AZURE DATA FACTORY

CAPSTONE – COVID USE CASE

Name: Majji Vijay Vamsi (Contractor)

Employee ID: 2320213

Cohort ID: CSDAIA24AZ005

Introduction:

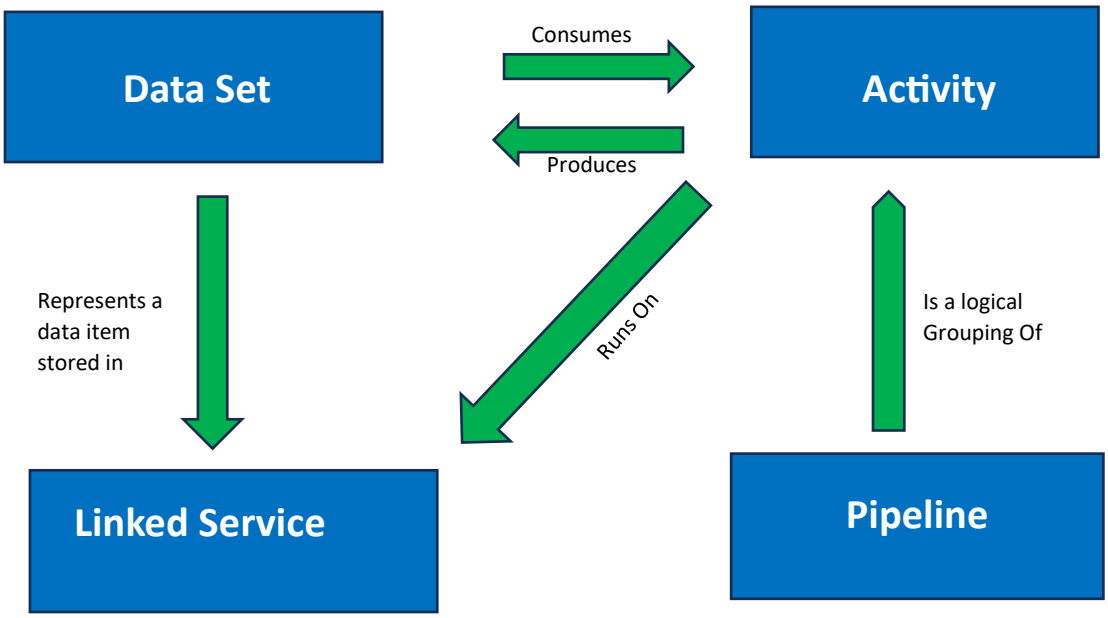
The Azure Tools which have been used for this project are Azure Data Factory, Azure Synapse Analytics and Azure Data Lake Storage " Azure Data Factory Capstone Covid Case Analysis". These components helps to extract data from its source storage (Data Lake), datasets are transformed in Data factory for analyzing the trends and insights which are in accordance with business needs, later when the interpretations are made then it is loaded to destination data warehouse. The "Azure Data Factory Capstone - Covid Use Case " system was created to learn how to build a real-world data pipeline in Azure Data Factory (ADF) to analyze the covid trend across the regions using Azure cloud data services. By performing this case study, we will learn.

- Ingestion of data from flat files into Azure Data Lake Gen2 and Azure Synapse using Azure Data Factory (ADF).
- Transforming the data using Data Flows in Azure Data Factory (ADF) and load into Azure Synapse.

Objective:

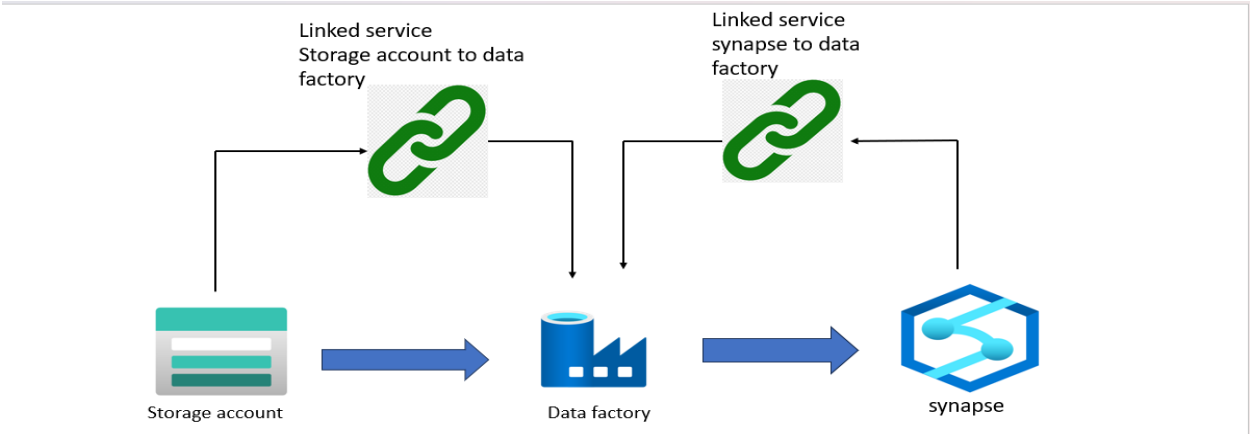
The main aim of this project (Covid Use Case Exercise) is that we will be having a hands-on experience on Storage, ADF Pipeline, Mapping Dataflow, Azure Synapse along with getting to know how to ingest data from flat files into Azure Data Lake Gen2 and Azure Synapse using Azure Data Factory (ADF) and also knowing how to transform data using Data Flows in Azure Data Factory (ADF) and load into Azure data lake. This report gives a summary of the entire project making us realize and interpret the use case scenario of Azure and its applications.

Overview of Data Factory Flow:

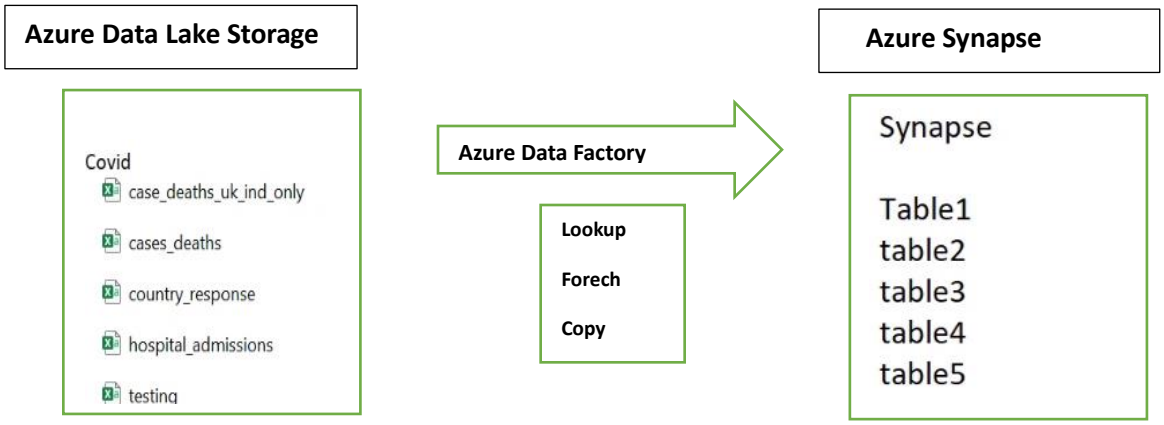


Project Requirements:

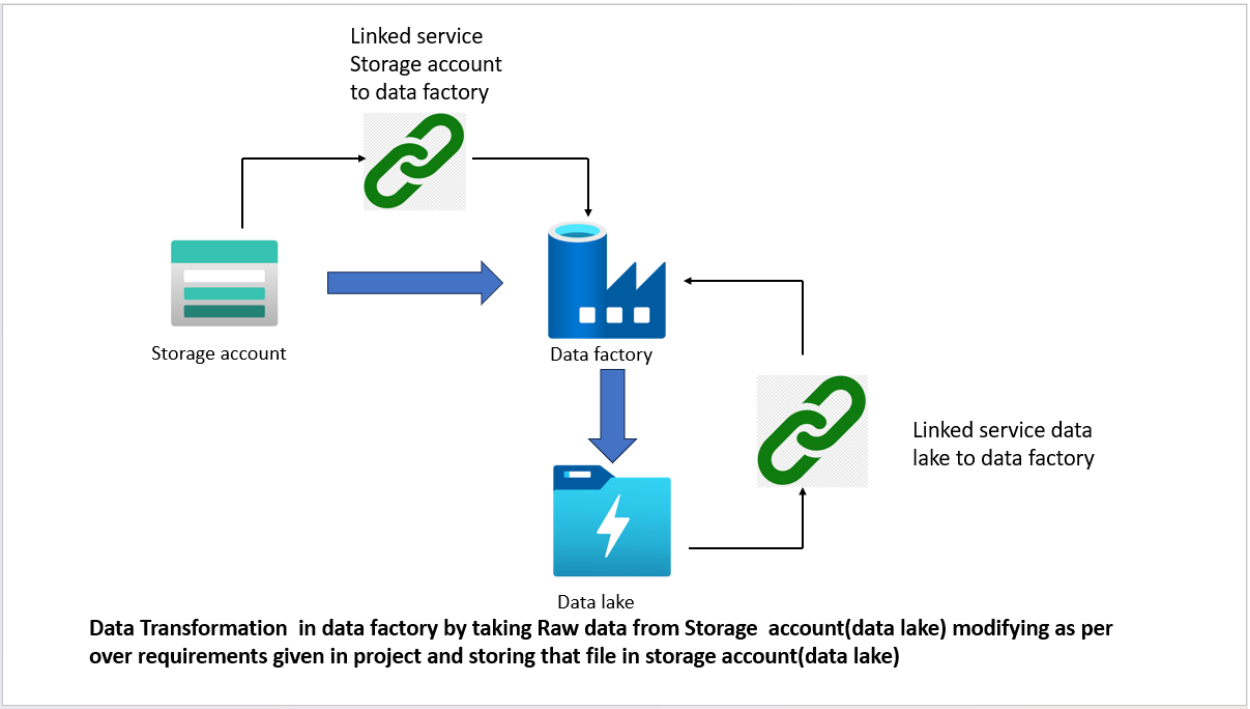
Requirement 1: Ingest raw flat files from data lake to synapse (data warehouse)

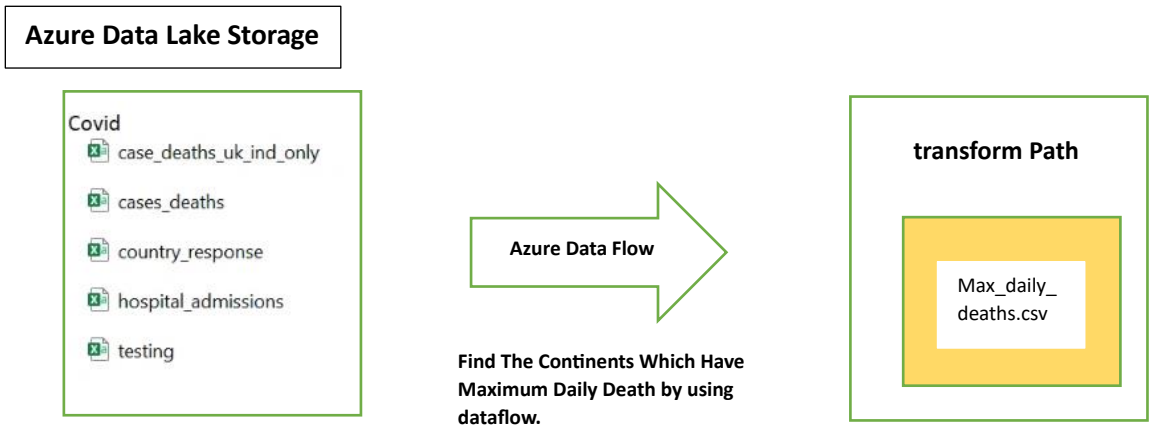


Ingesting process taken place in data factory by taking raw data from storage account(data lake) to relation database by creating tables in synapse(data warehouse)



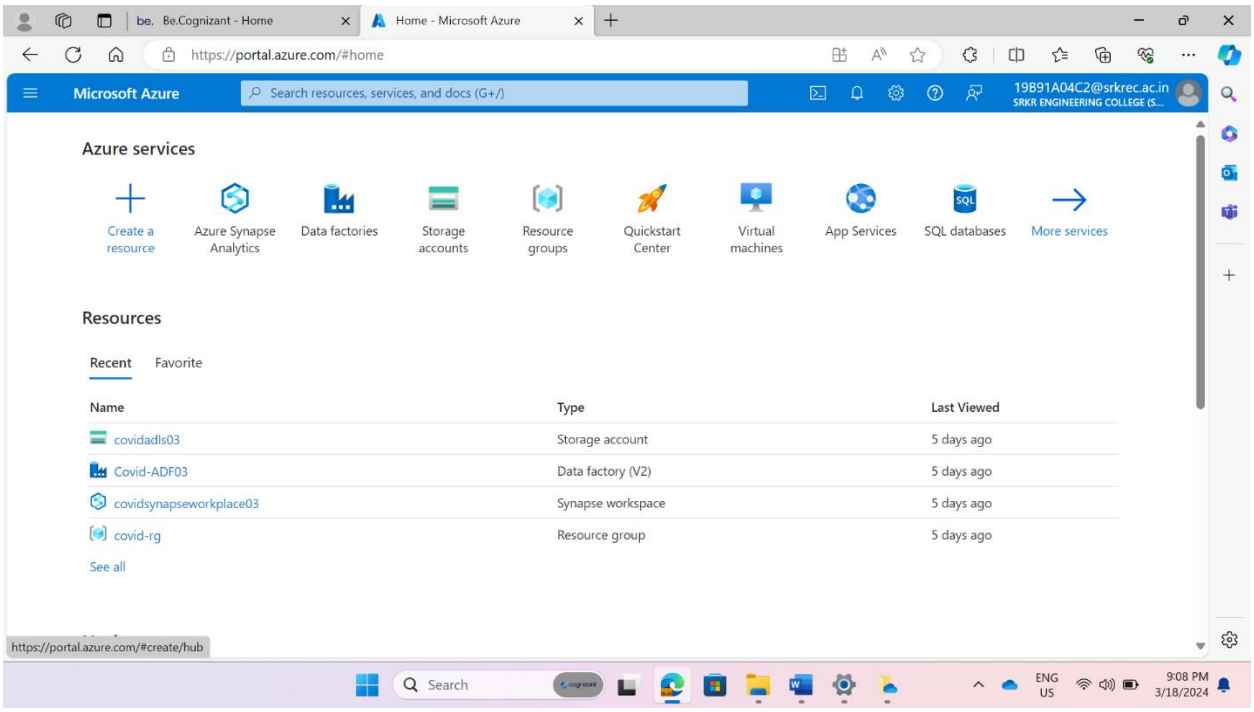
Requirement 2: Transform data using Data Flows in Azure Data Factory (ADF)



