Project on Azure Data factory – COVID 19 Reporting/Analysis

Contents

[Project overview 3](#_Toc171598139)

[Solution architecture 3](#_Toc171598140)

[Data Ingestion from Azure Blob 4](#_Toc171598141)

[Requirement: 4](#_Toc171598142)

[Azure Blob storage: 4](#_Toc171598143)

[Azure data factory 5](#_Toc171598144)

[Control flow activities: 5](#_Toc171598145)

[Creating an event Trigger 6](#_Toc171598146)

[ADF Activities used: 6](#_Toc171598147)

[Data Ingestion from HTTP 8](#_Toc171598148)

[Requirement: 8](#_Toc171598149)

[Azure Data Factory: 9](#_Toc171598150)

[Control Flow Activities (Look Up & For Each) 9](#_Toc171598151)

[ADF activities used 11](#_Toc171598152)

[Data transformation using data flow- Cases & Deaths 13](#_Toc171598153)

[Requirement 13](#_Toc171598154)

[Data transformation using Azure Data Bricks 16](#_Toc171598155)

[Creating a cluster 16](#_Toc171598156)

[Mounting Data Lake storage 17](#_Toc171598157)

[Transformation 18](#_Toc171598158)

[Copy data to Azure SQL 19](#_Toc171598159)

[Requirement 19](#_Toc171598160)

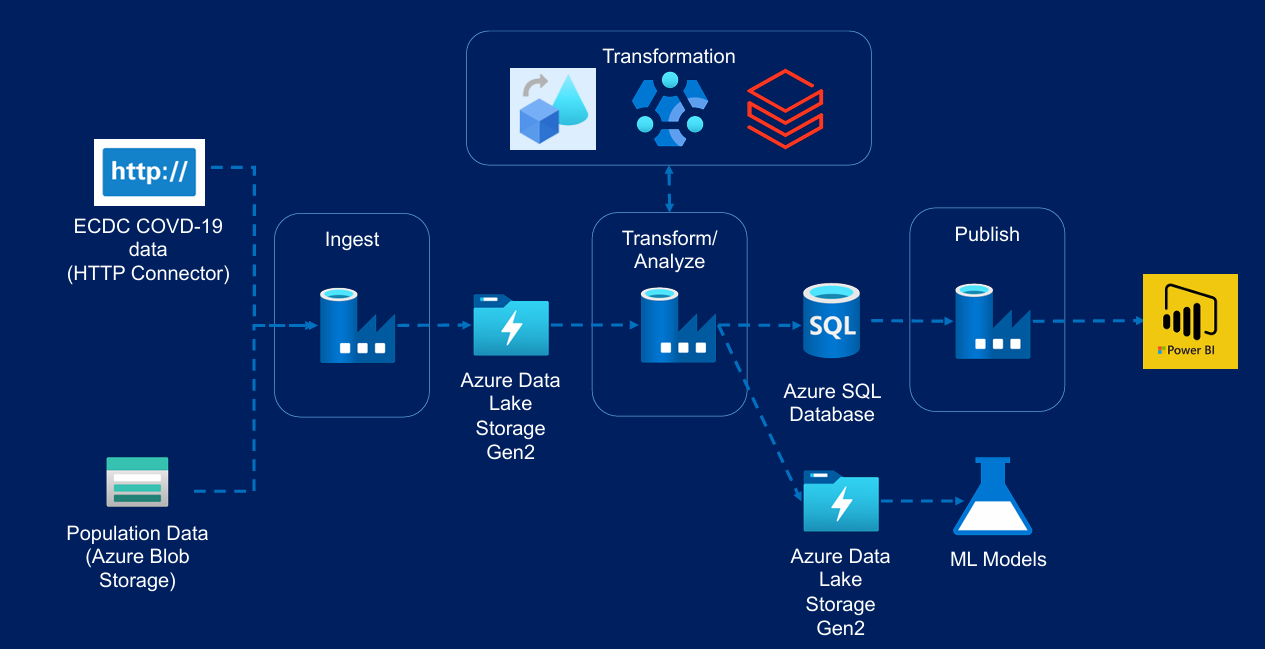
[Steps 19](#_Toc171598161)

[Building Power BI from SQL database 20](#_Toc171598162)

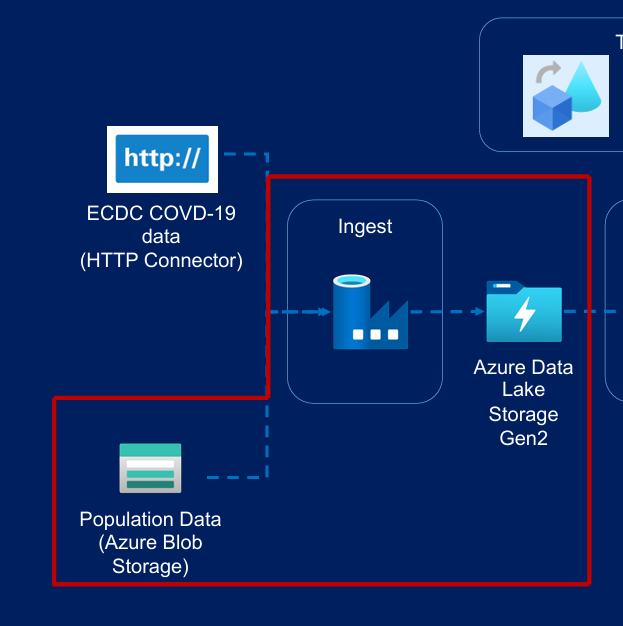
## Project overview

In this project, I used ECDC data to track COVID 19 metrics in Europe and further analyze it, using Azure Data Engineering service.

## Solution architecture

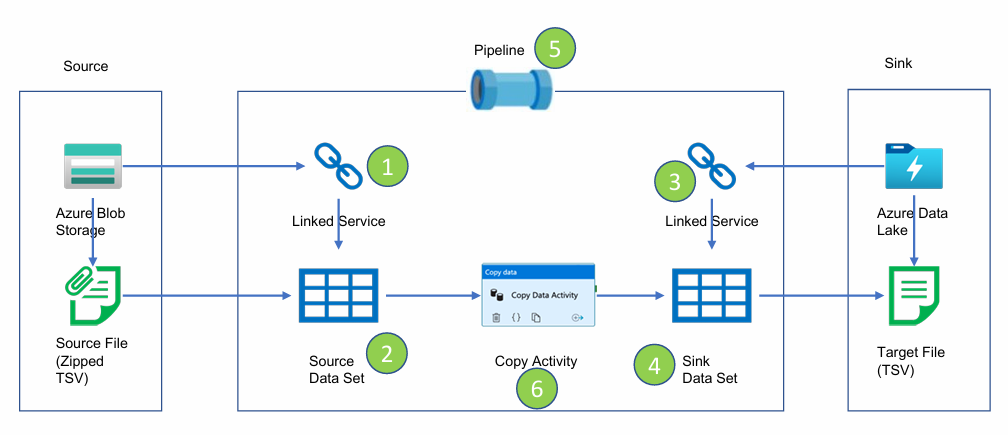


## Data Ingestion from Azure Blob



### Requirement:

Using the “Copy activity” copy the data which is present in the CSV file from Azure Blob storageàAzure Data Lake Storage.



### Azure Blob storage:

1. Create a blob container and upload the compressed csv file over here

### Azure data factory

1. Open the ADF.

2. Create a **Linked Service**

-2 linked service are created (Blob storage and Data Lake)

3. Create a **Data set (**When creating this, it will ask for the Linked Service that we created, followed by the path of the file**)**

-2 dataset’s are created for Source and Sink

4. Creating ADF pipeline

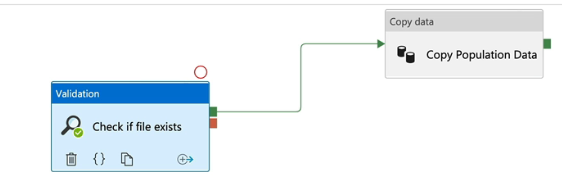
- Choose the “**Copy data**” activity

- Source as Blob and Sink as Data Lake

### Control flow activities:

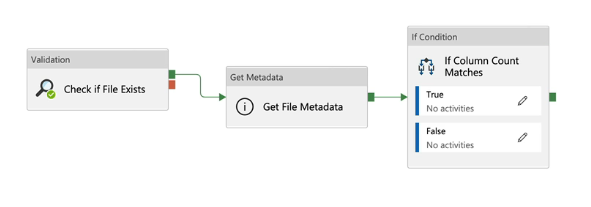
#### Execute the copy activity only when the file is available

1. Use **“Validation”** activity

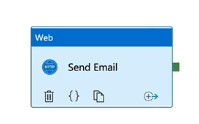


#### Execute Copy Activity only if file contents are as expected

1. Using the “**Get File Metadata**” and “**If Condition**” activity, in this project we are checking if the file contains 13 columns and then copying it.

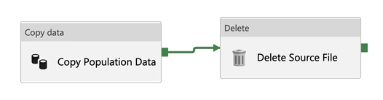


2. In the “**If Condition**” activity if the output if “**True**”, then the copy will be performed, if “False” using the “**Web**” activity , it will display that the condition is failed



#### Delete activity (Delete the source file on successful copy)

1. Once the copy activity is completed with the help of “**Delete**” activity, I will be deleting the file from the Source.



### Creating an event Trigger

A storage event rigger has been created, whenever a file is placed in the Azure blob (Source) this event will be triggered and the pipeline will be executed.

### ADF Activities used:

1. Copy

2. Validation

3. Get Metadata

4. If-Condition

5. Web

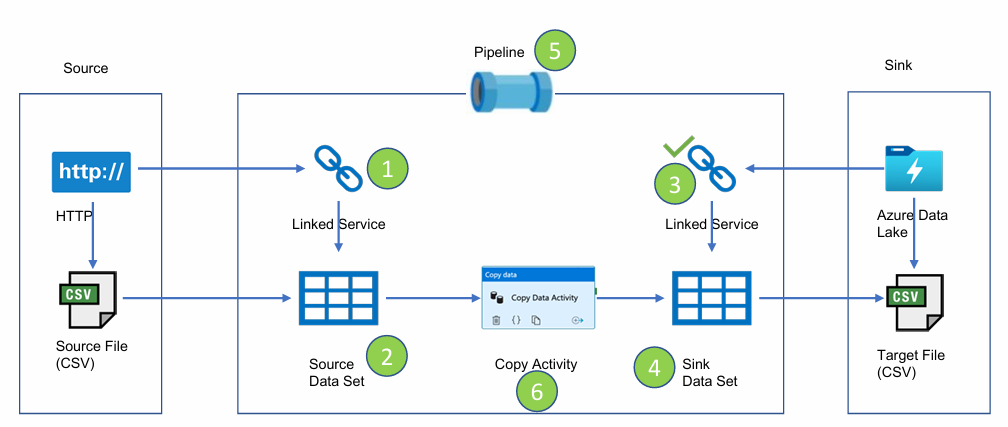
6. Delete

## Data Ingestion from HTTP

### Requirement:

Using the “Copy Activity” copy the data from https:// (ECDC data)[It contains 4 files] to Azure Data Lake.

Link: https://opendata.ecdc.europa.eu/covid19/nationalcasedeath/csv/data.csv

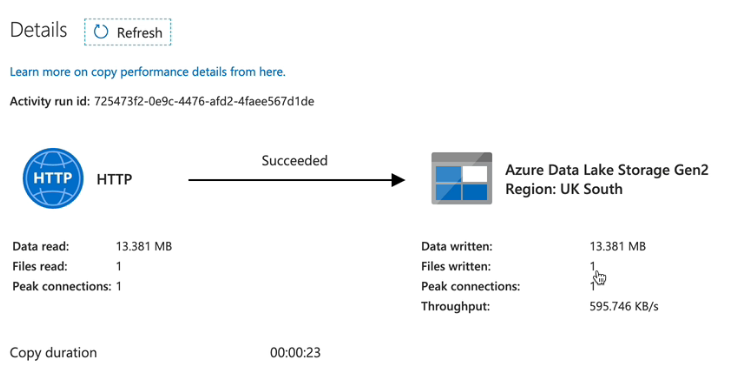


### Azure Data Factory:

1. Create a linked service for “http”

2. Create a dataset for “http” and to store the data create a dataset for “Data Lake”

3. Create a pipeline, use “Copy activity”

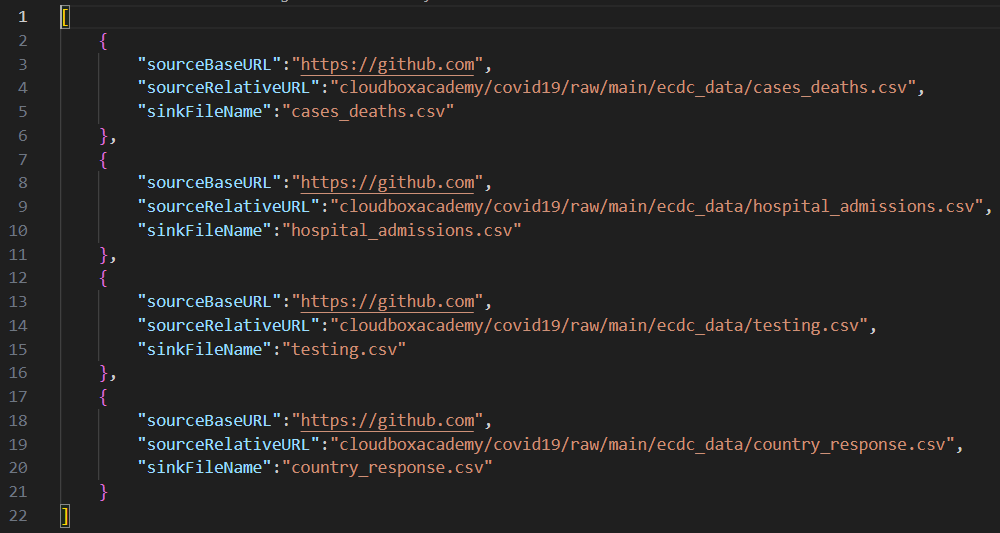


Note: Above method is used only for copying 1 file to destination, but we have still 3 files to be moved. If we do the same process it will not be recommended. So this can be solved using “**Control Flow Activities (Look Up & For Each)**”.

### 

### Control Flow Activities (Look Up & For Each)

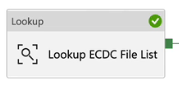
1. Under ADF create a new dataset for JSON file, and the file looks like below.



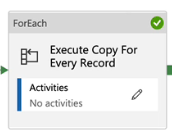
From here we will extract the data to destination.

#### Meta data driven pipeline

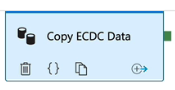
#### 1. Create a “Lookup” activity (Which is used to get you through the JSON file).



2. Create a “ForEach” activity (To get the items present inside the JSON file).

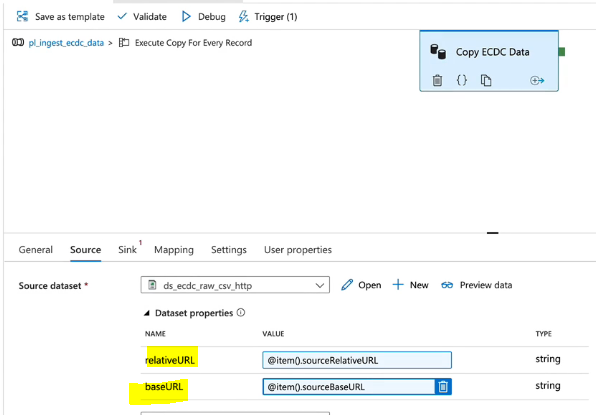


3. Inside ForEach activity just copy the “Copy” activity inside it(So that for 4 files also will be copying one by one) ,

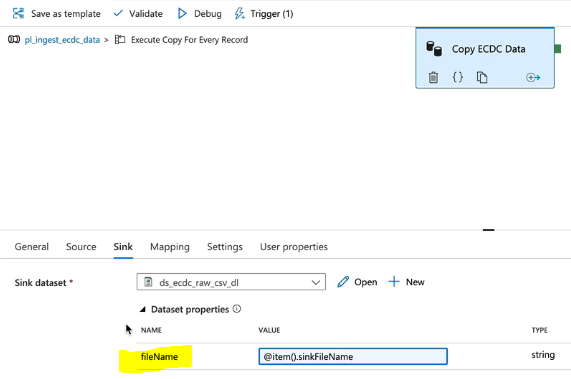


4. Then assign the parameters for Copy activity

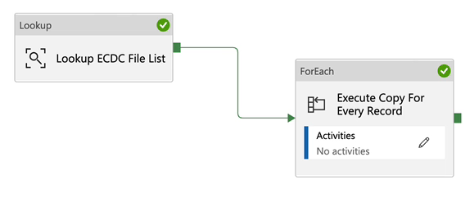
-In Source give (relativeURL,baseURL)



-In Sink give (fileName)



5. Once it is done, now from the json file it will be copying the data to destination with the help of Meta data



### ADF activities used

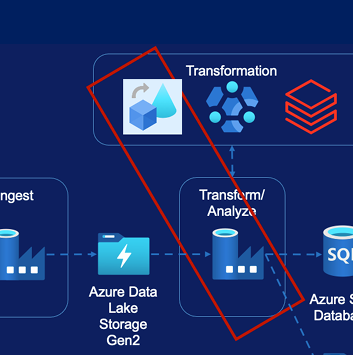
1. Web

2. Lookup

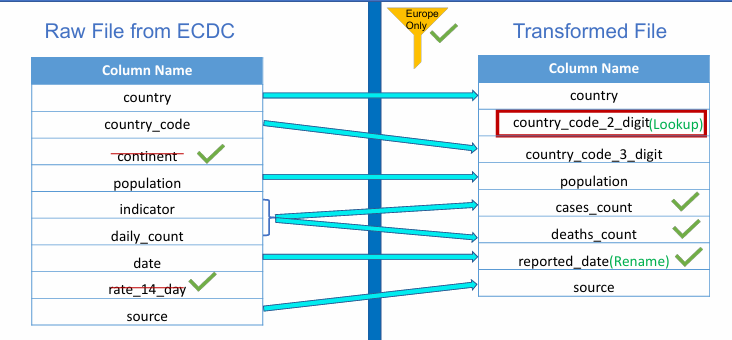
3. ForEach

4. Copy

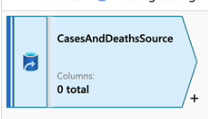
## Data transformation using data flow- Cases & Deaths



### Requirement



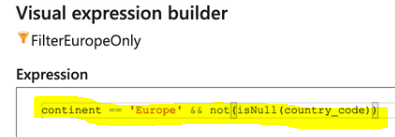
#### Source Transformation



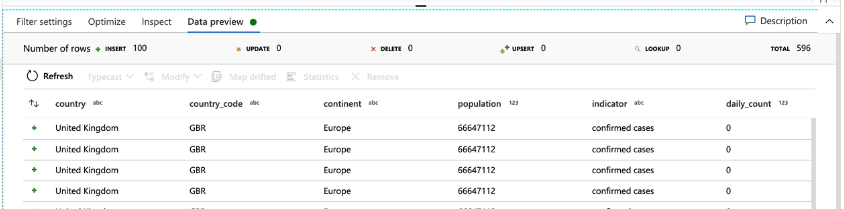
From here itself we will be able to change the datatype, rename it

#### Filter Transformation

With the help of this, we will be able to filter only the Europe data.

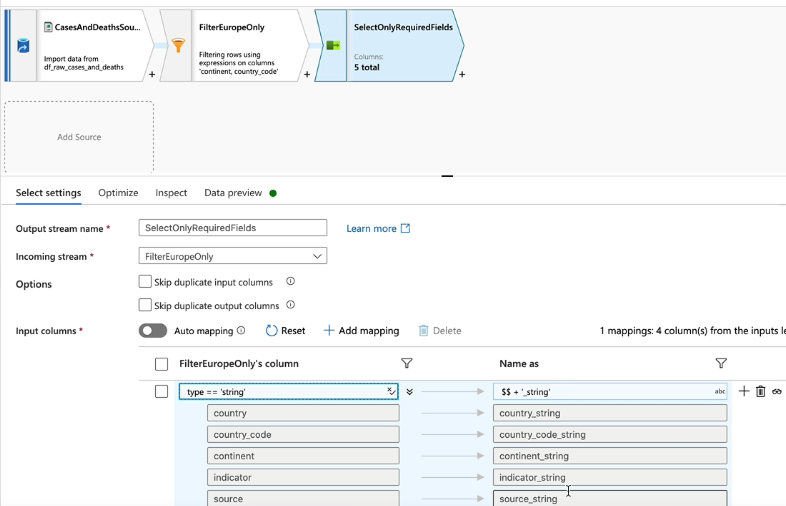
 

Data will be looking like below

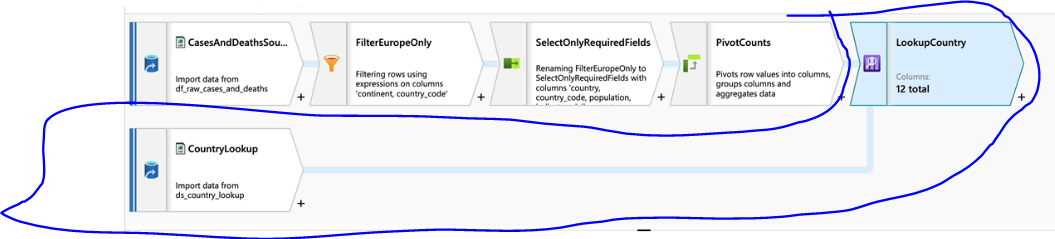


#### Select transformation

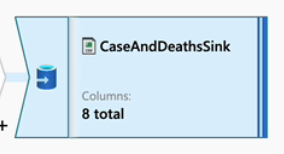
In here we can rename, delete the columns



#### Lookup transformation

 Sink transformation

Used to store the transformed data to the destination dataset

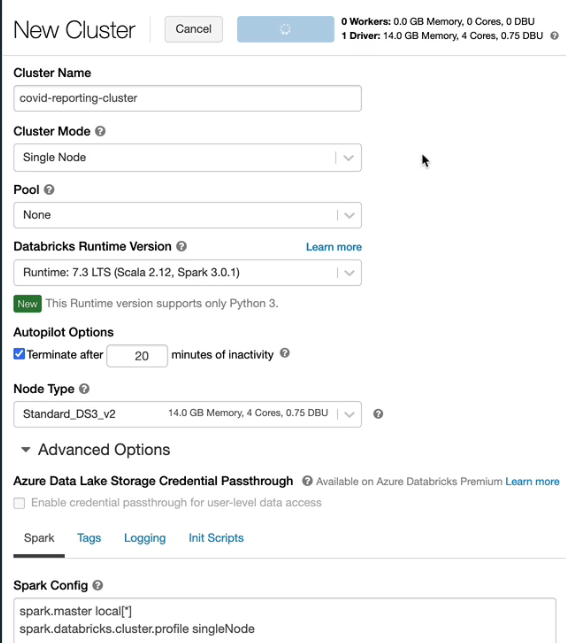


Once all the data flow is created, create a pipeline

## Data transformation using Azure Data Bricks

### Creating a cluster

In this project I have created an interactive cluster



### Mounting Data Lake storage

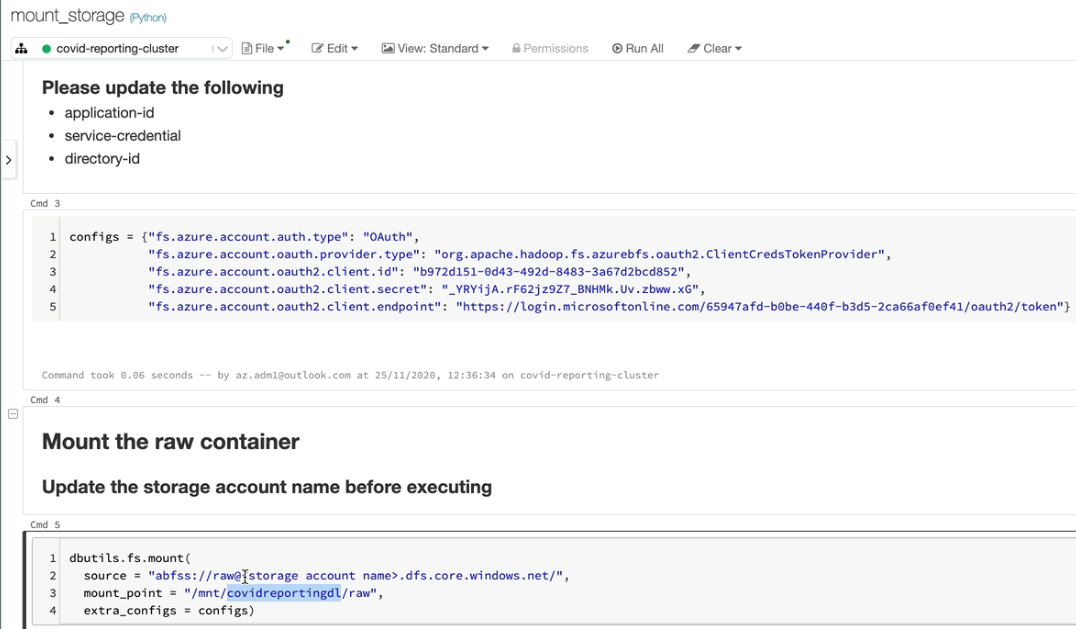
#### Create “Azure Service Principal”

Go to AAD->App registration

#### Grant access for “Data Lake” to “Azure Service Principal”

Go to Data Lake->Access Control (To grant access)

#### Create the mount in “data bricks” using Service Principal



### Transformation

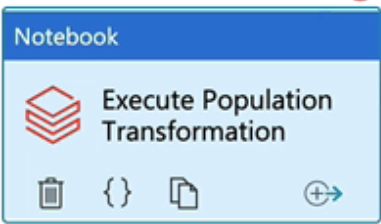
-Read the data and create a temp view

-Pivot the data by age group

-Using the join operation

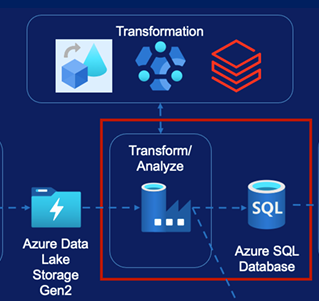
-Writing the output to the mount point

Once it’s done create a ADF to run the data bricks.



### 

## Copy data to Azure SQL



### Requirement

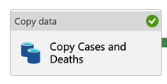
Copy the data from Data Lake to SQL

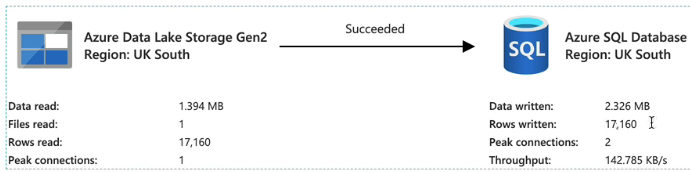
### Steps

1. Create a SQL database,

2. Create a pipeline, use “Copy Activity”

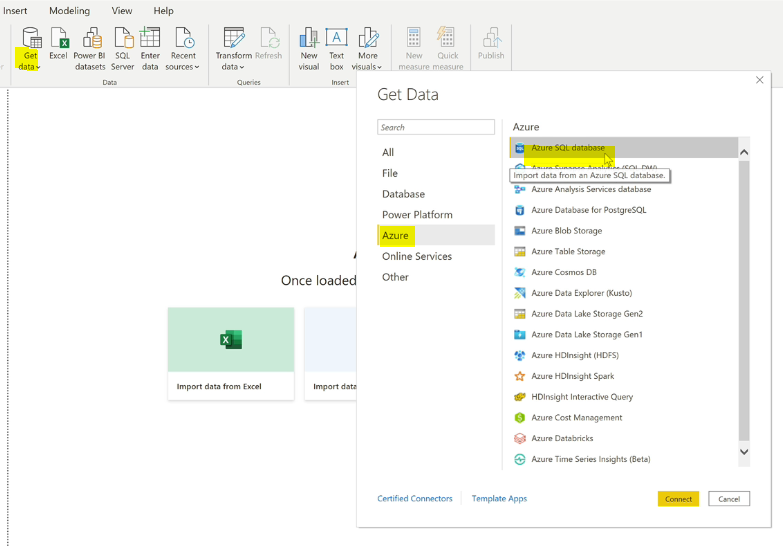
- Source will be a Dataset

 - Sink will be a SQL Database



## Building Power BI from SQL database

Download the Power BI in the desktop, once done go to Power BI and click on “Get Data” and select “Azure SQL” database.



Provide the server name and database name and finally provide the credentials and click connect once it is done we will get all our tables.

The final report will be looking like below:

