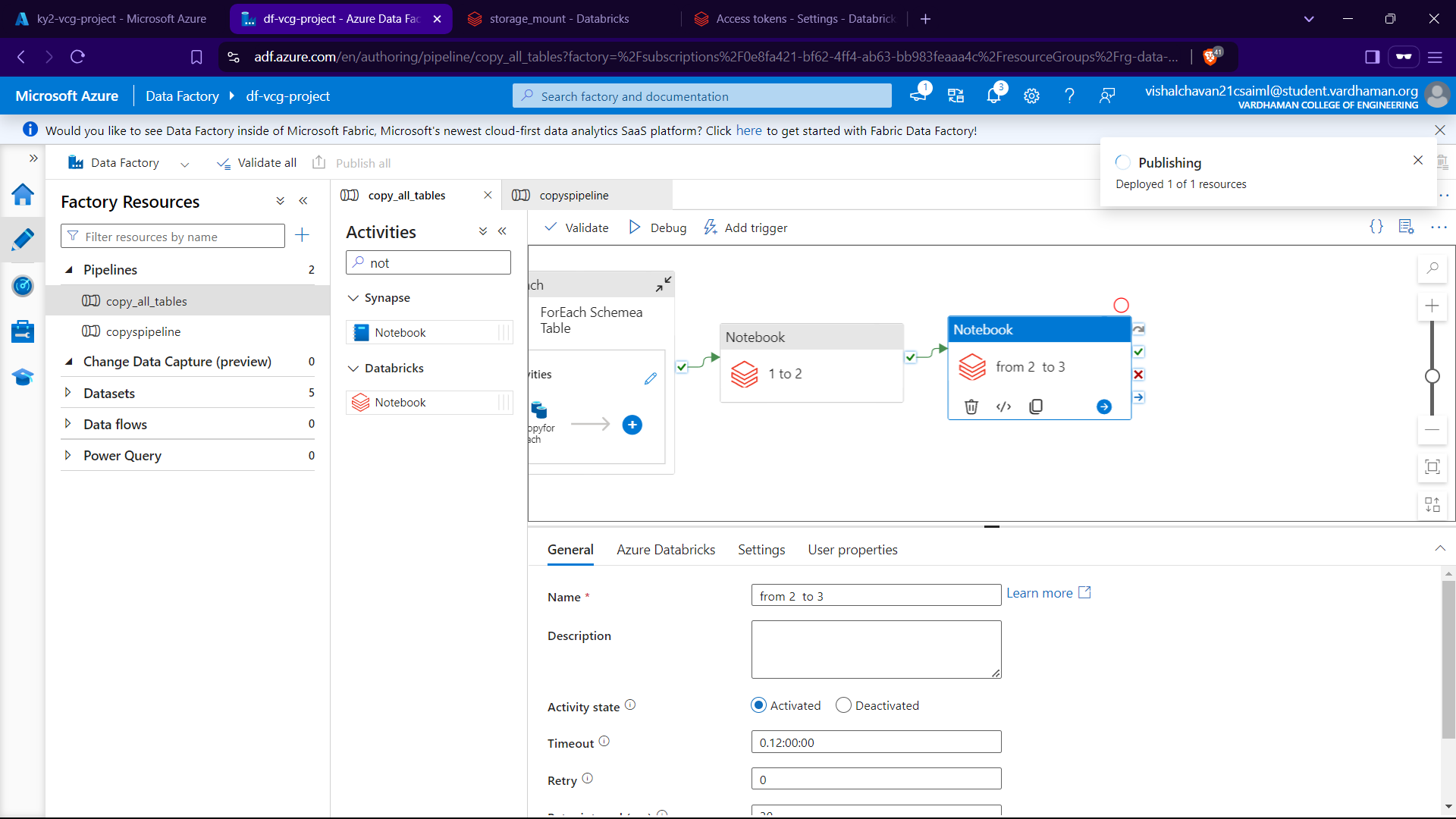


Now we go to ADF and make pipeline for level 1 – 2 and 2 – 3 tranformation

Create a linked service in ADF and go to compute->azuredatabricks



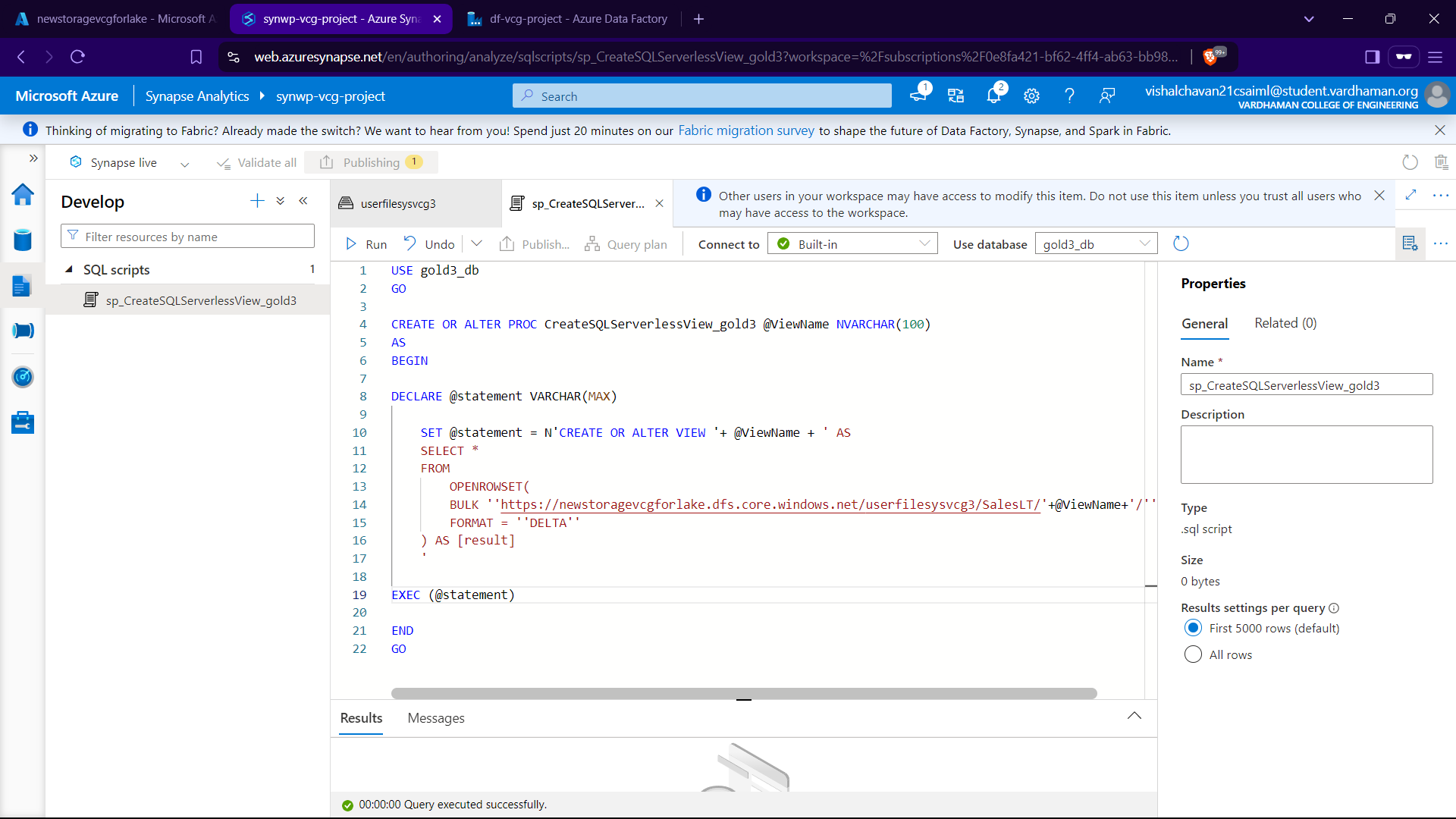
Now we create 2 nootebook with path given to specific notebooks from 1 – 2 and 2 – 3 and publishing the changes