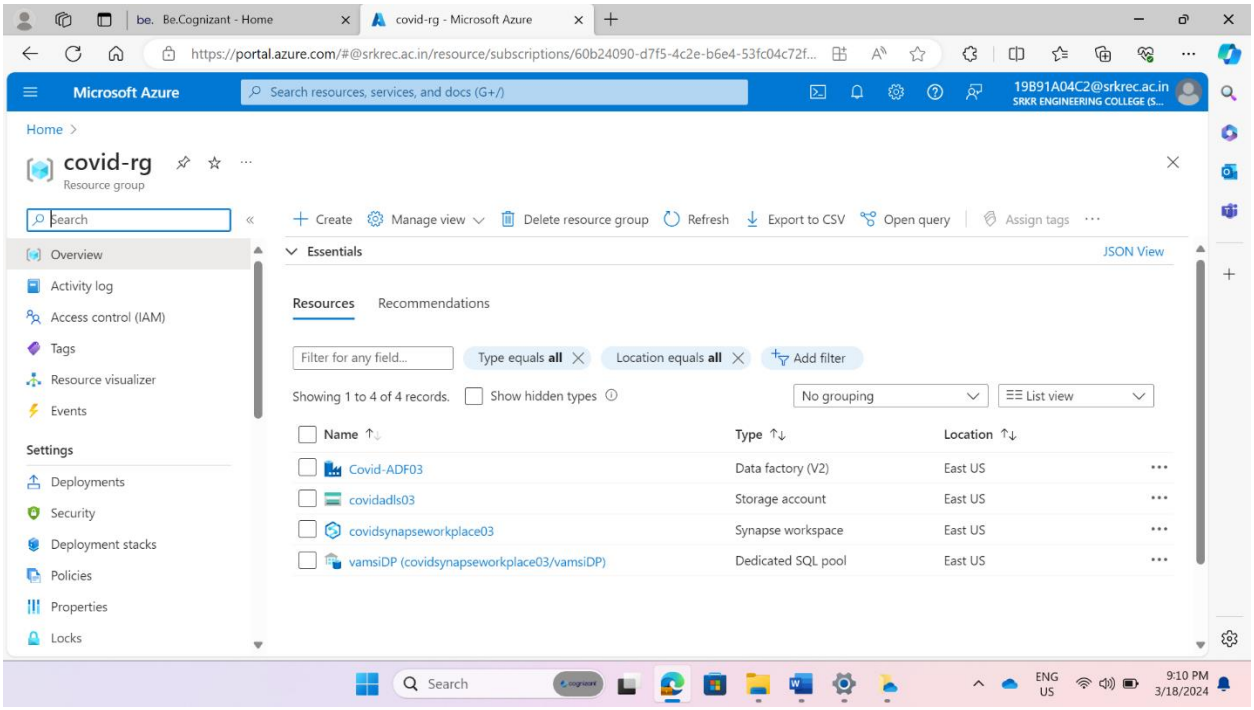


Procedure for Requirement 1:

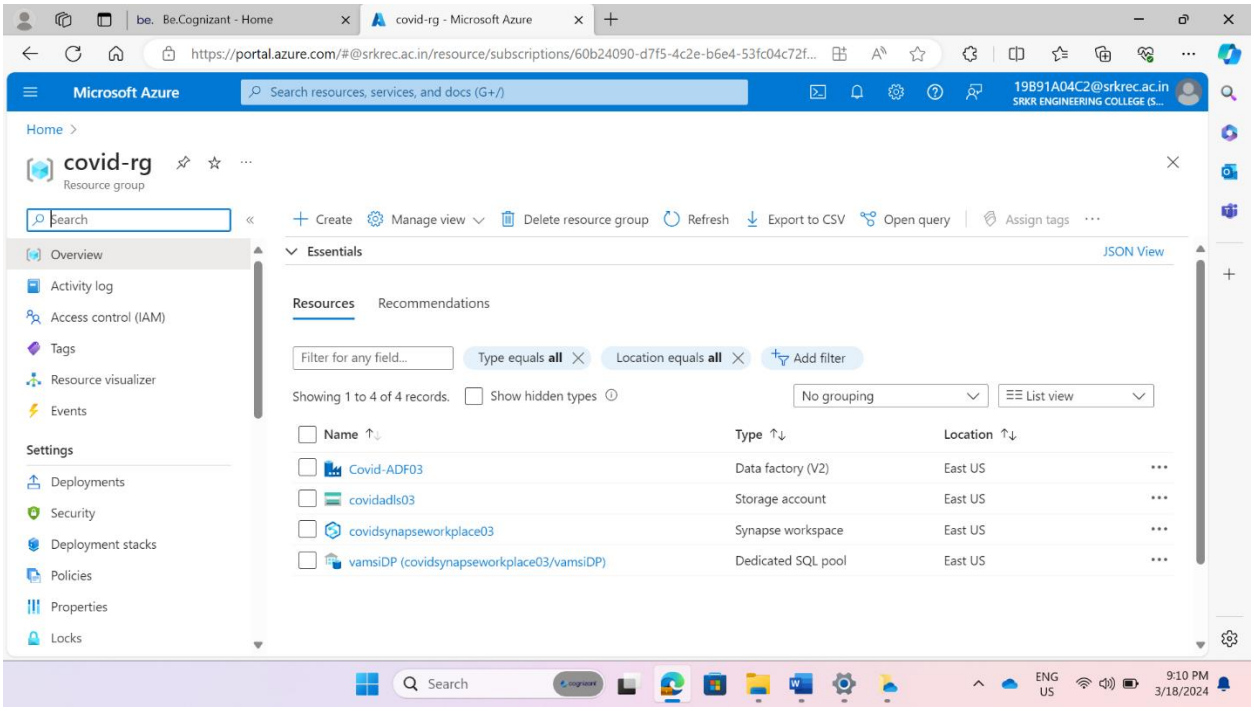
STEP 1: Open your Azure Account.



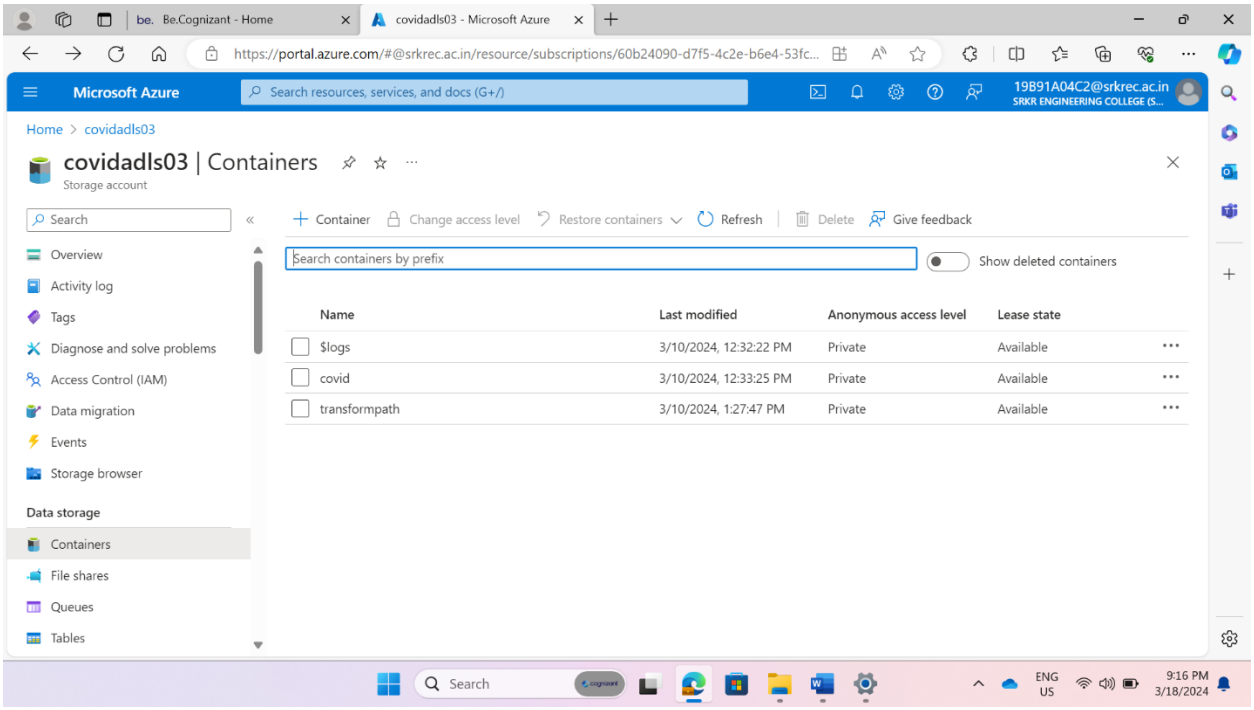
STEP 2: Create your own Resource Group.

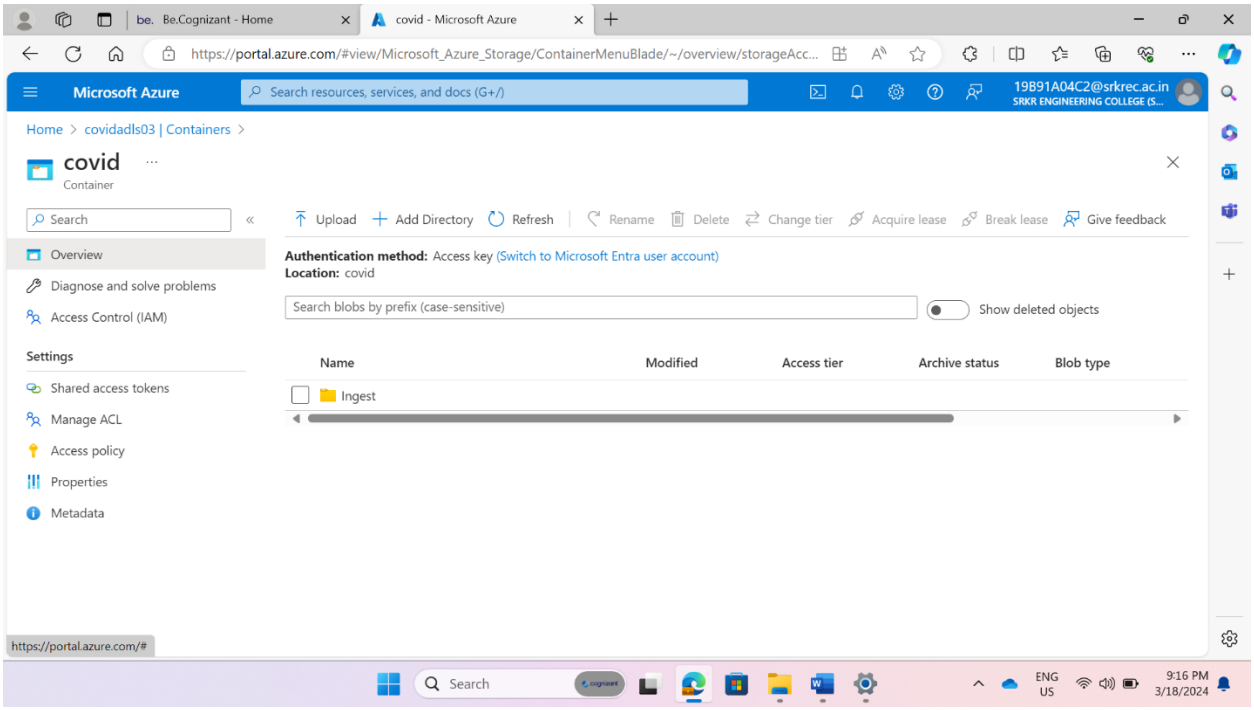


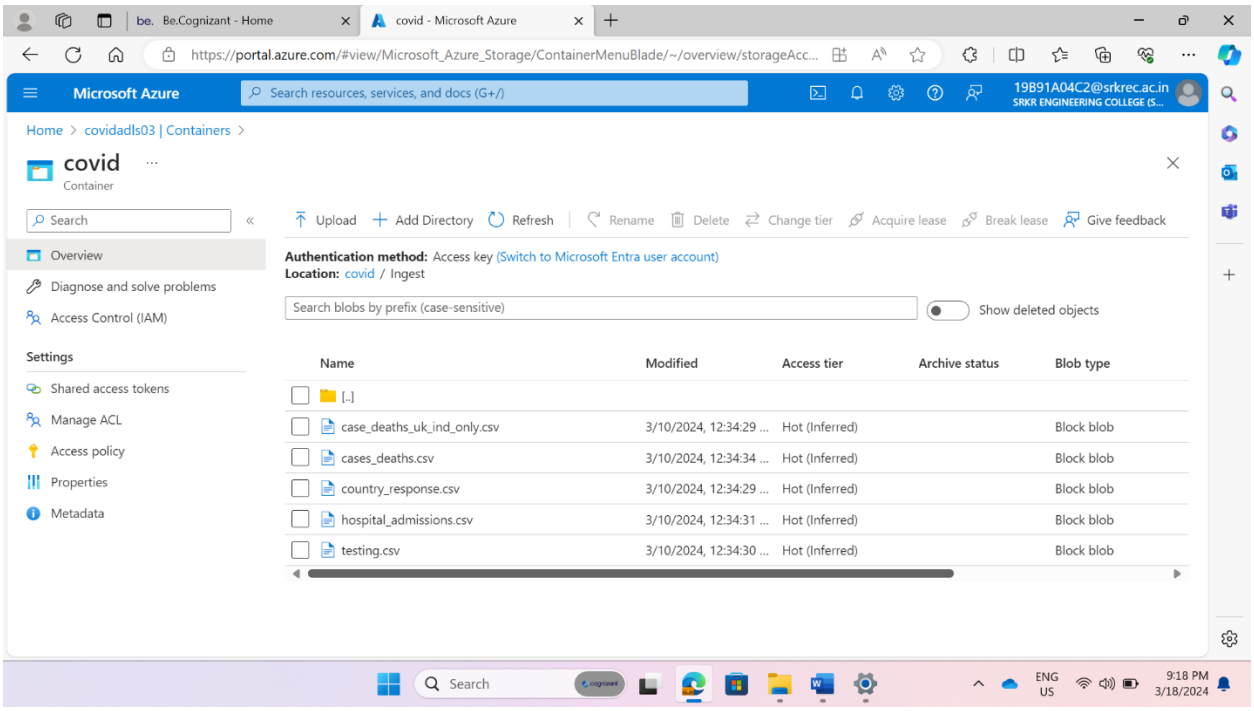
STEP 3: In our Resource Group, we need to create required resources for our project Like Storage Account, Synapse workspace (data warehouse), Azure Data Factory, DedicatedPool.



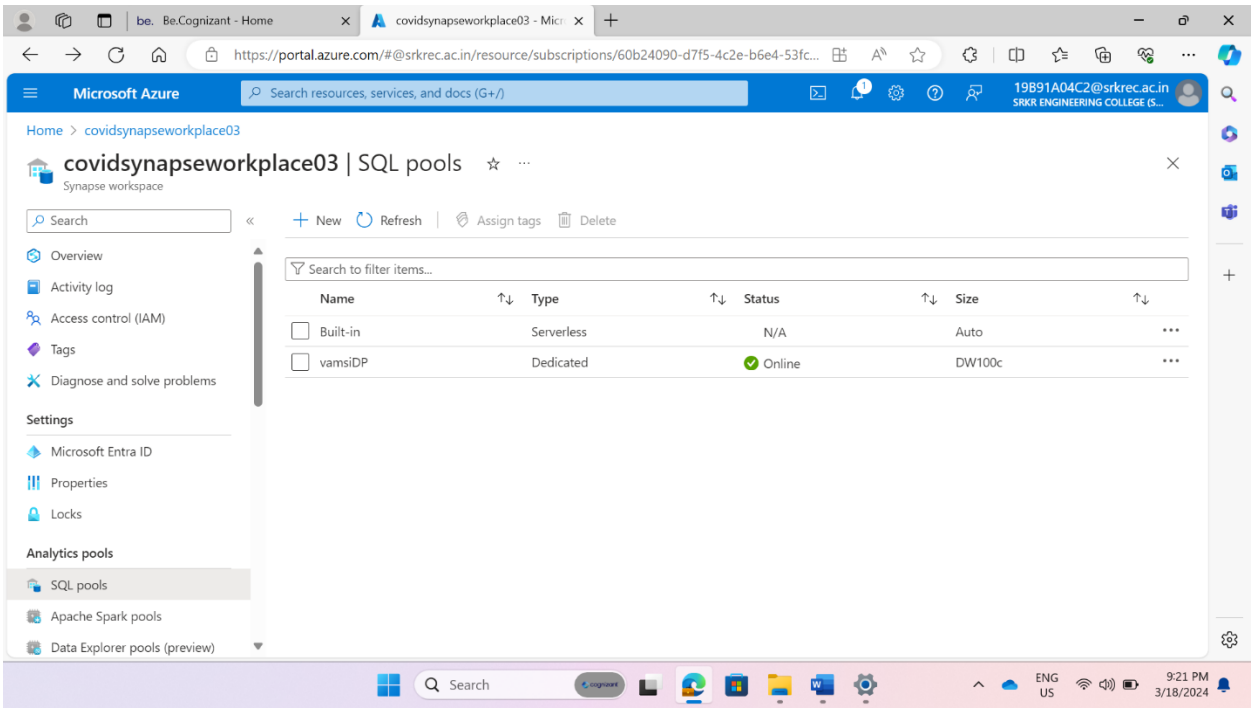
STEP 4: In storage account (ADLS) that we have created we need to create a container named **covid**. In this container we are creating a folder named **ingest** and in this folder we need to upload files.



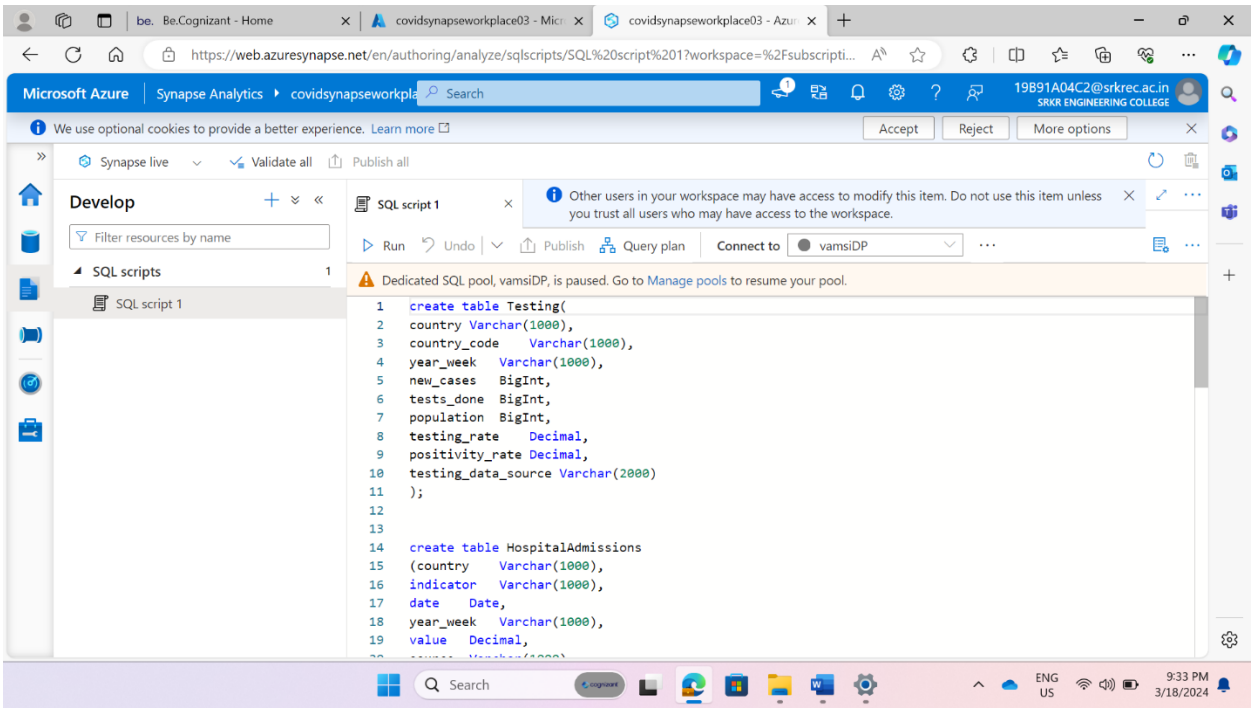


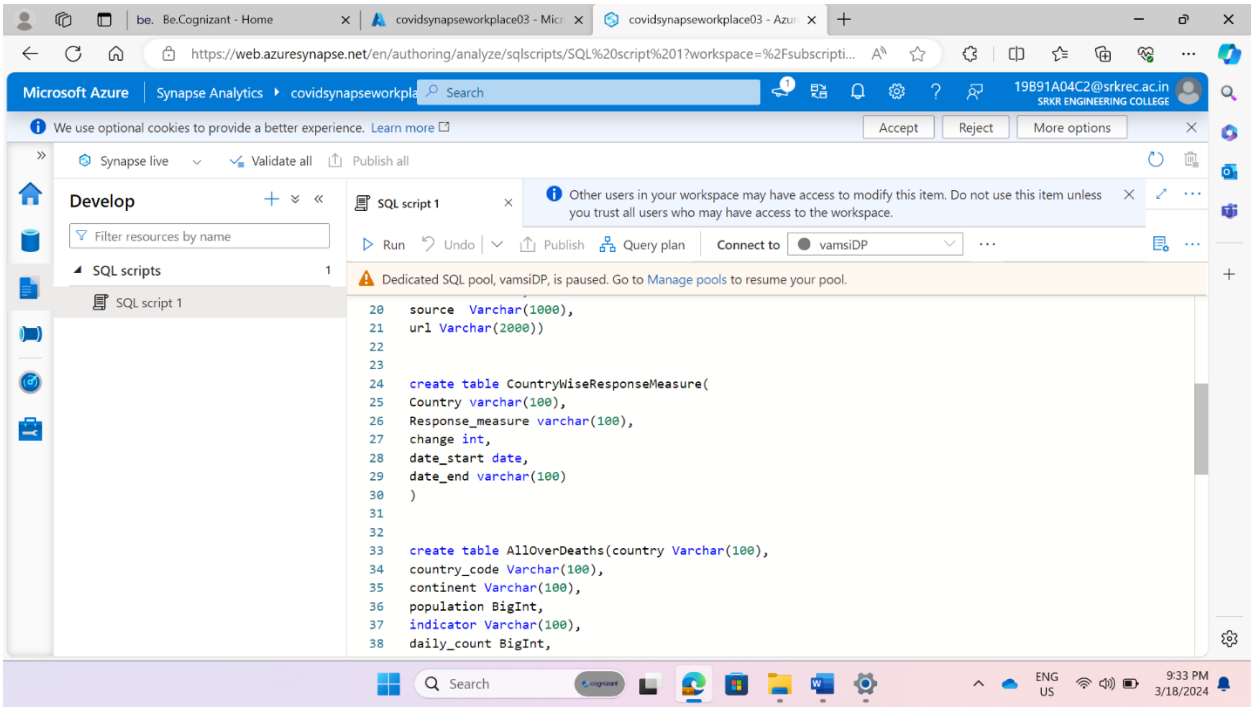


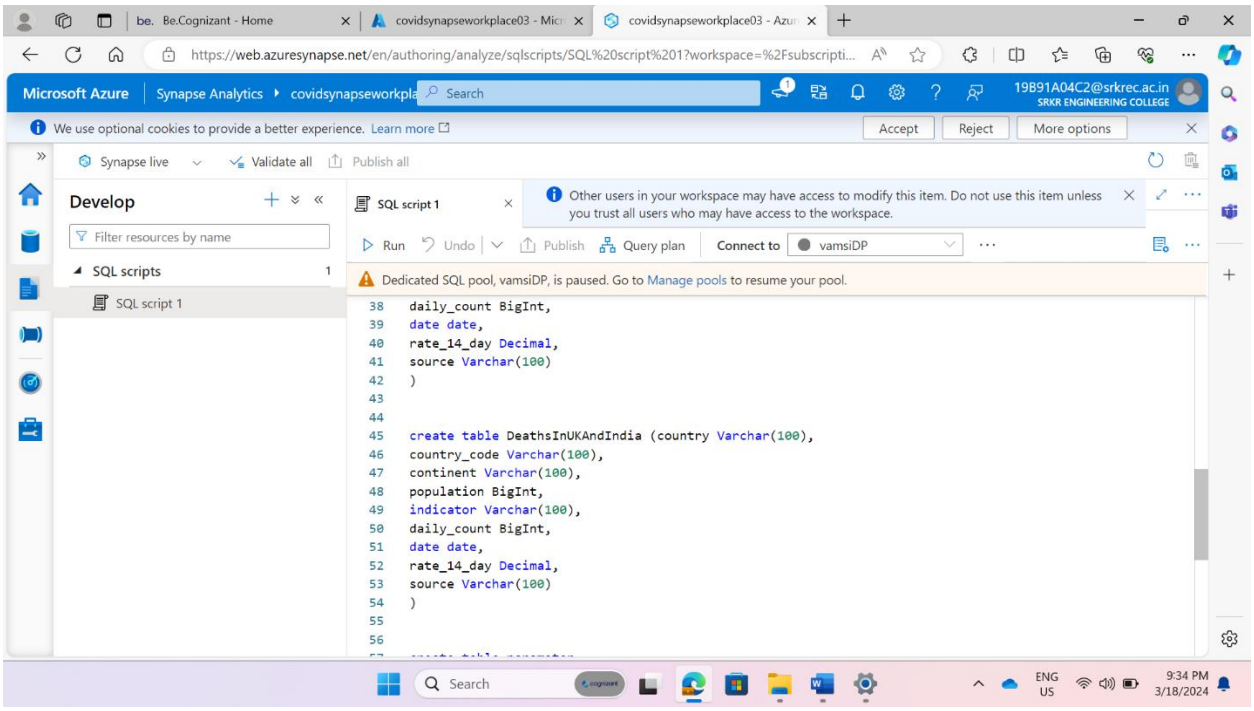
STEP 5: In Azure Synapse resource that we have created we need to create a dedicated pool and it should be turned on.



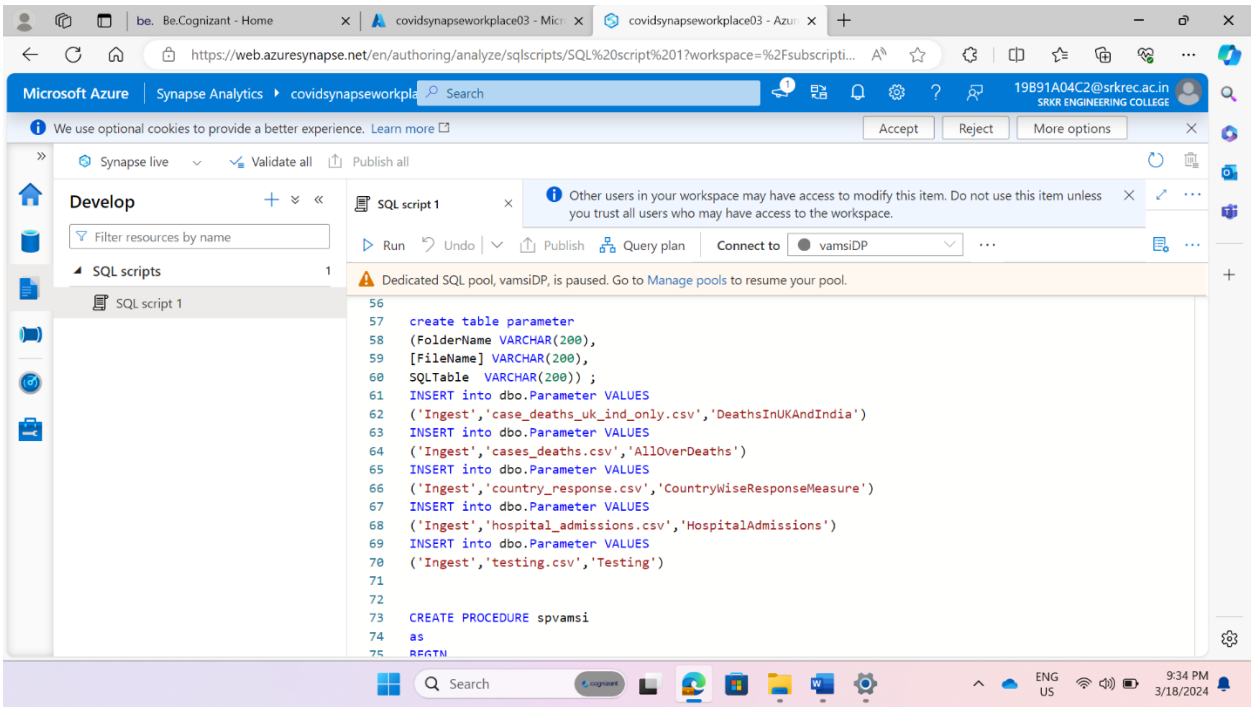
STEP 6: In Synapse Workspace, we need to create required tables using queries.



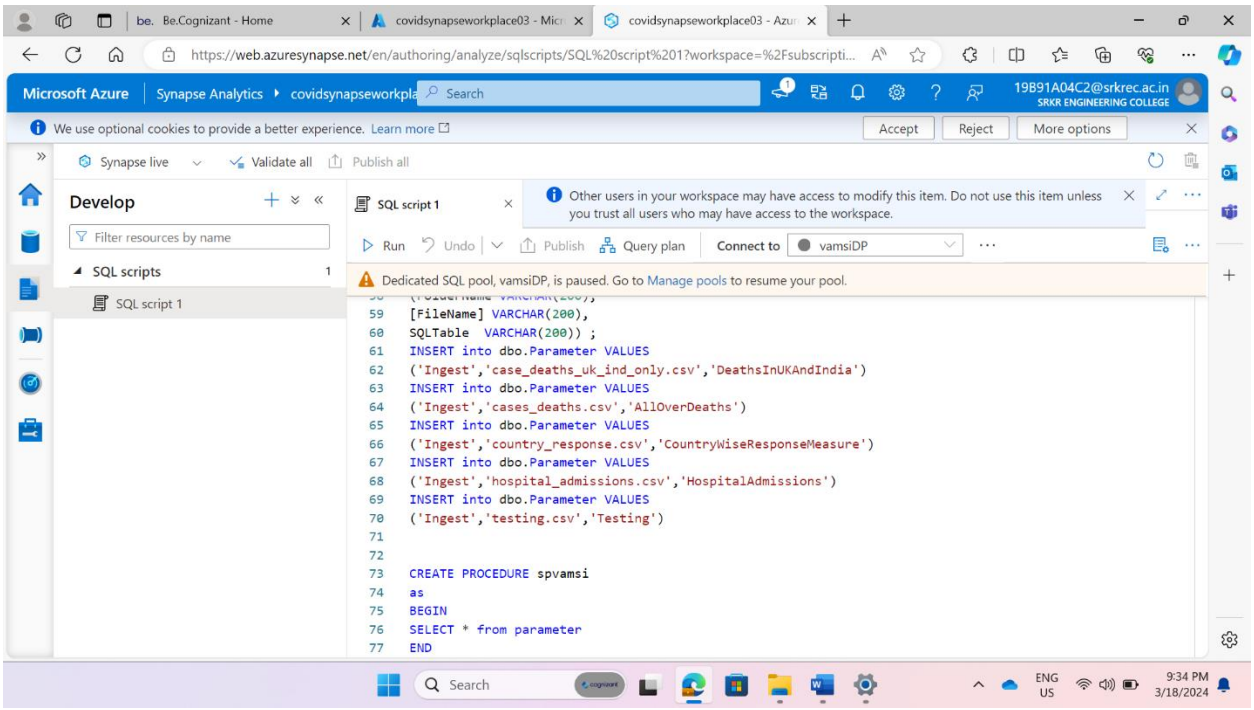




STEP 7: We need to create parameter Table and we need to Insert values to it.



