Azure Data Factory Project on Covid19

Azure Technology used in this project:

* Azure Storage Account.
* Azure Data Factory.
* Azure SQL Database.
* Azure Databricks
* PowerBI (For final Data visualisation)

Data Source:

ECDC Website.

Population Data gzip File.

**Activities:**

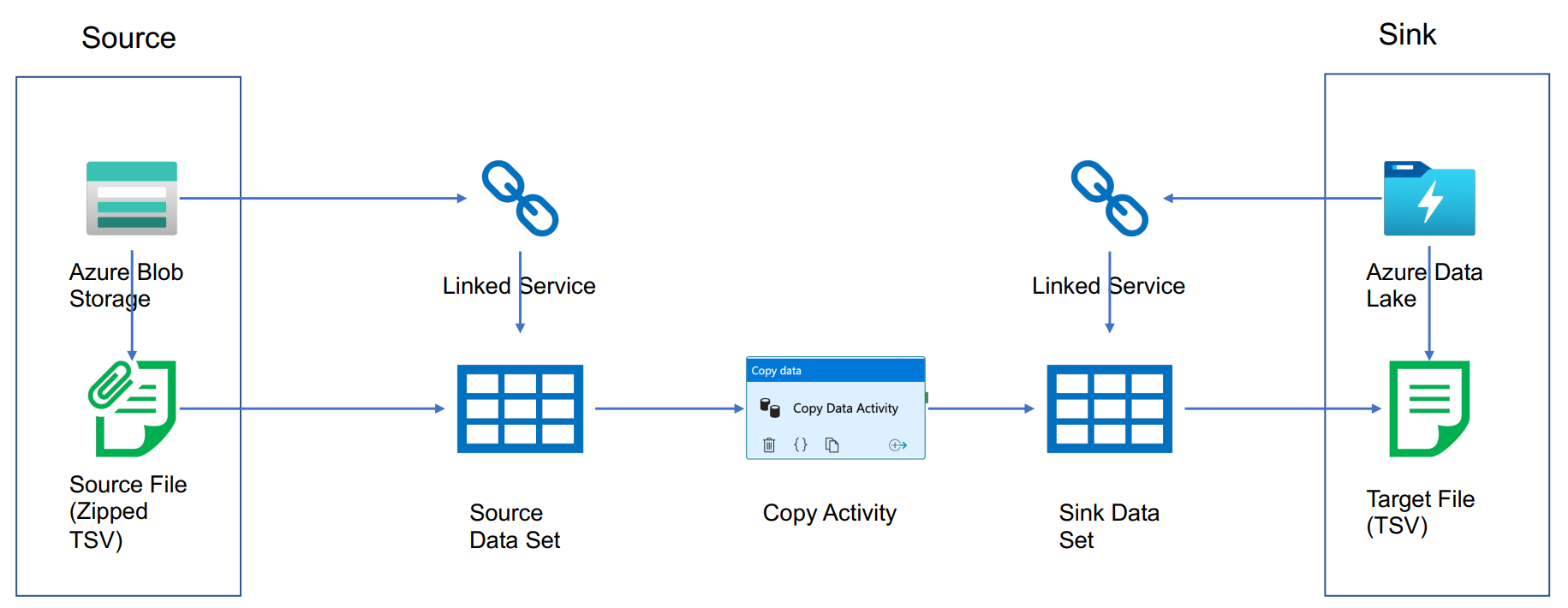
1.Copy Activity:

Azure Blob Storage Azure Data Lake

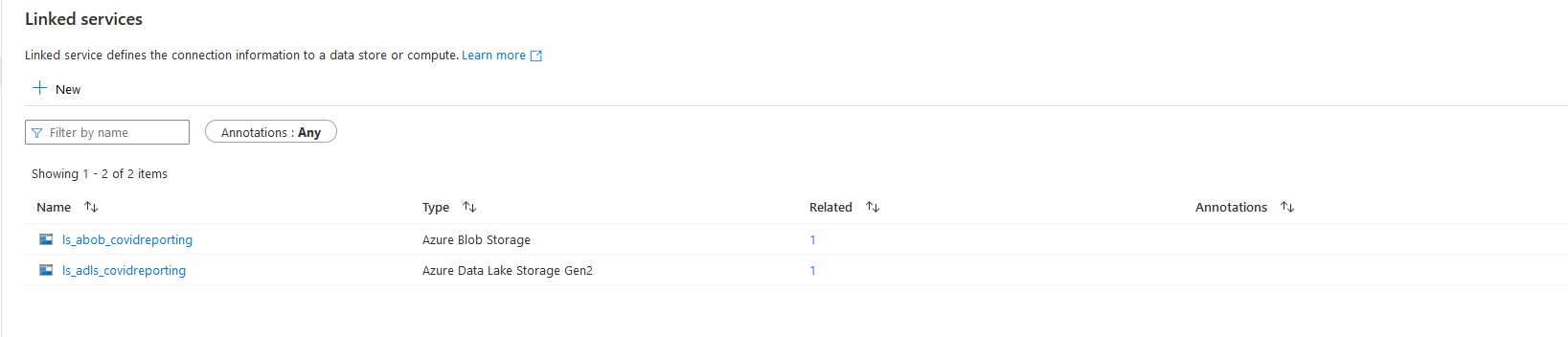
Source :Storage Account – covidreportingsa Container – population File - population\_by\_age.tsv.gz

Copy & Extract

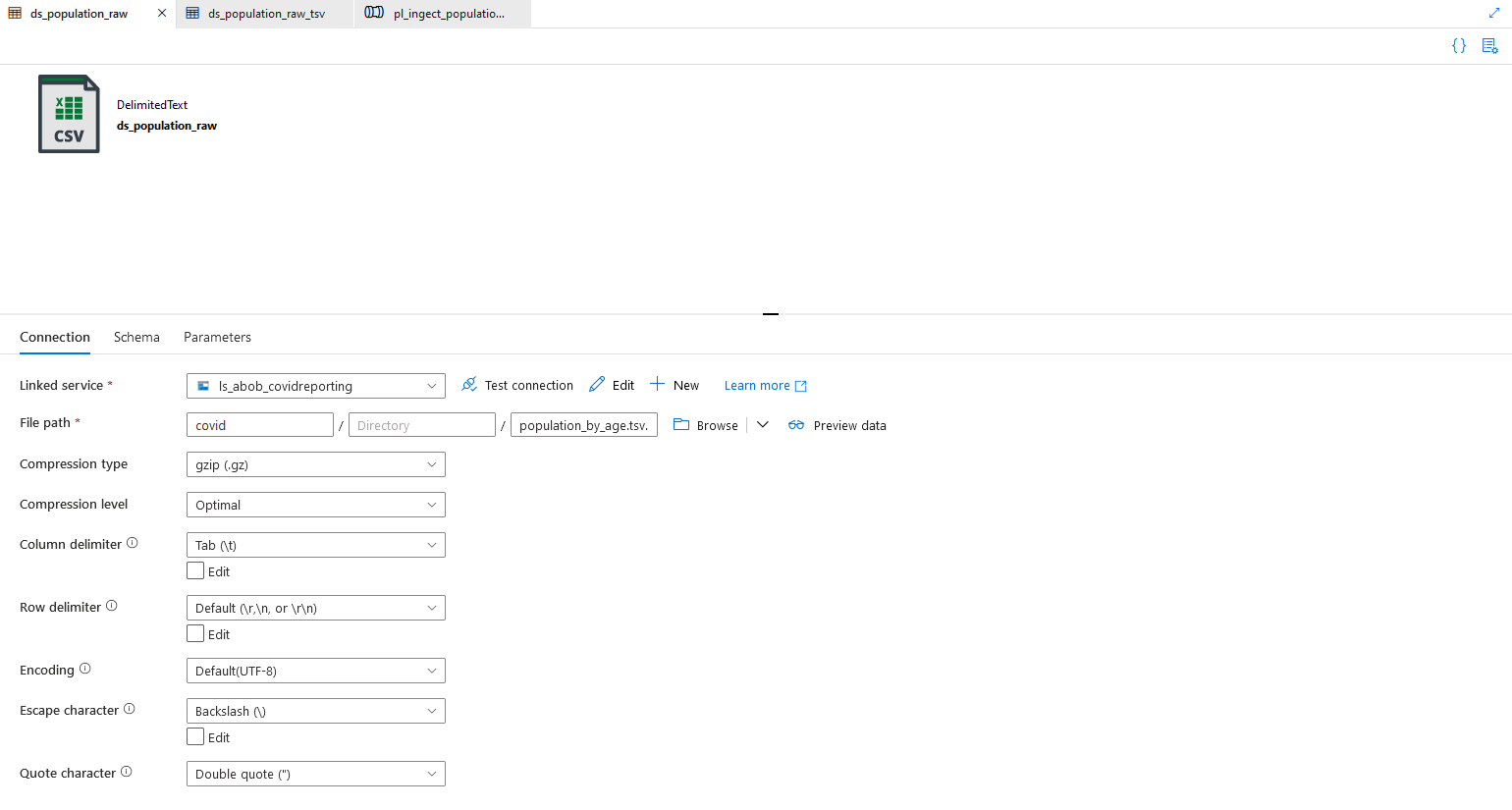
Target : Storage Account – covidreportingdl Container – raw File - population/population\_by\_age.tsv