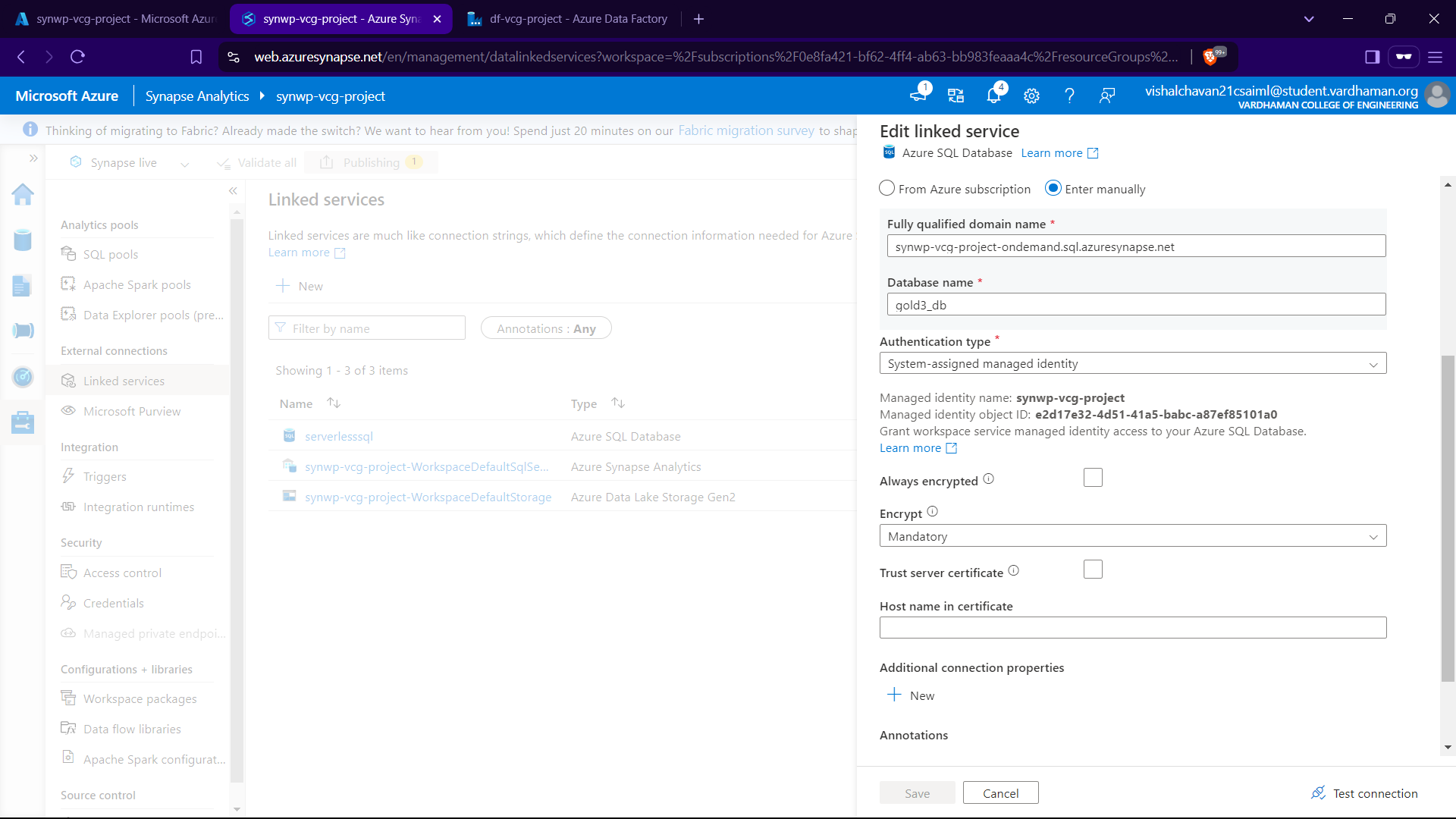
Section IV – Data Loading

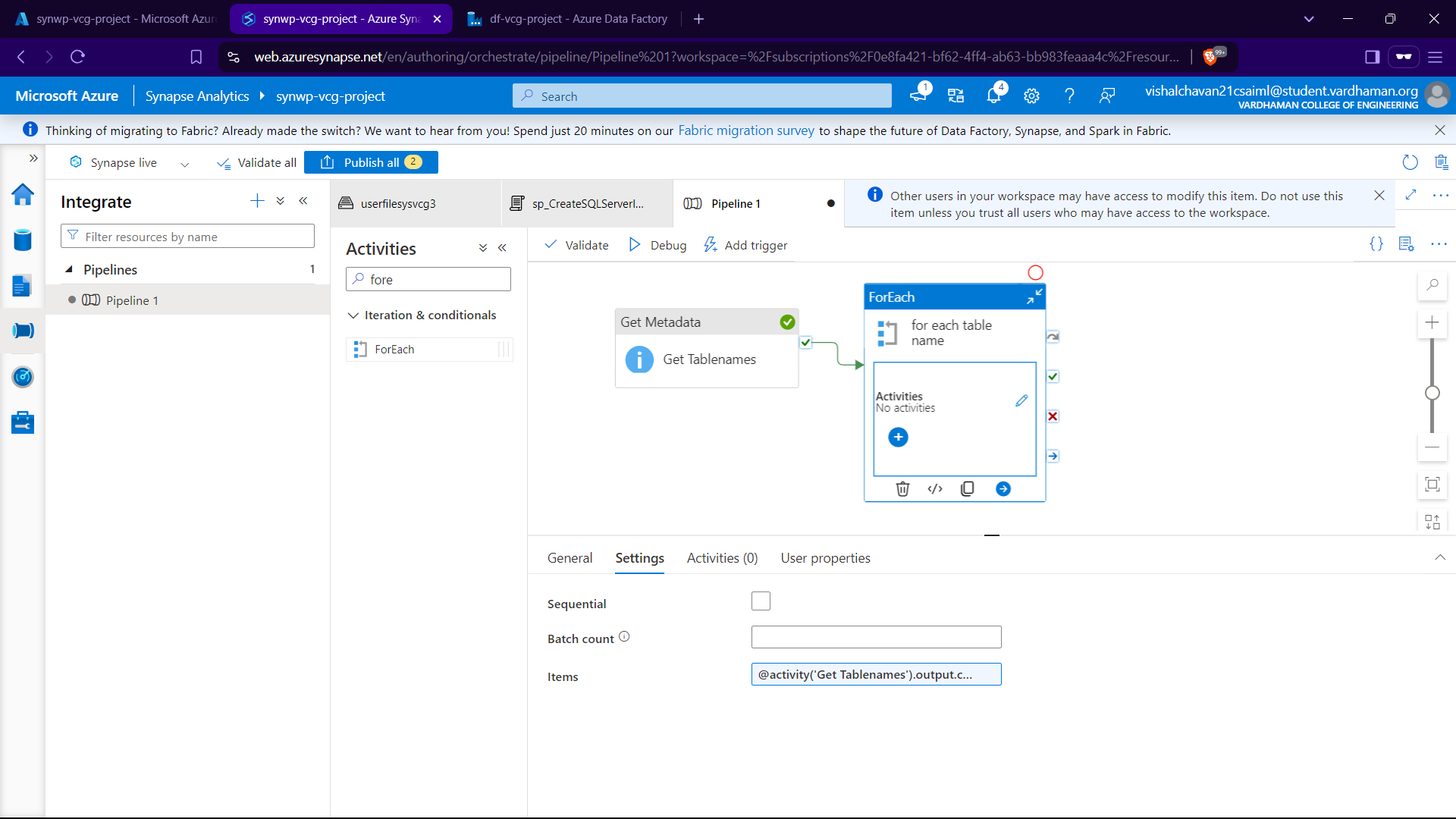
Now we have vlean/transformed data now we need to load in synapse analytics to load

Created a Database named gold3\_db and made a sql query to show the rows in the dataset in our device