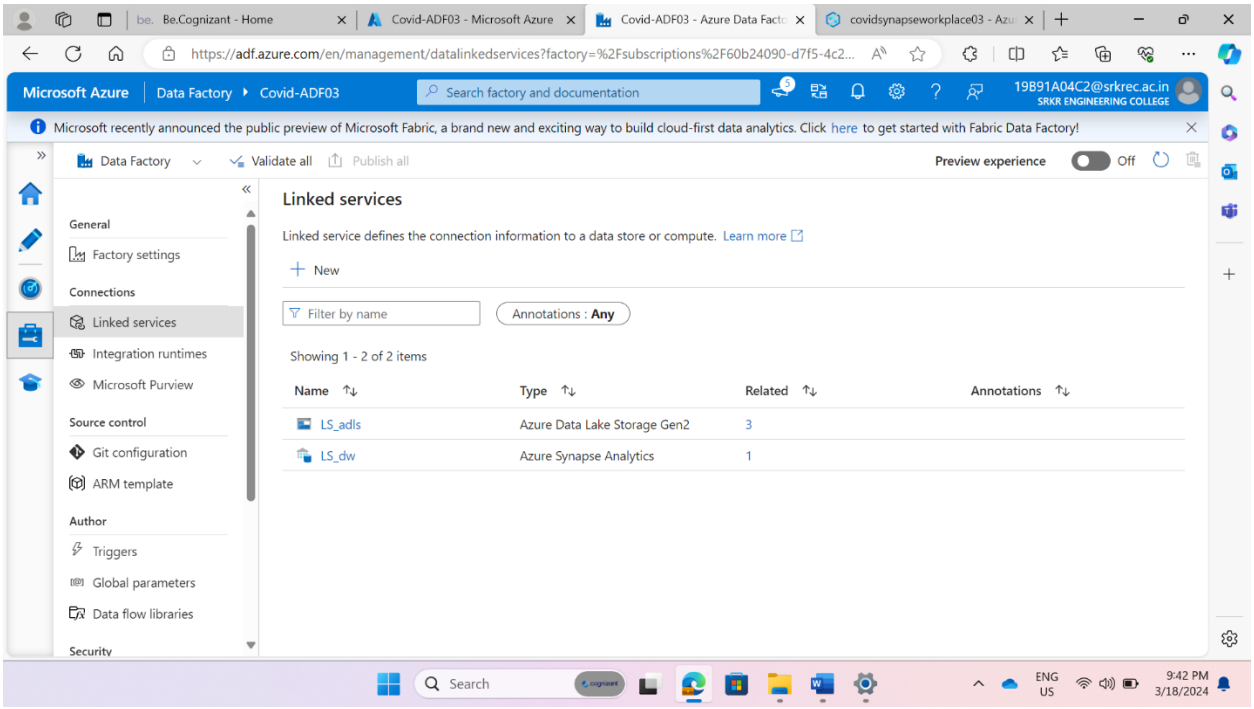
STEP 8: We need to create a Stored Procedure.



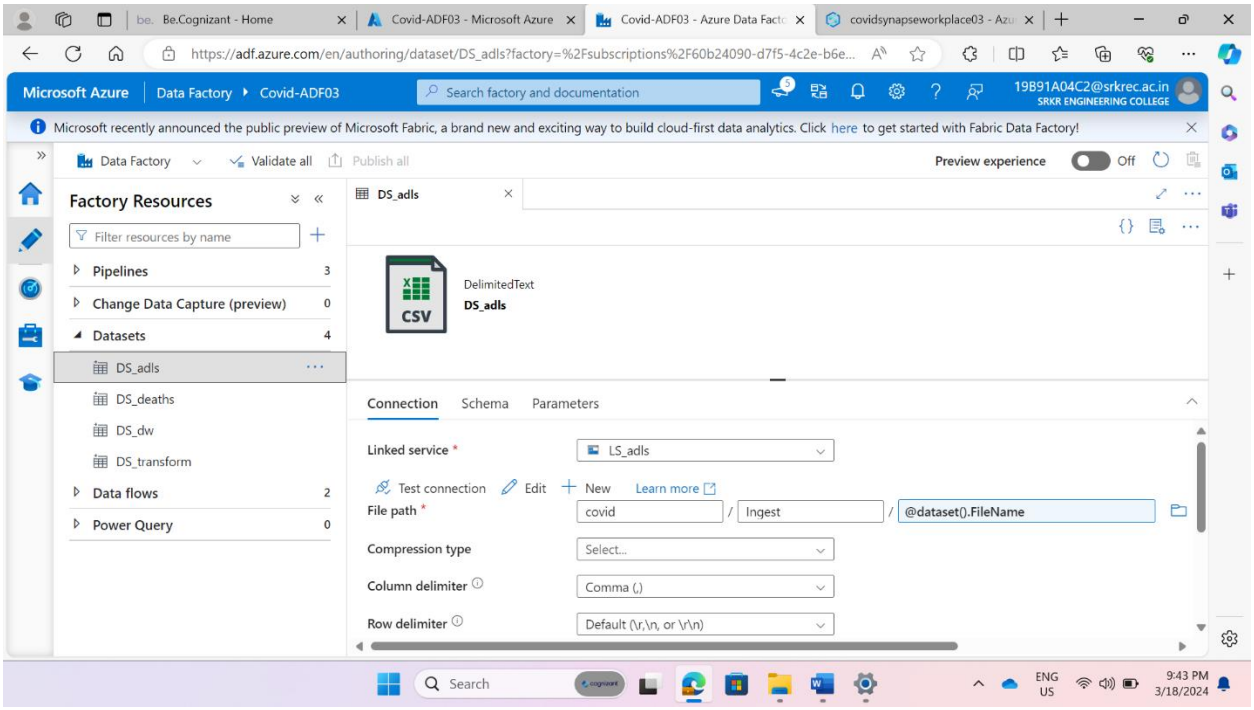
STEP 9: Creating **two linked services** as per the project requirement in datafactory

° **Storage account (data lake) to Azure Data factory.**

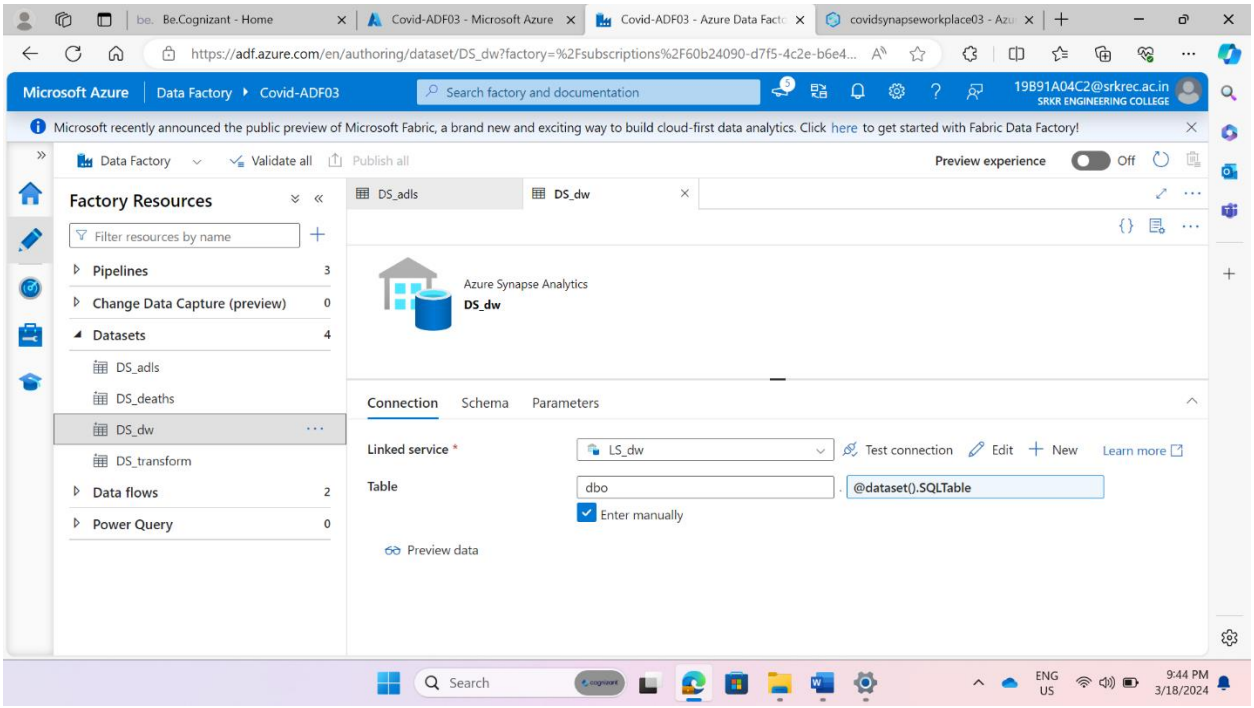
°**Azure synapse workspace to Azure Data factory.**



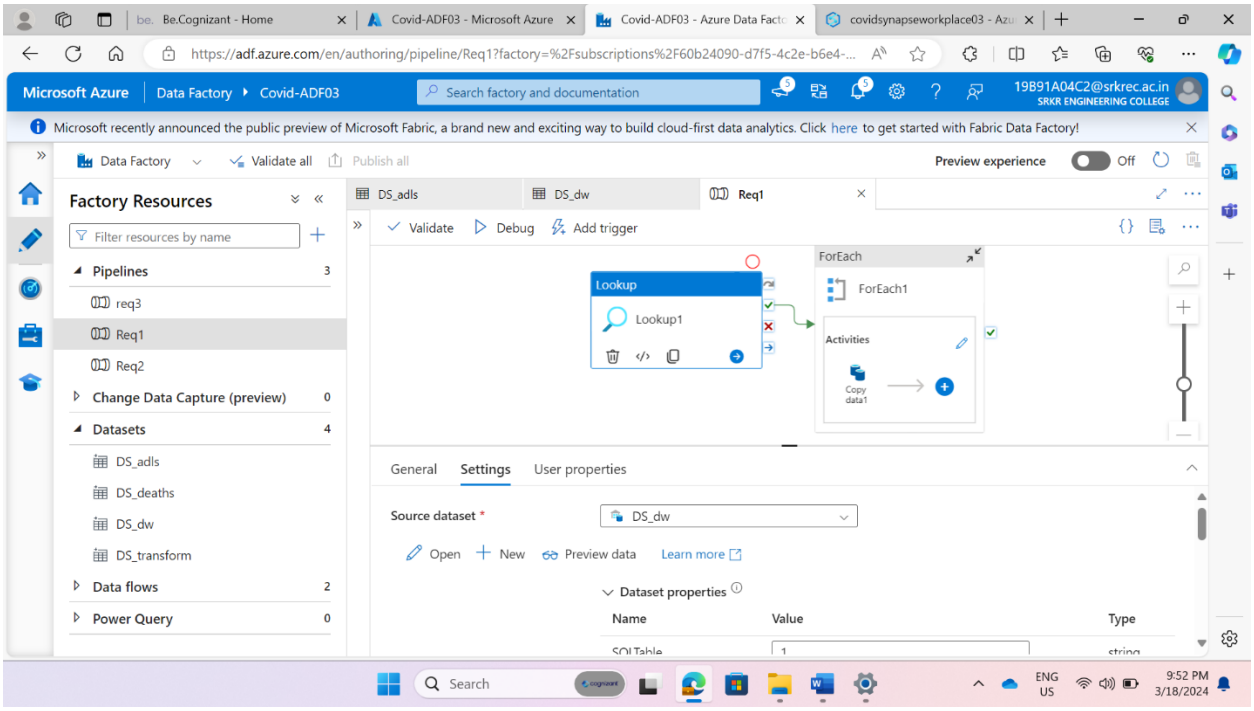
STEP 10: Create dataset **(DS_adls)** for fetching flat files from storage account (data lake) which are present in ingest folder inside covid container.

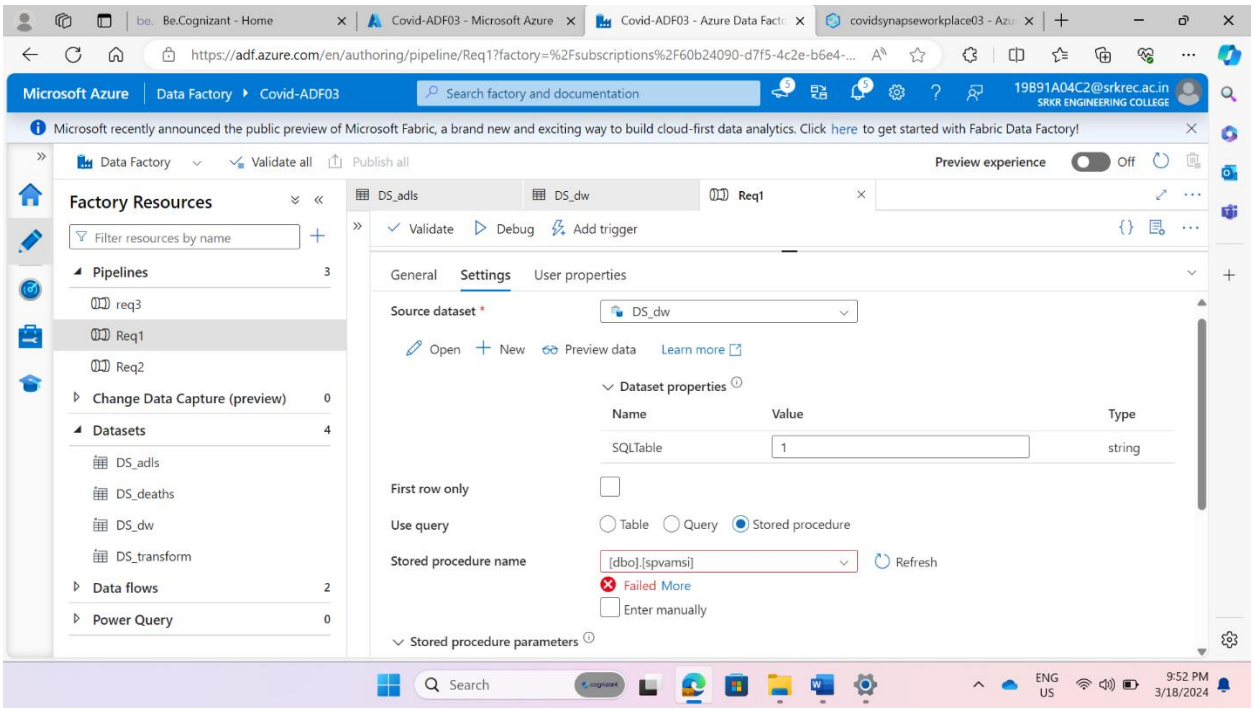


STEP 11: Create **dataset (DS_dw)** for inserting into SQL Tables created in synapse (data warehouse).

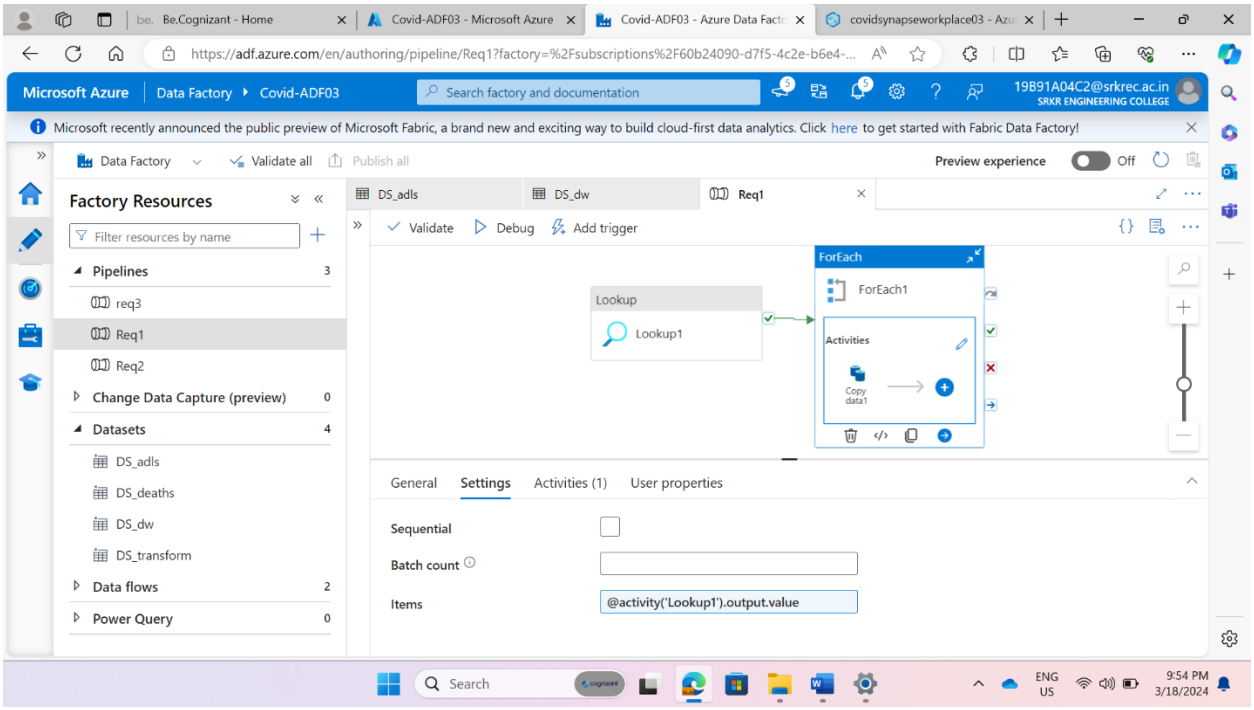


STEP-12: Create a pipeline and Drag and Drop the Look up Actvity into pipeline workspace and set the source dataset for Lookup and give the stored procedure name created in synapse (data warehouse).