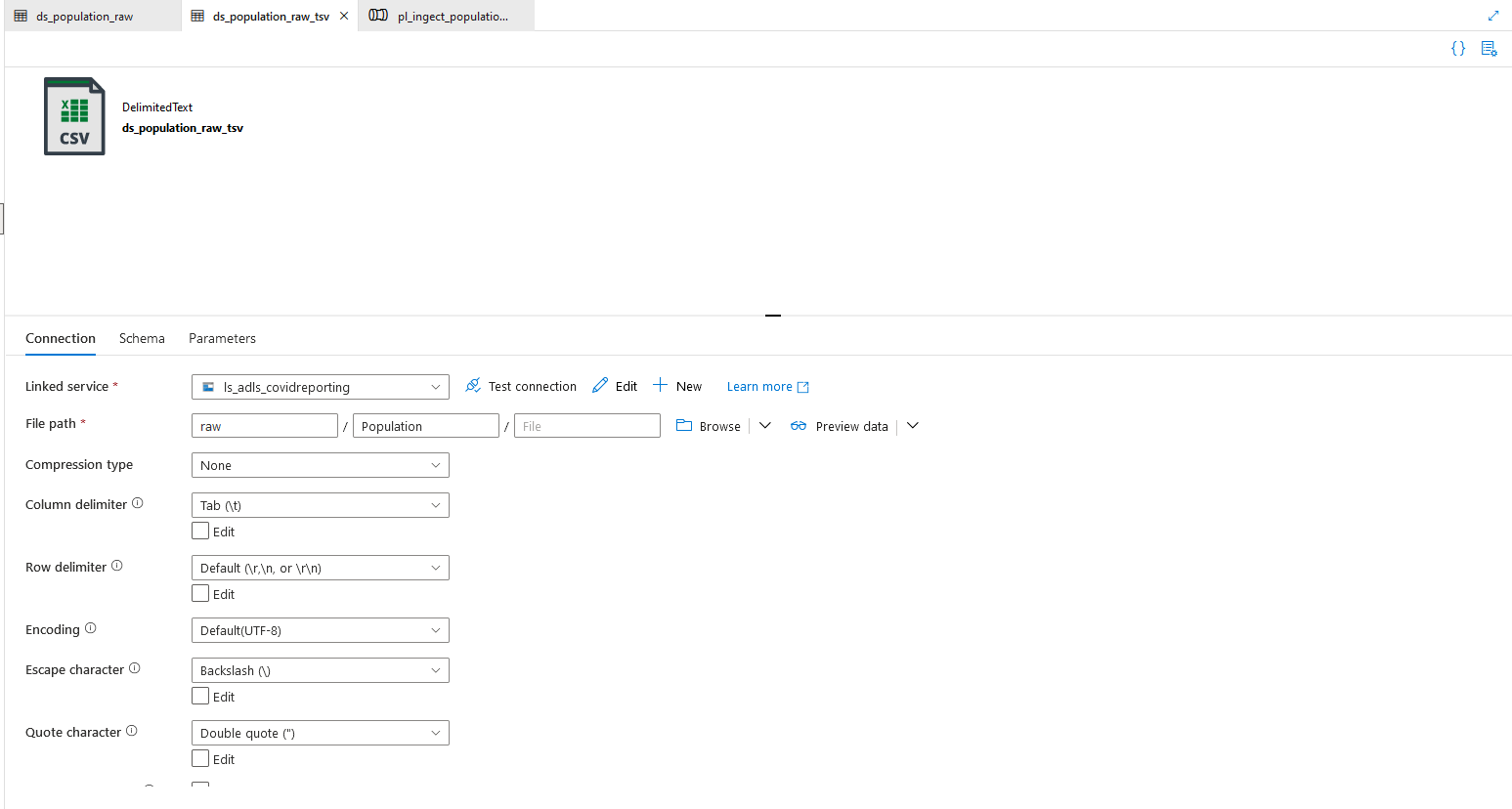
->Linked Services:



->Dataset:

Source Dataset: 

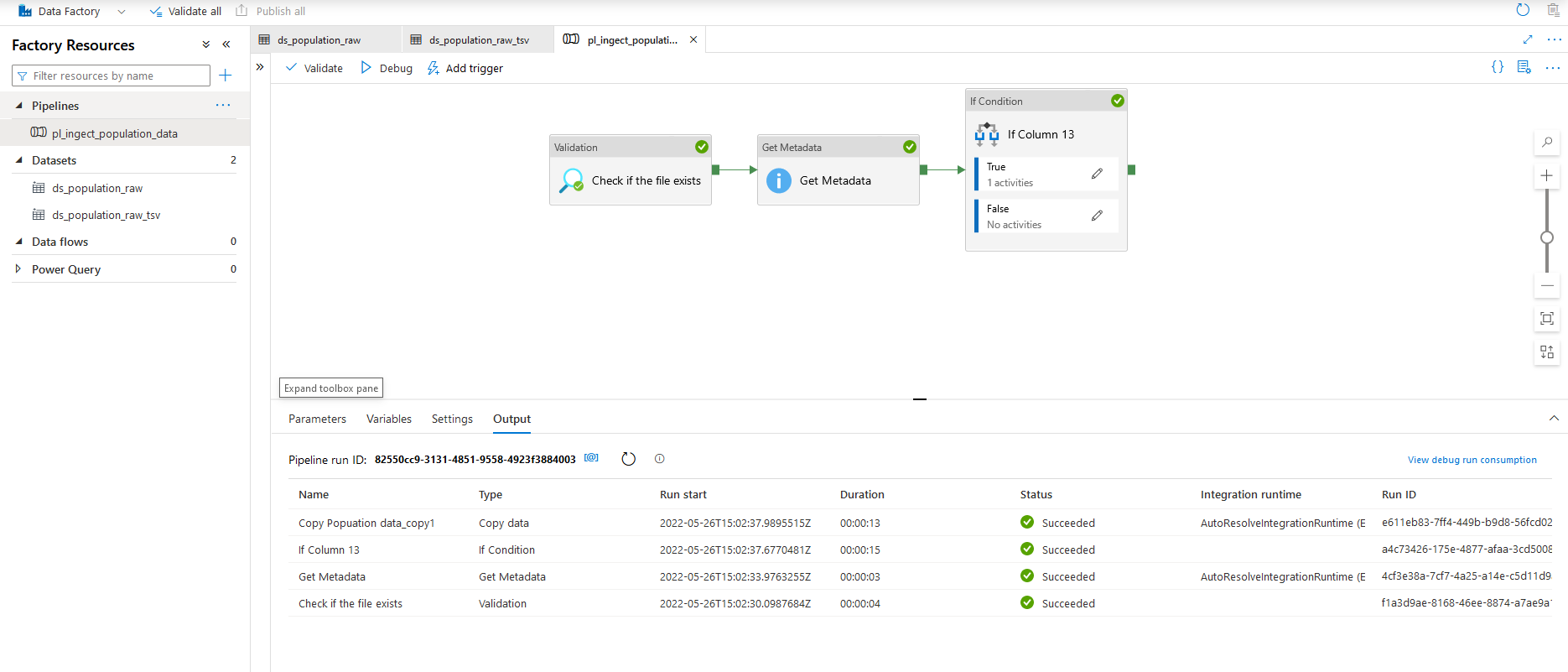
Destination Dataset:



->Pipeline:

In pipeline Check if file exists validation checks if the source has any file in it.

And then it sends this to Meta data where we use the output from the Get Metadata activity in conditional expressions to perform validation, or consume the metadata in if condition activity.



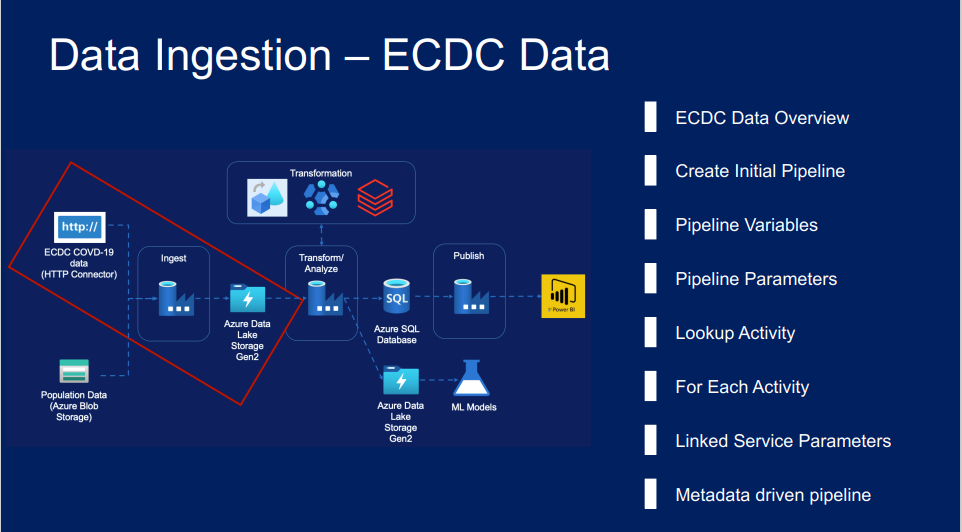
As the activity pipeline was successfully debug.The file was successfully copied from source (blob storage) to Destination(Data lake).

**2.Copying data from a HTTPS source:**

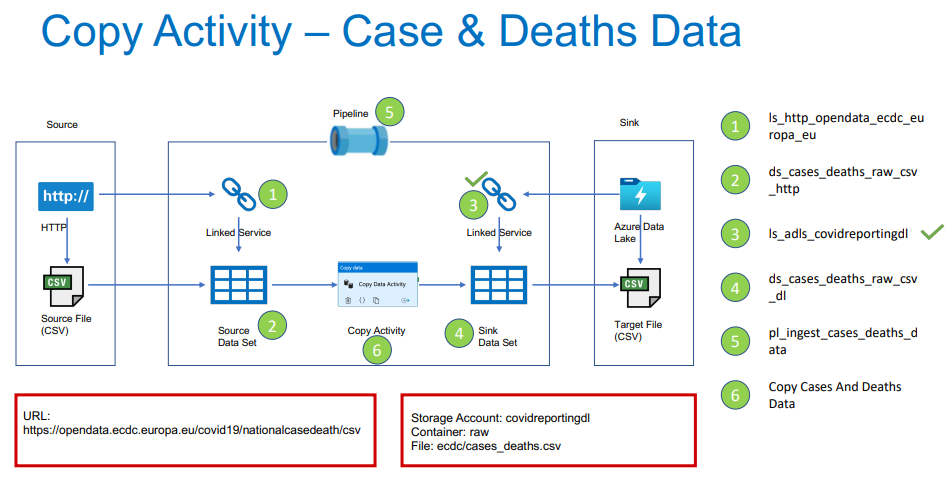
|  |  |
| --- | --- |
| sourceBaseURL | "<https://github.com>" |
| sourceRelativeURL | "cloudboxacademy/covid19/raw/main/ecdc\_data/cases\_deaths.csv" |
| sinkFileName | "cases\_deaths.csv" |
| 1 |  |
| sourceBaseURL | "<https://github.com>" |
| sourceRelativeURL | "cloudboxacademy/covid19/raw/main/ecdc\_data/hospital\_admissions.csv" |
| sinkFileName | "hospital\_admissions.csv" |
| 2 |  |
| sourceBaseURL | "<https://github.com>" |
| sourceRelativeURL | "cloudboxacademy/covid19/raw/main/ecdc\_data/testing.csv" |
| sinkFileName | "testing.csv" |
| 3 |  |
| sourceBaseURL | "<https://github.com>" |
| sourceRelativeURL | "cloudboxacademy/covid19/raw/main/ecdc\_data/country\_response.csv" |
| sinkFileName | "country\_response.csv" |

Above are the HTTPS links for the data from github.

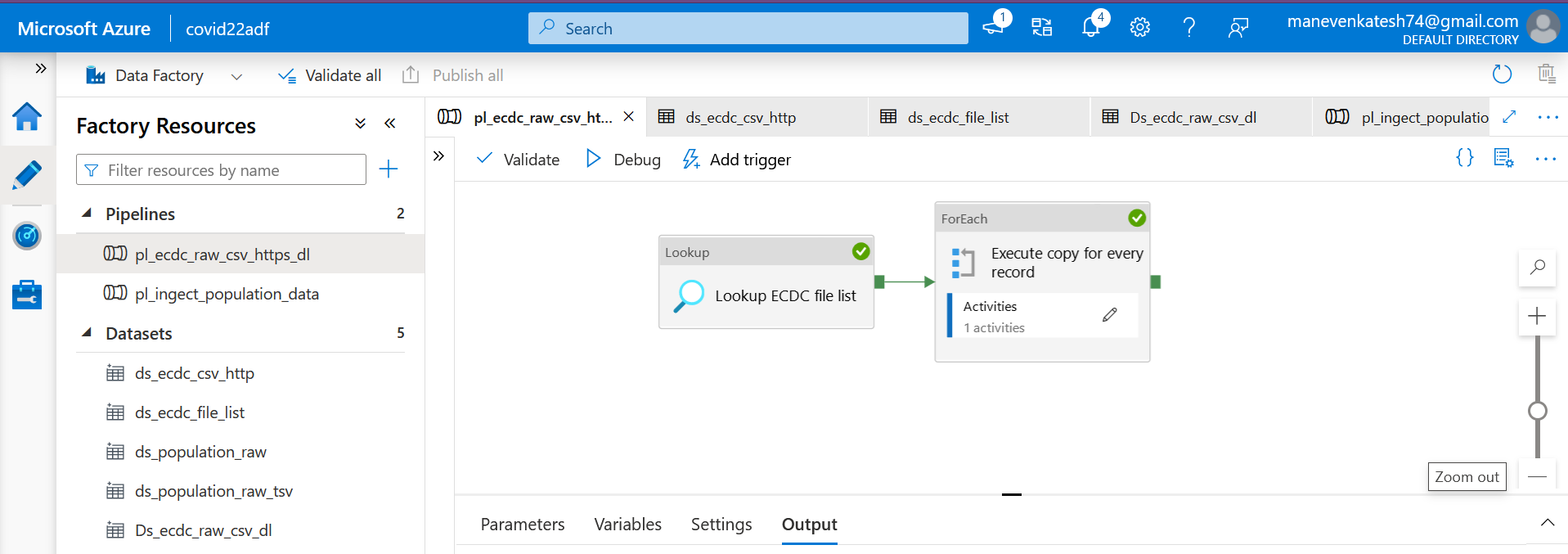
Now to get the data from https we have to follow below flow chart:



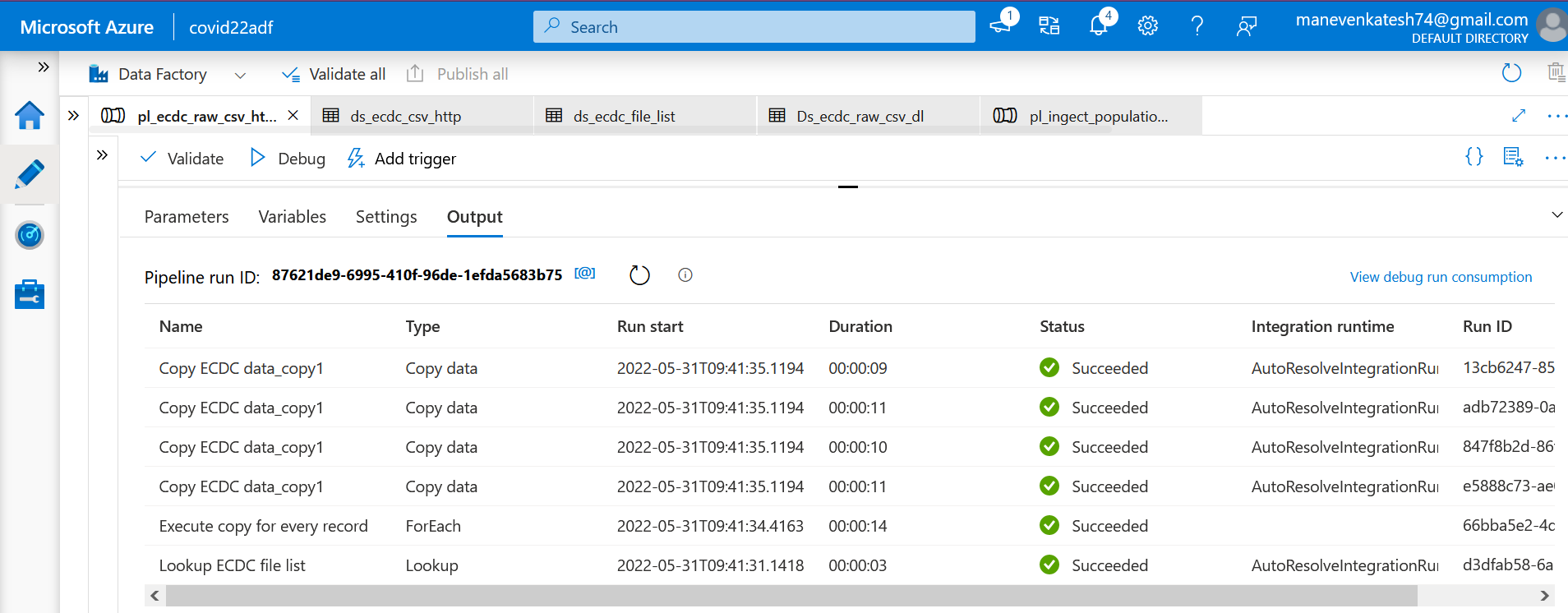
**HTTP 🡪 Azure Data Lake**



**We have to follow the same flow chart for case&deaths data, hospital admissions,testing, country response.**

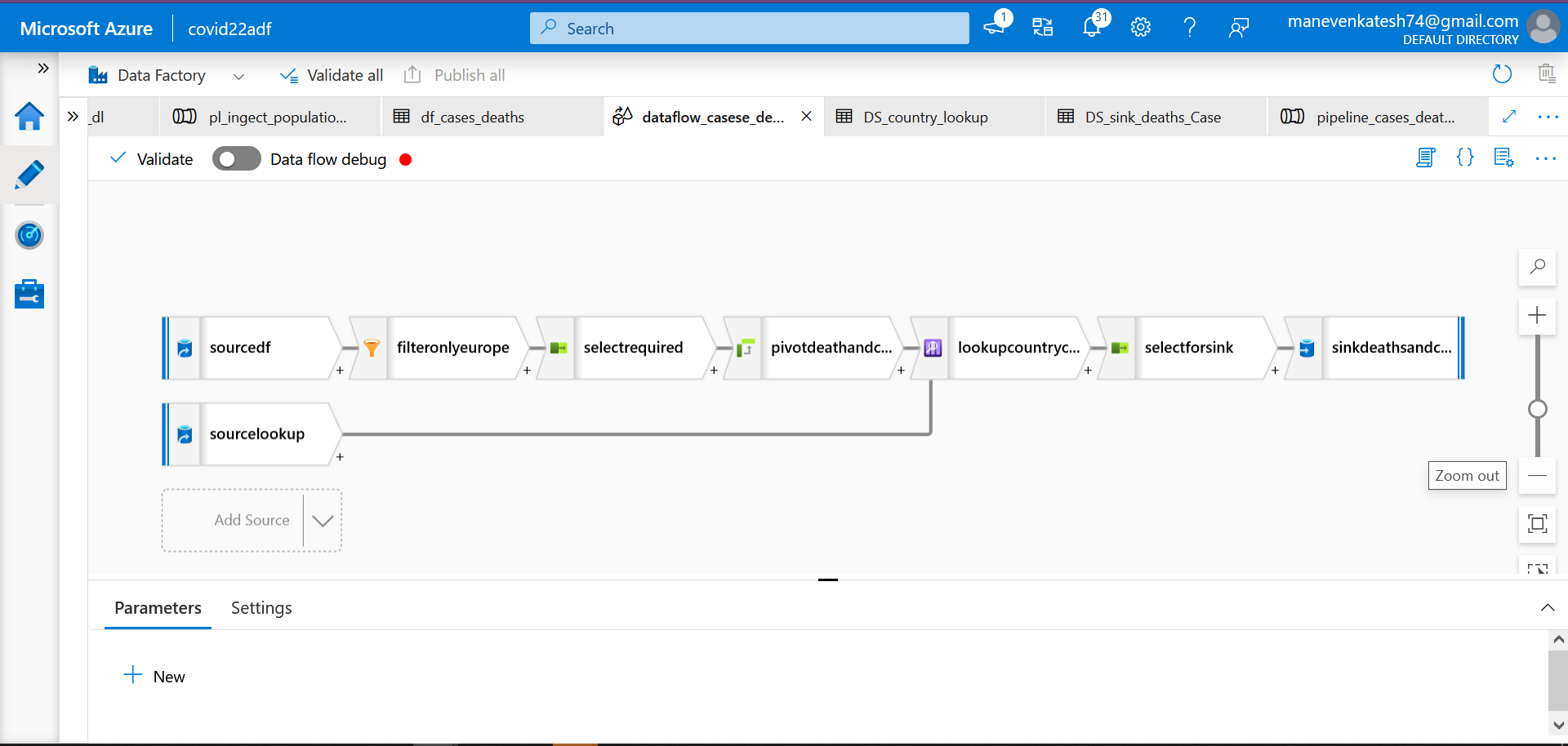


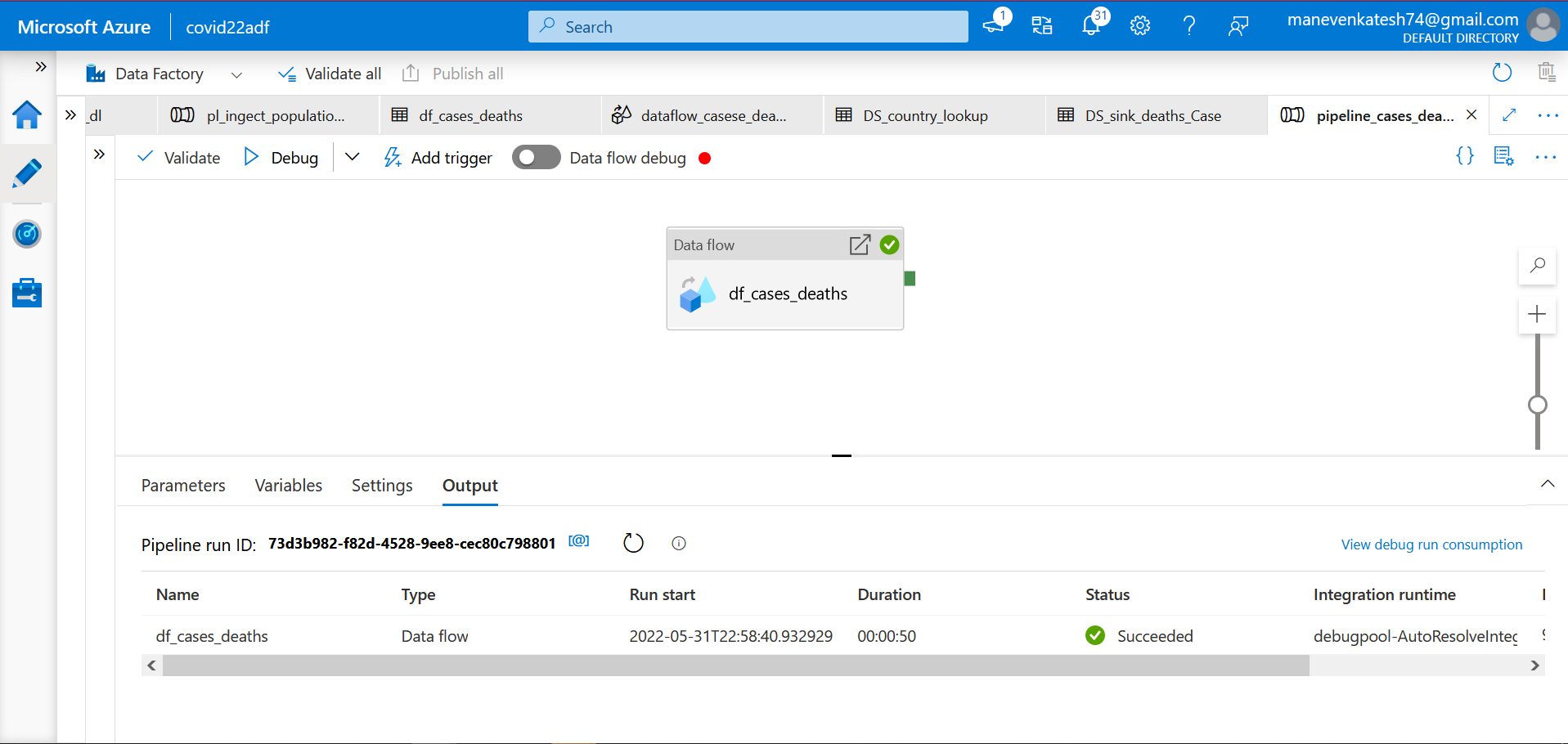
Debug:

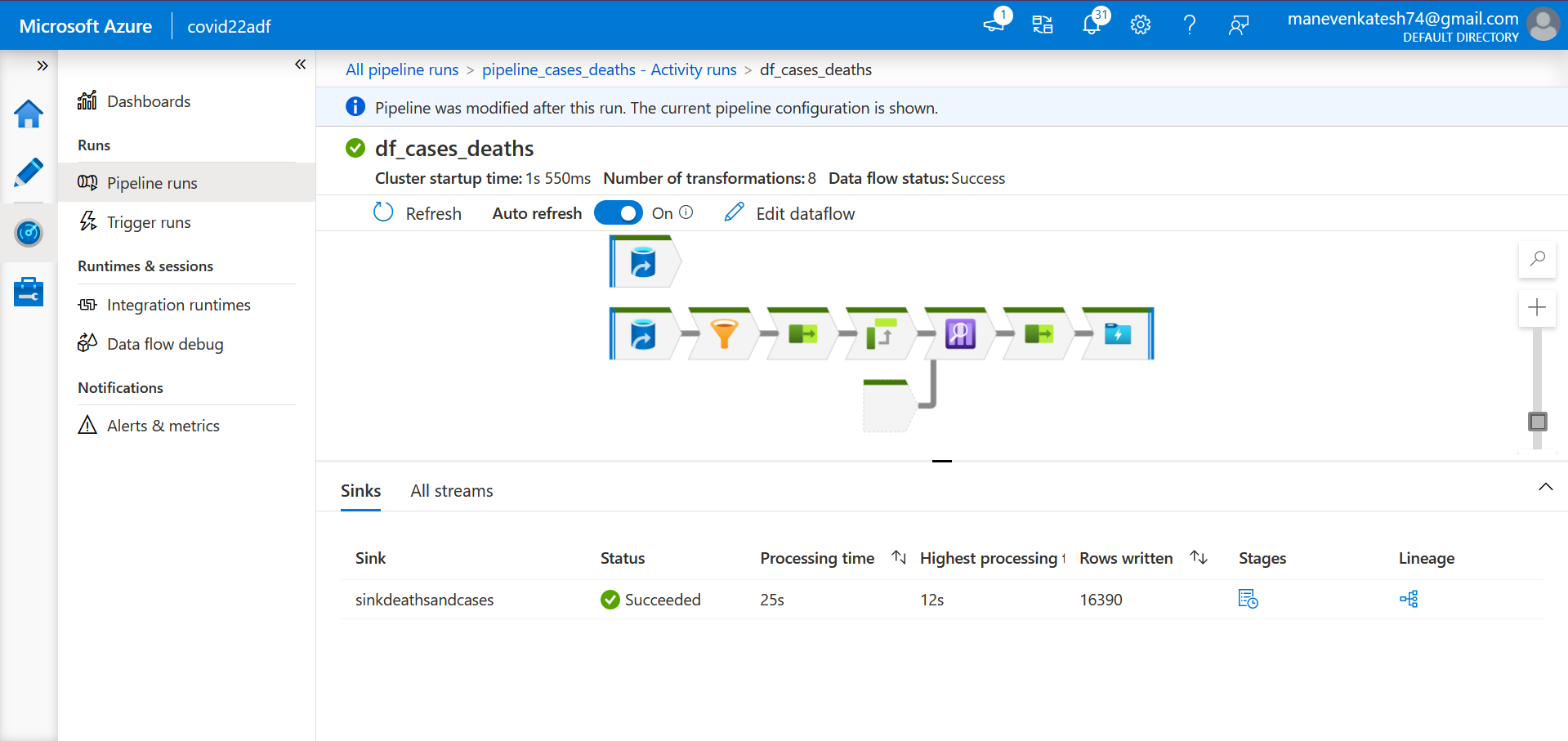


After storing the required data in the Azure Data lake .Next ,we will use the stored data for data transformation using data flow and data wrangling.

**3.Data flow :Data transformation :**







Using Data Flow :

We have transformed the cases and deaths information CSV file.

We used filter to just get the Europe Continent data.

Select statement to just view : country ,country\_code,Population,Indicator,Daily count ,Source ,Reported date.

Used Schema modifier Pivot for getting the confirmed cases and deaths through daily count and indicator column.

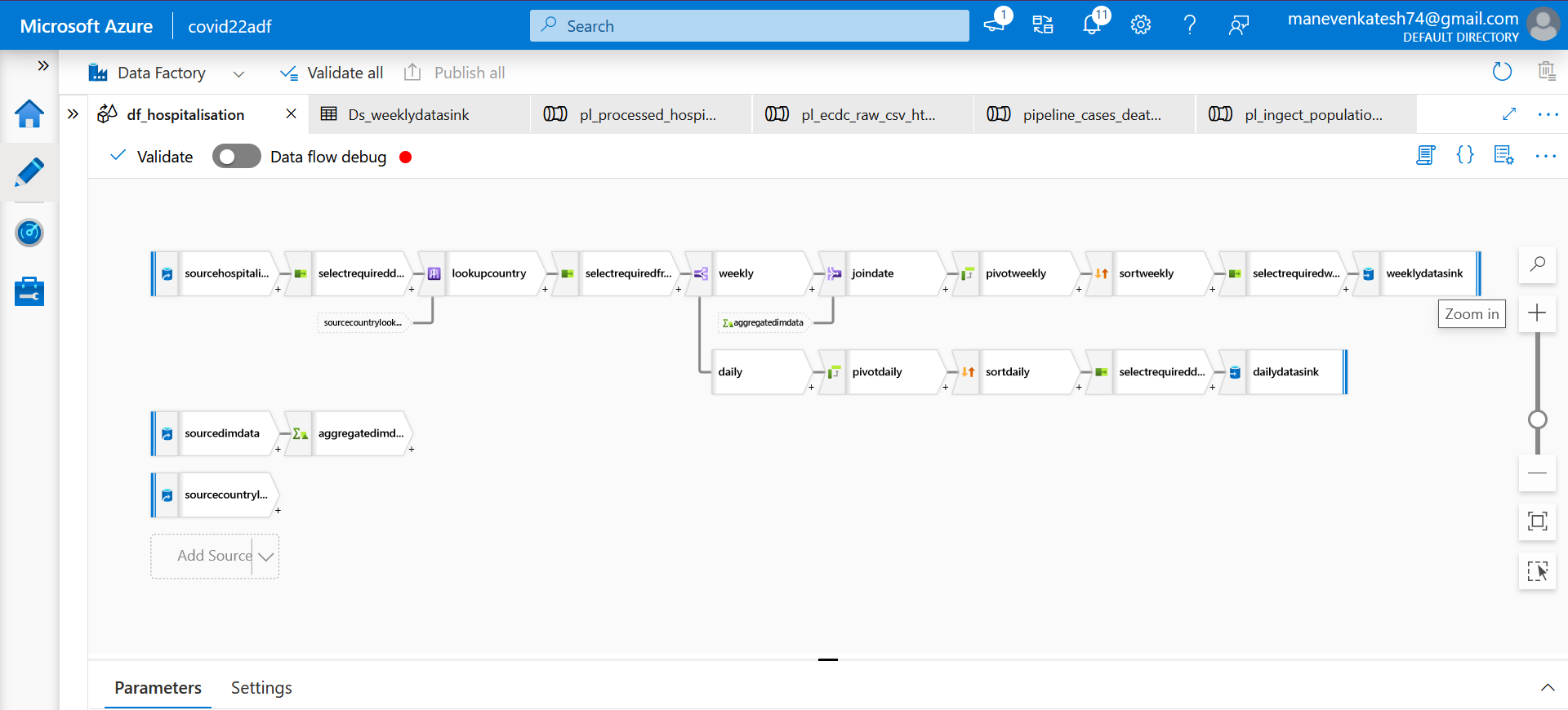
LookUp for getting Country code 2 and 3 Digit.And sink .

After validating creatd a pipeline DF activity and ran the debug to copy the data got to Data lake.

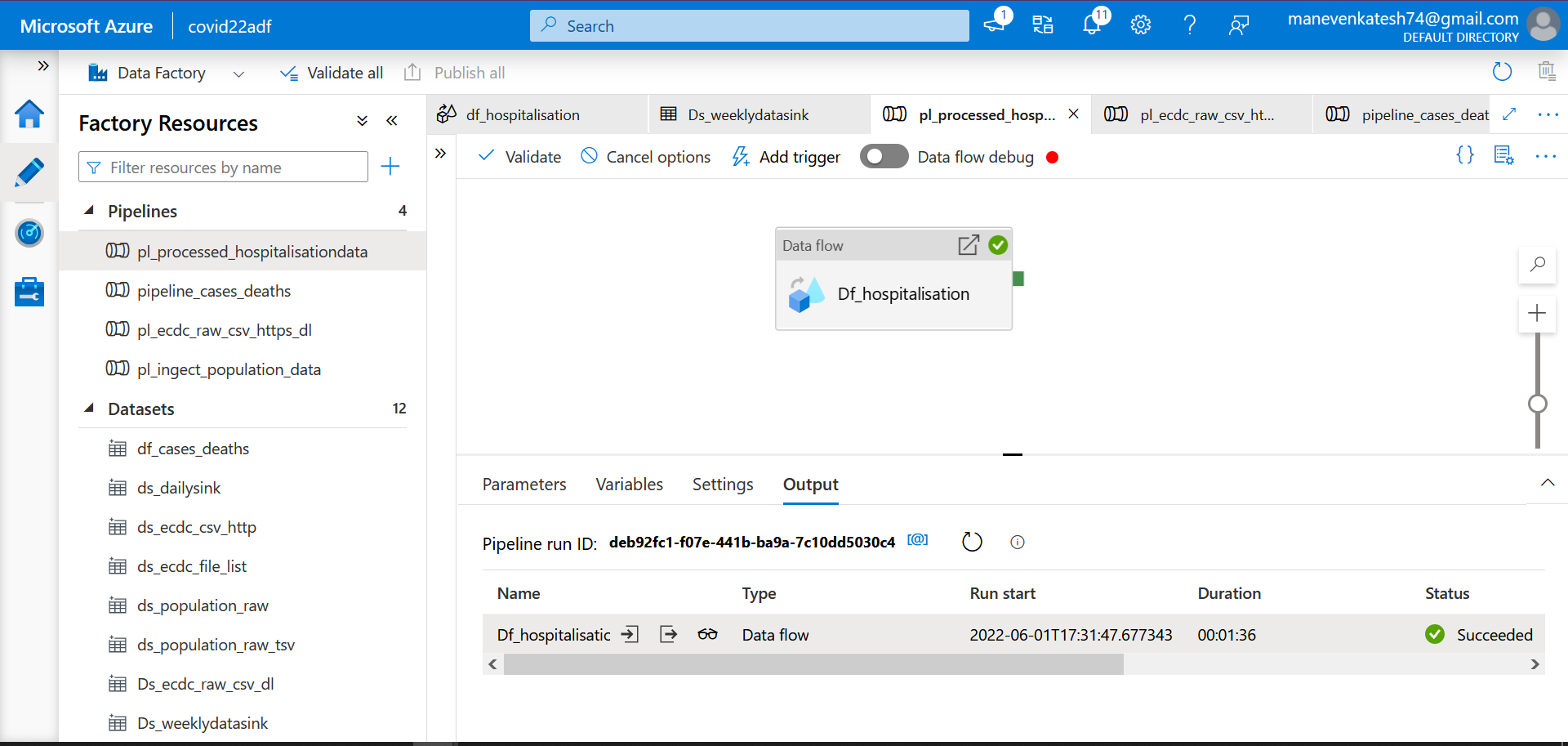
Same we performed for Hospitalisation data.

Hospitalisation data we divided that into two parts daily report and weekly report using the given column data.

We used Select ,Lookup,Conditional Split,Join,Pivot,Sort,Sink.



After Publishing the Data Flow .We created a Data flow pipeline and ran the instant trigger.To run the Data flow for whole file.



After successfully running the data transformed was moved to the data lakeusing the data set created.

Azure Databricks Activity:

* Create a Databricks in the Resource group created for covid reporting .
* Create a cluster in the Databricks.
* In Azure AD create a application and note down the Application ID and Directory ID and note down the secret key of application to connect the Azure Data Lake with the databricks .And go to Azure data lake and in access control create a storage blob data controller role.

**Mount the following data lake storage gen2 containers**

1. raw
2. processed
3. lookup

### Set-up the configs

#### Please update the following

* application-id
* service-credential
* directory-id

configs = {"fs.azure.account.auth.type": "OAuth",

"fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",

"fs.azure.account.oauth2.client.id": "fe4ed374-9a11-4e27-8304-93a05ee4a89e",

"fs.azure.account.oauth2.client.secret": "~O\_8Q~THSAAwavs\_rJDmA3q5l7hJRKaF~CLuSaXq",

"fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/4af5c1e0-c48d-4cf3-b5c9-b864ae68be1e/oauth2/token"}

🡪Raw container

dbutils.fs.mount(

source = "abfss://raw@covidadl22.dfs.core.windows.net/",

mount\_point = "/mnt/covidadl22/raw",

extra\_configs = configs)

🡪Processed container:

dbutils.fs.mount(

source = "abfss://processed@covidadl22.dfs.core.windows.net/",

mount\_point = "/mnt/covidadl22/processed",

extra\_configs = configs)

🡪lookup container

dbutils.fs.mount(

source = "abfss://lookup@covidadl22.dfs.core.windows.net/",

mount\_point = "/mnt/covidadl22/lookup",

extra\_configs = configs)

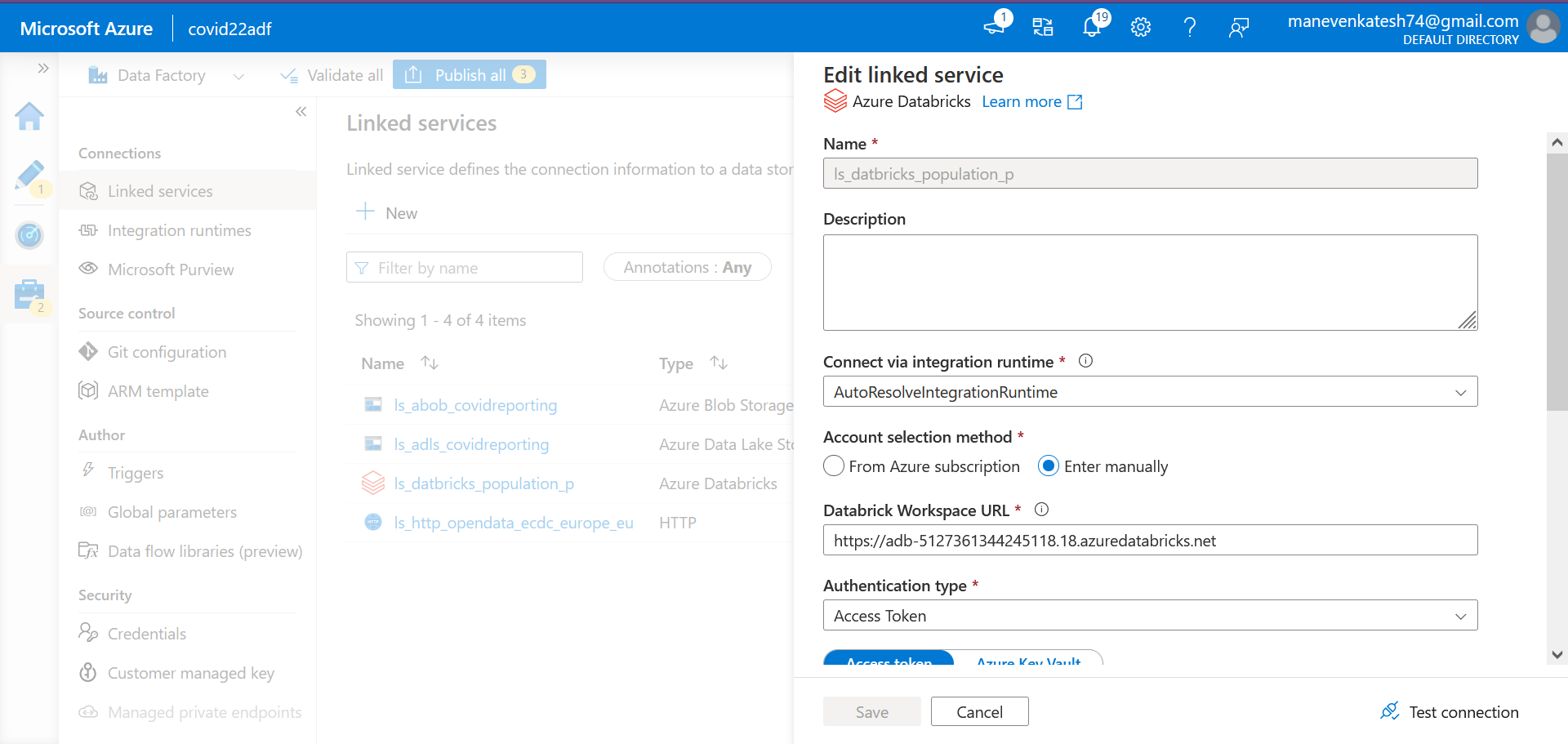
## Transform Population By Age data by performing the transformations below

#### -----------------------------------------------------------------------

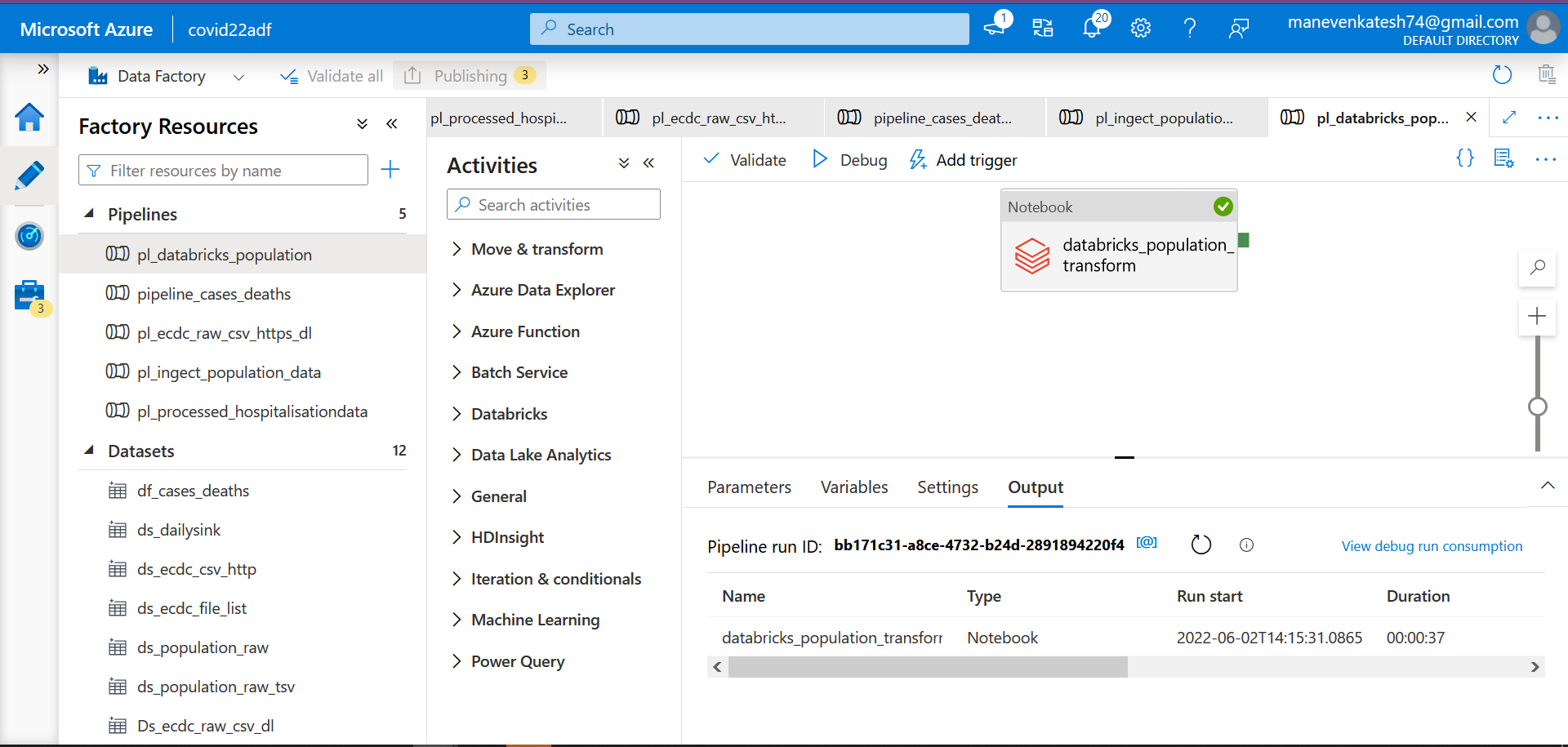
1. Split the country code & age group
2. Exclude all data other than 2019
3. Remove non numeric data from percentage
4. Pivot the data by age group
5. Join to dim\_country to get the country, 3 digit country code and the total population.

Created a cluster,and imported the py folder for transform.

Linked service for databricks



Created a pipeline for databricks:

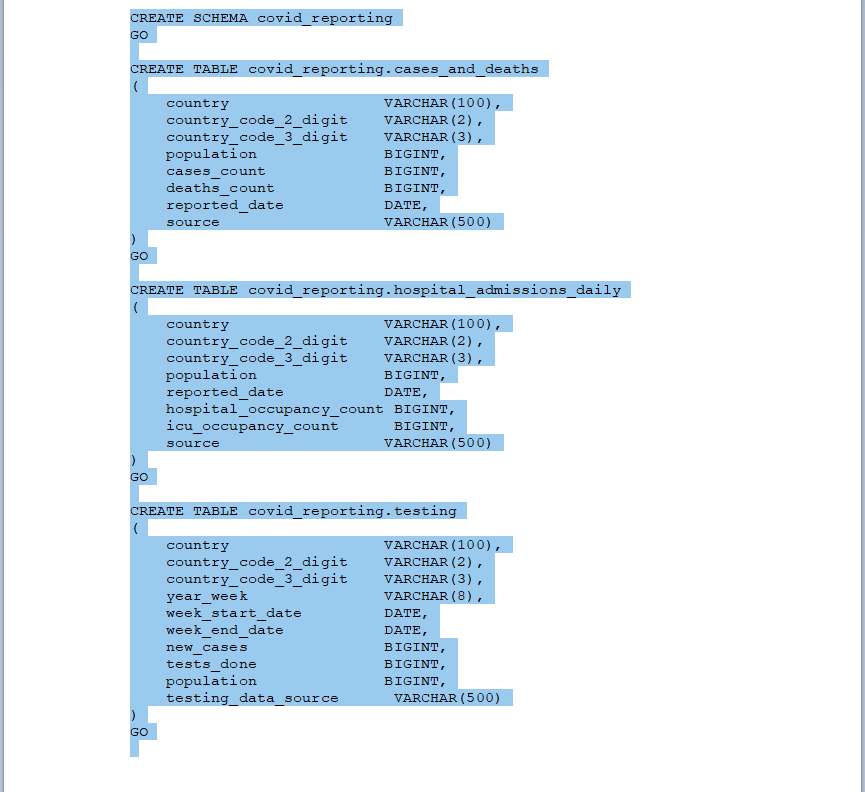


The transformed population file was moved to processed folder.

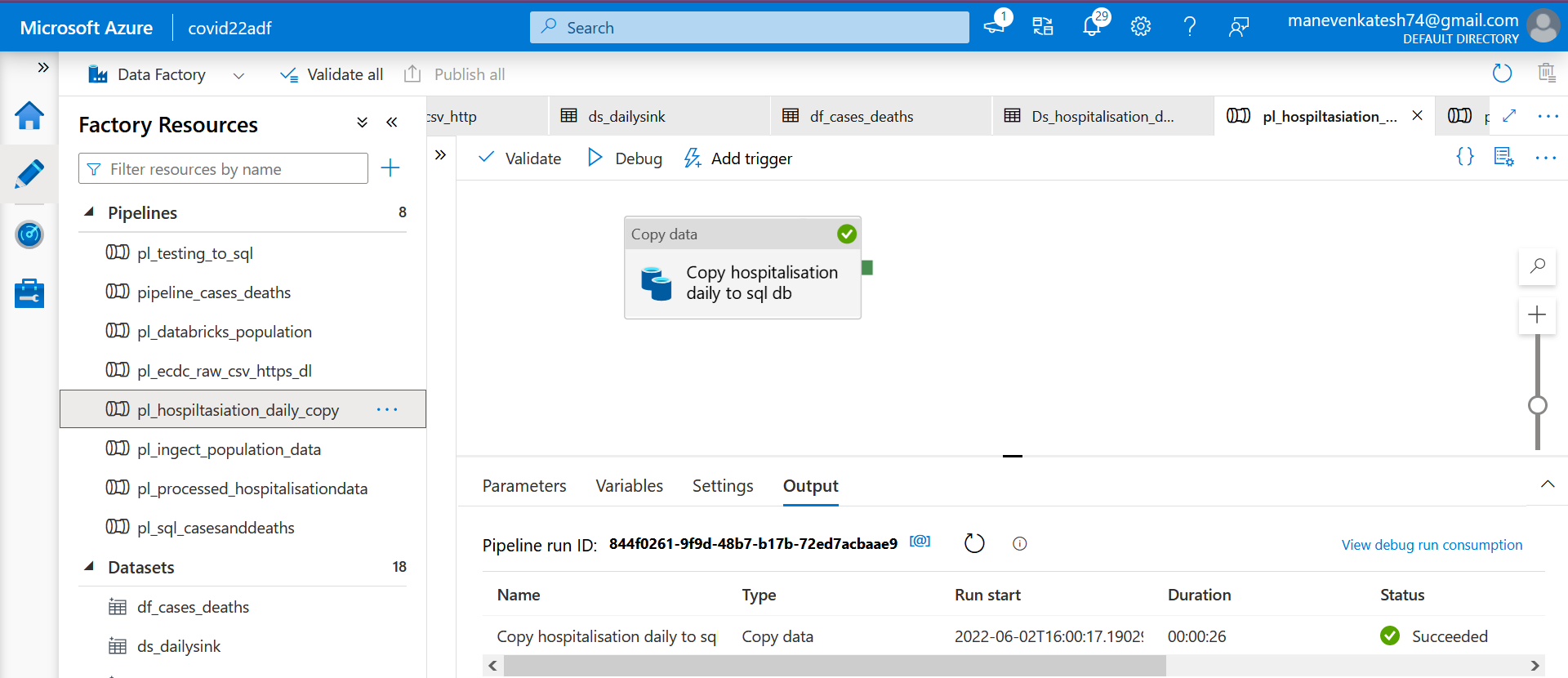
Now as all the data is transformed to the prcessed folder in the data lake .

Now we will move the data from processed folder from Data lake to SQL database .

We created 3 tables for Cases and deaths,Hspitalsation and Testing data files under schema **covid\_reporting.**



Graphical user interface, text, application, chat or text message

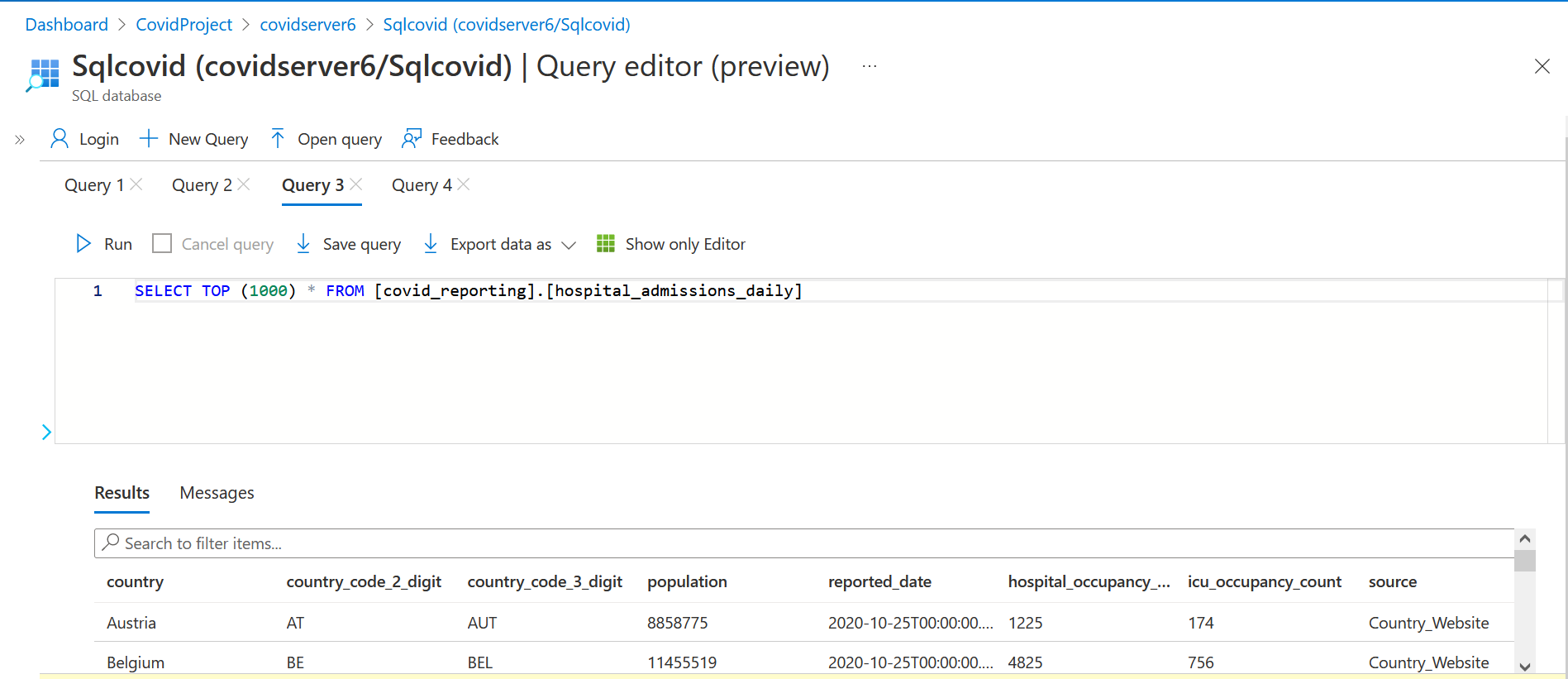
Description automatically generated

Graphical user interface, text, application

Description automatically generated

After all the pipelines were run successfully.The data was copied to the SQL tables created respectively.

Below are the output :



Now the task has been completed we now imported the SQL tables to Power BI for Visualisaton part.