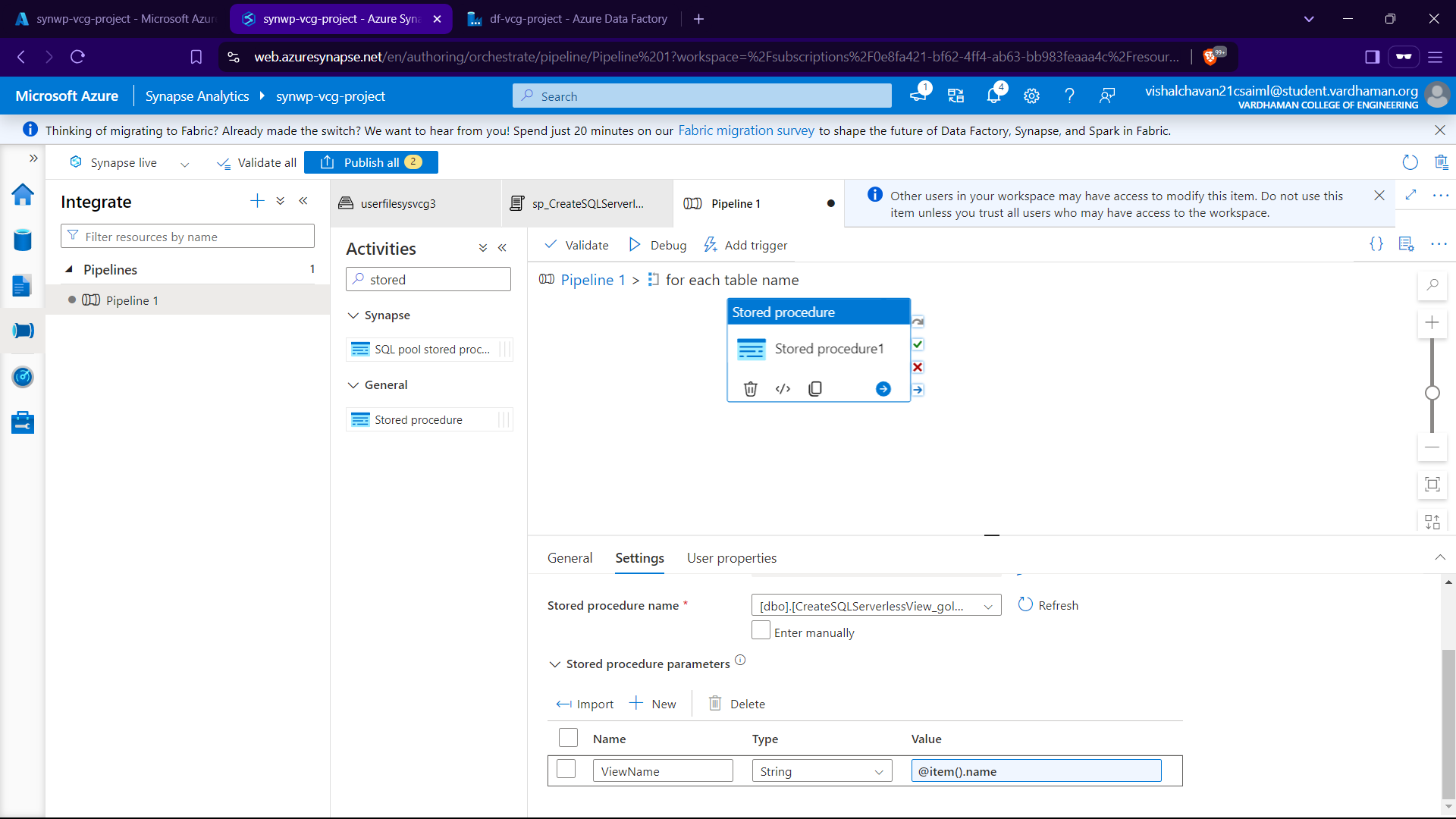


then we made a linked service

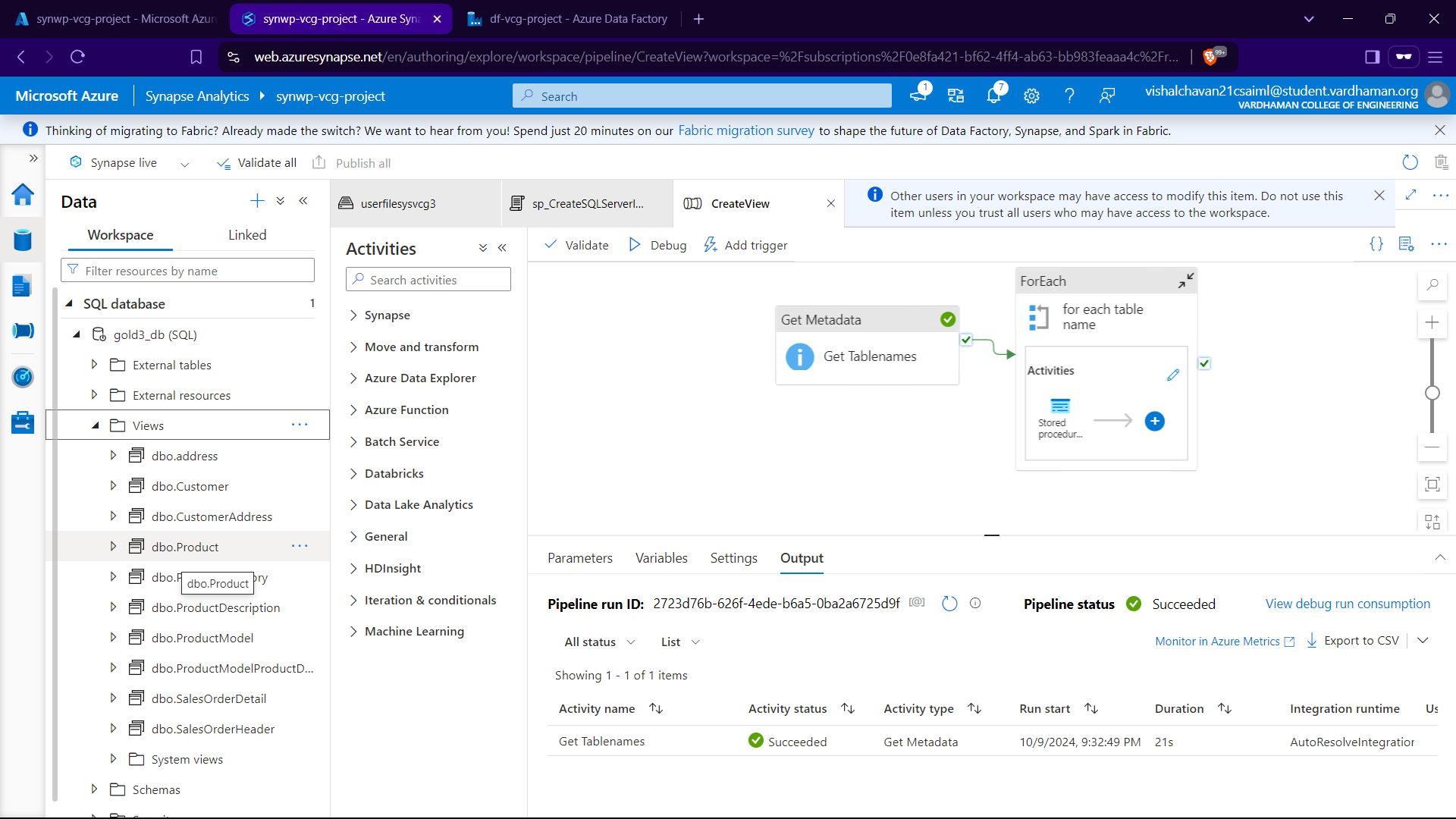


Then we make a pipe line where metadata gives tablenames as in put for each activity

Inside activity for ‘for each’ it gets the tablenames Views