STEP-13: Drag and Drop For each activity in pipeline workspace and configure the for each activity settings like Items with output of look up activity (@activity('Lookup1').output.value).



STEP-14: Click on copy data in foreach activity for copy of data from csv file to SQL Table. Configure settings at source side and sink side in copy data.

File Name (@ {item ().FileName}),

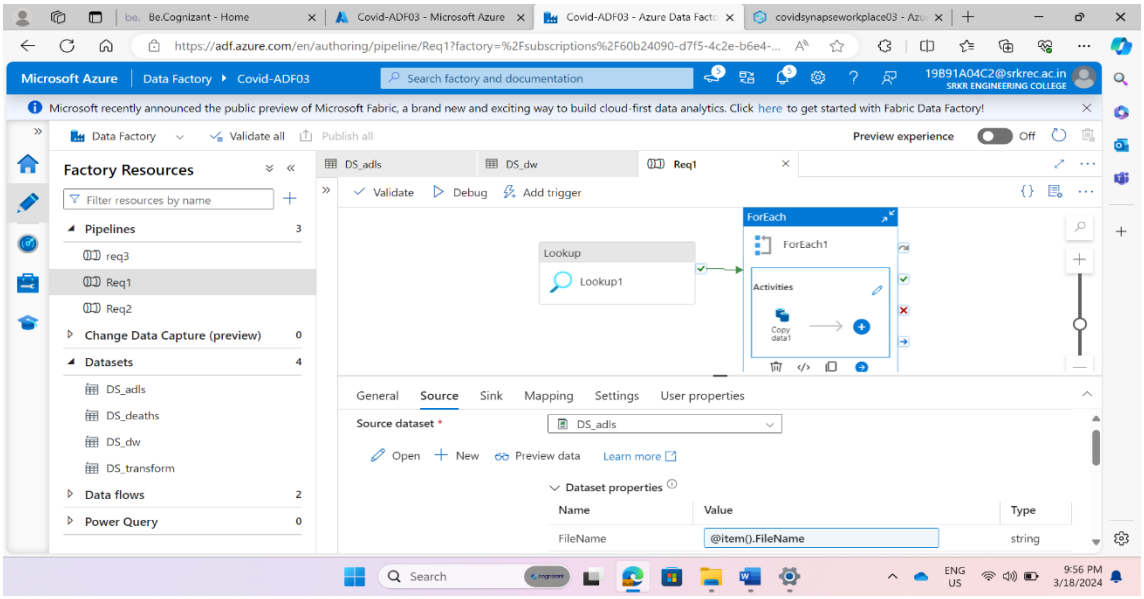
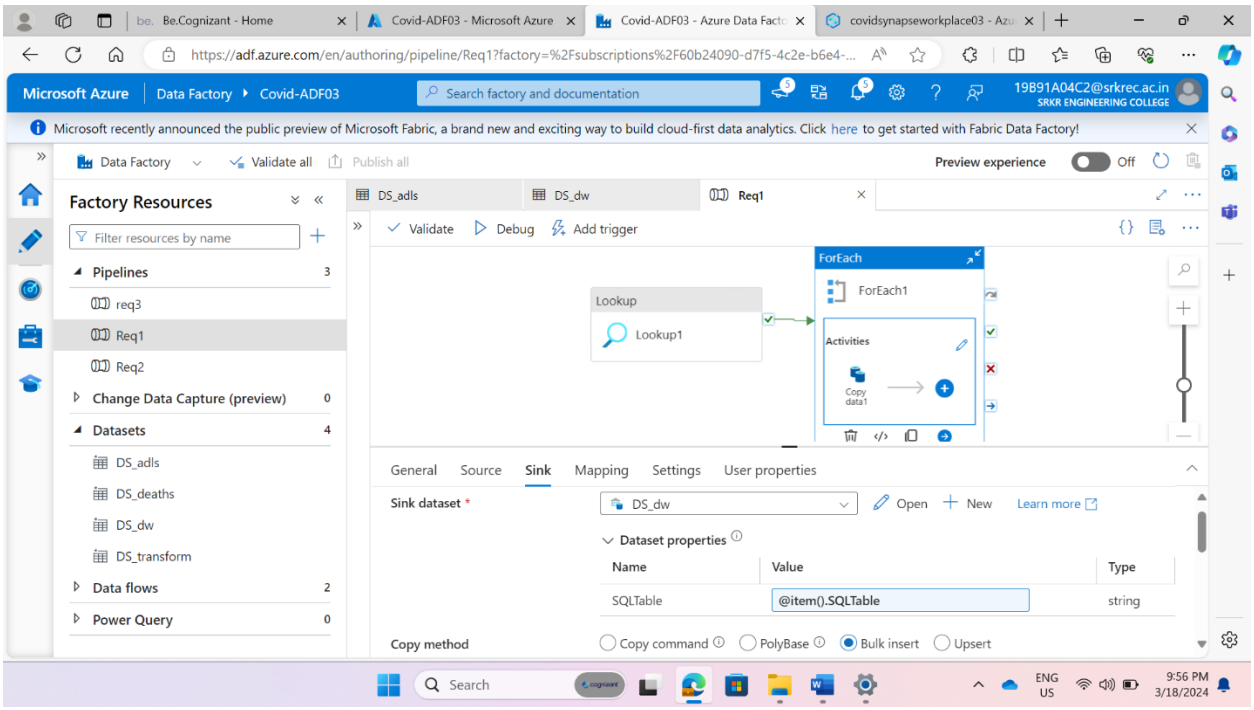
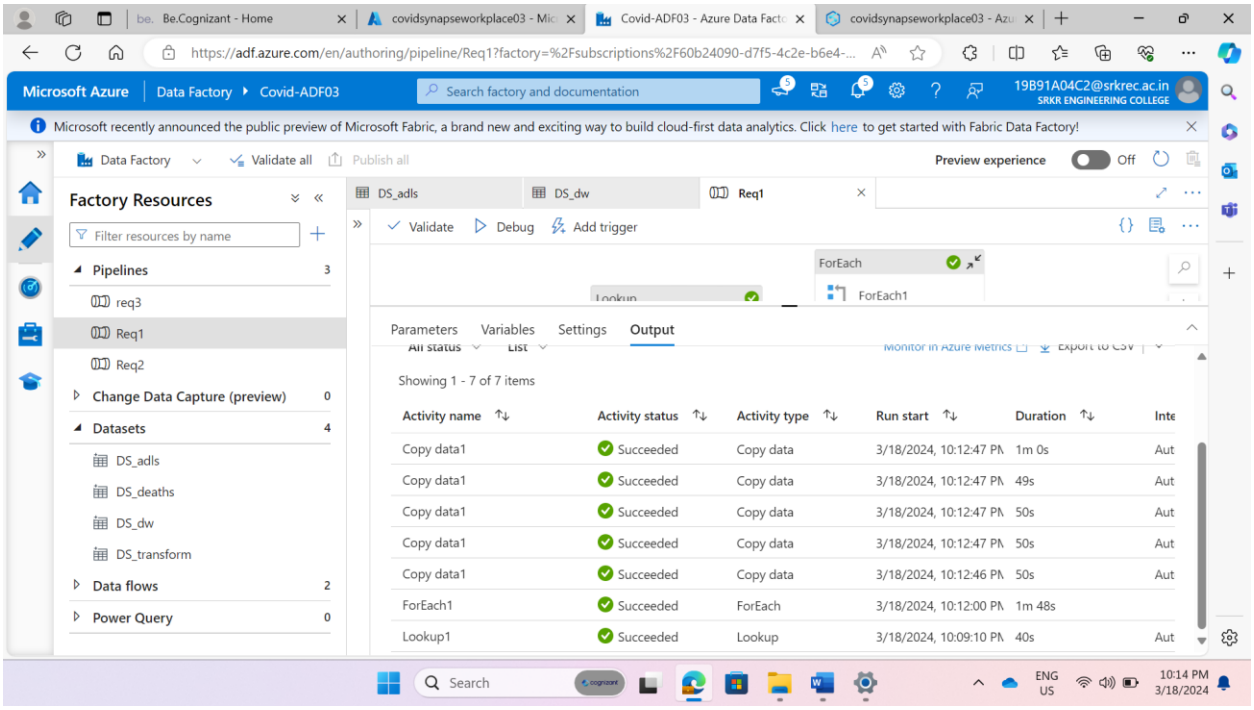


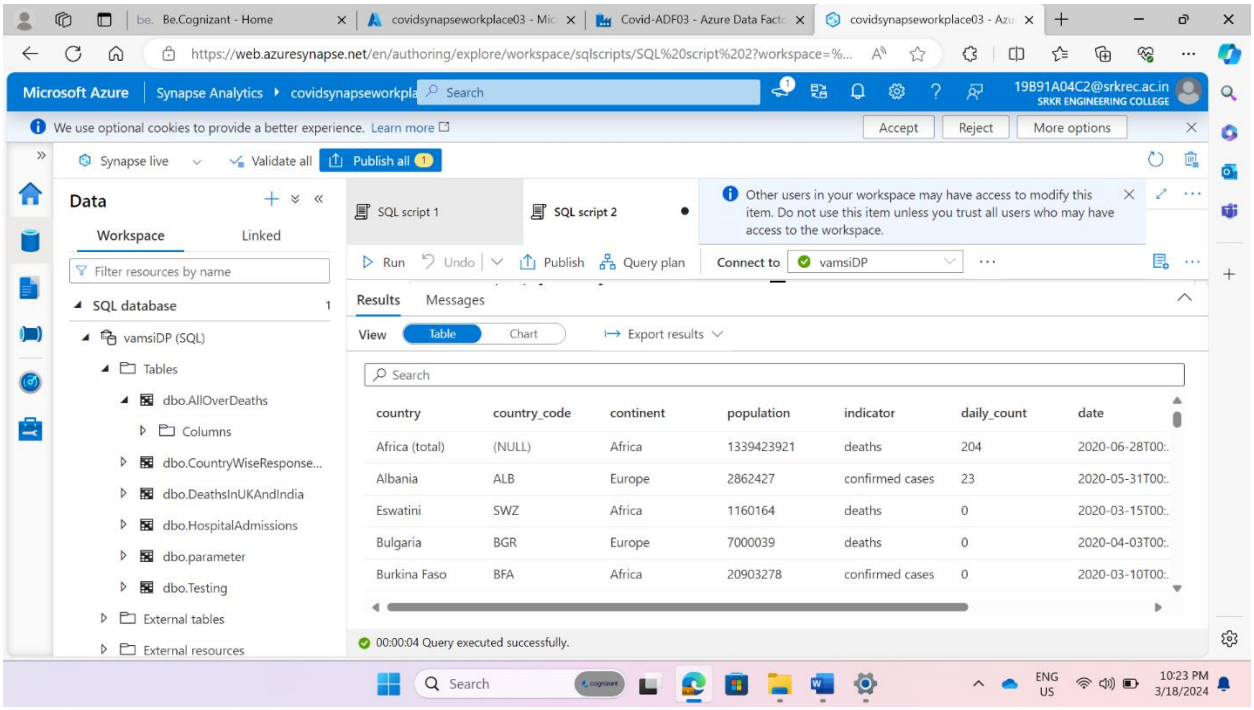
Table Name (@ {item().SQLTable})

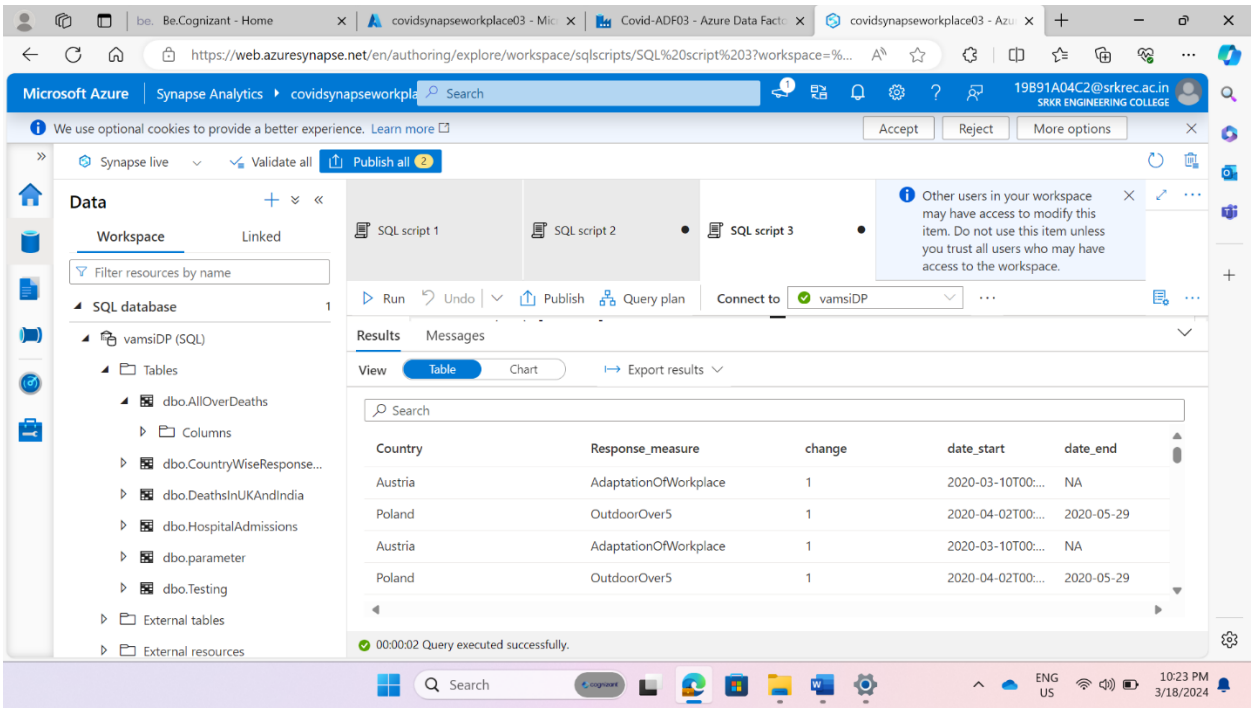


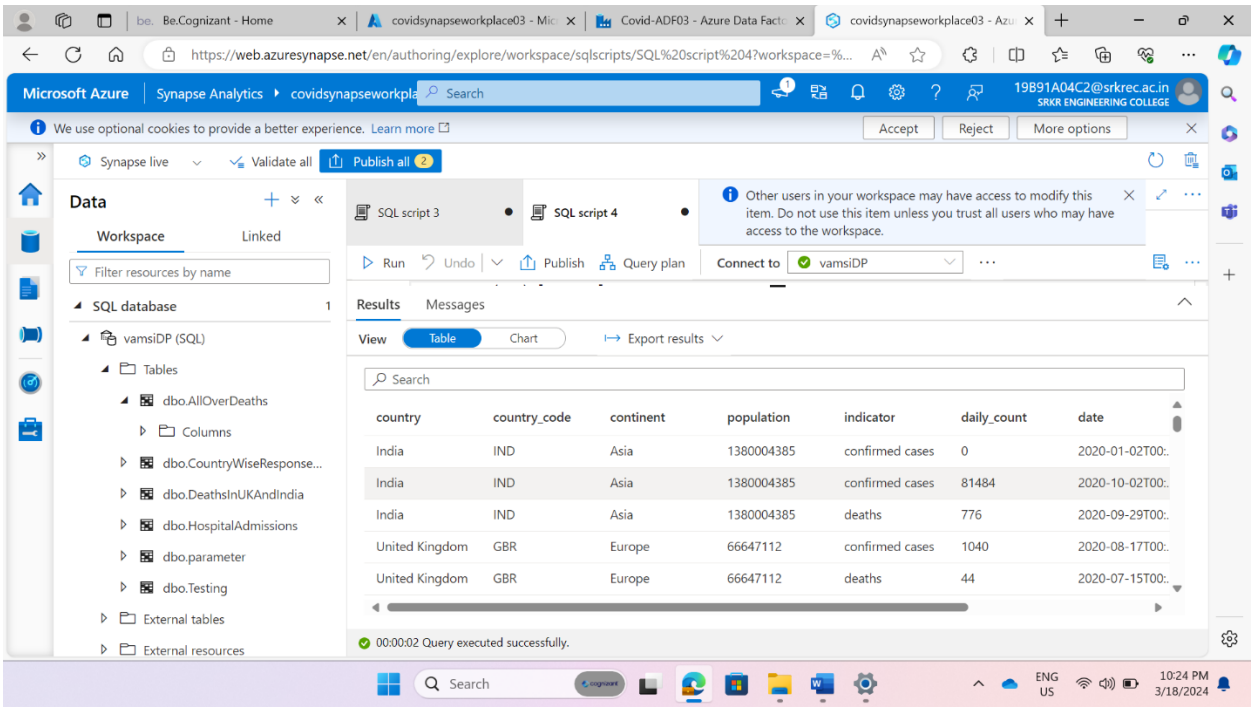
STEP-15: After setting whole pipeline by using Lookup and Foreach activity recheck all parameters, check the dedicated pool should be turn on and then turn on the debug option in pipeline. Finally, all the activities are successfully done.

