Final Project

Covid Use Case

Project By- Aniket Bhagwat Mohite.

Employee Id- 2261295

Domain- AIA-Azure

Cohort code- CDB23CSDDW006

I. Project Requirement:

Get the data from ingest path and after transformation using mapping dataflow store in a ADLS different storage container named ‘transformPath’. We need to do Aggregate and Rank transformation to find the continents which have max monthly death data in a single parquet file for each month in ADLS.

II. Resource Provided:

Data Given with Five CSV files containing covid information across the world.

III. Resource Used:

Azure Service	**Azure Service Name**
Resource Group	-default-
Azure Data Lake Storage Gen2 Account	Covidadls46
Azure Data Lake Storage Gen2 Container	covid
Azure Data Factory	Covid-ADF-46
Linked Service (adls gen2)	LS_adls

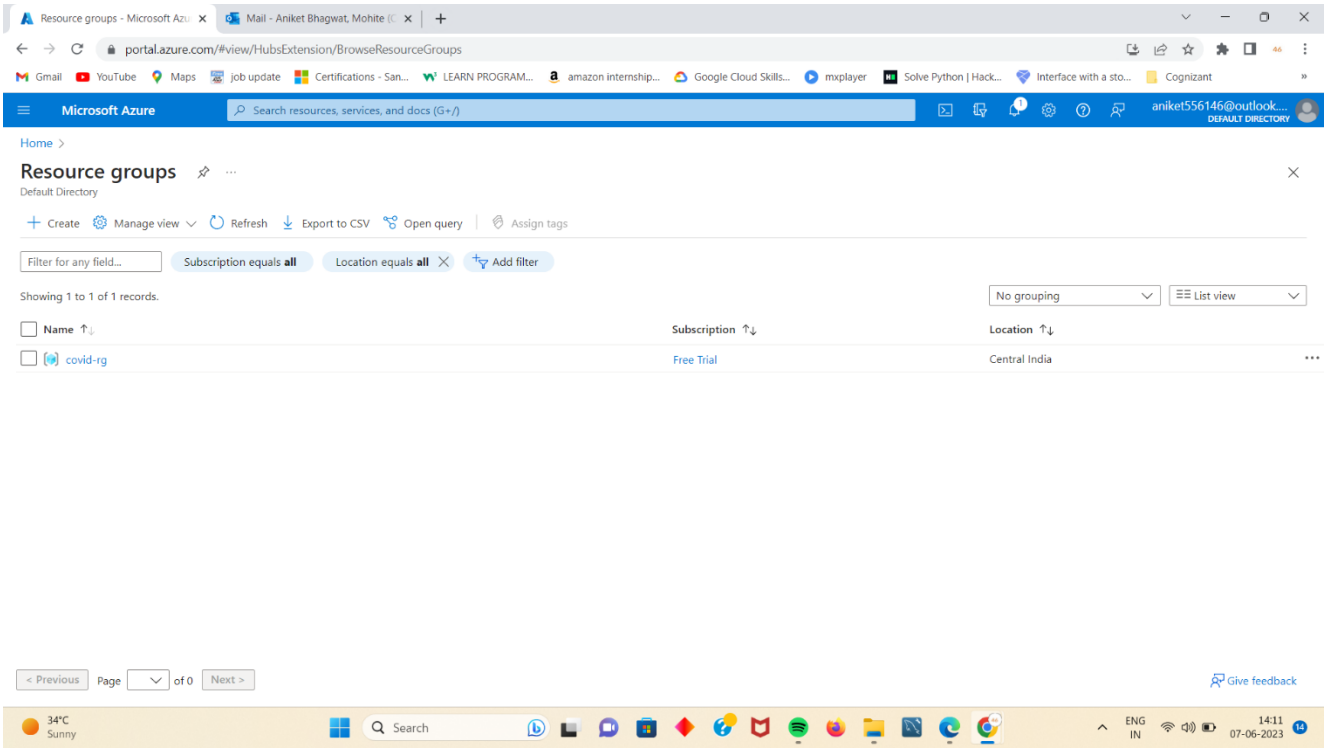
Process:

I have completed this project in three steps. Following are these three steps.

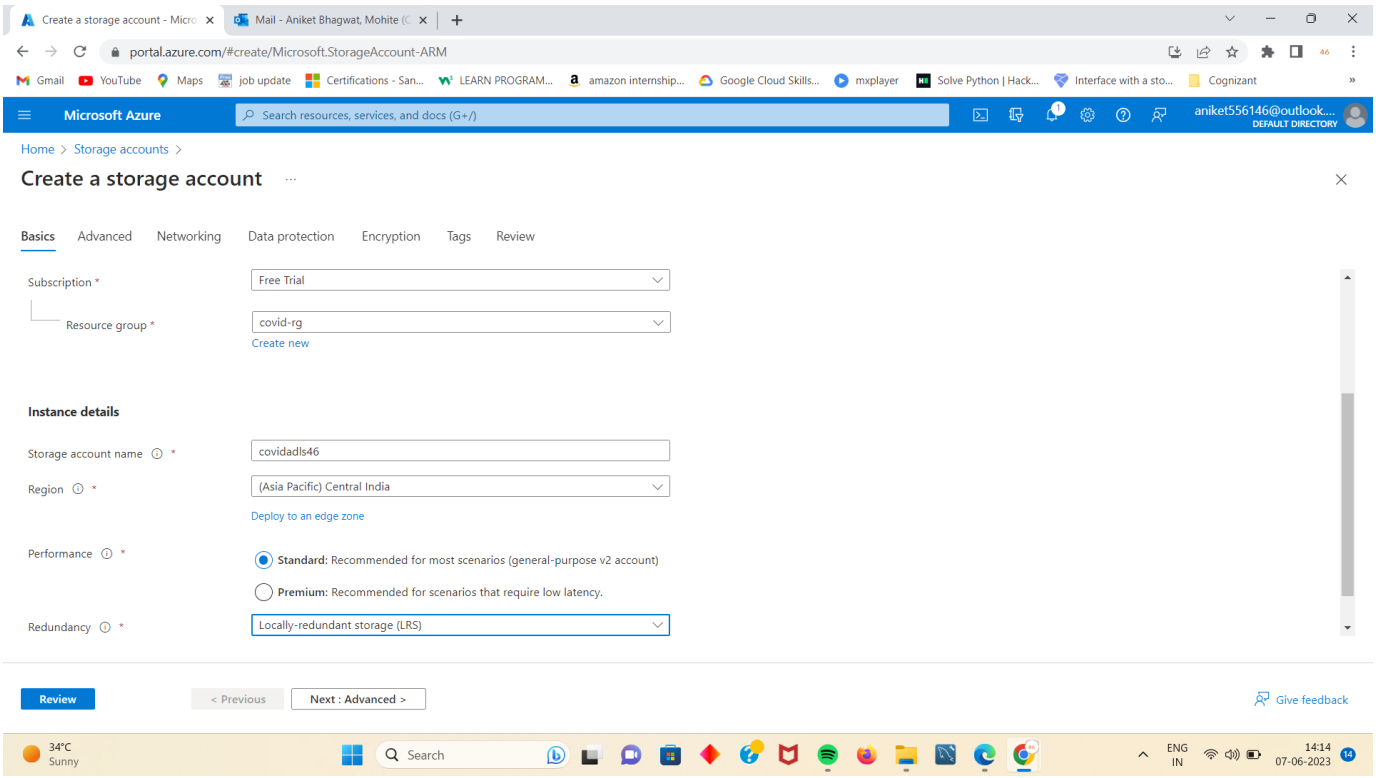
1) Copy Data from Source to Destination
2) Merge All files of Destination Folder
3) Find Ranking of Continent which have max daily death data in Month wise folders.

1.Copy Data from Source to Destination –

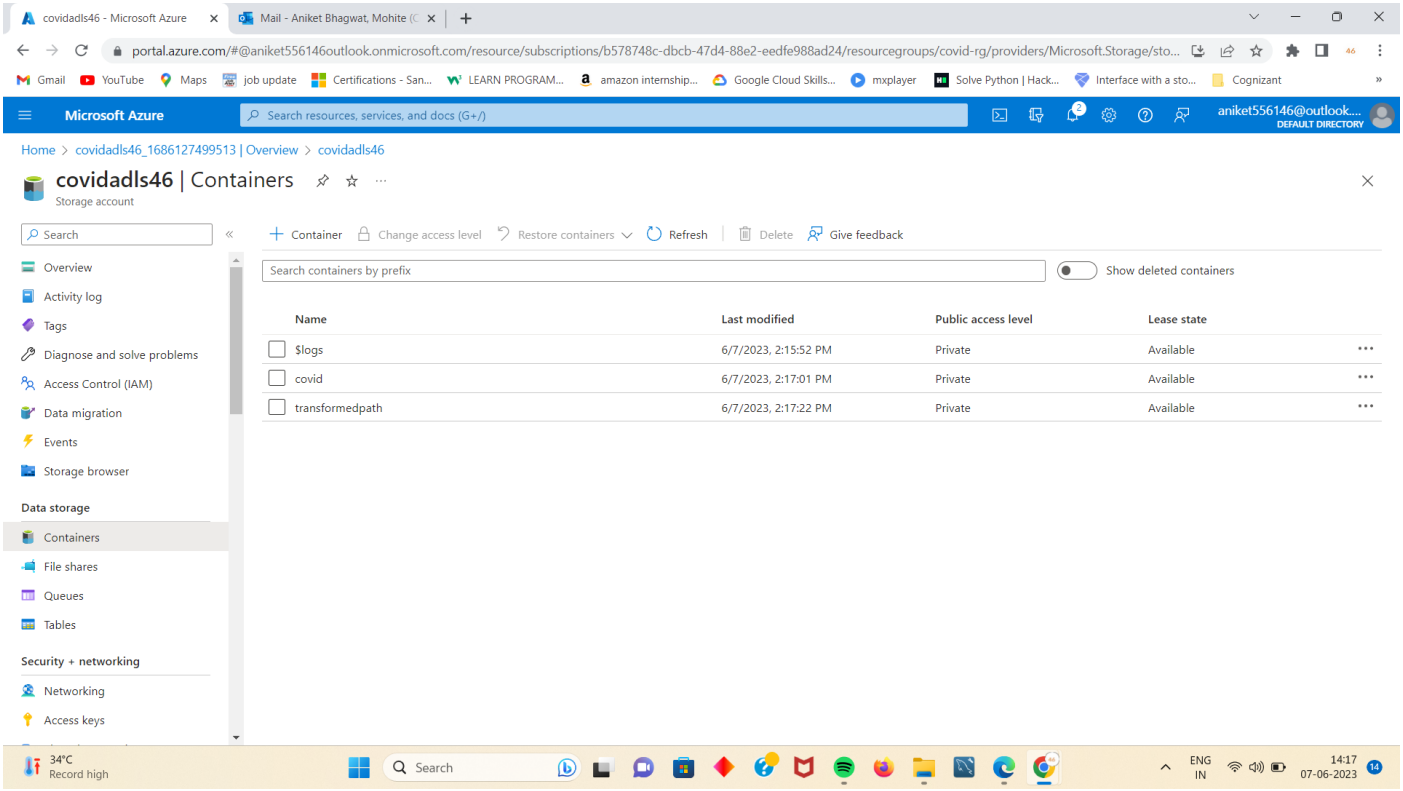
- Create one resource group named ‘Covid-rg’.



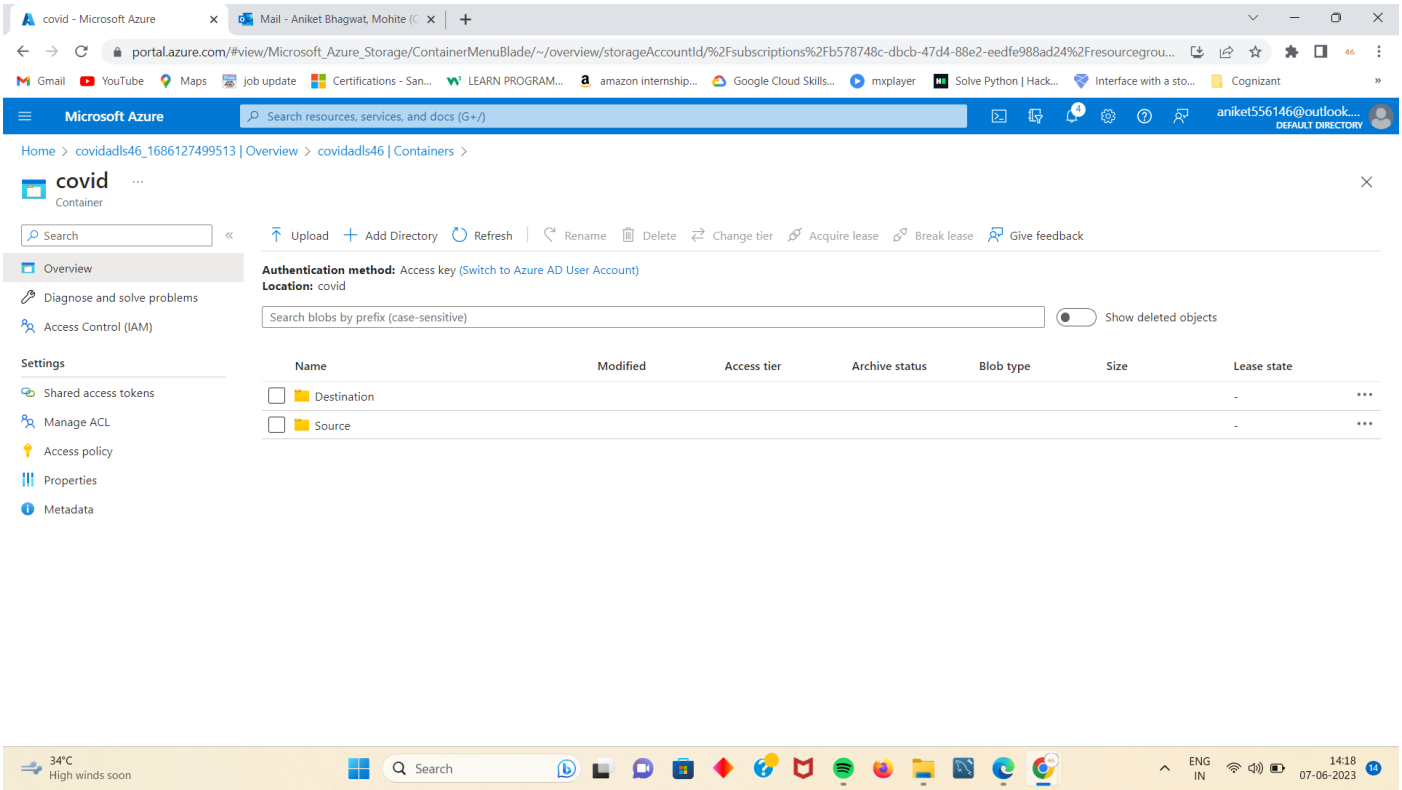
- Create storage account using same resource group.