Then after runing the pipieline when we refresh the database in view section the view of the top rows are now avaliable for us to view.for which we can create a new sql script to view.



Section V – Data Reporting

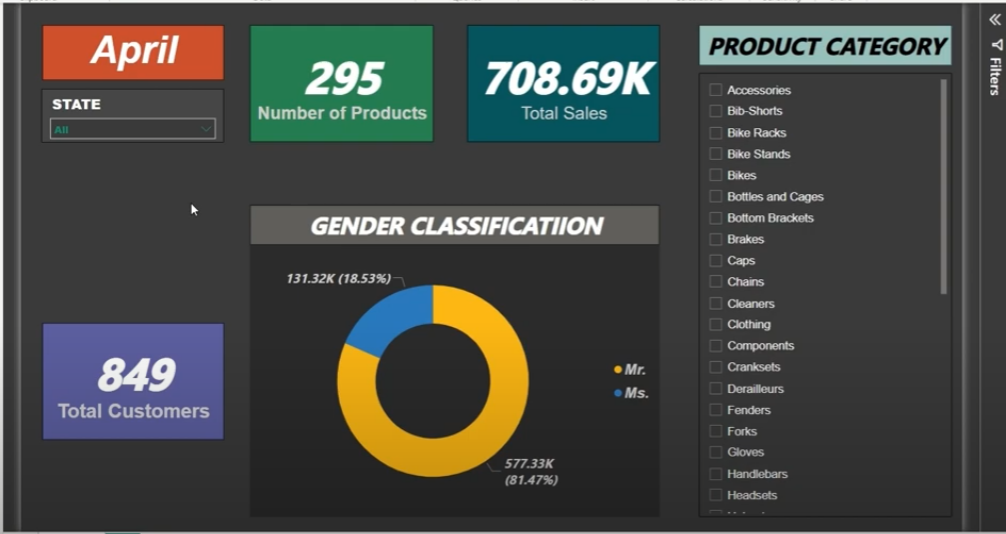
Now we use Power BI to fetch the views of the database.

