

AZURE DATA FACTORY

CAPSTONE – COVID USE CASE

Name: Majji Vijay Vamsi (Contractor)

Employee ID: 2320213

Cohort ID: CSDAIA24AZ005



Introduction:

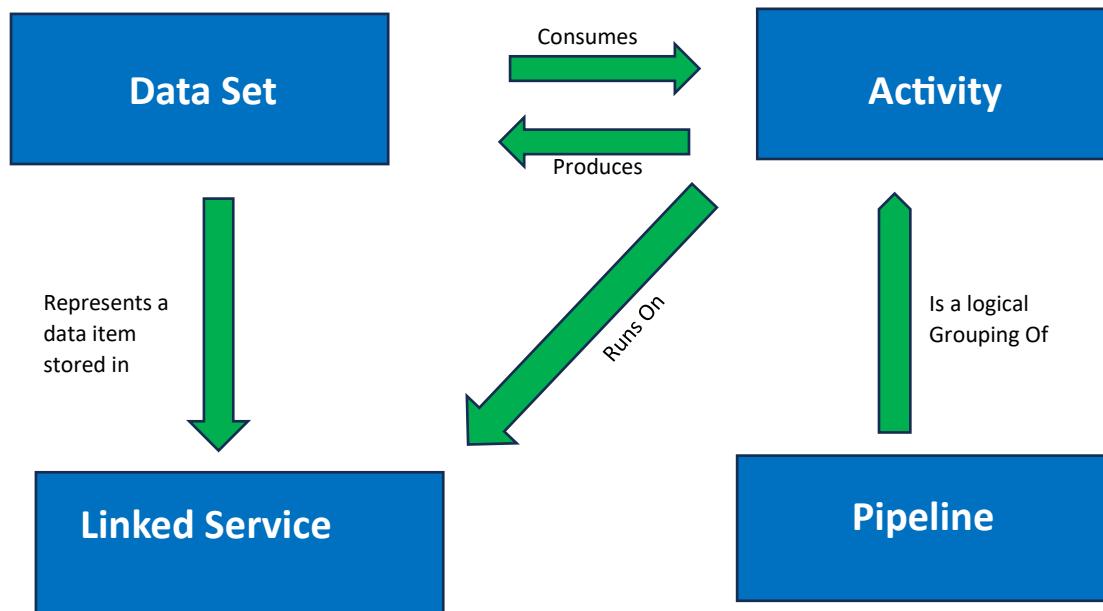
The Azure Tools which have been used for this project are Azure Data Factory, Azure Synapse Analytics and Azure Data Lake Storage " Azure Data Factory Capstone Covid Case Analysis". These components helps to extract data from its source storage (Data Lake), datasets are transformed in Data factory for 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:

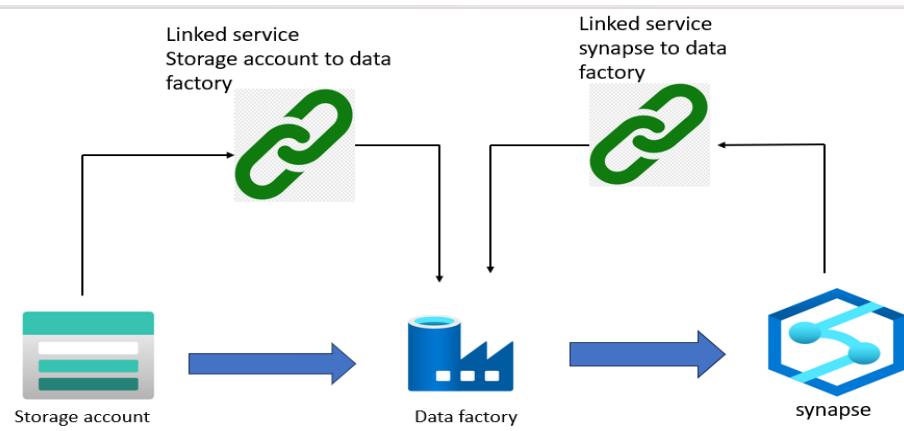
The main aim of this project (Covid Use Case Exercise) is that we will be having a hands-on experience on Storage, ADF Pipeline, Mapping Dataflow, Azure Synapse along with getting to know how to ingest data from flat files into Azure Data Lake Gen2 and Azure Synapse using Azure Data Factory (ADF) and 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:

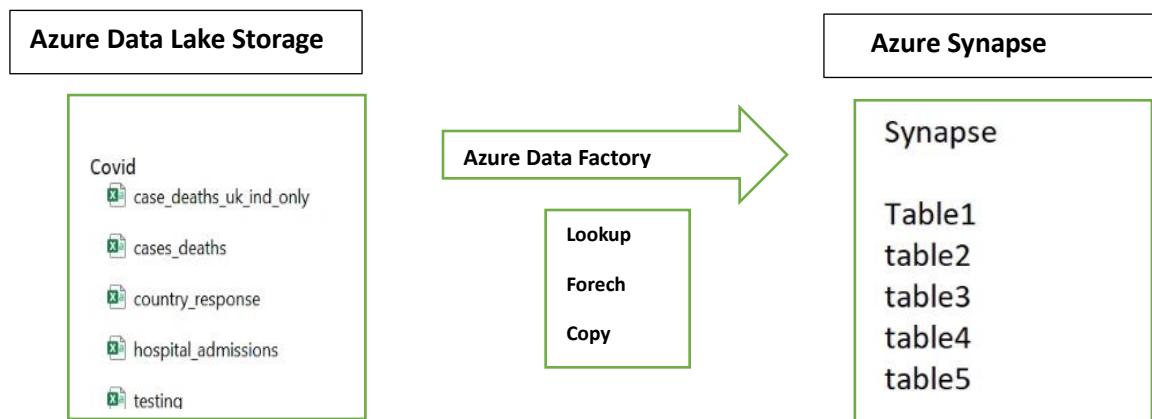


Project Requirements:

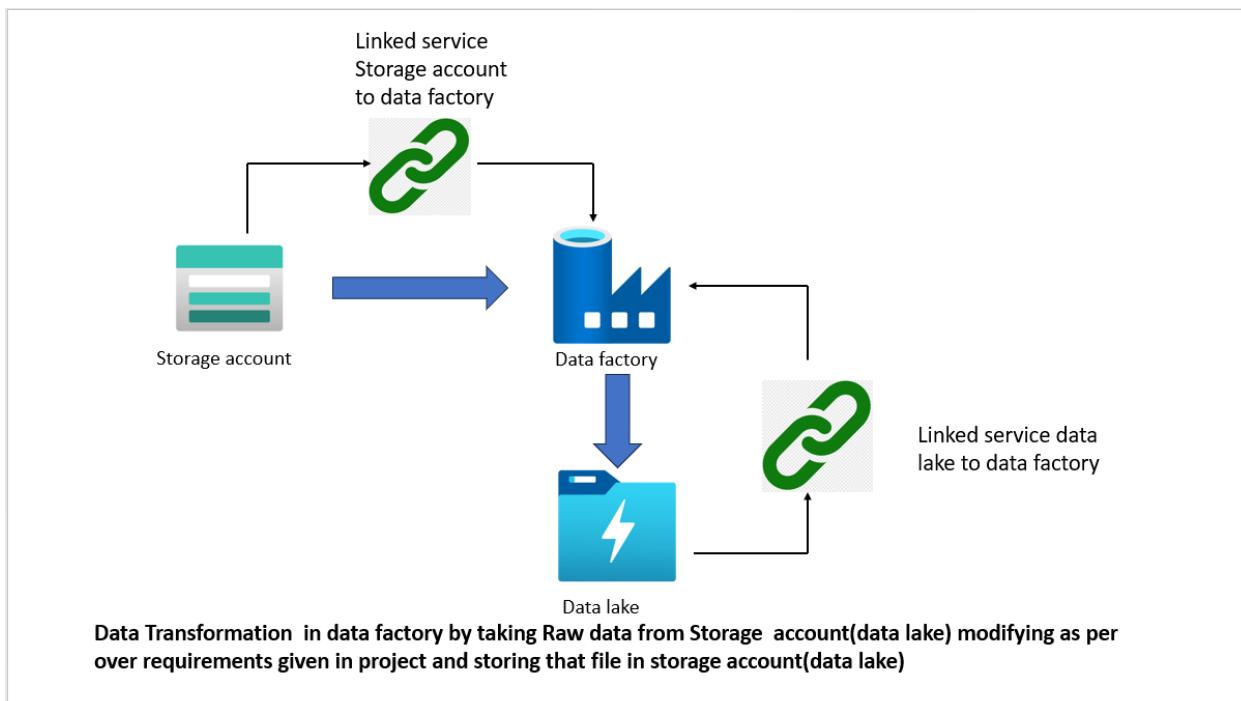
Requirement 1: Ingest raw flat files from data lake to synapse (data warehouse)

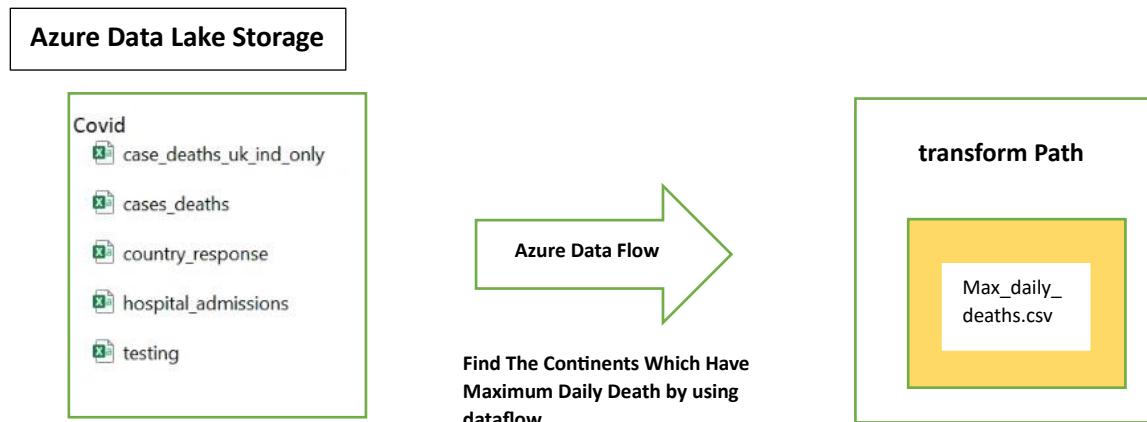


Ingesting process taken place in data factory by taking raw data from storage account(data lake) to relation database by creating tables in synapse(data warehouse)



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





Procedure for Requirement 1:

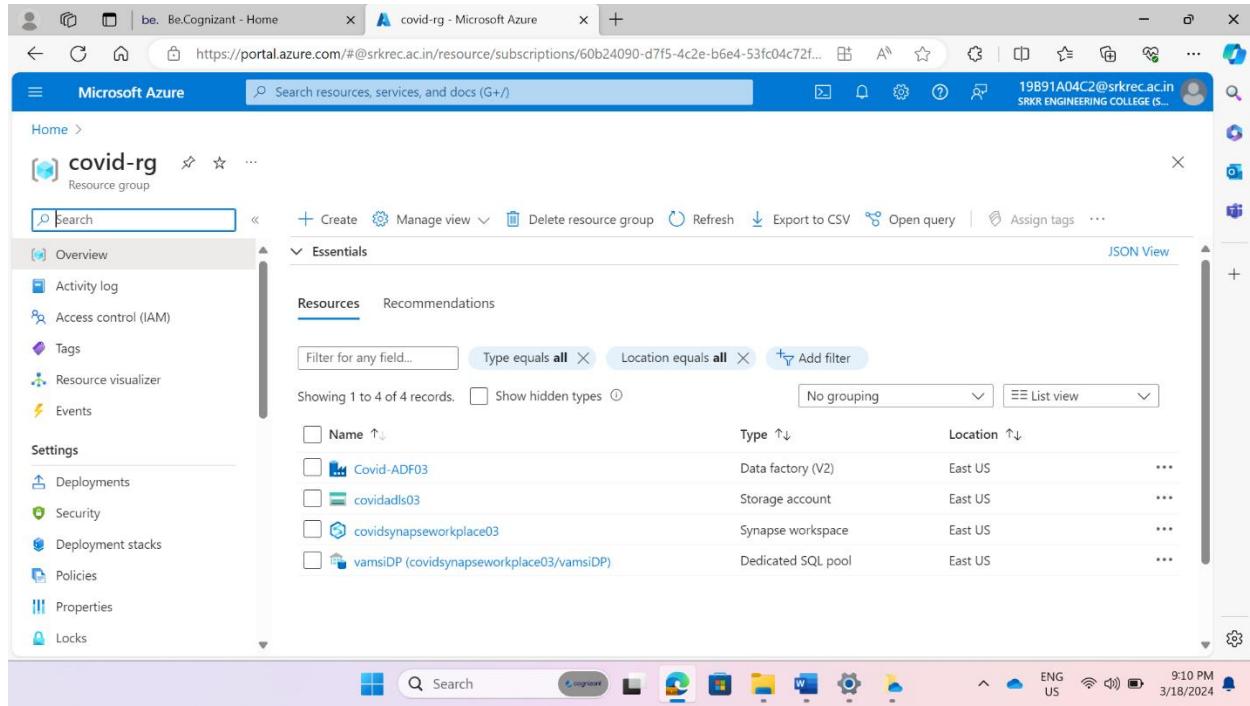
STEP 1: Open your Azure Account.

The screenshot shows the Microsoft Azure portal interface:

- Header:** Shows the URL <https://portal.azure.com/#home>, the user name **19B91A04C2@srkrec.ac.in SRKR ENGINEERING COLLEGE (S...)**, and a search bar.
- Left Sidebar:**
 - Azure services:** Includes icons for Create a resource, Azure Synapse Analytics, Data factories, Storage accounts, Resource groups, Quickstart Center, Virtual machines, App Services, SQL databases, and More services.
 - Resources:** A table showing recent resources:

Name	Type	Last Viewed
covidadls03	Storage account	5 days ago
Covid-ADF03	Data factory (V2)	5 days ago
covidsynapseworkspace03	Synapse workspace	5 days ago
covid-rg	Resource group	5 days ago
- Bottom Navigation:** Shows the URL <https://portal.azure.com/#create/hub>, the taskbar with various application icons, and system status indicators like battery level, signal strength, and date/time (9:08 PM, 3/18/2024).

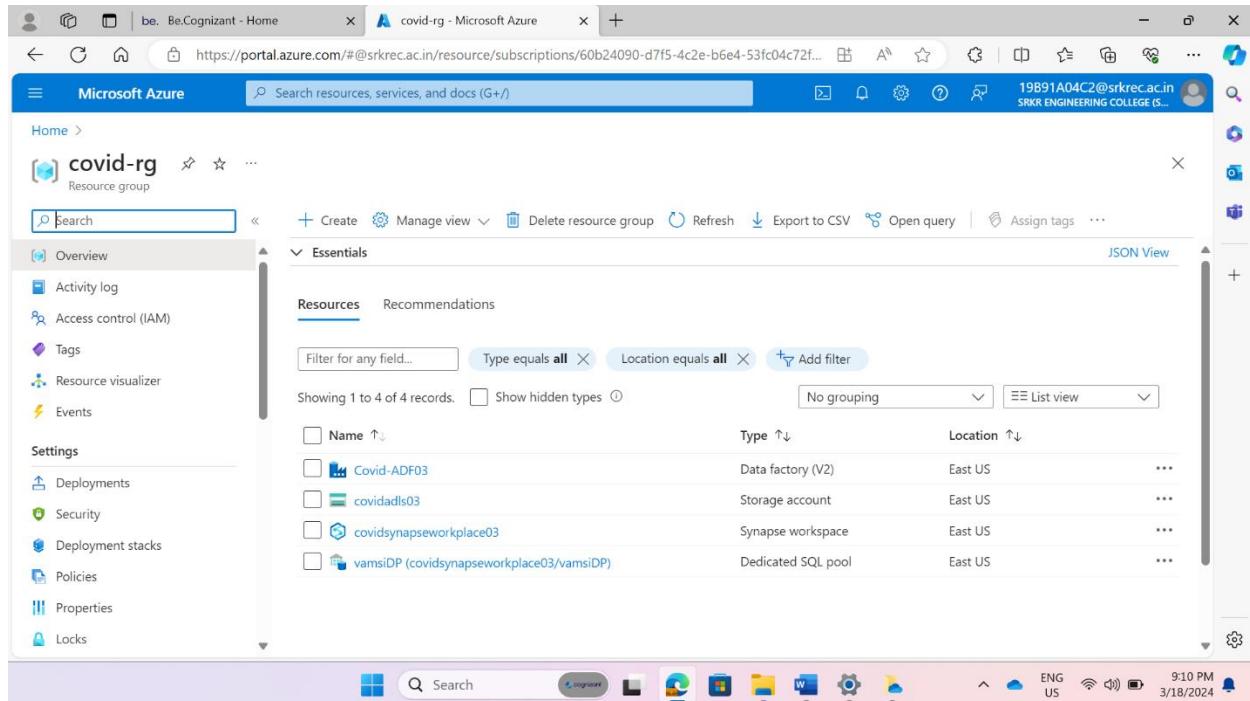
STEP 2: Create your own Resource Group.



The screenshot shows the Microsoft Azure portal interface. The left sidebar displays the 'covid-rg' resource group under 'Resource groups'. The main content area is titled 'Essentials' and shows a table of resources. The table has columns for Name, Type, and Location. The resources listed are:

Name	Type	Location
Covid-ADF03	Data factory (V2)	East US
covidadls03	Storage account	East US
covidsynapseworkspace03	Synapse workspace	East US
vamsiDP (covidsynapseworkspace03/vamsiDP)	Dedicated SQL pool	East US

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.



This screenshot is identical to the one above, showing the 'covid-rg' resource group overview in the Microsoft Azure portal. It displays the same list of resources: Covid-ADF03 (Data factory), covidadls03 (Storage account), covidsynapseworkspace03 (Synapse workspace), and vamsiDP (Dedicated SQL pool). The interface includes a search bar, filter options, and a JSON View link.

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.

The screenshot shows the Microsoft Azure portal interface for a storage account named 'covidadls03'. The left sidebar has 'Containers' selected under 'Data storage'. The main area displays a table of containers with columns: Name, Last modified, Anonymous access level, and Lease state. The 'covid' container is highlighted.

Name	Last modified	Anonymous access level	Lease state
\$logs	3/10/2024, 12:32:22 PM	Private	Available
covid	3/10/2024, 12:33:25 PM	Private	Available
transformpath	3/10/2024, 1:27:47 PM	Private	Available

The screenshot shows the Microsoft Azure portal interface for a container named 'covid' within the storage account 'covidadls03'. The left sidebar has 'Ingest' selected under 'Settings'. The main area displays a table of blobs with columns: Name, Modified, Access tier, Archive status, and Blob type. One blob named 'Ingest' is listed.

Name	Modified	Access tier	Archive status	Blob type
Ingest				

The screenshot shows the Microsoft Azure Storage Container blade for the 'covid' container. The left sidebar includes 'Overview', 'Diagnose and solve problems', 'Access Control (IAM)', 'Settings' (with 'Shared access tokens', 'Manage ACL', 'Access policy', 'Properties', and 'Metadata'), and a 'Search' bar. The main area displays a table of blobs with columns: Name, Modified, Access tier, Archive status, and Blob type. The blobs listed are:

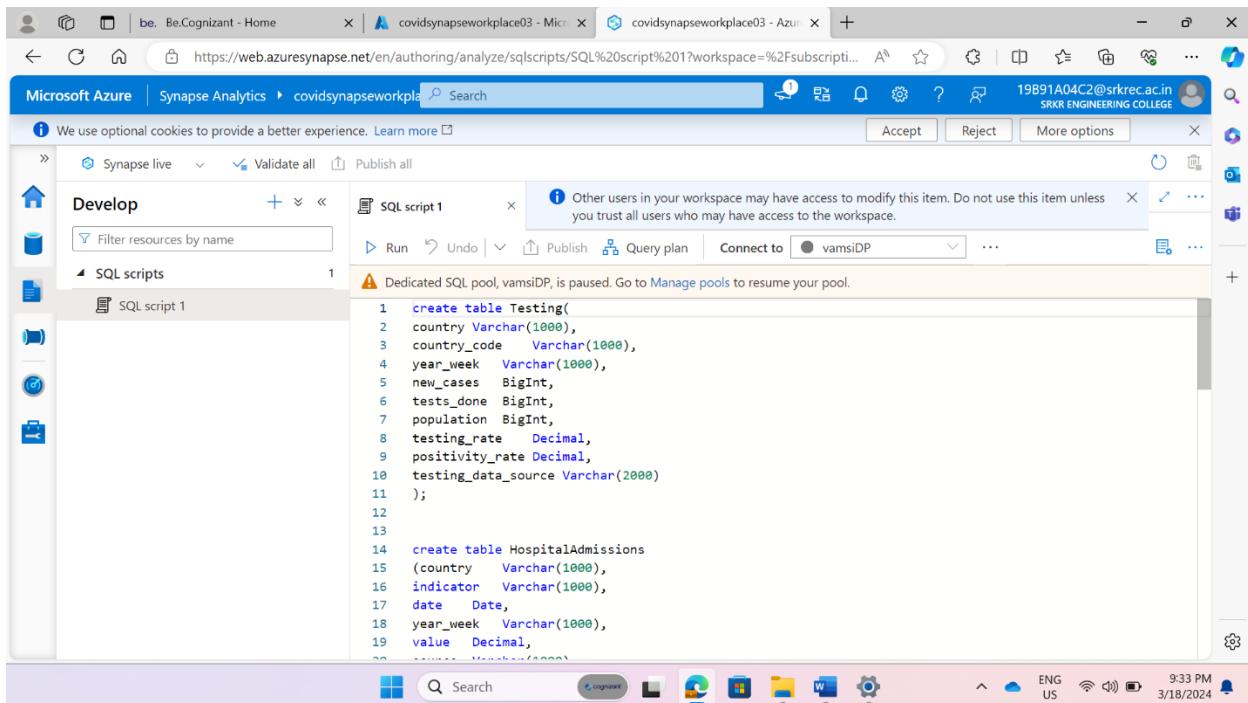
Name	Modified	Access tier	Archive status	Blob type
case_deaths_uk.ind_only.csv	3/10/2024, 12:34:29 ...	Hot (Inferred)		Block blob
cases_deaths.csv	3/10/2024, 12:34:34 ...	Hot (Inferred)		Block blob
country_response.csv	3/10/2024, 12:34:29 ...	Hot (Inferred)		Block blob
hospital_admissions.csv	3/10/2024, 12:34:31 ...	Hot (Inferred)		Block blob
testing.csv	3/10/2024, 12:34:30 ...	Hot (Inferred)		Block blob

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

The screenshot shows the Microsoft Azure Synapse workspace blade for 'covidsynapseworkspace03'. The left sidebar includes 'Overview', 'Activity log', 'Access control (IAM)', 'Tags', 'Diagnose and solve problems', 'Settings' (with 'Microsoft Entra ID', 'Properties', and 'Locks'), and 'Analytics pools' (with 'SQL pools' selected). The main area displays a table of SQL pools with columns: Name, Type, Status, and Size. The pools listed are:

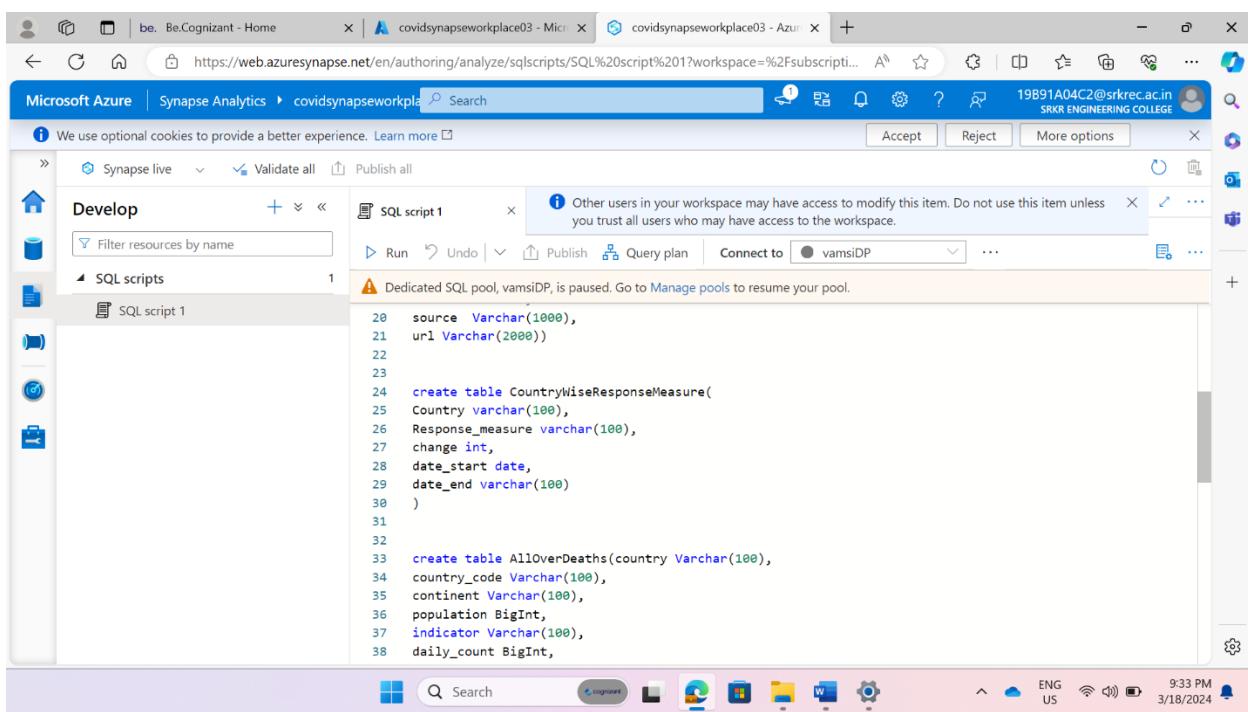
Name	Type	Status	Size
Built-in	Serverless	N/A	Auto
vamsiDP	Dedicated	Online	DW100c

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



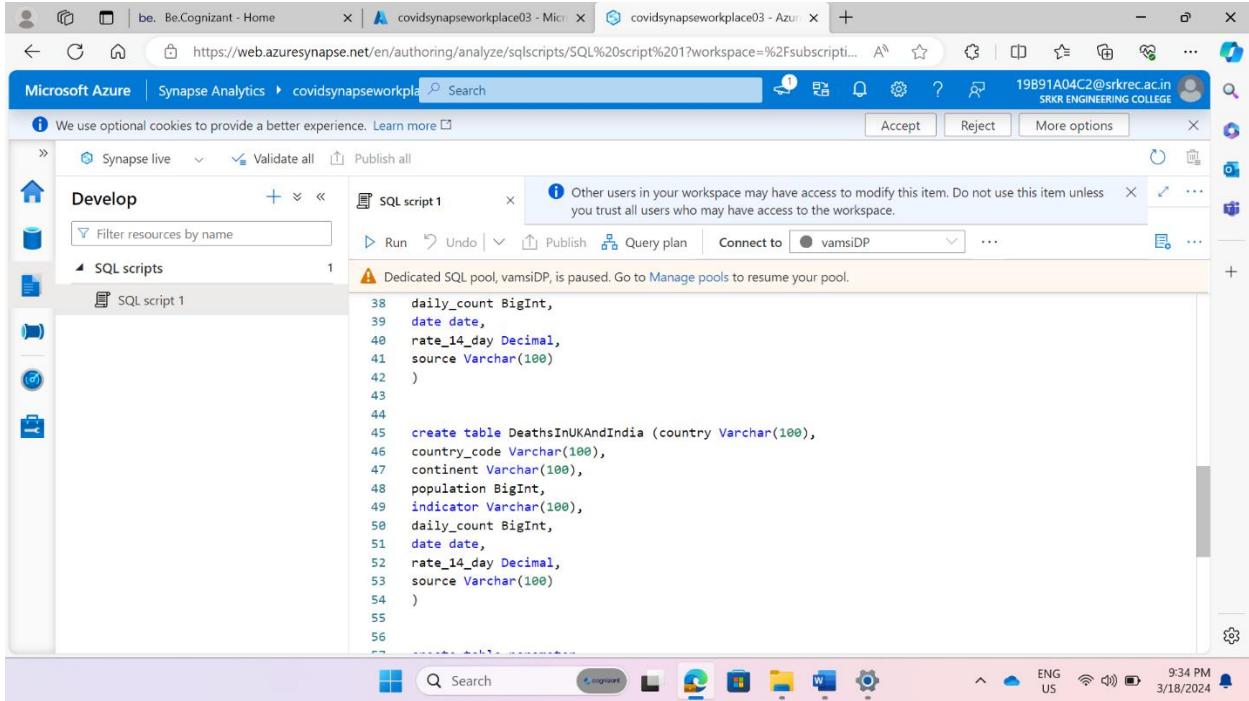
The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. The left sidebar has 'Develop' selected under 'SQL scripts'. A 'SQL script 1' tab is open, containing the following SQL code:

```
1  create table Testing(
2    country Varchar(1000),
3    country_code Varchar(1000),
4    year_week Varchar(1000),
5    new_cases BigInt,
6    tests_done BigInt,
7    population BigInt,
8    testing_rate Decimal,
9    positivity_rate Decimal,
10   testing_data_source Varchar(2000)
11 );
12
13
14  create table HospitalAdmissions
15  (country Varchar(1000),
16  indicator Varchar(1000),
17  date Date,
18  year_week Varchar(1000),
19  value Decimal,
```



The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. The left sidebar has 'Develop' selected under 'SQL scripts'. A 'SQL script 1' tab is open, containing the following SQL code:

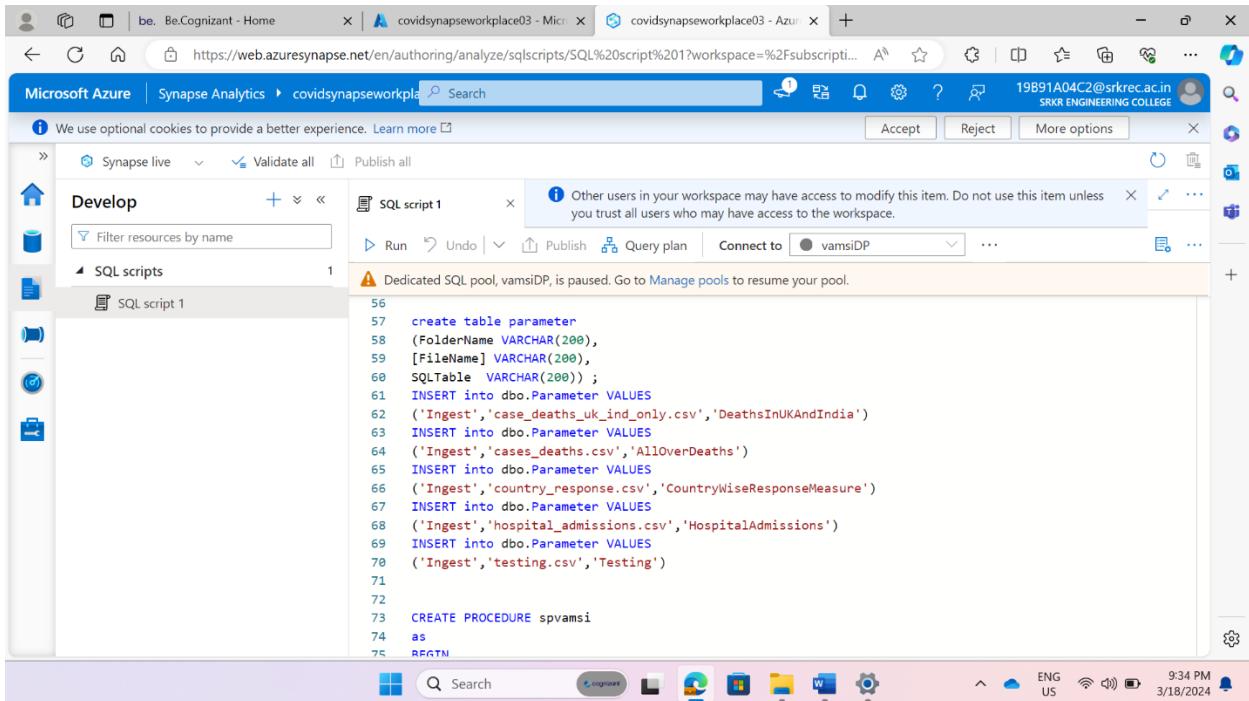
```
20  source Varchar(1000),
21  url Varchar(2000))
22
23
24  create table CountryWiseResponseMeasure(
25  Country varchar(100),
26  Response_measure varchar(100),
27  change int,
28  date_start date,
29  date_end varchar(100)
30 )
31
32
33  create table AllOverDeaths(country Varchar(100),
34  country_code Varchar(100),
35  continent Varchar(100),
36  population BigInt,
37  indicator Varchar(100),
38  daily_count BigInt,
```



The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. The main area displays a SQL script titled "SQL script 1". A warning message at the top right states: "Dedicated SQL pool, vamsiDP, is paused. Go to Manage pools to resume your pool." The script itself contains several lines of T-SQL code, including the creation of a table named "DeathsInUKAndIndia". The code is as follows:

```
38    daily_count BigInt,
39    date date,
40    rate_14_day Decimal,
41    source Varchar(100)
42  )
43
44
45  create table DeathsInUKAndIndia (country Varchar(100),
46  country_code Varchar(100),
47  continent Varchar(100),
48  population BigInt,
49  indicator Varchar(100),
50  daily_count BigInt,
51  date date,
52  rate_14_day Decimal,
53  source Varchar(100)
54  )
55
56
```

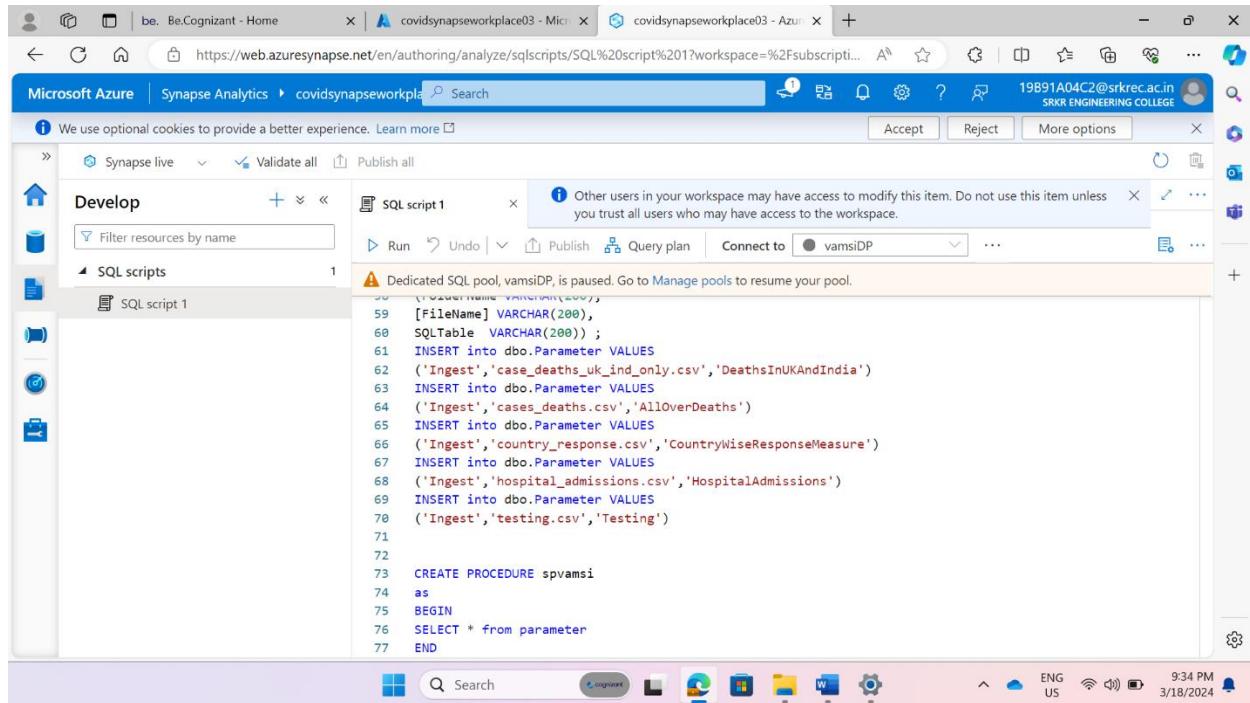
STEP 7: We need to create parameter Table and we need to Insert values to it.



The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. The main area displays a SQL script titled "SQL script 1". A warning message at the top right states: "Dedicated SQL pool, vamsiDP, is paused. Go to Manage pools to resume your pool." The script contains several lines of T-SQL code, including the creation of a table named "parameter" and multiple INSERT statements into it. The code is as follows:

```
56
57  create table parameter
58  ([FolderName] VARCHAR(200),
59  [FileName] VARCHAR(200),
60  SQLTable VARCHAR(200)) ;
61  INSERT into dbo.Parameter VALUES
62  ('Ingest','case_deaths_uk_ind_only.csv','DeathsInUKAndIndia')
63  INSERT into dbo.Parameter VALUES
64  ('Ingest','cases_deaths.csv','AllOverDeaths')
65  INSERT into dbo.Parameter VALUES
66  ('Ingest','country_response.csv','CountryWiseResponseMeasure')
67  INSERT into dbo.Parameter VALUES
68  ('Ingest','hospital_admissions.csv','HospitalAdmissions')
69  INSERT into dbo.Parameter VALUES
70  ('Ingest','testing.csv','Testing')
71
72
73  CREATE PROCEDURE spvamsi
74  as
75  RPTN
```

STEP 8: We need to create a Stored Procedure.



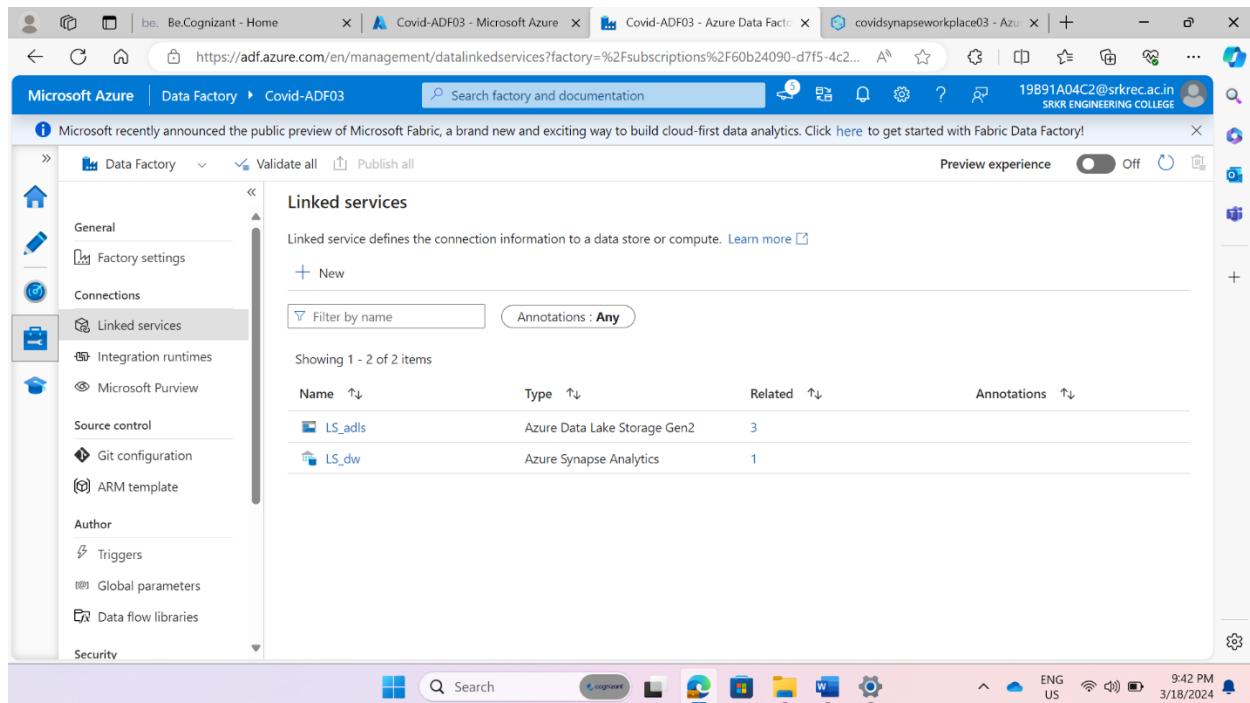
The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. The left sidebar has 'Develop' selected under 'SQL scripts'. A 'SQL script 1' tab is open, containing the following SQL code:

```
56  [vamsiName VARCHAR(200),
57  [FileName] VARCHAR(200),
58  SQLTable VARCHAR(200));
59  INSERT into dbo.Parameter VALUES
60  ('Ingest','case_deaths_uk_ind_only.csv','DeathsInUKAndIndia')
61  INSERT into dbo.Parameter VALUES
62  ('Ingest','cases_deaths.csv','AllOverDeaths')
63  INSERT into dbo.Parameter VALUES
64  ('Ingest','country_response.csv','CountryWiseResponseMeasure')
65  INSERT into dbo.Parameter VALUES
66  ('Ingest','hospital_admissions.csv','HospitalAdmissions')
67  INSERT into dbo.Parameter VALUES
68  ('Ingest','testing.csv','Testing')
69
70
71
72
73  CREATE PROCEDURE spvamsi
74  as
75  BEGIN
76  SELECT * from parameter
77  END
```

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.



The screenshot shows the Microsoft Azure Data Factory workspace interface. The left sidebar has 'Data Factory' selected under 'General'. Under 'Connections', 'Linked services' is selected. The main area shows the 'Linked services' page with the following details:

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

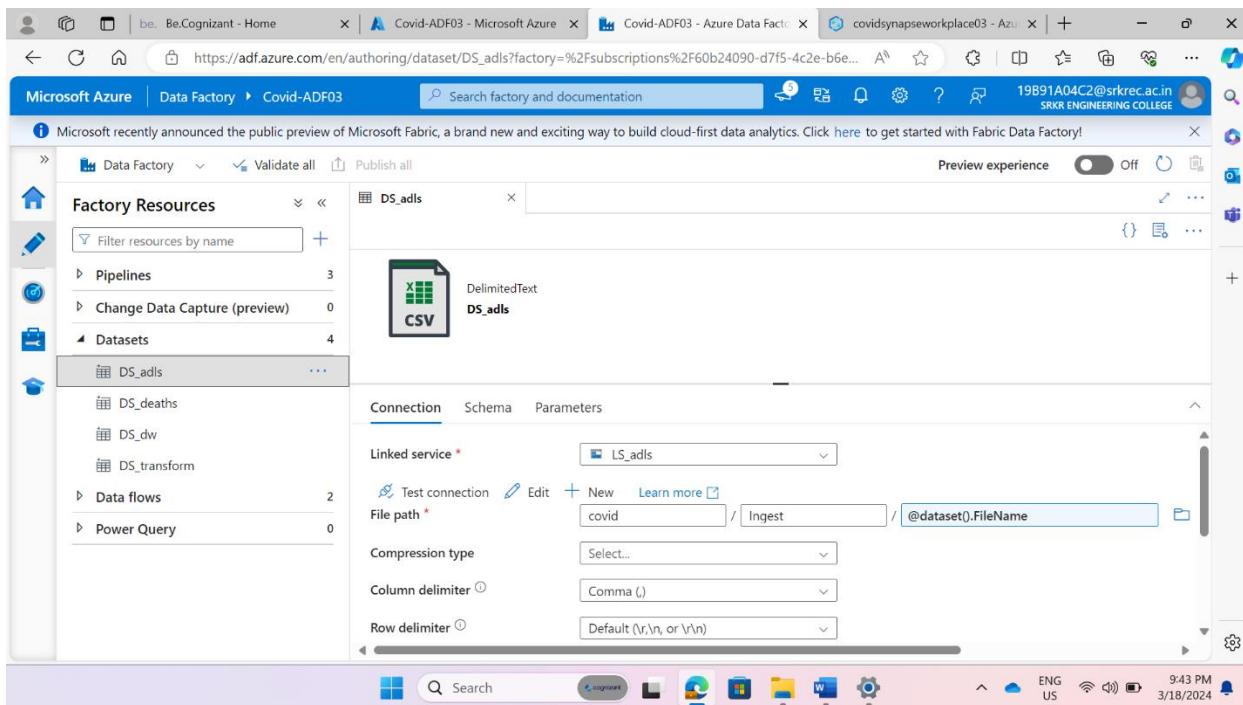
+ New

Filter by name Annotations : Any

Showing 1 - 2 of 2 items

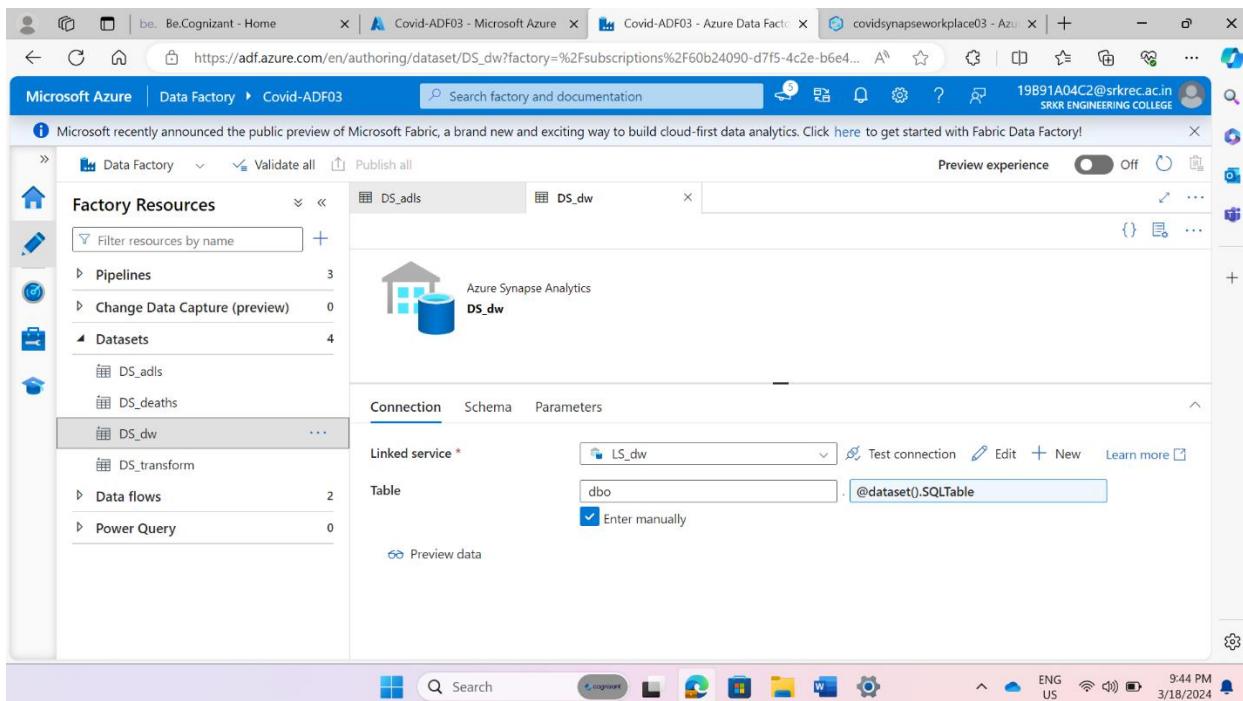
Name	Type	Related	Annotations
LS_adls	Azure Data Lake Storage Gen2	3	
LS_dw	Azure Synapse Analytics	1	

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



The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Change Data Capture (preview), Datasets, Data flows, and Power Query. Under 'Datasets', 'DS_adls' is selected. The main panel displays the 'DS_adls' dataset configuration. It shows a preview icon of a CSV file and the name 'DS_adls'. Below this, the 'Connection' tab is selected, showing 'Linked service' set to 'LS_adls'. The 'File path' field contains 'covid' / 'Ingest' / '@dataset().FileName'. Other settings include 'Compression type' (Select...), 'Column delimiter' (Comma (,), selected), and 'Row delimiter' (Default (\r\n, or \n\)). The top navigation bar shows the URL 'https://adf.azure.com/en/authoring/dataset/DS_adls?factory=%2Fsubscriptions%2F60b24090-d7f5-4c2e-b6e...', and the top right corner shows the user '19B91A04C2@srkrec.ac.in' and 'SRK ENGINEERING COLLEGE'.

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



The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Change Data Capture (preview), Datasets, Data flows, and Power Query. Under 'Datasets', 'DS_adls' is selected, followed by 'DS_dw'. The main panel displays the 'DS_dw' dataset configuration. It shows a preview icon of a database and the name 'DS_dw'. Below this, the 'Connection' tab is selected, showing 'Linked service' set to 'LS_dw'. The 'Table' field contains 'dbo' / '@dataset().SQLTable'. A checkbox 'Enter manually' is checked. The top navigation bar shows the URL 'https://adf.azure.com/en/authoring/dataset/DS_dw?factory=%2Fsubscriptions%2F60b24090-d7f5-4c2e-b6e4...', and the top right corner shows the user '19B91A04C2@srkrec.ac.in' and 'SRK ENGINEERING COLLEGE'.

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

The screenshot shows the Microsoft Azure Data Factory Pipeline Editor. A pipeline named "Req1" is selected. The pipeline structure is as follows:

```

graph TD
    Lookup1[Lookup] --> ForEach1[ForEach]
    ForEach1 --> CopyData1[Copy data]
  
```

The "Source dataset" for the Lookup activity is set to "DS_dw". The pipeline properties table shows:

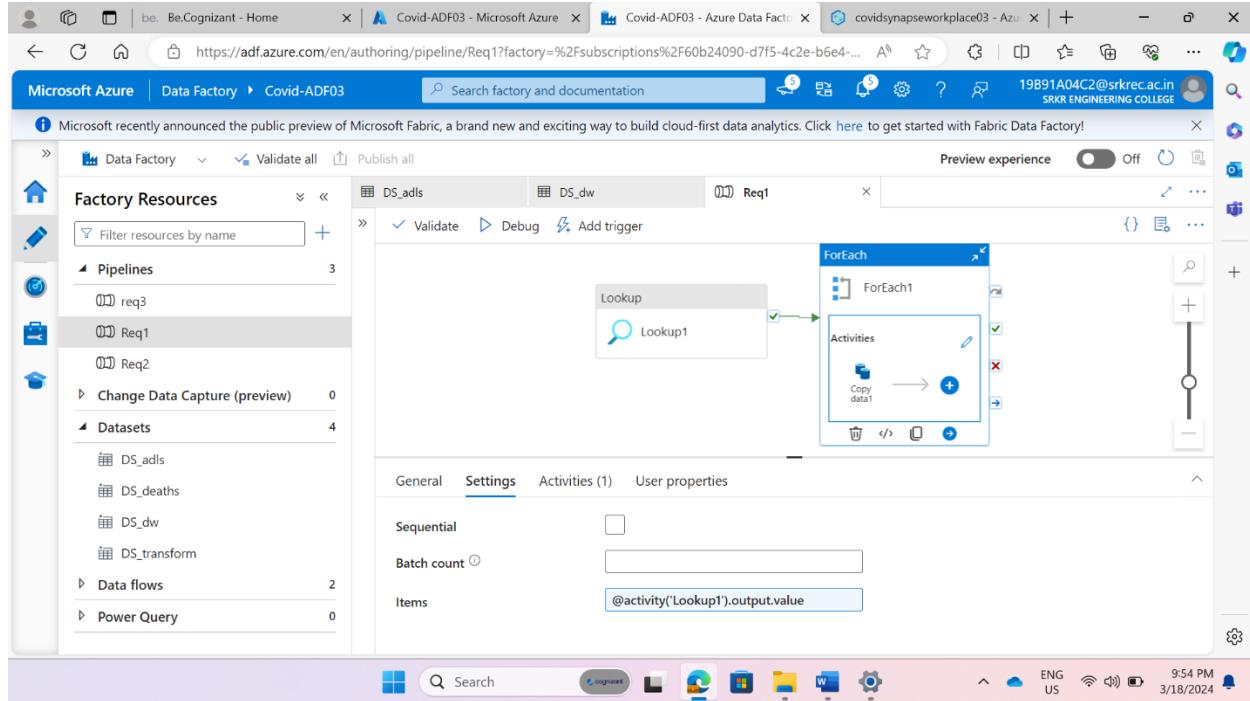
Name	Type
SQLTable	string

The screenshot shows the Microsoft Azure Data Factory Pipeline Editor with the "Settings" tab selected for pipeline "Req1". The "Source dataset" is set to "DS_dw". The "Dataset properties" section shows:

Name	Type
SQLTable	string

The "Stored procedure name" field is set to "[dbo].[spvamsi]". Below it, there is a note: "Failed More" and a checkbox "Enter manually".

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}),

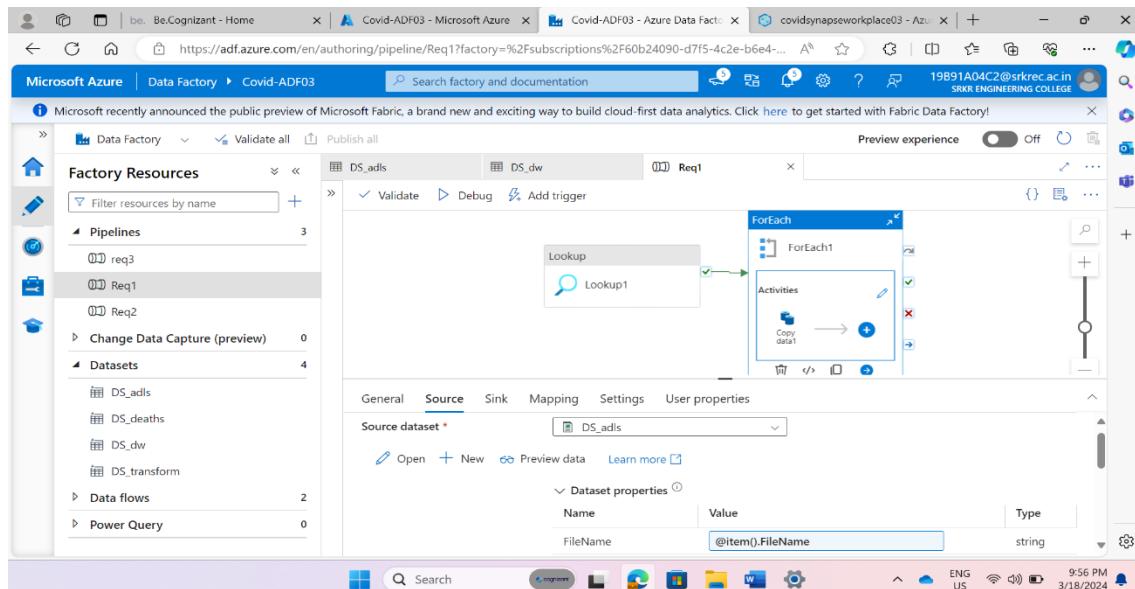


Table Name (@{item().SQLTable})

The screenshot shows the Microsoft Azure Data Factory Pipeline Editor. A pipeline named "Req1" is selected. The pipeline structure is as follows:

```

graph LR
    Lookup1[Lookup] --> ForEach1[ForEach]
    subgraph ForEach1 [ForEach1]
        direction TB
        subgraph Activities [Activities]
            CopyData1[Copy data1]
        end
    end

```

The "Sink" tab is selected in the pipeline editor. The "Sink dataset" is set to "DS_dw". In the "Sink dataset properties" table, the "Name" column is "SQLTable" and the "Value" column is "@{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.

The screenshot shows the Microsoft Azure Data Factory Pipeline Editor after the pipeline has run successfully. The pipeline status is "Succeeded". The output table shows the results of the activities:

Activity name	Activity status	Activity type	Run start	Duration	Inte
Copy data1	Succeeded	Copy data	3/18/2024, 10:12:47 PM	1m 0s	Aut
Copy data1	Succeeded	Copy data	3/18/2024, 10:12:47 PM	49s	Aut
Copy data1	Succeeded	Copy data	3/18/2024, 10:12:47 PM	50s	Aut
Copy data1	Succeeded	Copy data	3/18/2024, 10:12:47 PM	50s	Aut
Copy data1	Succeeded	Copy data	3/18/2024, 10:12:46 PM	50s	Aut
ForEach1	Succeeded	ForEach	3/18/2024, 10:12:00 PM	1m 48s	
Lookup1	Succeeded	Lookup	3/18/2024, 10:09:10 PM	40s	Aut

REQUIREMENT 1 OUTPUTS:

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace interface. The left sidebar displays the 'Data' section with 'Workspace' selected, showing a list of databases, tables, and external resources. The main area shows the results of 'SQL script 1'. The results table has columns: country, country_code, continent, population, indicator, daily_count, and date. The data includes records for Africa (total), Albania, Eswatini, Bulgaria, and Burkina Faso. A message at the top right indicates that other users in the workspace may have access to modify the item. The status bar at the bottom right shows the time as 10:23 PM and the date as 3/18/2024.

country	country_code	continent	population	indicator	daily_count	date
Africa (total)	(NULL)	Africa	1339423921	deaths	204	2020-06-28T00:00:00
Albania	ALB	Europe	2862427	confirmed cases	23	2020-05-31T00:00:00
Eswatini	SWZ	Africa	1160164	deaths	0	2020-03-15T00:00:00
Bulgaria	BGR	Europe	7000039	deaths	0	2020-04-03T00:00:00
Burkina Faso	BFA	Africa	20903278	confirmed cases	0	2020-03-10T00:00:00

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace interface. The left sidebar displays the 'Data' section with 'Workspace' selected, showing a list of databases, tables, and external resources. The main area shows the results of 'SQL script 2'. The results table has columns: Country, Response_measure, change, date_start, and date_end. The data includes records for Austria, Poland, and Austria again. A message at the top right indicates that other users in the workspace may have access to modify the item. The status bar at the bottom right shows the time as 10:23 PM and the date as 3/18/2024.

Country	Response_measure	change	date_start	date_end
Austria	AdaptationOfWorkplace	1	2020-03-10T00:00:00	NA
Poland	OutdoorOver5	1	2020-04-02T00:00:00	2020-05-29
Austria	AdaptationOfWorkplace	1	2020-03-10T00:00:00	NA
Poland	OutdoorOver5	1	2020-04-02T00:00:00	2020-05-29

Microsoft Azure | Synapse Analytics > covidsynapseworkspace03 | Search

We use optional cookies to provide a better experience. Learn more | Accept | Reject | More options

Synapse live | Validate all | Publish all

Data | Workspace | Linked

SQL database: vamsiDP (SQL)

- Tables
 - dbo.AllOverDeaths
 - dbo.CountryWiseResponse...
 - dbo.DeathsInUKAndIndia
 - dbo.HospitalAdmissions
 - dbo.parameter
 - dbo.Testing
- External tables
- External resources

SQL script 3 | SQL script 4

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	2020-01-02T00:00:00Z
India	IND	Asia	1380004385	confirmed cases	81484	2020-10-02T00:00:00Z
India	IND	Asia	1380004385	deaths	776	2020-09-29T00:00:00Z
United Kingdom	GBR	Europe	66647112	confirmed cases	1040	2020-08-17T00:00:00Z
United Kingdom	GBR	Europe	66647112	deaths	44	2020-07-15T00:00:00Z

00:00:02 Query executed successfully.

Search | Log out | Microsoft Edge | Task View | Start | File Explorer | This PC | File History | OneDrive | Taskbar | System tray | 10:24 PM | 3/18/2024

Microsoft Azure | Synapse Analytics > covidsynapseworkspace03 | Search

We use optional cookies to provide a better experience. Learn more | Accept | Reject | More options

Synapse live | Validate all | Publish all

Data | Workspace | Linked

SQL database: vamsiDP (SQL)

- Tables
 - dbo.AllOverDeaths
 - dbo.CountryWiseResponse...
 - dbo.DeathsInUKAndIndia
 - dbo.HospitalAdmissions
 - dbo.parameter
 - dbo.Testing
- External tables
- External resources

SQL script 4 | SQL script 5

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...	2020-04-02T00:00:00Z	2020-W14	1057	Surveillance	https://www.so...
Austria	Daily ICU occu...	2020-09-04T00:00:00Z	2020-W36	28	Country_Website	https://info.ges...
Belgium	Daily ICU occu...	2020-03-18T00:00:00Z	2020-W12	131	Country_Website	https://epistat...
Bulgaria	Daily hospital o...	2020-04-25T00:00:00Z	2020-W17	292	External_Github	https://github.c...
Bulgaria	Daily ICU occu...	2020-05-29T00:00:00Z	2020-W22	20	External_Github	https://github.c...

00:00:02 Query executed successfully.

Search | Log out | Microsoft Edge | Task View | Start | File Explorer | This PC | File History | OneDrive | Taskbar | System tray | 10:24 PM | 3/18/2024

The screenshot shows a Microsoft Azure Synapse Analytics workspace titled 'covidsynapseworkspace03'. A query has been run against a database named 'vamsiDP'. The results are displayed in a table format:

country	country_code	year_week	new_cases	tests_done	population	testing_rate
Austria	AT	2020-W15	2041	12339	8858775	139
France	FR	2020-W30	5854	459896	67012883	686
Luxembourg	LU	2020-W39	593	36818	613894	5997
Sweden	SE	2020-W32	1979	53772	10230185	525
Austria	AT	2020-W15	2041	12339	8858775	139

At the bottom of the results pane, a message indicates: "00:00:02 Query executed successfully."

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.

The screenshot shows the 'Containers' blade for a storage account named 'covidadls03'. The blade lists three containers: '\$logs', 'covid', and 'transformpath'. The 'transformpath' container was created specifically for storing transformed data files.

Name	Last modified	Anonymous access level	Lease state
\$logs	3/10/2024, 12:32:22 PM	Private	Available
covid	3/10/2024, 12:33:25 PM	Private	Available
transformpath	3/10/2024, 1:27:47 PM	Private	Available

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.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. Under Datasets, 'DS_adls' is expanded, showing 'DS_deaths'. The main workspace displays 'DS_deaths' as a DelimitedText CSV dataset. The 'Connection' tab is selected, showing 'LS_adls' as the linked service. The 'File path' field contains 'covid / Ingest / cases_deaths.csv'. Other settings like Compression type, Column delimiter (Comma (,),), and Row delimiter (Default (\r\n, or \n)) are visible. The top navigation bar shows the URL 'https://adf.azure.com/en/authoring/dataset/DS_deaths?factory=%2Fsubscriptions%2F60b24090-d7f5-4c2e-b...' and the title 'Covid-ADF03 - Azure Data Factory'.

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

The screenshot shows the Microsoft Azure Data Factory interface. The 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. Under Datasets, 'DS_adls' is expanded, showing 'DS_deaths' and 'DS_dw'. The main workspace displays 'DS_transform' as a DelimitedText CSV dataset. The 'Connection' tab is selected, showing 'LS_adls' as the linked service. The 'File path' field contains 'transformpath / Directory / File name'. Other settings like Compression type, Column delimiter (Comma (,),), and Row delimiter (Default (\r\n, or \n)) are visible. The top navigation bar shows the URL 'https://adf.azure.com/en/authoring/dataset/DS_transform?factory=%2Fsubscriptions%2F60b24090-d7f5-4c2e-b...' and the title 'Covid-ADF03 - Azure Data Factory'.

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 | Preview experience Off

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!

Data Factory | Validate all | Publish all

Factory Resources

- 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 | DS_deaths | DS_transform | req3 | dataflow1

Aggregate settings | Optimize | Inspect | Data preview | Learn more

Output stream name: aggregate1 | Description: Aggregating data by 'continent' producing columns 'daily_count'

Incoming stream: filter1

Group by | Aggregates

Columns	Name as
abc continent	continent

10:41 PM 3/18/2024

Microsoft Azure | Data Factory | Covid-ADF03 | Search factory and documentation | Preview experience Off

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!

Data Factory | Validate all | Publish all

Factory Resources

- 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 | DS_deaths | DS_transform | req3 | dataflow1

Aggregate settings | Optimize | Inspect | Data preview | Learn more

Incoming stream: filter1

Grouped by: continent

Add | Clone | Delete | Open expression builder

Column	Expression
daily_count	max(daily_count)

10:41 PM 3/18/2024

Microsoft Azure | Data Factory > Covid-ADF03 | Search factory and documentation

1 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!

Preview experience Off

Factory Resources

Pipelines: req3, Req1, Req2

Change Data Capture (preview): 0

Datasets: DS_adls, DS_deaths, DS_dw, DS_transform

Data flows: dataflow1, dataflowCovid

DS_deaths DS_transform req3 dataflow1

Validate Data flow debug

Rank settings Optimize Inspect Data preview

Incoming stream: aggregate1

Options: Case insensitive, Dense

Rank column: Rank

Sort conditions: aggregate1's column, daily_count, Descending

Microsoft Azure | Data Factory > Covid-ADF03 | Search factory and documentation

1 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!

Preview experience Off

Factory Resources

Pipelines: req3, Req1, Req2

Change Data Capture (preview): 0

Datasets: DS_adls, DS_deaths, DS_dw, DS_transform

Data flows: dataflow1, dataflowCovid

DS_deaths DS_transform req3 dataflow1

Validate Data flow debug Debug Settings

Sink Settings Errors Mapping Optimize Inspect Data preview

Incoming stream: rank1

Sink type: Dataset, Inline, Cache

Dataset: DS_transform

Skip line count:

Options: Allow schema drift (checked), Validate schema

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

The screenshot shows the Microsoft Azure Data Factory pipeline editor. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (req3, Req1, Req2), 'Datasets' (DS_adls, DS_deaths, DS_dw, DS_transform), and 'Data flows' (dataflow1, dataflowCovid). The main workspace displays a dataflow named 'dataflow1' with the following stages:

- source1: Import data from DS_deaths
- filter1: Filtering rows using expression on columns 'Indicator'
- aggregate1: Aggregating data by column producing column 'daily_count'
- rank1: Ranking rows on columns 'daily_count'
- sink1: Sink

The 'Data preview' tab is selected, showing the following data table:

continent	abc ↑ daily_count abc ↑	Rank abc ↑↓
America	7740	1
Europe	5363	2
Asia	2500	3
Africa	698	4
Oceania	60	5

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

The screenshot shows the Microsoft Azure Data Factory pipeline editor. The 'Activities' pane on the left lists various services: Azure Data Explorer, Azure Function, Batch Service, Databricks, Data Lake Analytics, General, HDInsight, Iteration & conditionals, Machine Learning, and Power Query. The main workspace shows the pipeline 'req3' with its activities:

- DS_deaths
- DS_transform
- req3
- dataflow1

The 'req3' activity is currently running, indicated by a green progress bar. The 'Pipeline status' is shown as 'Succeeded'. The 'Output' section displays the following details:

- Pipeline run ID: 87ef30dd-250e-4519-9da6-fc41dc4f72ec
- Pipeline status: Succeeded
- Activity name: dataflow1
- Activity status: Succeeded
- Activity type: Data flow
- Run start: 3/18/2024, 10:56:39 F

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.

The screenshot shows the Microsoft Azure Storage Blob Properties page. The URL in the address bar is https://portal.azure.com/#view/Microsoft_Azure_Storage/BlobPropertiesBladeV2/storageAccountId/%23. The page displays a table with the following data:

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

MY AZURE DETAILS:

The screenshot shows the Microsoft My Account page. The URL in the address bar is <https://myaccount.microsoft.com/?ref=MeControl>. The page displays the following information:

- My Account** dropdown menu:
 - MAJJI VENKATA...
 - 19B91A04C2@srkrec.ac.in
- Overview** section:
 - Security info
 - Devices
 - Password
 - Organizations
 - Settings & Privacy
 - My sign-ins
 - My Apps
 - My Groups
- MAJJI VENKATA ANANTHA VIJAY VAMSI** section:
 - Student
 - Email: 19B91A04C2@srkrec.ac.in
 - Why can't I edit?
- Security info** section:
 - Keep your verification methods and security info up to date.
 - UPDATE INFO >
- Devices** section (partially visible)