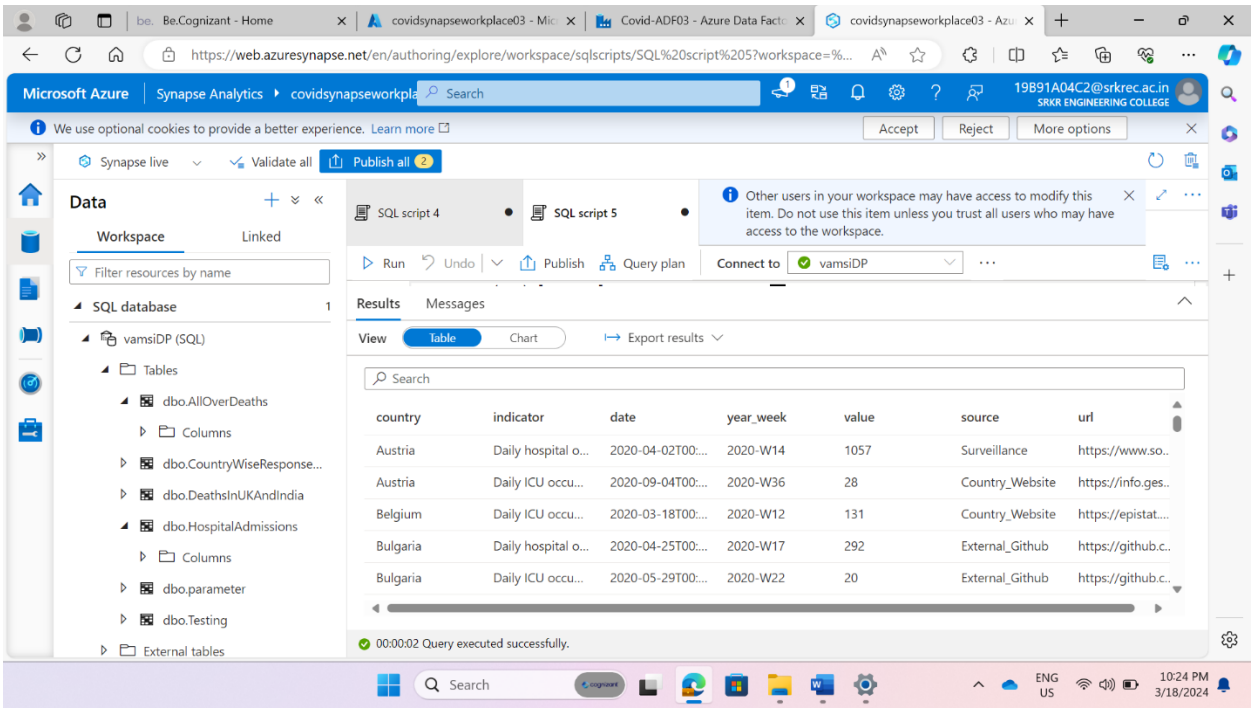


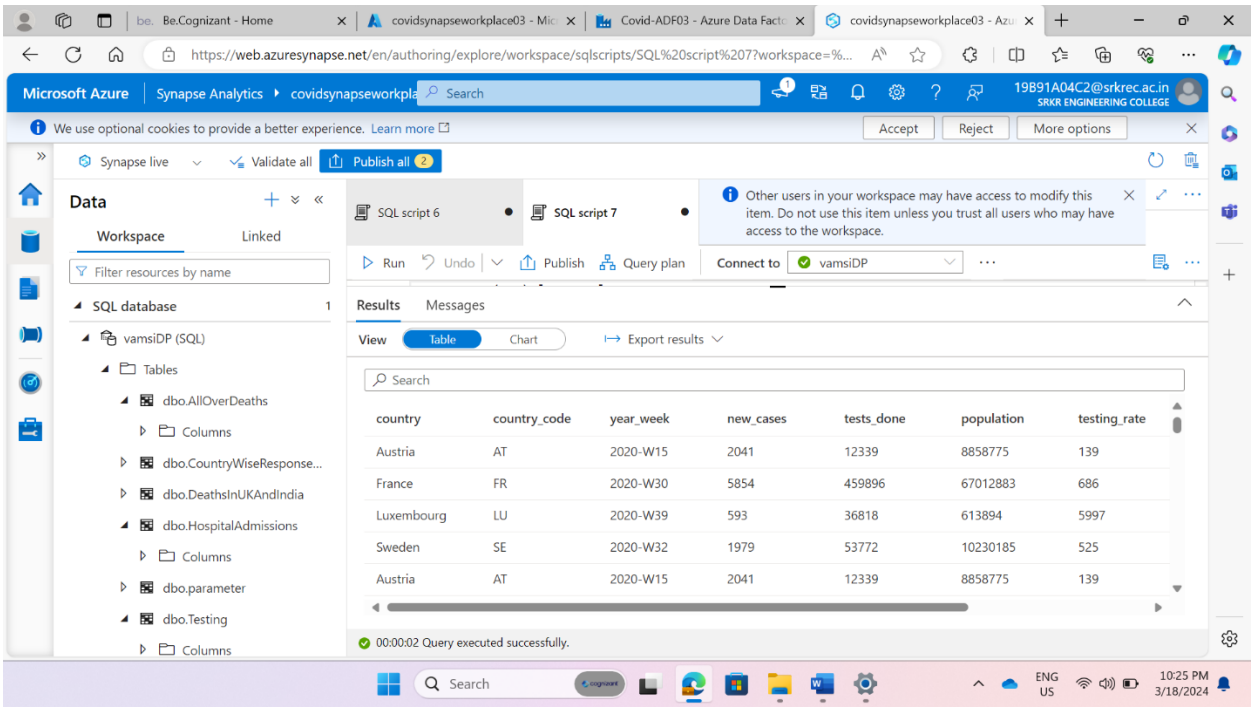
REQUIREMENT 1 OUTPUTS:





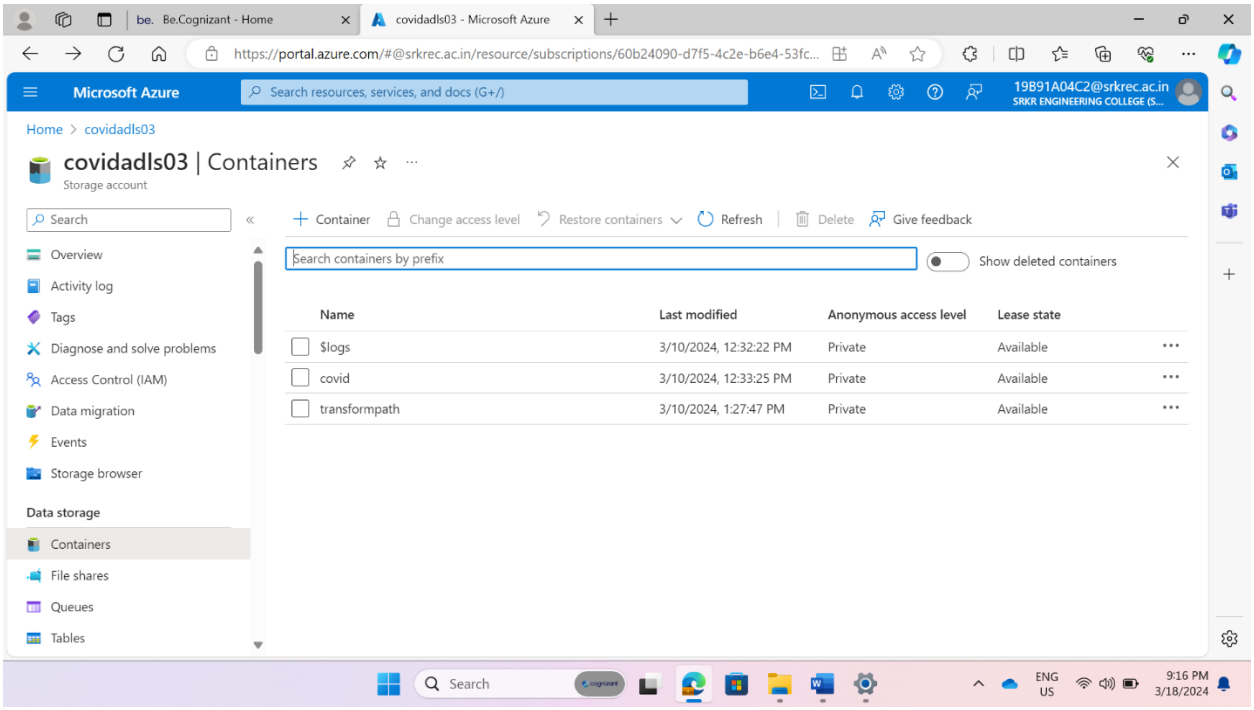

Microsoft Azure
Synapse Analytics
Search
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse live
Validate all
Publish all 2
Data
Workspace
Linked
Filter resources by name
SQL database
vamsiDP (SQL)
Tables
dbo.AllOverDeaths
Columns
dbo.CountryWiseResponse...
dbo.DeathsInUKAndIndia
dbo.HospitalAdmissions
dbo.parameter
dbo.Testing
External tables
External resources
SQL script 3
SQL script 4
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.
Run
Undo
Publish
Query plan
Connect to
vamsiDP
Results
Messages
View
Table
Chart
Export results
Search
country
country_code
continent
population
indicator
daily_count
date
India
IND
Asia
1380004385
confirmed cases
0
2020-01-02T00:..
India
IND
Asia
1380004385
confirmed cases
81484
2020-10-02T00:..
India
IND
Asia
1380004385
deaths
776
2020-09-29T00:..
United Kingdom
GBR
Europe
66647112
confirmed cases
1040
2020-08-17T00:..
United Kingdom
GBR
Europe
66647112
deaths
44
2020-07-15T00:..
00:00:02 Query executed successfully.
Search
ENG
US
10:24 PM
3/18/2024


Microsoft Azure
Synapse Analytics
Search
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse live
Validate all
Publish all 2
Data
Workspace
Linked
Filter resources by name
SQL database
vamsiDP (SQL)
Tables
dbo.AllOverDeaths
Columns
dbo.CountryWiseResponse...
dbo.DeathsInUKAndIndia
dbo.HospitalAdmissions
Columns
dbo.parameter
dbo.Testing
External tables
SQL script 4
SQL script 5
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.
Run
Undo
Publish
Query plan
Connect to
vamsiDP
Results
Messages
View
Table
Chart
Export results
Search
country
indicator
date
year_week
value
source
url
Austria
Daily hospital o...
2020-04-02T00:...
2020-W14
1057
Surveillance
https://www.so..
Austria
Daily ICU occu...
2020-09-04T00:...
2020-W36
28
Country_Website
https://info.ges..
Belgium
Daily ICU occu...
2020-03-18T00:...
2020-W12
131
Country_Website
https://epistat....
Bulgaria
Daily hospital o...
2020-04-25T00:...
2020-W17
292
External_Github
https://github.c..
Bulgaria
Daily ICU occu...
2020-05-29T00:...
2020-W22
20
External_Github
https://github.c..
00:00:02 Query executed successfully.
Search
ENG
US
10:24 PM
3/18/2024

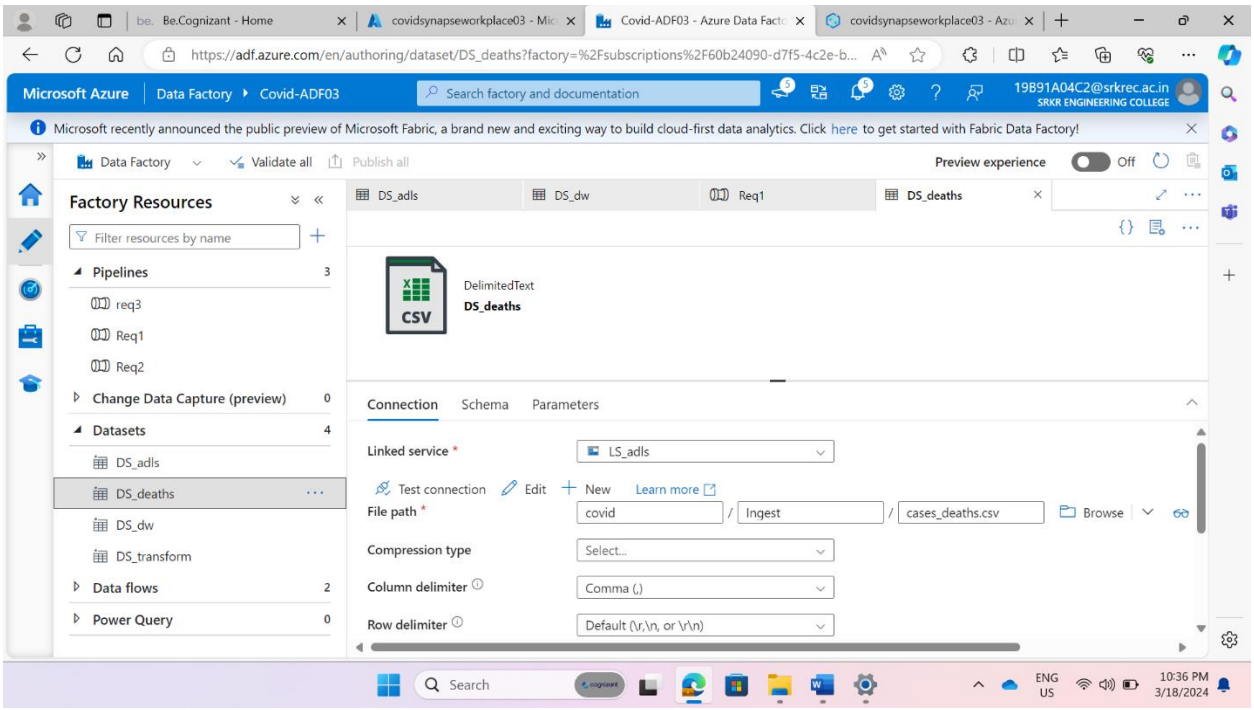


Procedure for Project Requirement 2:

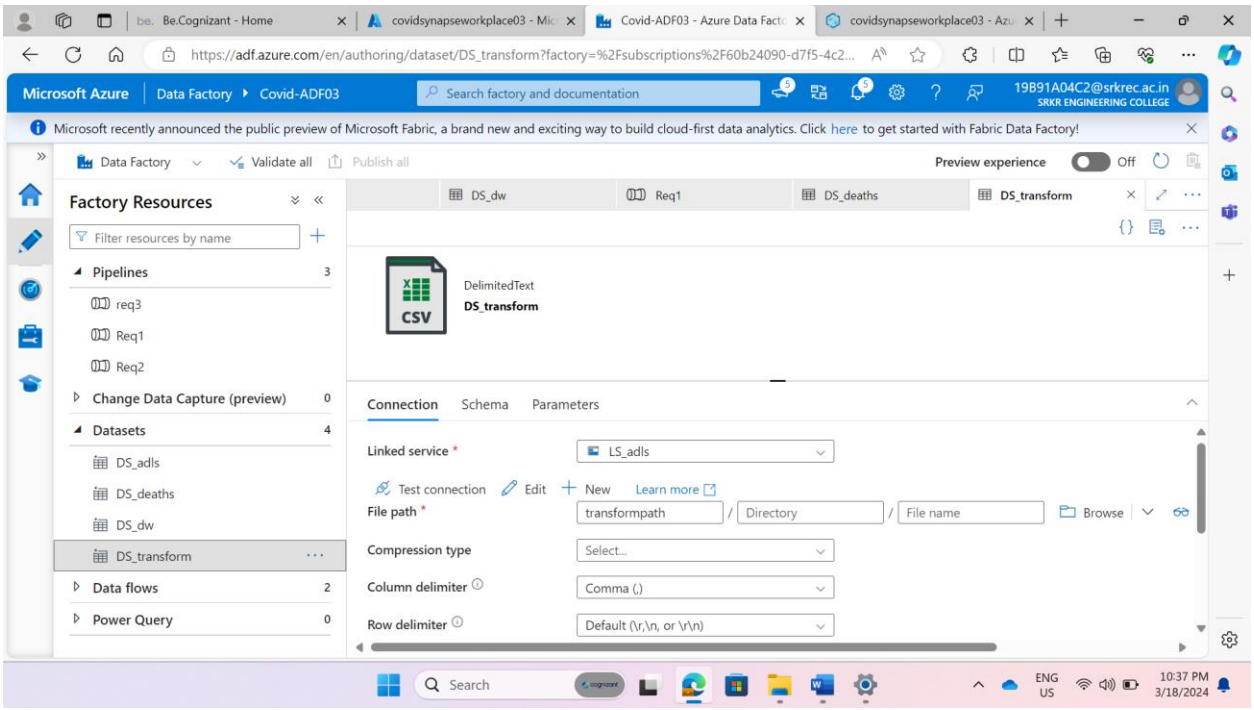
STEP 1: Create another container with name **"transformpath"** for second requirement given in project for storing transformed data file by using data flow.



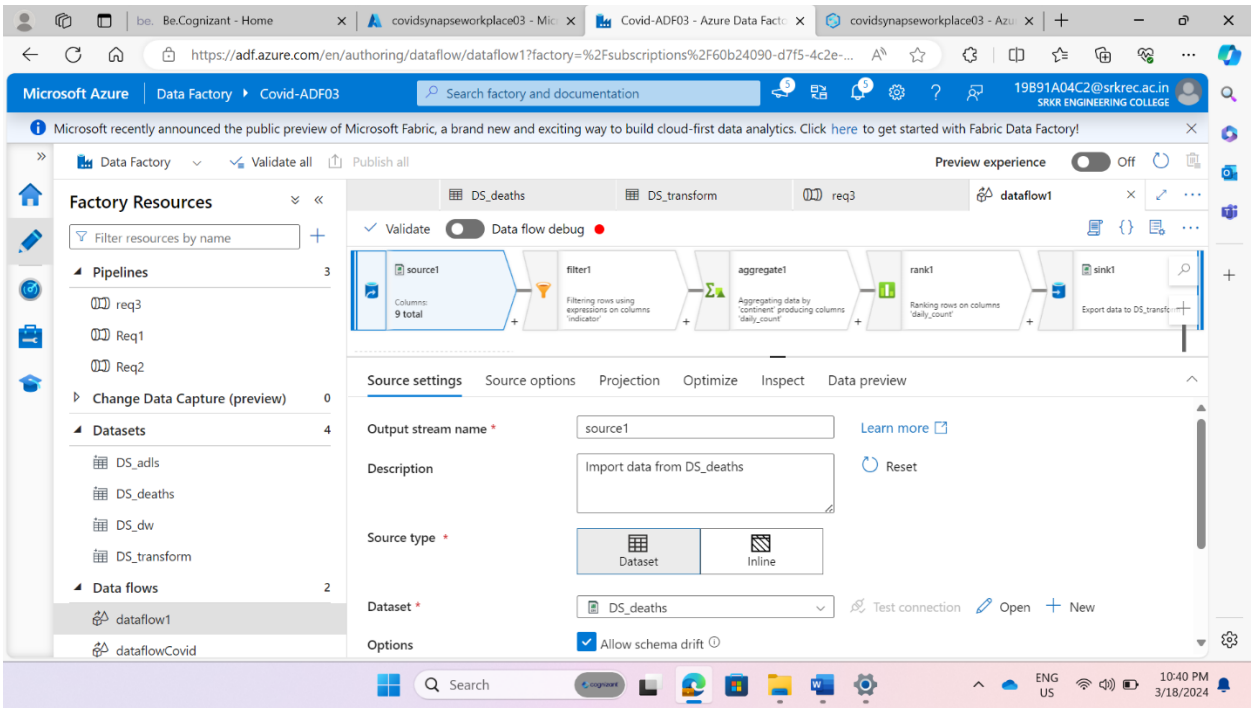
STEP 2: Create source dataset (**DS_deaths**) for dataflow by giving a file specific file name on which data transformation need to be taken place as per project requirement.

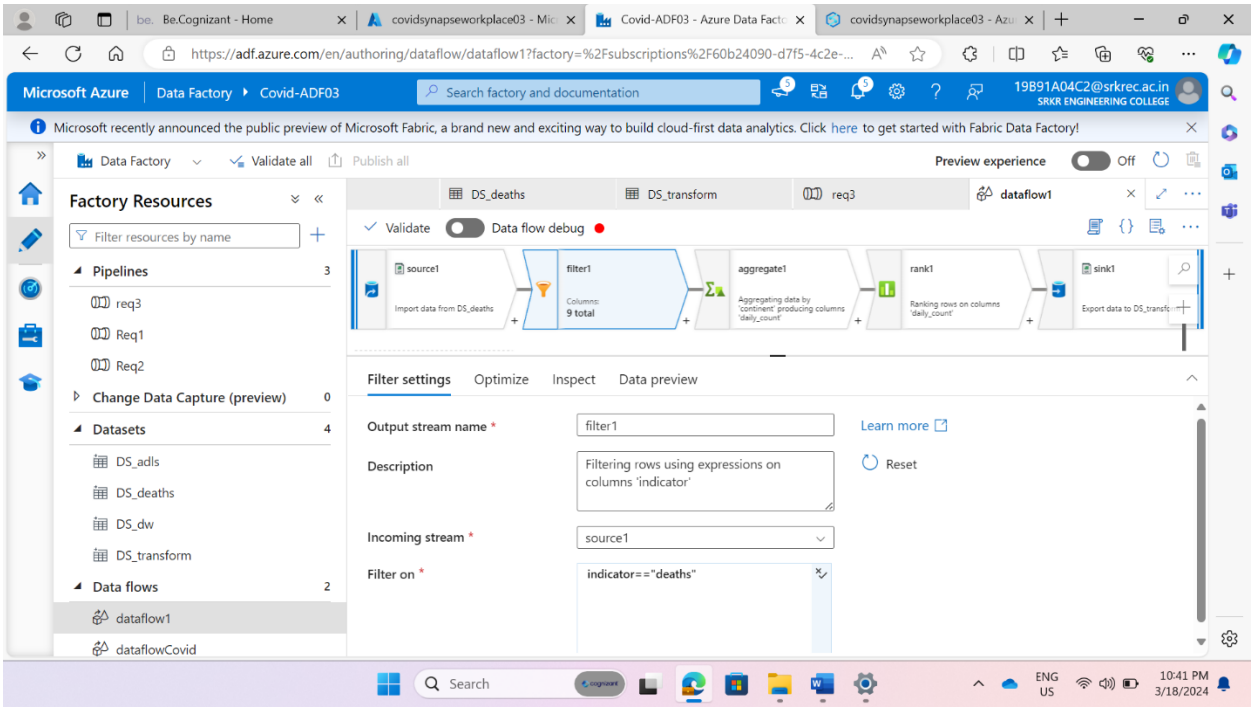


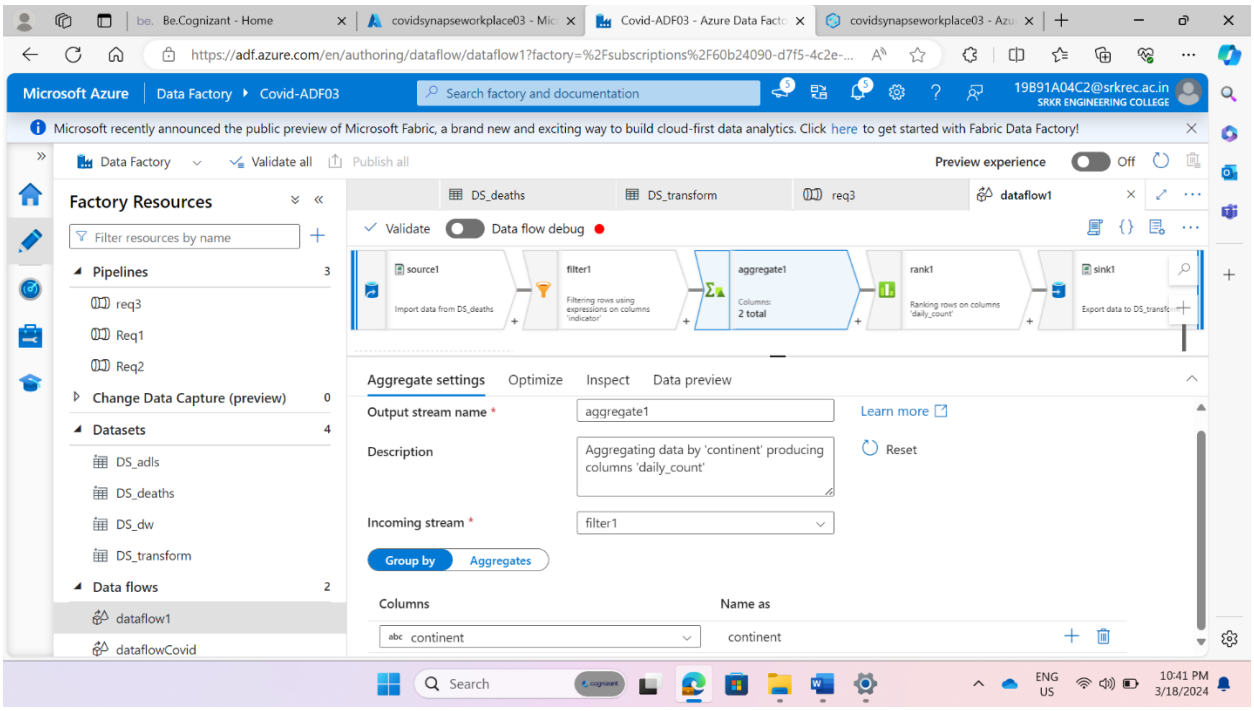
STEP 3: Create target dataset (**DS_transform**) for dataflow to keep that data transformed file in specific place for further use.

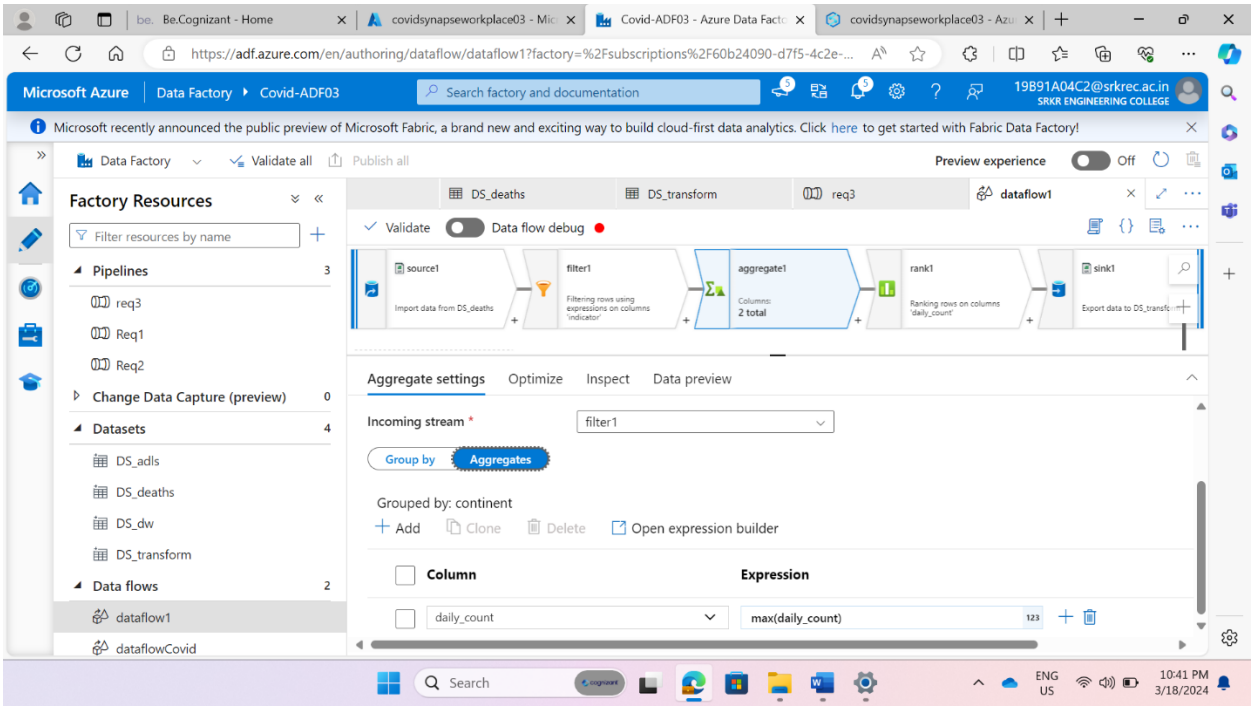


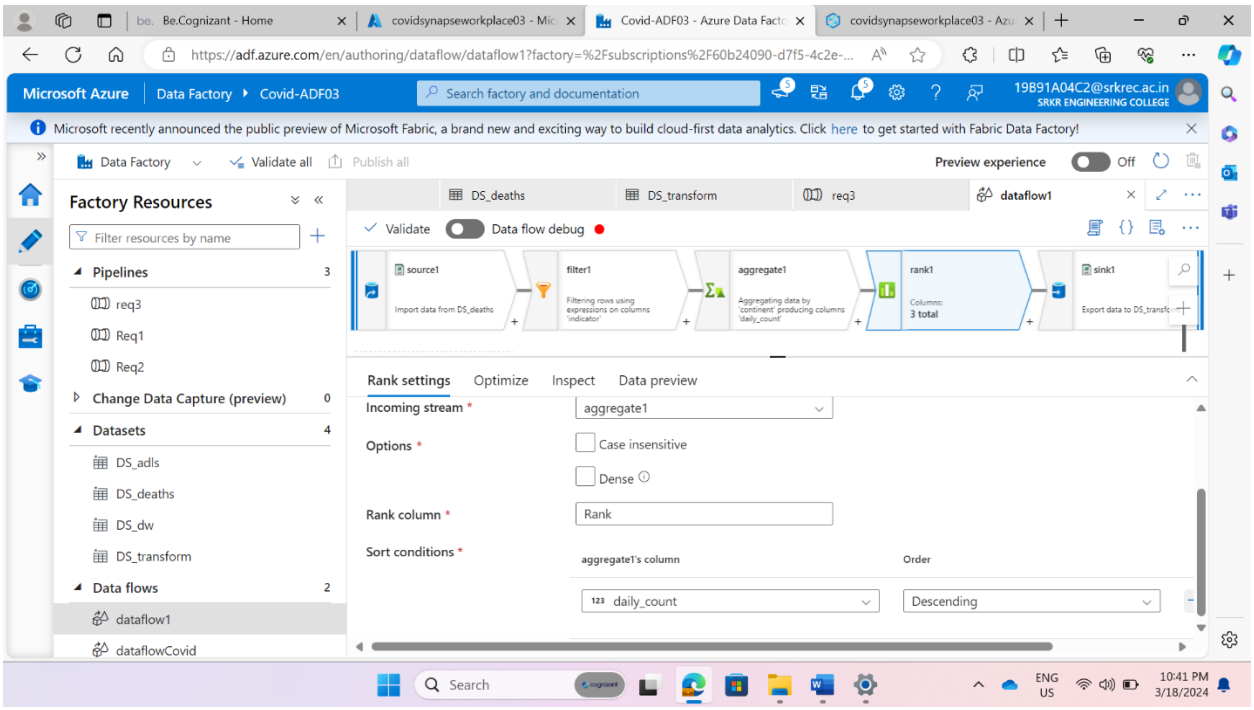
STEP 4: Develop dataflow by using some transformations like source, filter, aggregate, rank and sink as per the question given in project documentation.