Go to GetData->More->Azure->synaptic analytics->connect

Give the serverless endpoint URL to the pop up window and click ok then go to microsoft account as secuity and connect.

Select all tables and load

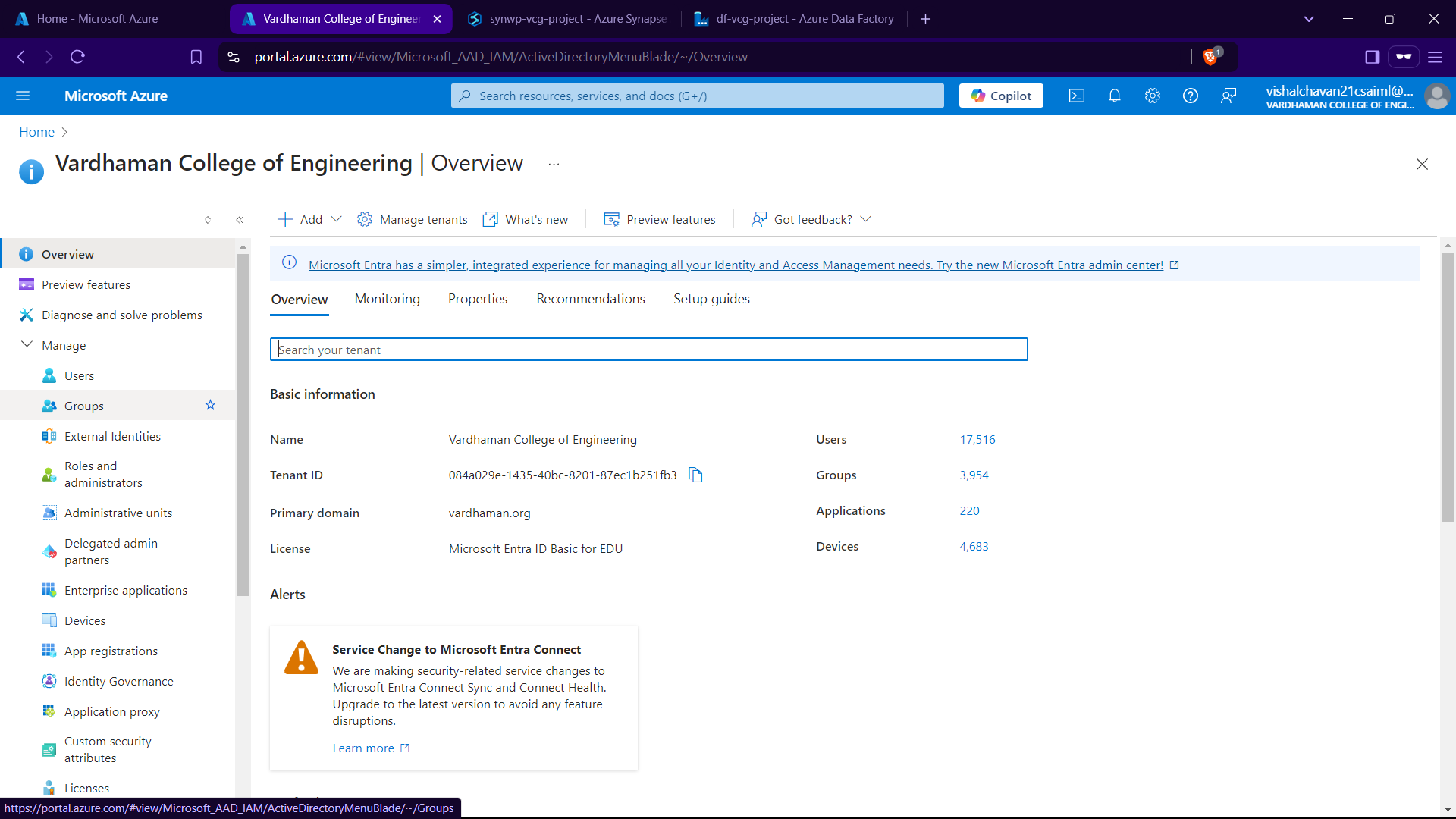
In Model View Section by Manage Relationship we can mangae them accordingly. and proceed with the reporting



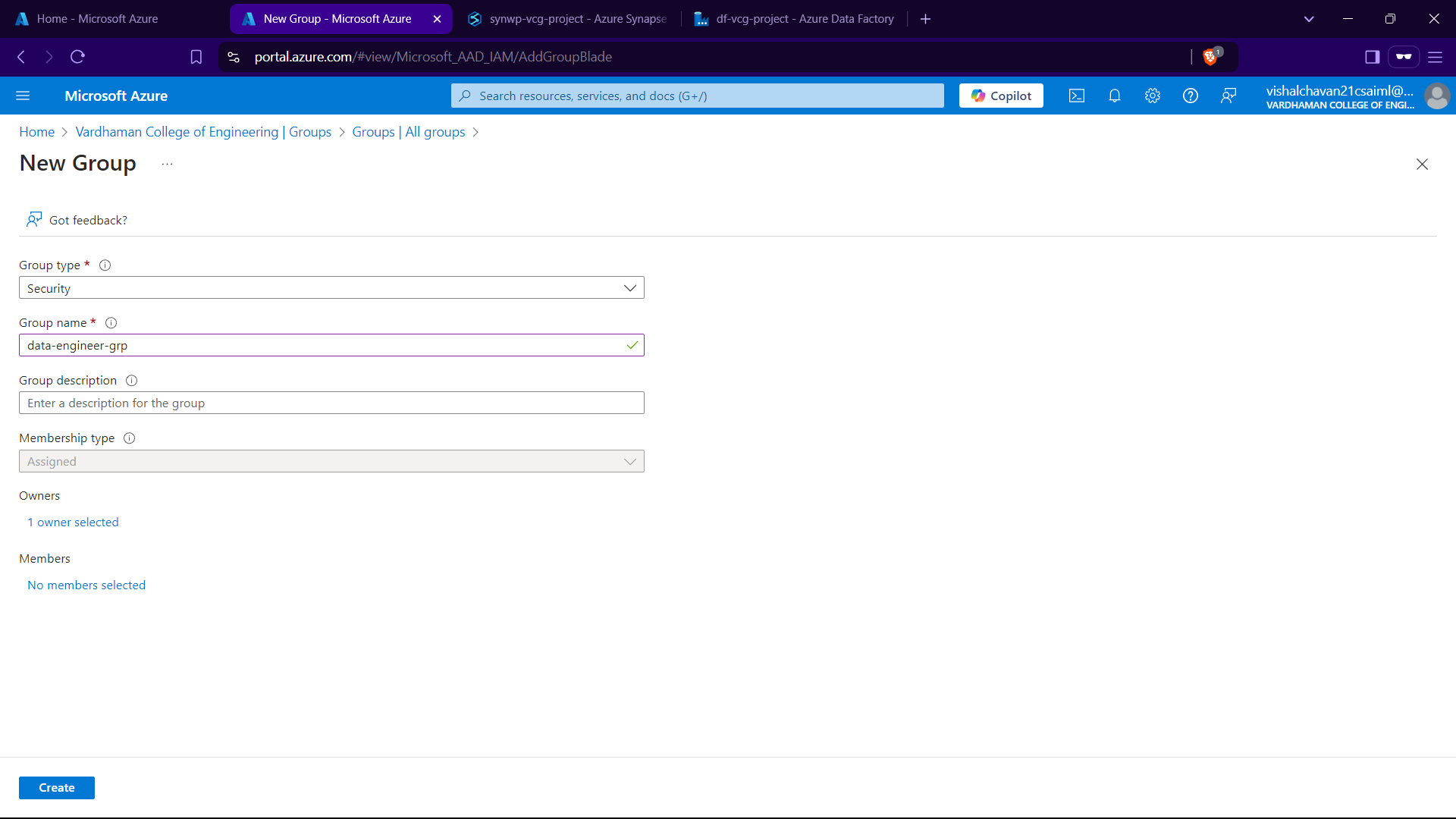
Section VI - Security and Governance (AAD)

As I am the sole owner of this only I as a user can access it but to add someone else is really annoying like going to resource grp and giving access control etc we can make security grp and anyone on security grp can use the resource.

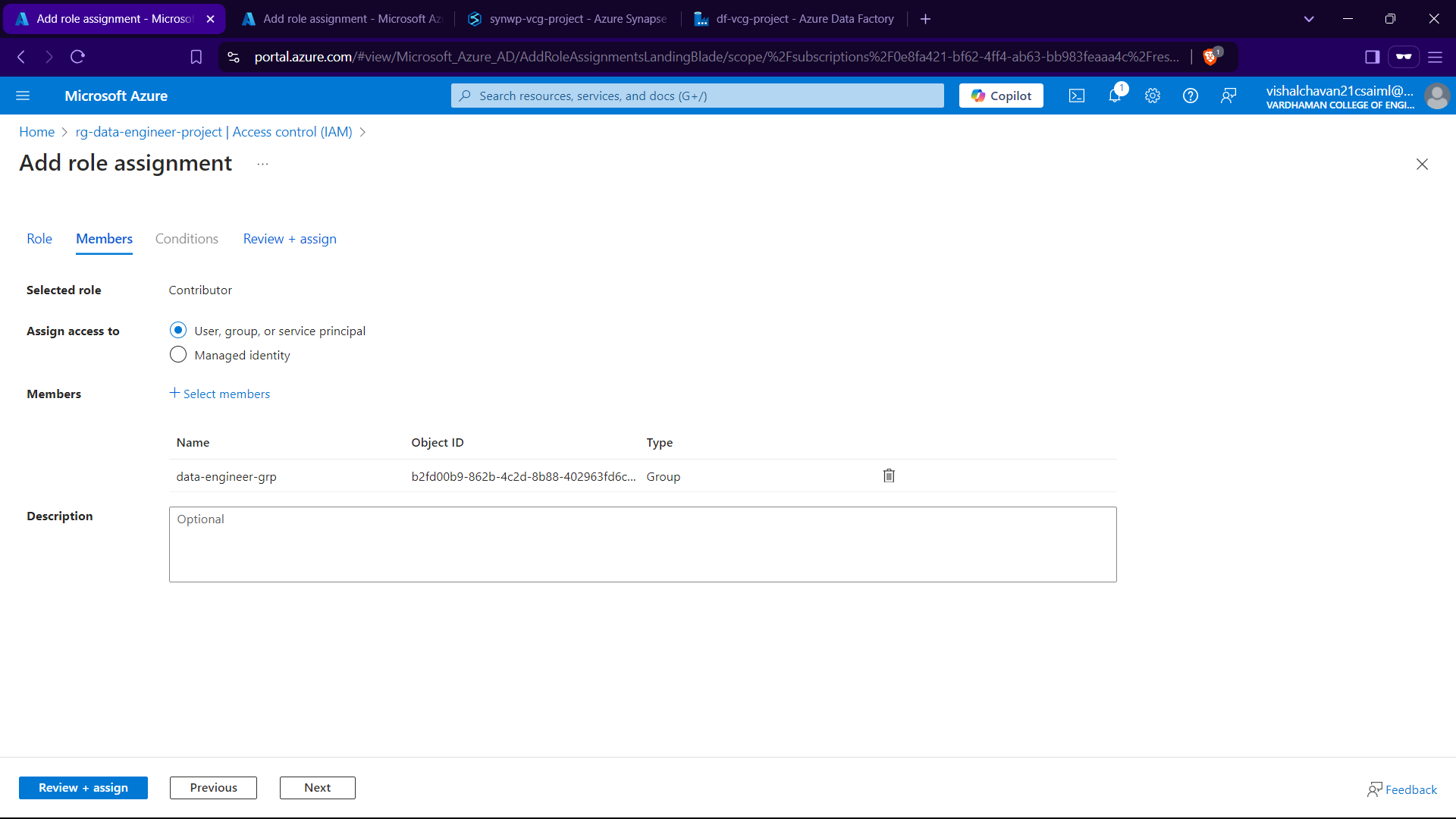
Go to Microsoft Entra ID and in groups



Give the grp owner and members I have only given owner as myself.



Go to role assignments in resource grp- privilegd roles->selct members and add grp name

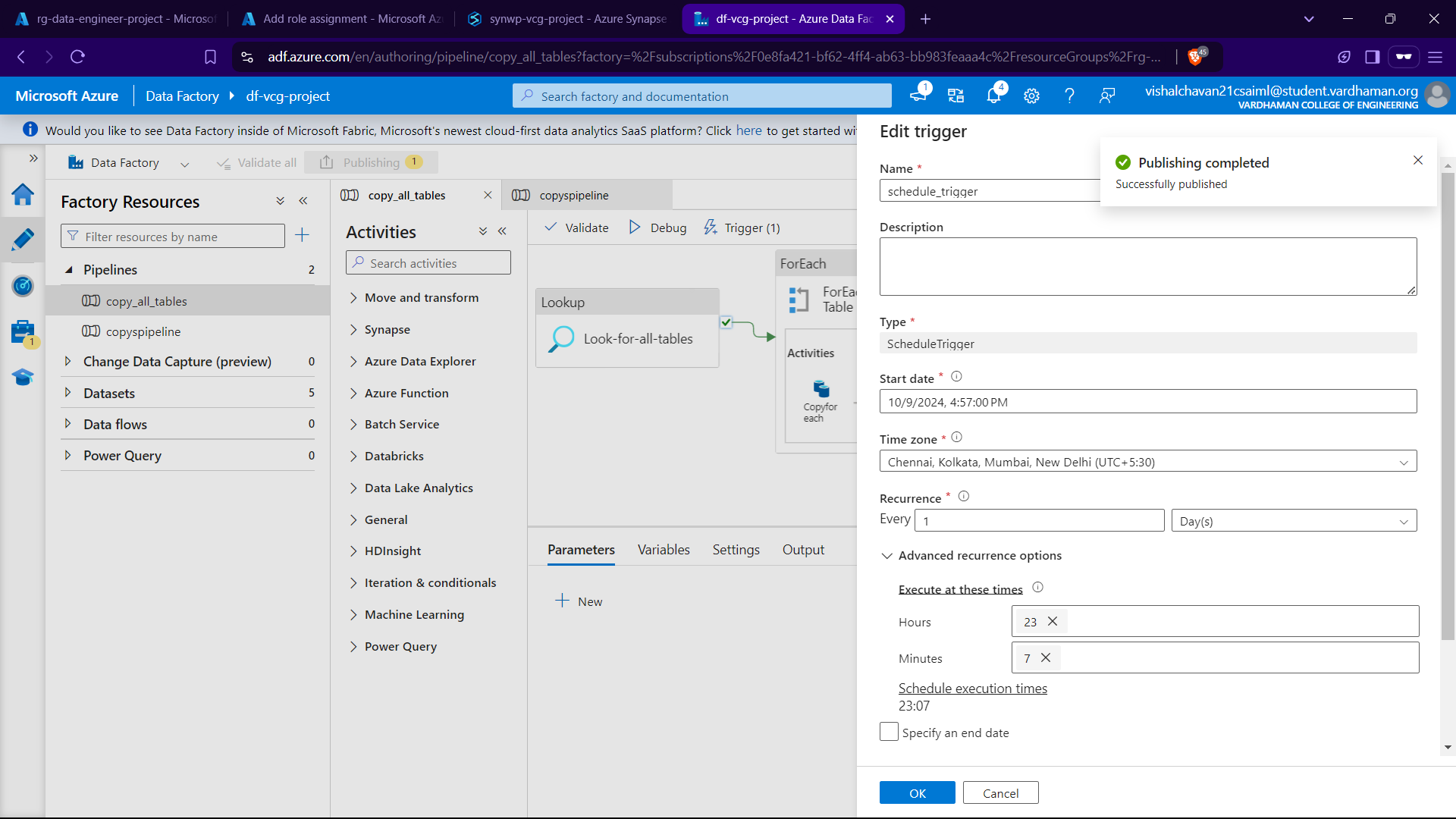


When we give access to any other accound and in that account we can go to resource grp there and when we refresh the page they have been granted the resource grp which we used.

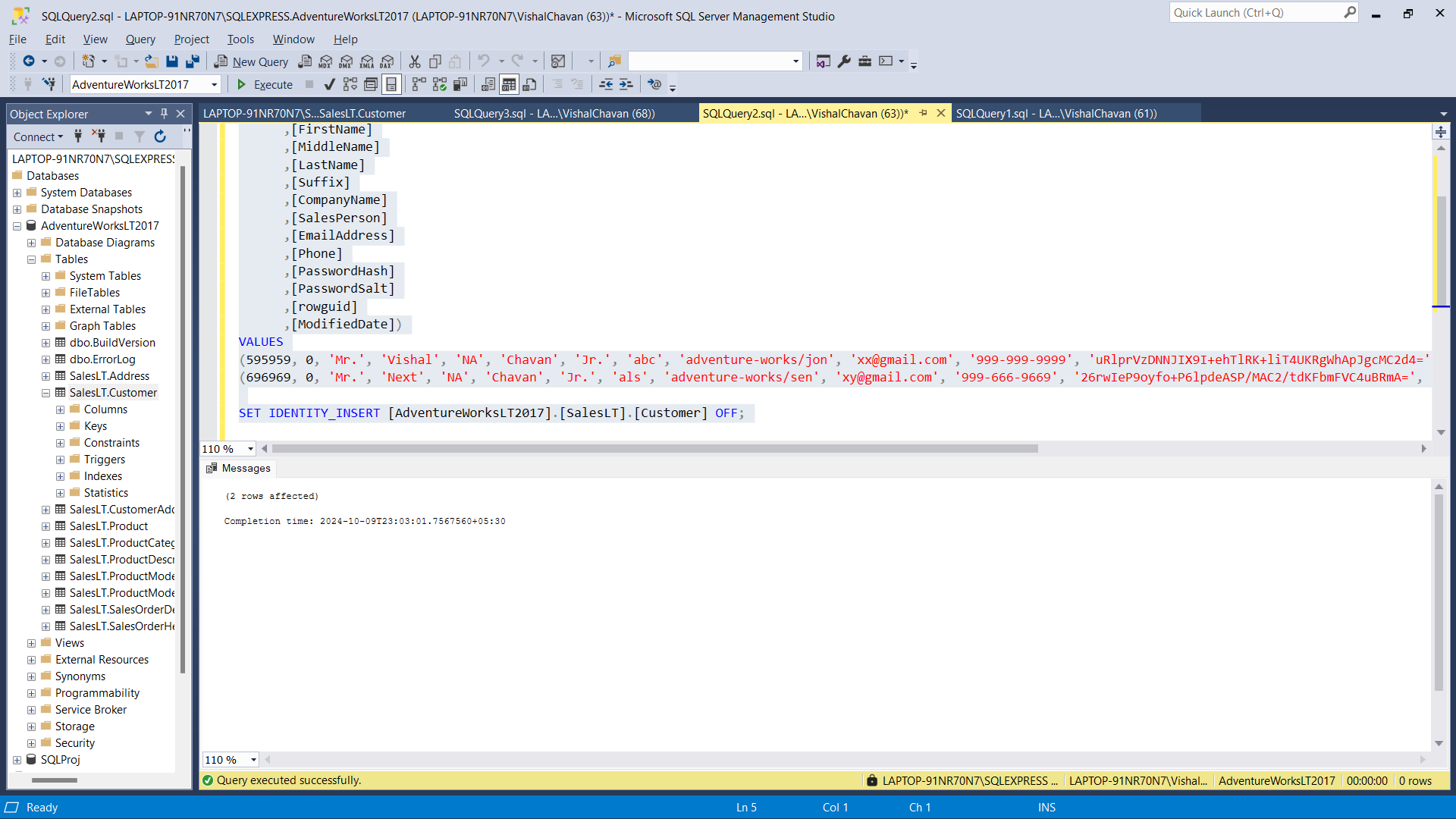
Section VII – Pipeline Testing

Here when we add a row in our premise database it should be replected in the report.

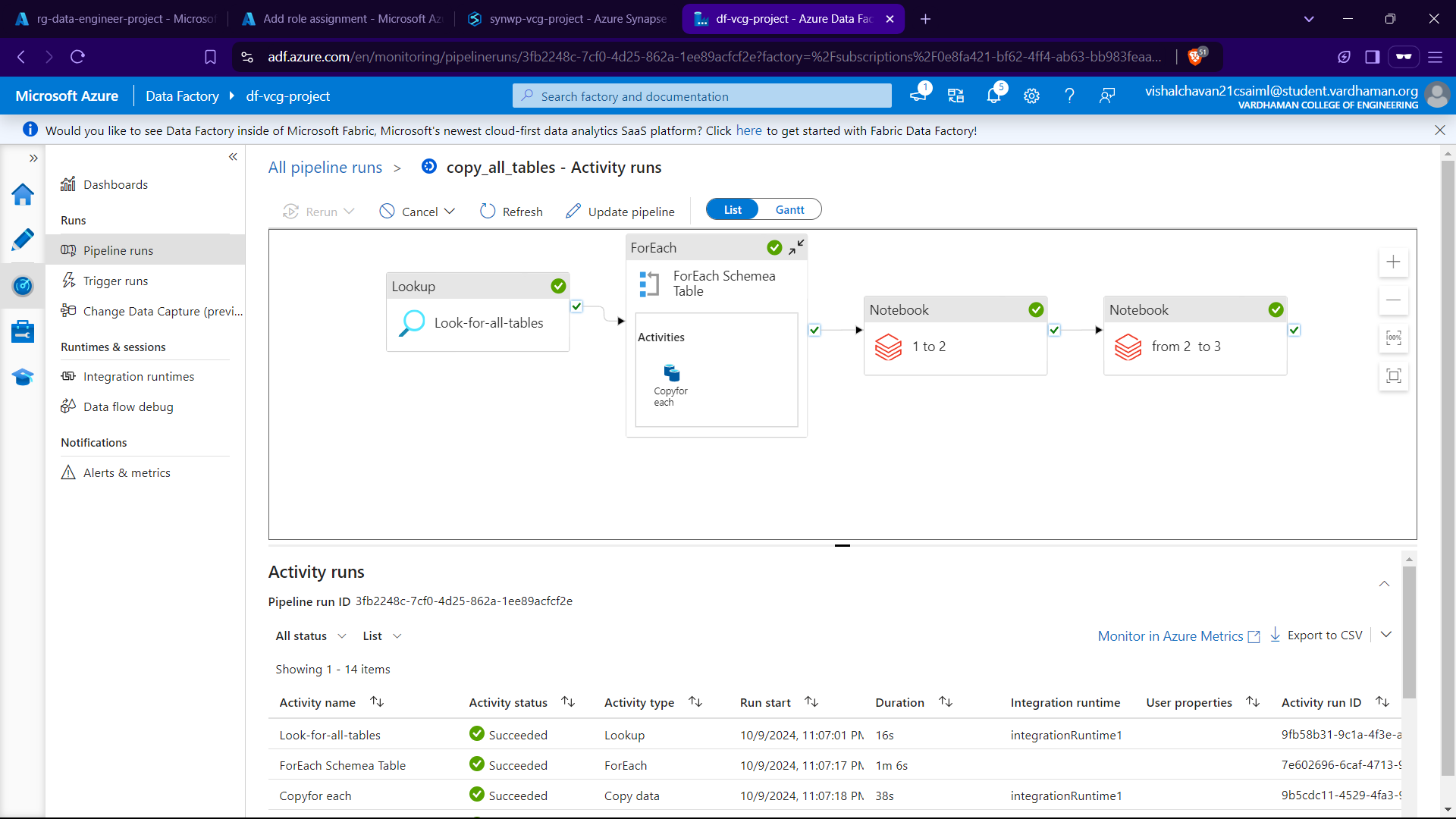
We can do this by creating triggers from ADF



Now before the ‘trigger’ triggers we update the Customer Table by inserting 2 new rows in to it.



Then after sucessfully runnning sql we then monitor the pipeline as it gets triggered at the specified time.



Now we can see that the no of customers have increased by two in the Power BI dashboard.

