AZURE DATA FACTORY

CAPSTONE – COVID USE CASE

Emp id : 2320074

Cohort id : CSDAIA24AZ003

Introduction

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.

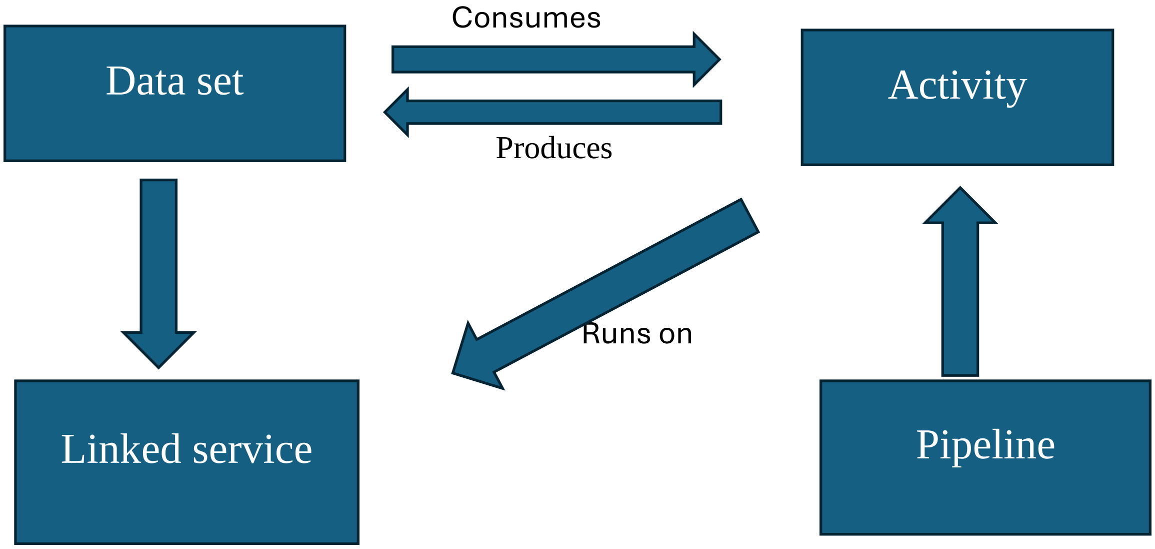
- How to ingest data from flat files into Azure Data Lake Gen2 and Azure Synapse using Azure Data Factory (ADF)
- How to transform data using Data Flows in Azure Data Factory (ADF) and load into Azure Synapse

Through this exercise, you will be having a hands-on experience on Storage, ADF Pipeline, Mapping Dataflow, Azure Synapse.

Purpose of the project

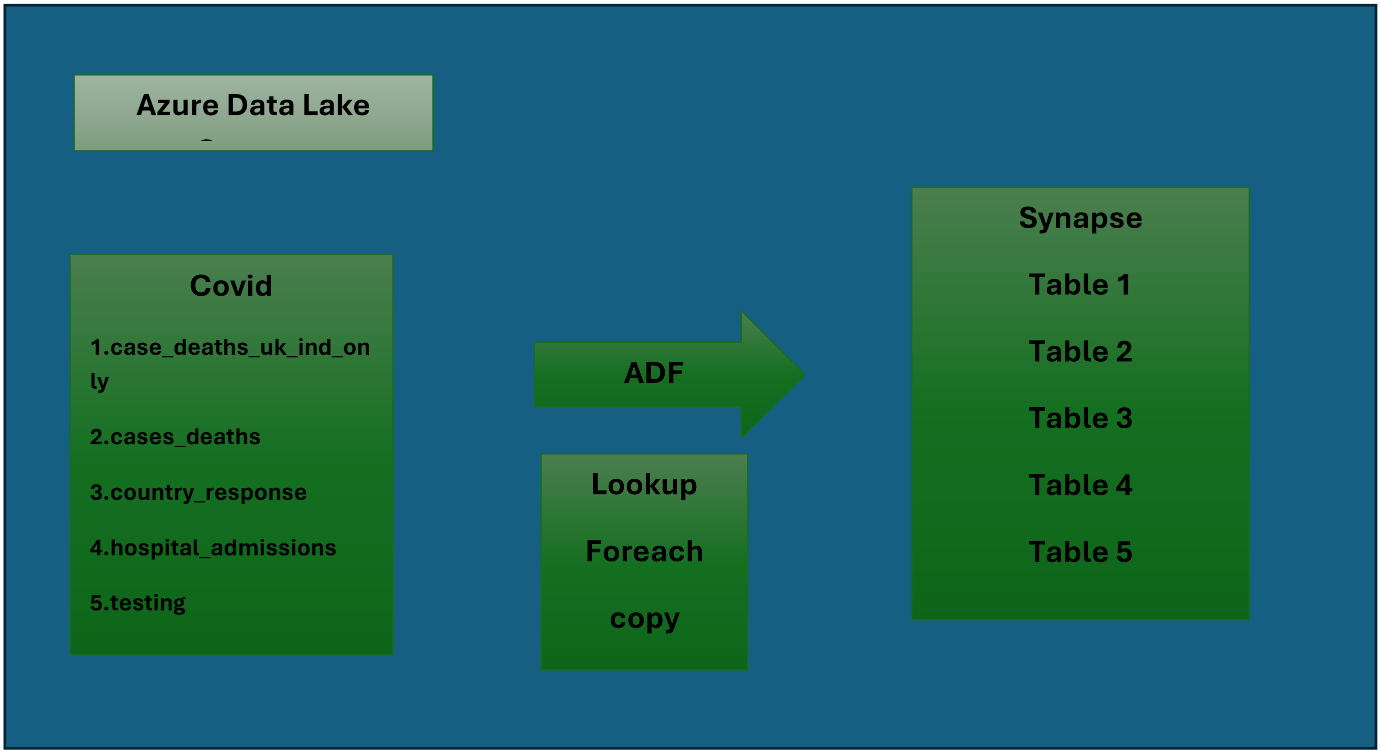
The main aim of this project (Covid Use Case Exercise) is that we will be having a hands-on experience on Storage, ADF Pipeline, Mapping Dataflow, Azure Synapse along with getting to know how to ingest data from flat files into Azure Data Lake Gen2 and Azure Synapse using Azure Data Factory (ADF) and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

