Name: Joshi Satya Vardan, Adina (Contractor)

Employee ID:2320097

Cohort Code: CSDAIA24AZ003

**Azure Data Factory Capstone - Covid Use Case**

**Introduction:**

Azure Data Factory, Synapse and Azure Data Lake Storage are some of the tools used to complete the project " Azure Data Factory Capstone - Covid Use Case " in accordance with business requirements. These Azure components help to extract data from its source (Data Lake), convert it in accordance with business needs, and load it into a 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.

* 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

**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:**

Consumes

Produces

Represents a data item stored in

Runs On

**Data Set**

**Activity**

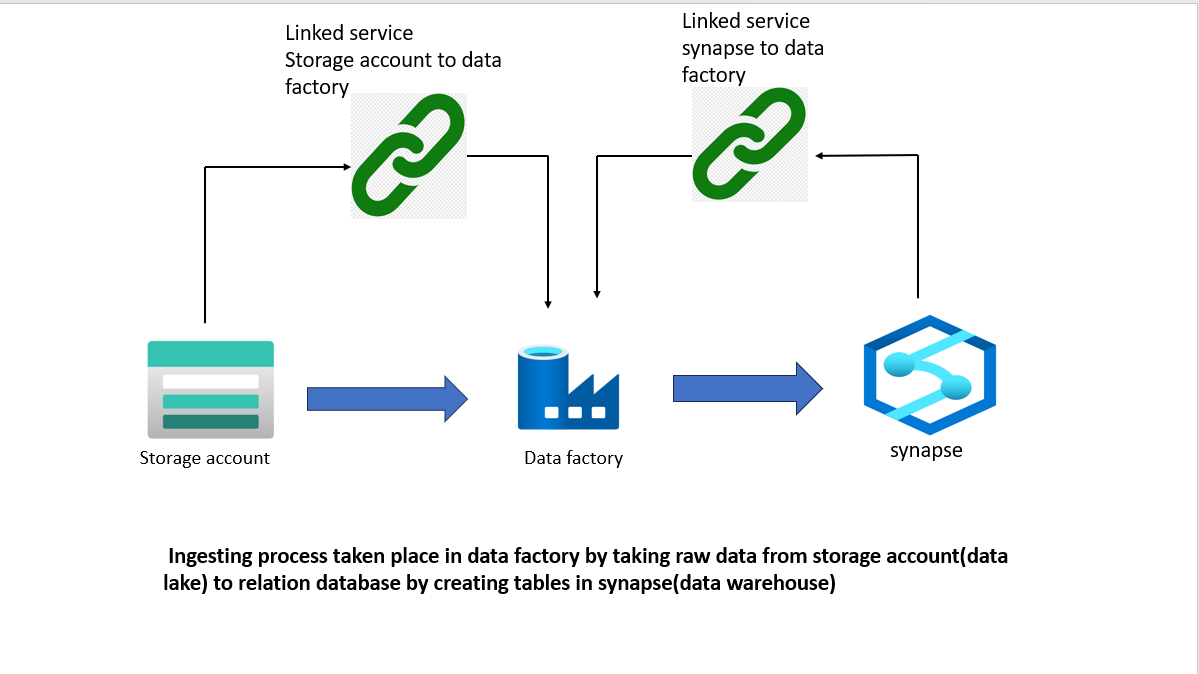
**Linked Service**

Is a logical Grouping Of

**Pipeline**

**Project Requirements:**

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



A table with black text

Description automatically generated

**Azure Synapse**

**ge**

**Azure Data Lake Storage**

**ge**

**SQL database**

Table 1

Table 2

Table 3

Table 4

Table 5

A screenshot of a computer

Description automatically generated

**Azure Data Factory**

**Lookup**

**Forech**

**Copy**

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

A diagram of data storage

Description automatically generated

**Azure Data Lake Storage**

A screenshot of a computer

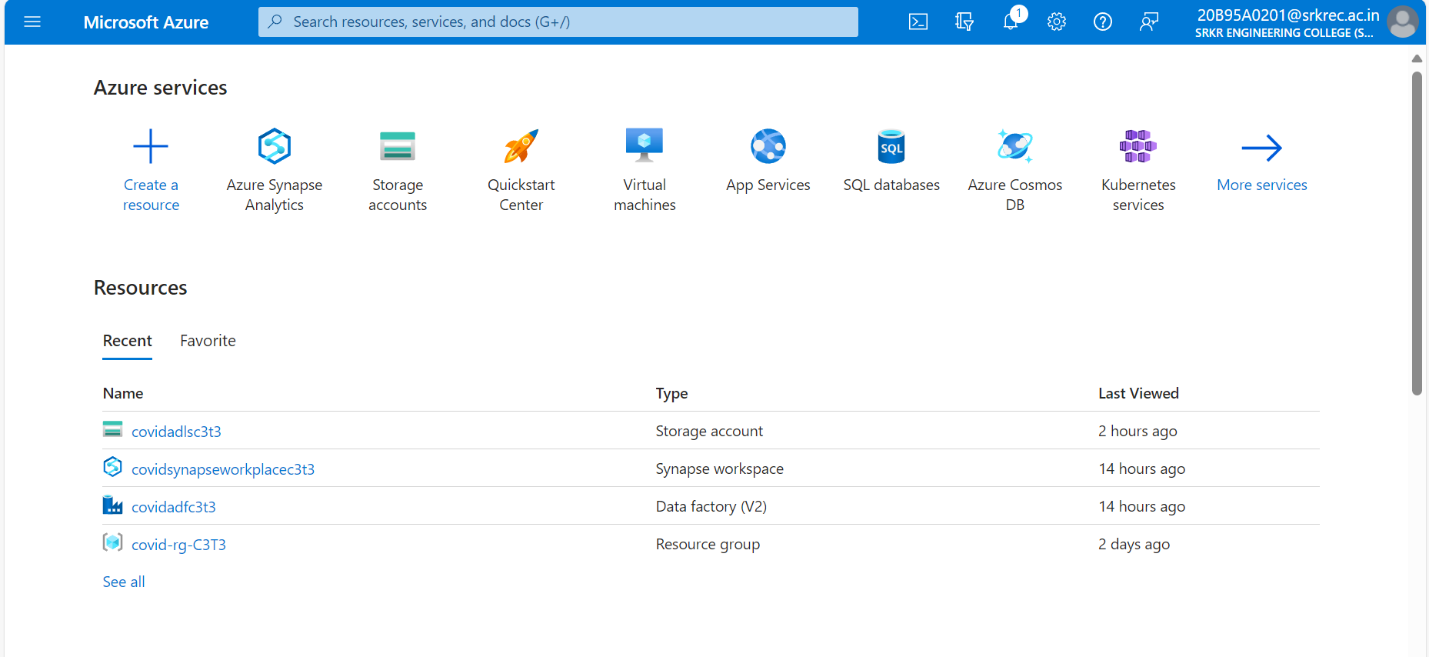
Description automatically generated

**transformPath**

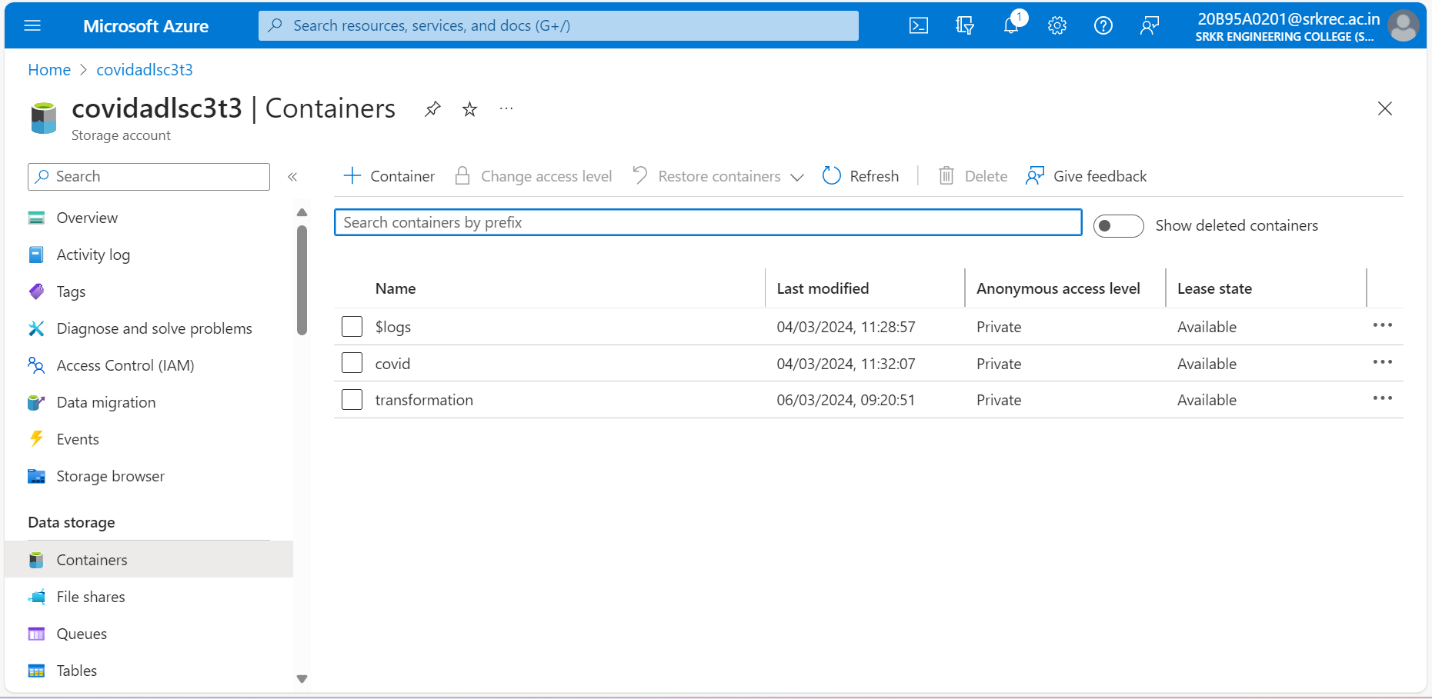
**Azure Data Flow**

Max\_daily\_deaths.csv

**Find The Continents Which Have Maximum Daily Death by using dataflow.**



**Step 1:** Created one Resource group(**covid-rg-C3T3**) and required resources for project like Storage account(**covidadlsc3t3**), Synapse workspace (data warehouse) (**covidsynapseworkplacec3t3**), Azure Data factory(**covidadfc3t3**).



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

A screenshot of a computer

Description automatically generated

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

. A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

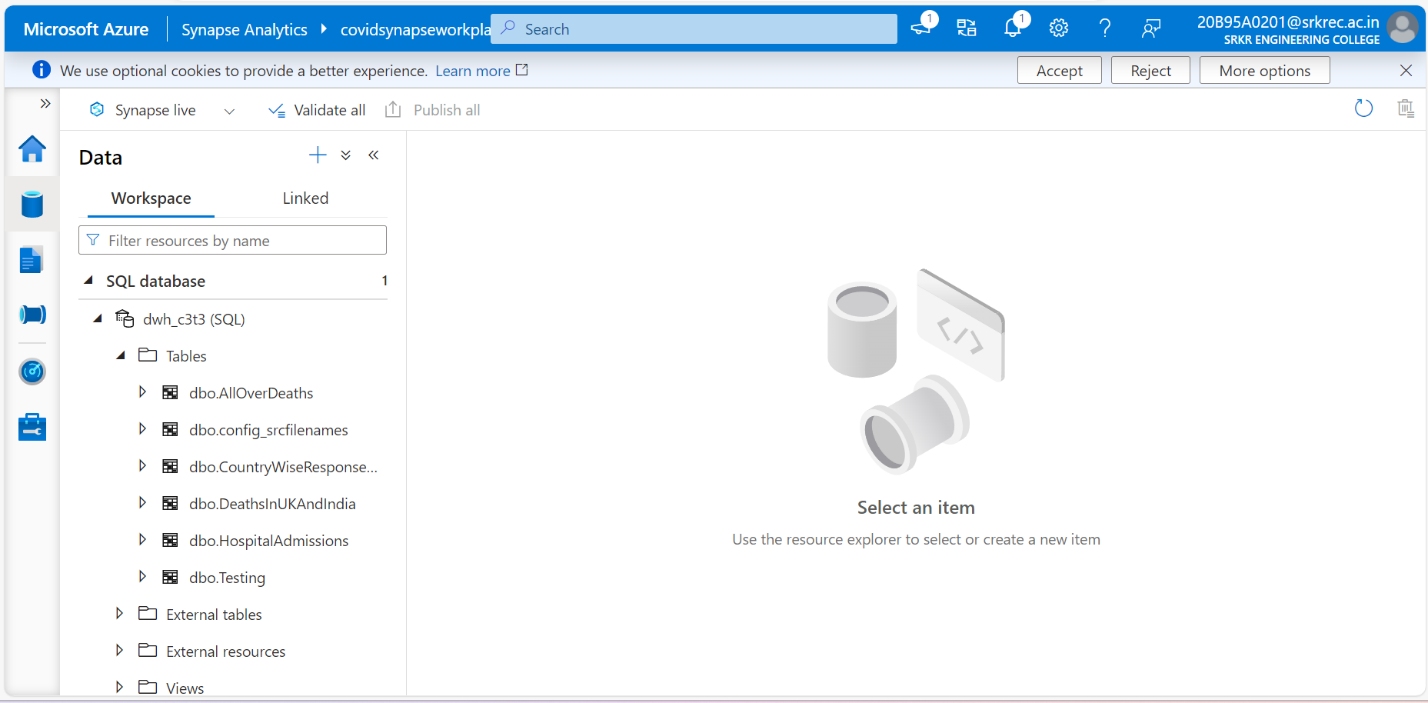
**Step 5:** Creating **two linked services** as per the project requirement in Azure data factory.

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

A screenshot of a computer

Description automatically generated

**Step 6:** Created **dataset (DS\_adlsV2Dataset)** for fetching flat files from storage account (data lake) which are present in ingest folder inside covid container.



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

**A computer screen shot of a computer

Description automatically generatedStep 8:** Created **dataset (DS\_dwdataset)** for inserting into SQL Tables created in synapse (data warehouse).

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

**Step 10:** Created **dataset(DS\_configDataset)** for fetching parameters table form synapse(data warehouse).

A screenshot of a computer

Description automatically generated

**Step 11:** Creating required datasets in Azure data factory.

* To fetch covid flat files from Storage account (data lake) **🡪(DS\_adlsV2Dataset).**
* To fetch parameters table from Synapse (data warehouse) **🡪(DS\_configDataset).**
* To insert data into SQL tables in Synapse (data warehouse) **🡪(DS\_dw\_dataset).**
* To fetch cases\_deaths.csv file from storage account (data lake) **🡪(DS\_casesanddeaths\_target).**
* To insert data transformation done file by using dataflow into storage account (data lake)**🡪(DS\_max\_daily\_deaths\_target)**

A screenshot of a computer

Description automatically generated

**Step 12:** Before creating pipeline in 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.

A screenshot of a computer

Description automatically generated

**Step 13:** Create a **pipeline** (**PL\_ ADFCovidUseCaseDataFactory**) and Drag and Drop the **Look up Activity** into pipeline workspace and set the **source dataset** **(DS\_configDataset)** for Lookup and choose the option **stored procedure** andgiven the stored procedure name created in synapse (data warehouse).

A computer screen shot of a computer

Description automatically generated

**Step 14:** 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).**

A screenshot of a computer

Description automatically generated

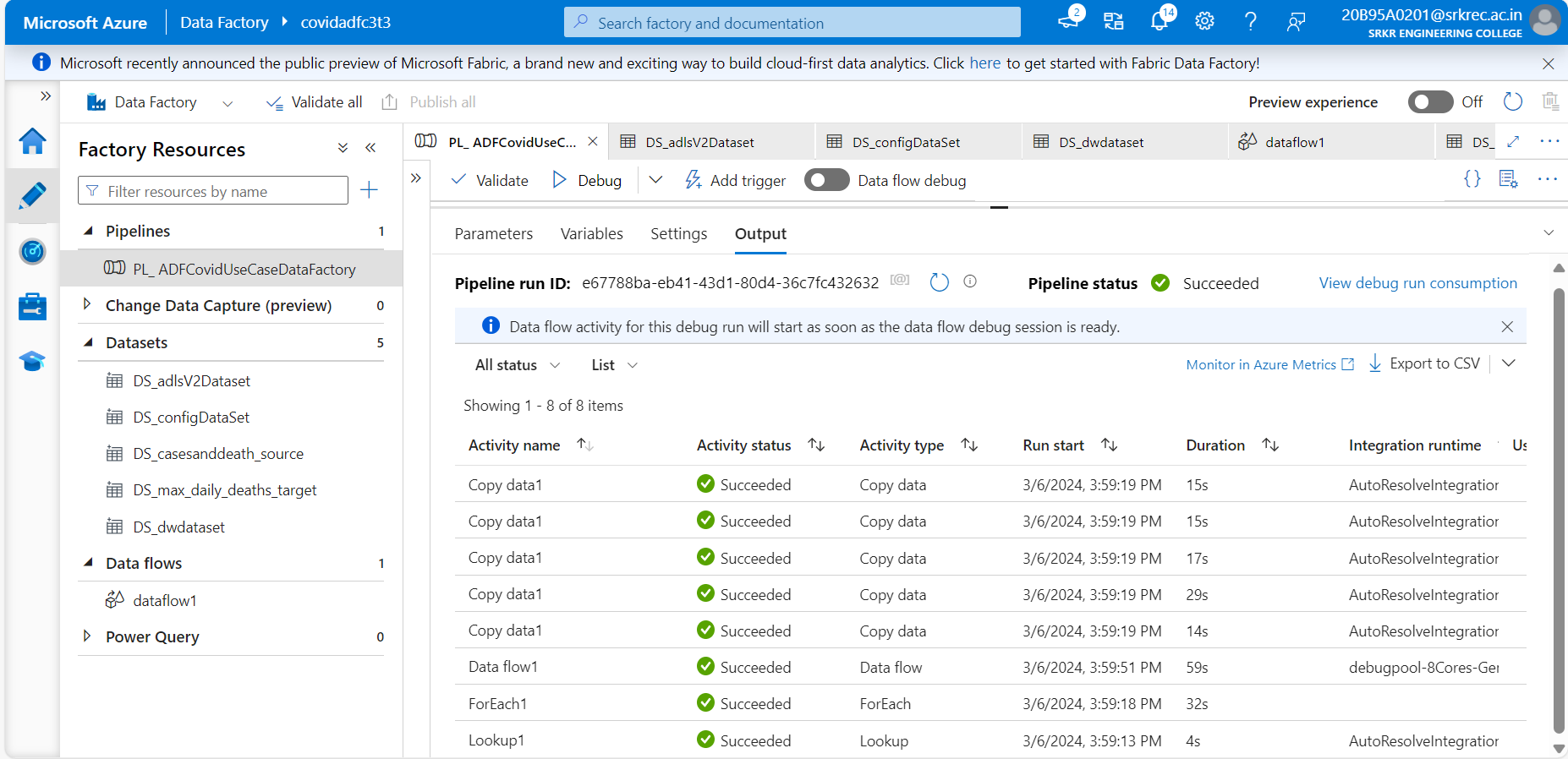
**Step 15:** Click on add activity symbol present on foreach activity and inside foreach activity add a **Copy activity** for copy data from CVS file into SQL table.

Configure settings at source side in copy activity by giving dataset **(DS\_adlsV2Dataset)** and giving folder name and file name dynamic by taking from foreach activity by **item.**

A screenshot of a computer

Description automatically generatedFolder Name **(@ {item (). FolderName}),** File Name **(@ {item (). FileName})**

**Step 16:** Configure setting in copy activity at sink side by giving dataset **(DS\_dwdataset)** and giving sqltableName dynamically by taking from foreach activity by item🡪sqltableName **(@ {item (). sqltableName})**



**Step 17:** 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.

A computer screen shot of a message

Description automatically generated

**Step 18:** After successfully run of pipeline now we need check the data inserted into tables in data warehouse by preforming two SQL queries operation given in project documentation.

A screenshot of a computer

Description automatically generated

**Step 19:** After running first SQL Query written in SQL script and it is successfully ruined and given output as per the query.

**A screenshot of a computer

Description automatically generatedStep 20:** compare the output appear for previous SQL query written in SQL script with original CSV file in excel sheet.

A computer screen shot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

**Step 22:** Successfully query is ruined and given output as per the query.

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

**Step 24:** Create another container with name **“transformation”** for second requirement given in project for storing transformed data file by using data flow.

**A screenshot of a computer

Description automatically generatedStep 25:** Create source dataset (**DS\_casesanddeath\_source**) for dataflow by giving a file specific filename on which data transformation need to be taken place as per project requirement.

A screenshot of a computer

Description automatically generated

**Step 26:** Create target dataset (**DS\_max\_daily\_deaths\_target**) for dataflow to keep that data transformed file in specific place for further use.

A screenshot of a computer

Description automatically generated

**Step 27:** Develop dataflow by using some transformations like source, filter, pivot, aggregate, rank and sink as per the question given in project documentation.

A computer screen shot of a computer

Description automatically generated

**Step 28:** After successfully creating dataflow, I click on dataflow debug output appear as per the question.

A screenshot of a computer

Description automatically generated

**Step 29:** Drag and drop the data flow into the pipeline workspace and connect with before activity to run after successfully complication of before activity.

A screenshot of a computer

Description automatically generated

**Step 30:** Dataflow is ruined successfully after click on debug as how in above picture.

A screenshot of a computer

Description automatically generated

**Step31:** After that we need to check file is appear in the transformation container in storage account (data lake). I successfully got that file in my container as per the question given in the project documentation.

**CONCLUSION:**

**Requirement 1 Output:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Requirement 2 Output:**

A computer screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**MY AZURE SUBSCRIPTION PROFILE**