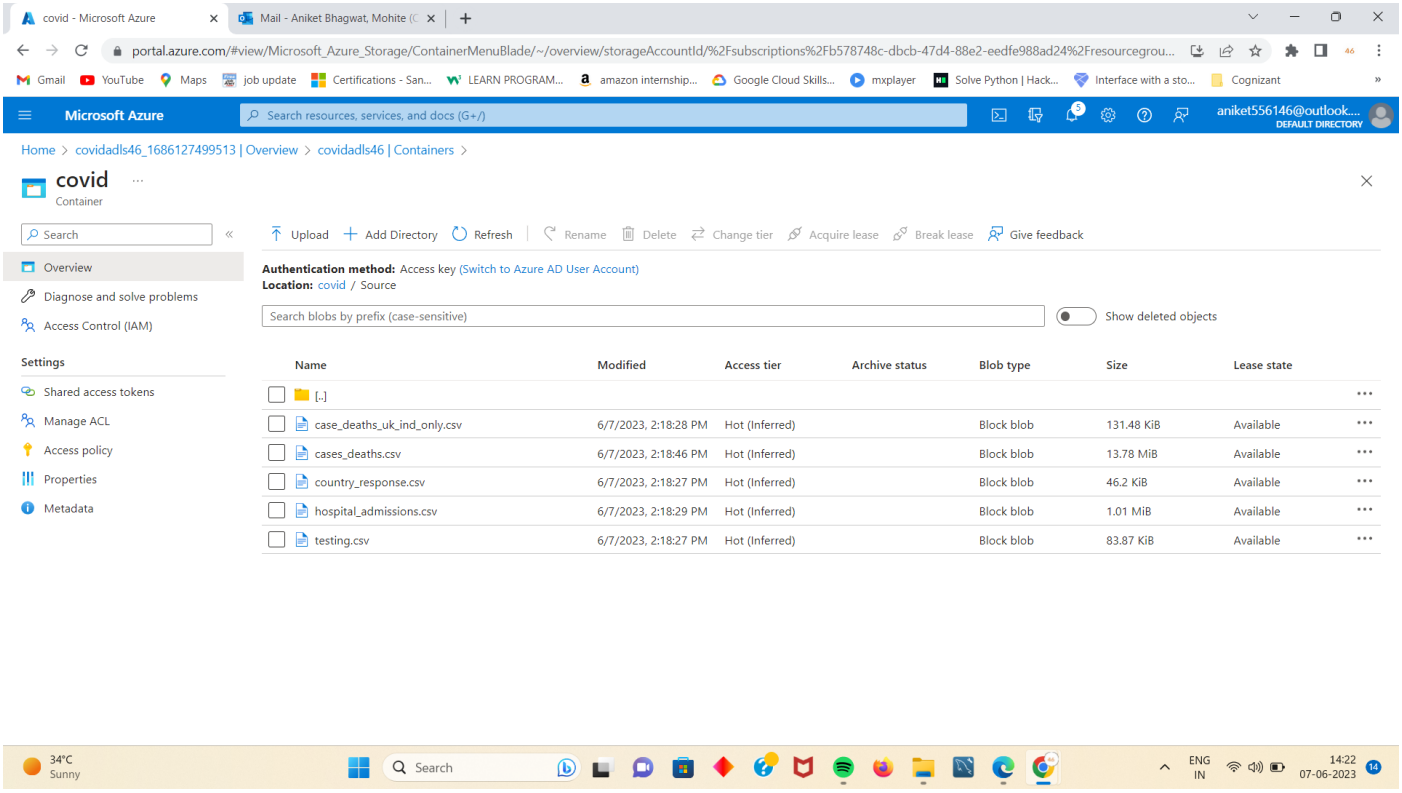
- create two containers named covid and transformedpath.



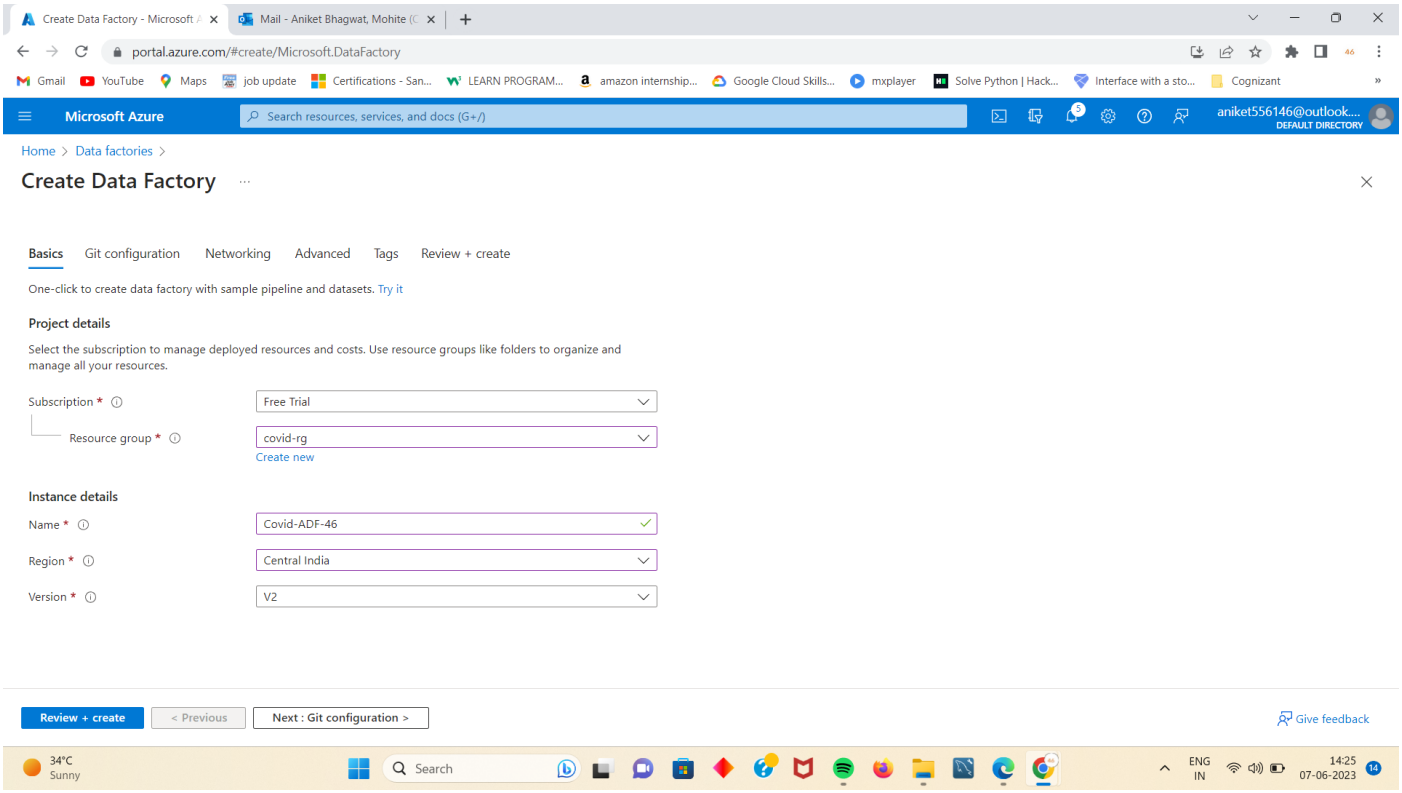
- In covid container create two directories 1) Source 2) Destination



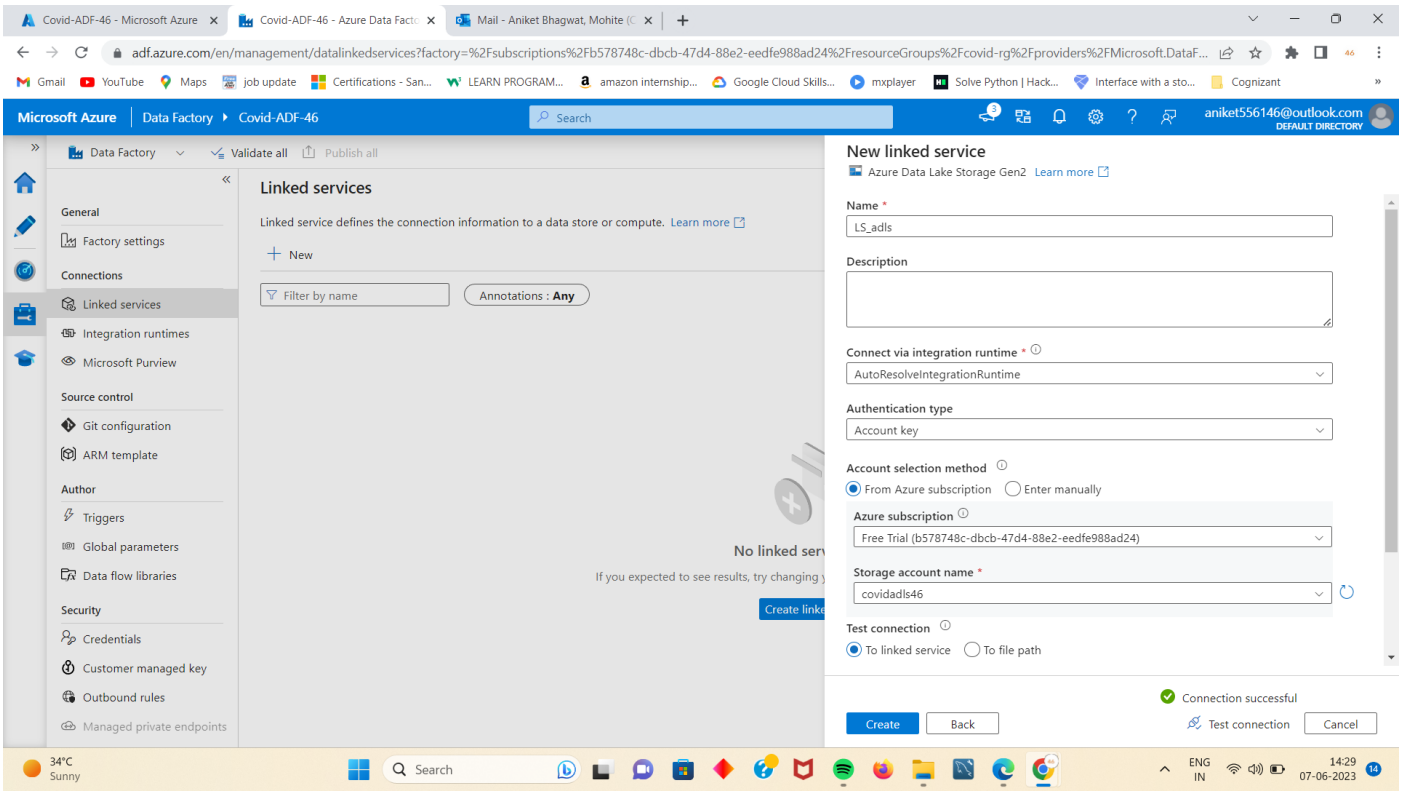
- Upload all given files to source directory.



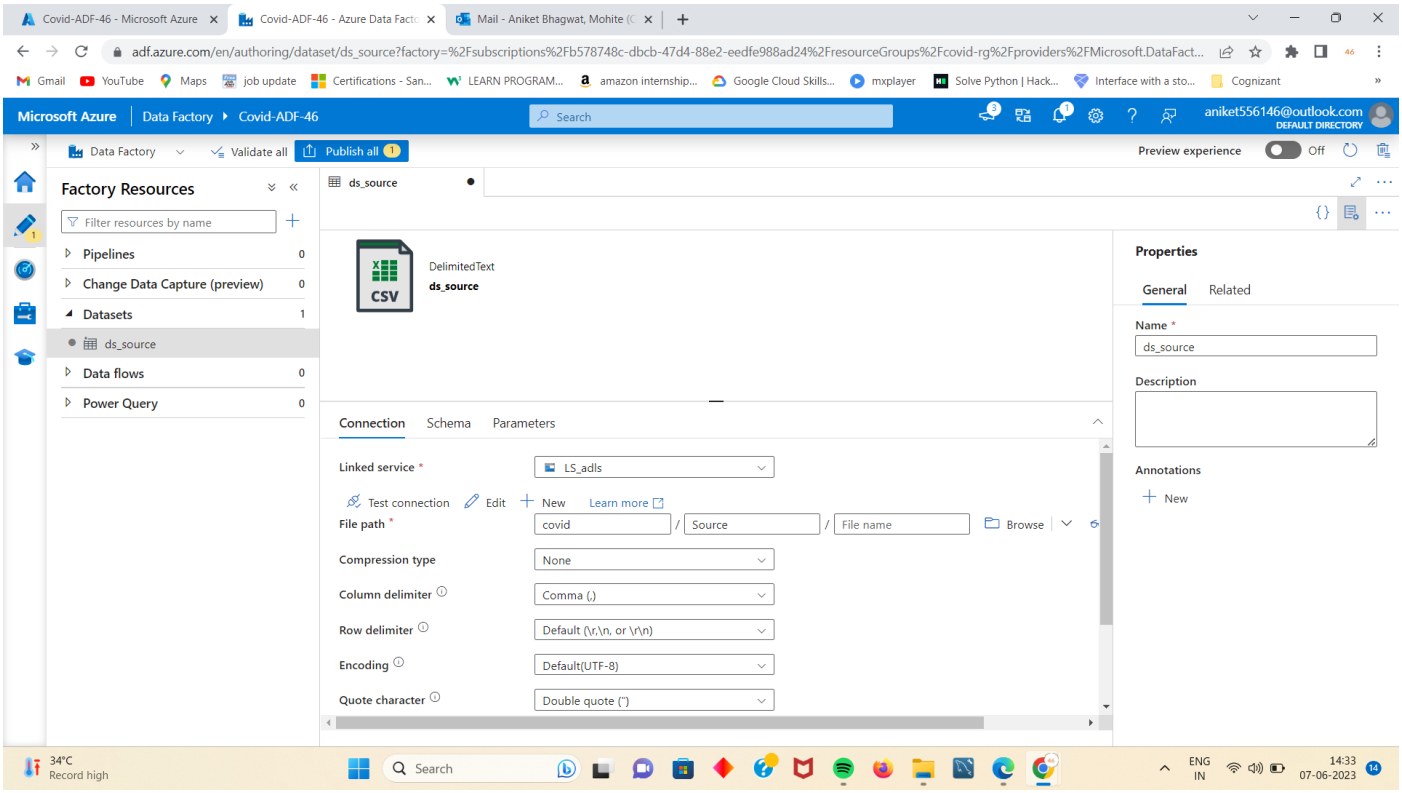
- Create Azure Data Factory named ‘Covid-ADF-Project46’ in same resource group.



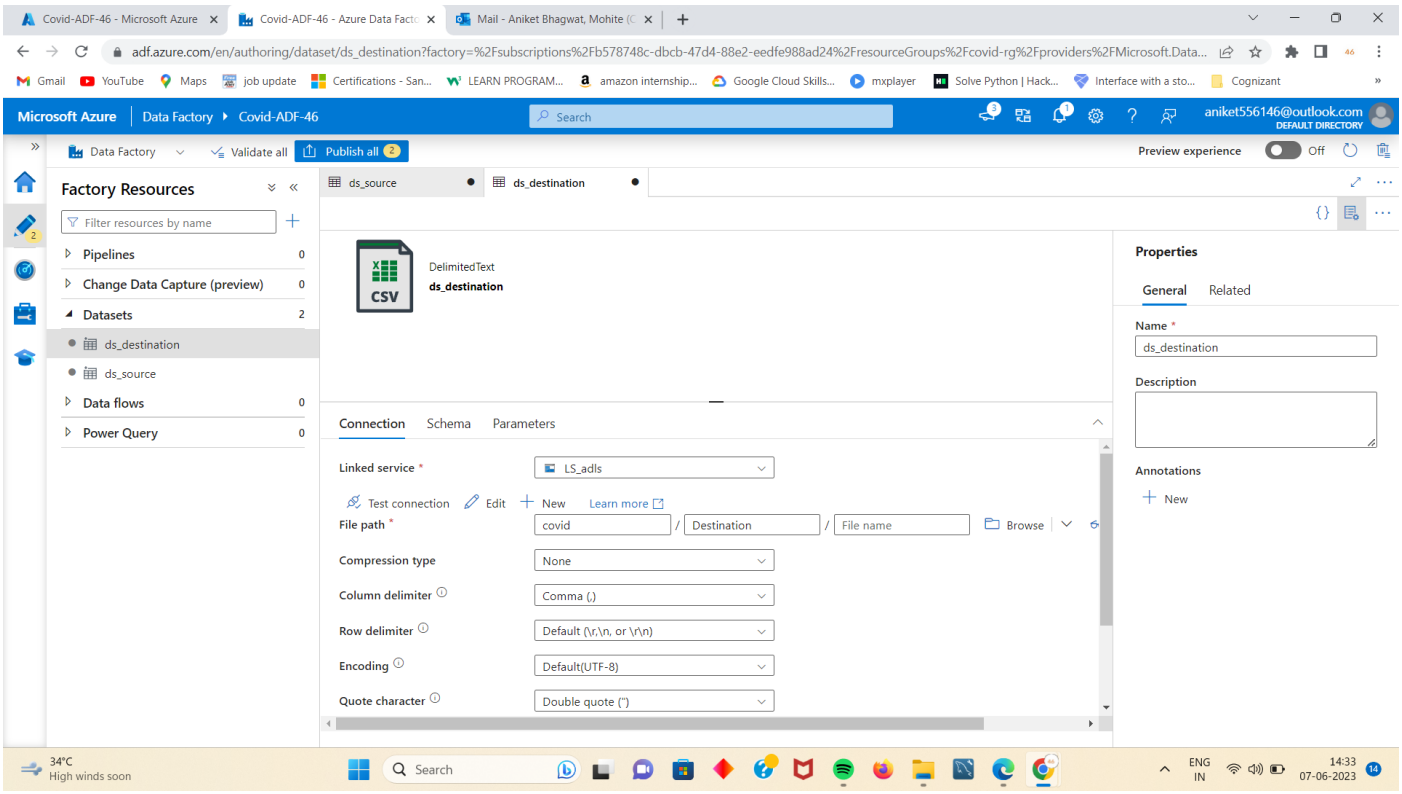
- Create Linked Service in Azure Data Factory using storage account.



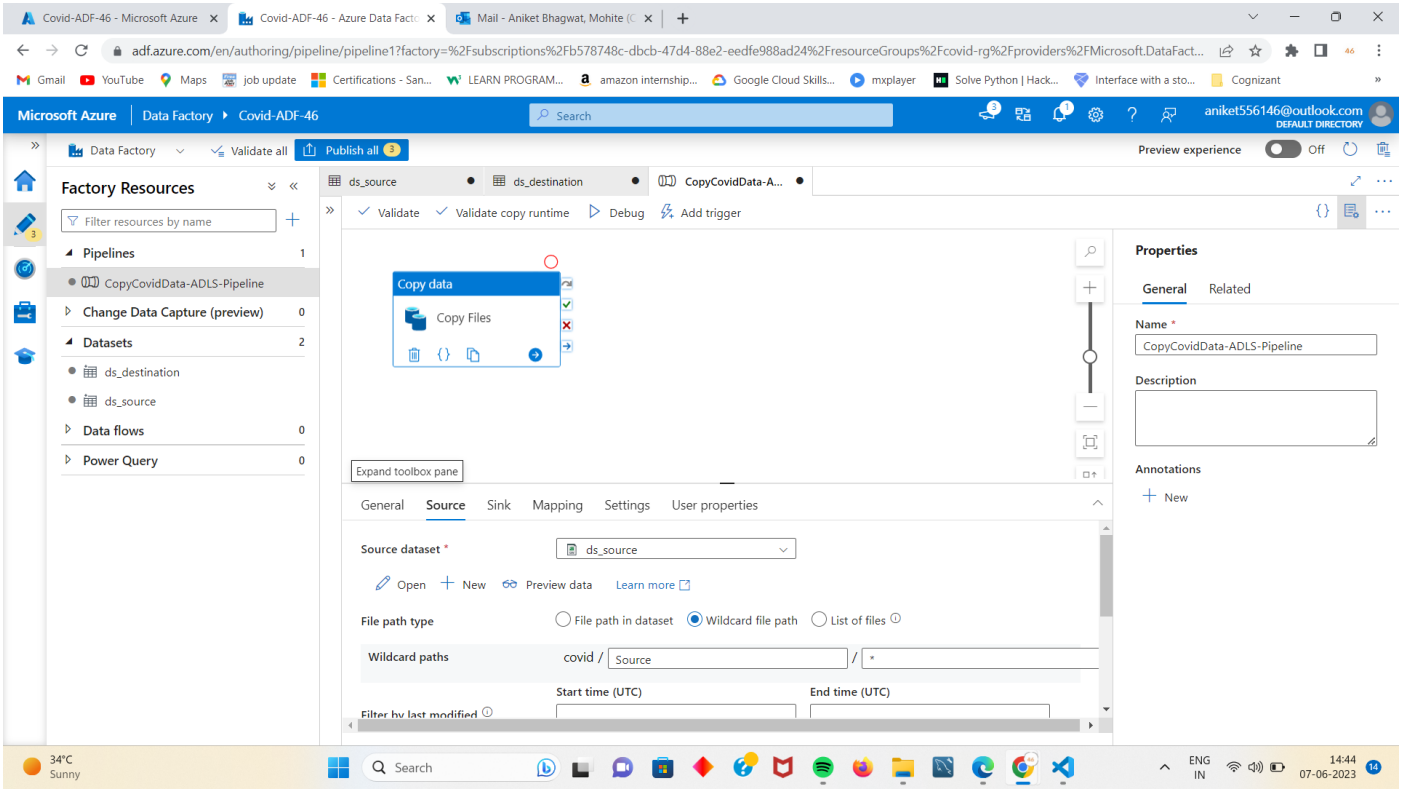
- Create Data Sets to Copy all data from Source and Destination
  1) Create Data Set ds_source for source data

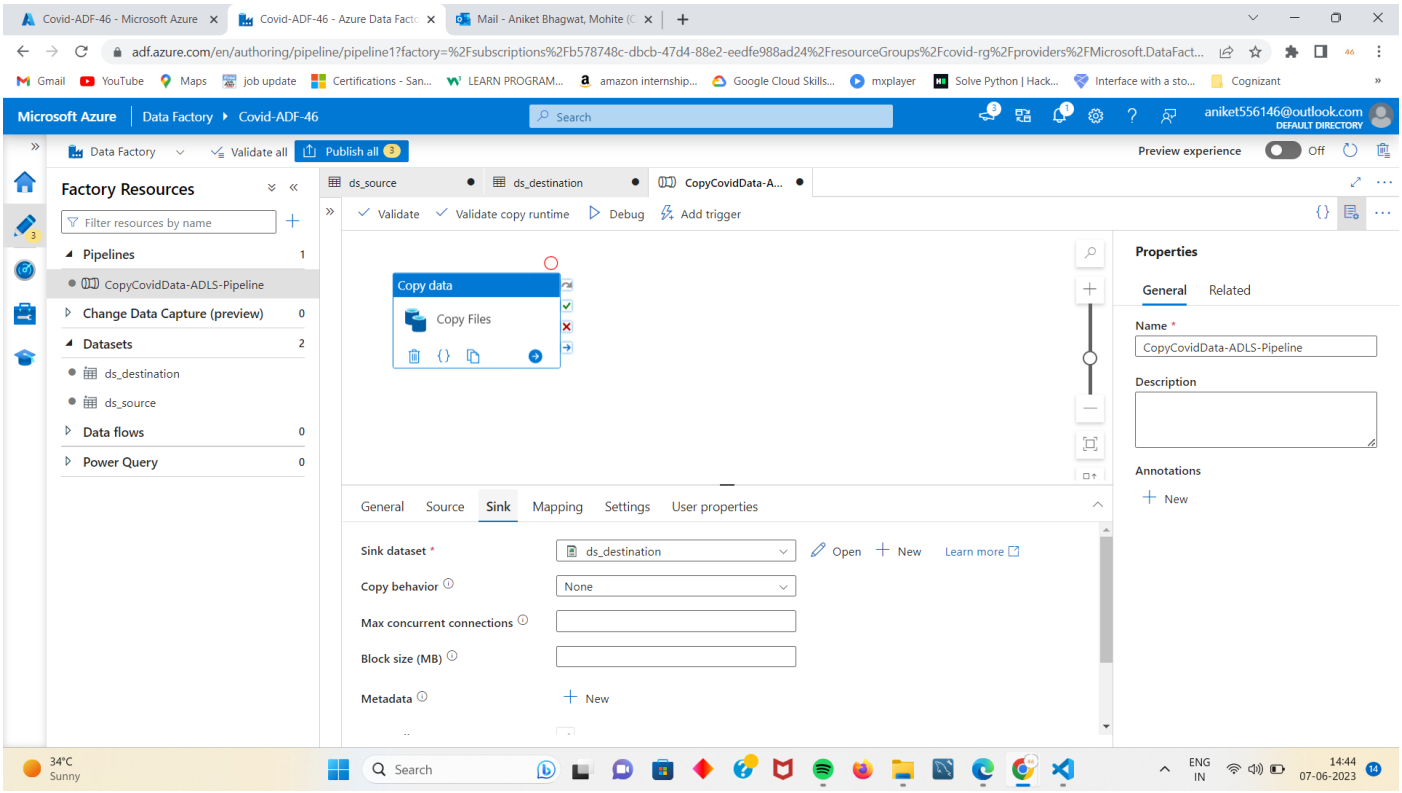


2) Create Data Set ds_destination to get data in Destination directory.

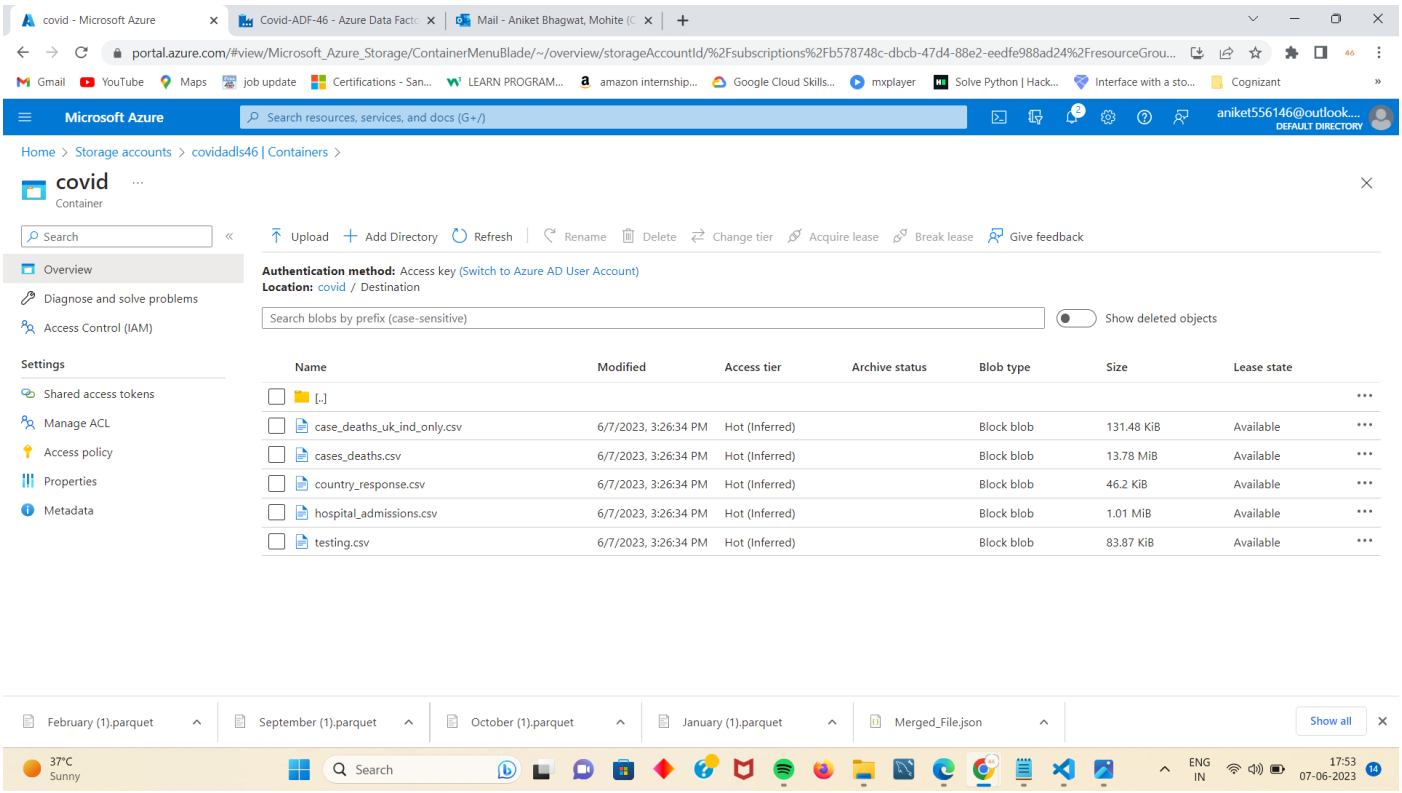


- Use copy data activity to copy data from Source to Destination directory.




Microsoft Azure
Data Factory
Covid-ADF-46
Factory Resources
Pipelines
CopyCovidData-ADLS-Pipeline
Change Data Capture (preview)
Datasets
ds_destination
ds_source
Data flows
Power Query
Copy data
Copy Files
Validate
Validate copy runtime
Debug
Add trigger
Properties
General
Related
Name *
CopyCovidData-ADLS-Pipeline
Description
Annotations
New
General
Source
Sink
Mapping
Settings
User properties
Sink dataset *
ds_destination
Open
New
Learn more
Copy behavior
None
Max concurrent connections
Block size (MB)
Metadata
New

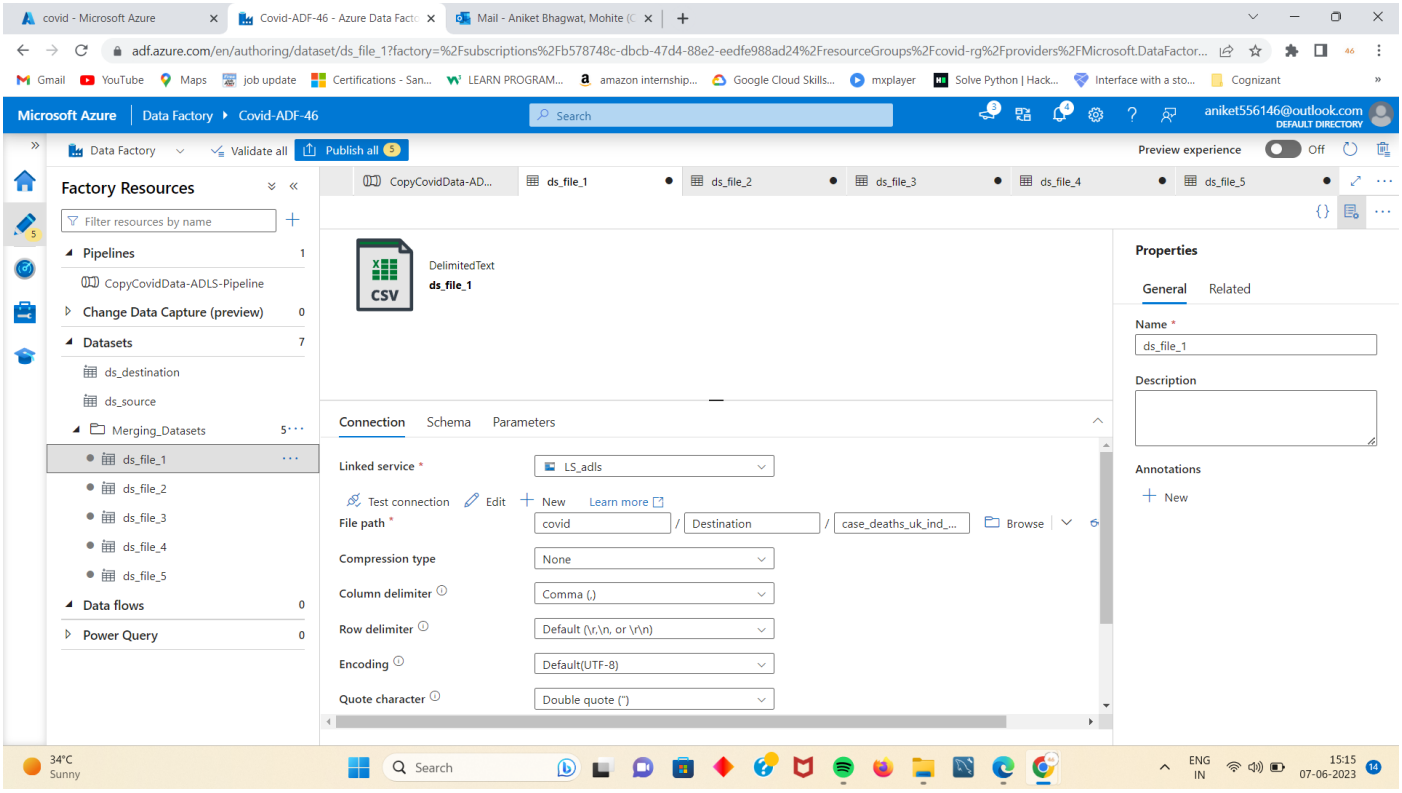
Output:


Microsoft Azure
Home > Storage accounts > covidadls46 | Containers >
covid
Container
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Authentication method: Access key (Switch to Azure AD User Account)
Location: covid / Destination
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
[..]
case_deaths_uk_ind_only.csv 6/7/2023, 3:26:34 PM Hot (Inferred) Block blob 131.48 KiB Available
cases_deaths.csv 6/7/2023, 3:26:34 PM Hot (Inferred) Block blob 13.78 MiB Available
country_response.csv 6/7/2023, 3:26:34 PM Hot (Inferred) Block blob 46.2 KiB Available
hospital_admissions.csv 6/7/2023, 3:26:34 PM Hot (Inferred) Block blob 1.01 MiB Available
testing.csv 6/7/2023, 3:26:34 PM Hot (Inferred) Block blob 83.87 KiB Available
February (1).parquet
September (1).parquet
October (1).parquet
January (1).parquet
Merged_File.json
Show all

2.Merge All files of Destination Folder

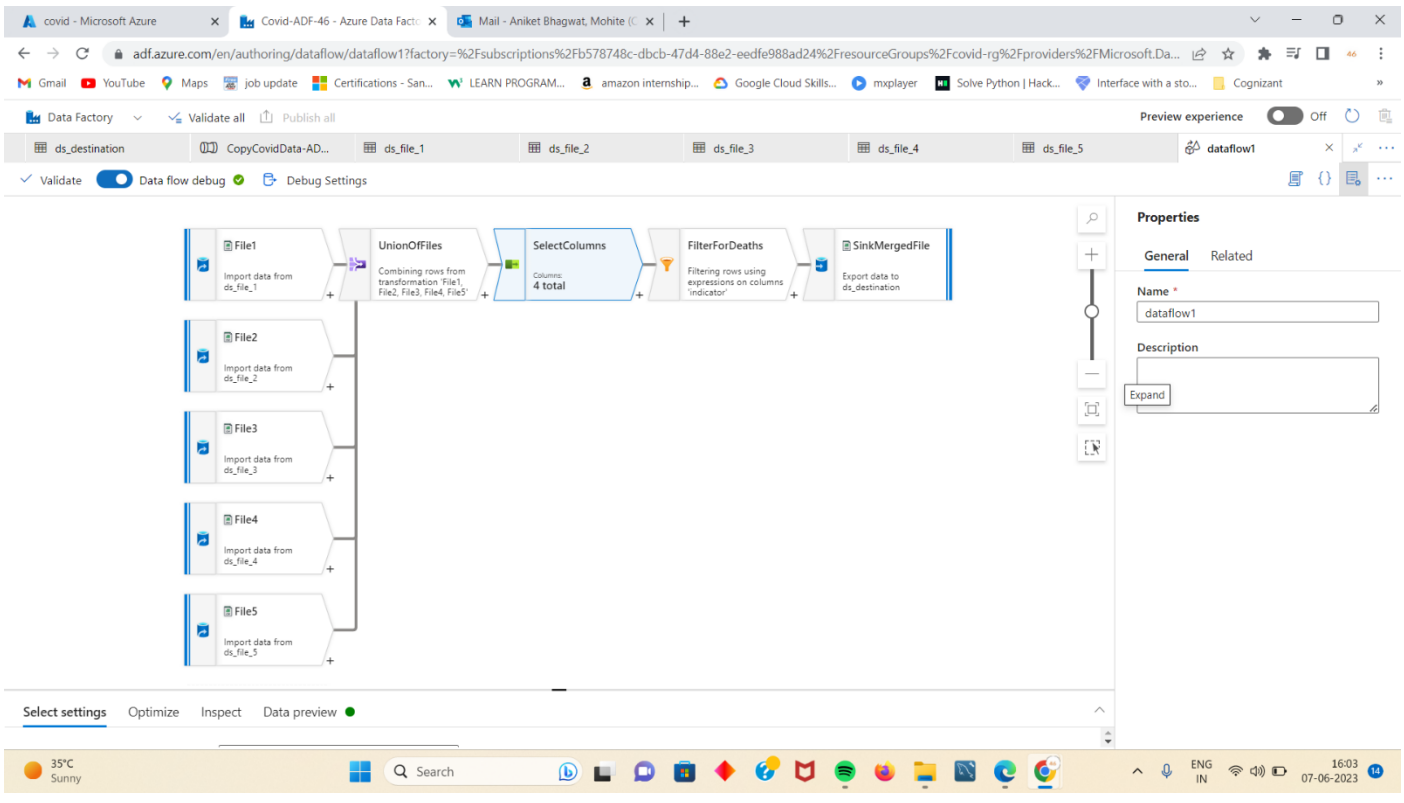
Here I have used 5 datasets for each file to get as source in Dataflow.

1)Create Datasets for each file give path of Destination folder.

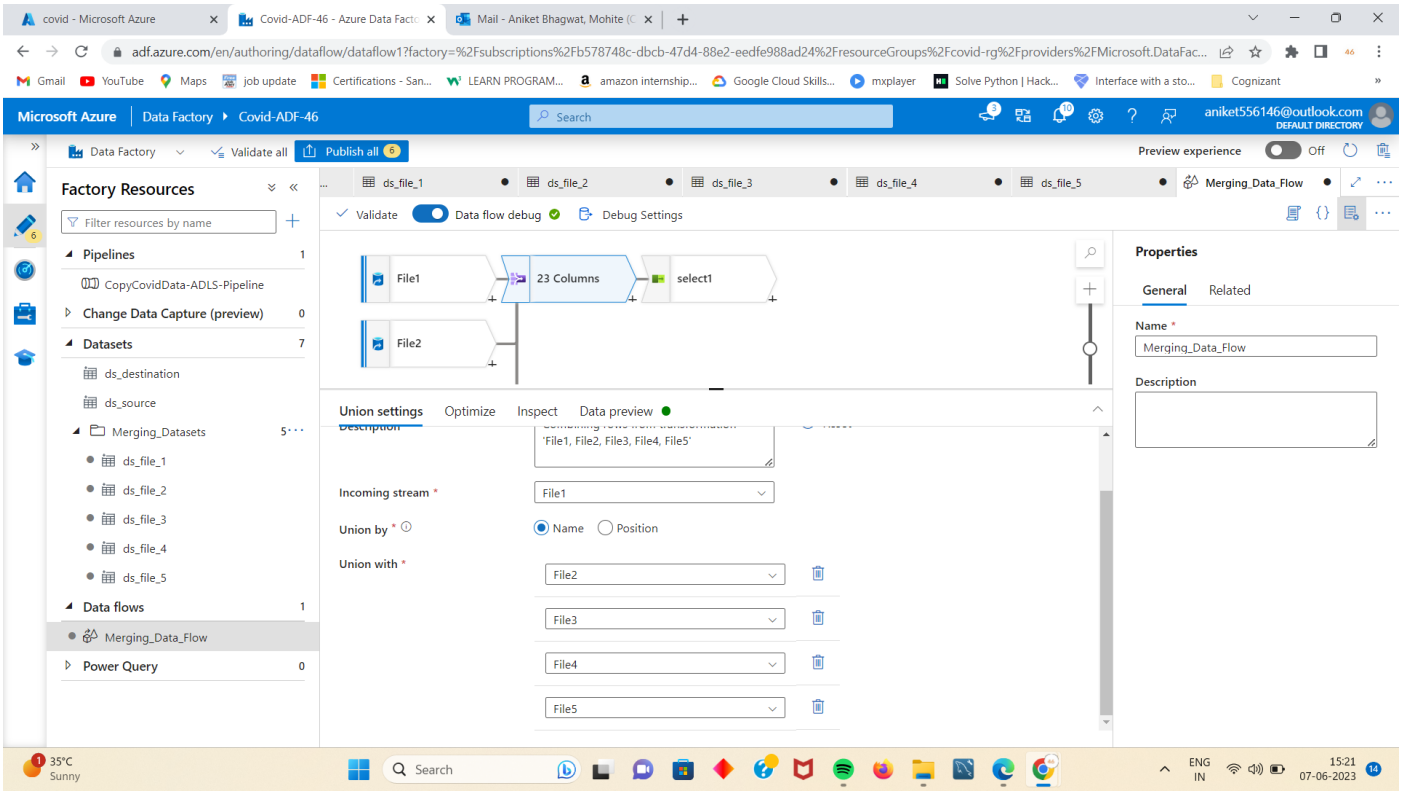


2)Create Dataflow

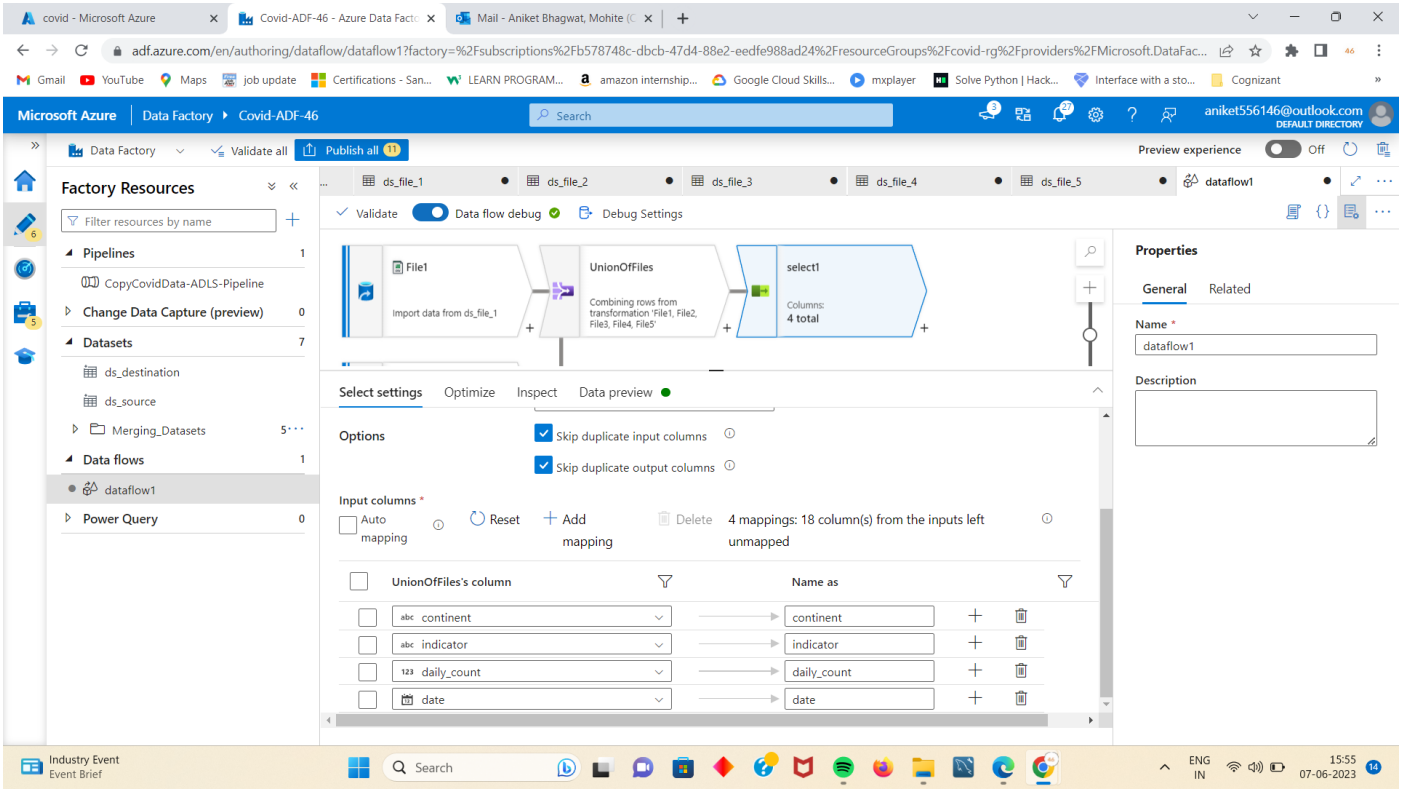
1) Select each file Dataset as Source



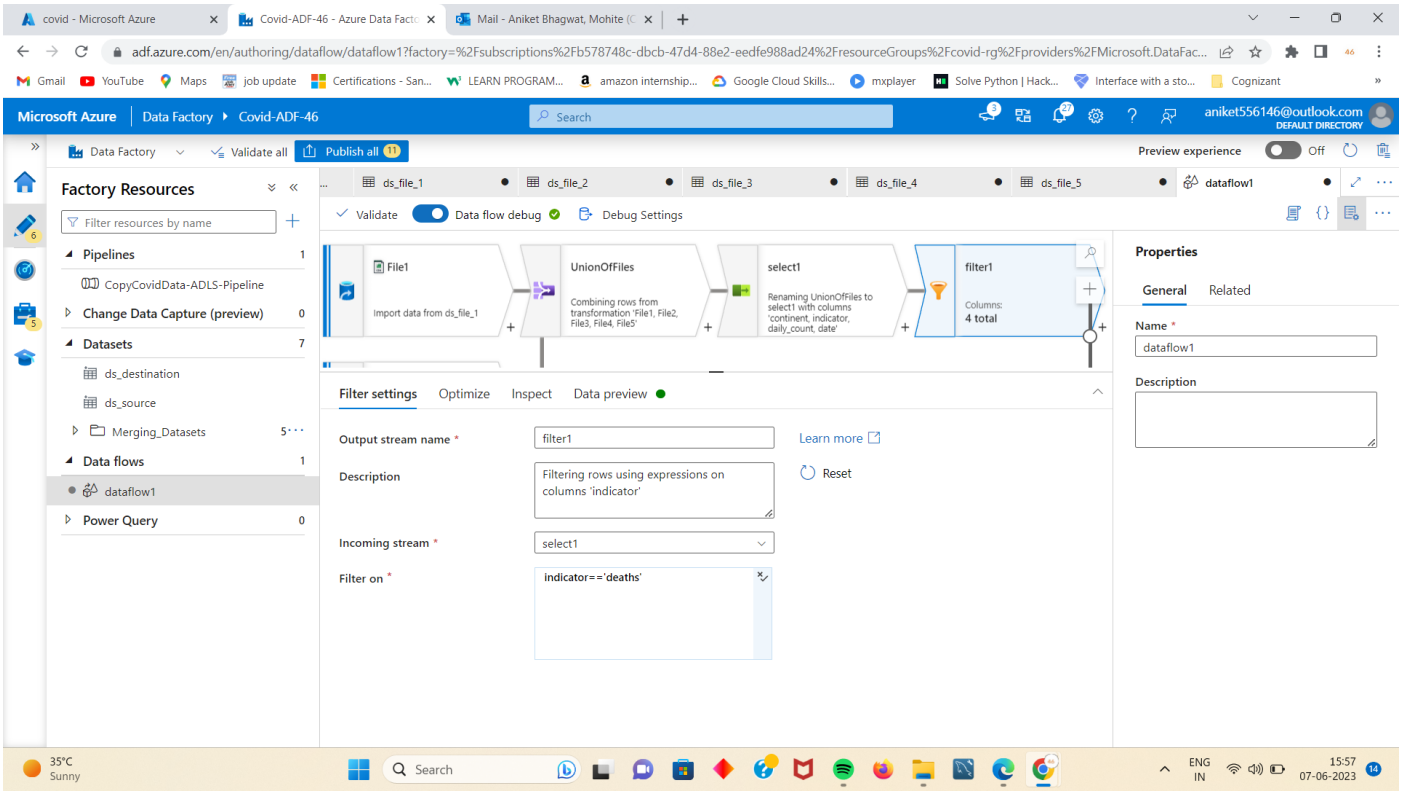
2) Use Union Activity to combine each file.



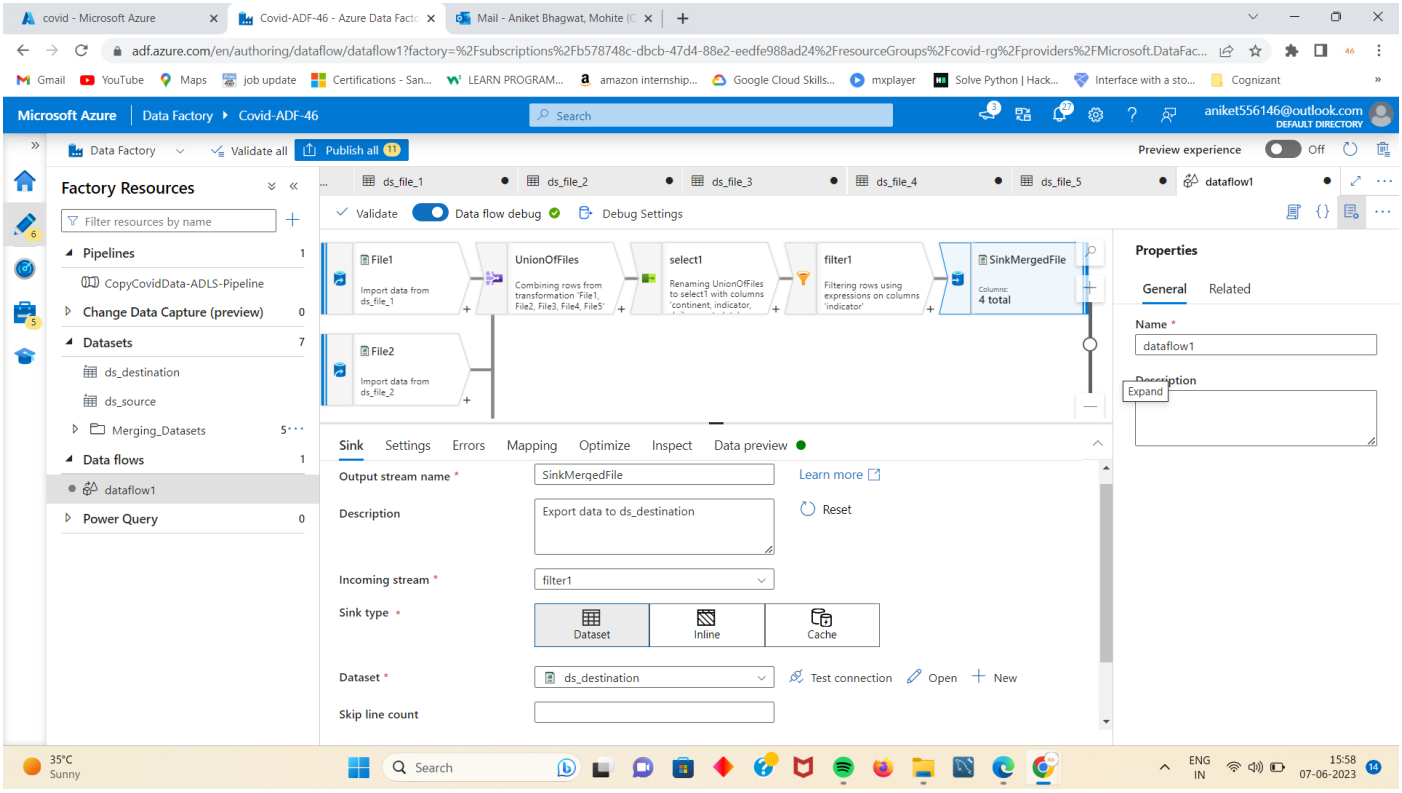
3) Use Select activity to Select particular columns required as output.



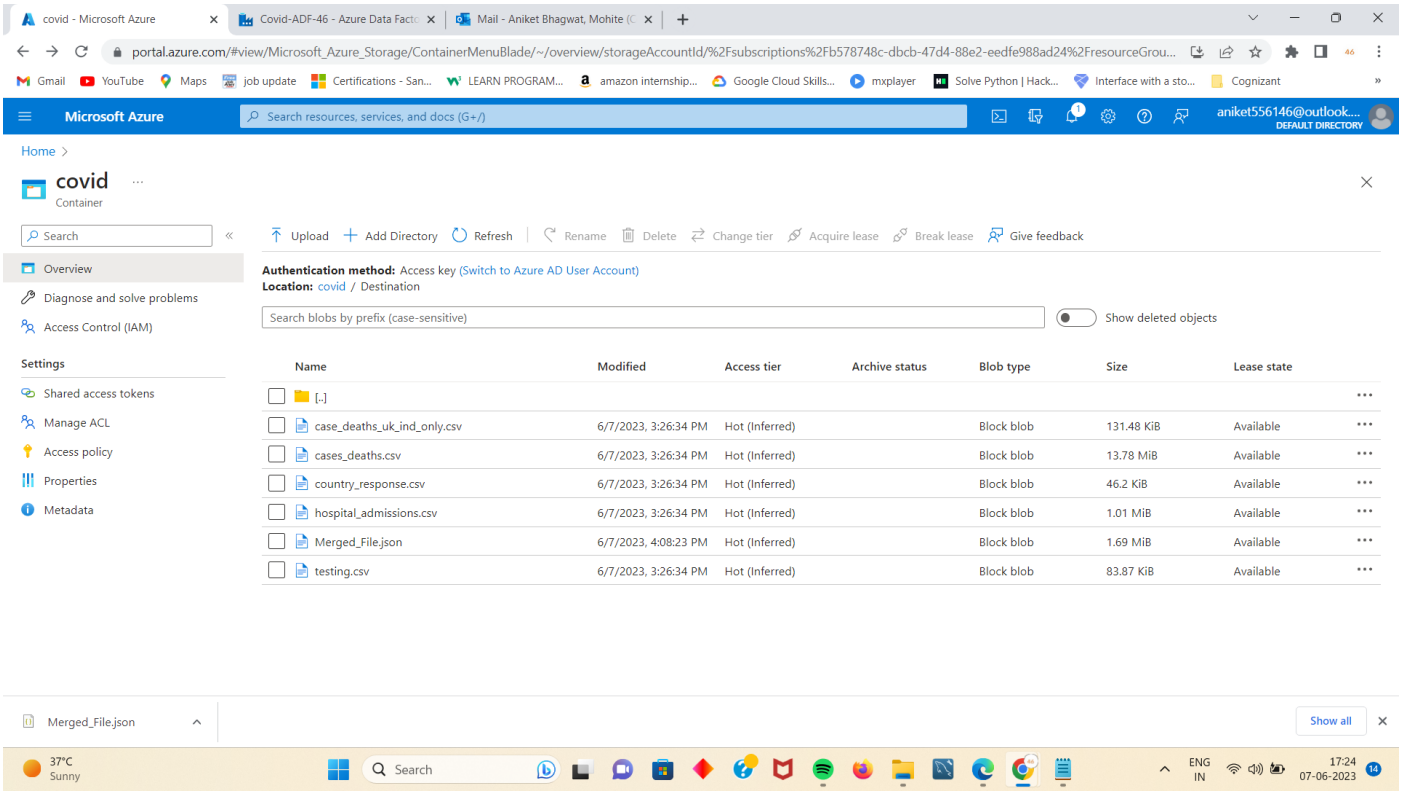
4) Use Filter to get only death cases in merged file.



5) Use sink to get output file as merged file, In setting choose output as single file and give name to file.

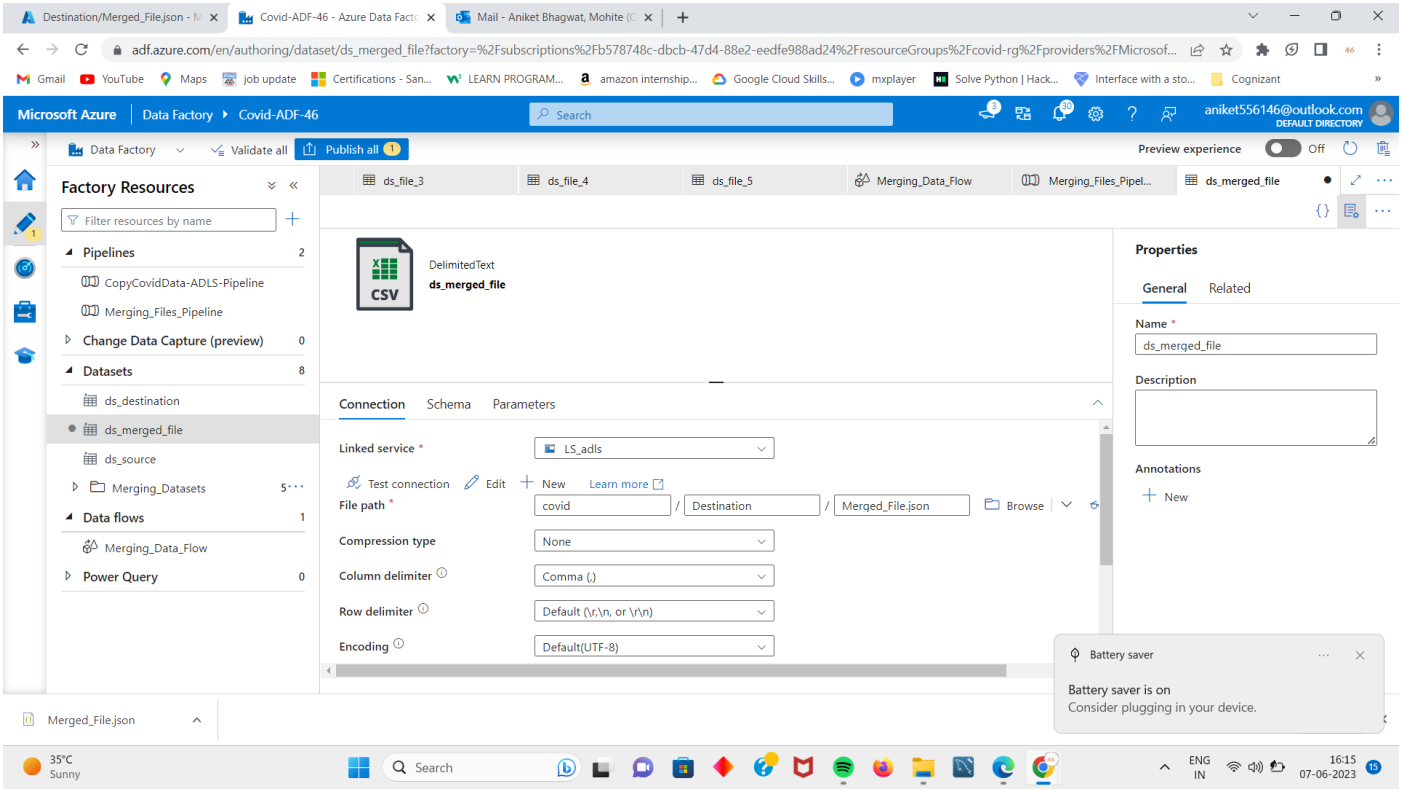


Output File -

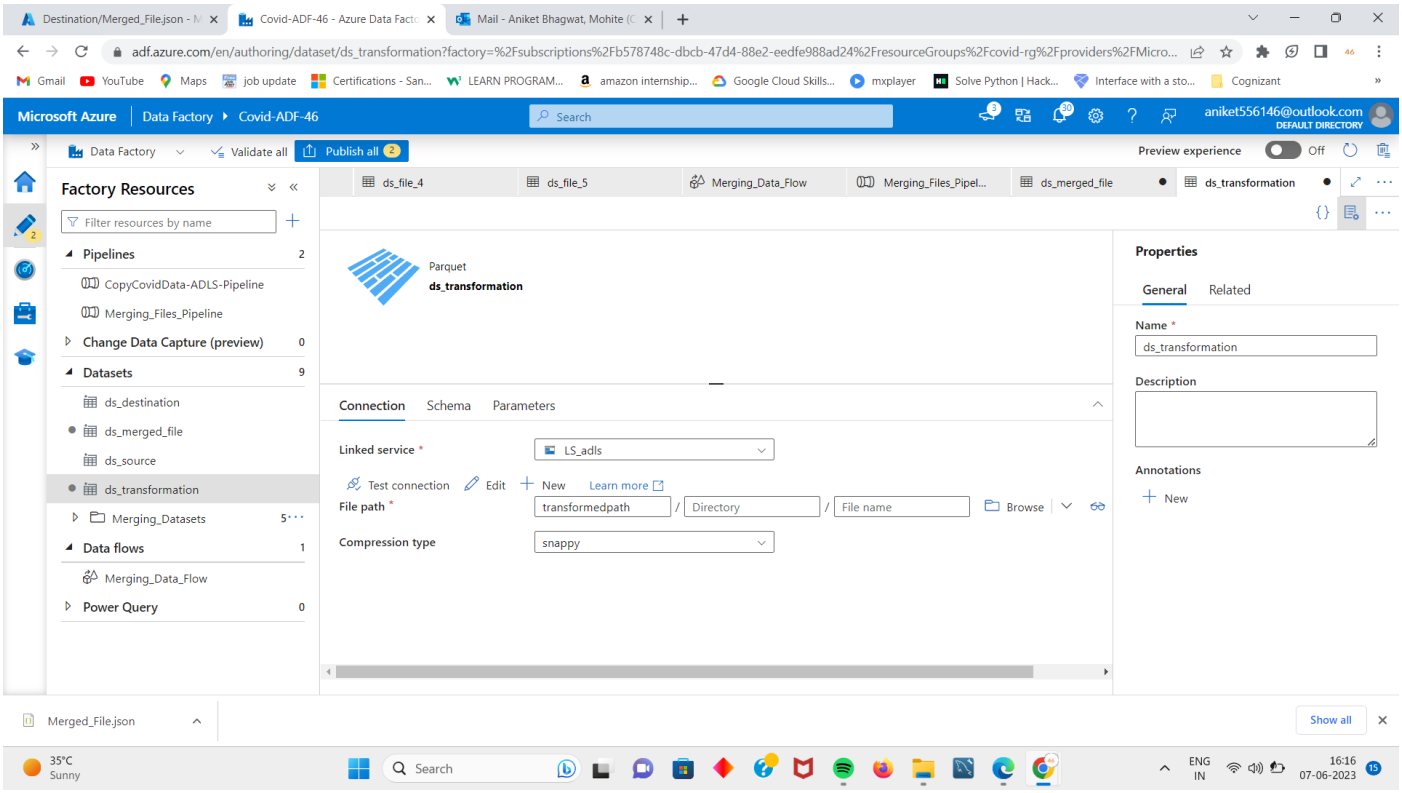


3. Find Ranking of Continent which have max daily death data

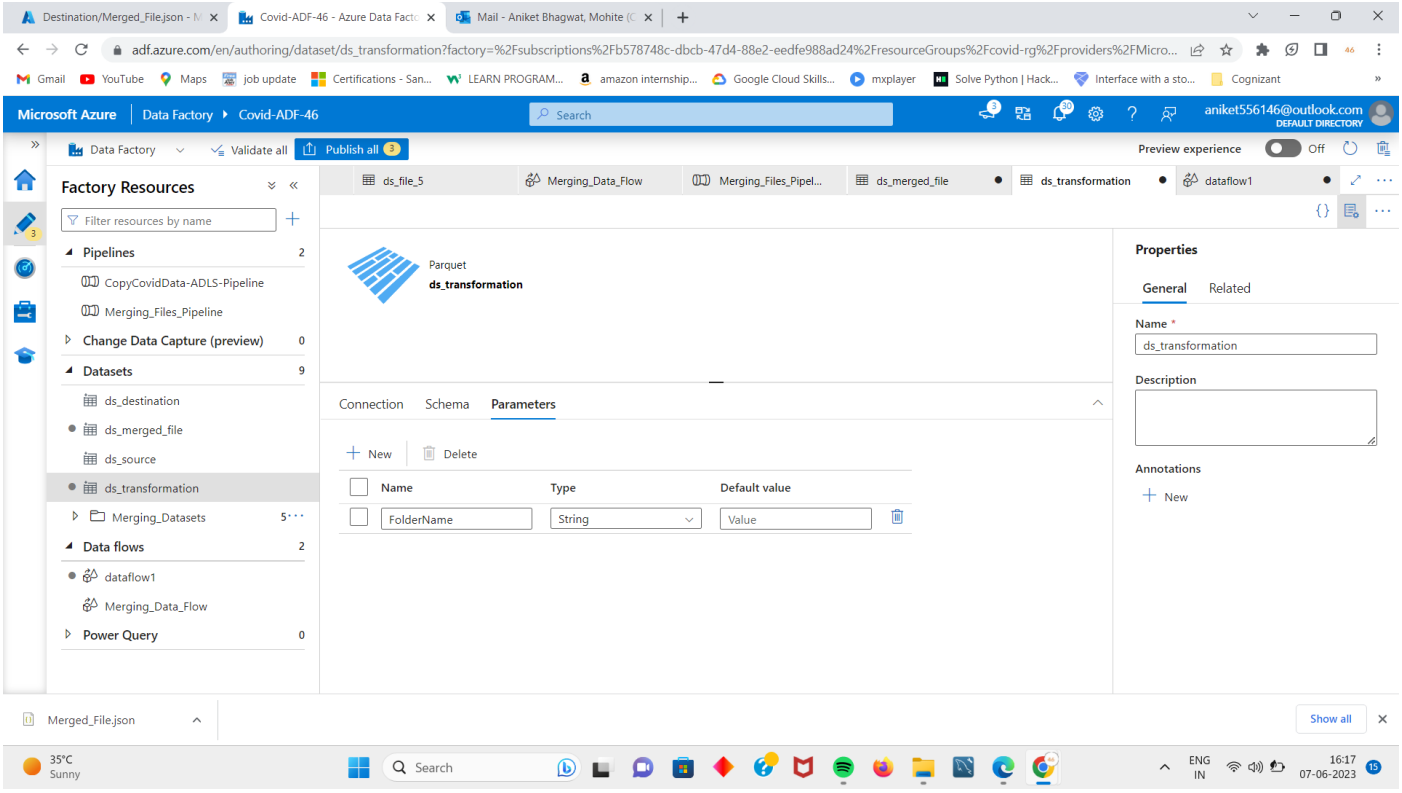
1) create Dataset for merged file

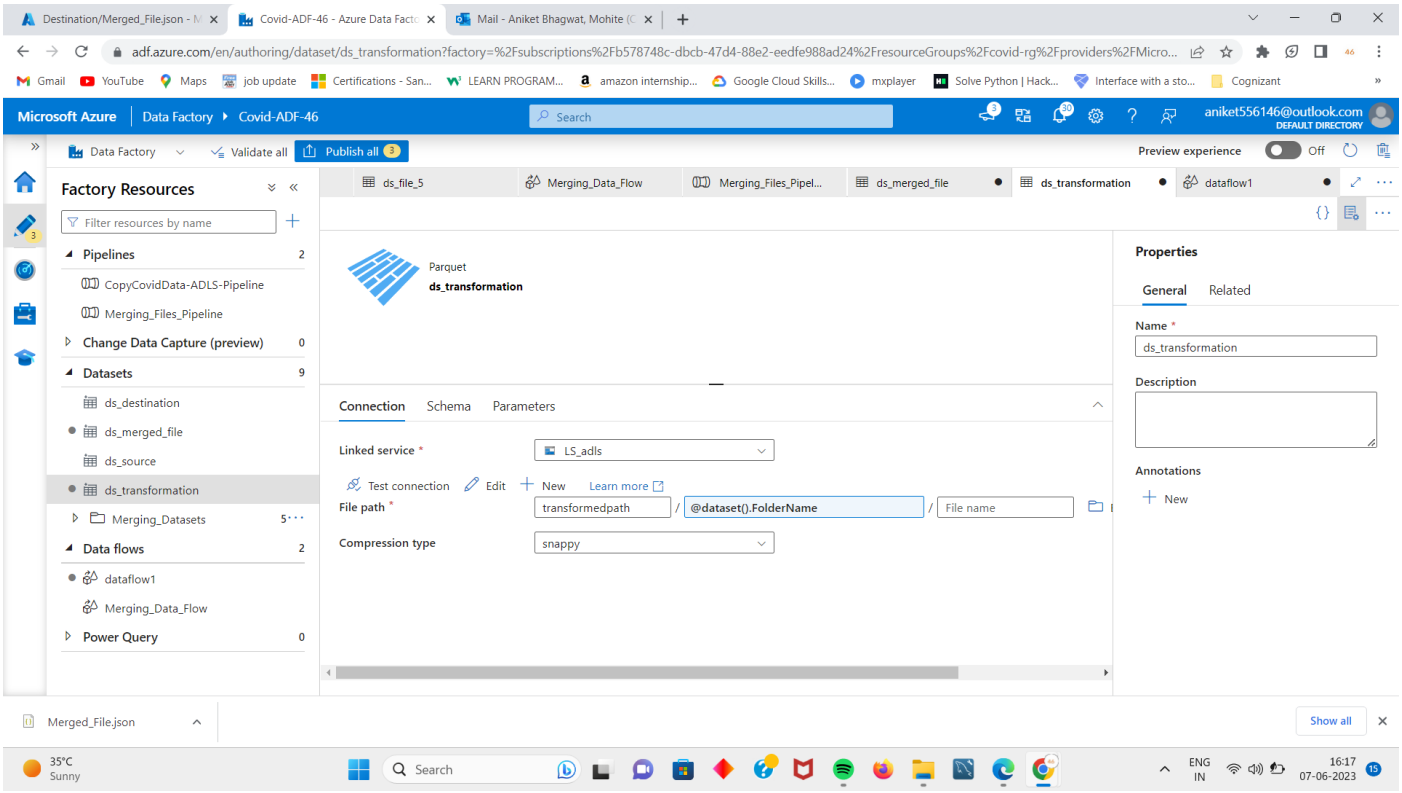


2) Create Dataset for Transformpath container to store Month wise files in Parquet Format.

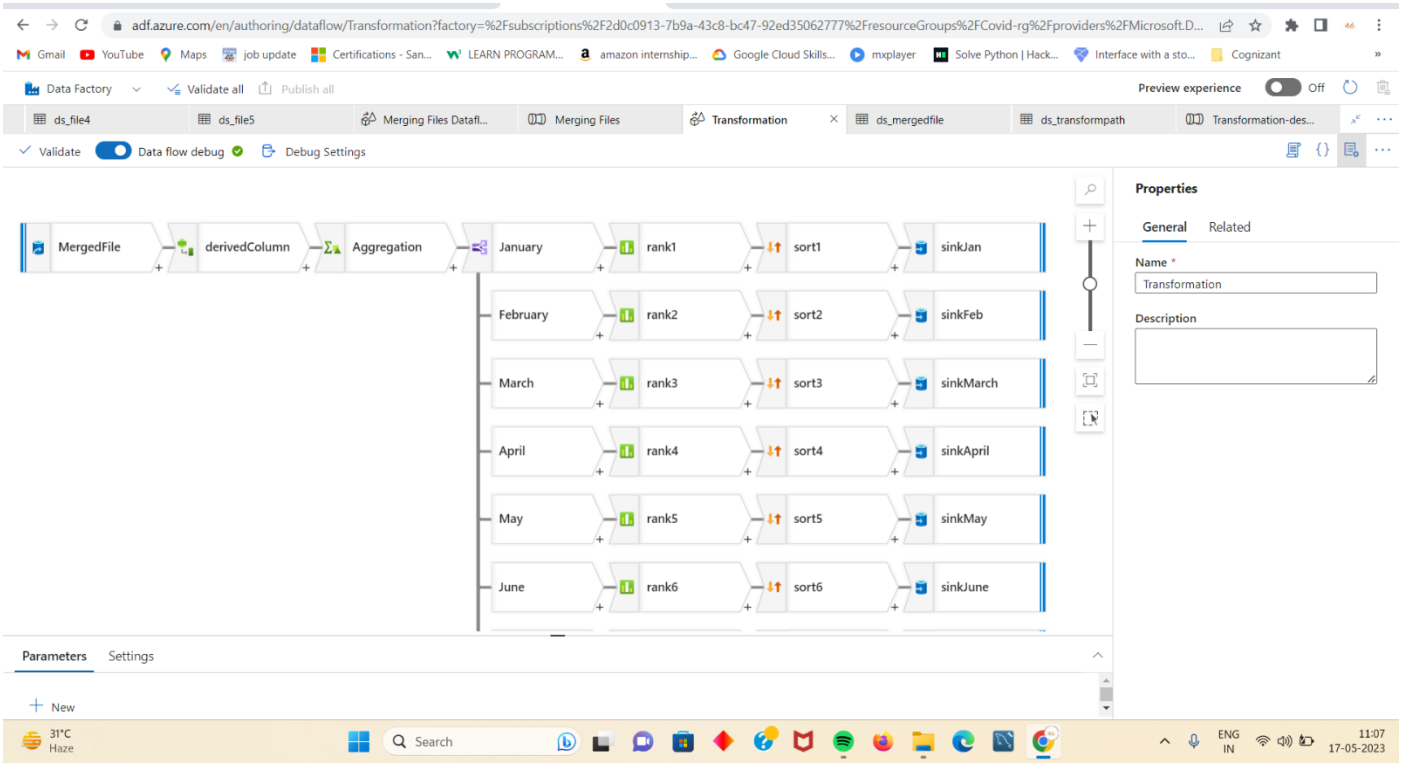


3) To give dynamic folder name, create parameter and add dynamic path in connection.

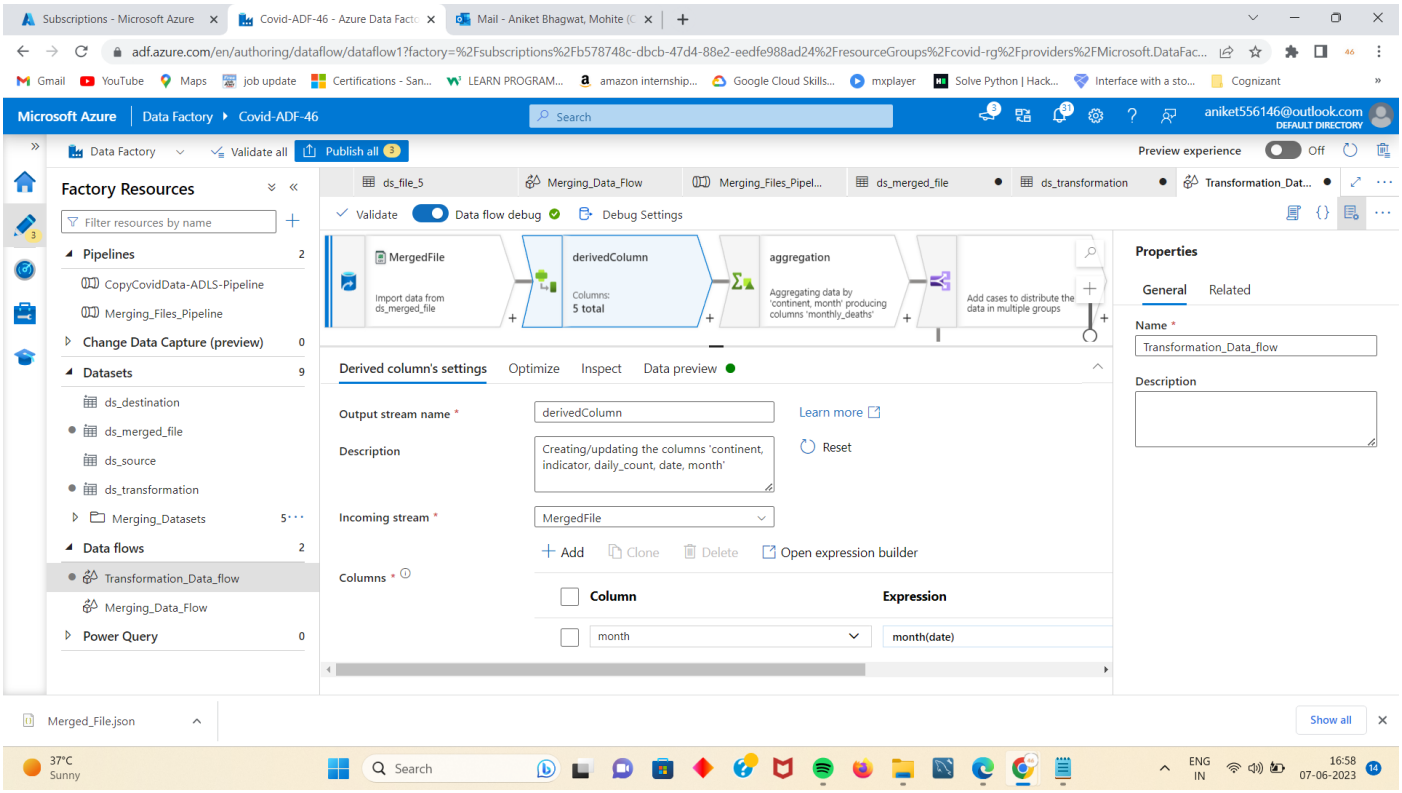




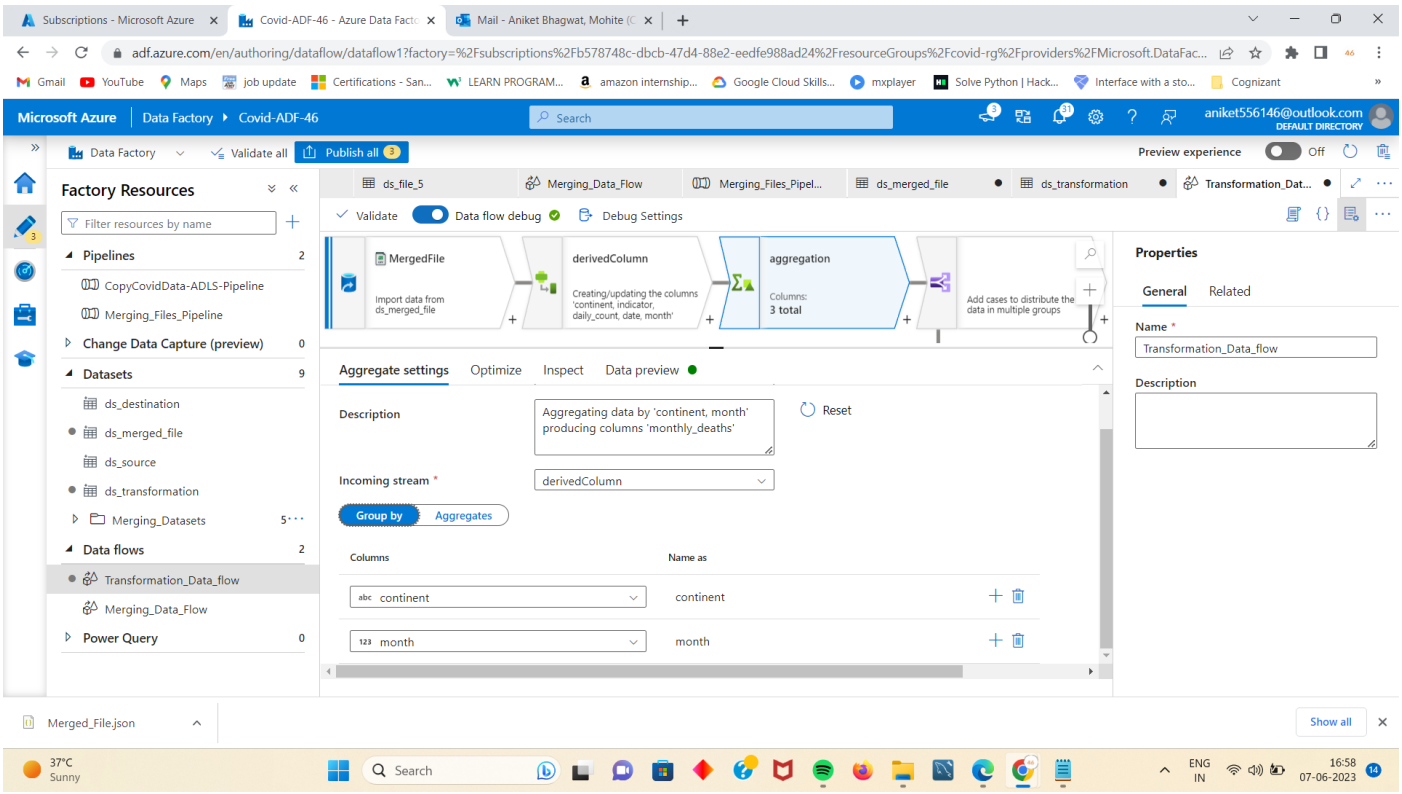
4) Create Dataflow for Transformation.
    1) Select Source for Dataflow.



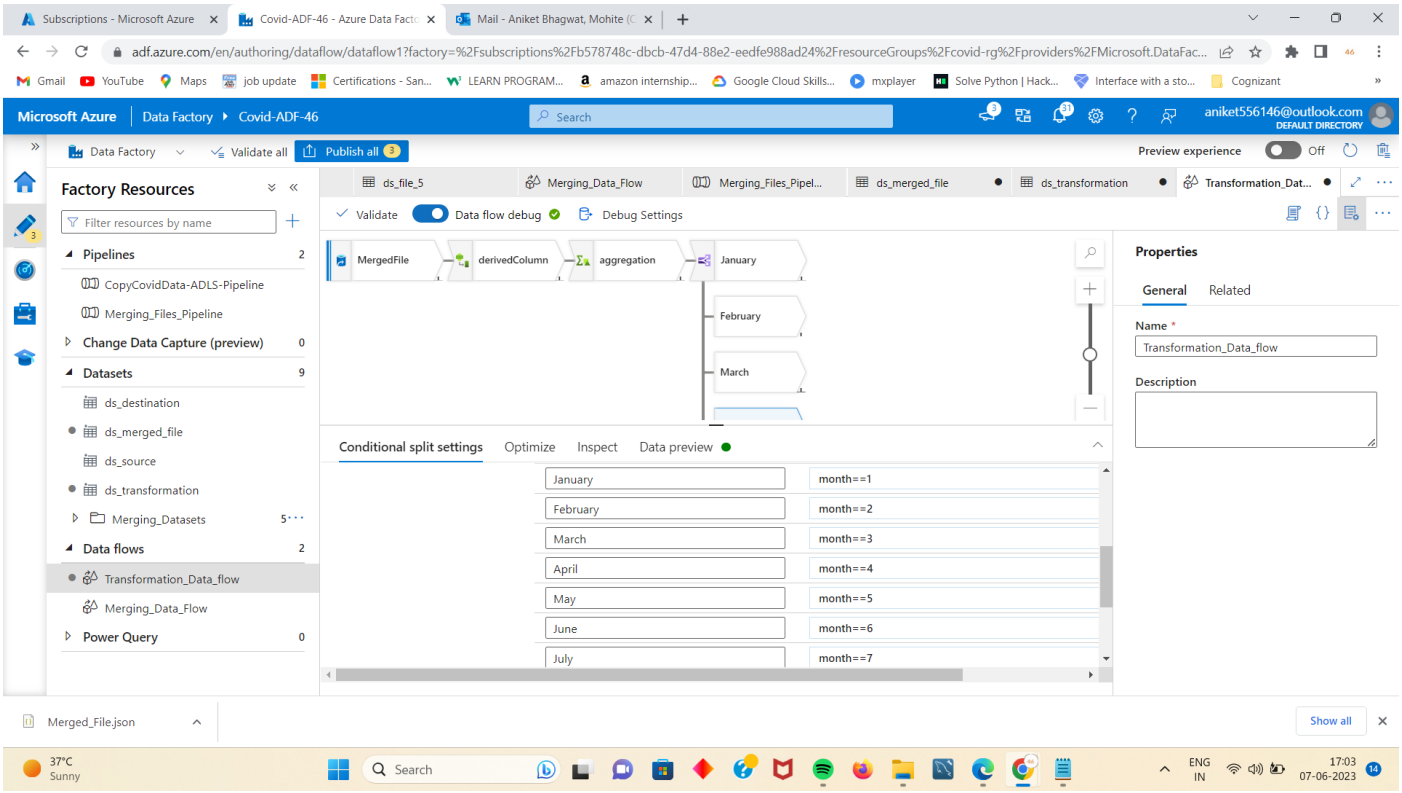
2) Use Derived Column activity to extract date in new column for further process.



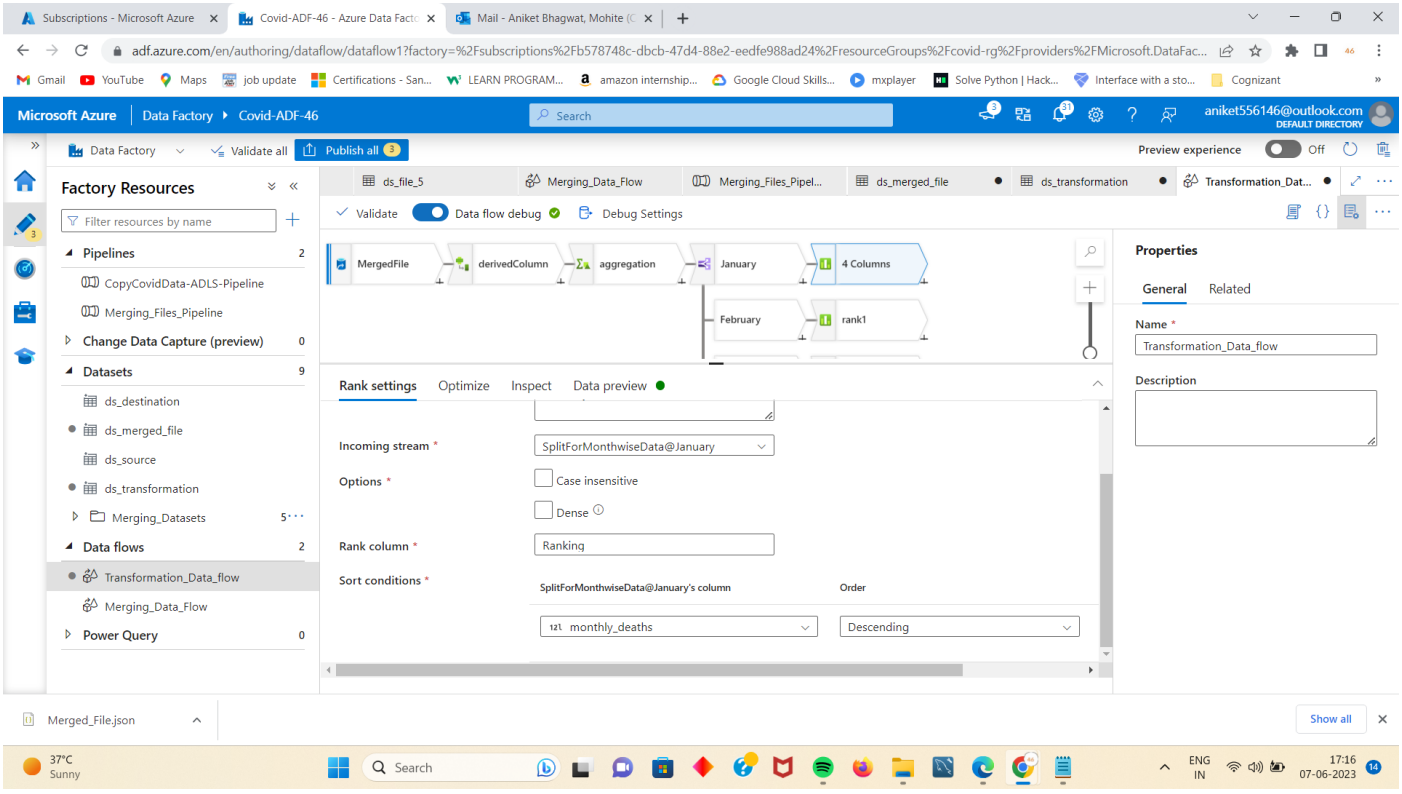
3) Use Aggregation activity to get Maximum deaths in month.



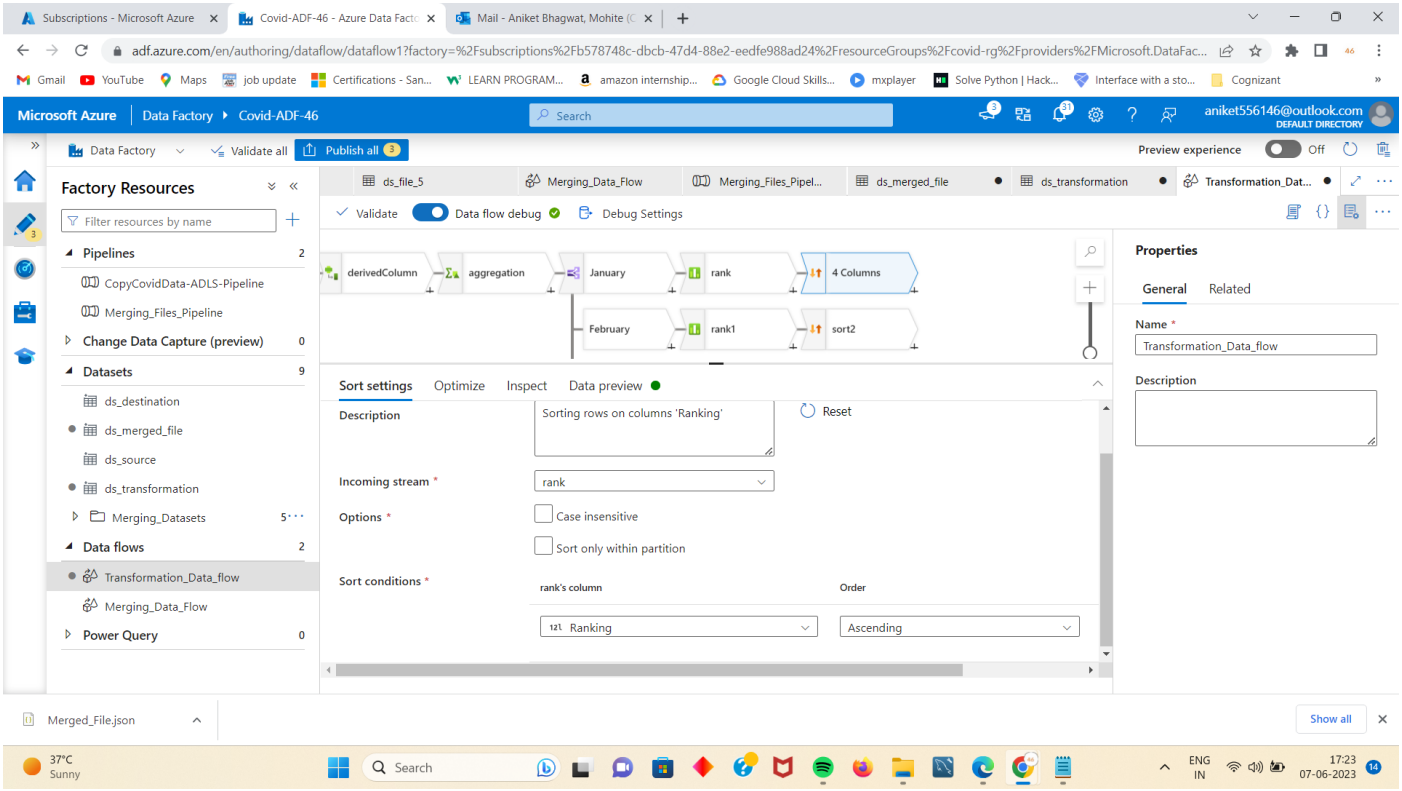
4) Use Conditional Split to get data month wise.



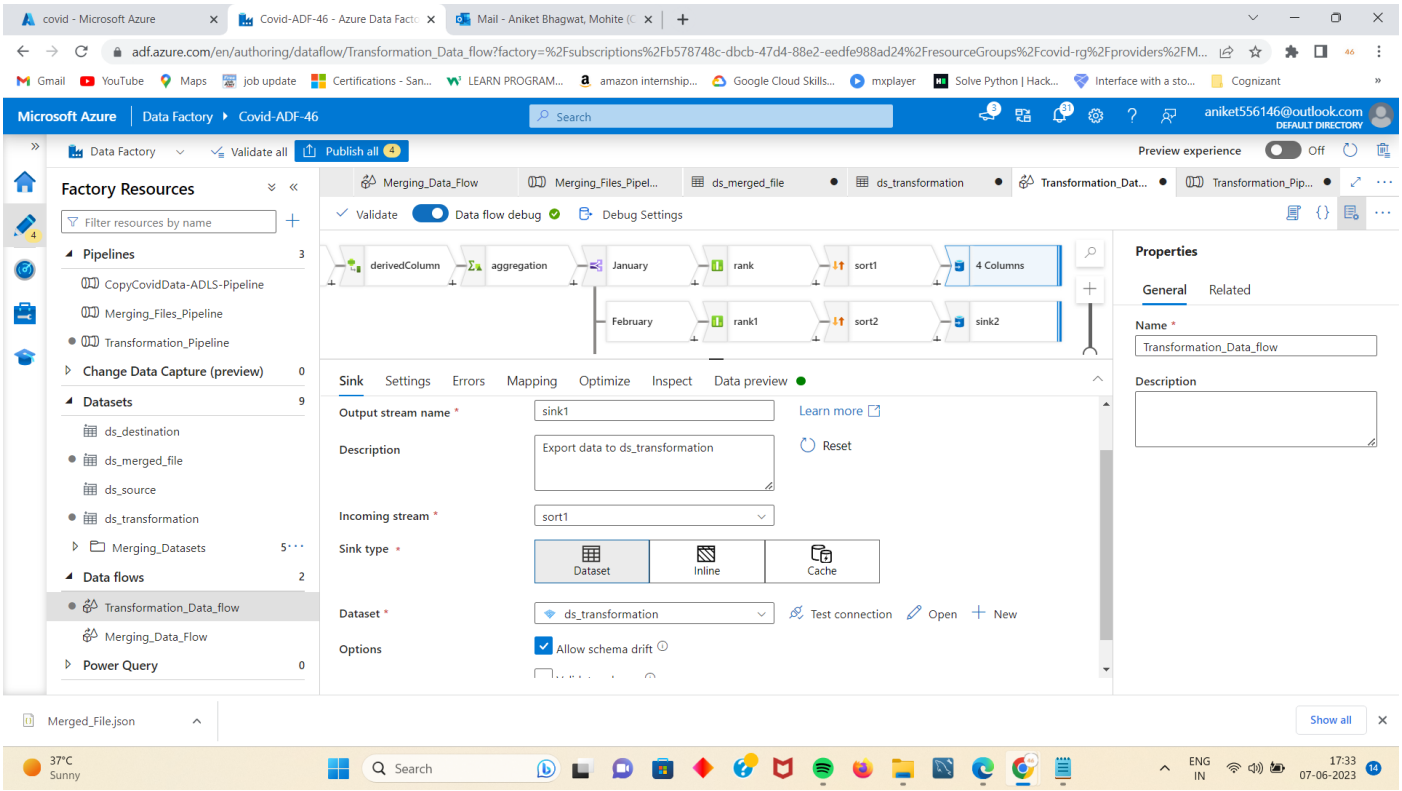
5) Use Rank activity to give ranks to each continent.

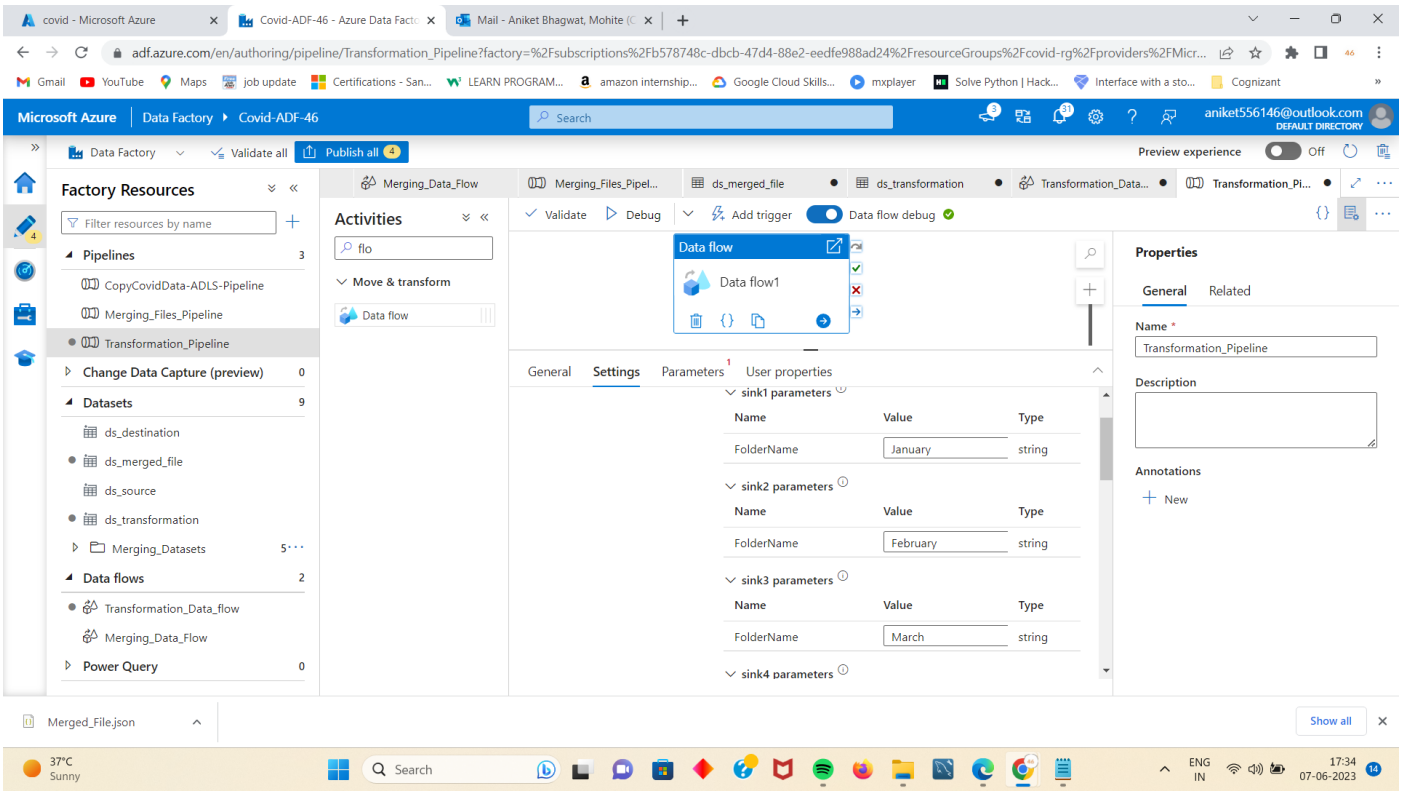


6) Use sort to get ranks sequentially in output file.

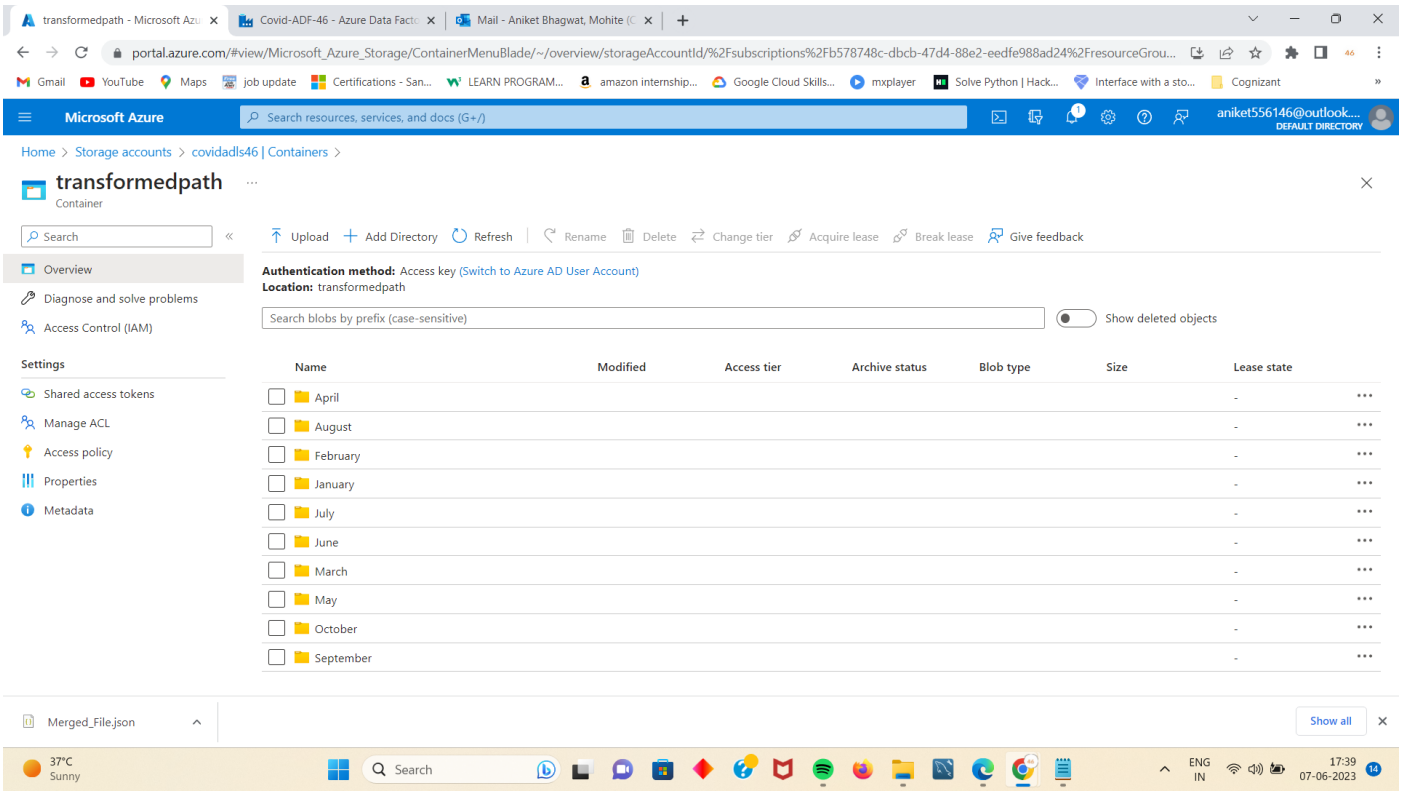


7) Use Sink to get output in transformedpath container, select dataset created for it. And in setting select output to single file.

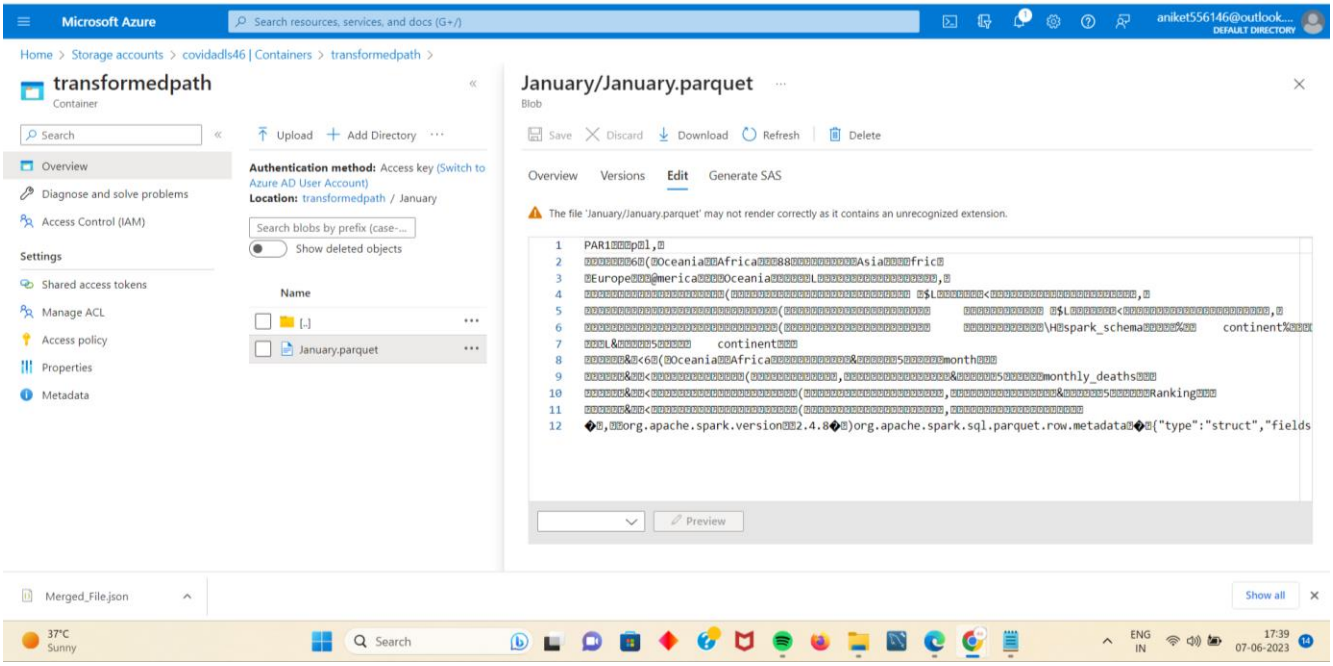




Output of this project:



January Output In Parquet Format:



January Output:

Parquet File Viewer For Web

Your data will NOT be uploaded anywhere! All operations happen locally in your web browser!

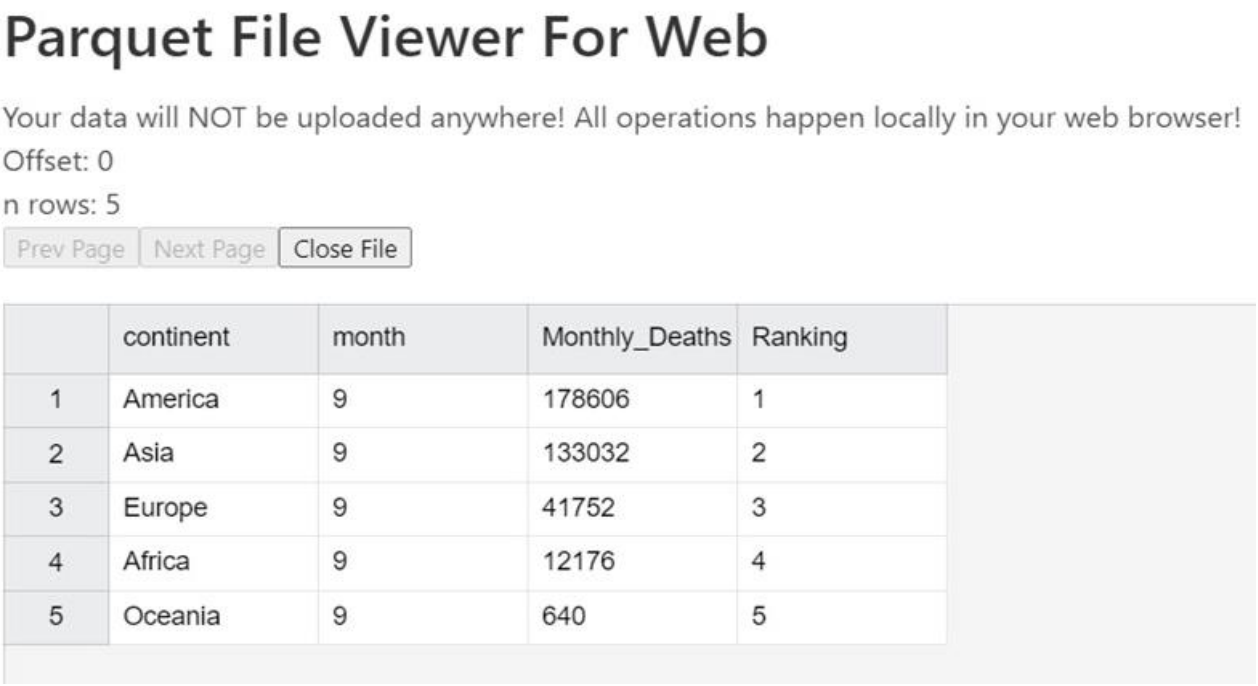
Offset: 0

n rows: 5

Prev Page | Next Page | Close File

	continent	month	monthly_deaths	Ranking
1	Asia	1	426	1
2	Africa	1	0	2
3	Europe	1	0	2
4	America	1	0	2
5	Oceania	1	0	2

September output:



	continent	month	Monthly_Deaths	Ranking
1	America	9	178606	1
2	Asia	9	133032	2
3	Europe	9	41752	3
4	Africa	9	12176	4
5	Oceania	9	640	5

Conclusion:

Upon completion of the Covid use case project I got required output in parquet file. The objective of this project is achieved. I learn to build a real-world data pipeline in Azure Data Factory (ADF) and use different activities in it.