AZURE DATA FACTORY CAPSTONE – COVID USE CASE

Emp Id: 2319933
Name: Bandi Yuktha Reddy
Cohort Id: CSDAIA24AZ002

CONTENTS

Project Architecture Flow
Resource Requirements
Project work flow
Data Verification
Project Outcome

INTRODUCTION

In the wake of the global COVID-19 pandemic, the need for reliable data integration and analysis has become paramount. This capstone project leverages Azure Data Factory (ADF), a cloud-based data integration service, to orchestrate and automate the movement and transformation of COVID-19 data. The project aims to provide insightful analytics that can aid in understanding the spread of the virus, its impact on healthcare systems, and the effectiveness of public health interventions.

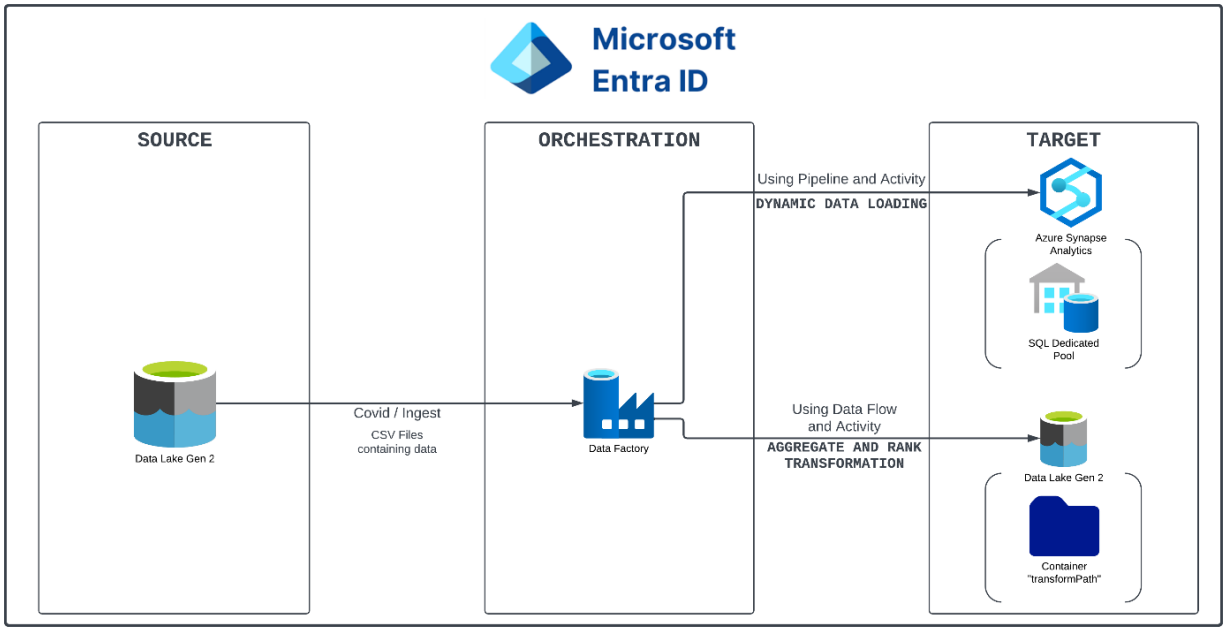
Utilizing ADF's robust capabilities, we will create a pipeline that sources up-to-date COVID-19 data from various repositories, processes it through custom transformations, and loads it into a centralized data store for analysis. This end-to-end solution will enable stakeholders to make informed decisions based on the latest trends and statistics of the pandemic.

This introduction sets the stage for the project, outlining its purpose and the technology used. It can be further expanded to detail specific objectives, data sources, and expected outcomes.

The purpose of the Covid use case exercise is 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, you will learn.

- How to ingest data from flat files into Azure Data Lake Gen2 and Azure Synapse using Azure Data Factory (ADF)
- How to transform data using Data Flows in Azure Data Factory (ADF) and load into Azure Synapse.

PROJECT ARCHITECTURE FLOW



RESOURCE REQUIREMENTS

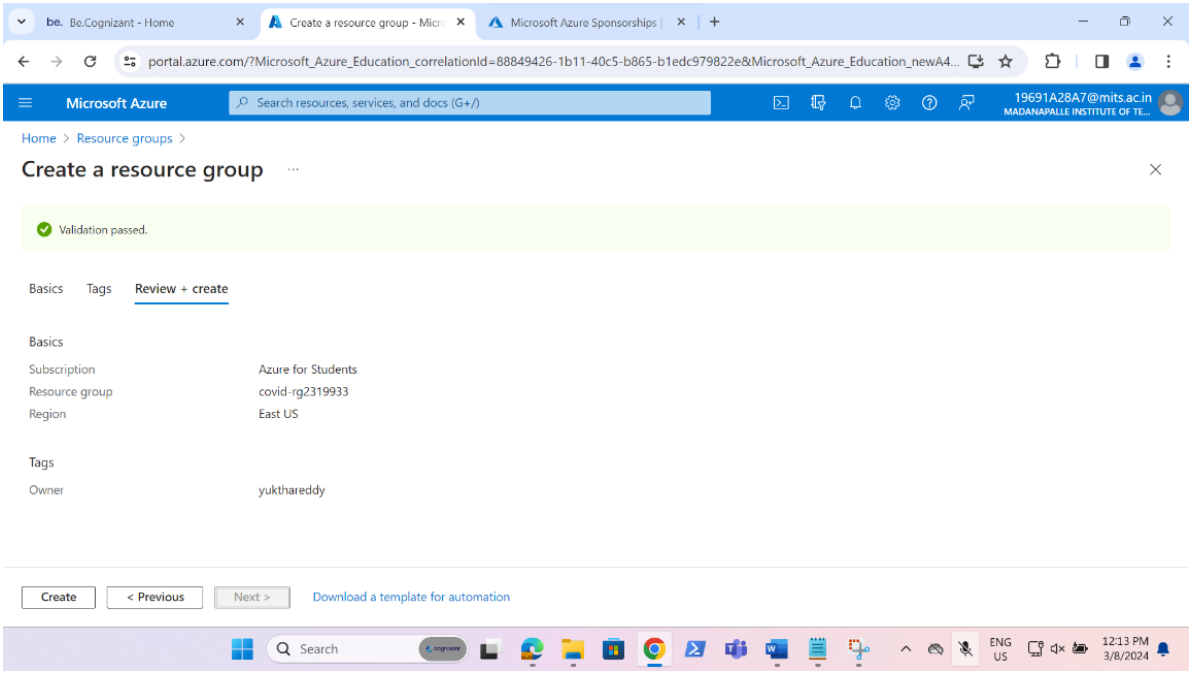
Azure Service	Azure Service Name
Resource Group	covid-rg
Azure Data Lake Storage Gen2 Account	Covid-adls
Azure Data Lake Storage Gen2 Container	Covid
Azure Synapse Workspace	Covid-Synapse-Workplace
Azure Data Factory	Covid-ADF

PROJECT WORKFLOW

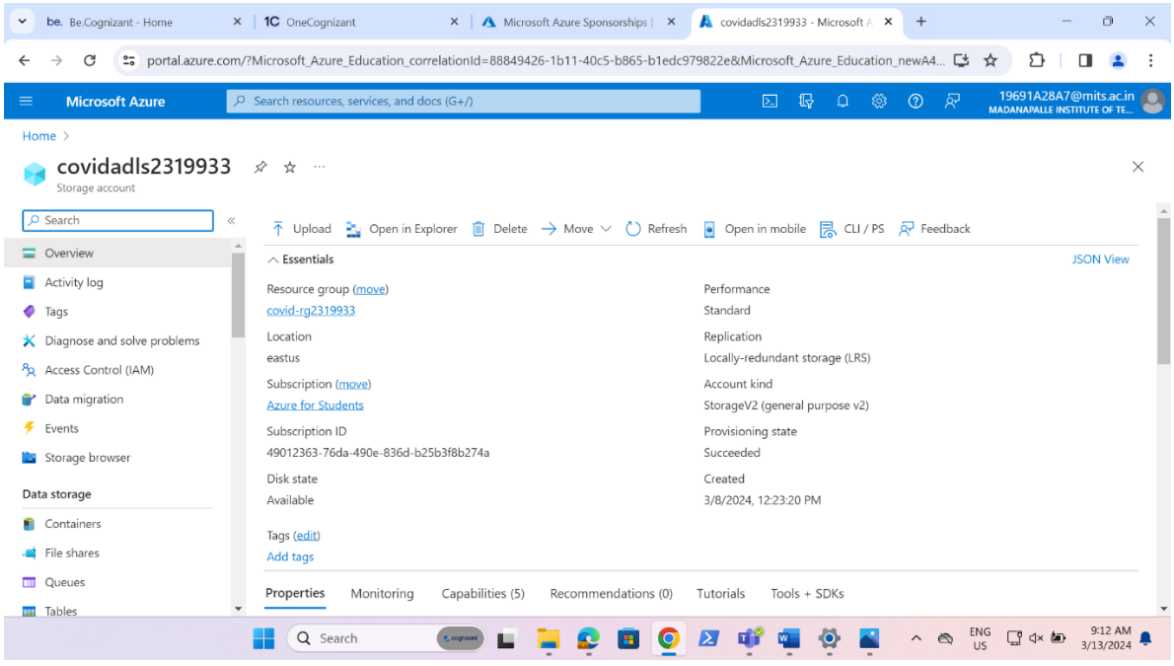
- This project consists of two requirements.

Procedure for Requirement 1:

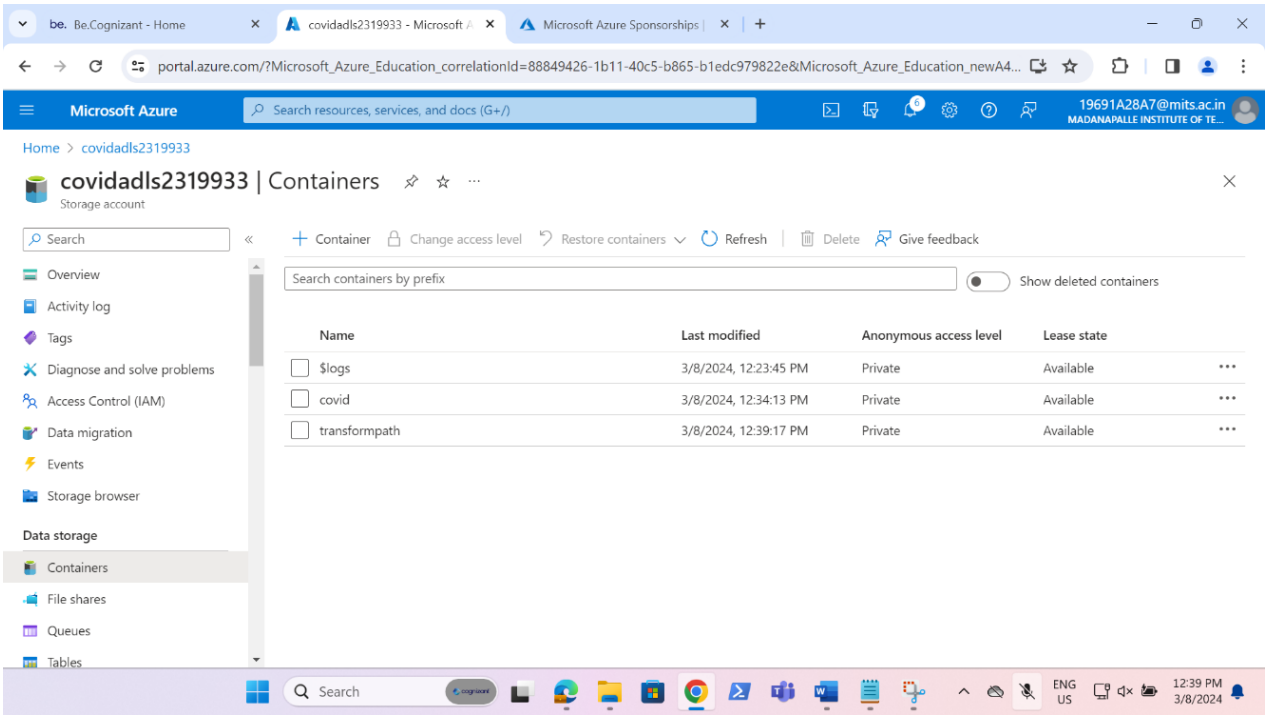
Step1: Create a Resource group(covid-rg2319933)



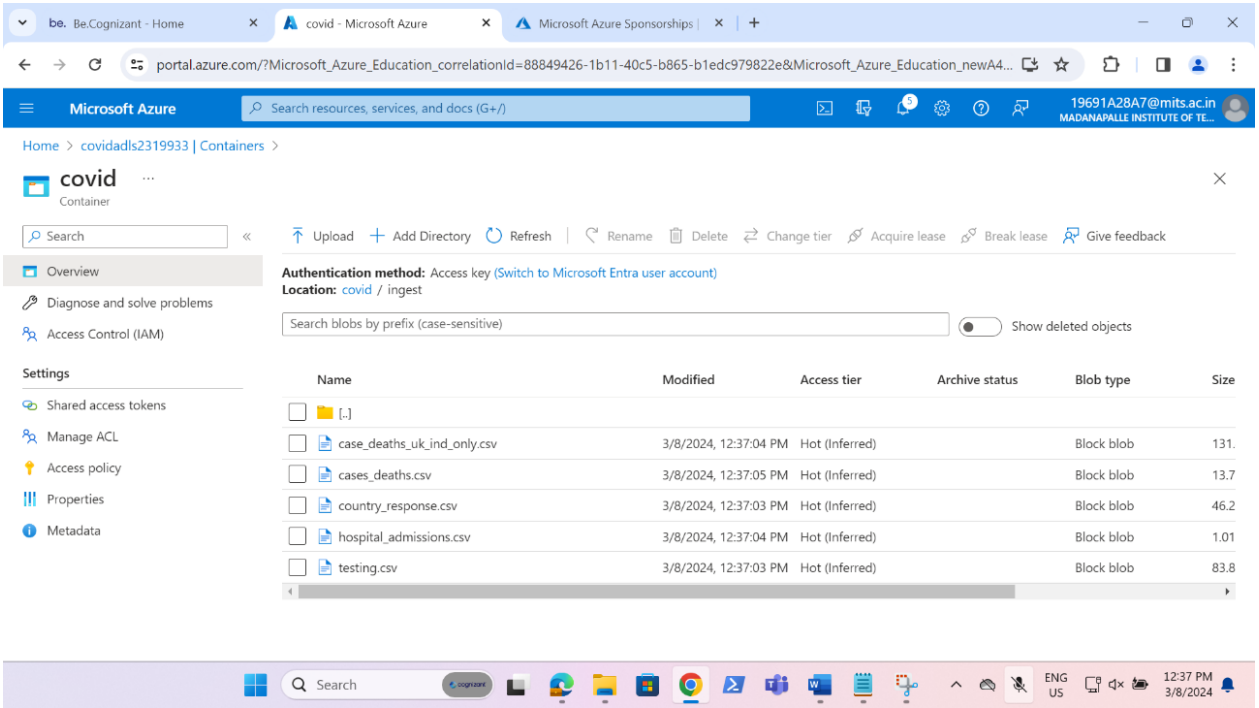
Step2: Create an Azure Datalake storage (covidadls2319933)



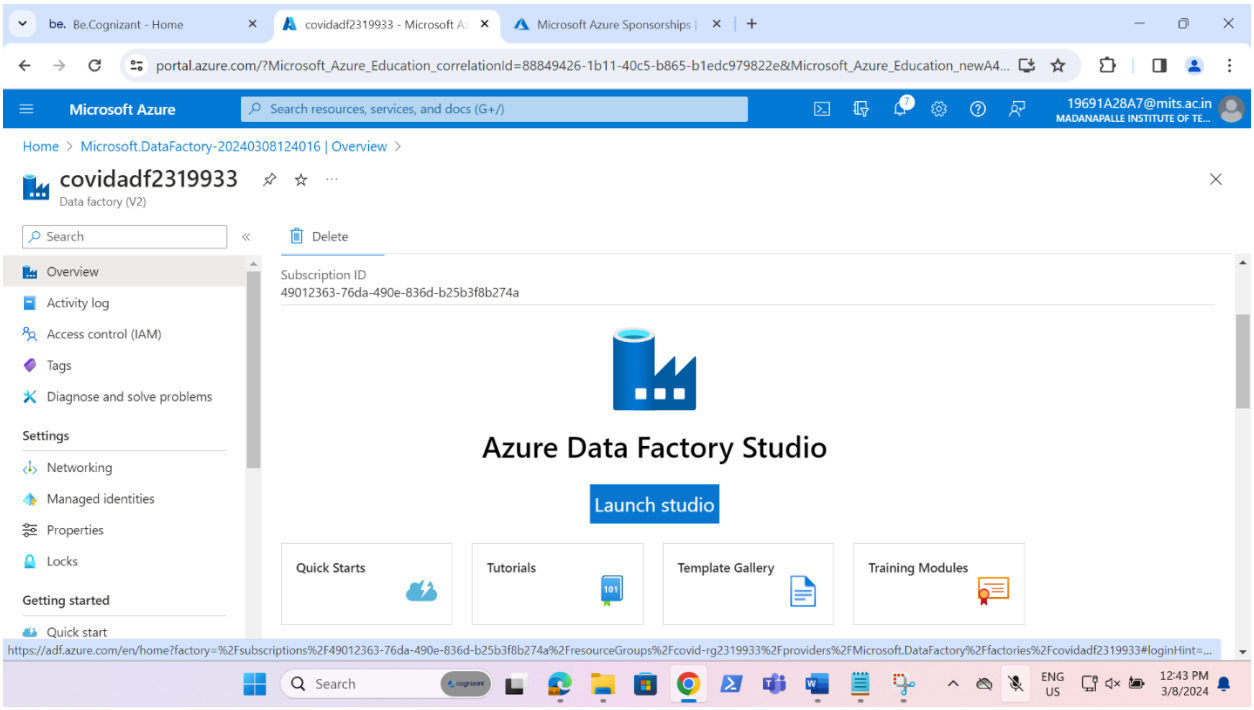
Step3: Create two containers(covid and transformpath) within adls.



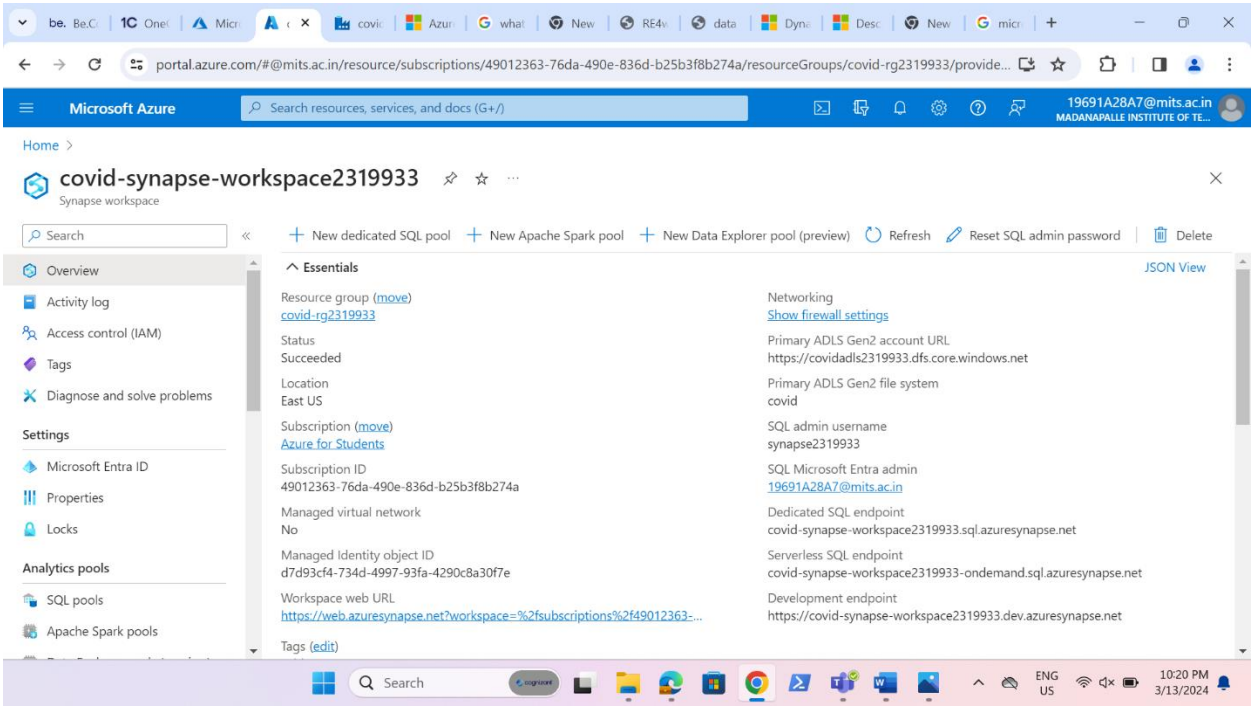
Step4: Create a directory(ingest) in covid container and add files into the container.



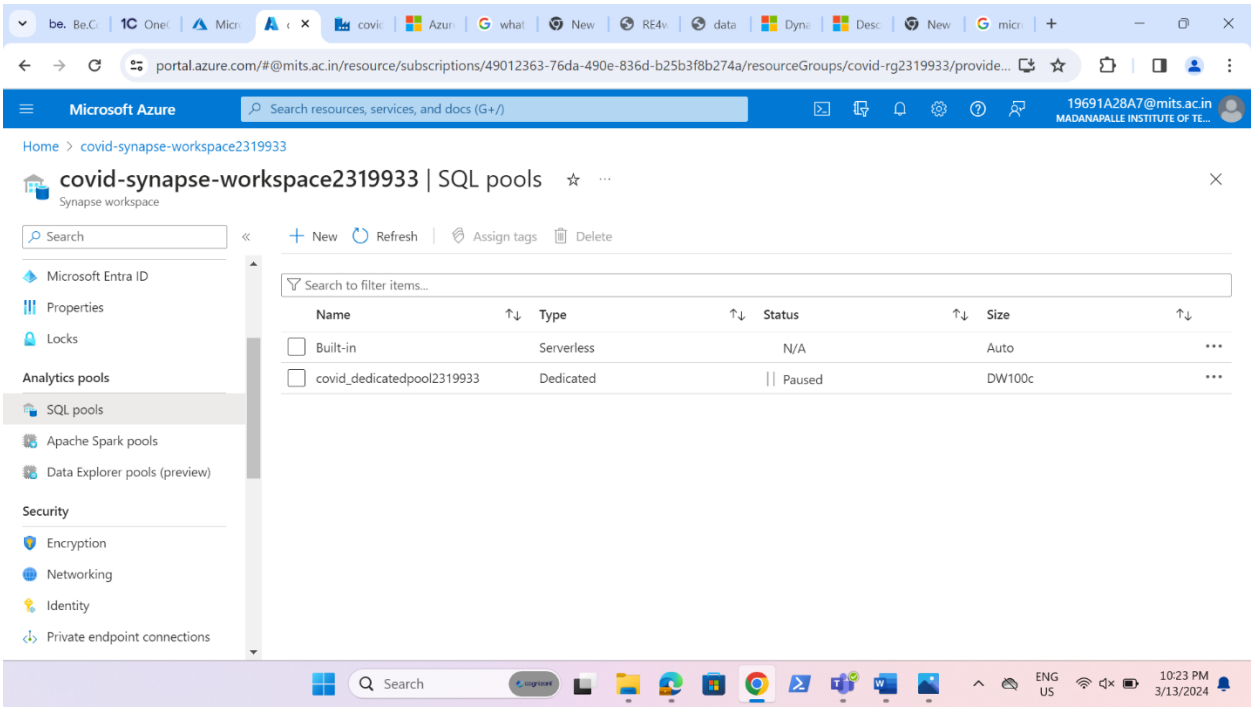
Step5: Create an Azure Data Factory (covidadf2319933)



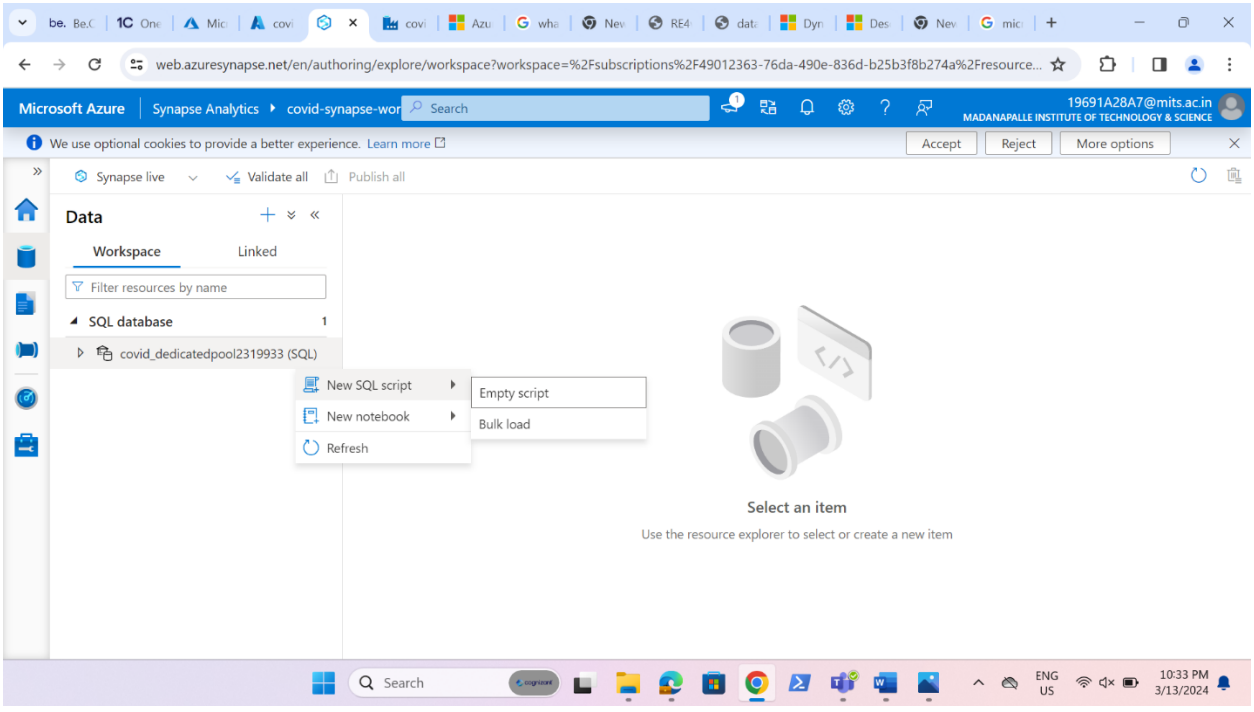
Step6: Create a Synapse workspace(covid-synapse-workspace2319933)



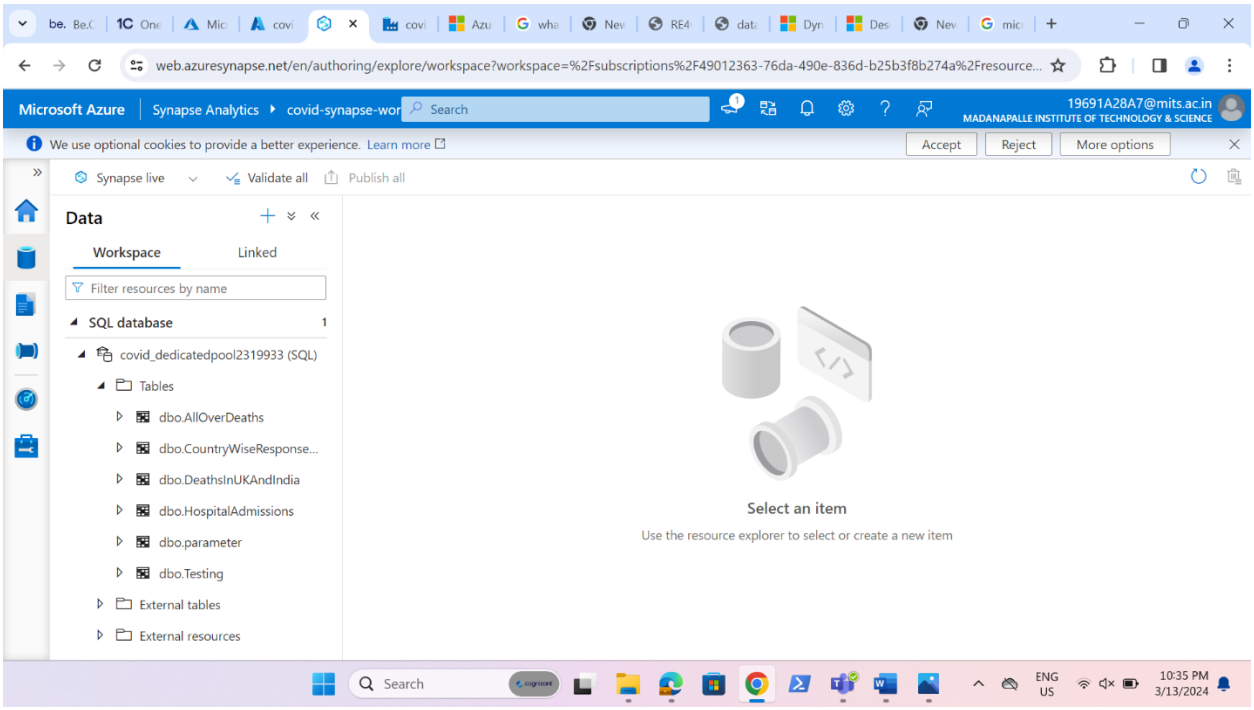
Step7: Create a dedicated pool(covid_dedicatedpool2319933).



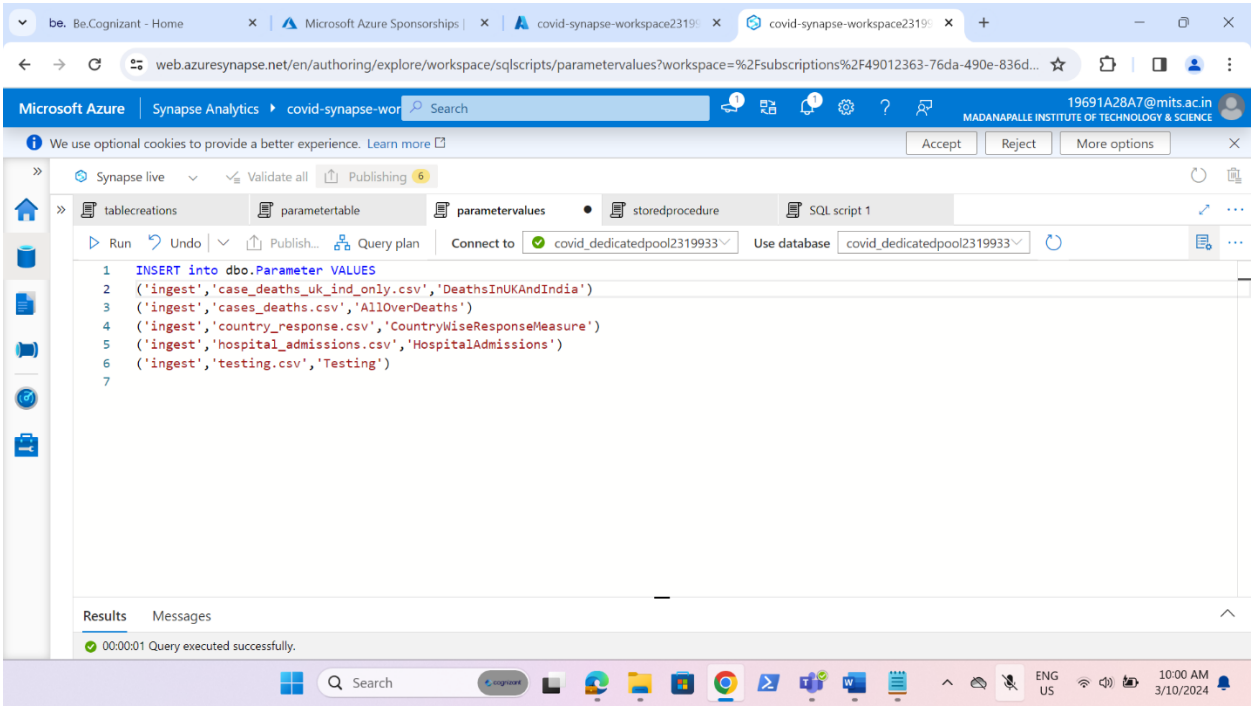
Step8: Open the dedicated pool.Open the data in the left panel and select new SQL script to create table within the dedicated pool.



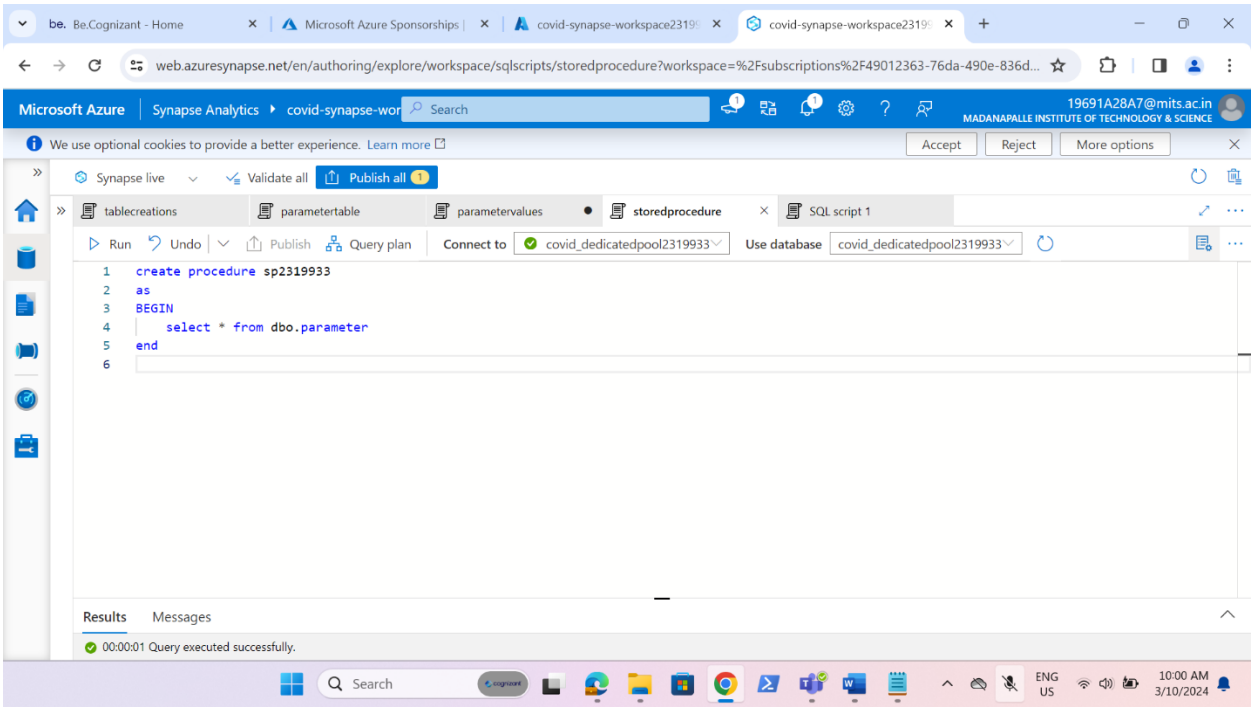
Step9: Create five tables in the sql scripts and run each of them.



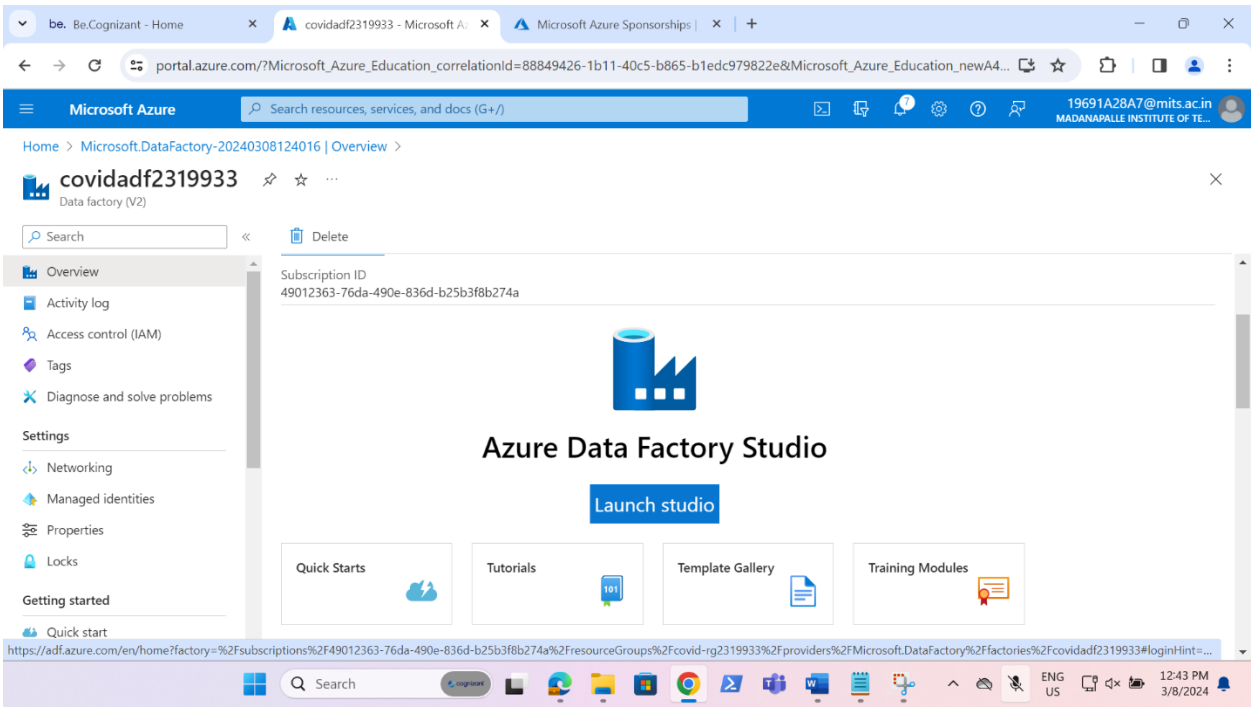
Step10: Create a table(parameter) to store all these 5 table schemas which is used for dynamic loading of data.



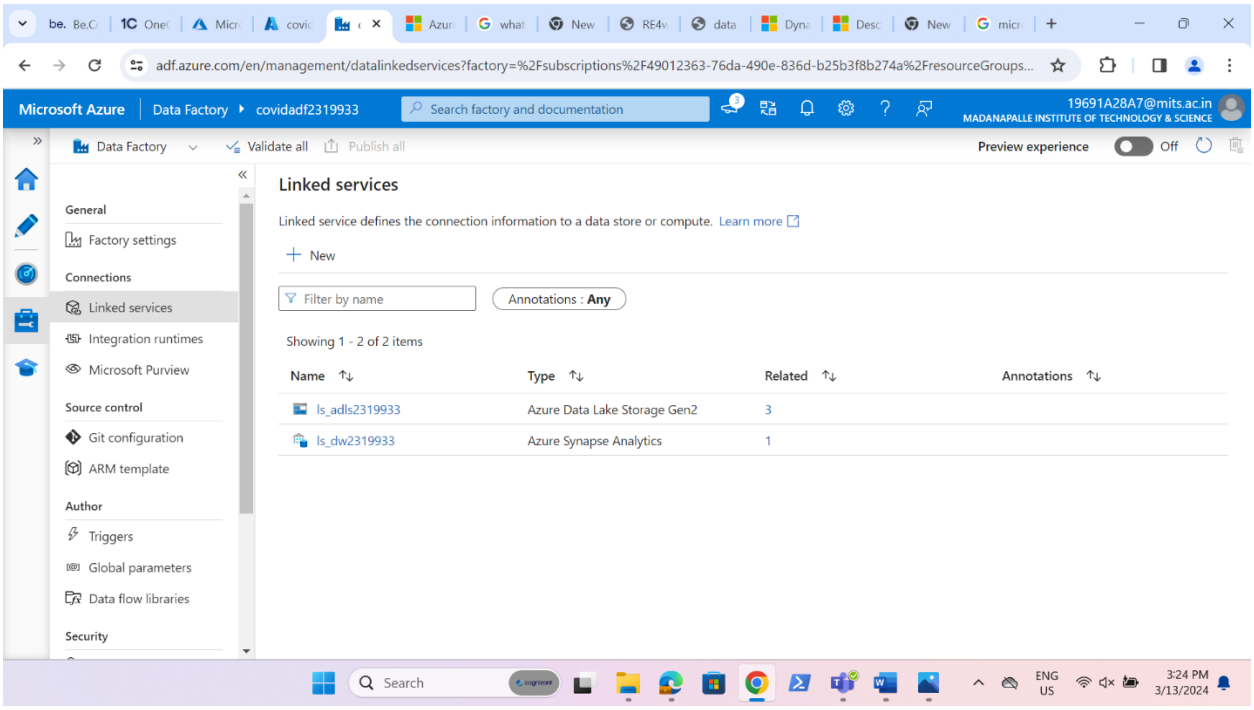
Step11: Create a stored procedure in Programability which is present in left panel.



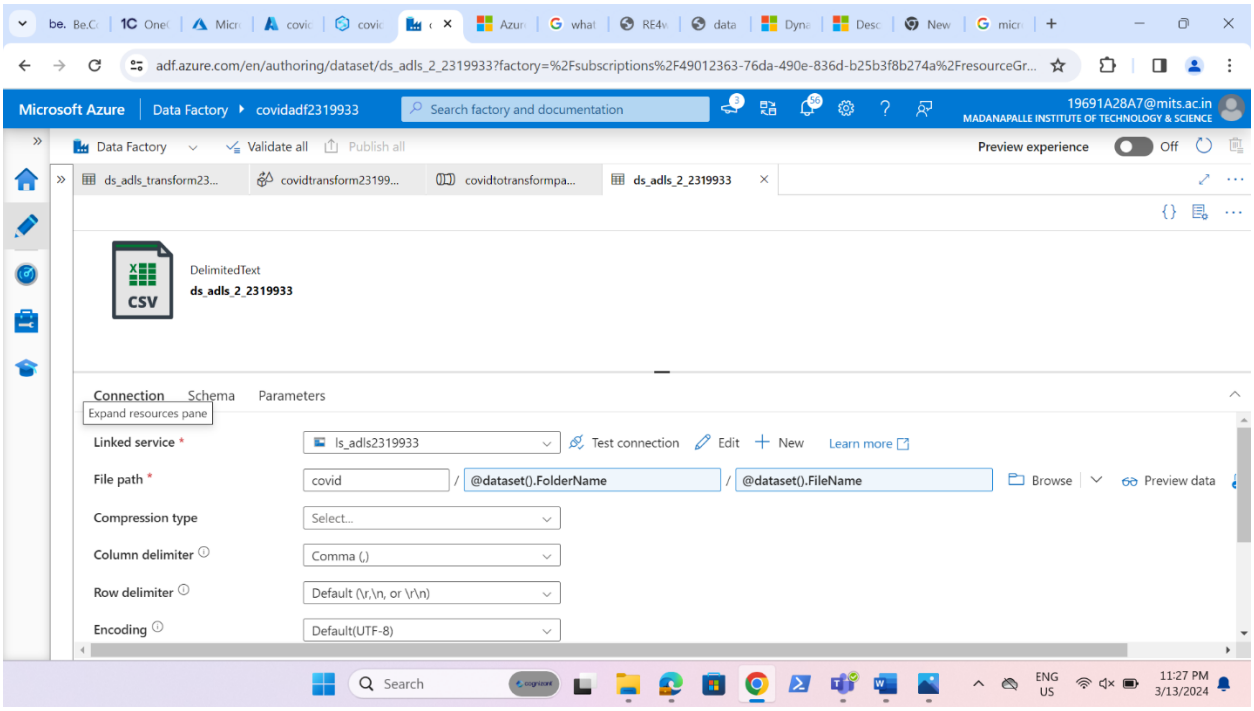
Step12: Launch the Azure Data Factory Studio (covidadf2319933)

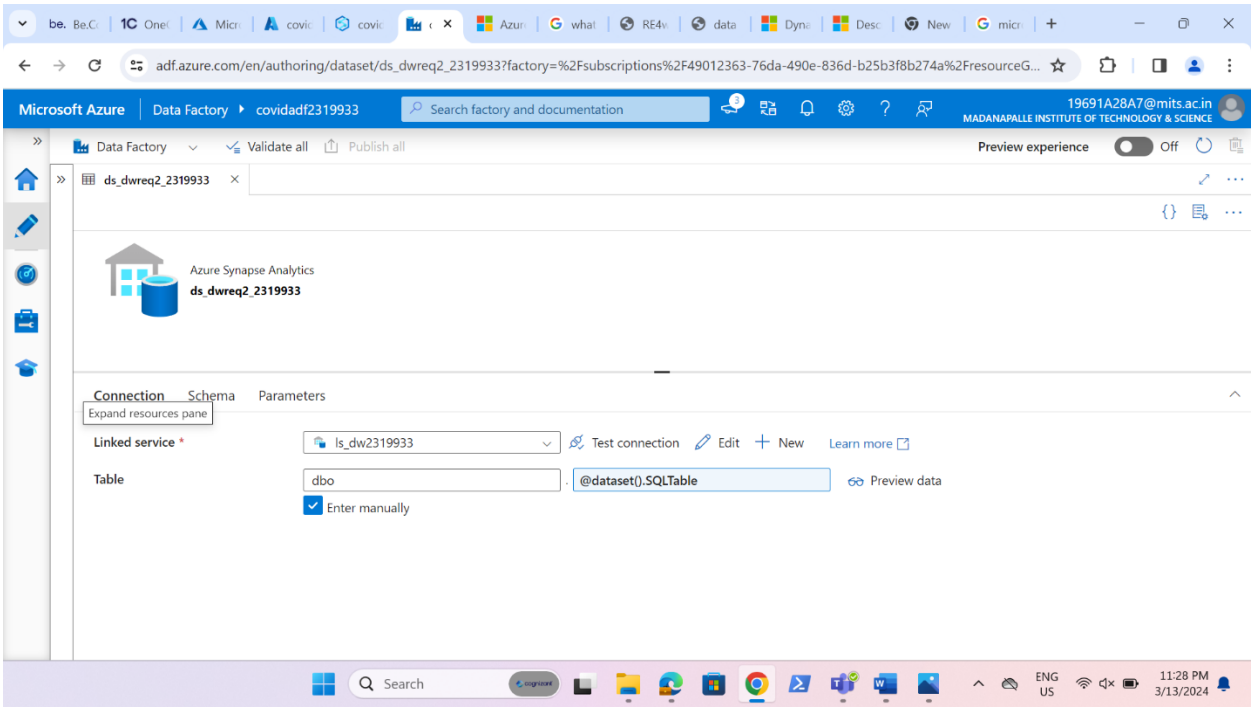


Step13: Create one Linked service (ls_dw2319933) for synapse workspace in manage.

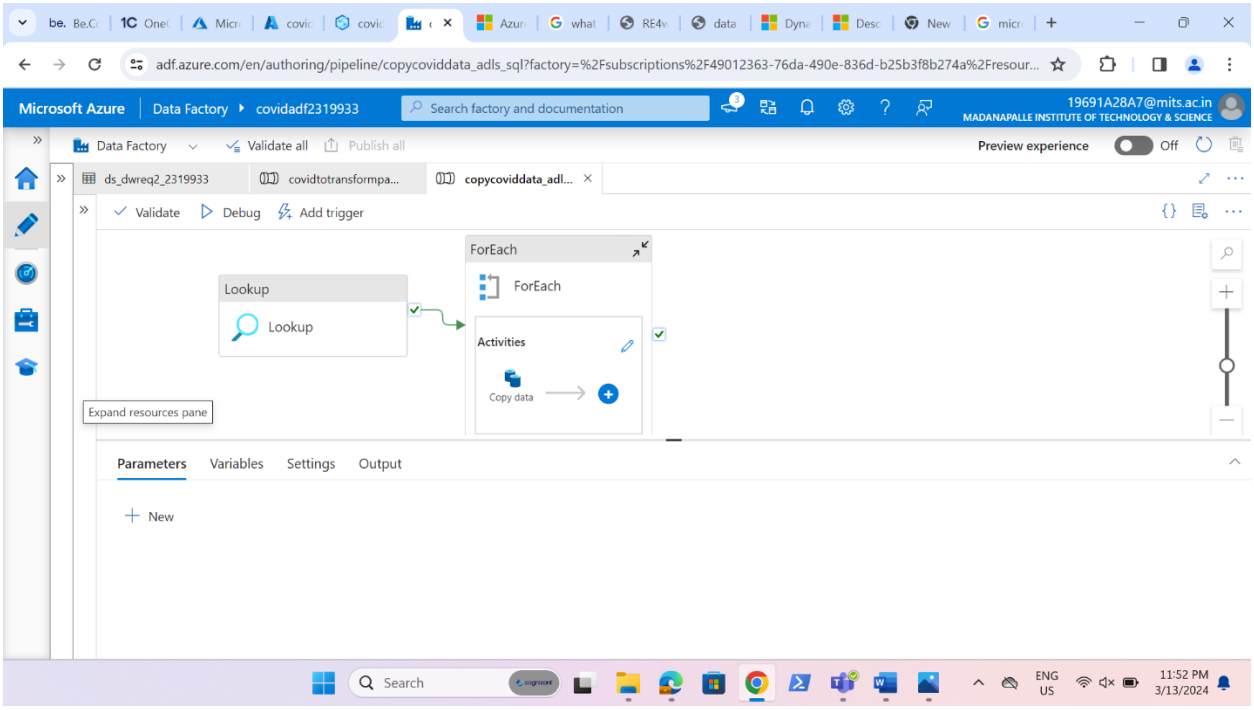


Step14: Create two datasets one for adls(ds_adls_2_2319933) and for synapse Workspace (ds_dwreq2_2319933).

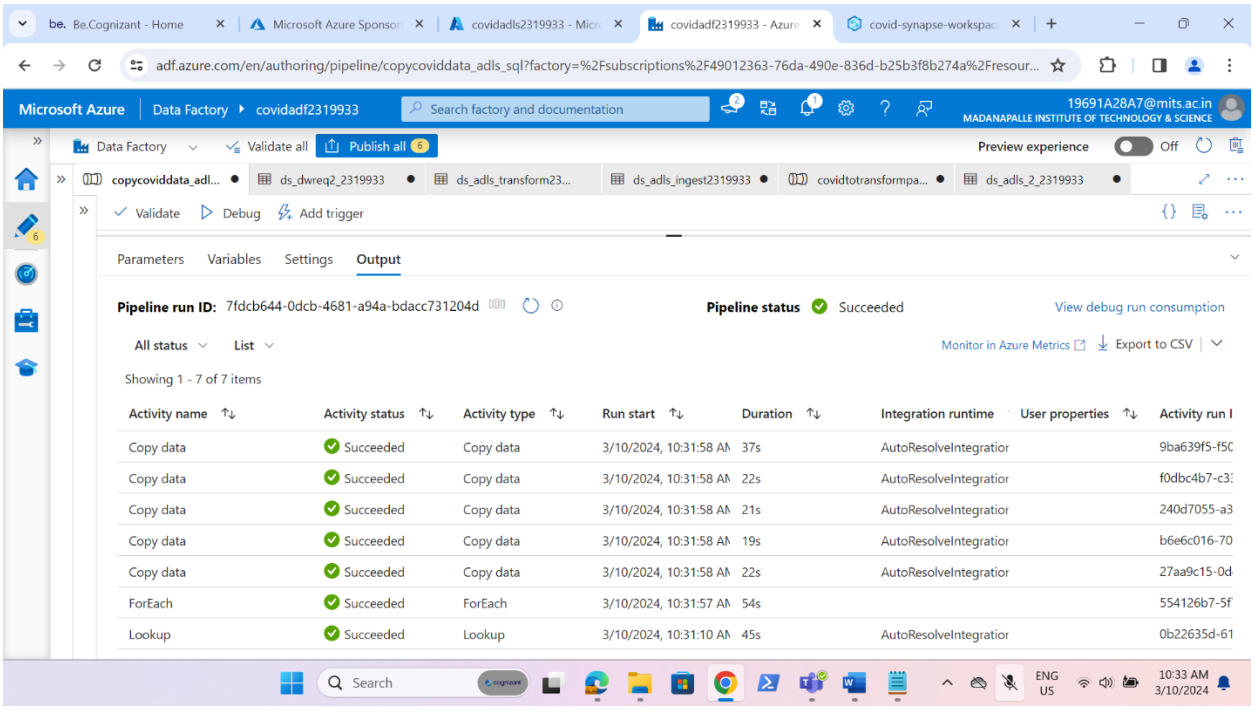




Step15: Create a pipeline(copycoviddata_adls_synapse) that will able to copy the datasets in ADLSGen2 dynamically in a loop.

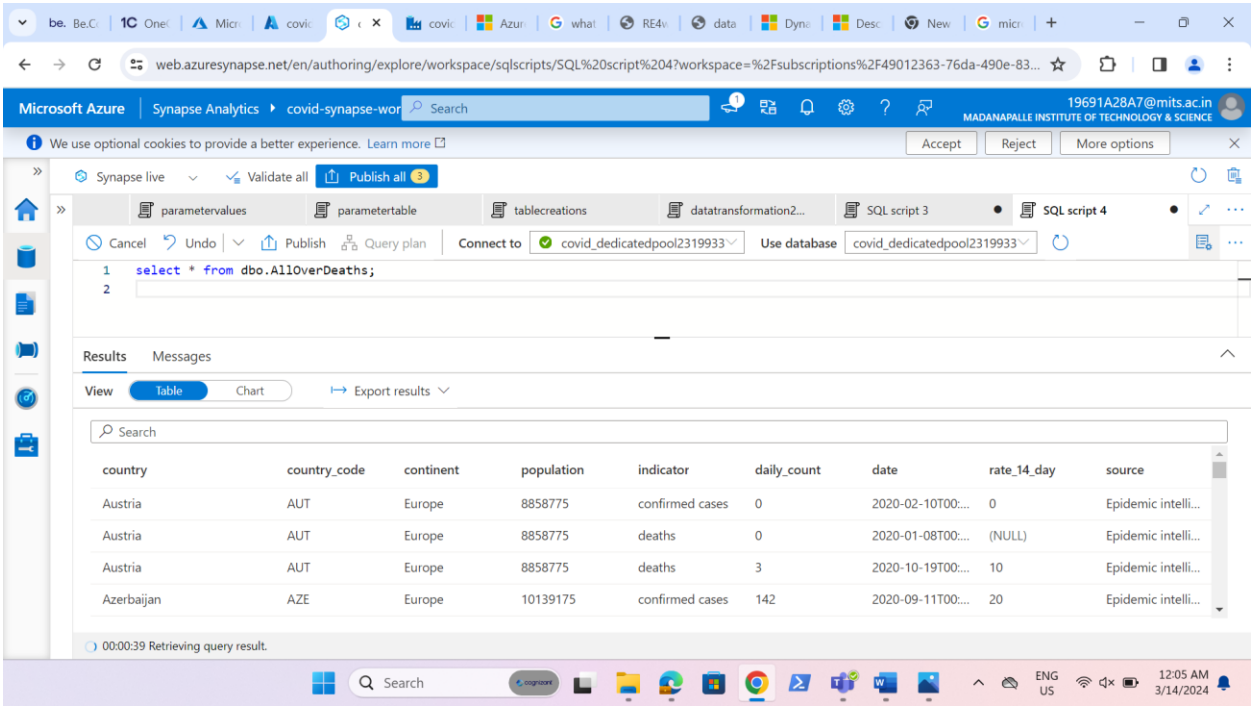


Step16: Now debug the pipeline and make sure the dedicated pool which was created before should not pause it should be on online mode.



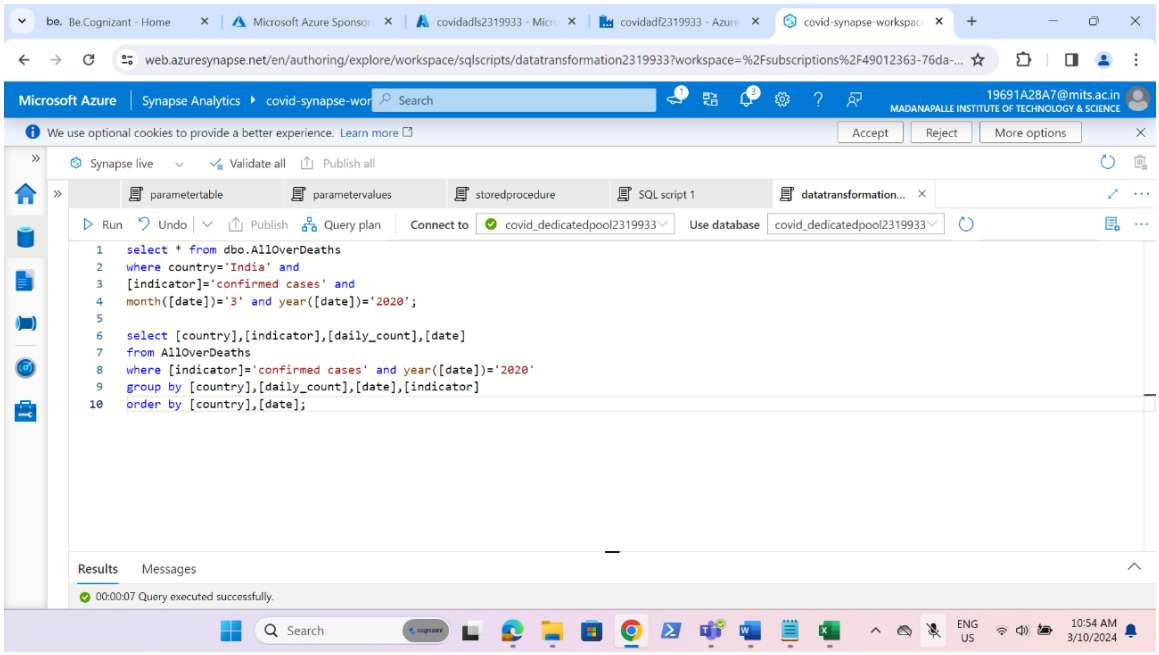
Step17: On successful completion of debugging check whether all the files are loaded in the sink or not.

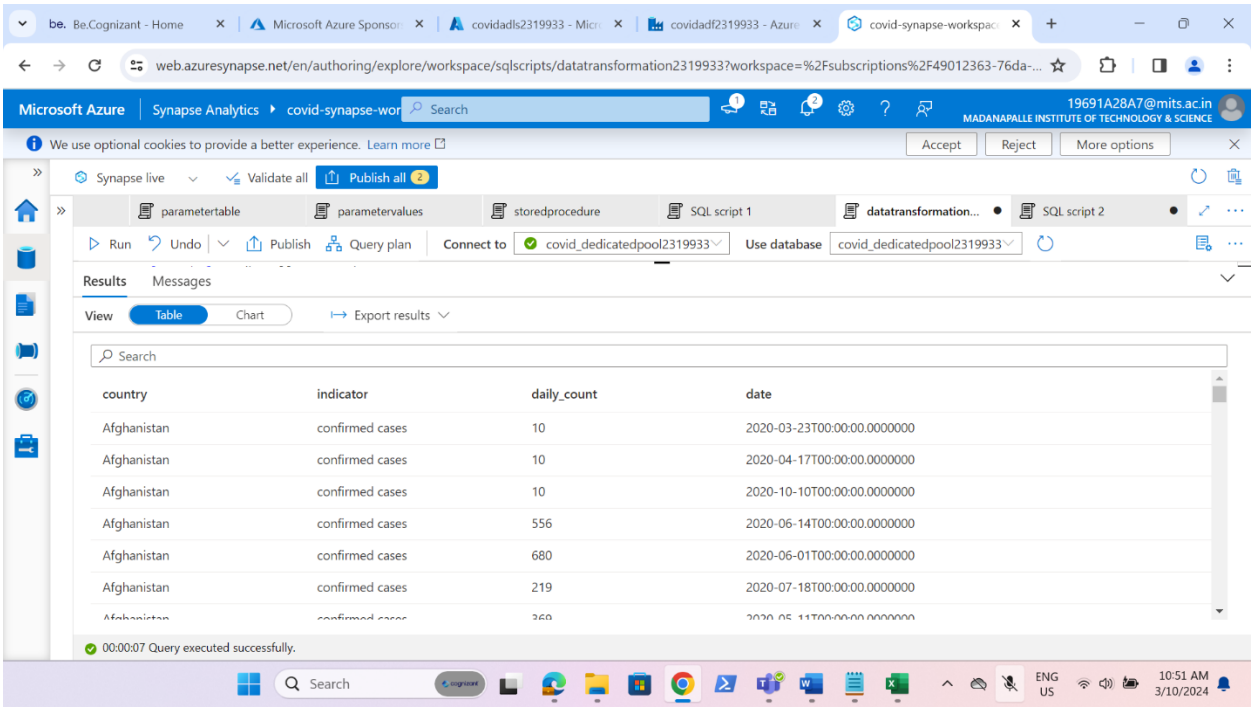
To check this write some queries in synapse workspace.

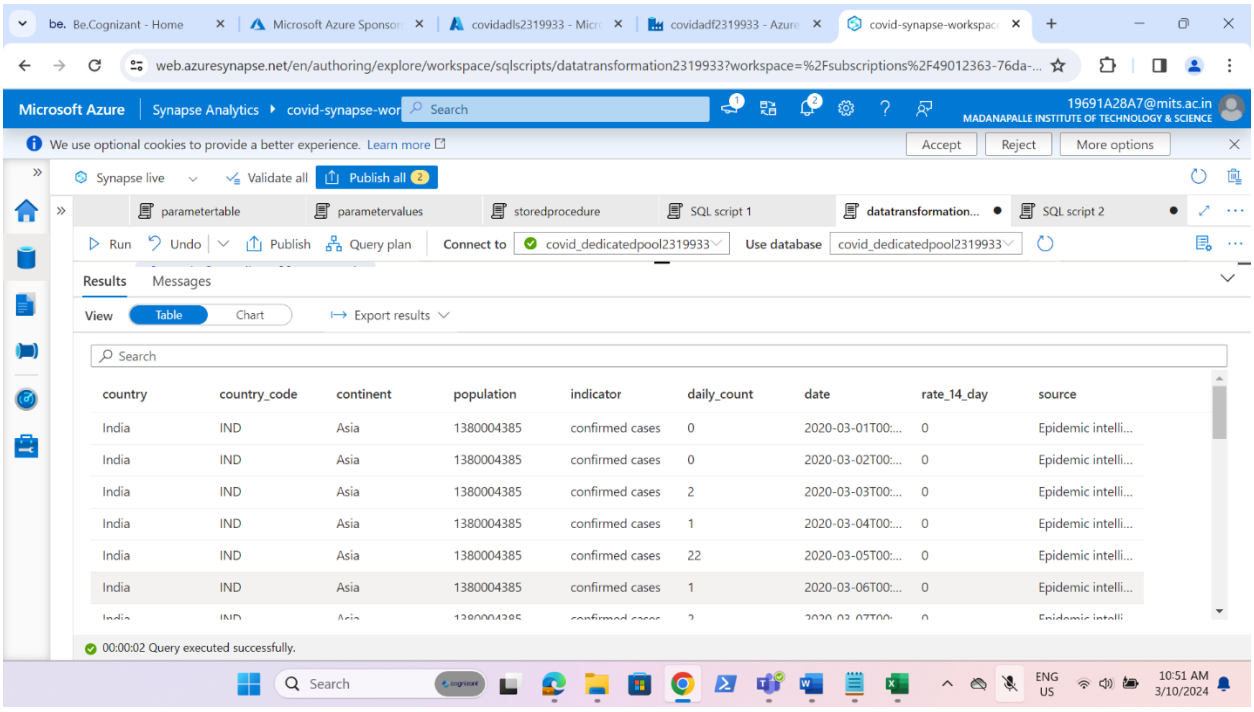


DATA VERFICATION:

- Write a Query to check the confirmed cases in India in March'2020 and verify with source data in excel.
- Write a query to show country wise confirmed case in 2020.

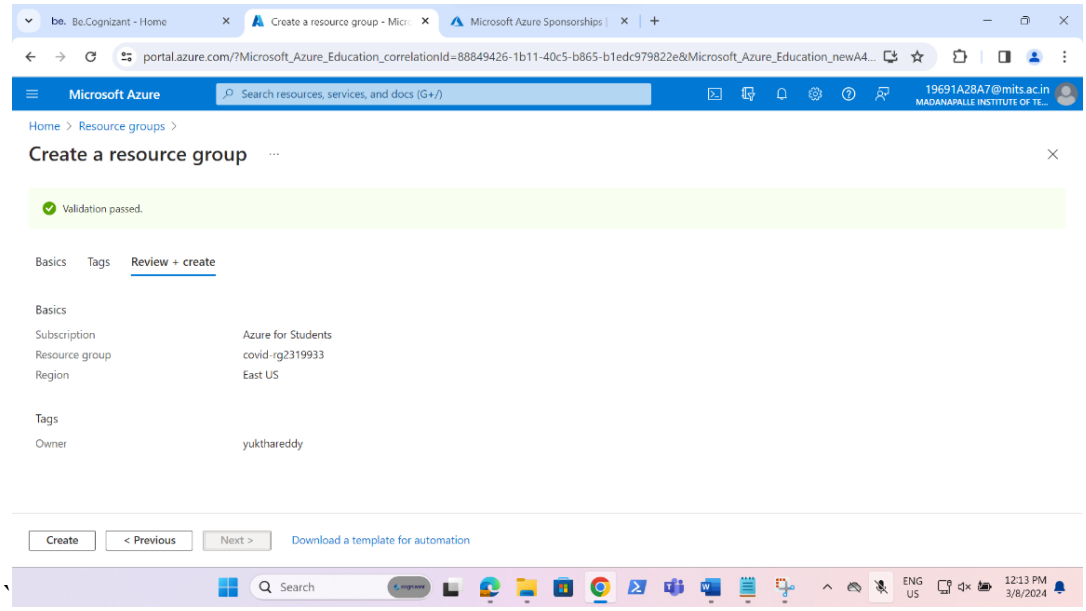




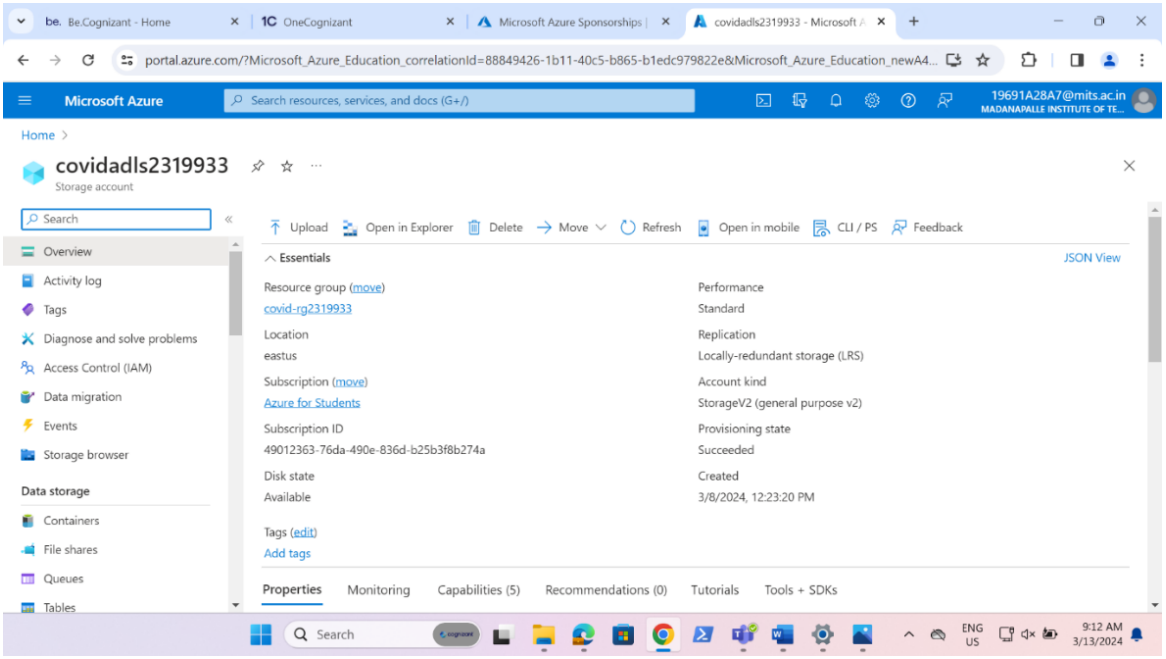

Microsoft Azure
Synapse Analytics
covid-synapse-wor
Search
19691A28A7@mits.ac.in
MADANAPALLE INSTITUTE OF TECHNOLOGY & SCIENCE
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse live
Validate all
Publish all 2
parametertable
parametervalues
storedprocedure
SQL script 1
datatransformation...
SQL script 2
Run
Undo
Publish
Query plan
Connect to
covid_dedicatedpool2319933
Use database
covid_dedicatedpool2319933
Results
Messages
View
Table
Chart
Export results
Search
country
country_code
continent
population
indicator
daily_count
date
rate_14_day
source
India IND Asia 1380004385 confirmed cases 0 2020-03-01T00:... 0 Epidemic intelli...
India IND Asia 1380004385 confirmed cases 0 2020-03-02T00:... 0 Epidemic intelli...
India IND Asia 1380004385 confirmed cases 2 2020-03-03T00:... 0 Epidemic intelli...
India IND Asia 1380004385 confirmed cases 1 2020-03-04T00:... 0 Epidemic intelli...
India IND Asia 1380004385 confirmed cases 22 2020-03-05T00:... 0 Epidemic intelli...
India IND Asia 1380004385 confirmed cases 1 2020-03-06T00:... 0 Epidemic intelli...
00:00:02 Query executed successfully.
Search
ENG
US
10:51 AM
3/10/2024

Procedure of Requirement 2:

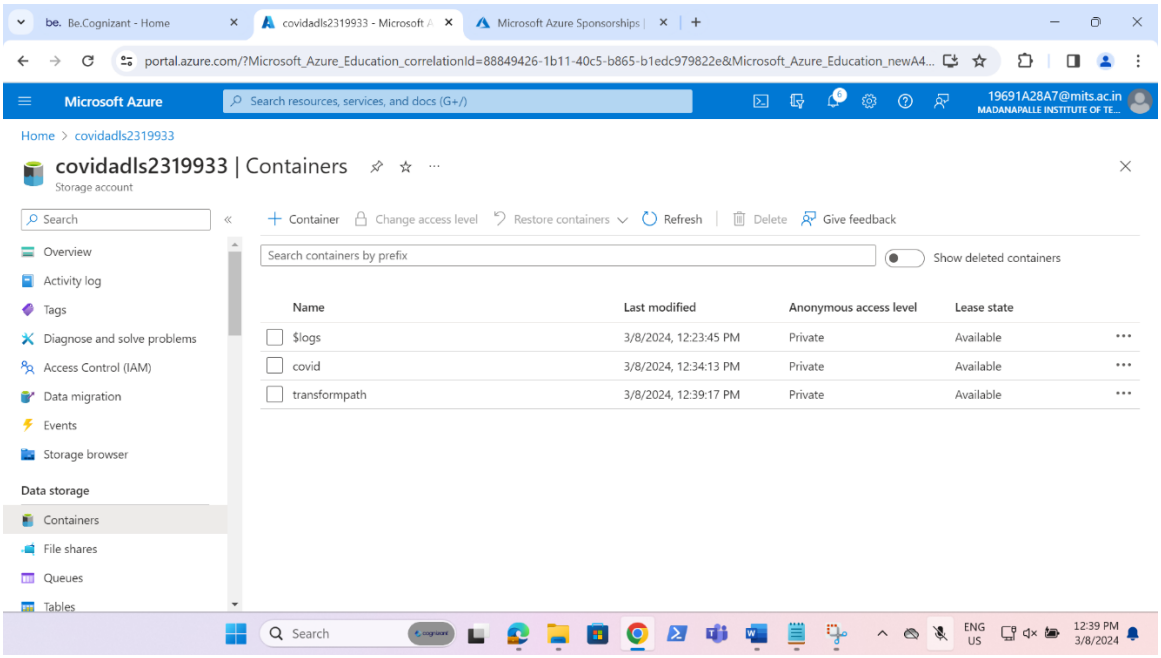
Step1: Create a Resource group(covid-rg2319933)



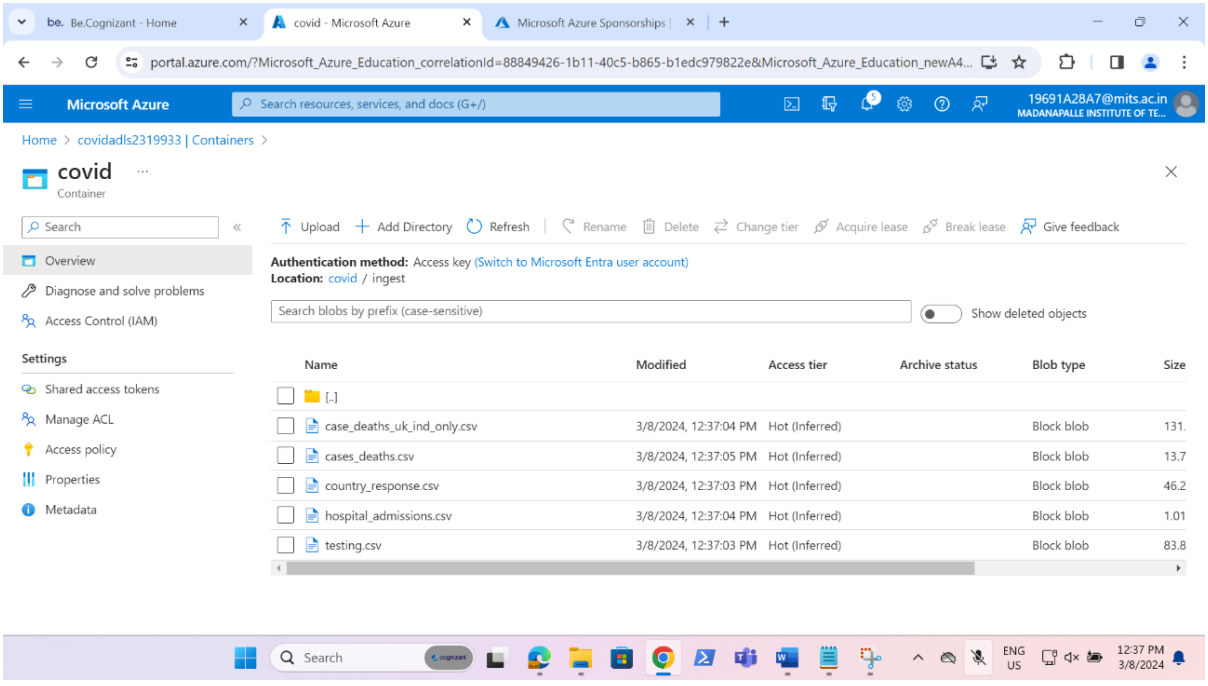
Step2: Create an Azure Datalake storage (covidadls2319933)



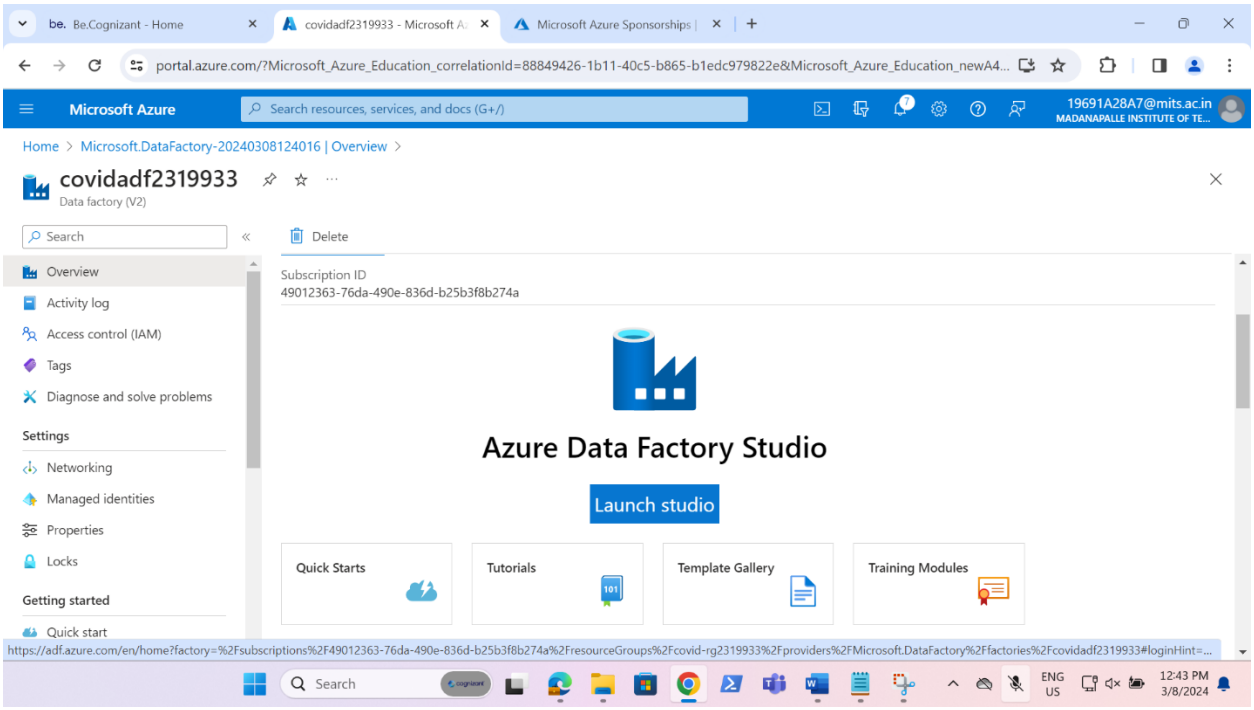
Step3: Create a container(covid) within adls storage.



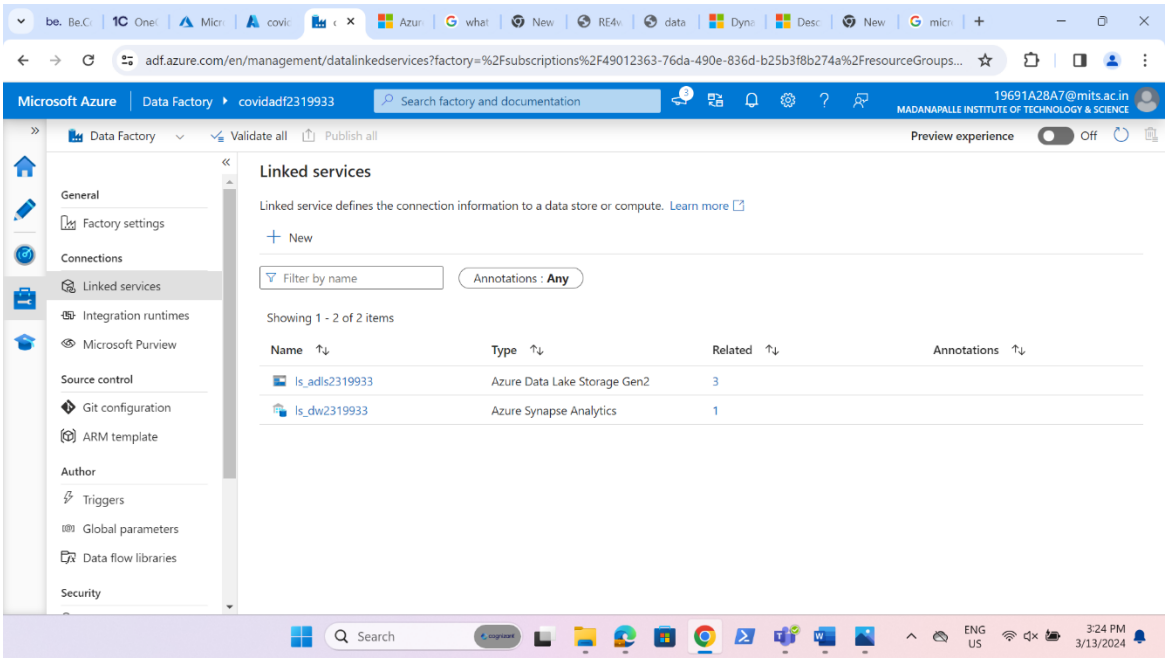
Step4: Create a directory(ingest) in covid container and add files into the container.



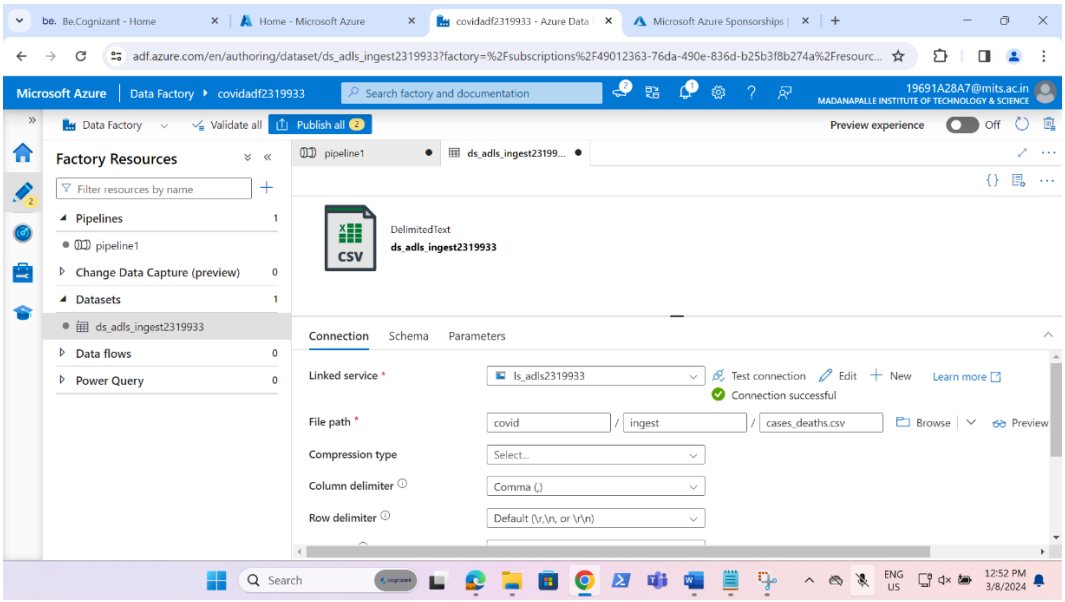
Step5: Create an Azure Data Factory (covidadf2319933)

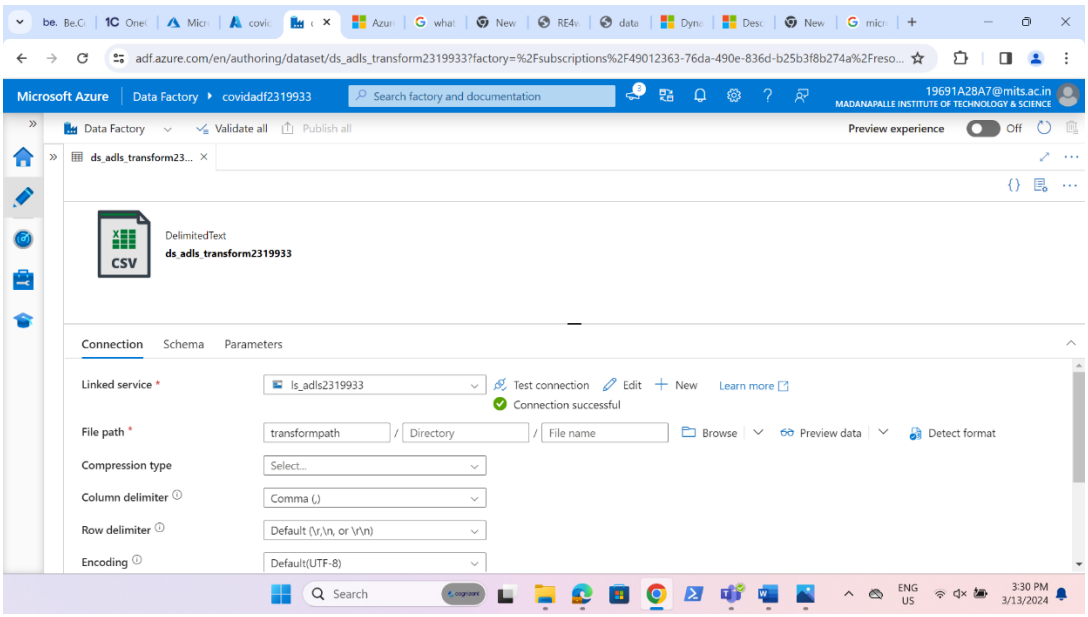


Step6: Open Azure Data Factory Studio create one Linked service (ls_adls2319933) for Adls in manage.

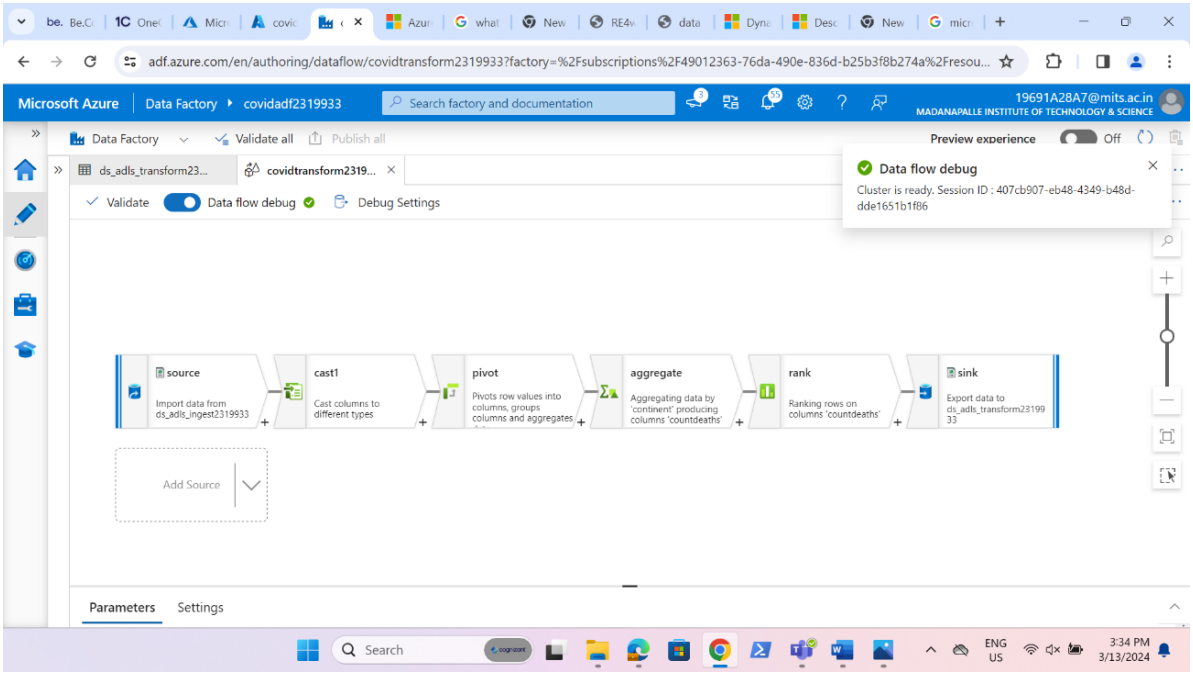


Step7: Create two datasets,one for source path and one for sink path(ds_adls_ingest2319933) and (ds_adls_transform2319933) in author.

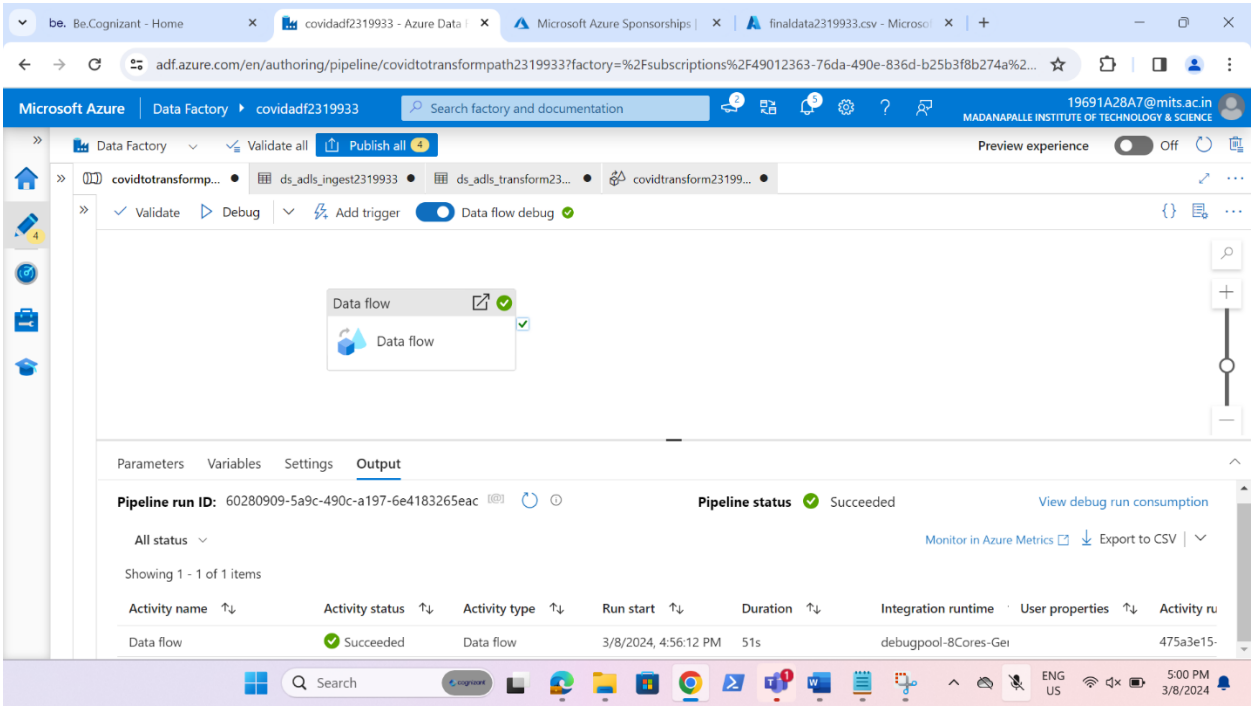




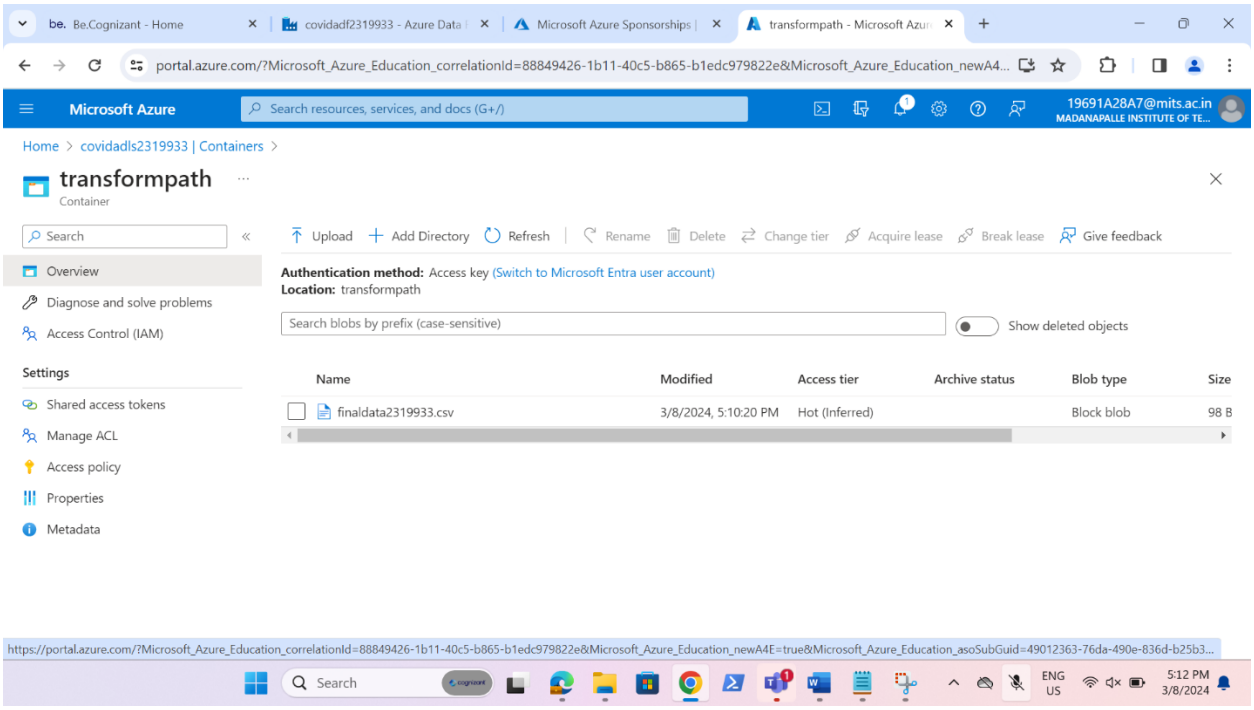
Step8: Create a Dataflow(covidtransform2319933) and after creating turn on the dataflow debug.



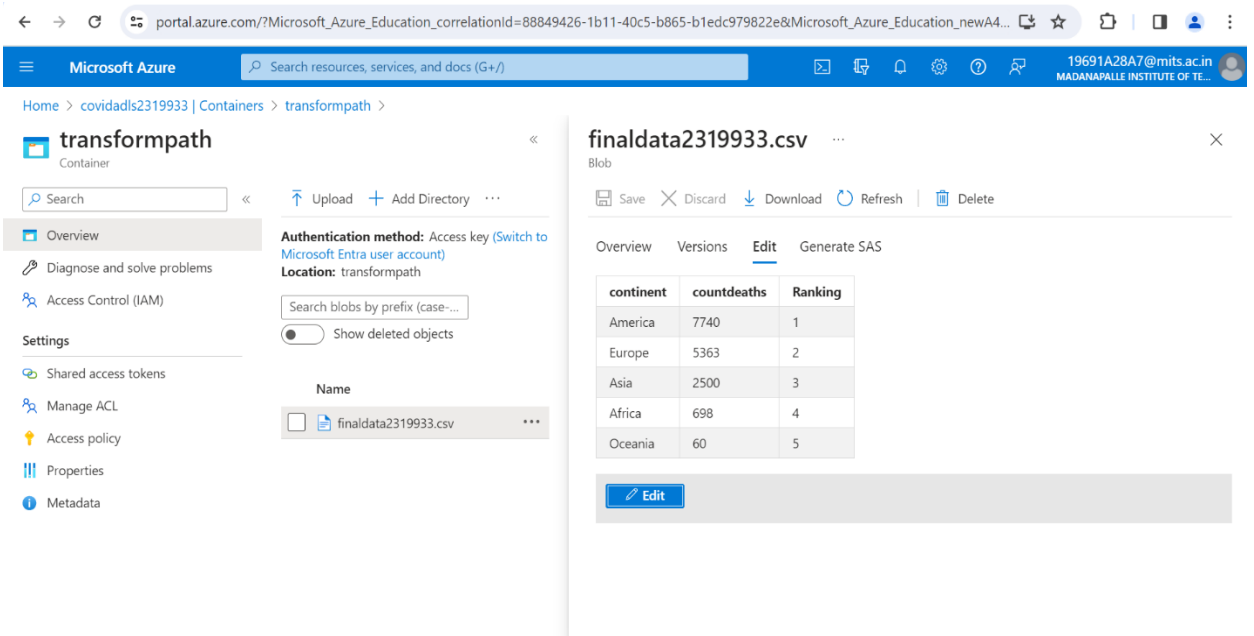
Step9: Create a Pipeline(covidtotransformpath2319933), drag and drop the dataflow activity and select the above dataflow and debug it.



Step10: After successful completion of debugging check whether the file uploaded in transformpath or not.



DATA VERFICATION:



continent	countdeaths	Ranking
America	7740	1
Europe	5363	2
Asia	2500	3
Africa	698	4
Oceania	60	5

Queries used in Project:

5 SQL QUERIES:

```
create table Testing(
country Varchar(1000),
country_code Varchar(1000),
year_week Varchar(1000),
new_cases BigInt,
tests_done BigInt,
population BigInt,
testing_rate Decimal,
positivity_rate Decimal,
testing_data_source Varchar(2000)
);

create table HospitalAdmissions
(country Varchar(1000),
indicator Varchar(1000),
date Date,
year_week Varchar(1000),
value Decimal,
source Varchar(1000),
url Varchar(2000)
);
```

```
create table CountryWiseResponseMeasure(
Country varchar(100),
Response_measure varchar(100),
change int,
date_start date,
date_end varchar(100)
);

create table AllOverDeaths(country Varchar(100),
country_code Varchar(100),
continent Varchar(100),
population BigInt,
indicator Varchar(100),
daily_count BigInt,
date date,
rate_14_day Decimal,
source Varchar(100)
);

create table DeathsInUKAndIndia (country Varchar(100),
country_code Varchar(100),
continent Varchar(100),
population BigInt,
indicator Varchar(100),
daily_count BigInt,
```

```
date date,
rate_14_day Decimal,
source Varchar(100)
);
```

Parameter Table schema:

```
create table parameter
(
   FolderName VARCHAR(200),
   [FileName] VARCHAR(200),
   SQLTable  VARCHAR(200)
);
INSERT into dbo.Parameter VALUES
('ingest','case_deaths_uk_ind_only.csv','DeathsInUKAndIndia')
('ingest','cases_deaths.csv','AllOverDeaths')
('ingest','country_response.csv','CountryWiseResponseMeasure')
('ingest','hospital_admissions.csv','HospitalAdmissions')
('ingest','testing.csv','Testing')
```

Stored Procedure Schema:

```
create procedure sp3final
as
BEGIN
   select * from dbo.parameter
end
```

Data Verification queries:

```
select * from dbo.AllOverDeaths
where country='India' and
[indicator]='confirmed cases' and
month([date])='3' and year([date])='2020';

select [country],[indicator],[daily_count],[date]
from AllOverDeaths
where [indicator]='confirmed cases' and year([date])='2020'
group by [country],[daily_count],[date],[indicator]
order by [country],[date];
```