Architecture of the project



Requirements of project

Requirement 1



Azure Data Lake

Covid

1.case_deaths_uk_ind_only

2.cases_deaths

3.country_response

4.hospital_admissions

5.testing

ADF

Lookup

Foreach

copy

Synapse

Table 1

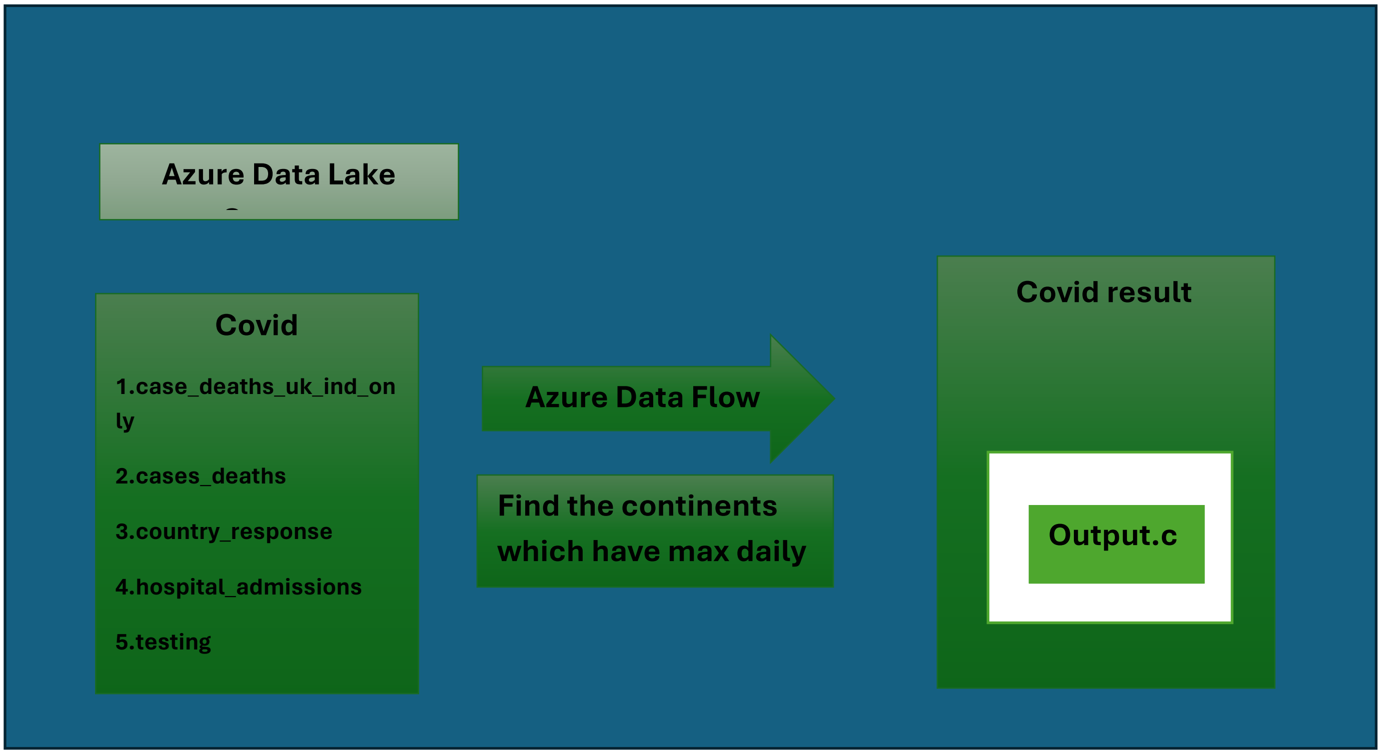
Table 2

Table 3

Table 4

Table 5

Requirement 2



Azure Data Lake

Covid

1.case_deaths_uk_ind_only

2.cases_deaths

3.country_response

4.hospital_admissions

5.testing

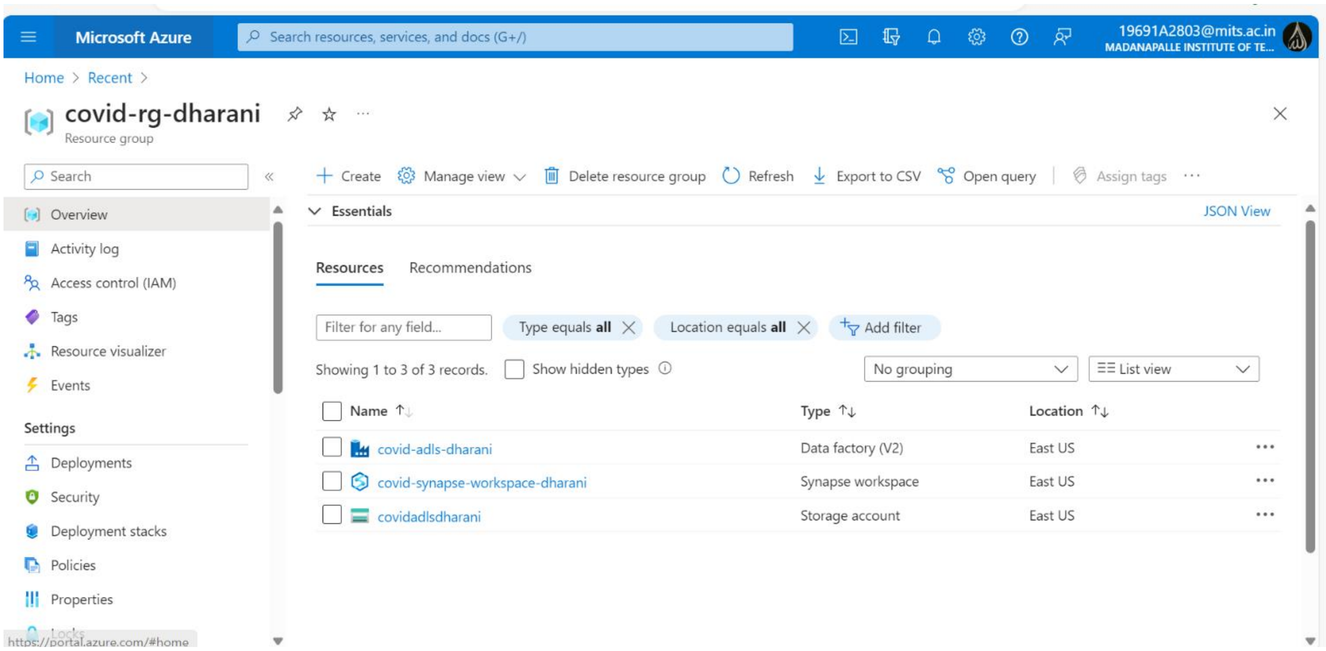
Azure Data Flow

Find the continents which have max daily

Covid result

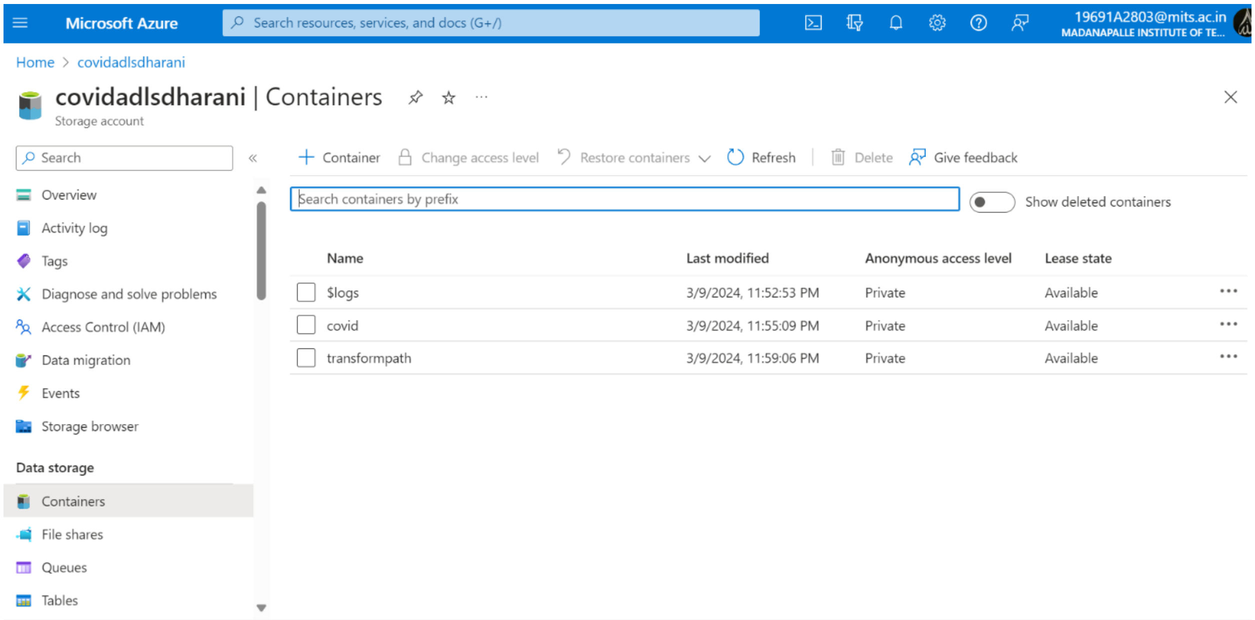
Output.c

Procedure of Requirement 1



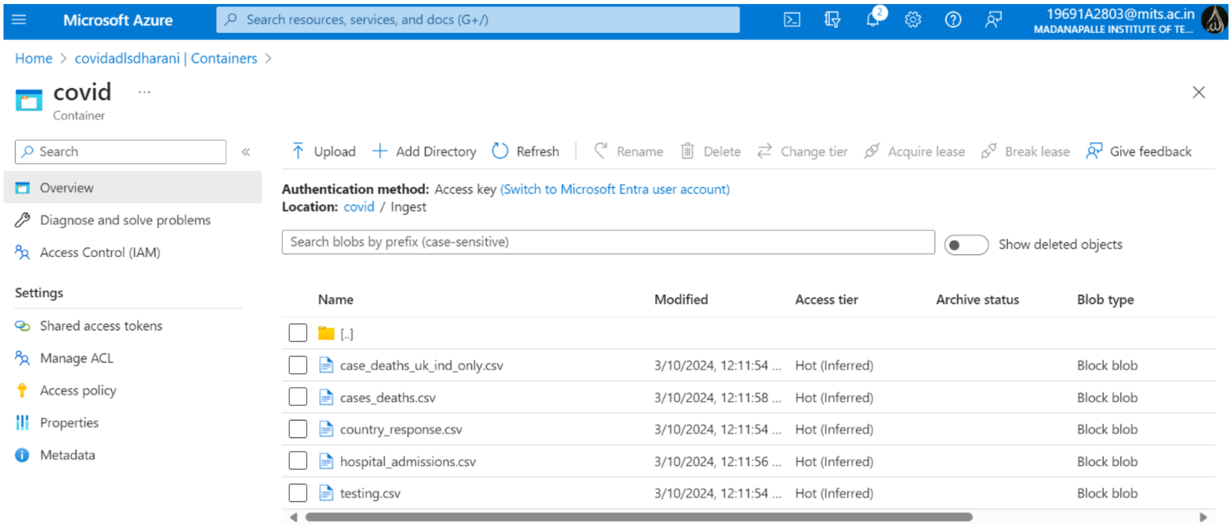
Name	Type	Location
covid-adls-dharani	Data factory (V2)	East US
covid-synapse-workspace-dharani	Synapse workspace	East US
covidadlsdharani	Storage account	East US

Step 1: created one Resource group and required resources for project like storage account, Synapse workspace (data warehouse), Azure Data Factory.



Name	Last modified	Anonymous access level	Lease state
$logs	3/9/2024, 11:52:53 PM	Private	Available
covid	3/9/2024, 11:55:09 PM	Private	Available
transformpath	3/9/2024, 11:59:06 PM	Private	Available

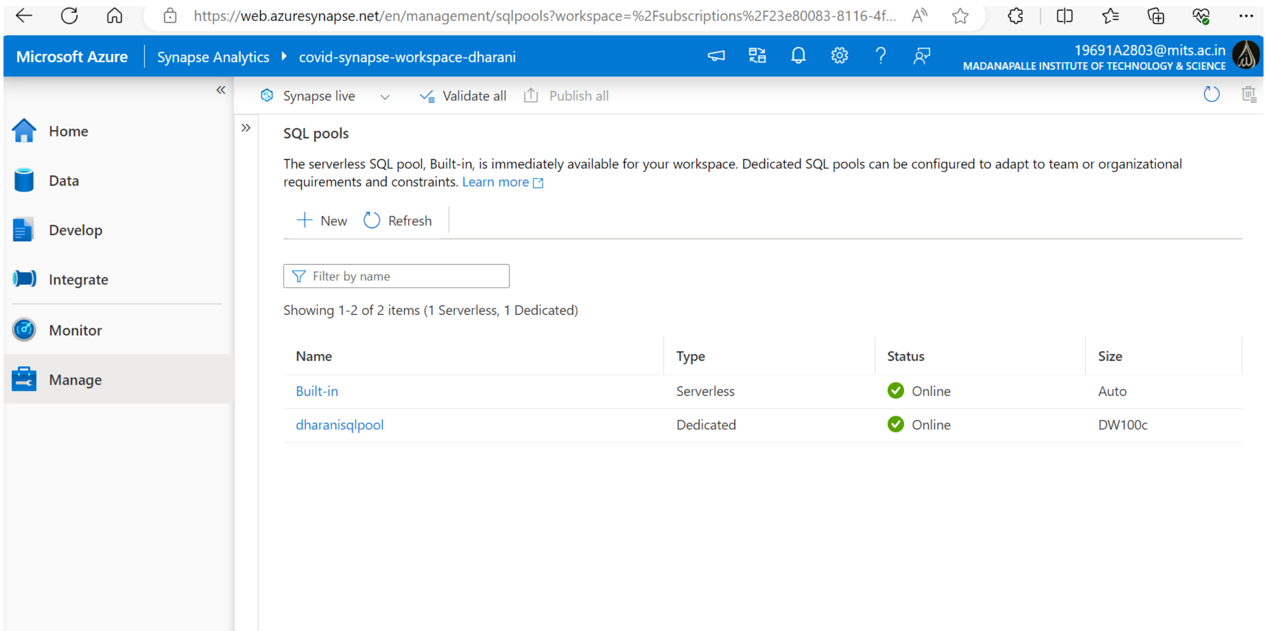
Step 2 : Created Container with name ”**covid**” in Storage account(data lake) for holding folder which contains flat files.



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: covid / Ingest

	Name	Modified	Access tier	Archive status	Blob type
□	[..]				
□	case_deaths_uk_ind_only.csv	3/10/2024, 12:11:54 ...	Hot (Inferred)		Block blob
□	cases_deaths.csv	3/10/2024, 12:11:58 ...	Hot (Inferred)		Block blob
□	country_response.csv	3/10/2024, 12:11:54 ...	Hot (Inferred)		Block blob
□	hospital_admissions.csv	3/10/2024, 12:11:56 ...	Hot (Inferred)		Block blob
□	testing.csv	3/10/2024, 12:11:54 ...	Hot (Inferred)		Block blob

Step 3 : Create Folder with name **"Ingest"** inside the container "covid" and uploaded csv files(data sets) from personal computer which are there in zip file given in project document.


https://web.azuresynapse.net/en/management/sqlpools?workspace=%2Fsubscriptions%2F23e80083-8116-4f...
Microsoft Azure
Synapse Analytics
covid-synapse-workspace-dharani
19691A2803@mits.ac.in
MADANAPALLE INSTITUTE OF TECHNOLOGY & SCIENCE
Synapse live
Validate all
Publish all
Home
Data
Develop
Integrate
Monitor
Manage
New
Refresh
Filter by name

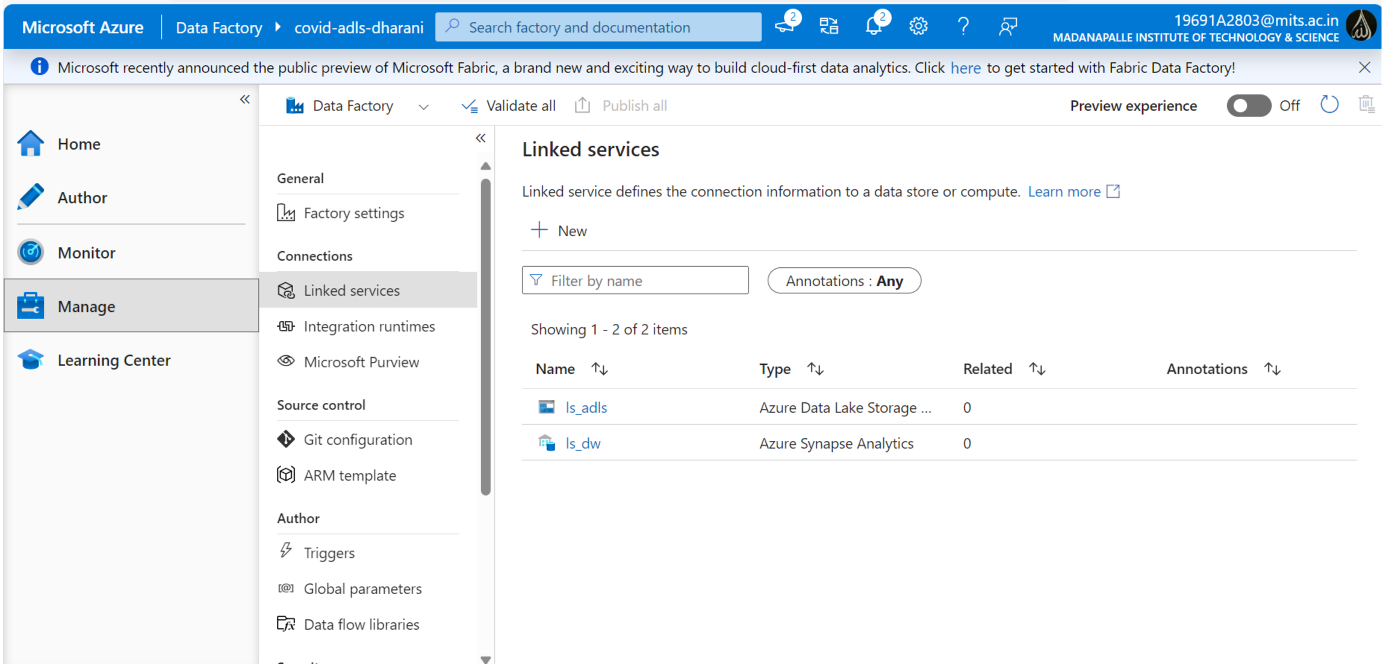
SQL pools

The serverless SQL pool, Built-in, is immediately available for your workspace. Dedicated SQL pools can be configured to adapt to team or organizational requirements and constraints. Learn more

Showing 1-2 of 2 items (1 Serverless, 1 Dedicated)

Name	Type	Status	Size
Built-in	Serverless	✓ Online	Auto
dharanisqlpool	Dedicated	✓ Online	DW100c

Step 4 : Create Azure Synapse resource and one dedicated pool inside the azure synapse for data warehouse creation and it should be turn on.



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!

Linked services

Linked service defines the connection information to a data store or compute. Learn more

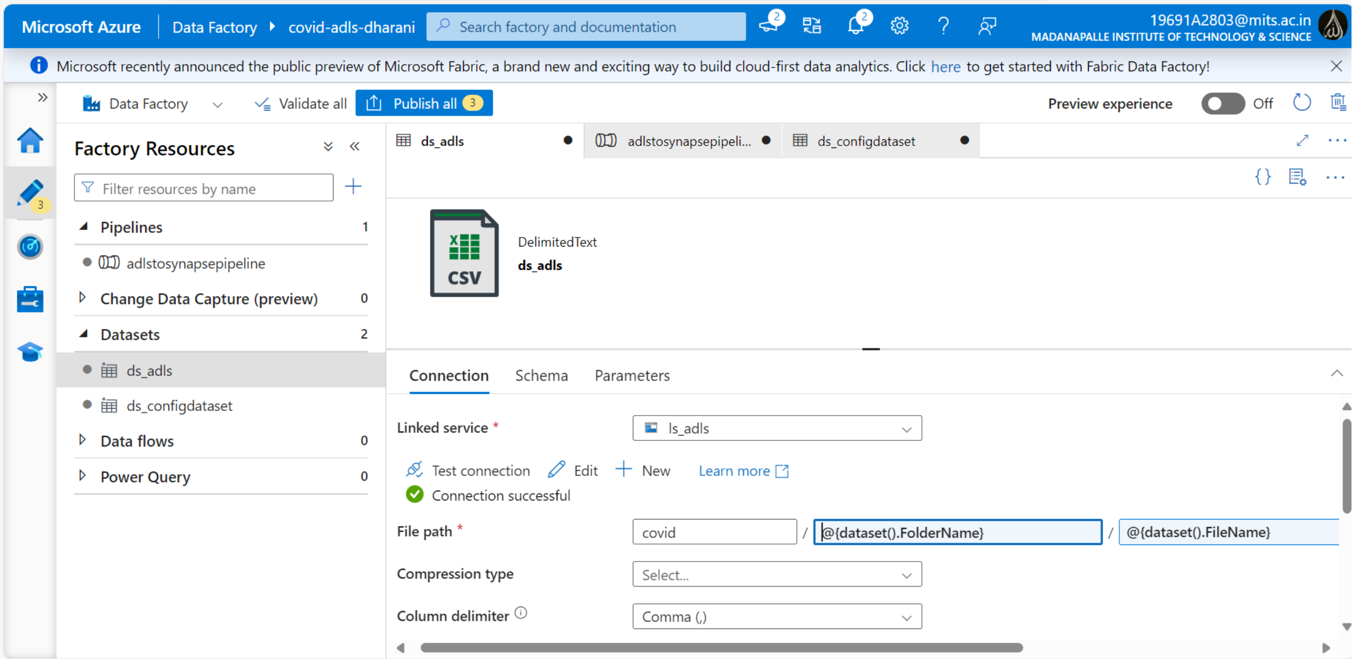
Showing 1 - 2 of 2 items

Name	Type	Related	Annotations
ls_adls	Azure Data Lake Storage ...	0	
ls_dw	Azure Synapse Analytics	0	

Step 5 : Creating **two linked services** as per the project requirement in Azure data factory

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

°Azure synapse workspace to Azure Data factory.



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!

Factory Resources	
Pipelines	1
adlstosynapsepipeline	
Change Data Capture (preview)	0
Datasets	2
ds_adls	
ds_configdataset	
Data flows	0
Power Query	0

DelimitedText
ds_adls

Linked service * ls_adls

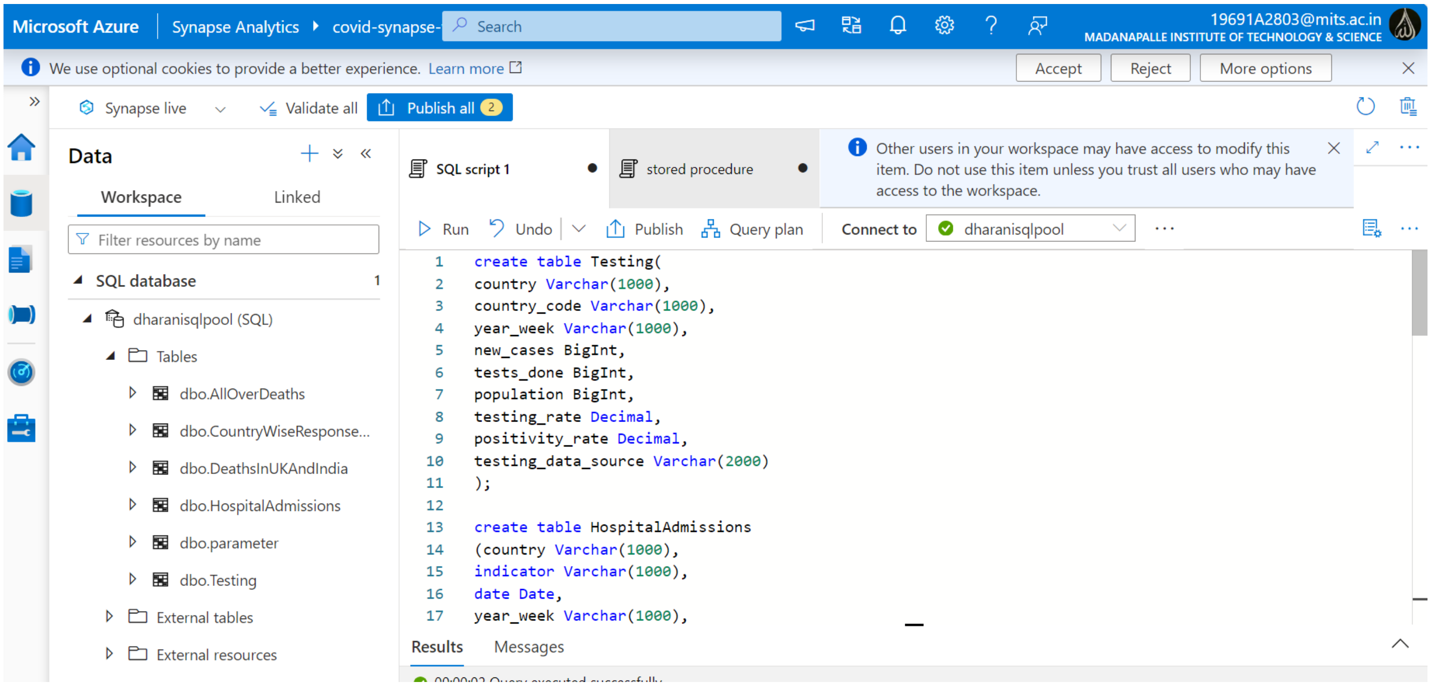
Connection successful

File path * covid / @{dataset().FolderName} / @{dataset().FileName}

Compression type Select...

Column delimiter Comma (,)

Step 6 : Created dataset **(ds_adls)** for fetching flat files from storage account(data lake) which are present in ingest folder inside covid container.



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

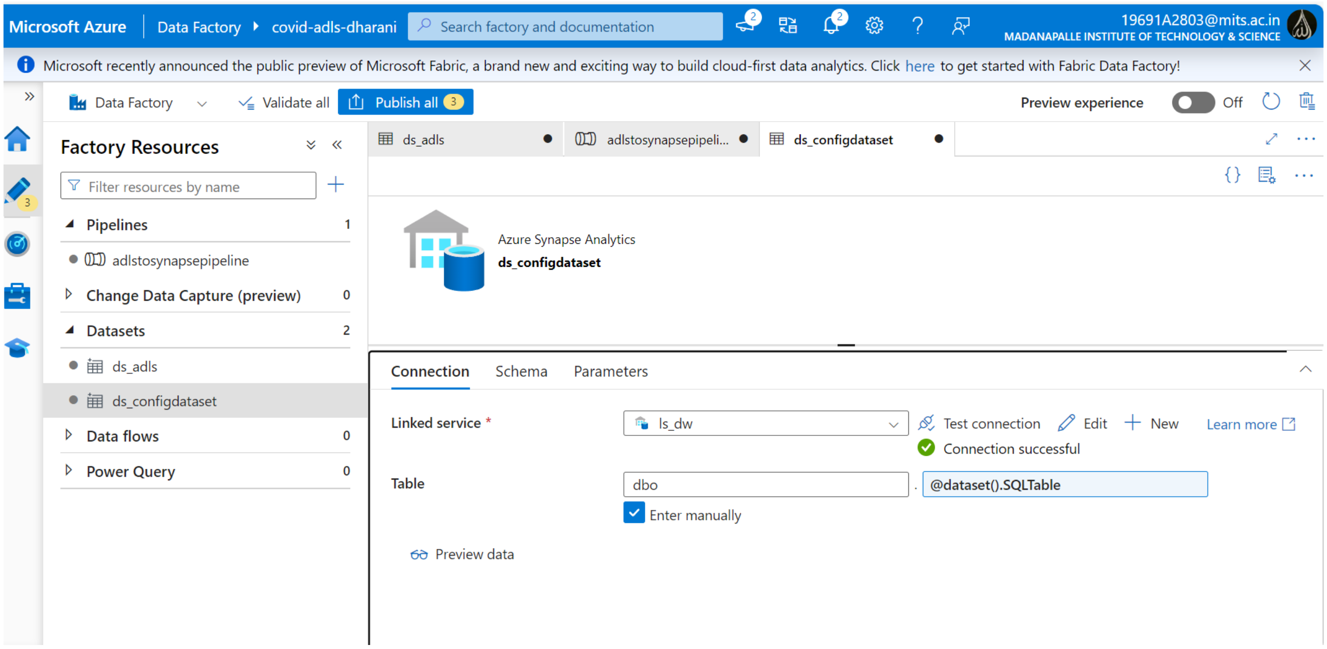
Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Connect to ✓ dharanisqlpool

```
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),
```

Step 7 : Created required SQL Tables in Synapse SQL database by writing create table queries in Synapse workspace (SQL Script).



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!

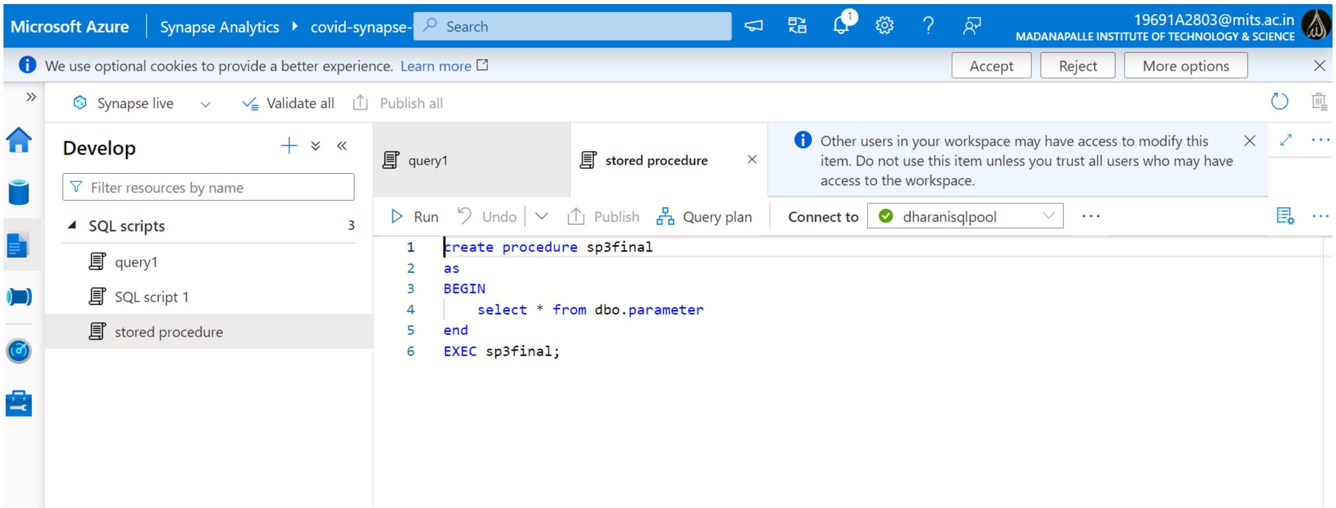
Linked service * ls_dw

✓ Connection successful

Table dbo . @dataset().SQLTable

☑ Enter manually

Step 8 : Created **dataset(ds_configdataset)** for inserting into SQL Tables created in synapse(data warehouse).

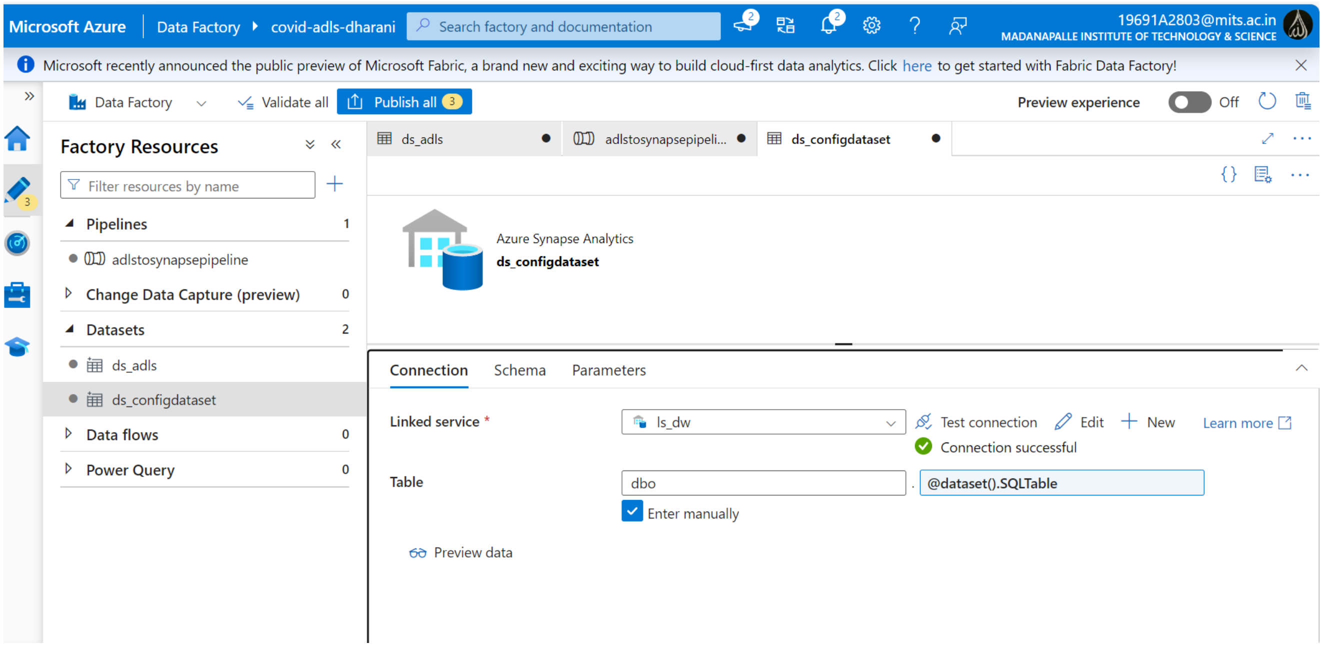


We use optional cookies to provide a better experience. Learn more

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.

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

Step 9 : Created **stored procedure(sp3final)** for fetching records from parameters table.



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!

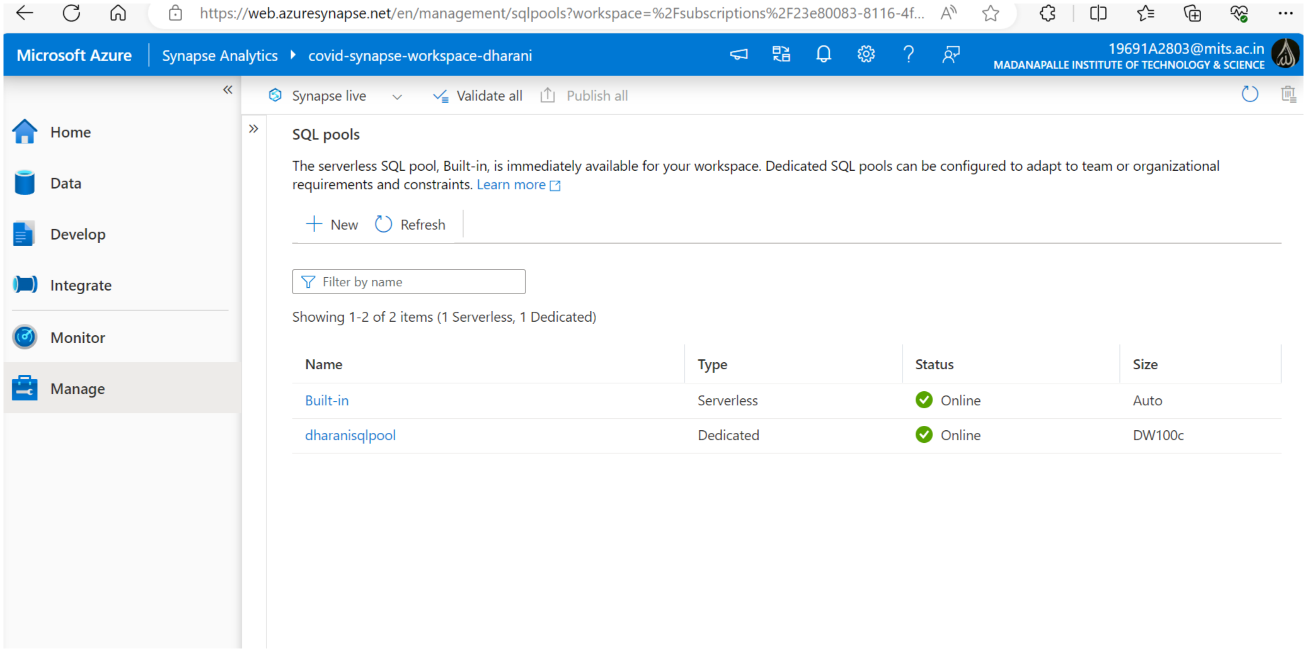
Linked service * ls_dw

Connection successful

Table dbo . @dataset().SQLTable

☑ Enter manually

Step 10 : Created dataset(ds_configdataset) for fetching parameters table from synapse(data warehouse).



https://web.azuresynapse.net/en/management/sqlpools?workspace=%2Fsubscriptions%2F23e80083-8116-4f...
Microsoft Azure
Synapse Analytics
covid-synapse-workspace-dharani
19691A2803@mits.ac.in
MADANAPALLE INSTITUTE OF TECHNOLOGY & SCIENCE
Synapse live
Validate all
Publish all
Home
Data
Develop
Integrate
Monitor
Manage
New
Refresh
Filter by name

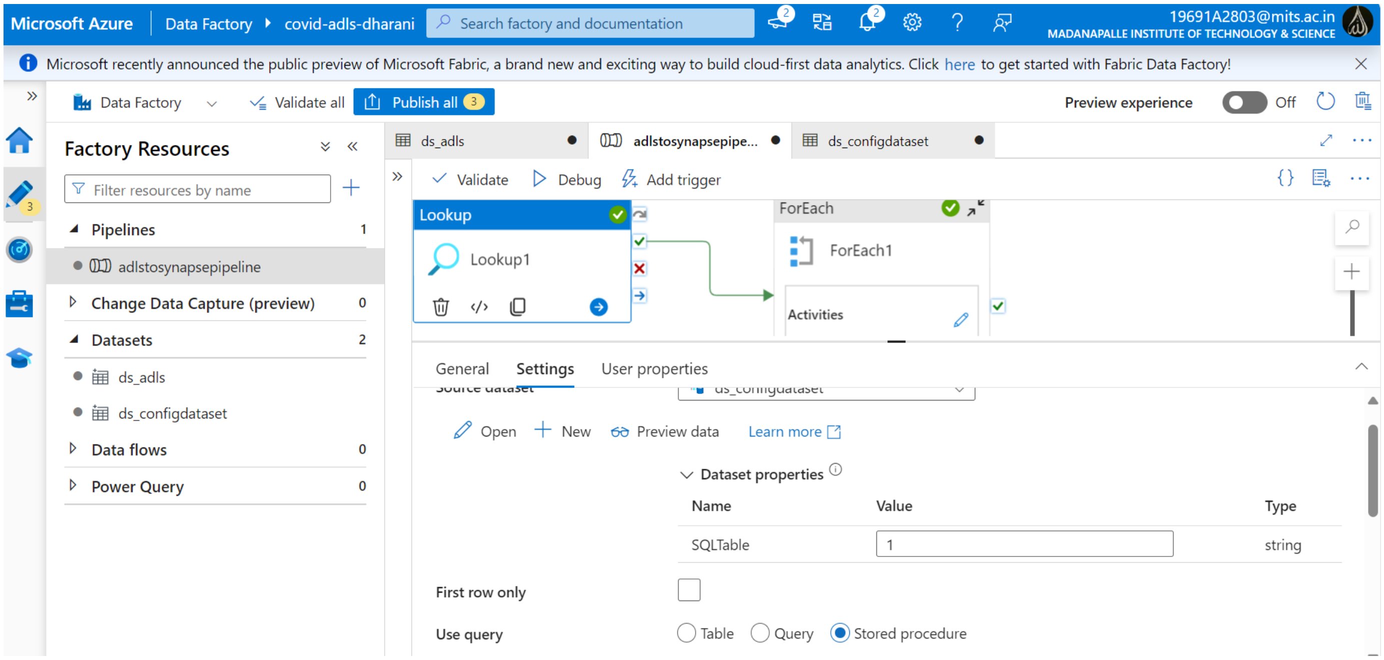
SQL pools

The serverless SQL pool, Built-in, is immediately available for your workspace. Dedicated SQL pools can be configured to adapt to team or organizational requirements and constraints. Learn more

Showing 1-2 of 2 items (1 Serverless, 1 Dedicated)

Name	Type	Status	Size
Built-in	Serverless	Online	Auto
dharanisqlpool	Dedicated	Online	DW100c

Step 12 : Before creating the pipeline in the data factory, we need to turn on the dedicated pool. we need to check these two to three times while moving on to creation of pipeline.



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!

Source dataset: ds_configdataset

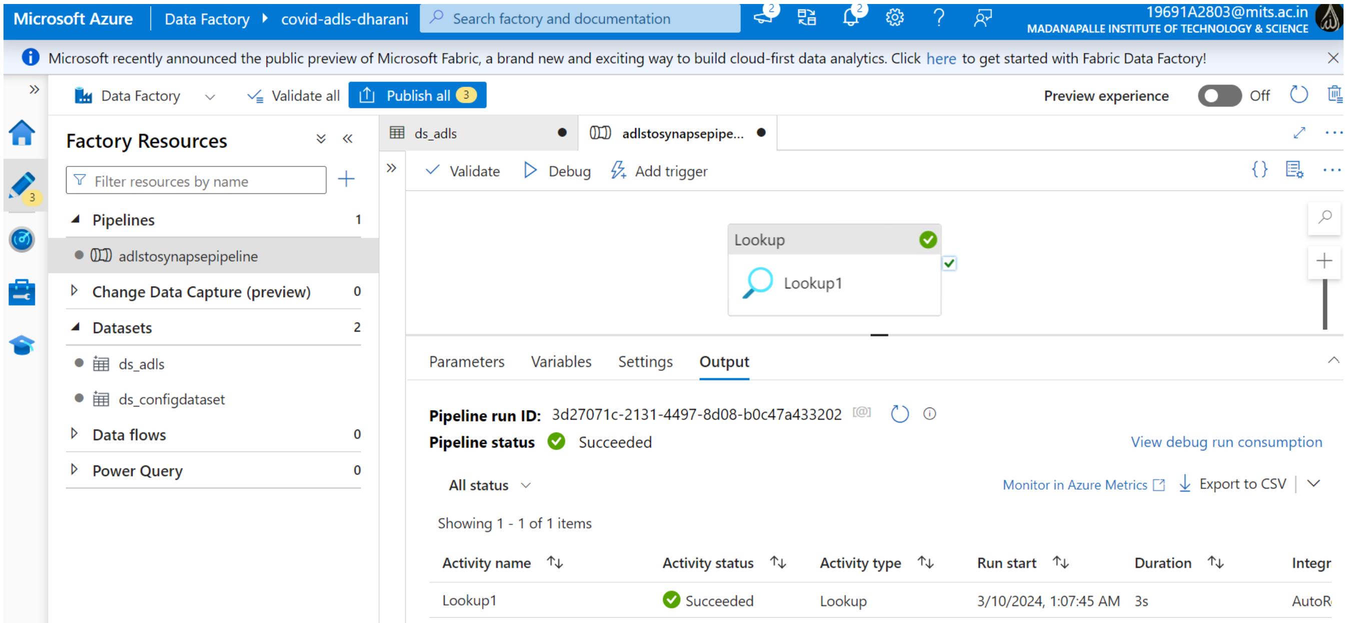
Dataset properties

Name	Value	Type
SQLTable	1	string

First row only □

Use query ○ Table ○ Query ● Stored procedure

Step 13 : Drag and Drop the LOOK UP activity into pipeline workspace and set the source dataset (ds_configdataset) for lookup activity and choose the option **Stored procedure** and give the stored procedure name created in synapse(data warehouse).



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!

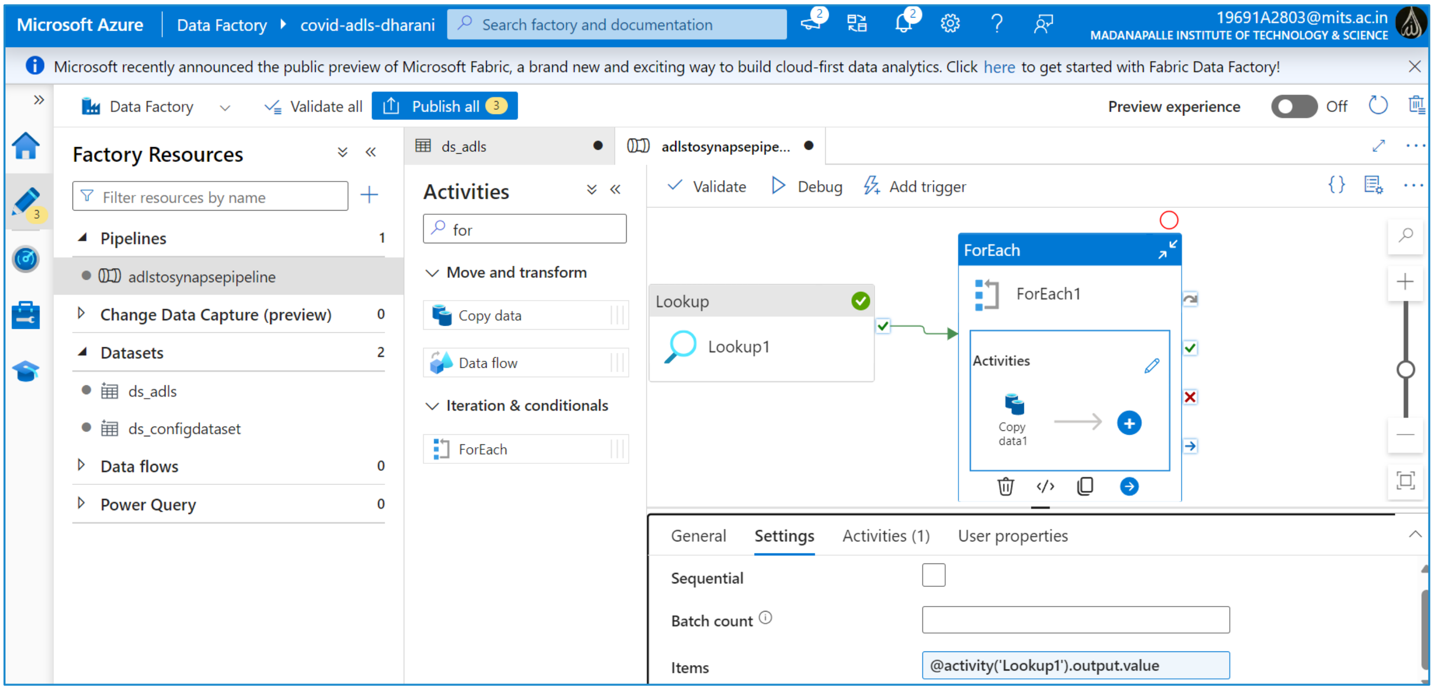
Pipeline run ID: 3d27071c-2131-4497-8d08-b0c47a433202

Pipeline status ✓ Succeeded

Showing 1 - 1 of 1 items

Activity name	Activity status	Activity type	Run start	Duration	Integr
Lookup1	✓ Succeeded	Lookup	3/10/2024, 1:07:45 AM	3s	AutoR

Step 14 : After clicking on debug, lookup will be run successfully.



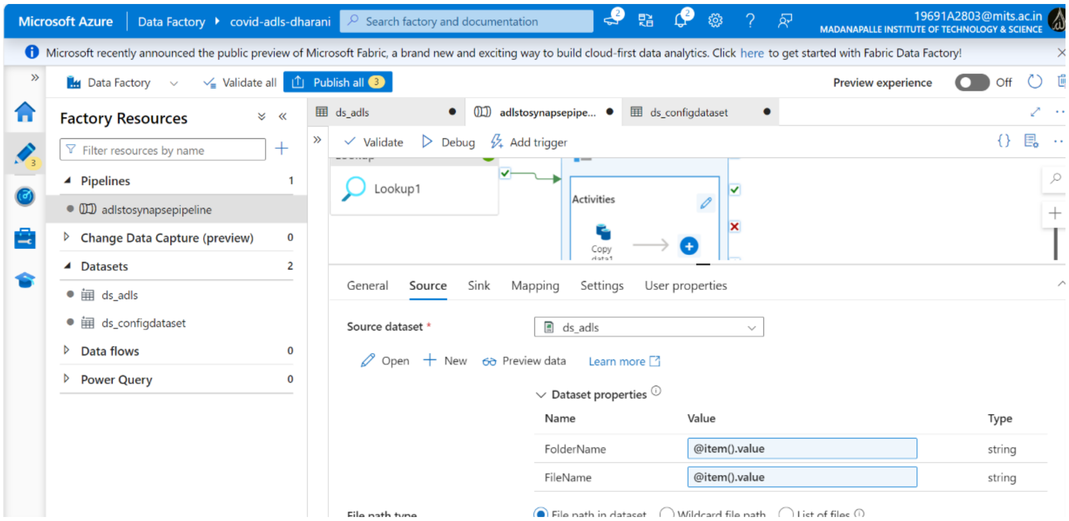
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Sequential □

Batch count

Items @activity('Lookup1').output.value

Step 15 : Drag and Drop **For each activity** in pipeline workspace and configure the for each activity settings like **Items** with output of lookup activity (@activity('Lookup1').Output.value).



Source dataset * ds_adls

Dataset properties

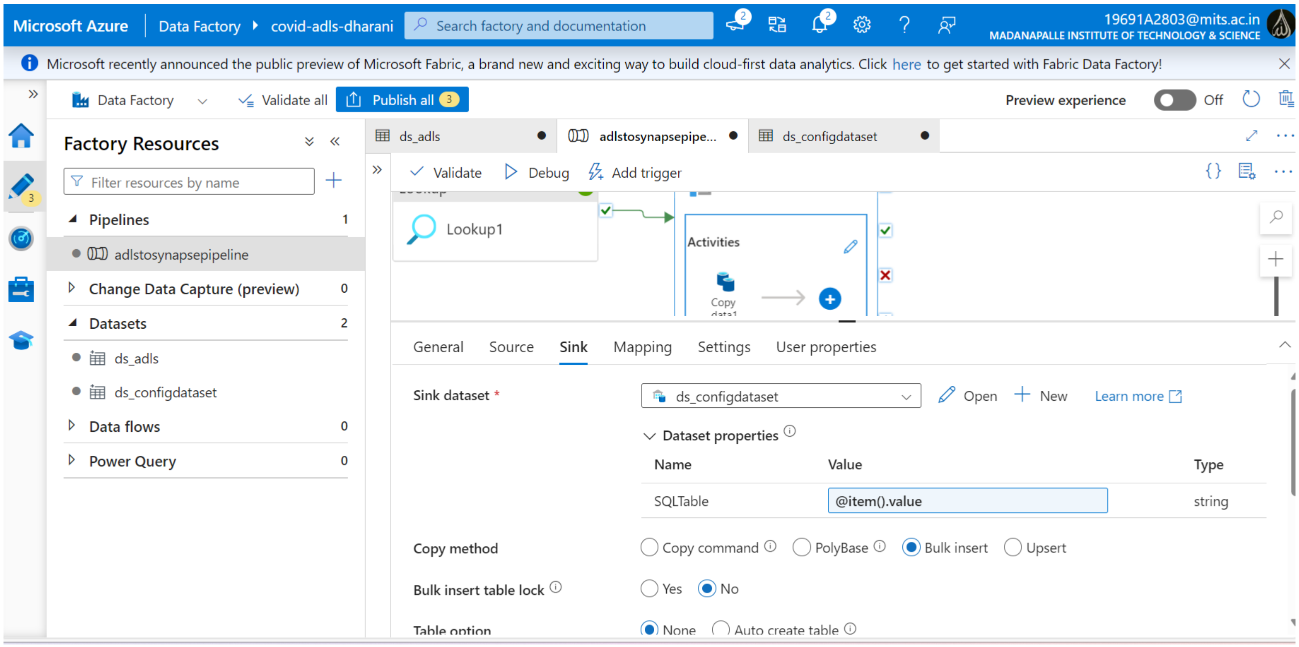
Name	Value	Type
FolderName	@item().value	string
FileName	@item().value	string

File path type ● File path in dataset ○ Wildcard file path ○ List of files

Step 16 : Click on add activity symbol present on foreach activity and inside foreach activity add a **copy activity** for copy data from CSV file into SQL Table.

Configure settings at source side in copy activity by giving dataset**(ds_adls)** and giving folder name and file name dynamically by taking from foreach activity by **item**.

Folder Name(@{item().FolderName})Name(@item().FileName})



Sink dataset * ds_configdataset

Dataset properties

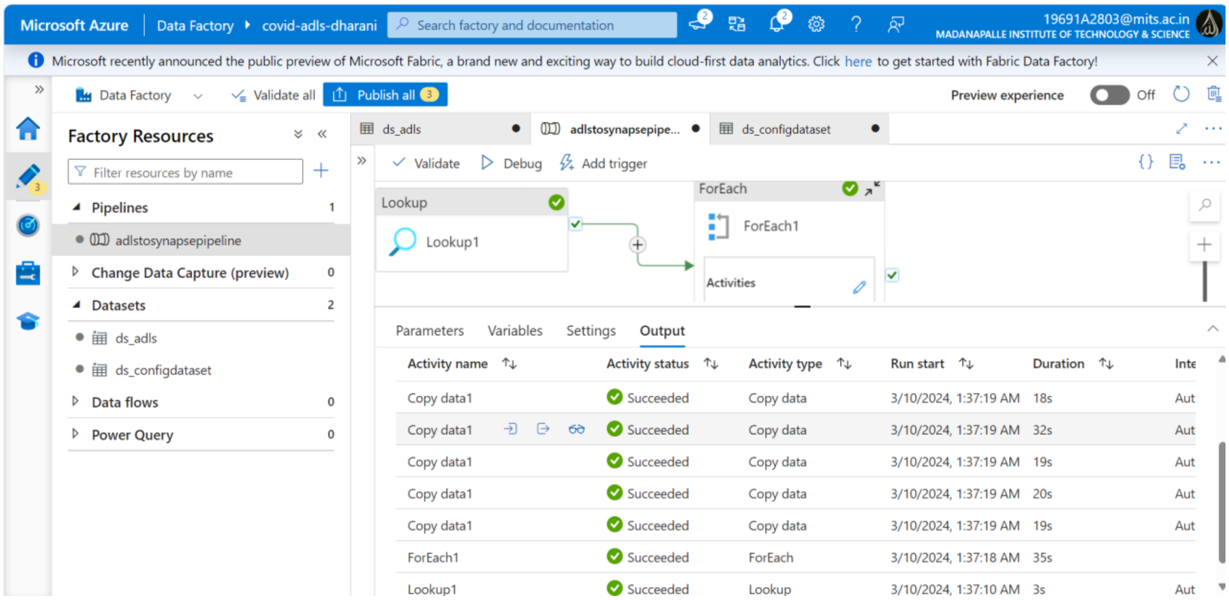
Name	Value	Type
SQLTable	@item().value	string

Copy method ○ Copy command ○ PolyBase ● Bulk insert ○ Upsert

Bulk insert table lock ○ Yes ● No

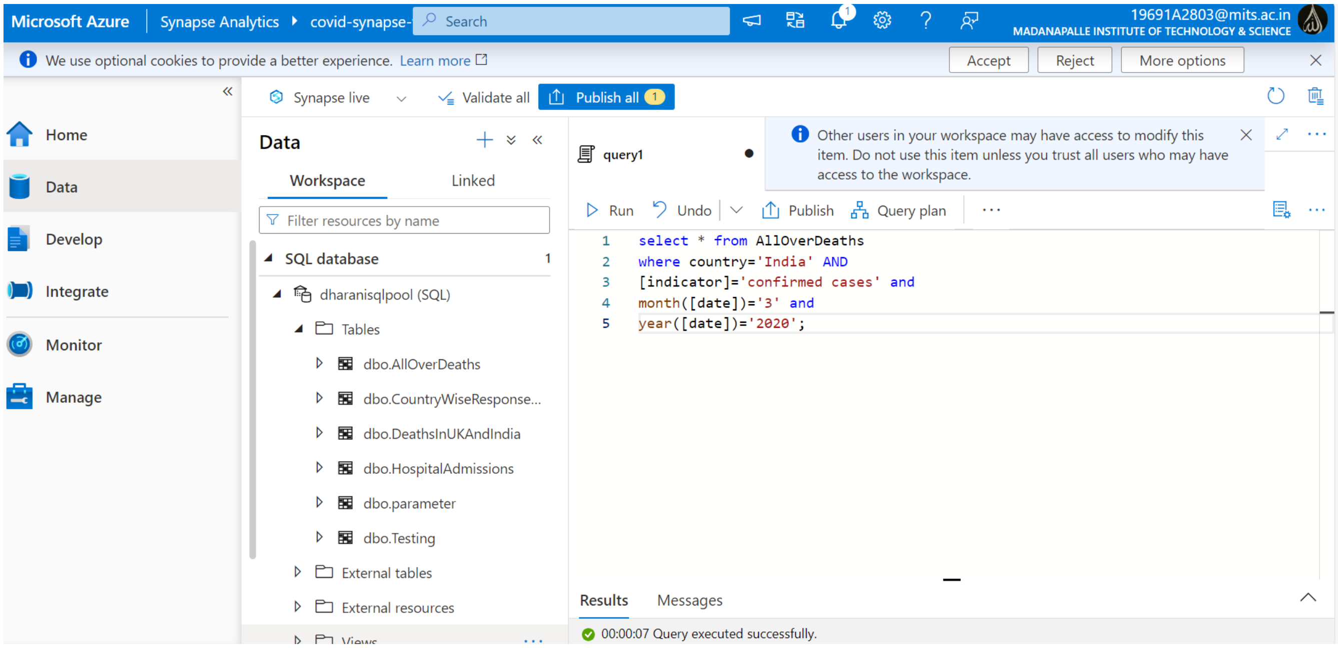
Table option ● None ○ Auto create table

Step 17 : Configure setting in copy activity at sink by giving dataset(ds_configdataset) and giving sqltableName dynamically by taking from foreach activity by **item->sqltableName({@item().sqltableName})**


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!

Activity name	Activity status	Activity type	Run start	Duration	Inte
Copy data1	Succeeded	Copy data	3/10/2024, 1:37:19 AM	18s	Aut
Copy data1	Succeeded	Copy data	3/10/2024, 1:37:19 AM	32s	Aut
Copy data1	Succeeded	Copy data	3/10/2024, 1:37:19 AM	19s	Aut
Copy data1	Succeeded	Copy data	3/10/2024, 1:37:19 AM	20s	Aut
Copy data1	Succeeded	Copy data	3/10/2024, 1:37:19 AM	19s	Aut
ForEach1	Succeeded	ForEach	3/10/2024, 1:37:18 AM	35s	
Lookup1	Succeeded	Lookup	3/10/2024, 1:37:10 AM	3s	Aut

Step 18 : After setting whole pipeline by using lookup and foreach activity recheck all parameters given in each configuration setting, check the dedicated pool is turn on and then turn on the debug option in pipeline. Finally, all the activities are successfully ruined.

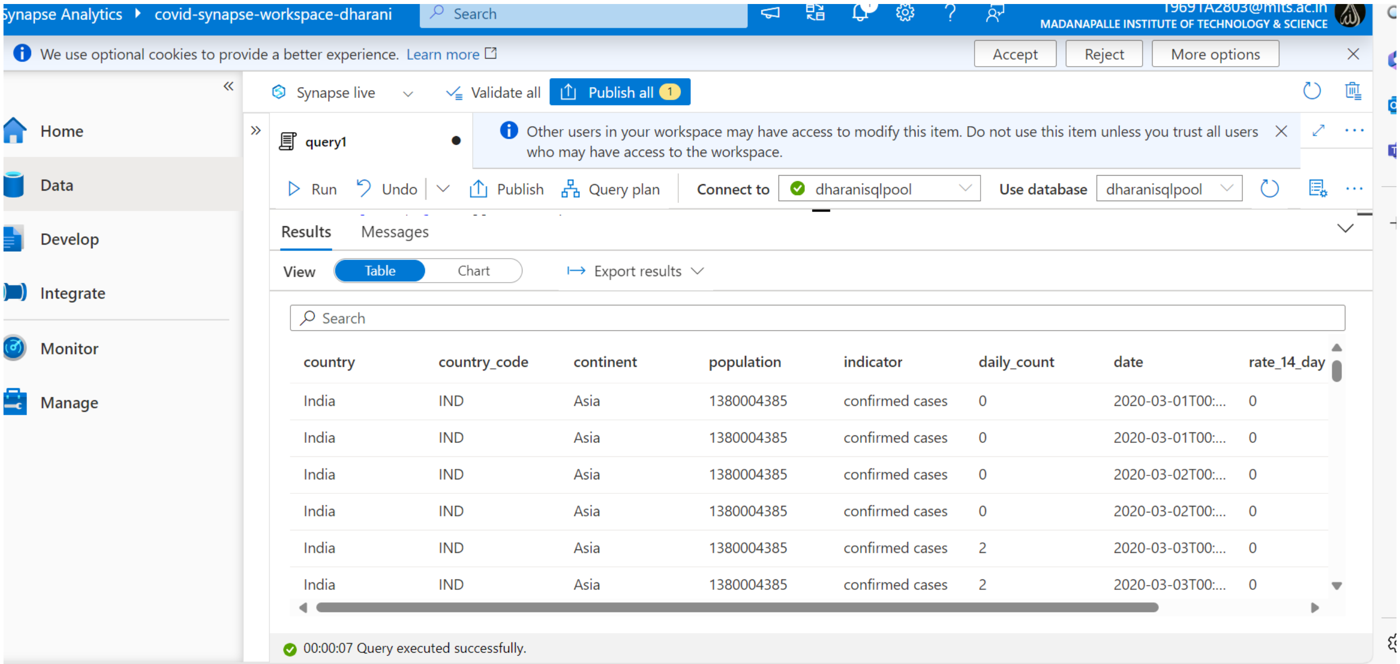

We use optional cookies to provide a better experience. Learn more

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.

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

00:00:07 Query executed successfully.

Step 19 : After successfully run of pipeline ,now we need to check the data inserted into tables in data warehouse by performing two SQL queries operation given in project documentation.



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

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.

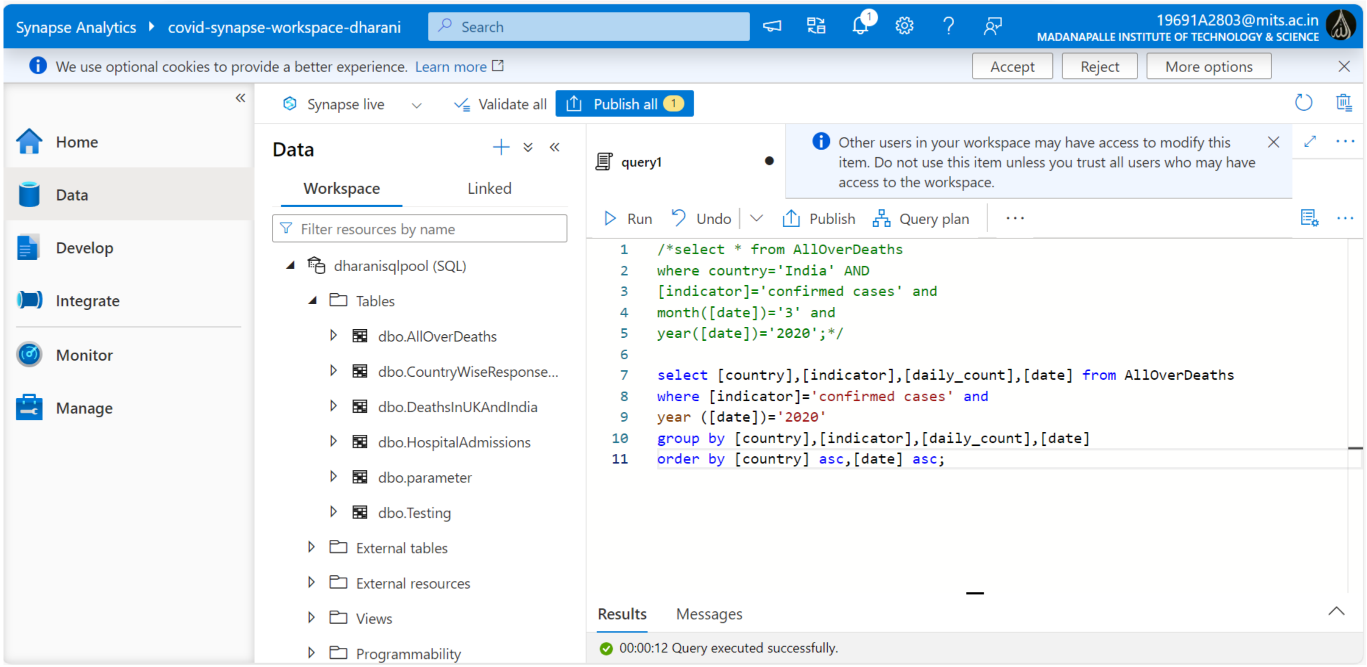
Connect to dharanisqlpool

Use database dharanisqlpool

country	country_code	continent	population	indicator	daily_count	date	rate_14_day
India	IND	Asia	1380004385	confirmed cases	0	2020-03-01T00:...	0
India	IND	Asia	1380004385	confirmed cases	0	2020-03-01T00:...	0
India	IND	Asia	1380004385	confirmed cases	0	2020-03-02T00:...	0
India	IND	Asia	1380004385	confirmed cases	0	2020-03-02T00:...	0
India	IND	Asia	1380004385	confirmed cases	2	2020-03-03T00:...	0
India	IND	Asia	1380004385	confirmed cases	2	2020-03-03T00:...	0

00:00:07 Query executed successfully.

Step 20 : After running first SQL query written in SQL script and it is successfully has ran and given output as per the query.



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

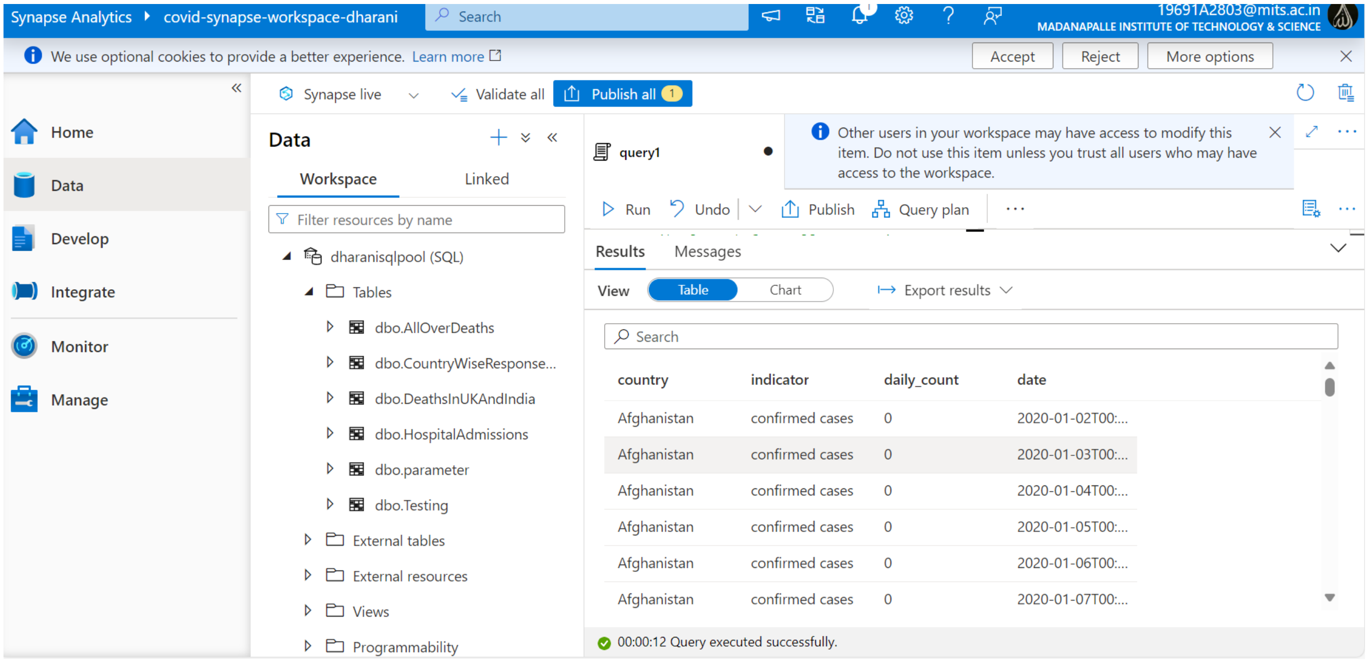
Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

```
/*select * from 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],[indicator],[daily_count],[date]
order by [country] asc,[date] asc;
```

00:00:12 Query executed successfully.

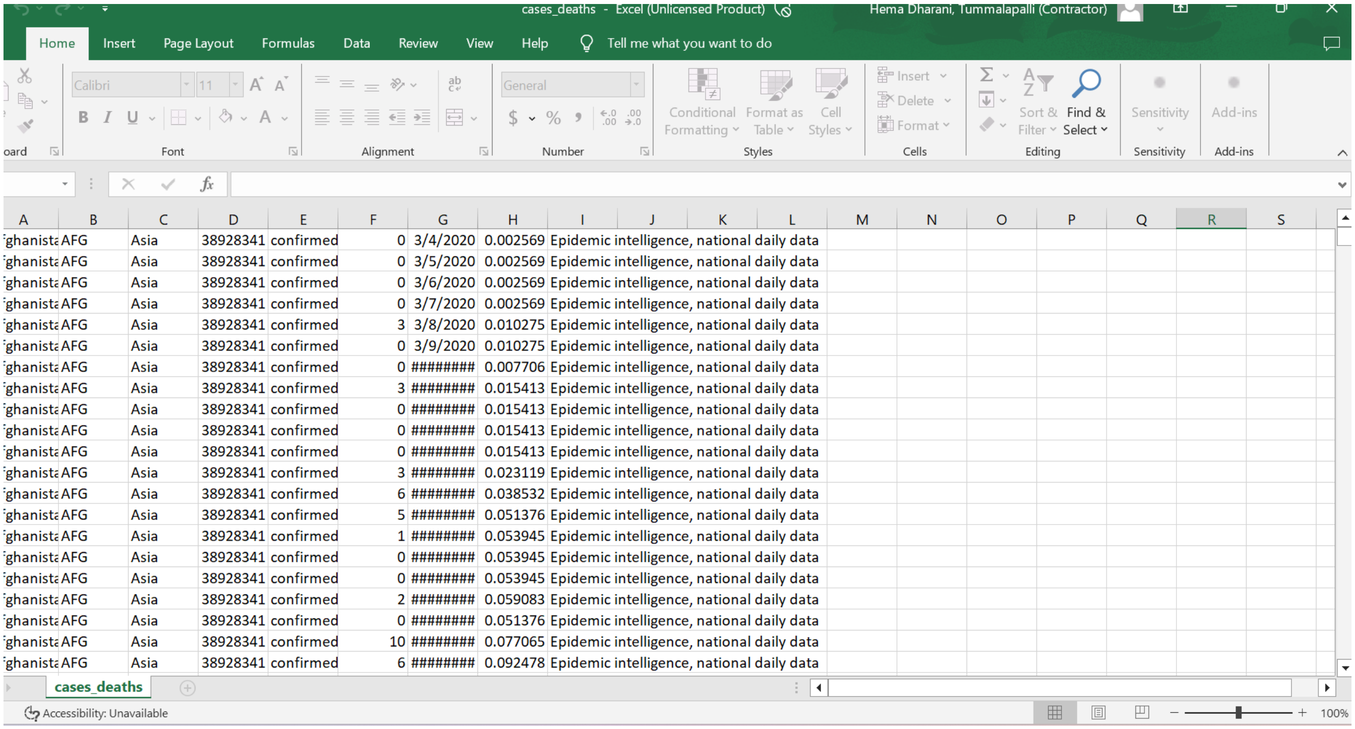
Step 21 : Wrote the second SQL query as per the question given in the project documentation and click on run.



country	indicator	daily_count	date
Afghanistan	confirmed cases	0	2020-01-02T00:...
Afghanistan	confirmed cases	0	2020-01-03T00:...
Afghanistan	confirmed cases	0	2020-01-04T00:...
Afghanistan	confirmed cases	0	2020-01-05T00:...
Afghanistan	confirmed cases	0	2020-01-06T00:...
Afghanistan	confirmed cases	0	2020-01-07T00:...

00:00:12 Query executed successfully.

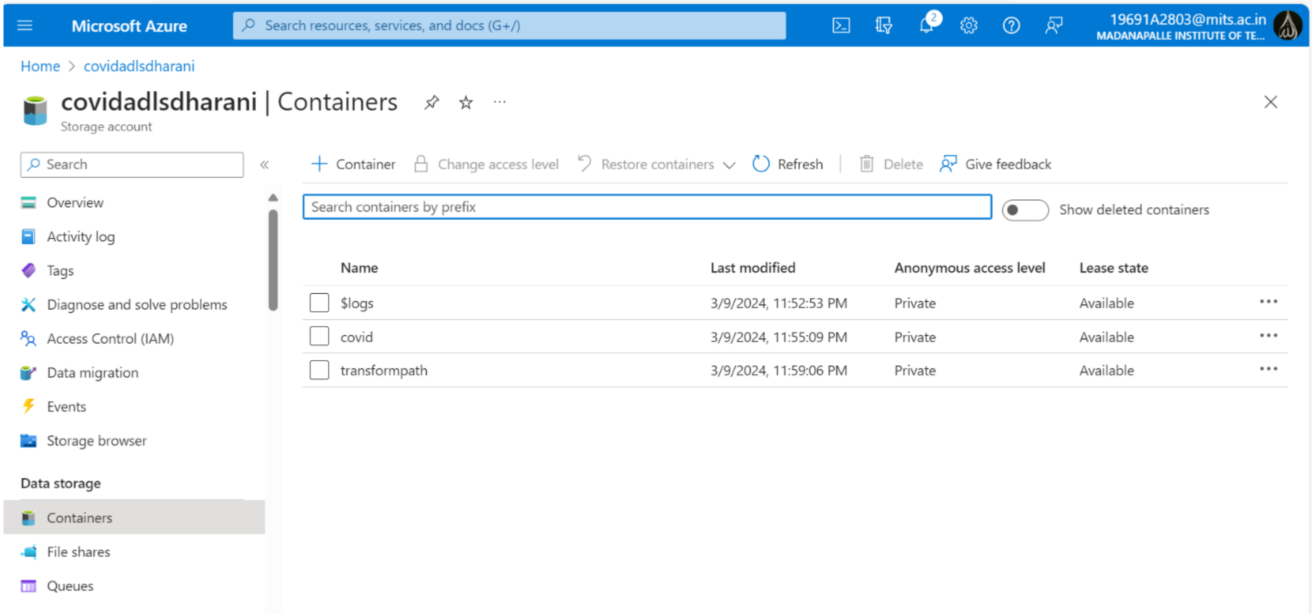
Step 22 : Successfully query has ran and given output as per the query.



A	B	C	D	E	F	G	H	I	J	K	L
ghanista	AFG	Asia	38928341	confirmed	0	3/4/2020	0.002569	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	0	3/5/2020	0.002569	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	0	3/6/2020	0.002569	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	0	3/7/2020	0.002569	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	3	3/8/2020	0.010275	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	0	3/9/2020	0.010275	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	0	########	0.007706	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	3	########	0.015413	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	0	########	0.015413	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	0	########	0.015413	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	0	########	0.015413	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	3	########	0.023119	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	6	########	0.038532	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	5	########	0.051376	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	1	########	0.053945	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	0	########	0.053945	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	0	########	0.053945	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	2	########	0.059083	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	0	########	0.051376	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	10	########	0.077065	Epidemic intelligence, national daily data			
ghanista	AFG	Asia	38928341	confirmed	6	########	0.092478	Epidemic intelligence, national daily data			

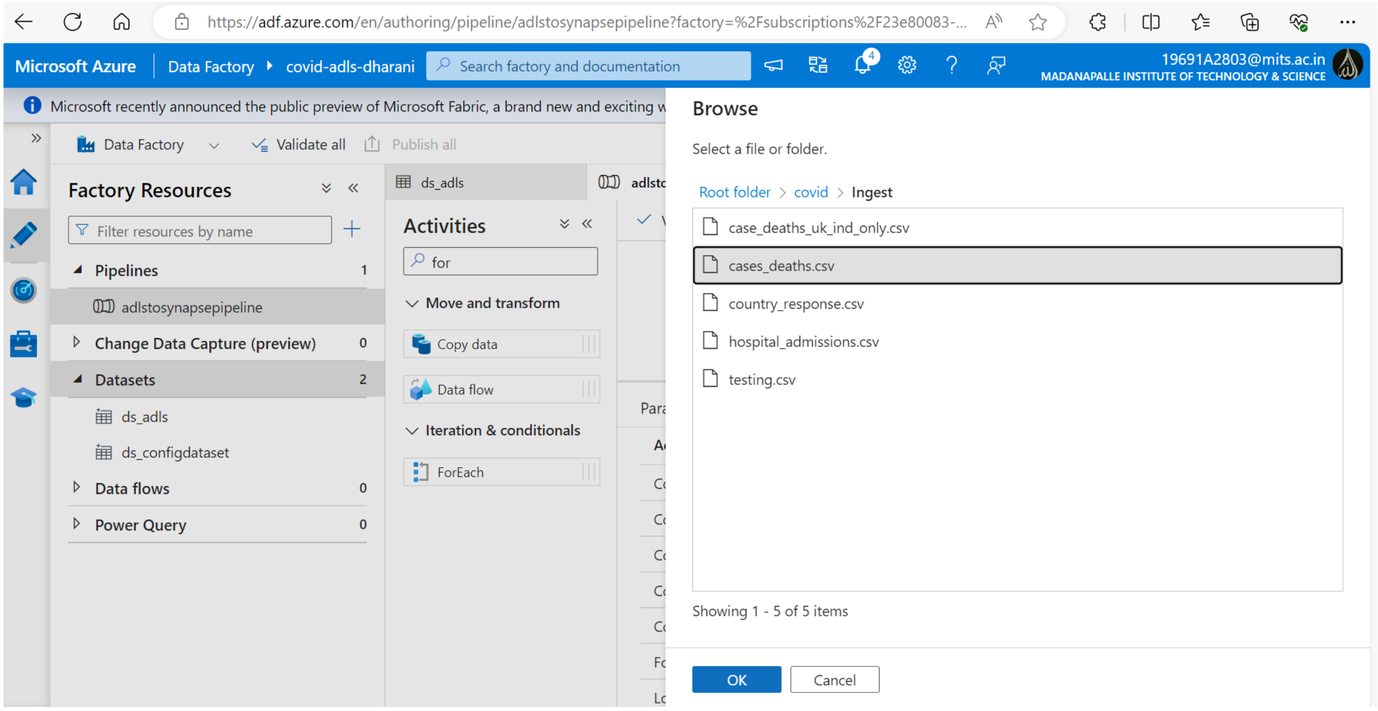
Step 23 : Compare the above SQL query output with data which is appear from csv file which is open in excel for reference check.

Procedure for Project Requirement 2 :



Name	Last modified	Anonymous access level	Lease state
$logs	3/9/2024, 11:52:53 PM	Private	Available
covid	3/9/2024, 11:55:09 PM	Private	Available
transformpath	3/9/2024, 11:59:06 PM	Private	Available

Step 23 : Create another container with name **”transformpath”** for second requirement given in project for storing transformed data file by using data flow.

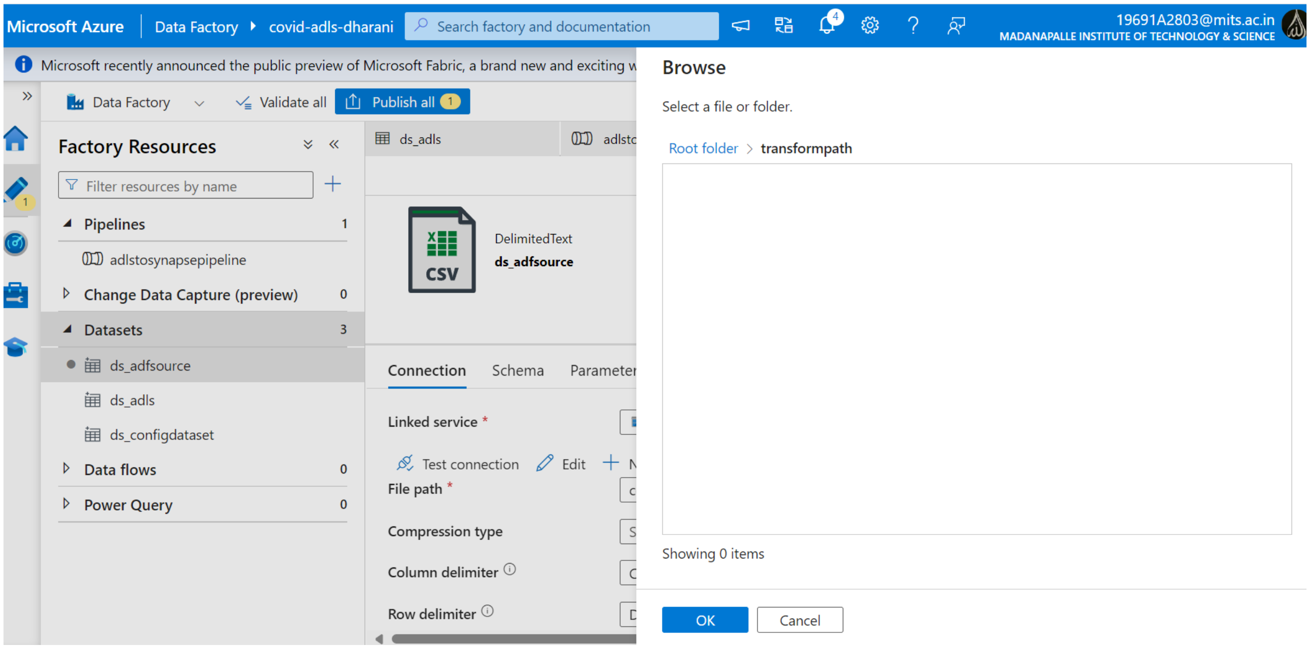


Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting w

Select a file or folder.

Showing 1 - 5 of 5 items

Step 24 : Create source dataset (ds_adls) for dataflow by giving a file specific file name on which data transformation need to be taken place as per project requirement.



Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting w

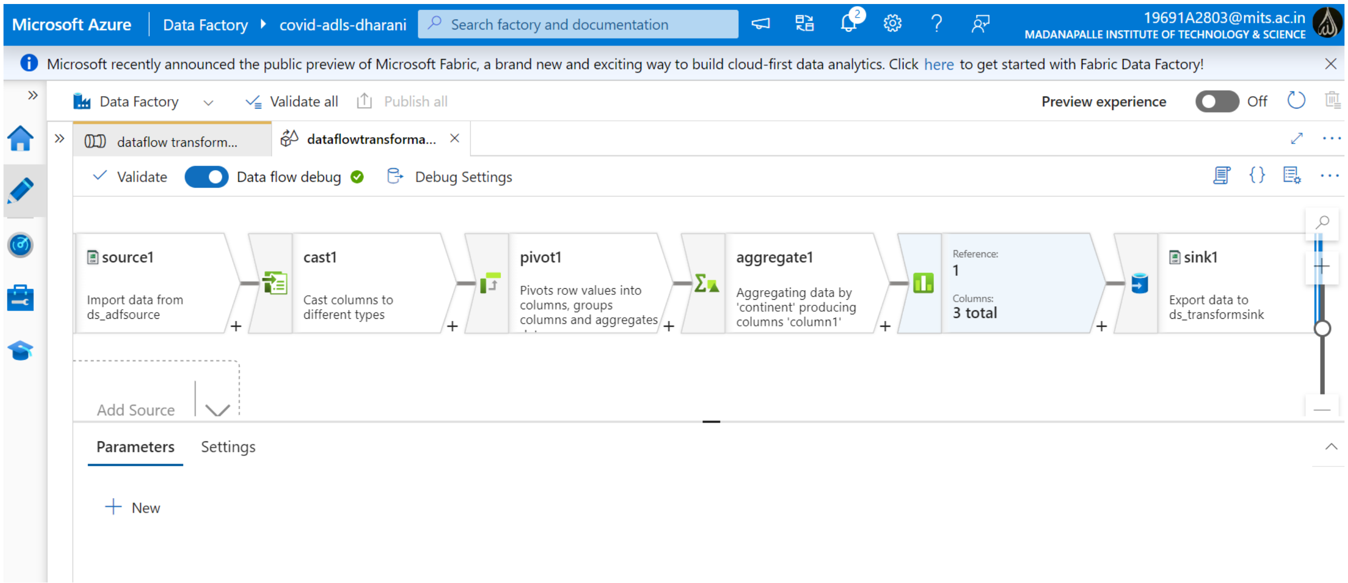
Browse

Select a file or folder.

Root folder > transformpath

Showing 0 items

Step 25 : Create target dataset (ds_adfsource) for dataflow to keep that data transformed file in specific place for further use.



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!

source1 — Import data from ds_adfsource

cast1 — Cast columns to different types

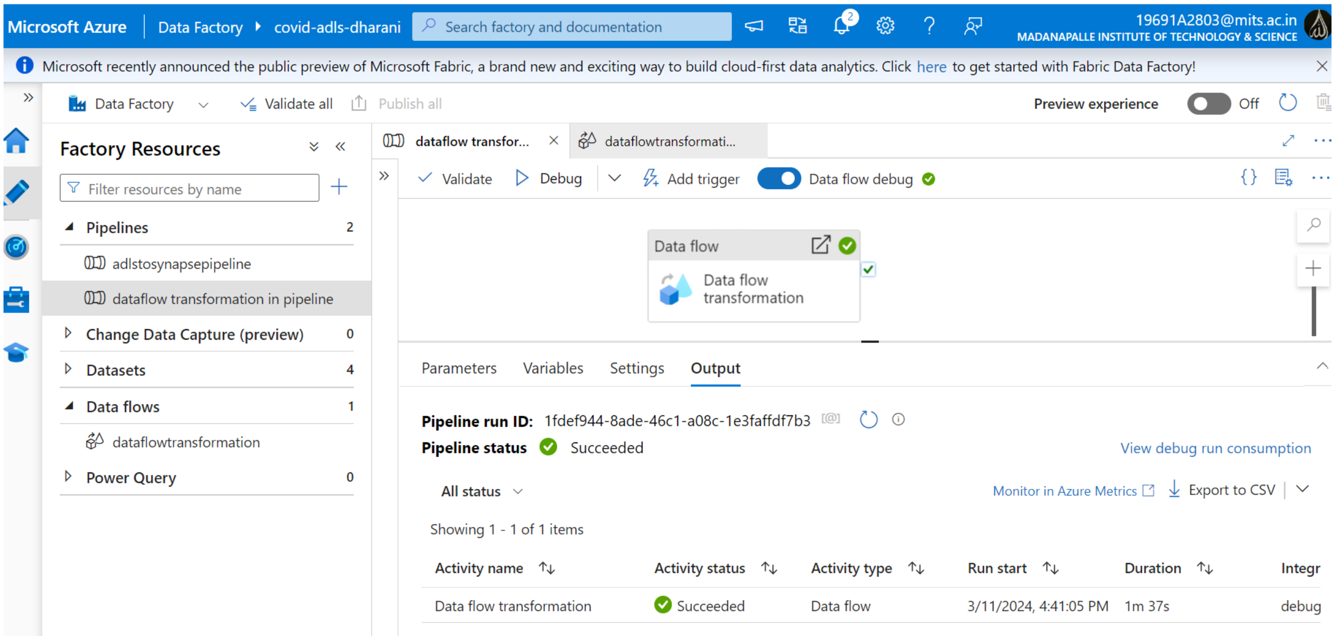
pivot1 — Pivots row values into columns, groups columns and aggregates

aggregate1 — Aggregating data by 'continent' producing columns 'column1'

Reference: 1 Columns: 3 total

sink1 — Export data to ds_transformsink

Step 26 : After successfully creating dataflow I click on dataflow debug.



Pipeline run ID: 1fdef944-8ade-46c1-a08c-1e3faffdf7b3

Pipeline status ✓ Succeeded

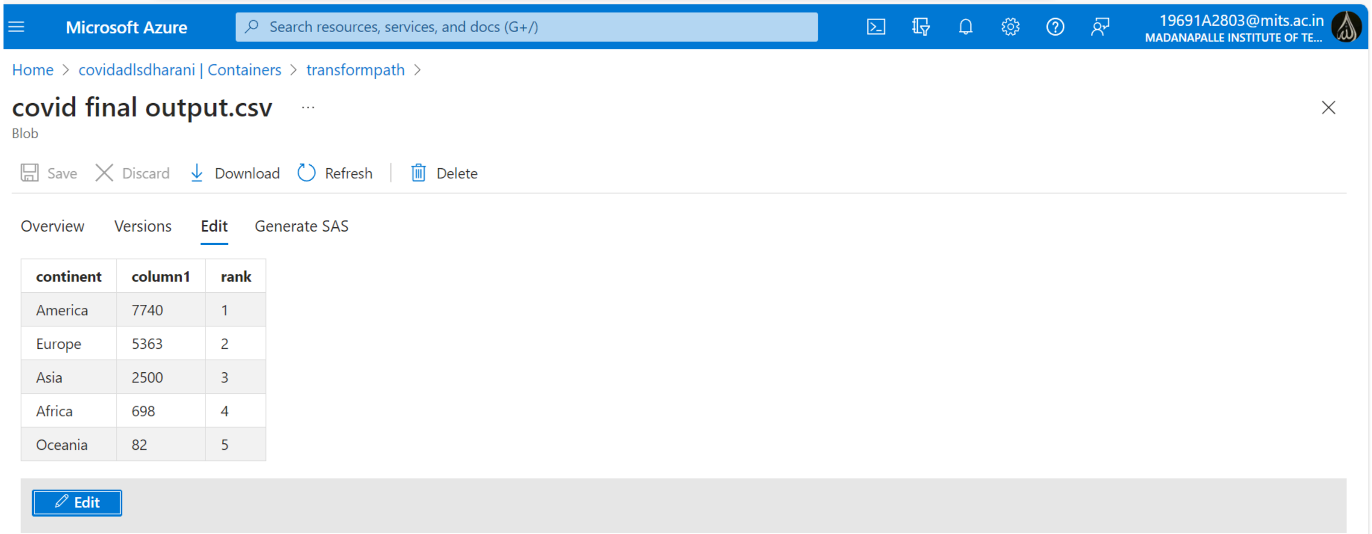
Showing 1 - 1 of 1 items

Activity name	Activity status	Activity type	Run start	Duration	Integr
Data flow transformation	✓ Succeeded	Data flow	3/11/2024, 4:41:05 PM	1m 37s	debug

Step 27 : After clicking debug, drag the dataflow in pipeline and run the pipeline. you will get the output once it is run successfully.

v

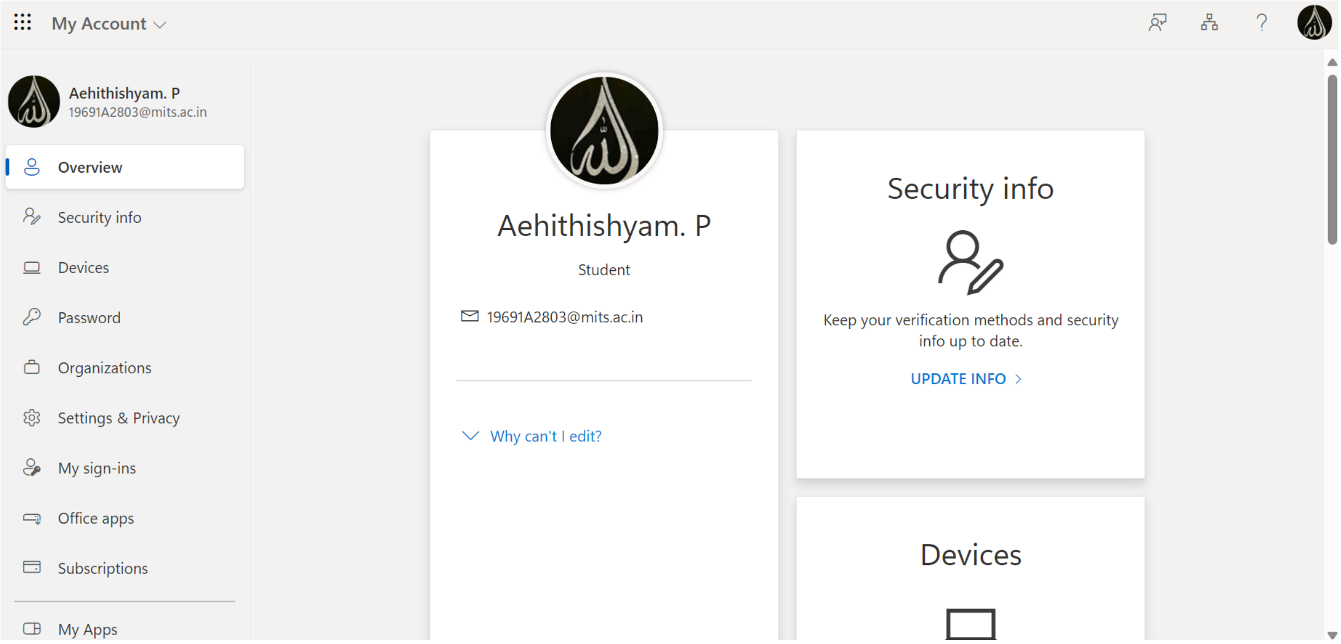
Step 28 : After that we need to check file appear in the transformpath container in storage account (data lake). I successfully got that file in my container as per the question given in the project documentation.


Microsoft Azure
Search resources, services, and docs (G+/)
19691A2803@mits.ac.in
MADANAPALLE INSTITUTE OF TE...
Home > covidadlsdharani | Containers > transformpath >
Blob
Save
Discard
Download
Refresh
Delete
Overview
Versions
Edit
Generate SAS
Edit

covid final output.csv

continent	column1	rank
America	7740	1
Europe	5363	2
Asia	2500	3
Africa	698	4
Oceania	82	5

MY DETAILS:



Aehithishyam. P
19691A2803@mits.ac.in

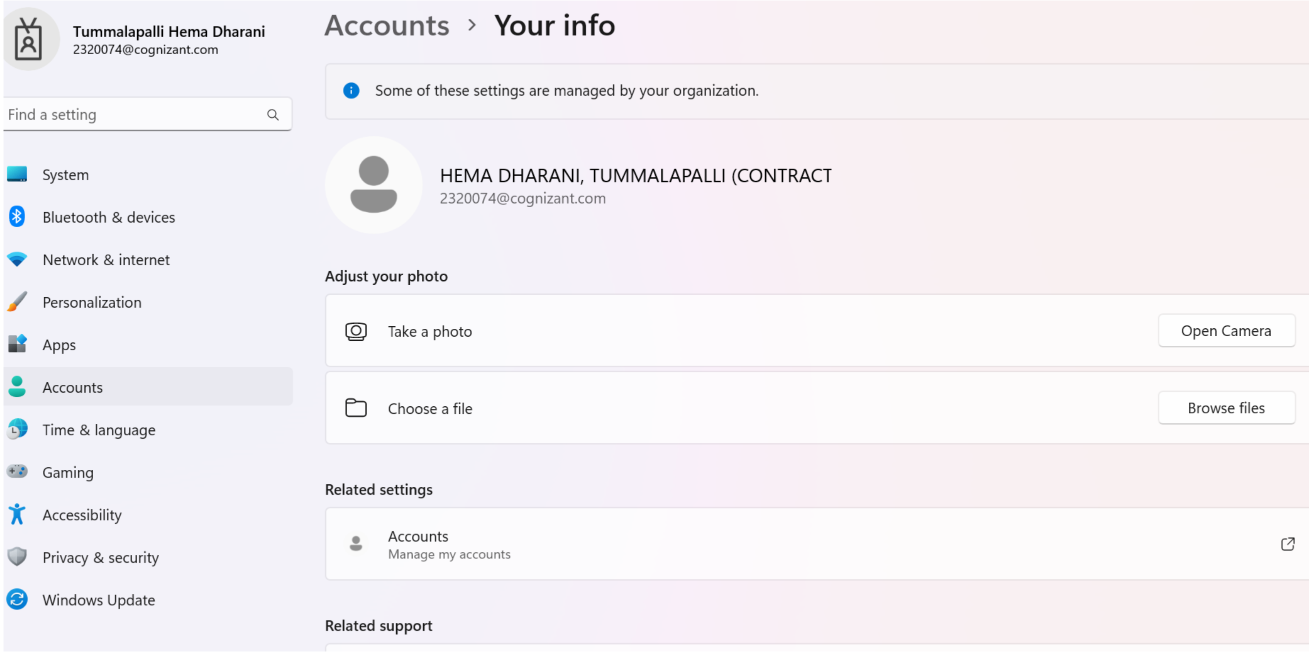
Aehithishyam. P

Student

19691A2803@mits.ac.in

Security info

Keep your verification methods and security info up to date.



Tummalapalli Hema Dharani
2320074@cognizant.com

Accounts > Your info

Some of these settings are managed by your organization.

HEMA DHARANI, TUMMALAPALLI (CONTRACT
2320074@cognizant.com

Adjust your photo

Take a photo

Choose a file

Related settings

Accounts
Manage my accounts