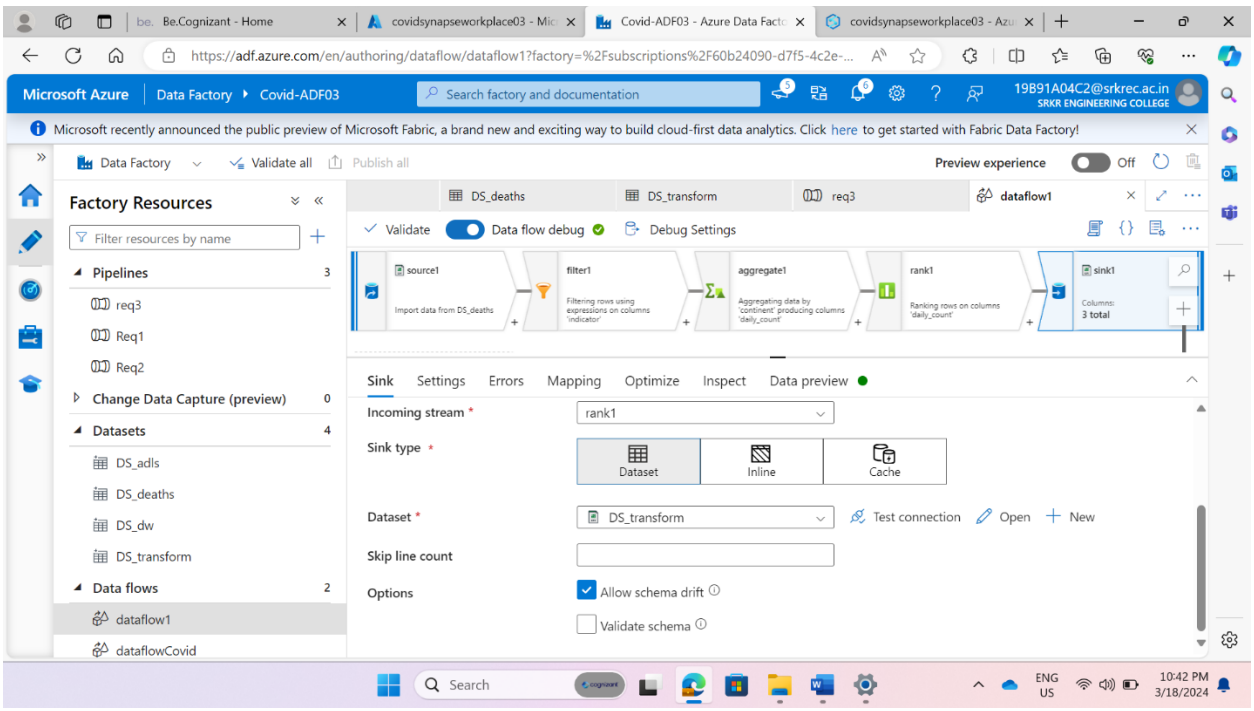




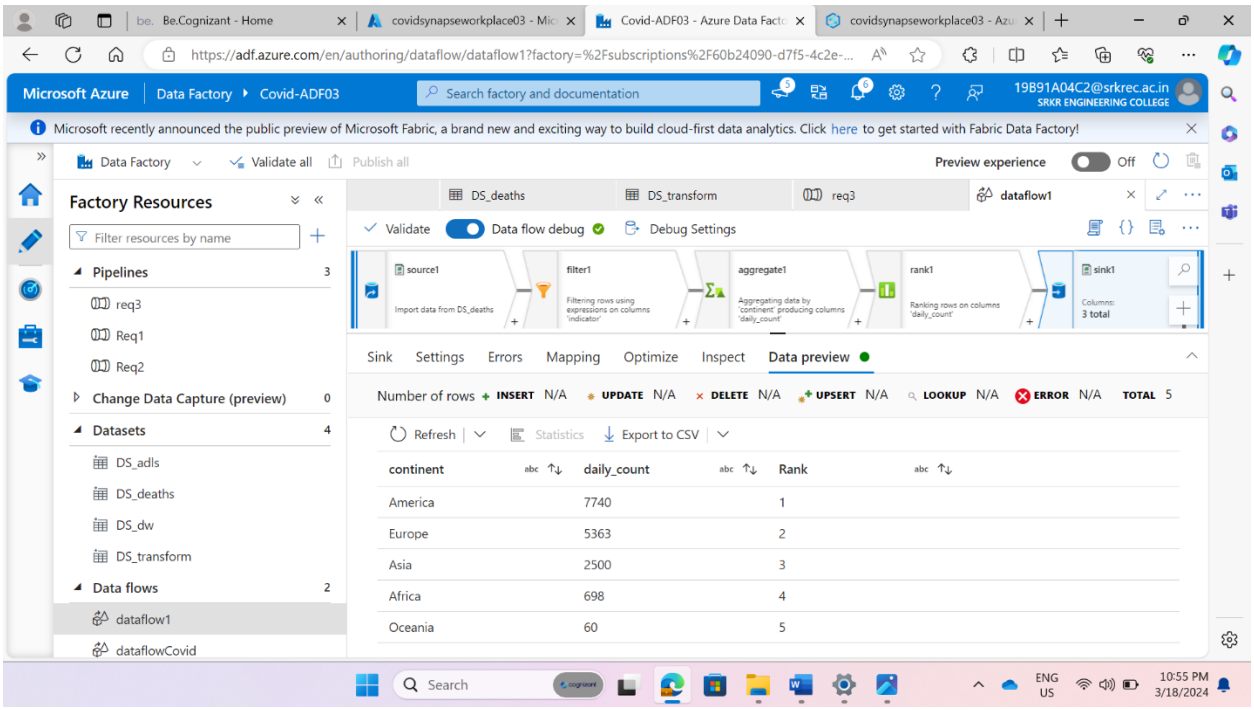

Microsoft Azure
Data Factory
Covid-ADF03
Search factory and documentation
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Validate all
Publish all
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines
req3
Req1
Req2
Change Data Capture (preview)
Datasets
DS_adls
DS_deaths
DS_dw
DS_transform
Data flows
dataflow1
dataflowCovid
DS_deaths
DS_transform
req3
dataflow1
Validate
Data flow debug
source1
Import data from DS_deaths
filter1
aggregate1
Columns:
2 total
rank1
sink1
Aggregate settings
Optimize
Inspect
Data preview
Output stream name *
aggregate1
Learn more
Description
Aggregating data by 'continent' producing columns 'daily_count'
Reset
Incoming stream *
filter1
Group by
Aggregates
Columns
Name as
continent
continent
Search
ENG
US
10:41 PM
3/18/2024


Microsoft Azure
Data Factory
Covid-ADF03
Search factory and documentation
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!
Validate all
Publish all
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines
req3
Req1
Req2
Change Data Capture (preview)
Datasets
DS_adls
DS_deaths
DS_dw
DS_transform
Data flows
dataflow1
dataflowCovid
DS_deaths
DS_transform
req3
dataflow1
Validate
Data flow debug
source1
Import data from DS_deaths
filter1
aggregate1
Columns:
2 total
rank1
sink1
Aggregate settings
Optimize
Inspect
Data preview
Incoming stream *
filter1
Group by
Aggregates
Grouped by: continent
Add
Clone
Delete
Open expression builder
Column
Expression
daily_count
max(daily_count)
Search
ENG
US
10:41 PM
3/18/2024

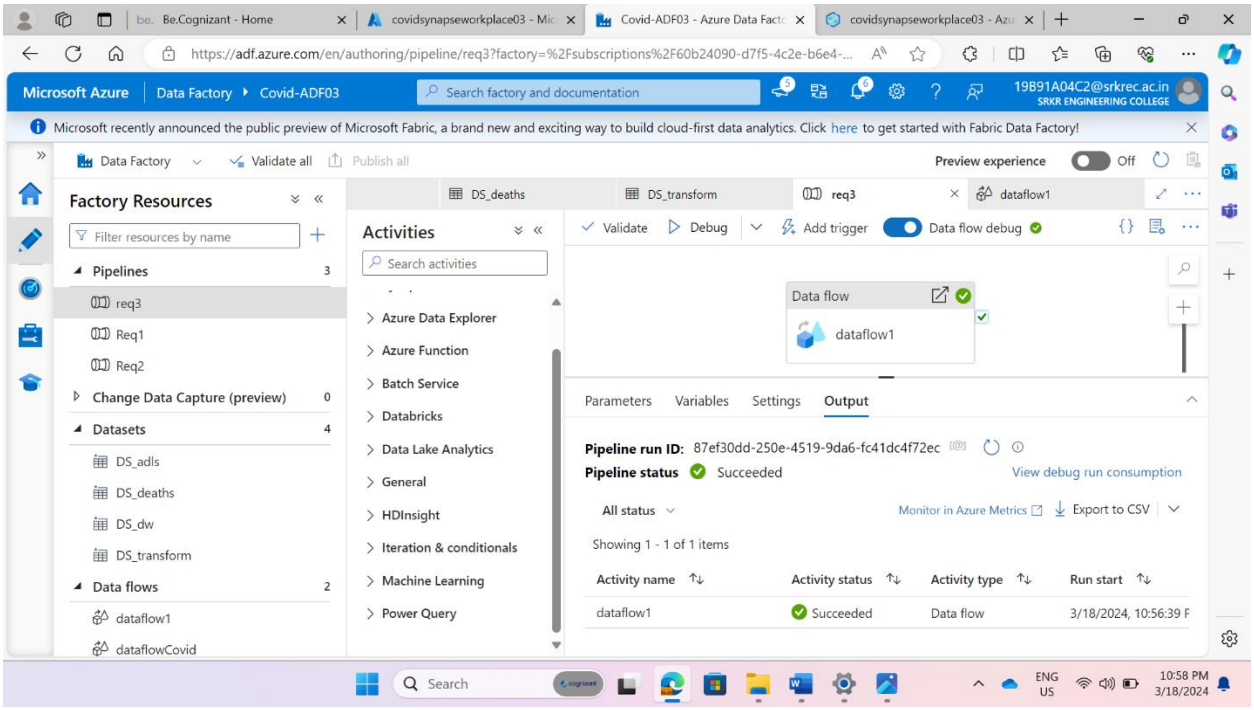

Microsoft Azure
Data Factory
Covid-ADF03
Search factory and documentation
Factory Resources
Filter resources by name
Pipelines
req3
Req1
Req2
Change Data Capture (preview)
Datasets
DS_adls
DS_deaths
DS_dw
DS_transform
Data flows
dataflow1
dataflowCovid
Validate
Data flow debug
source1
filter1
aggregate1
rank1
sink1
Rank settings
Optimize
Inspect
Data preview
Incoming stream *
aggregate1
Options *
Case insensitive
Dense
Rank column *
Rank
Sort conditions *
aggregate1's column
Order
daily_count
Descending


Microsoft Azure
Data Factory
Covid-ADF03
Search factory and documentation
Factory Resources
Filter resources by name
Pipelines
req3
Req1
Req2
Change Data Capture (preview)
Datasets
DS_adls
DS_deaths
DS_dw
DS_transform
Data flows
dataflow1
dataflowCovid
Validate
Data flow debug
Debug Settings
source1
filter1
aggregate1
rank1
sink1
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Incoming stream *
rank1
Sink type *
Dataset
Inline
Cache
Dataset *
DS_transform
Test connection
Open
New
Skip line count
Options
Allow schema drift
Validate schema

STEP 5: After successfully creating dataflow click on dataflow debug.

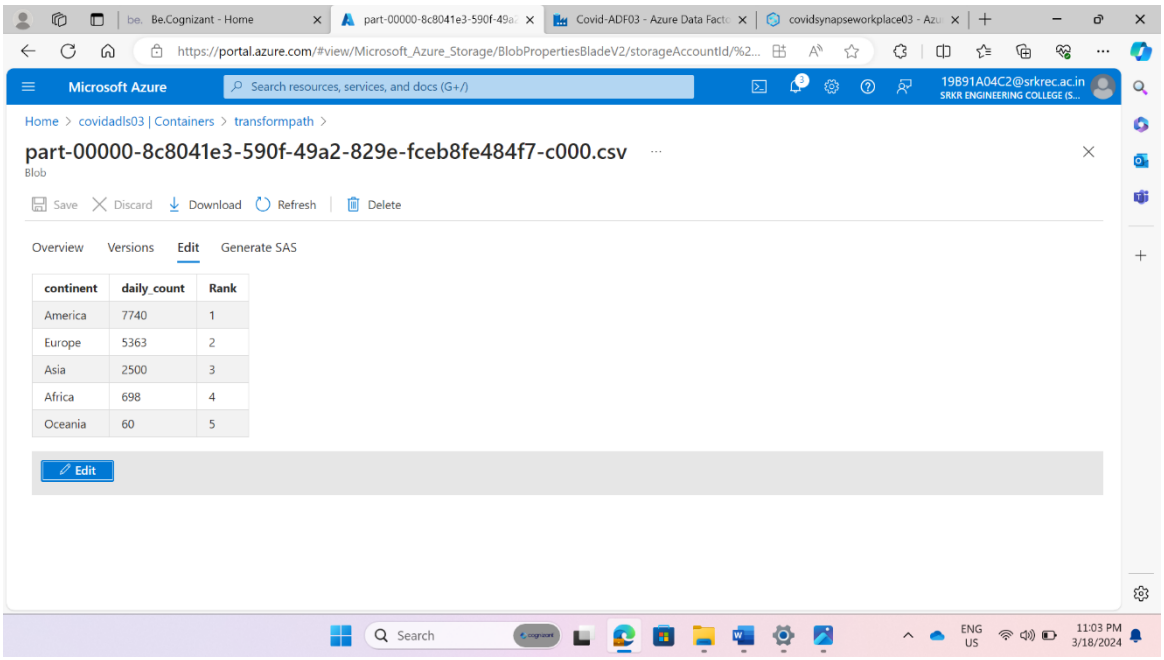


STEP 6: After clicking debug, drag the dataflow in pipeline and run the pipeline. you will get the output once it is run successfully.



REQUIREMENT 2 OUTPUT:

STEP 7: After that we need to check file appear in the transformpath container in storage account (data lake). I successfully got that file in my container as per the question given in the project documentation.



MY AZURE DETAILS: