

# **Streamlining Analytics Pipeline with Azure Data Services**

## **CLOUD COMPUTING – PROJECT**



Hrushika Papani (16354873)

Viswanth Tammana (16352538)

Chandra Haas (16355595)

Dheeraj Golla (16348914)

10<sup>th</sup> May, 2024

### **Roles and Contribution:**

- Hrushika had done the data transformation by creating dataflows in the azure data factory. Performed different transformation activities like Source, filter, select, pivot, lookup and sink, then created ADF pipeline for the above transformation.
- Viswa, ingested the data from HTTP(ECDC website).Created the control flow activities and the linked service parameters. Monitored pipelines and integration runtimes, created alerts for failures or successes, and visualized metrics using plots and Kusto Query Language in Azure.
- ChandraHaas had ingested the population data to the datalake using data factory. Performed copy activity, created linked services and datasets and pipeline. Worked on control flow activities like Validation, Get Metadata, If Condition, Web, Delete, Fail. Created trigger to ingest the data to the data lake.
- Dheeraj manages the integration of transformed data, including cases and deaths, hospital admissions, and testing data, into an Azure SQL database. He oversees the setup of SQL tables, data pipelines, and copy activities to enable efficient reporting and analysis.

### **Introduction and Purpose:**

- Real-world data engineering workflow on Microsoft Azure, specifically targeting COVID-19 data analysis due to the pandemic's extensive data generation.
- Address the need for efficient COVID-19 data handling by enabling organizations to collect, integrate, and analyze data effectively through an end-to-end data pipeline on Azure
- Empower data engineers with scalable pipeline development skills, integrating CI/CD for automation and reliability, aiding real-time pandemic response decisions.

## **Background and Motive:**

- The COVID-19 pandemic has inundated organizations with vast amounts of data from diverse sources, necessitating robust data engineering solutions to efficiently manage and analyze this information.
- Leveraging Microsoft Azure's suite of tools enables the development of a comprehensive data pipeline for COVID-19 data analysis, addressing the need for seamless data transfer, advanced analytics, and automation.
- By building an end-to-end data pipeline utilizing Azure Data Factory, Azure Data Lake Storage, and Databricks, the project aims to empower organizations with timely insights into COVID-19 data, facilitating informed decision-making for response and recovery strategies.
- This project serves to equip data engineers with the necessary skills to construct and manage scalable data pipelines tailored for COVID-19 data analysis, ultimately enabling organizations to leverage data-driven approaches for more effective pandemic response efforts.

## **Project Implementation:**

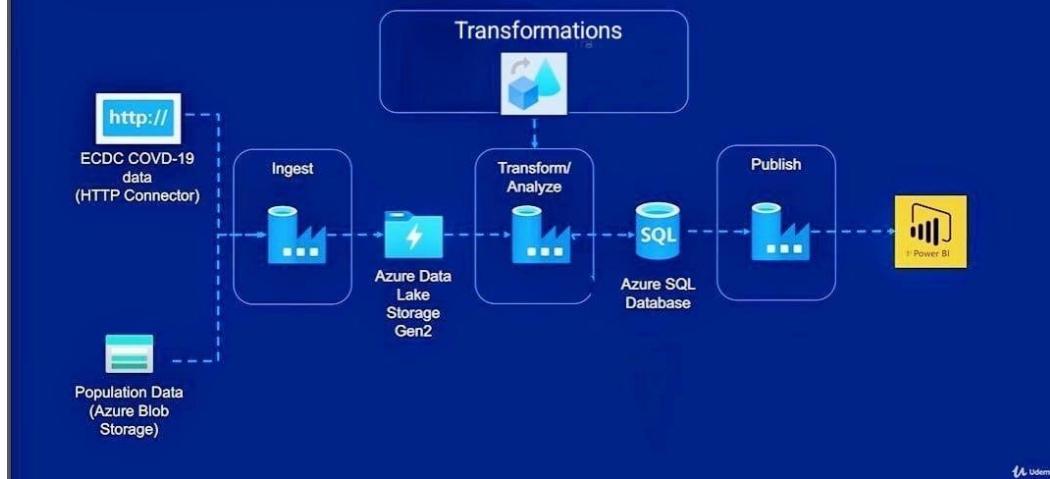
The data comprises population data ranging period from 2008 to 2019.

Data is based on country code and age (0-max).

**Step 1 :** We need to setup the environment

Solution Architecture

# Solution Architecture



It includes Creating an Azure Data factory

In the home we could see “create a resource” click on it and create data factory

## Create Data Factory

**⚠ Changes on this step may reset later selections you have made. Review all options prior to deployment.**

**Basics**   Git configuration   Networking   Advanced   Tags   Review + create

One-click to create data factory with sample pipeline and datasets. [Try it](#)

### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \* ⓘ

Azure subscription 1

Resource group \* ⓘ

(New) covid-reporting-rg4

[Create new](#)

### Instance details

Name \* ⓘ

covid-reporting-adfgroup4

Region \* ⓘ

UK South

Version \* ⓘ

V2

**covid-reporting-adfgroup4**

Resource group (move) : covid-reporting-rg4  
Status : Succeeded  
Location : UK South  
Subscription (move) : Azure subscription\_1  
Subscription ID : 3da58d65-e067-4baa-8e6c-e59781aed060

Type : Data factory (V2)  
Getting started : Quick start

Azure Data Factory Studio

Launch studio

## Step 2: Also create azure storage account:

**covidreportingsagroup4**

Resource group (move) : covid-reporting-rg4  
Location : uksouth  
Subscription (move) : Azure subscription\_1  
Subscription ID : 3da58d65-e067-4baa-8e6c-e59781aed060  
Disk state : Available  
Created : 4/18/2024, 7:55:29 PM

Tags (edit) : Add tags

Properties Monitoring Capabilities (7) Recommendations (0) Tutorials Tools + SDKs

Blob service		Security	
Hierarchical namespace	Disabled	Require secure transfer for REST API operations	Enabled
Default access tier	Hot	Storage account key access	Enabled
Blob anonymous access	Disabled	Minimum TLS version	Version 1.2
Blob soft delete	Enabled (7 days)	Infrastructure encryption	Disabled
Container soft delete	Enabled (7 days)		
Versioning	Disabled		
Change feed	Disabled		

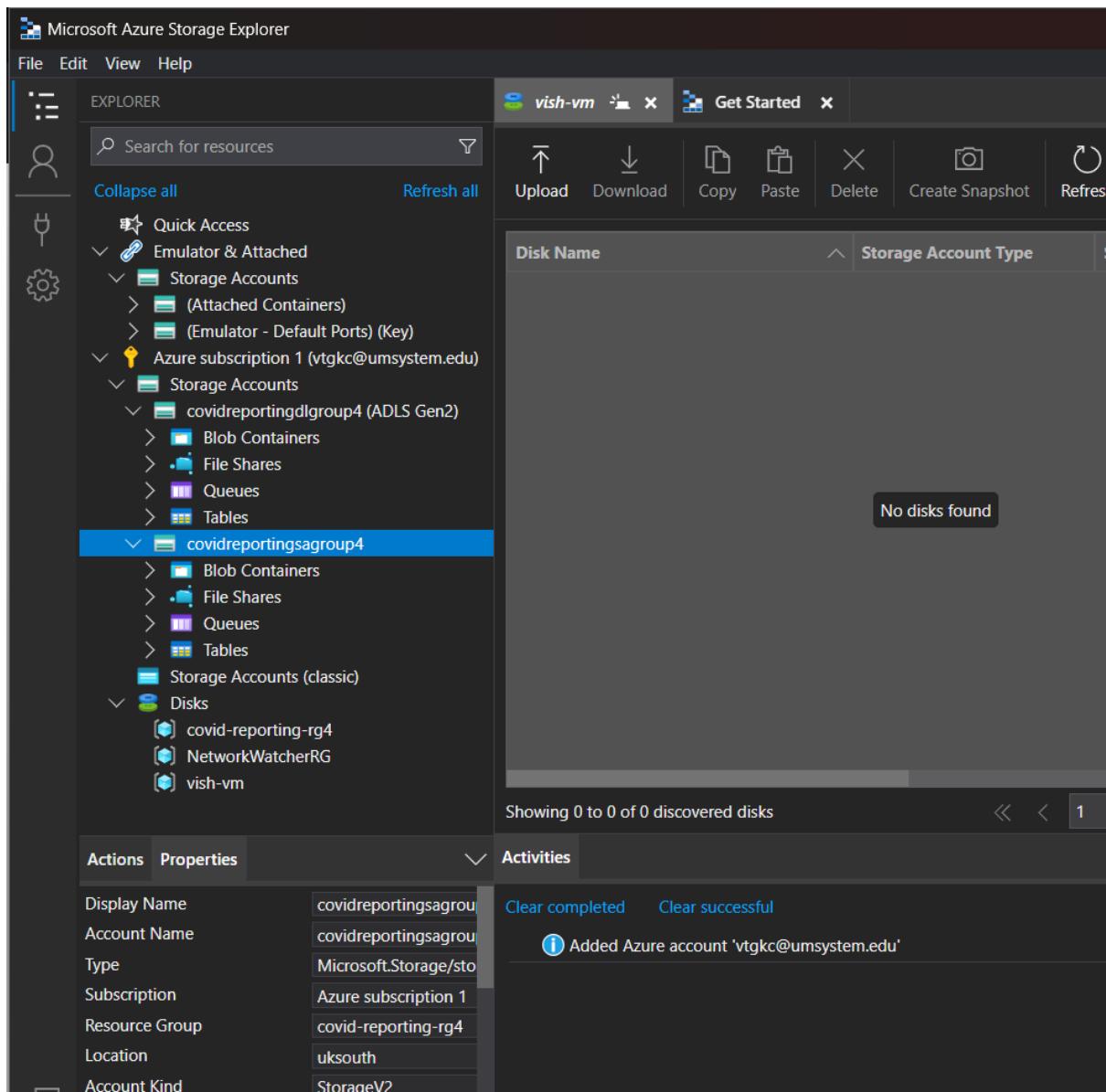
Networking

## Step 3: Installing Azure Storage Explorer

From the below link we can install the azure storage explorer.

<https://azure.microsoft.com/en-us/products/storage/storage-explorer>

After successful installation launch the explorer.



#### Step 4: Now creating Azure data lake storage:

The screenshot shows the Azure portal's storage account configuration page for 'covidreportingdlgroup4'. The left sidebar lists 'Overview', 'Activity log', 'Tags', 'Diagnose and solve problems', 'Access Control (IAM)', 'Data migration', 'Events', 'Storage browser', 'Data storage' (with 'Containers', 'File shares', 'Queues', 'Tables'), and 'Security + networking' (with 'Networking'). The main content area shows the account's properties. Under 'Properties', the 'Data Lake Storage' section includes 'Hierarchical namespace' (Enabled), 'Default access tier' (Hot), 'Blob anonymous access' (Enabled), 'Blob soft delete' (Enabled (7 days)), 'Container soft delete' (Enabled (7 days)), and 'Versioning' (Disabled). The 'Security' section includes 'Require secure transfer for REST API operations' (Enabled), 'Storage account key access' (Enabled), 'Minimum TLS version' (Version 1.2), and 'Infrastructure encryption' (Disabled). The 'Networking' section is currently empty.

## Creating Azure SQL Database

While doing so it allows us to configure Username and password.

Username: admingroup4

Pwd: CCgroup4

Configure ...

Feedback

### Service and compute tier

Select from the available tiers based on the needs of your workload. The vCore model provides a wide range of configuration controls and offers Hyperscale and Serverless to automatically scale your database based on your workload needs. Alternately, the DTU model provides set price/performance packages to choose from for easy configuration. [Learn more](#)

SQL Database Hyperscale: Low price, high scalability, and best feature set. [Learn more](#)

Service tier

Basic (For less demanding workloads)

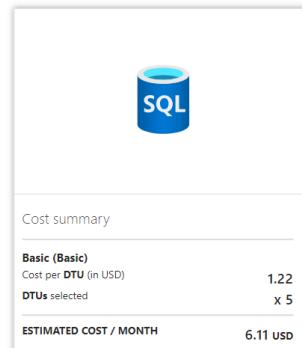
[Compare service tiers](#)

DTUs [Compare DTU options](#)

**5 (Basic)**

Data max size (GB)

2



**covid-dbgroup4 (covid-srv4/covid-dbgroup4)**

Overview

Resource group (move) : covid-reporting-rg4

Status : Online

Location : UK South

Subscription (move) : Azure subscription 1

Subscription ID : 3da58d65-e067-4baa-8e6c-e59781aed060

Tags (edit) : Add tags

Getting started

Start working with your database

Configure access

Connect to application

Start developing

Installed Azure Data Studio and connected database:

**Step 5:** Now we need to link the database to Azure Date Studio

	name	object_id	principal_id	schema_id	parent_object_id
1	sp_drop_trusted_assembly	-1072667163	NULL	4	0
2	TABLE_PRIVILEGES	-1072372588	NULL	3	0
3	sp_help_spatial_geometry_index_xml	-1068265529	NULL	4	0
4	sp_get_migration_vlf_state	-1064940705	NULL	4	0
5	sp_bindsession	-1064199433	NULL	4	0
6	sp_xa_prepare_ex	-1063740264	NULL	4	0
7	dm_os_hosts	-1061705188	NULL	4	0
8	sp_xtp_force_gc	-1060571671	NULL	4	0
9	sp_rename	-1058549068	NULL	4	0
10	sp_cdc_disable_table	-1057038550	NULL	4	0
11	dm_os_memory_brokers	-1055124494	NULL	4	0
12	sp_process_memory_leak_record	-1054537646	NULL	4	0
13	sp_procedure_params_100_rowset2	-1052345683	NULL	4	0
14	dm_db_stats_properties_internal	-1052007962	NULL	4	0
15	sp_autostats	-1051705964	NULL	4	0

**Step 6:** Ingest population data (blob storage) to datalake using Azure Data factory.

Copy activity, Create Linked services, datasets, Execute Pipeline

Control flow activities (Validation, if condition, web, get metadata, delete)

Triggers (to automate the workflow for the pipelines)

## COPY ACTIVITY

Ingest population data (blob storage) to datalake using Azure Data factory.

## 1. Add container in the blob storage > storage account

The screenshot shows the Microsoft Azure Storage browser interface for the 'covidreportingsgroup4' storage account. On the left, the navigation menu includes 'Overview', 'Activity log', 'Tags', 'Diagnose and solve problems', 'Access Control (IAM)', 'Data migration', 'Events', 'Storage browser' (which is selected), and 'Storage Mover'. Under 'Data storage', there are 'Containers', 'File shares', 'Queues', and 'Tables'. The main pane displays 'Blob containers' with a single item, '\$logs'. The container details show it was last modified on 4/18/2024 at 7:56:00 PM, has an 'Anonymous access level' of 'Private', and is in an 'Available' lease state.

## 2. Create container

The screenshot shows the Microsoft Azure Storage browser interface for the 'covidreportingsgroup4' storage account. The 'New container' dialog is open on the right, prompting for a container name ('population') and anonymous access level ('Private (no anonymous access)'). A note indicates that the access level is set to private because anonymous access is disabled on the storage account. The main pane shows the existing '\$logs' container.

## 3. Upload files

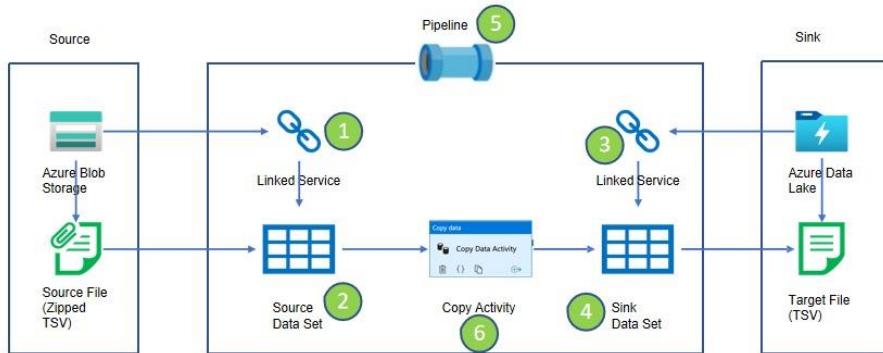
The screenshot shows the Microsoft Azure Storage browser interface for the 'covidreportingsgroup4' storage account. The 'New container' dialog is open on the right, prompting for a container name ('population') and anonymous access level ('Private (no anonymous access)'). A note indicates that the access level is set to private because anonymous access is disabled on the storage account. The main pane shows the existing '\$logs' container.

The screenshot shows the Azure Storage browser interface for a storage account named 'covidreportingsagroup4'. The left sidebar includes links for Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, Storage browser, Storage Mover, Containers, File shares, and Queues. The main area displays a list of blobs under the 'population' container. A single blob, 'population\_by\_age (3).tsv.gz', is selected. The details pane shows the blob's name, last modified (4/18/2024, 10:11:08 PM), access tier (Hot (Inferred)), blob type (Block blob), size (5.05 KiB), and lease state (Available).

#### 4. Create a raw container in the data lake.

Naming standards

## Copy Activity



Linked services and datasets

Create linked services for storage account

## Create linked services for datalake storage gen 2

## Create a corresponding dataset:

Microsoft Azure | Data Factory > covid-reporting-adfgroup4

Validate all Publish all

Search factory and documentation

Factory Resources

Filter resources by name

- Pipelines 0
- Change Data Capture (preview) 0
- Datasets 2
  - ds\_population\_raw\_gz
  - ds\_population\_raw\_tsv
- Data flow 0
- Power Query 0

ds\_population\_raw\_gz x ds\_population\_raw\_tsv

DelimitedText  
ds\_population\_raw\_gz

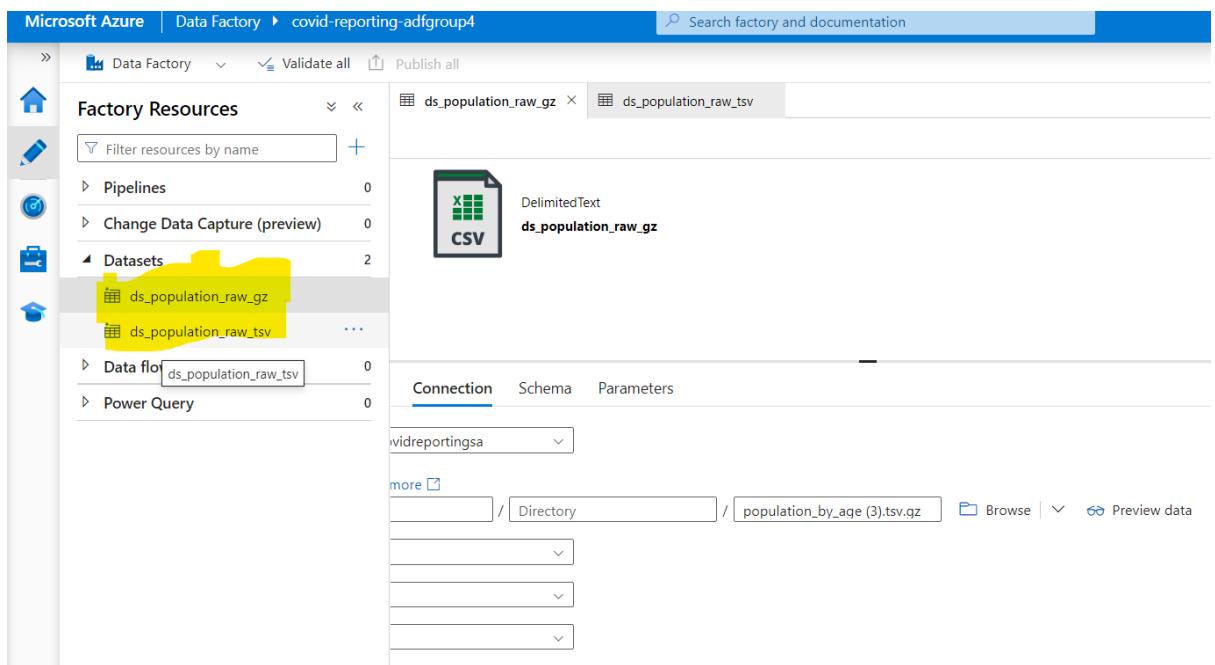
CSV

Connection Schema Parameters

avidreportingsa

more / Directory / population\_by\_age (3).tsv.gz

Browse | Preview data



Microsoft Azure | Data Factory > covid-reporting-adfgroup4

Validate all Publish all

Search factory and documentation

Factory Resources

Filter resources by name

- Pipelines 0
- Change Data Capture (preview) 0
- Datasets 2
  - ds\_population\_raw\_gz
  - ds\_population\_raw\_tsv
- Data flow 0
- Power Query 0

ds\_population\_raw\_gz x ds\_population\_raw\_tsv

DelimitedText  
ds\_population\_raw\_gz

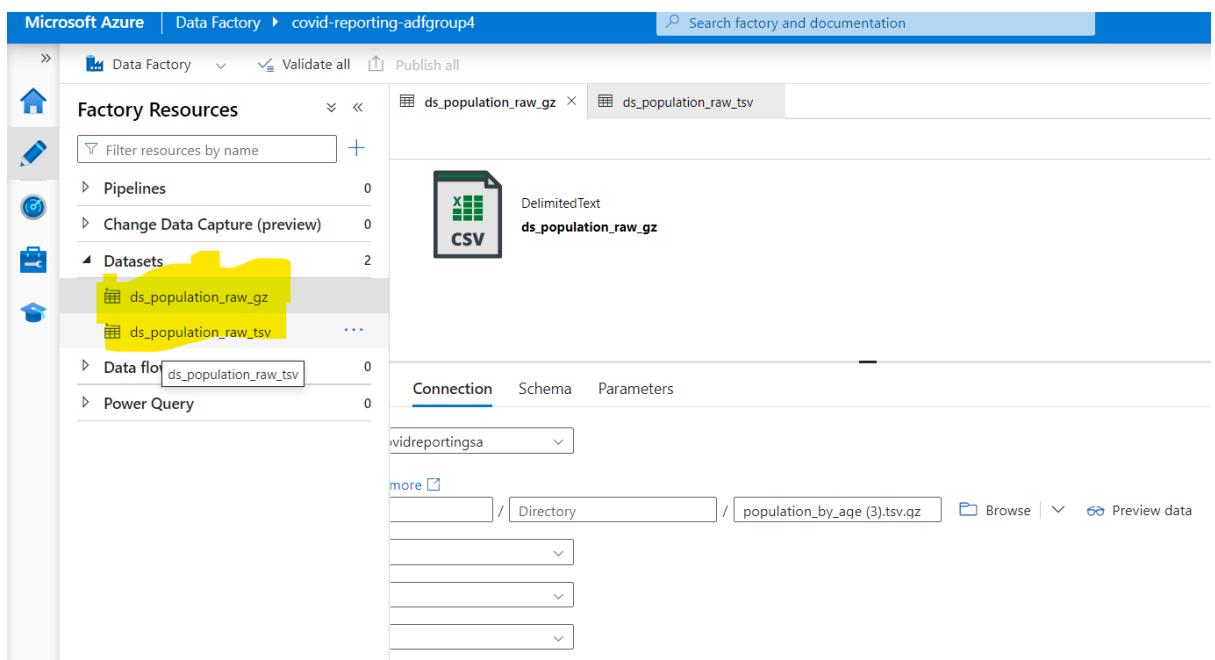
CSV

Connection Schema Parameters

avidreportingsa

more / Directory / population\_by\_age (3).tsv.gz

Browse | Preview data



Creating pipeline ADF

The screenshot shows the Azure Data Factory pipeline editor. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. A pipeline named 'pl\_ingerst\_population\_data' is selected. The main workspace displays a 'Copy data' activity under the 'Activities' tab. The 'Output' tab is selected, showing a table of pipeline run details:

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity
Copy Population Data	Succeeded	Copy data	4/18/2024, 10:55:15 PM	15s	AutoResolveIntegration		5af67d4

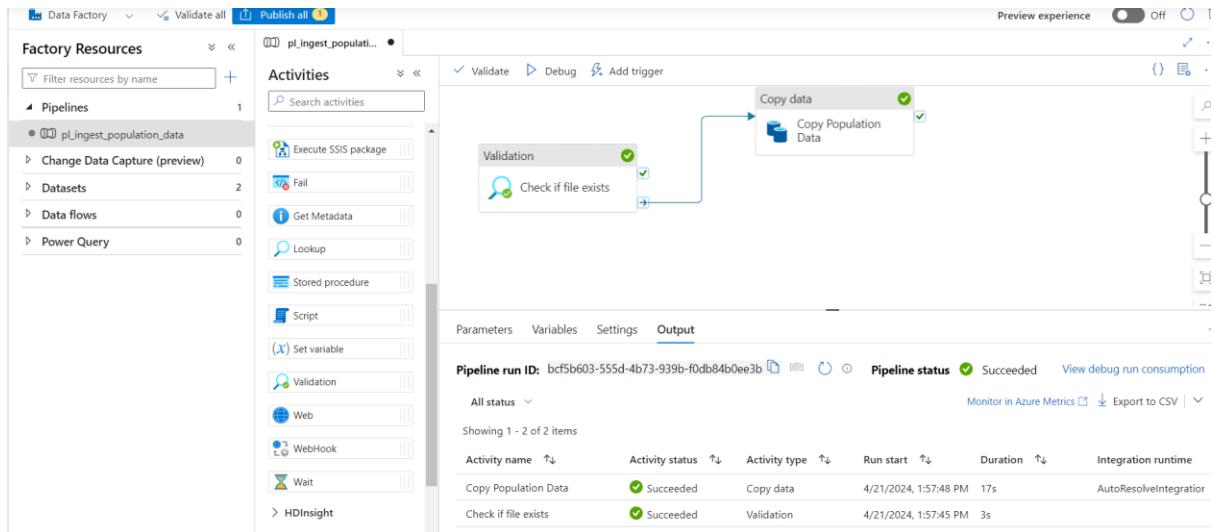
The 'Properties' pane on the right shows the pipeline's name as 'pl\_ingerst\_population\_data'. The 'Output' section of the pipeline run details table is also visible.

The screenshot shows the 'Details' page for a pipeline run. The left sidebar lists Pipelines, Datasets, Data flows, and Power Query. The main area shows the 'Copy Population Data' activity with a status of 'Succeeded'. It details the transfer from 'Azure Blob Storage' to 'Azure Data Lake Storage Gen2', both in 'Region: UK South'. Key metrics include:

- Azure Blob Storage:**
  - Data read: 5.986 KB
  - Files read: 1
  - Peak connections: 1
- Azure Data Lake Storage Gen2:**
  - Data written: 26.07 KB
  - Files written: 1
  - Peak connections: 1
- Copy duration:** 00:00:13
- Throughput:** 2.993 KB/s
- Transfer Details:**
  - Start time: 4/18/2024, 10:55:16 PM
  - Used DIUs: 4
  - Used parallel copies: 1
  - Duration: 00:00:13
  - Working duration: 00:00:08
  - Total duration: 00:00:08
  - Transfer steps: Listing source (00:00:00), Reading from source (00:00:00), Writing to sink (00:00:00)
- Data consistency verification:** Unsupported

At the bottom, there is a satisfaction survey with a 5-star rating icon and the question: "How satisfied or dissatisfied are you with the performance of this copy activity?"

## Control Flow Activities 1 -Validation



## Control Flow Activities 2 – Get Metadata, If Condition, Web Activities

### Get File Metadata copy

```
{
    "size": 5986,
    "exists": true,
    "columnCount": 13,
    "effectiveIntegrationRuntime": "AutoResolveIntegrationRuntime (UK South)",
    "executionDuration": 0,
    "durationInQueue": {
        "integrationRuntimeQueue": 1
    },
    "billingReference": {
        "activityType": "PipelineActivity",
        "billableDuration": [

```

```
        {
            "meterType": "AzureIR",
            "duration": 0.01666666666666666,
            "unit": "Hours"
        }
    ]
}

}
```

## Output

```
 Copy to clipboard
```

```
{
    "size": 5986,
    "exists": true,
    "columnCount": 13,
    "effectiveIntegrationRuntime": "AutoResolveIntegrationRuntime (UK South)",
    "executionDuration": 0,
    "durationInQueue": {
        "integrationRuntimeQueue": 1
    },
    "billingReference": {
        "activityType": "PipelineActivity",
        "billableDuration": [
            {
                "meterType": "AzureIR",
                "duration": 0.01666666666666666,
                "unit": "Hours"
            }
        ]
    }
}
```

If condition

Factory Resources

Pipelines

- pl\_ingest\_population\_data

Activities

- Move and transform
- Synapse
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDIInsight
- Iteration & conditionals
- Filter
- ForEach
- If Condition
- Switch

Validation

Check if file exists → Get Metadata → If Condition → Copy Population Data

If Condition

If Column Count Matches

True: Copy Population Data

False:

Pipeline run ID: 330c8b9b-4cf4-4c1d-8514-2ac4f92b0260

Pipeline status: Succeeded

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime
Copy Population Data	Succeeded	Copy data	4/21/2024, 2:17:49 PM	15s	AutoResolveInteg
If Column Count Matches	Succeeded	If Condition	4/21/2024, 2:17:49 PM	17s	
Get File Metadata	Succeeded	Get Metadata	4/21/2024, 2:17:46 PM	2s	AutoResolveInteg
Check if file exists	Succeeded	Validation	4/21/2024, 2:17:40 PM	6s	

Testing the condition with 17, it returns false

Validate all

Activities

Validation

Check if file exists

Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@equals(activity('Get File Metadata').output.columnCount, 17)
```

Clear contents

Activity outputs Parameters System variables Functions Variables

Count did not match, and send email failed since the request is not possible with url provided.

Factory Resources

Pipelines

- pl\_ingest\_population\_data

Activities

- Move and transform
- Synapse
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- Append variable
- Delete
- Execute Pipeline
- Execute SSIS package
- Fail
- Get Metadata

Validation

Check if file exists → Get Metadata → If Condition → Copy Population Data

If Condition

If Column Count Matches

True: Copy Population Data

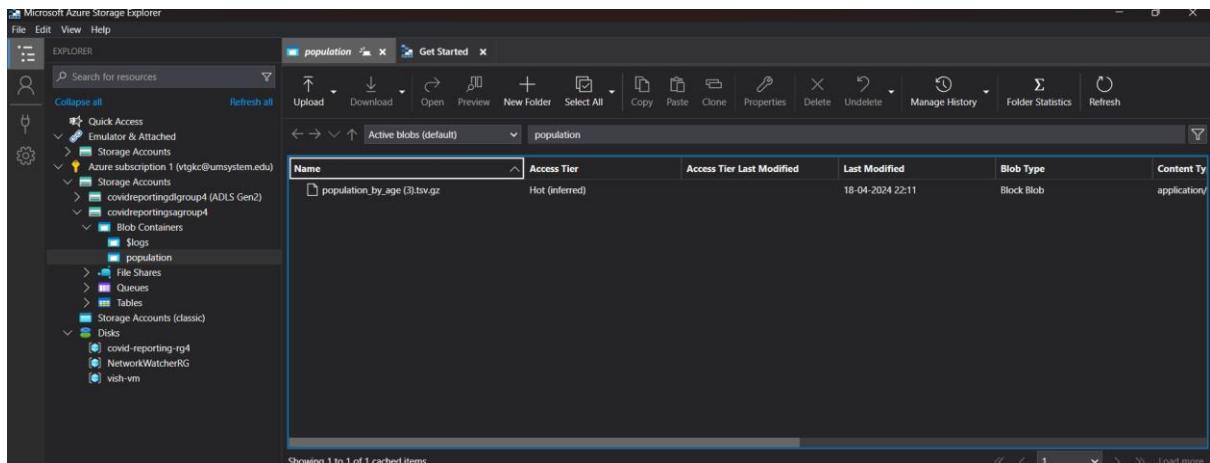
False:

Pipeline run ID: f2cd0708-6d73-4257-bf2d-092344b414a3

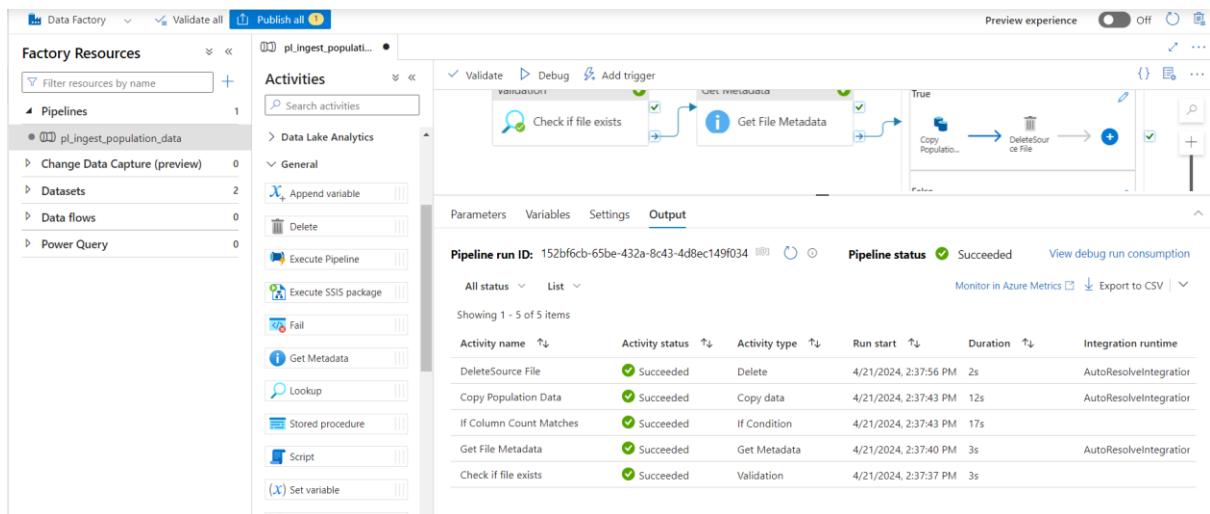
Pipeline status: Failed

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime
Send Email	Failed	Web	4/21/2024, 2:30:27 PM	8s	AutoResolveIntegrator
If Column Count Matches	Failed	If Condition	4/21/2024, 2:30:26 PM	10s	
Get File Metadata	Succeeded	Get Metadata	4/21/2024, 2:30:22 PM	3s	AutoResolveIntegrator
Check if file exists	Succeeded	Validation	4/21/2024, 2:30:17 PM	5s	

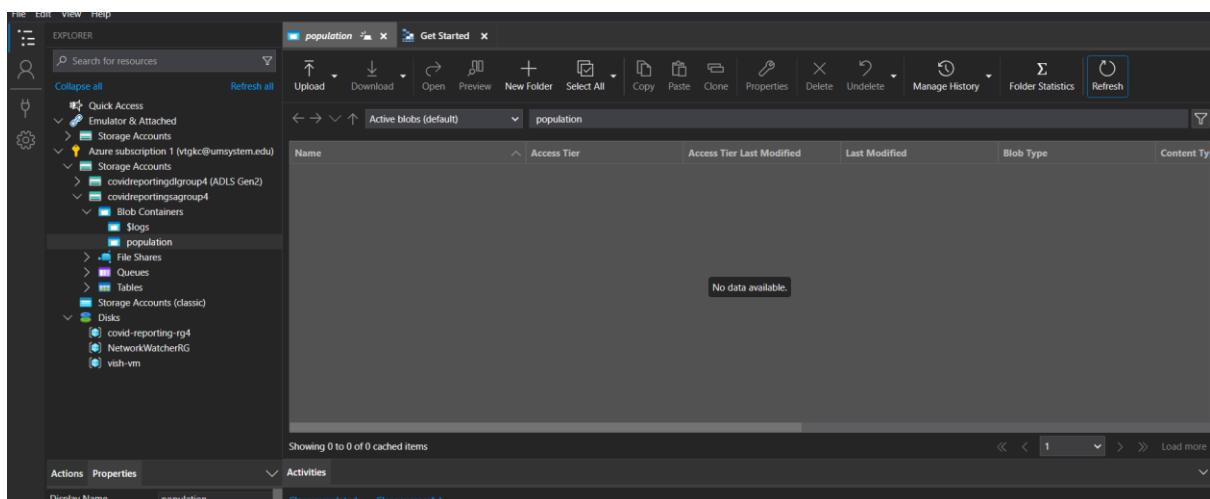
We could see the gz file in the corresponding path.



Delete file option successfully carried out



And now we could see the gz file got vanished



## Created an event Trigger

The screenshot shows the 'Triggers' section in the Azure Data Factory interface. On the left, there's a navigation sidebar with options like General, Author, Connections, Microsoft Purview, Source control, Author, Triggers (which is selected), Global parameters, Data flow libraries, Security, and Credentials. The main area is titled 'Triggers' with the sub-instruction 'To execute a pipeline set the trigger. Triggers represent a unit of processing that determines when a pipeline execution needs to be kicked off.' Below this is a table with one item:

Name	Type	Status	Related	Annotations
tr_ingest_population_data	Storage events	Started	0	

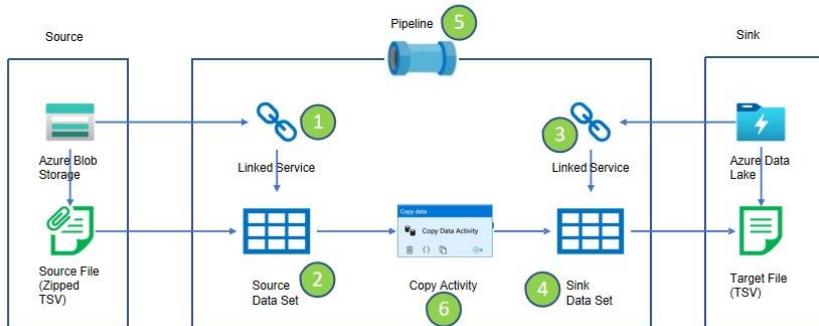
## Data Ingestion from https

Data is from ECDC website

<https://github.com/cloubboxacademy/covid19>

Copy Activity – Case & Deaths Data

## Copy Activity



Create pipeline:

1. URL:

[https://raw.githubusercontent.com/cloubboxacademy/covid19/main/raw/main/ecdc\\_data/cases\\_deaths.csv](https://raw.githubusercontent.com/cloubboxacademy/covid19/main/raw/main/ecdc_data/cases_deaths.csv)

Linked service.

Base URL : <https://raw.githubusercontent.com/>

The screenshot shows the 'Linked services' section of the Azure Data Factory blade. It lists three items:

- ls\_ablob\_covidreportinga (Type: Azure Blob Storage)
- ls\_adls\_covidreportingd (Type: Azure Data Lake Storage Gen2)
- ls\_http\_opendata\_ecdc\_europa\_eu (Type: HTTP)

## DL dataset

The screenshot shows the 'Datasets' blade in the Azure Data Factory interface. A 'DelimitedText' dataset named 'ds\_cases\_deaths\_raw\_csv\_dl' is selected. The properties pane on the right shows:

- Name: ds\_cases\_deaths\_raw\_csv\_dl
- Description: (empty)
- Annotations: (empty)

The 'Connection' tab of the dataset configuration shows:

- Linked service: ls\_adls\_covidreportingd
- File path: raw / ecdc / cases\_deaths\_csv
- Compression type: Select...
- Column delimiter: Comma (,)
- Row delimiter: Default (\r\n, or \n\r)
- Encoding: Default(UTF-8)
- Quote character: Double quote (")
- Escape character: Backslash (\)
- First row as header: checked
- Null value: (empty)

## http dataset

Relative url:

cloudboxacademy/covid19/main/raw/main/ecdc\_data/cases\_deaths.csv

## Pipeline

Details of http can be seen in the below screenshot

**Details** Refresh

Learn more on copy performance details from here.

Activity run id: 9357583f-7cee-461a-9ffa-f88895545310

**HTTP** → **Azure Data Lake Storage Gen2**  
Region: UK South

Succeeded

Data read: 14.445 MB  
Files read: 1  
Peak connections: 1

Data written: 14.445 MB  
Files written: 1  
Peak connections: 1

Copy duration: 00:00:10  
Throughput: 7.223 MB/s

HTTP → Azure Data Lake Storage Gen2

	Start time	Used DILUs	Used parallel copies	Duration	Working duration	Total duration
Listing source	4/21/2024, 7:36:15 PM	4	1	00:00:10	00:00:06	
Reading from source					00:00:00	
Writing to sink					00:00:02	

Data consistency verification: Unsupported

How satisfied or dissatisfied are you with the performance of this copy activity?

Properties

Name: pl\_ingest\_cases\_deaths\_data

Description:

Annotations

Integration runtime: User properties

Autofilesolve/integration

Now in the storage explorer you can see a new folder is created (ecdc).

Microsoft Azure Storage Explorer

File Edit View Help

EXPLORER

Search for resources Refresh all

Quick Access

Emulator & Attached

Storage Accounts

Storage Accounts

covidreportingdlgroup4 (ADLS Gen2)

Blob Containers

Slogs

raw

File Shares

Queues

Tables

covidreportingsagroup4

Storage Accounts (classic)

Disks

covid-reporting-rg

NetworkWatcherRG

vst-h-vm

Get Started

row Active blobs (default) raw

Name Access Tier Access Tier Last Modified Last Modified Blob Type Content Type

ecdc 21-04-2024 19:36 Folder

population 18-04-2024 23:09 Folder

Showing 1 to 2 of 2 cached items

Actions Properties Activities

Display Name: raw  
URL: https://covidreportingdlrg...  
Custom Domain:  
Type: Blob Container (ADLS Gen2)  
HNS Enabled: true  
DFS Endpoint: https://covidreportingdlrg...  
Lease State: available  
Lease Status: unlocked

Clear completed Clear successful

Transfer of 'D:\AzureStorageExplorer\CC\_data\_Files\Env\_preparation\population\_by\_age.tsv.gz' to 'population' complete: 1 item transferred (used SAS, discovery completed)  
Started at: 21-04-2024 15:32, Duration: 7 seconds

Copy As Copy Command to Clipboard

## Pipeline variable

## Pipeline source-sink hospital

The screenshot shows the Azure Data Factory pipeline editor. On the left, the 'Factory Resources' sidebar lists pipelines, datasets, and other resources. In the center, a pipeline named 'pl\_ingest\_cases\_deaths\_data' is selected, showing its activities. One activity, 'Copy Cases and Deaths Data', is highlighted with a green checkmark. The 'Properties' pane on the right shows the pipeline's name as 'pl\_ingest\_cases\_deaths\_data'. The 'Output' tab of the pipeline details shows a successful run ID: 9c3aa502-f1b6-41ce-ac0b-61f5dc67eeb1, with a status of 'Succeeded'. Below this, a table provides detailed activity statistics.

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties
Copy Cases and Deaths Data	Succeeded	Copy data	4/21/2024, 8:17:46 PM	11s	AutoResolveIntegration	

This screenshot displays the detailed run statistics for the 'Copy Cases and Deaths Data' activity. It shows the source as 'HTTP' and the sink as 'Azure Data Lake Storage Gen2, Region: UK South'. Key metrics include data read (1.059 MB), files read (1), peak connections (1), copy duration (00:00:09), and throughput (529.27 KB/s). The 'Properties' pane on the right is identical to the one in the first screenshot, showing the pipeline's name as 'pl\_ingest\_cases\_deaths\_data'.

This is hospital info and the file has been added

The screenshot shows the Microsoft Azure Storage Explorer interface. In the left sidebar, under 'Storage Accounts', there are entries for 'Azure subscription 1 (vtgk@umsystem.edu)' and 'covidreportingadgroup4 (ADLS Gen2)'. Under 'covidreportingadgroup4', there are 'Blob Containers' containing 'raw', 'Logs', and 'File Shares'. There are also 'Queues' and 'Tables' listed. The 'raw' container is selected, showing two CSV files: 'cases\_deaths.csv' and 'hospital\_admissions.csv'. The table view provides details such as Access Tier (Hot (inferred)), Last Modified (21-04-2024 19:36 and 21-04-2024 20:17), Blob Type (Block Blob), and Content Type (application/). Below the table, the 'Activities' section shows two completed transfers: one from 'raw/ecdc/cases\_deaths.csv' to a local path and another from 'D:\AzureStorageExplorer\CC\_data\_Files\Env\_preparation' to 'population\_by\_age.tsv.gz'.

## Pipeline parameters and schedule trigger

### Hospital admissions link

[https://raw.githubusercontent.com/cloubboxacademy/covid19/main/raw/main/ecdc\\_data/hospital\\_admissions.csv](https://raw.githubusercontent.com/cloubboxacademy/covid19/main/raw/main/ecdc_data/hospital_admissions.csv)

Relative url : cloubboxacademy/covid19/main/raw/main/ecdc\_data/hospital\_admissions.csv

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the navigation pane includes 'Data Factory', 'Factory Resources', 'Pipelines', 'Datasets', 'Data flows', and 'Power Query'. The 'pipelines' section is expanded, showing a pipeline named 'p\_inges'. A specific activity named 'HTTP' is selected, which has succeeded. The activity details show the source as 'HTTP' and the sink as 'Azure Data Lake Storage Gen2, Region: UK South'. Key metrics include Data read: 1.059 MB, Files read: 1, Peak connections: 1, Copy duration: 00:00:18, Throughput: 529.27 KB/s, and a detailed breakdown of the transfer process. The right side of the screen displays the success status and various monitoring and export options.

The screenshot shows the Microsoft Azure Data Factory interface for a pipeline named 'pl\_ingest\_ecdc\_data'. The pipeline has two activities: 'Copy data' and 'Copy Cases and Deaths Data'. The 'Copy data' activity is selected, showing its parameters, variables, settings, and output. The output table displays one item: 'Copy Cases and Deaths Data' with a status of 'Succeeded'. Pipeline run ID: bed2b028-1d74-4231-9de3-069537bcd497. Pipeline status: Succeeded.

## Hospital trigger run

The screenshot shows the Microsoft Azure Data Factory interface for trigger runs. The 'Trigger runs' section is selected, displaying a table of trigger runs. One run is listed: 'tr\_ingroup\_hospital\_admissions...' triggered by a 'Schedule trigger' on 4/23/2024 at 12:45:59 PM, with a status of 'Succeeded'. Run ID: 08584877125254819146848.

## Goto settings

The screenshot shows the Microsoft Azure Data Factory interface for trigger settings. The 'Properties' section is displayed, listing the following properties:

Name	Value
TriggerTime	4/23/24, 12:45:59 PM
ScheduleTime	4/23/24, 12:45:59 PM
triggerObject	{"name": "trigger_421B8CAF-BE66-42CF-81DA-E3028693F304", "startTime": "2024-04-23T17:45:59.9956925Z", "endTime": "2024-04-23T17:45:59.9956925Z"}

Screenshot of Microsoft Azure Data Factory Pipeline Runs page showing a successful activity run.

**Activity runs**

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID	Log
Copy ECDC Data	Succeeded	Copy data	4/23/2024, 12:46:01 PM	12s	AutoResolveIntegrationR		cd13217d-b807-47c9-9031-7d3692630464	

**Details** Refresh

Learn more on copy performance details from here.

Activity run id: cd13217d-b807-47c9-9031-7d3692630464

**HTTP** → **Azure Data Lake Storage Gen2**

Succeeded  
Azure IR region: UK South

Data read: 1.059 MB  
Files read: 1  
Peak connections: 1

Data written: 1.059 MB  
Files written: 1  
Peak connections: 1

Copy duration: 00:00:10  
Throughput: 529.27 KB/s

HTTP → Azure Data Lake Storage Gen2

- Start time: 4/23/2024, 12:46:02 PM
- Used DIUs: 4
- Used parallel copies: 1
- Duration: 00:00:10

Details	Working duration	Total duration
Queue	00:00:06	00:00:06
Transfer	Listing source 00:00:00 Reading from source 00:00:00 Writing to sink 00:00:00	00:00:02

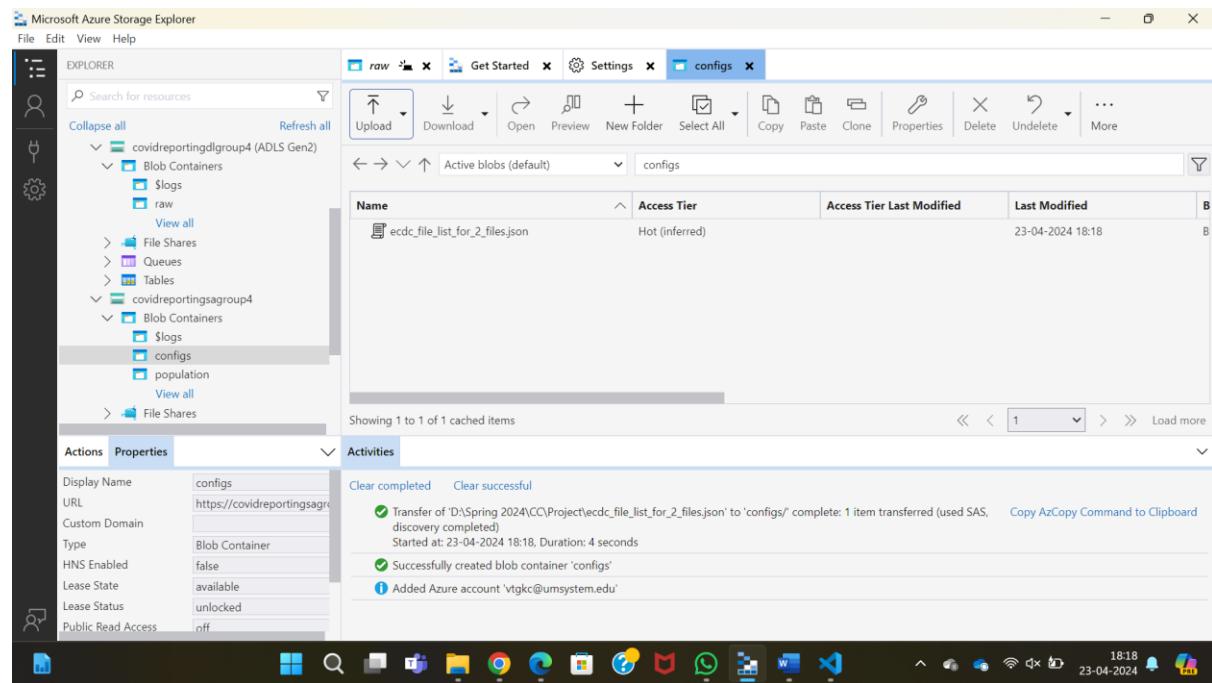
Data consistency verification: Unsupported

How satisfied or dissatisfied are you with the performance of this copy activity?

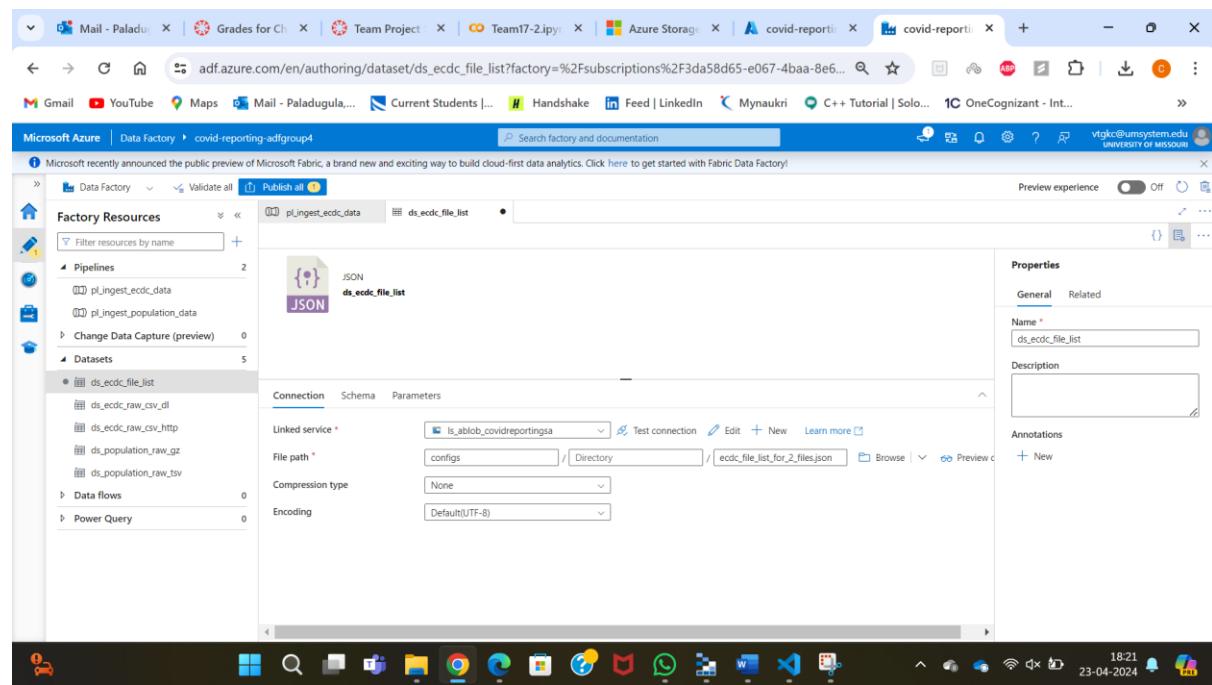
## Control flow Activities:

Json file: ecdc\_file\_list\_for\_2\_files

Upload the file in the storage account in configs.



Upload the same json in the datasets in data factory.



Make a new pipeline the lookup activity.

5:39 mins

In manage, Linked services update the base url,

The screenshot shows the Microsoft Azure Data Factory interface. On the left, a sidebar lists various settings like General, Connections, Source control, and Security. The main area is titled 'Linked services' and shows a table of existing connections. A new connection is being created with the name 'ls\_http\_opendata\_ecdc\_europa\_eu'. The 'Type' is set to 'HTTP'. In the 'Base URL' field, the expression '@{linkedService().sourceBaseUrl}' is entered. Other configuration options like 'Server certificate validation' and 'Authentication type' (set to 'Anonymous') are also visible.

In pipelines, create the parameters for the base url.

The screenshot shows the Microsoft Azure Data Factory interface under the 'Factory Resources' section. It displays a pipeline named 'pl\_ingest\_ecdc\_data'. The 'Activities' pane shows a 'Copy data' activity. Below it, the 'Parameters' tab is open, listing three parameters: 'sourceRelativeURL', 'sinkFileName', and 'sourceBaseUrl', each with a string type and a value placeholder. The pipeline structure also includes other activities like 'Move and transform' and 'Synapse'.

In datasets, create the parameters for the base url.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, and Data flows. Under Pipelines, there are two entries: 'pl\_ingest\_ecdc\_data' and 'ds\_ecdc\_file\_list'. Under Datasets, there are five entries: 'ds\_ecdc\_file\_list', 'ds\_ecdc\_raw\_csv\_dl', 'ds\_ecdc\_raw\_csv\_http', 'ds\_population\_raw\_gz', and 'ds\_population\_raw\_tsv'. Under Data flows, there is one entry: 'pl\_ingest\_ecdc\_data'. The main workspace displays the 'ds\_ecdc\_raw\_csv\_http' dataset, which is a DelimitedText type with a CSV icon. Below the dataset, the 'Parameters' tab is selected, showing two parameters: 'relativeURL' (String type) and 'baseURL' (String type). The 'Connection' and 'Schema' tabs are also visible. The top navigation bar includes links for Mail, YouTube, Maps, Gmail, and various Microsoft services like Azure Storage and Power BI. The bottom taskbar shows the Windows Start button, Task View, File Explorer, and other system icons.

This screenshot is similar to the first one but focuses on the 'Linked service' configuration for the 'ds\_ecdc\_raw\_csv\_http' dataset. In the 'Parameters' tab, the 'sourceBaseURL' parameter is set to '@dataset().baseURL'. The 'Linked service' dropdown is expanded to show 'ls\_http\_opendata\_ecdc\_europa\_eu' as the selected service. Below it, the 'Linked service properties' section shows the 'Name' and 'Value' for 'sourceBaseURL'. Further down, the 'Relative URL' field contains '@dataset().relativeURL'. Other settings like 'Compression type', 'Column delimiter', and 'Row delimiter' are also visible. The rest of the interface and taskbar are identical to the first screenshot.

Relative url: cloudboxacademy/covid19/main/raw/main/ecdc\_data/country\_response.csv

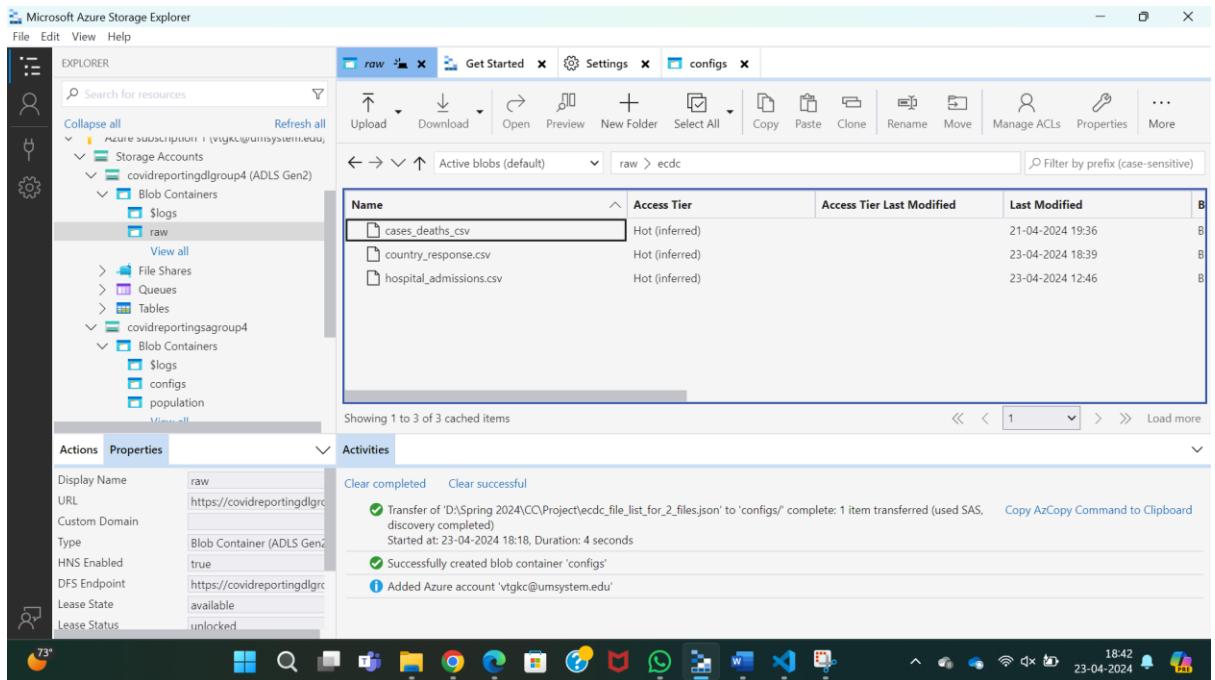
Base url: <https://raw.githubusercontent.com/>

The screenshot shows the Microsoft Azure Data Factory pipeline configuration interface. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (pl\_inge...), 'Datasets' (ds\_ecdc\_file\_list, ds\_ecdc\_raw\_csv\_dl, ds\_ecdc\_raw\_csv\_http, ds\_population\_raw\_gz, ds\_population\_raw\_tsv), and 'Power Query'. In the main pane, a pipeline named 'pl\_inge...' is selected. The 'Activities' section shows a sequence of steps: 'Move and transform' (highlighted in blue), 'Synapse', 'Azure Data Explorer', 'Batch Service', 'Databricks', 'Data Lake Analytics', 'General', 'HDInsight', 'Iteration & conditionals', 'Machine Learning' (with three sub-options: Machine Learning Batch, Machine Learning Up, Machine Learning Exec), and 'Power Query'. Below the activities, the 'Parameters' tab is active, showing three parameters: sourceRelativeURL (string, value: cloudboxacademy/covid19/...), sinkFileName (string, value: country\_response.csv), and sourceBaseUrl (string, value: https://raw.githubusercontent.com/). The bottom status bar shows the date and time as 23-04-2024 18:39.

Copied the country response from ecdc website to data lake.

The screenshot shows the details of a completed pipeline activity. The activity type is 'HTTP' and it succeeded, pointing to 'Azure Data Lake Storage Gen2' in 'Region: UK South'. Key performance metrics are displayed: Data read: 47.308 KB, Files read: 1, Peak connections: 1; Data written: 47.308 KB, Files written: 1, Peak connections: 1. The 'Copy duration' was 00:00:09 with a throughput of 47.308 KB/s. The 'HTTP → Azure Data Lake Storage Gen2' section shows the start time as 4/23/2024, 6:39:47 PM, and a total duration of 00:00:09. The breakdown of the duration is: Queue (00:00:05) and Transfer (00:00:04). The status bar at the bottom shows the date and time as 23-04-2024 18:40.

We have verified the same thing in the explorer.



We publish it to save the changes.

### Meta data driven pipeline:

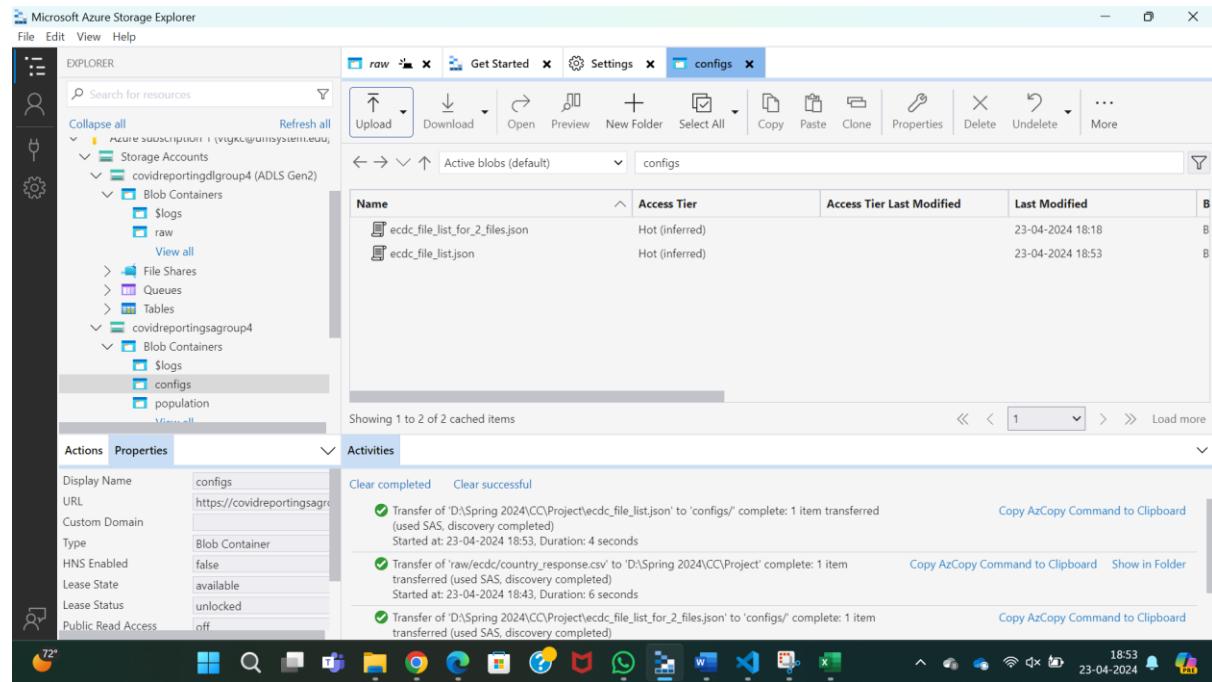
File name : ecdc\_file\_list

Relative URL: [cloubboxacademy/covid19/main/raw/main/ecdc\\_data/testing.csv](https://cloubboxacademy/covid19/main/raw/main/ecdc_data/testing.csv)

Base URL: <https://raw.githubusercontent.com/>

```
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More
Team17-2.ipynb {} ecdc_file_list_for_2_files.json {} ecdc_file_list.json X
D: > Spring 2024 > CC > Project > {} ecdc_file_list.json > ...
1 [
2   {
3     "sourceBaseUrl": "https://raw.githubusercontent.com/",
4     "sourceRelativeURL": "cloubboxacademy/covid19/main/raw/main/ecdc_data/cases_deaths.csv",
5     "sinkFileName": "cases_deaths.csv"
6   },
7   {
8     "sourceBaseUrl": "https://raw.githubusercontent.com/",
9     "sourceRelativeURL": "cloubboxacademy/covid19/main/raw/main/ecdc_data/hospital_admissions.csv",
10    "sinkFileName": "hospital_admissions.csv"
11  },
12  {
13    "sourceBaseUrl": "https://raw.githubusercontent.com/",
14    "sourceRelativeURL": "cloubboxacademy/covid19/main/raw/main/ecdc_data/testing.csv",
15    "sinkFileName": "testing.csv"
16  },
17  {
18    "sourceBaseUrl": "https://raw.githubusercontent.com/",
19    "sourceRelativeURL": "cloubboxacademy/covid19/main/raw/main/ecdc_data/country_response.csv",
20    "sinkFileName": "country_response.csv"
21  }
22 ]
23 
```

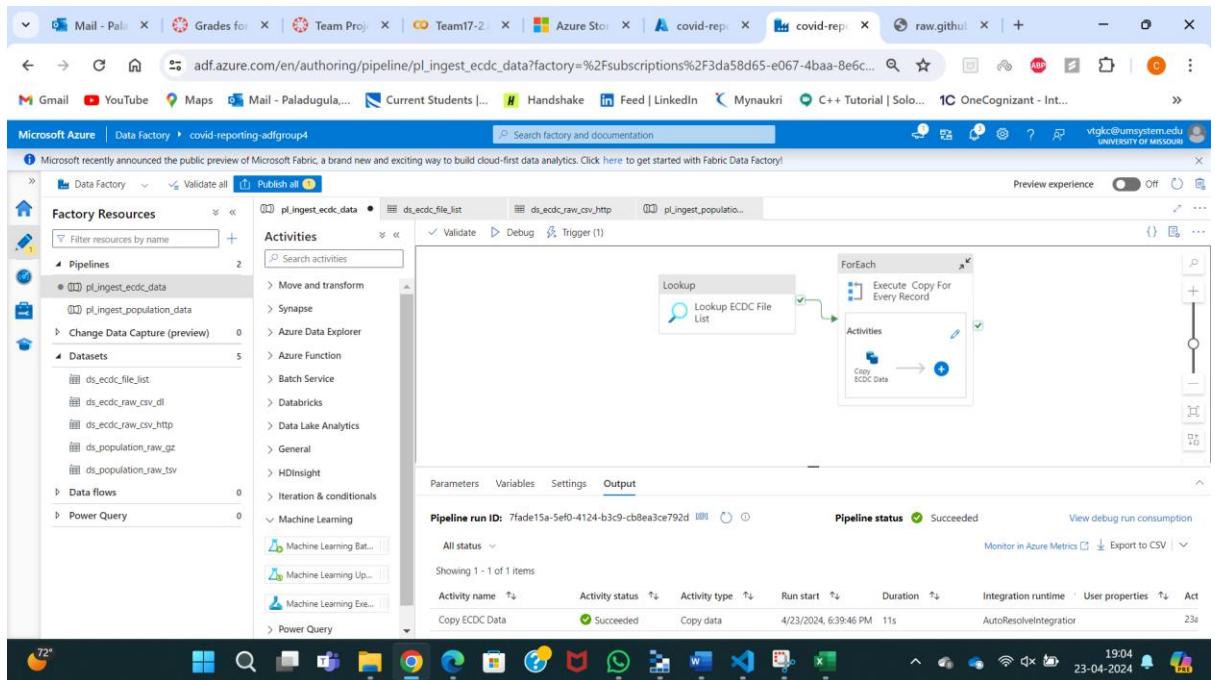
## Upload the file in Explorer



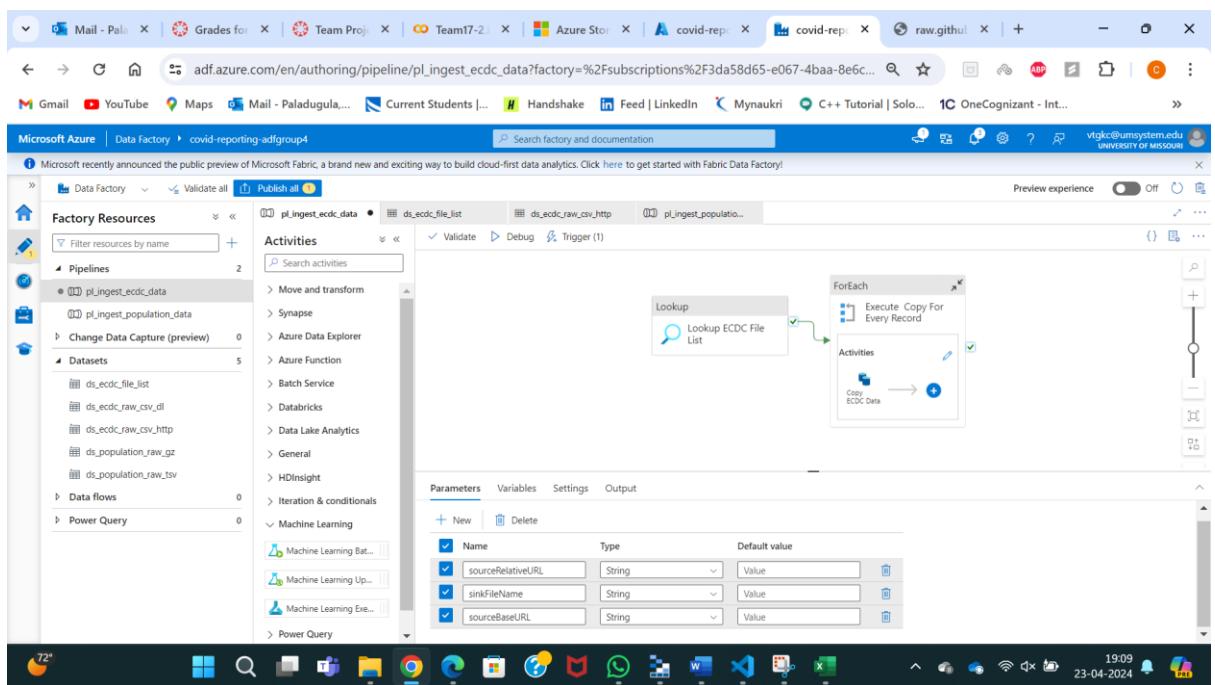
## Creating a lookup activity in the ecdc pipeline,

The screenshot shows the Microsoft Data Factory portal. The left sidebar lists factory resources: Pipelines, Datasets, Data flows, and Power Query. The 'Activities' tab is selected in the center, showing a list of available activities including 'Move and transform', 'Synapse', 'Azure Data Explorer', etc. A 'Lookup' activity is currently being configured. The configuration pane shows the 'Source dataset' set to 'ds\_ecdc\_file\_list'. Other settings include 'File path type' (set to 'File path in dataset'), 'Start time (UTC)', and 'End time (UTC)'. The pipeline editor on the right shows a pipeline named 'pl\_ingest\_ecdc\_data' with several stages connected.

Create a for each activity and connect to the lookup activity, then move the copy data to from the pipeline to for each activity.

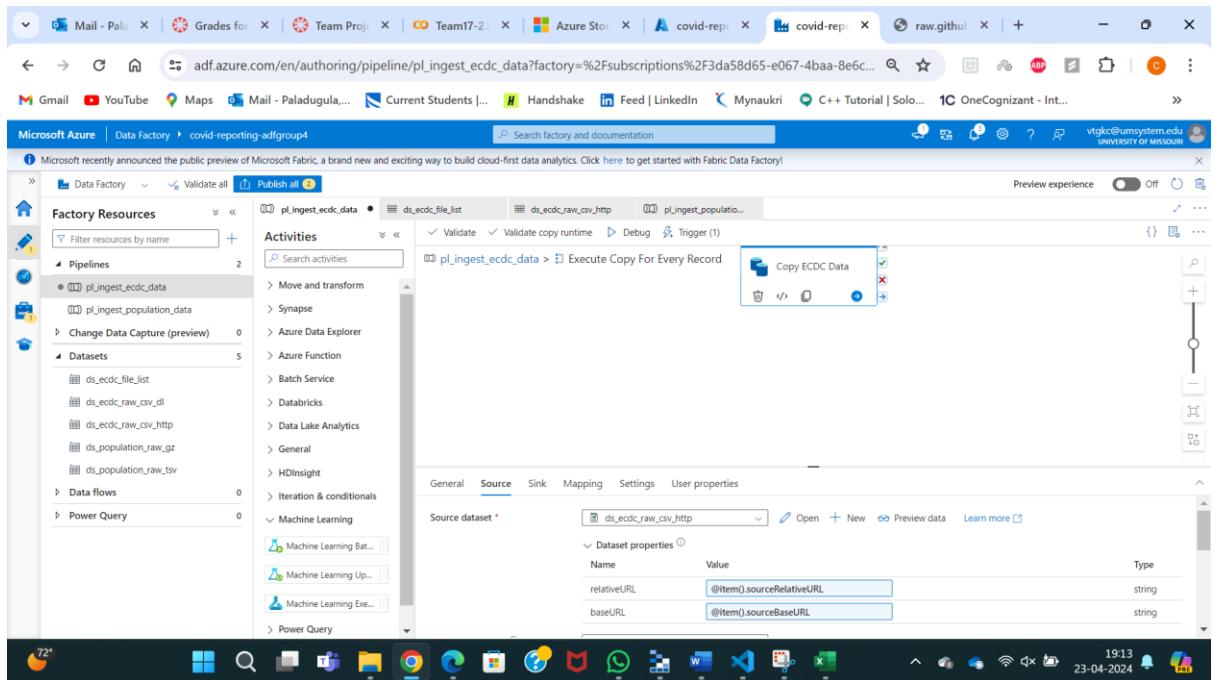


Then delete the parameters in the pipeline,

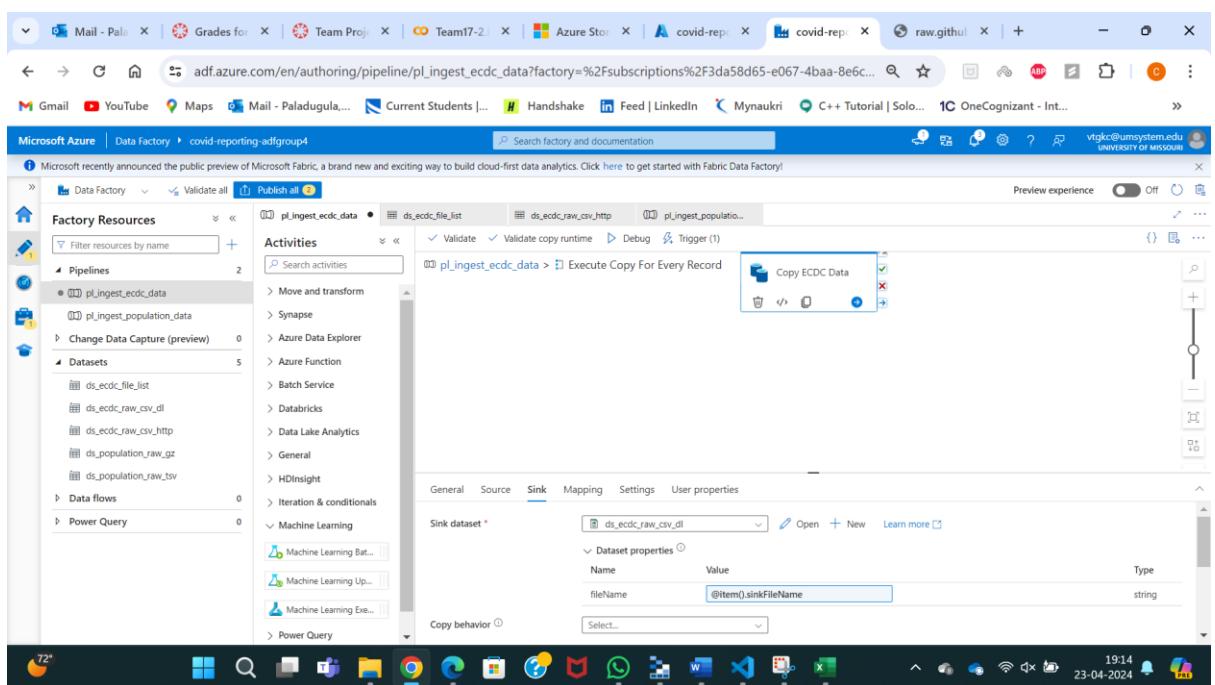


Open activity in the for each activity, Open the copy data.

Update the relative url and base url, in the source



In sink, update the sink filename



The pipeline is executed successfully.

The screenshot shows the Microsoft Azure Data Factory pipeline editor. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. The main workspace displays a pipeline named 'pl\_ingest\_ecdc\_data' with a 'For Each' loop activity. The 'Activities' pane shows a 'Lookup' activity followed by an 'Execute Copy For Every Record' activity. Below the workspace is a table of pipeline run history:

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
Copy ECDC Data	Succeeded	Copy data	4/23/2024, 7:16:39 PM	14s	AutoResolveintegrator		93ad0558-3d66-4dd6-a6b0-a5f08cd0b70
Copy ECDC Data	Succeeded	Copy data	4/23/2024, 7:16:39 PM	12s	AutoResolveintegrator		89ba99e1-f6ae-4eb5-b312-ed3ba4049ef6
Copy ECDC Data	Succeeded	Copy data	4/23/2024, 7:16:39 PM	15s	AutoResolveintegrator		f2164ac3-bc76-4881-b9d1-99121ac264e
Copy ECDC Data	Succeeded	Copy data	4/23/2024, 7:16:39 PM	14s	AutoResolveintegrator		709997-d502-4140-9f50-20200981609
Execute Copy For Every Record	Succeeded	ForEach	4/23/2024, 7:16:38 PM	17s			10f8aa4d-30be-4568-b3b5-8e5d9d9cfec
Lookup ECDC File List	Succeeded	Lookup	4/23/2024, 7:16:33 PM	4s	AutoResolveintegrator		a61b9ffa-cb5c-49bf-9a6d-21955d0d70fd

Now we can check the same in the explorer.

The screenshot shows the Microsoft Azure Storage Explorer. The left sidebar shows storage accounts and containers. The main pane displays blobs in the 'raw' container of the 'ecdc' blob container. The table lists the blobs:

Name	Access Tier	Access Tier Last Modified	Last Modified
cases_deaths.csv	Hot (inferred)		23-04-2024 19:16
country_response.csv	Hot (inferred)		23-04-2024 19:16
hospital_admissions.csv	Hot (inferred)		23-04-2024 19:16
testing.csv	Hot (inferred)		23-04-2024 19:16

The bottom section shows recent activities:

- Clear completed
- Clear successful
- Deletion of 'cases\_deaths.csv' from 'raw/ecdc/' completed: 1 completed (used name and key) Started at: 23-04-2024 19:18, Duration: 4 seconds
- Transfer of 'D:\Spring 2024\CC\Project\ecdc\_file\_list.json' to 'configs/' complete: 1 item transferred (used SAS, discovery completed) Started at: 23-04-2024 18:53, Duration: 4 seconds
- Transfer of 'raw/ecdc/country\_response.csv' to 'D:\Spring 2024\CC\Project' complete: 1 item transferred (used SAS, discovery completed) Started at: 23-04-2024 18:43, Duration: 6 seconds

Create a new trigger, To update all the data automatically

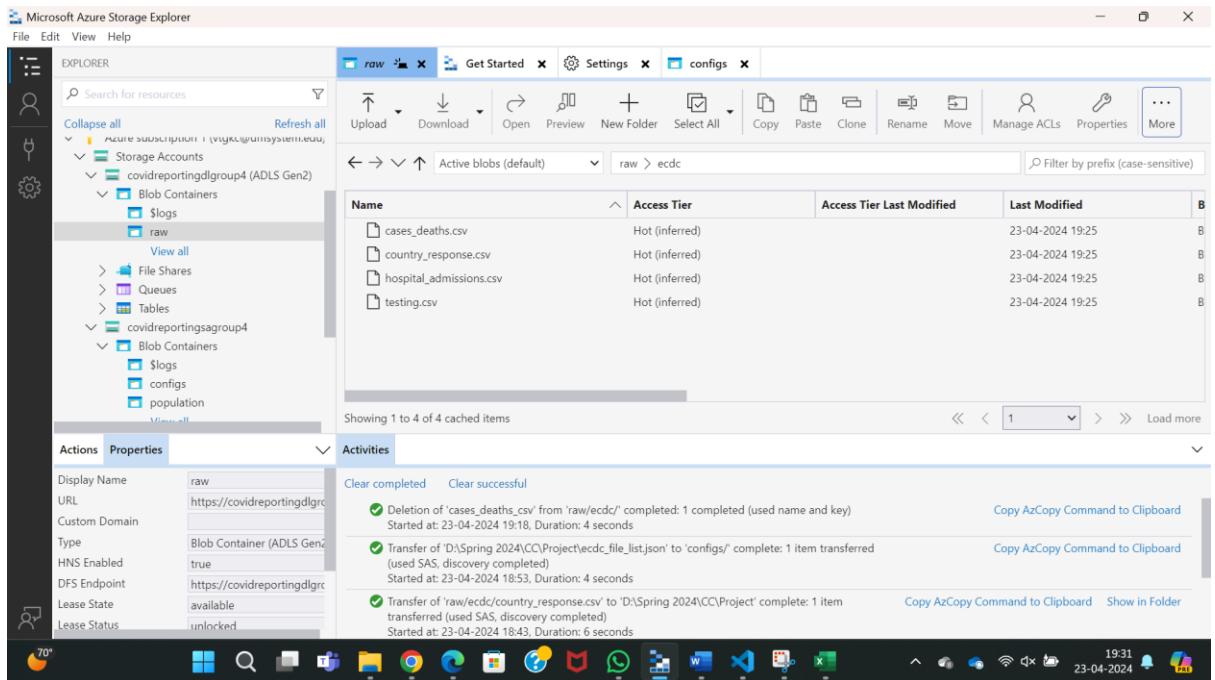
The screenshot shows the Microsoft Azure Data Factory interface. On the left, a sidebar lists various sections like General, Connections, Source control, Author, and Triggers. The Triggers section is selected. In the main area, there's a table titled 'Triggers' with one item: 'tr\_ingest\_population\_data'. A modal window titled 'New trigger' is open on the right, showing the configuration for a 'Schedule' trigger. The 'Start date' is set to '4/23/2024 7:28:04 PM'. The 'Time zone' is 'Central Time (US & Canada) (UTC-6)'. The 'Recurrence' is set to 'Every 1 Day(s)'. Under 'Advanced recurrence options', there are fields for 'Hours' and 'Minutes', and a 'Schedule execution times' dropdown set to '19:23'. There are also checkboxes for 'Specify an end date' and 'Annotations'. At the bottom of the modal are 'OK' and 'Cancel' buttons.

Add the trigger in the pipeline and publish all.

The screenshot shows the Microsoft Azure Data Factory interface. The sidebar is similar to the previous screen, with the Triggers section selected. The main area is titled 'Trigger runs' and displays a table of trigger executions. The table has columns: Trigger name, Trigger type, Trigger time, Status, Pipelines, Run, Message, Properties, and Run ID. Two entries are listed:

Trigger name	Trigger type	Trigger time	Status	Pipelines	Run	Message	Properties	Run ID
tr_ingest_ecdc_data	Schedule trigger	4/23/2024, 7:25:00 PM	Succeeded	1	Original			08584676885847118364426
tr_ingest_hospital_admissions...	Schedule trigger	4/23/2024, 12:45:59 PM	Succeeded	1	Original			0858477125248191466848

In the file explorer, we can see the data is updated.



## Data – flow: Data transformation

Transform cases and deaths data, only performing for country Europe, Created country lookup file by doing the lookup activity on cases deaths and testing file, to bring two digit and three-digit country code's together.

The screenshot shows a Microsoft Excel spreadsheet titled 'country\_lookup'. The table has the following structure:

	A	B	C	D	E	F	G
1	country	country_code_2_digit	country_code_3_digit	continent	population		
2	Aruba	AW	ABW	America	106766		
3	Afghanistan	AF	AFG	Asia	38928341		
4	Angola	AO	AGO	Africa	32866268		
5	Anguilla	AI	AIA	America	15002		
6	Albania	AL	ALB	Europe	2862427		
7	Andorra	AD	AND	Europe	76177		
8	United Arab Emirates	AE	ARE	Asia	9890400		
9	Argentina	AR	ARG	America	45195777		
10	Armenia	AM	ARM	Europe	2963234		
11	Antigua and Barbuda	AG	ATG	America	97928		
12	Australia	AU	AUS	Oceania	25499881		

Create a new data flow, Enable the data flow debug,

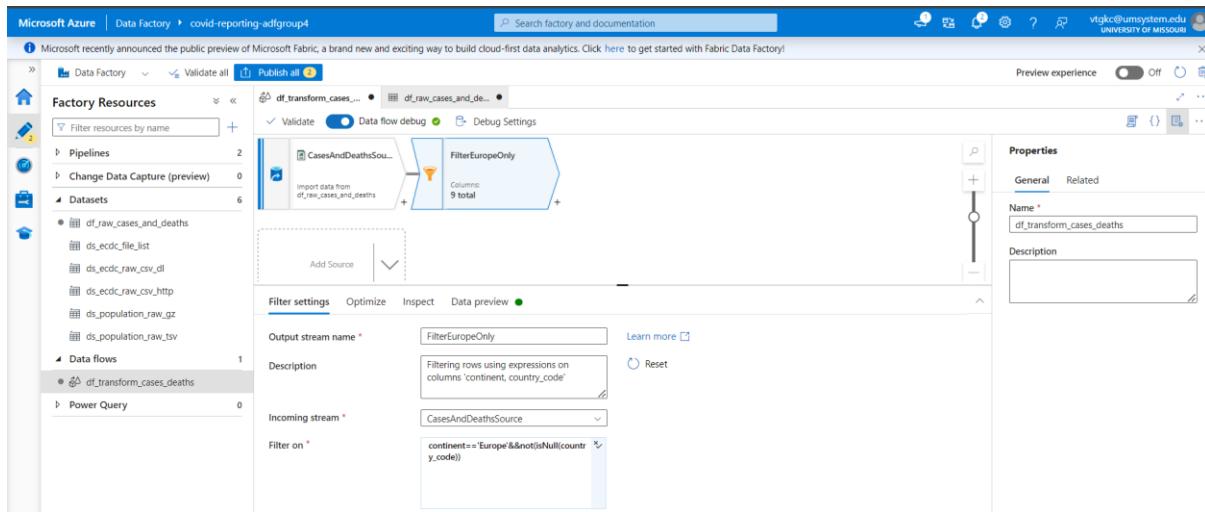
The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists Pipelines (2), Change Data Capture (preview) (0), Datasets (5), and Data flows (1). The 'dataflow1' item under Data flows is selected. The main workspace shows a 'dataflow1' data flow with a single 'Add Source' step. The 'Properties' panel on the right shows the 'Name' field set to 'dataflow1'. The 'Data flow debug' switch is turned on. The 'Parameters' tab is selected.

Create a source and give the cases death file as source, you can see the new file in the datasets

The screenshot shows the Microsoft Azure Data Factory interface. The 'Factory Resources' sidebar lists Pipelines (2), Change Data Capture (preview) (0), Datasets (5), and Data flows (1). A dataset named 'df\_transform\_cases\_deaths' is selected. The main workspace shows a 'df\_transform\_cases...' data flow with a 'CasesAndDeathsSource' source. The 'Properties' panel on the right shows the 'Name' field set to 'df\_transform\_cases\_deaths'. The 'Dataset' dropdown in the 'Source settings' section is highlighted with a yellow box. Other visible fields include 'Output stream name' (CasesAndDeathsSource), 'Description' (Import data from df\_raw\_cases\_and\_deaths), and 'Source type' (Dataset).

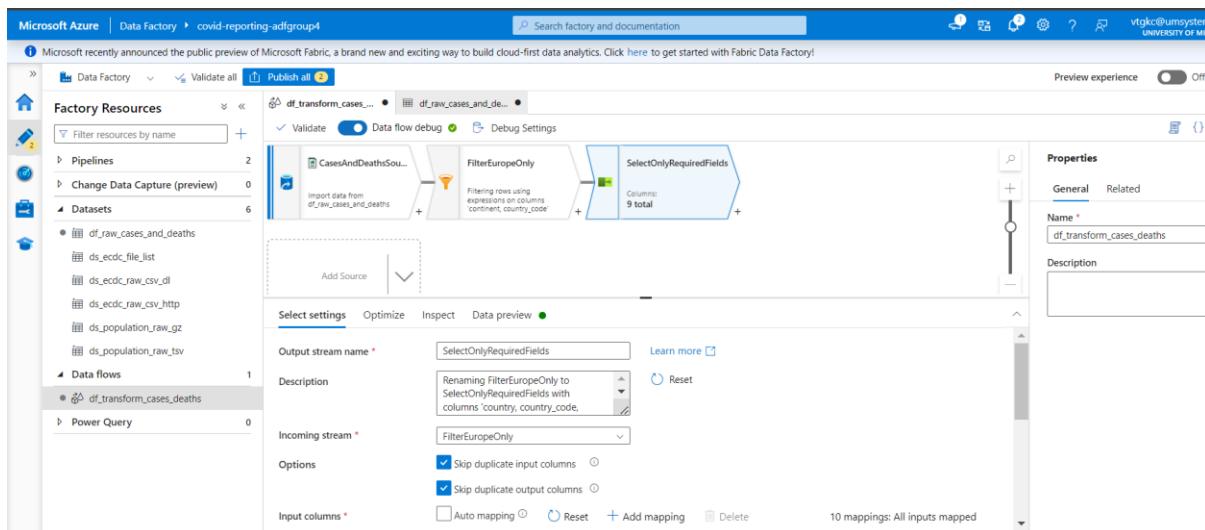
Filter transformation:

Add the filter to the source using the + icon, filter condition as Europe country. Also went through other options in Filter to explore.



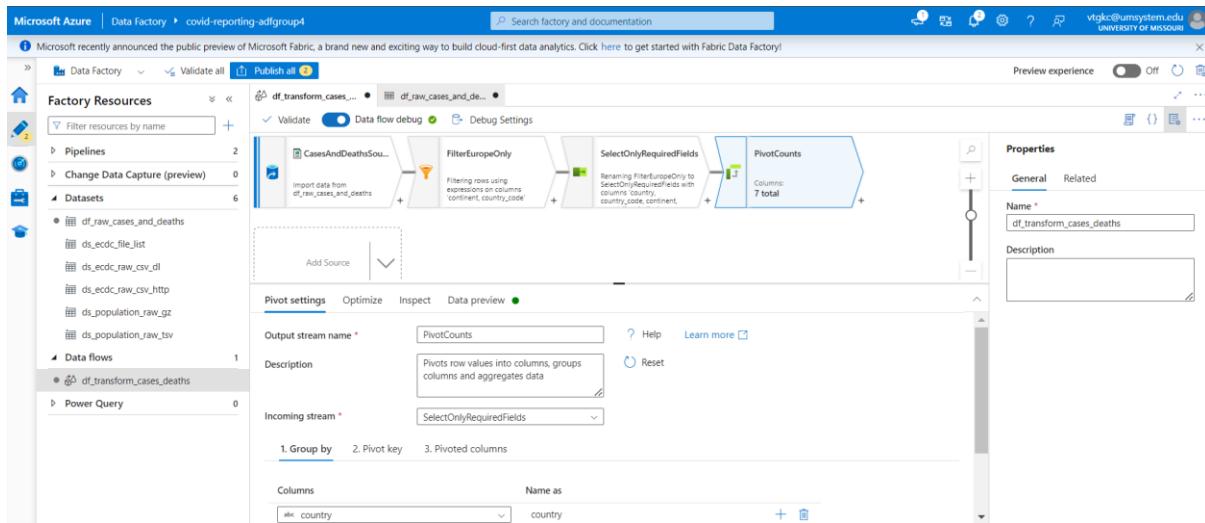
Select transformation:

create a select transform, using + icon. Also went through other options in select to explore.

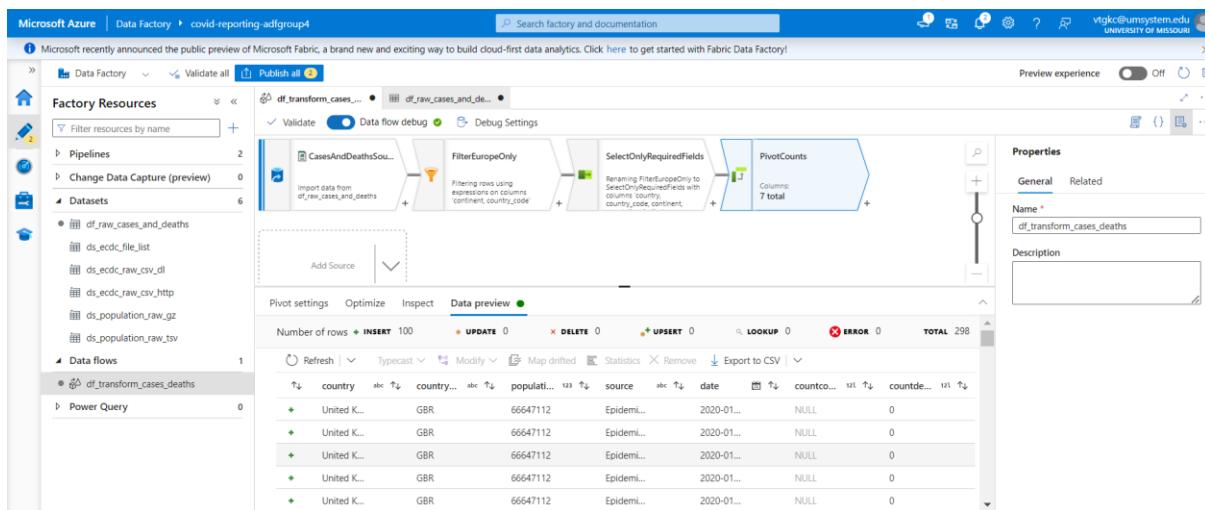


Pivot transformation:

create a pivot transform, using + icon.



In Data preview tab, two new columns as countdeaths and countconfirmed cases



Lookup transformation:

Create a new dataset for lookup, --- countrylookup

Microsoft Azure | Data Factory > covid-reporting-adfgroup4

Factory Resources

Pipelines: 2

Change Data Capture (preview): 0

Datasets: 7

- ds\_raw\_cases\_and\_deaths
- ds\_country\_lookup**
- ds\_ecdc\_file\_list
- ds\_ecdc\_raw\_csv\_dl
- ds\_ecdc\_raw\_csv\_http
- ds\_population\_raw\_gz
- ds\_population\_raw\_tsv

Data flows: 1

- df\_transform\_cases\_deaths**

Power Query: 0

Properties

Name: ds\_country\_lookup

Description:

## Create a new source for the country lookup,

Microsoft Azure | Data Factory > covid-reporting-adfgroup4

Factory Resources

Pipelines: 2

Change Data Capture (preview): 0

Datasets: 7

- ds\_raw\_cases\_and\_deaths
- ds\_country\_lookup**
- ds\_ecdc\_file\_list
- ds\_ecdc\_raw\_csv\_dl
- ds\_ecdc\_raw\_csv\_http
- ds\_population\_raw\_gz
- ds\_population\_raw\_tsv

Data flows: 1

- df\_transform\_cases\_deaths**

Power Query: 0

Properties

Name: df\_transform\_cases\_deaths

Description:

## Join the lookup with the countrylookup source

Microsoft Azure | Data Factory > covid-reporting-adfgroup4

Factory Resources

Pipelines: 2

Change Data Capture (preview): 0

Datasets: 7

- ds\_raw\_cases\_and\_deaths
- ds\_country\_lookup**
- ds\_ecdc\_file\_list
- ds\_ecdc\_raw\_csv\_dl
- ds\_ecdc\_raw\_csv\_http
- ds\_population\_raw\_gz
- ds\_population\_raw\_tsv

Data flows: 1

- df\_transform\_cases\_deaths**

Power Query: 0

Properties

Name: df\_transform\_cases\_deaths

Description:

In data preview you can see two new columns two digit and three digit country code's

The screenshot shows the Microsoft Azure Data Factory interface for a pipeline named "covid-reporting-adgroup4". The "Data flows" section contains one item, "df\_transform\_cases\_deaths". The "Data preview" tab is selected, displaying a table with 298 rows. The columns are: rce, date, reported..., countconfirmed..., countdeaths, country\_code\_2\_digit, country\_code\_3\_digit, continent. The last two columns, "country\_code\_2\_digit" and "country\_code\_3\_digit", are highlighted in green, indicating they are newly created columns from the transformation step.

Create the select, after lookup as selectforsink,

The screenshot shows the Microsoft Azure Data Factory interface for the same pipeline. The "Data flows" section shows the "df\_transform\_cases\_deaths" data flow. The "Select settings" tab is selected. In the "Input columns" section, there are two columns: "abc: PivotCounts@country" and "abc: country\_code\_2\_digit". Mappings are defined: "country" is mapped to "abc: PivotCounts@country" and "country\_code\_2\_digit" is mapped to "abc: country\_code\_2\_digit". The "Properties" pane on the right shows the transformation is named "SelectForSink".

In data preview, we can see the columns,

Create a Sink after the select, create a new dataset in the explorer with name processed,

In the sink create a new dataset,

## The data preview in sink,

Microsoft Azure | Data Factory > covid-reporting-adgroup4

Factory Resources

- Pipelines
- Change Data Capture (preview)
- Datasets
- ds\_country\_lookup
- ds\_ecdc\_file\_list
- ds\_ecdc\_raw\_csv\_dl
- ds\_ecdc\_raw\_csv\_http
- ds\_population\_raw\_gz
- ds\_population\_raw\_tsv
- ds\_processed\_cases\_and\_deaths
- ds\_raw\_cases\_and\_deaths
- Data flows
- df\_transform\_cases\_deaths
- Power Query

Properties

- General
- Related
- Name: df\_transform\_cases\_deaths
- Description

Pipeline Steps:

- lectOnlyRequiredFields
- PivotCounts
- LookupCountry
- SelectForSink
- CaseAndDeathsSink

Data preview

country	country_code	population	cases_count	deaths_count	reported_date
United Kin...	UK	GBR	66647112	NULL	0
United Kin...	UK	GBR	66647112	NULL	2020-01-03
United Kin...	UK	GBR	66647112	NULL	2020-01-04
United Kin...	UK	GBR	66647112	NULL	2020-01-05
United Kin...	UK	GBR	66647112	NULL	2020-01-06

Now give validate and publish,

Microsoft Azure | Data Factory > covid-reporting-adgroup4

Factory Resources

- Pipelines
- Change Data Capture (preview)
- Datasets
- ds\_country\_lookup
- ds\_ecdc\_file\_list
- ds\_ecdc\_raw\_csv\_dl
- ds\_ecdc\_raw\_csv\_http
- ds\_population\_raw\_gz
- ds\_population\_raw\_tsv
- ds\_processed\_cases\_and\_deaths
- ds\_raw\_cases\_and\_deaths
- Data flows
- df\_transform\_cases\_deaths
- Power Query

Properties

- General
- Related
- Name: df\_transform\_cases\_deaths
- Description

Pipeline Steps:

- lectOnlyRequiredFields
- PivotCounts
- LookupCountry
- SelectForSink
- CaseAndDeathsSink

Data preview

country	country_code	population	cases_count	deaths_count	reported_date
United Kin...	UK	GBR	66647112	NULL	0
United Kin...	UK	GBR	66647112	NULL	2020-01-03
United Kin...	UK	GBR	66647112	NULL	2020-01-04
United Kin...	UK	GBR	66647112	NULL	2020-01-05
United Kin...	UK	GBR	66647112	NULL	2020-01-06

Data flow validation output

Your data flow has been validated.  
No errors were found.

Microsoft Azure | Data Factory > covid-reporting-adgroup4

Factory Resources

- Capture (preview)
- ds\_country\_lookup
- ds\_ecdc\_file\_list
- ds\_ecdc\_raw\_csv\_dl
- ds\_ecdc\_raw\_csv\_http
- ds\_population\_raw\_gz
- ds\_population\_raw\_tsv
- ds\_processed\_cases\_and\_deaths
- ds\_raw\_cases\_and\_deaths
- df\_transform\_cases\_deaths

Properties

- General
- Related
- Name: df\_transform\_cases\_deaths
- Description

Pipeline Steps:

- lectOnlyRequiredFields
- PivotCounts
- LookupCountry
- SelectForSink
- CaseAndDeathsSink

Data preview

country	country_code	population	cases_count	deaths_count	reported_date
United Kin...	UK	GBR	66647112	NULL	0
United Kin...	UK	GBR	66647112	NULL	2020-01-03
United Kin...	UK	GBR	66647112	NULL	2020-01-04
United Kin...	UK	GBR	66647112	NULL	2020-01-05
United Kin...	UK	GBR	66647112	NULL	2020-01-06

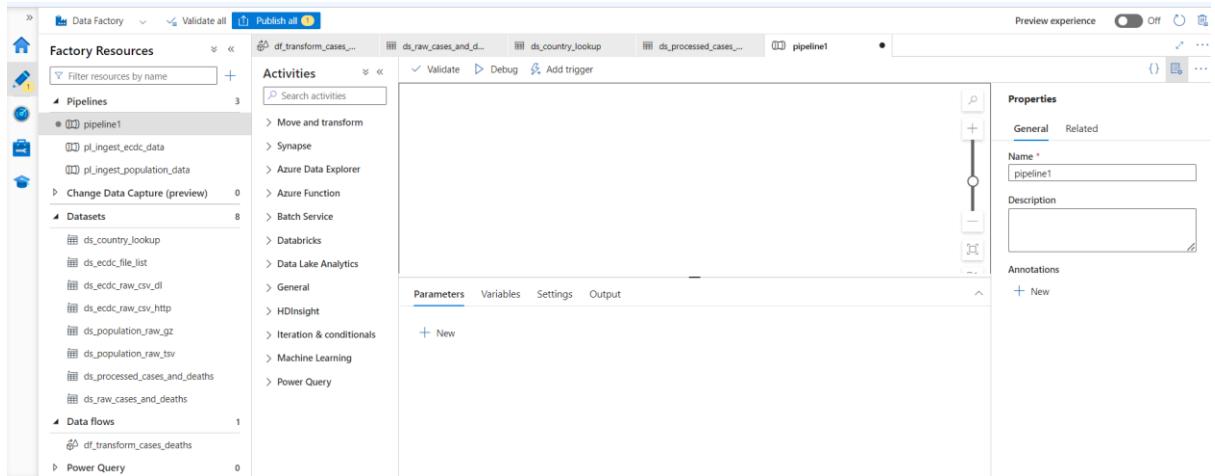
Preview experience

Data

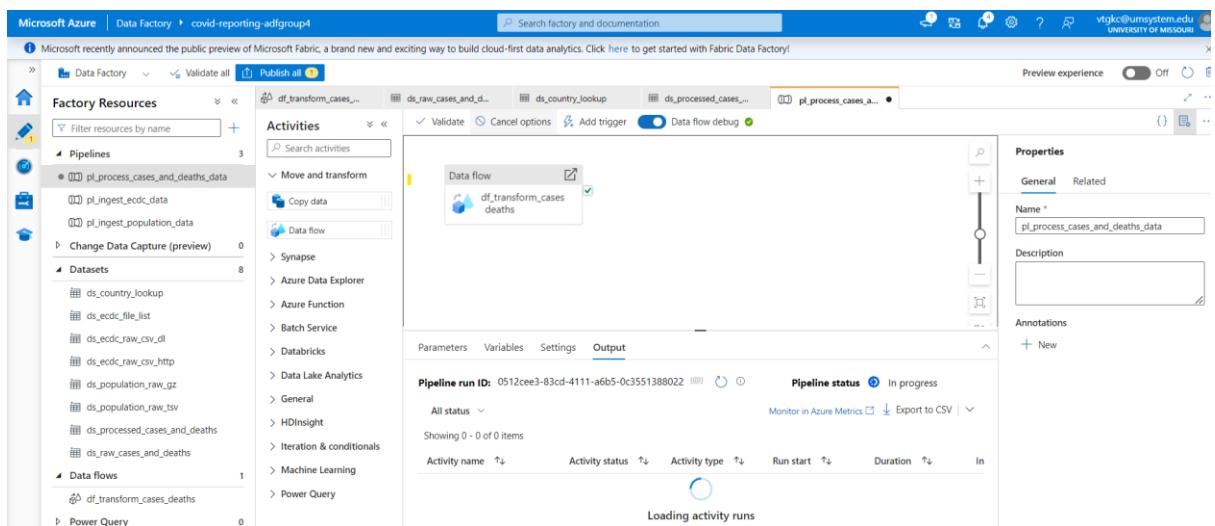
Publishing completed  
Successfully published

Your data flow has been validated.  
No errors were found.

## Create ADF pipeline:



In activity, add data flow, select the data flow and hit debug



Wait till success,

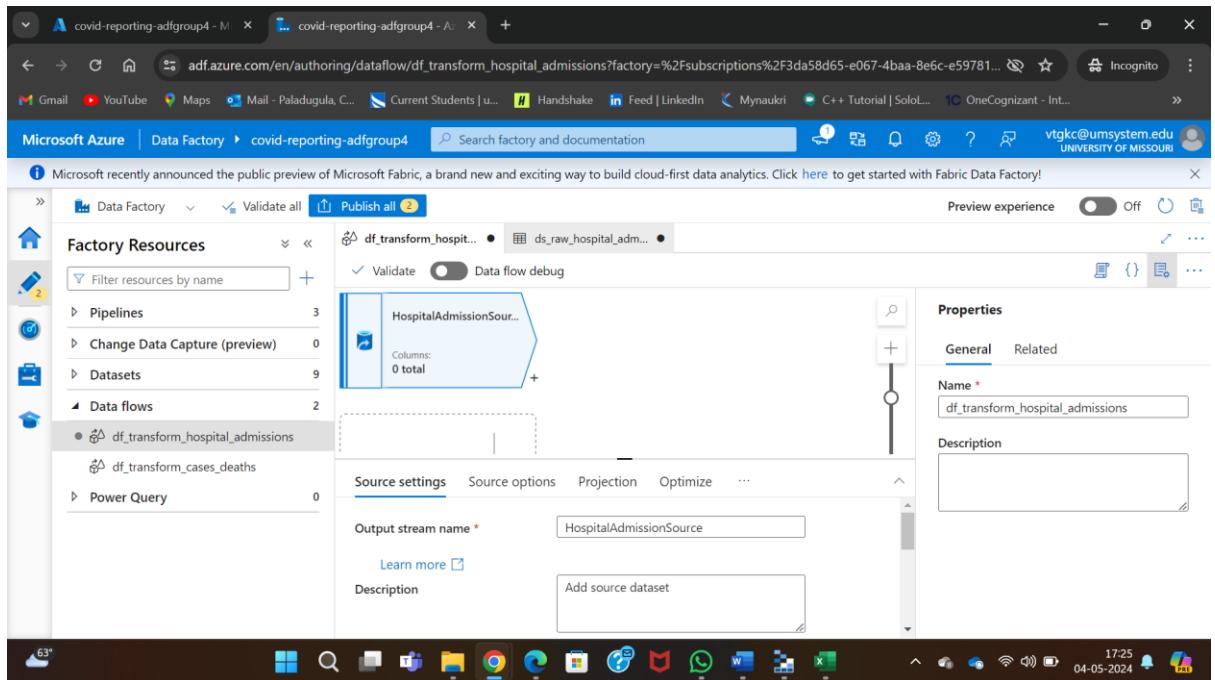
The screenshot shows the Microsoft Azure Data Factory interface. On the left, the navigation pane lists 'Factory Resources' including Pipelines, Datasets, and Data flows. The 'Activities' section is expanded, showing options like Copy data, Data flow, Synapse, and Azure Data Explorer. A specific pipeline named 'df\_transform\_cases\_and\_deaths' is selected. The main workspace displays the pipeline's structure with a data flow activity named 'df\_transform\_cases\_deaths'. The 'Output' tab is selected, showing a table of pipeline runs. One run is listed with the status 'Succeeded'. The properties panel on the right shows the pipeline's name as 'pl\_process\_cases\_and\_deaths\_data'.

## Pipeline preview,

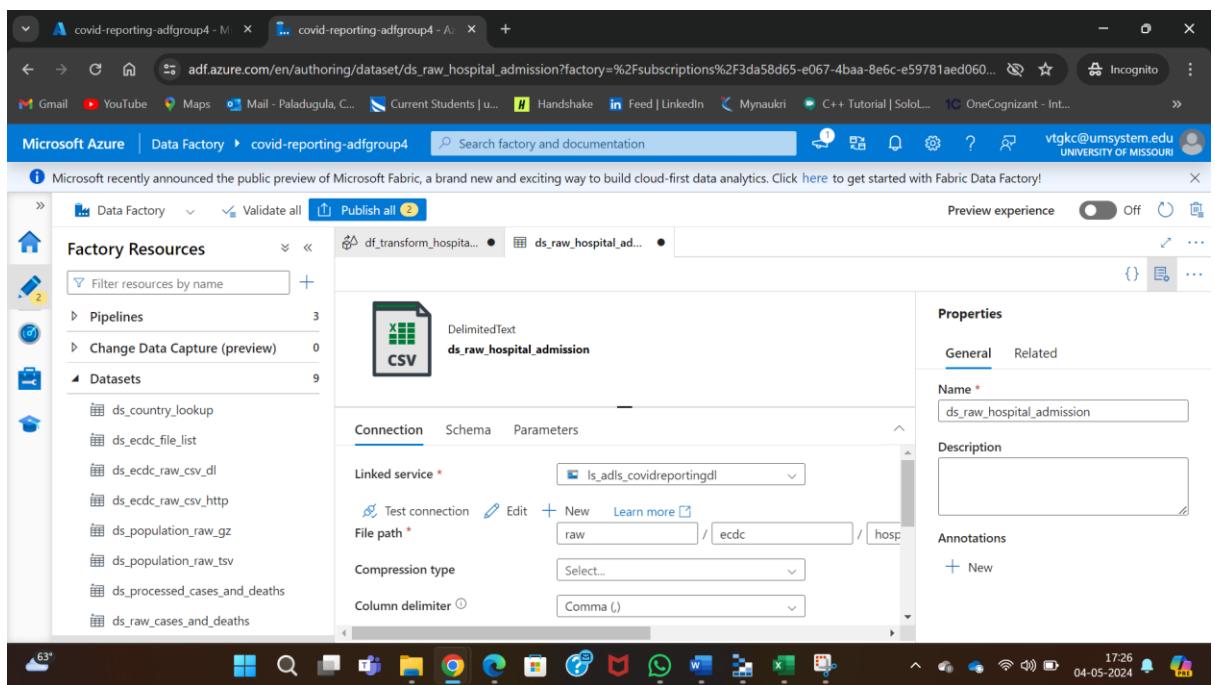
This screenshot shows the 'Pipeline runs' section of the Azure Data Factory interface. It highlights a specific run for the pipeline 'df\_transform\_cases\_deaths'. The run status is 'Succeeded'. Below the run details, the data flow visualization shows the flow of data from various sources through various stages of transformation. The 'Sink' table at the bottom provides a summary of the processing results, indicating a total of 16390 rows written.

## Data Flow – hospital admissions and data transformation

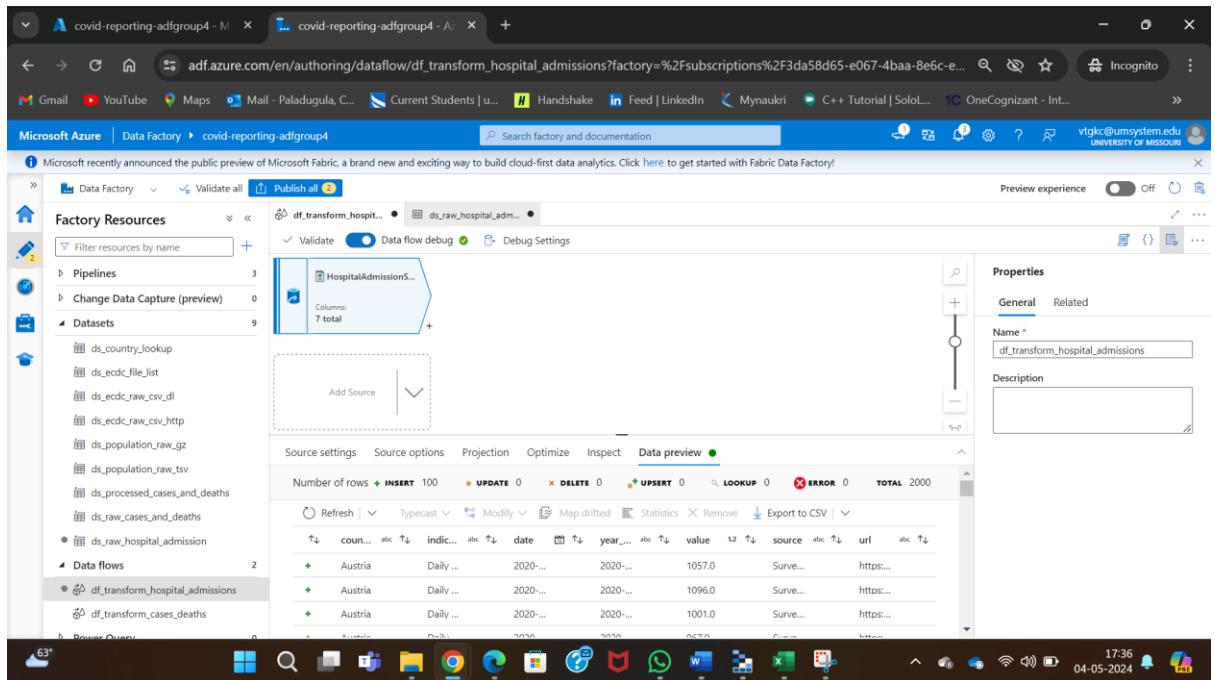
Create a new DataFlow, name – df\_transform\_hospital\_admissions



Create a new data set,

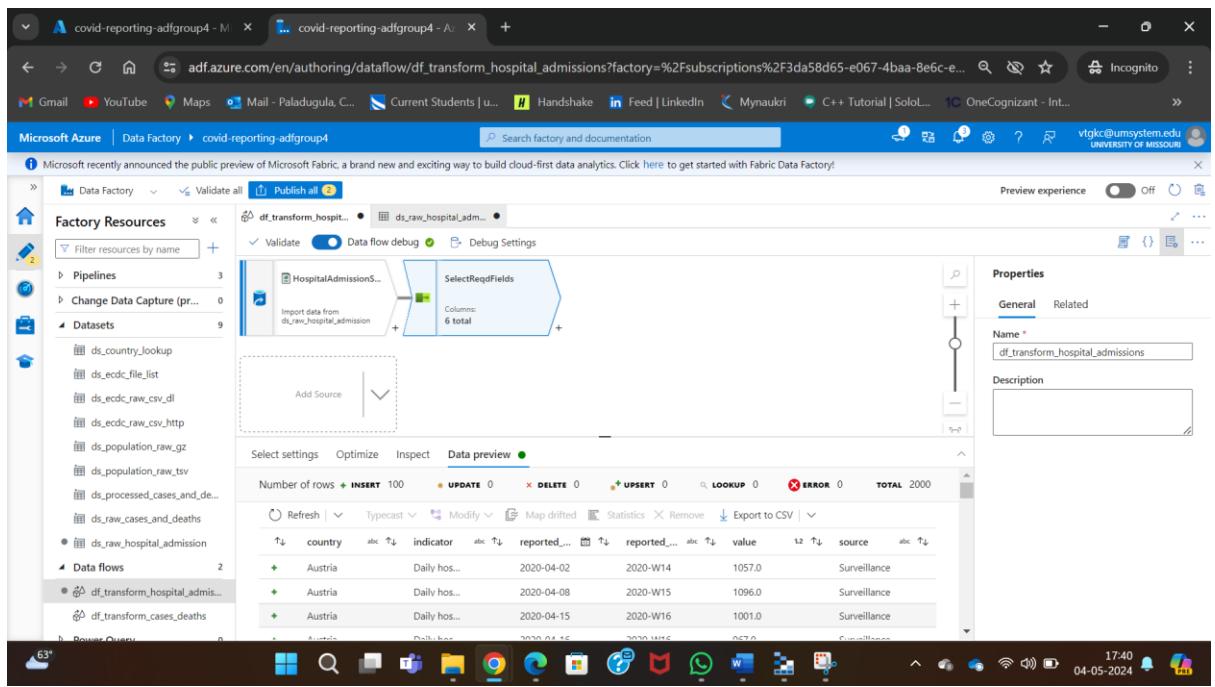


Create a source transformation, you can see the Data preview,

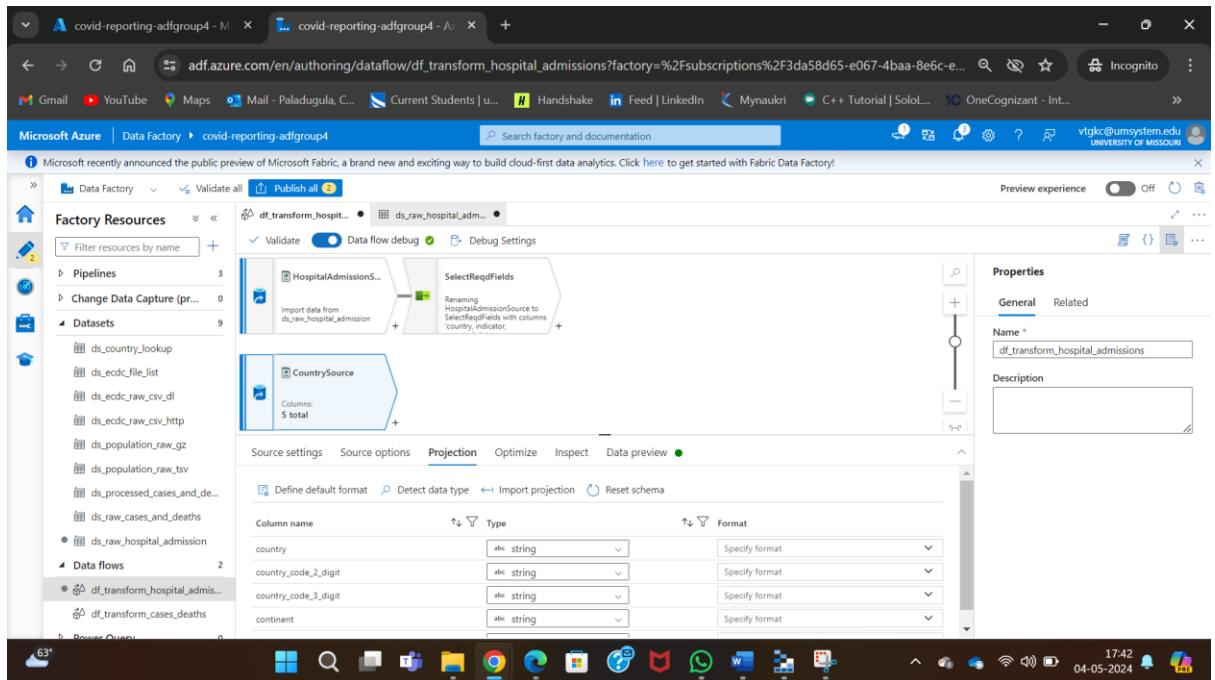


Create a select transformation to remove some fields, also to rename some fields,

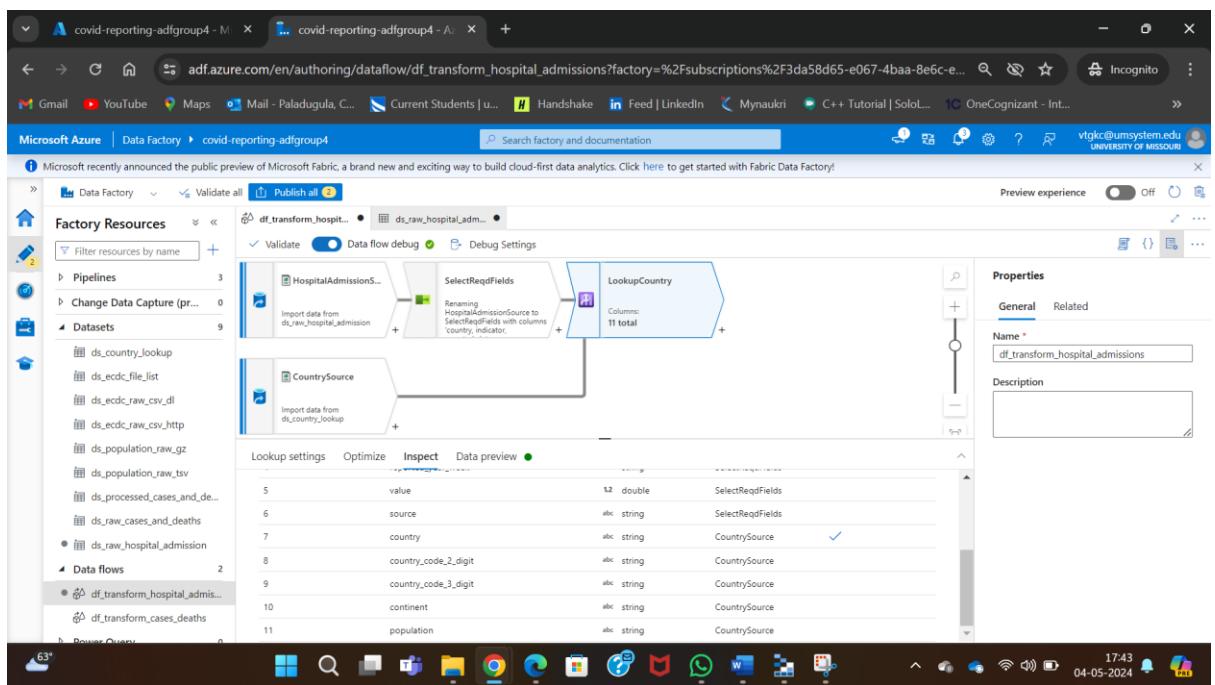
You can see the data preview the new fields and name changes



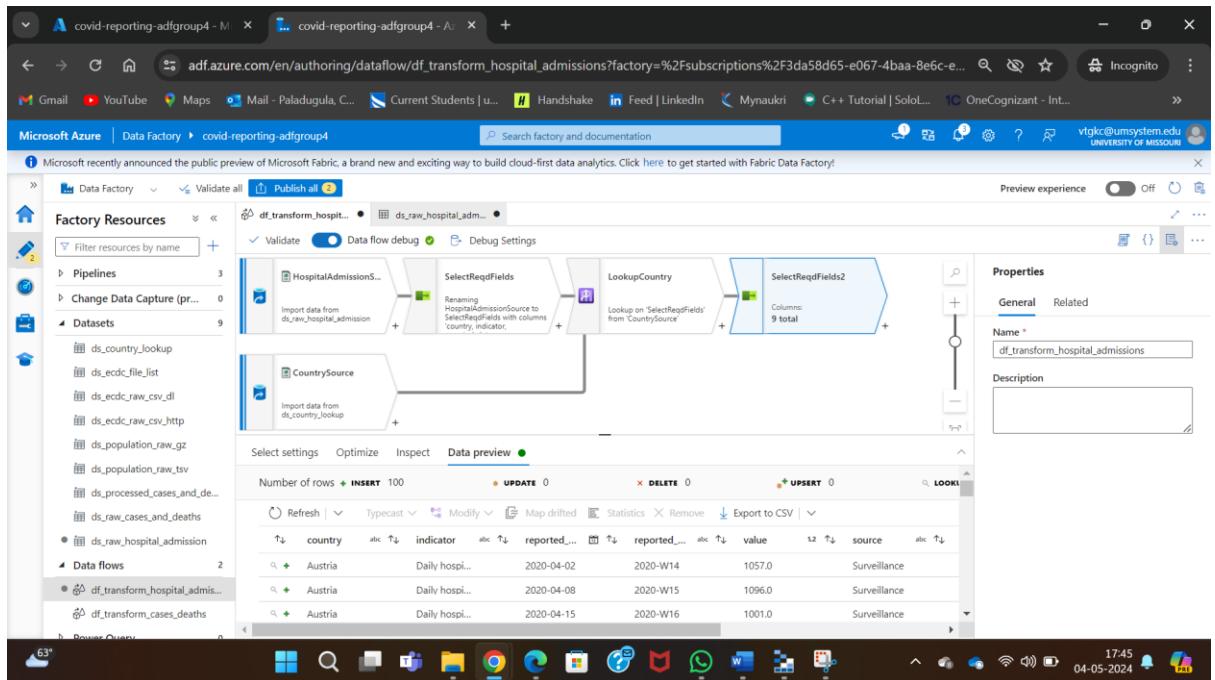
Create new source for the lookup ,



Now create the lookup,



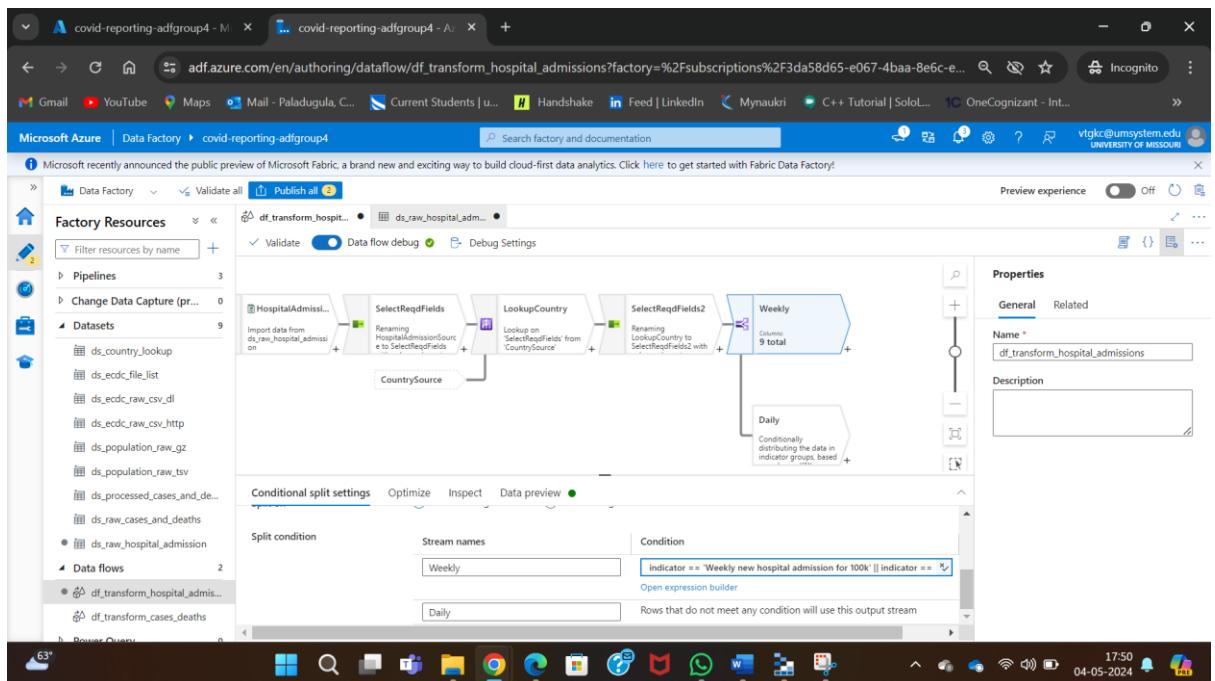
Create a new select to remove some fields, to remove the duplicate columns and continents



Now we split the data into 2 streams, one for the daily info and the other one for weekly,

We create a conditional split, here is the condition given for the split.

Weekly condition = "indicator == 'Weekly new hospital admissions per 100k' || indicator == 'Weekly new ICU admissions per 100k'"



Here is the data preview, for the week

The screenshot shows the Microsoft Azure Data Factory Data Flow preview interface. The pipeline 'df\_transform\_hospital\_admissions' is displayed with a 'Weekly' output stream. The preview pane shows three rows of data for Belgium:

country	indicator	reported...	reported...	value	source
Belgium	Weekly ne...	NULL	2020-W06	NULL	TESSy COV...
Belgium	Weekly ne...	NULL	2020-W07	NULL	TESSy COV...
Belgium	Weekly ne...	NULL	2020-W08	NULL	TESSy COV...

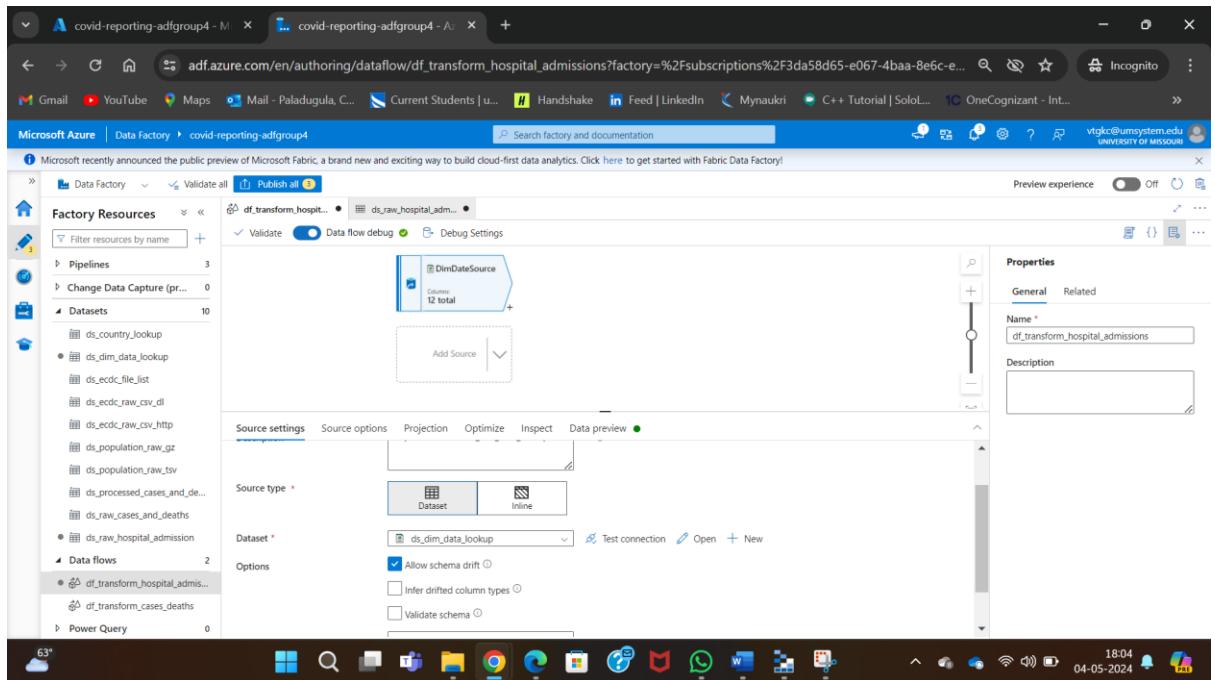
For the daily,

The screenshot shows the Microsoft Azure Data Factory Data Flow preview interface. The pipeline 'df\_transform\_hospital\_admissions' is displayed with a 'Daily' output stream. The preview pane shows three rows of data for Austria:

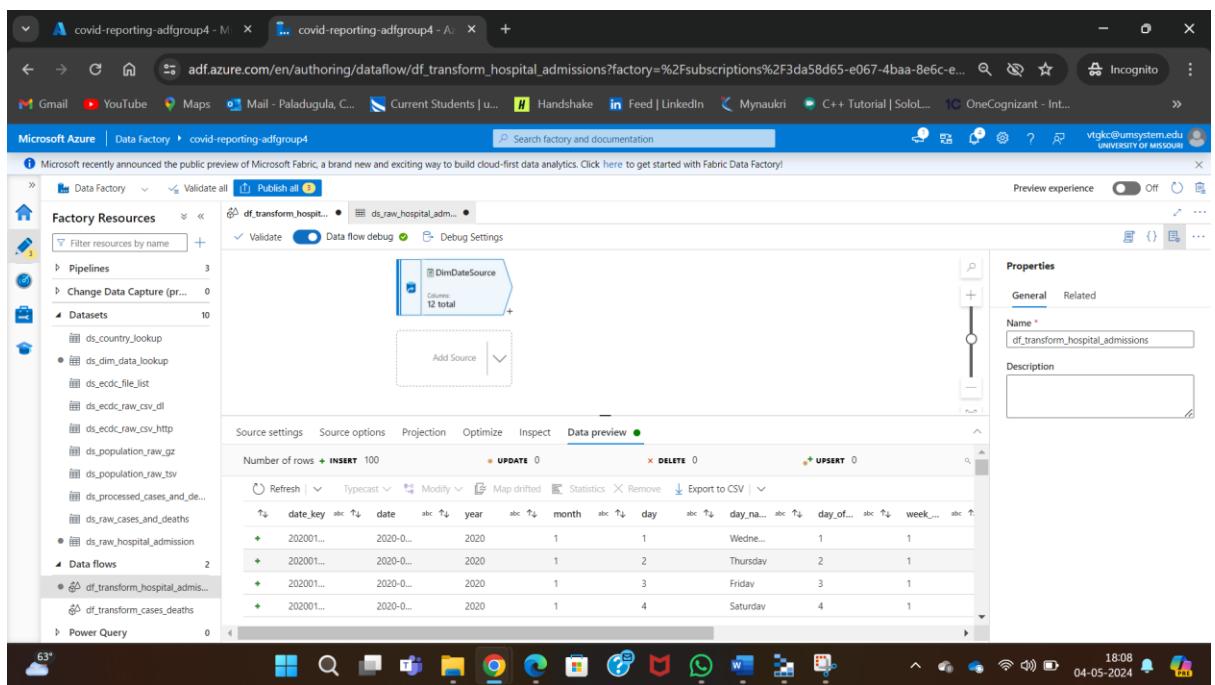
country	indicator	reported...	reported...	value	source
Austria	Daily hospit...	2020-04-02	2020-W14	1057.0	Surveillance
Austria	Daily hospit...	2020-04-08	2020-W15	1096.0	Surveillance
Austria	Daily hospit...	2020-04-15	2020-W16	1001.0	Surveillance

Create a Source transformation, for dim date

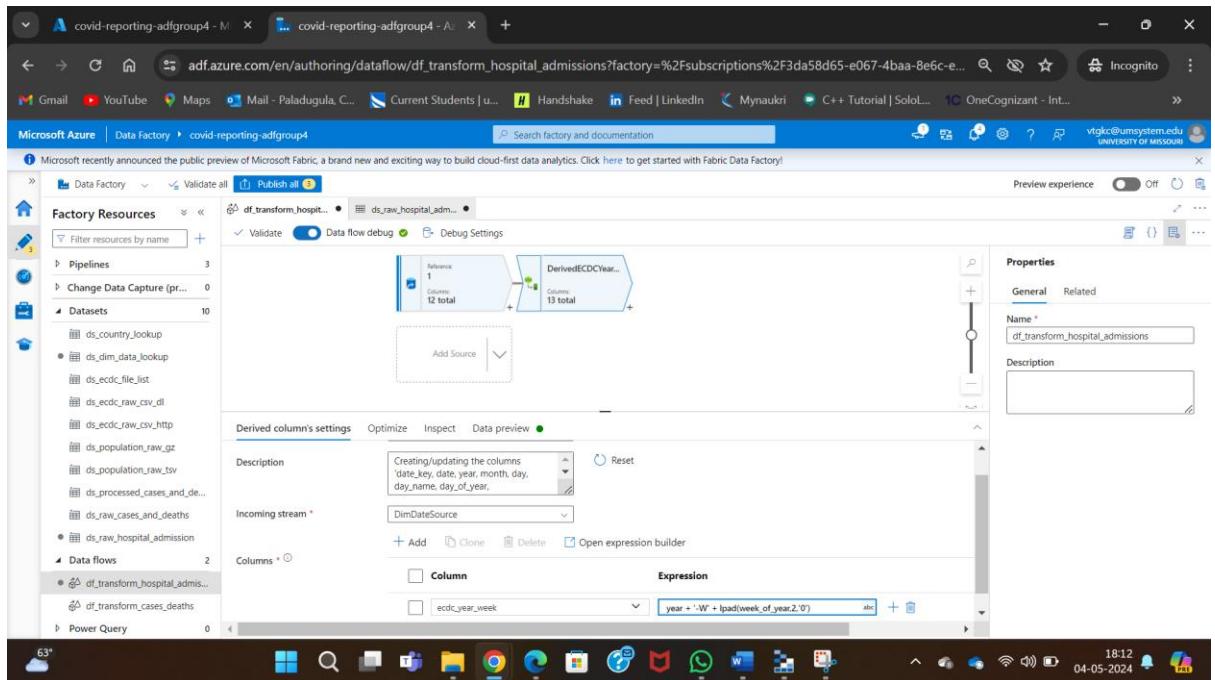
Create a new Source , and add a new dataset as well



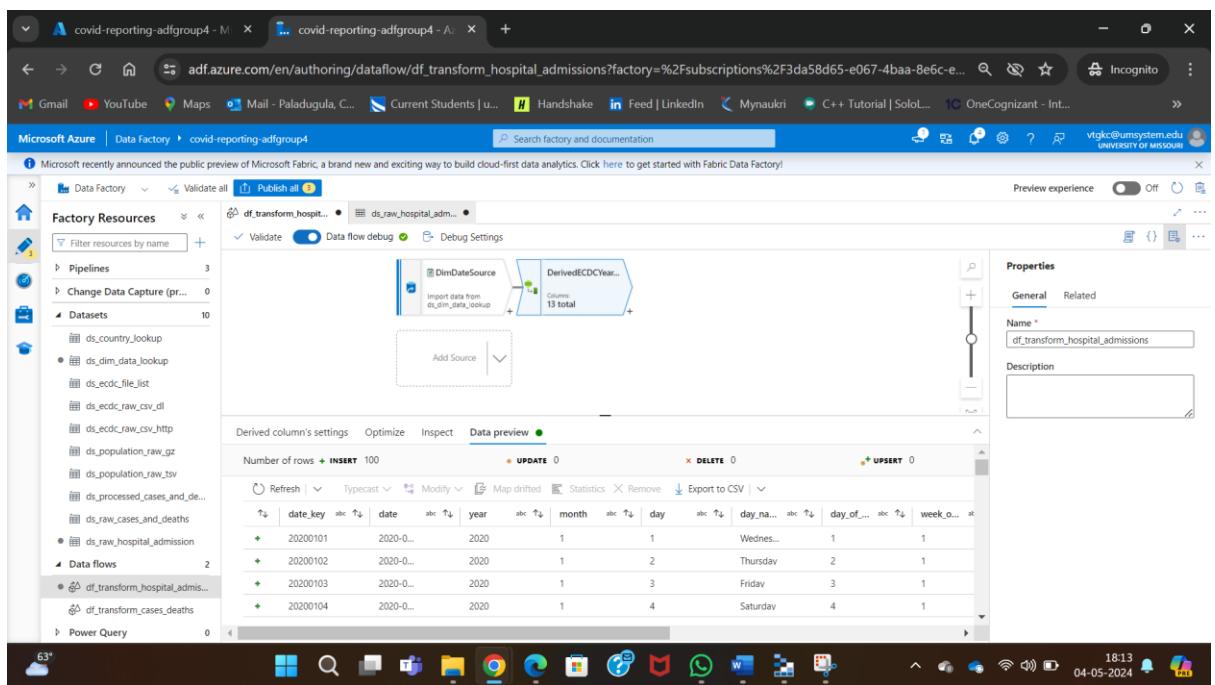
Here is the data preview,



Now we combine two columns (year , week of the year as ecdc\_year\_week) using derive transformation, you see the condition for combining them



Here is the data preview,



Now we create a single record to get one record for week,

Create a aggregate transformation, We used both group by and aggregate conditions,

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. The data flow named "df\_transform\_hospital\_admissions" is displayed. The flow starts with a "DimDataSource" step, followed by a "DerivedCCDCYear..." step with a condition "year + '-W' + lpad(week\_of\_year,2,'0')", and ends with an "AggDimDate" step. The schema table shows three columns: "ecdc\_year\_week" (Type: string, Group by), "week\_start\_date" (Type: string, Aggregate), and "week\_end\_date" (Type: string, Aggregate). The properties pane on the right shows the name is "df\_transform\_hospital\_admissions".

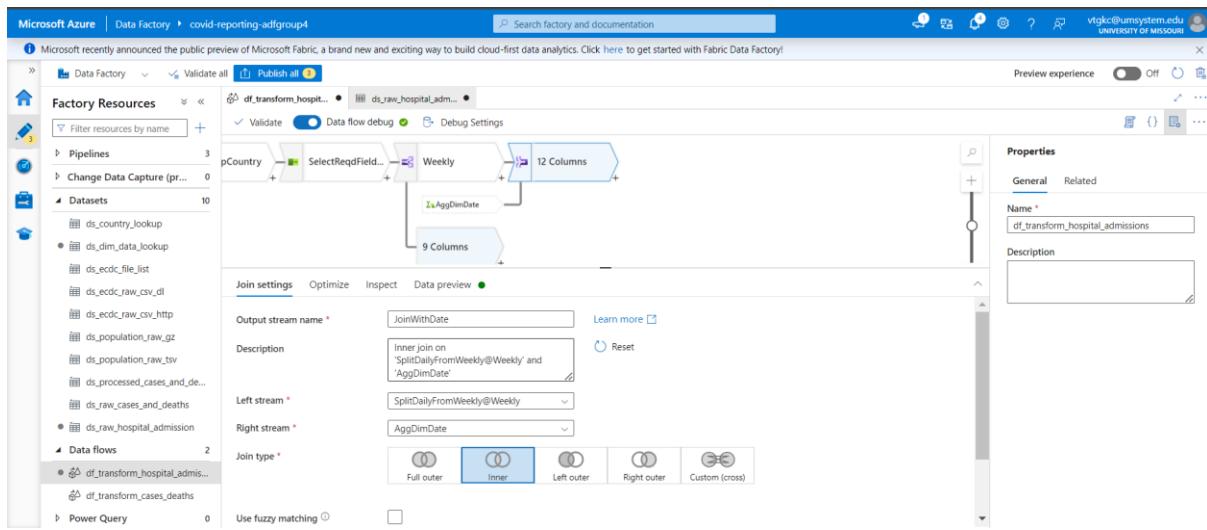
Here is the data preview, after performing the derived column condition in the aggregate condition, year + '-W' + lpad(week\_of\_year,2,'0')

We get the same results even after deleting the derived column,

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. The data flow named "df\_transform\_hospital\_admissions" is displayed. The flow starts with a "DimDataSource" step and ends with an "AggDimDate" step. The schema table shows three columns: "ecdc\_year\_week" (Type: string, Group by), "week\_start\_date" (Type: string, Aggregate), and "week\_end\_date" (Type: string, Aggregate). The properties pane on the right shows the name is "df\_transform\_hospital\_admissions".

Now we join the data from hospital weekly and dimdata source stream

## Now create join transformation

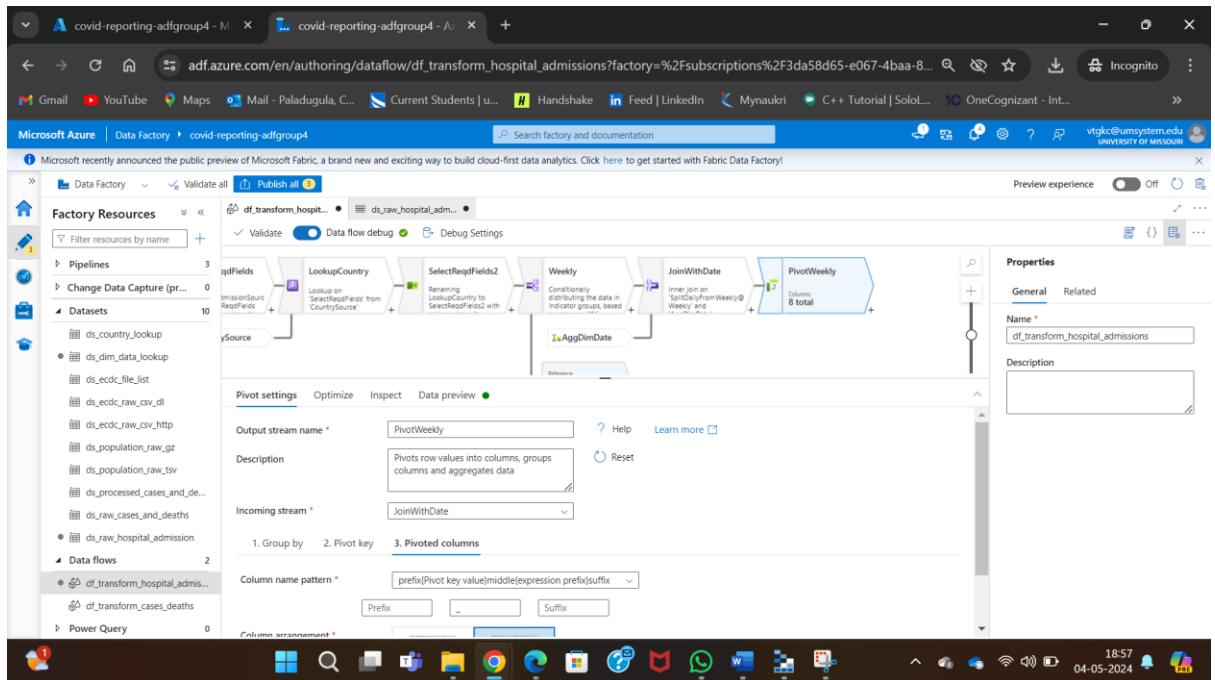


## data preview

Join settings		Optimize	Inspect	Data preview
★ UPSERT	0	🔍 LOOKUP	0	✖ ERROR 0
				TOTAL 144
source	abc ↑↓	country_co... abc ↑↓	country_co... abc ↑↓	population abc ↑↓
TESSy COVI...	CZ	CZE	10649800	2020-W05
TESSy COVI...	CY	CYP	875899	2020-W05
TESSy COVI...	CZ	CZE	10649800	2020-W06
TESSy COVI...	BE	BEL	11455519	2020-W06
TESSy COVI...	CZ	CZE	10649800	2020-W07
TESSy COVI...	BE	BEL	11455519	2020-W07

Now we perform pivot transformation, to combine daily and weekly streams

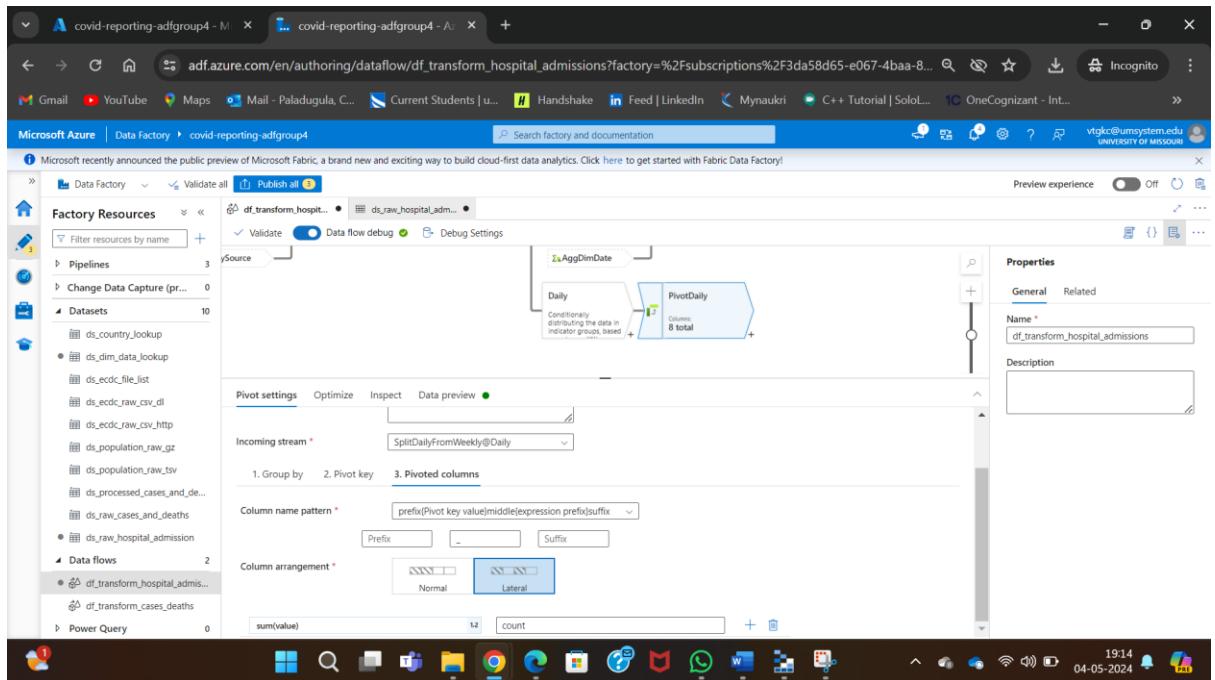
To get the hospital admission count from indicator and value.



Here is the data preview,

country	country_code	population	reported_year	source
Belgium	BE	11455519	2020-W06	TESSy COVID-...
Belgium	BE	11455519	2020-W07	TESSy COVID-...
Belgium	BE	11455519	2020-W08	TESSy COVID-...
Belgium	BE	11455519	2020-W09	TESSy COVID-...
Belgium	BE	11455519	2020-W10	TESSy COVID-...
Belgium	BE	11455519	2020-W11	TESSy COVID-...

Now we create another pivot for daily,

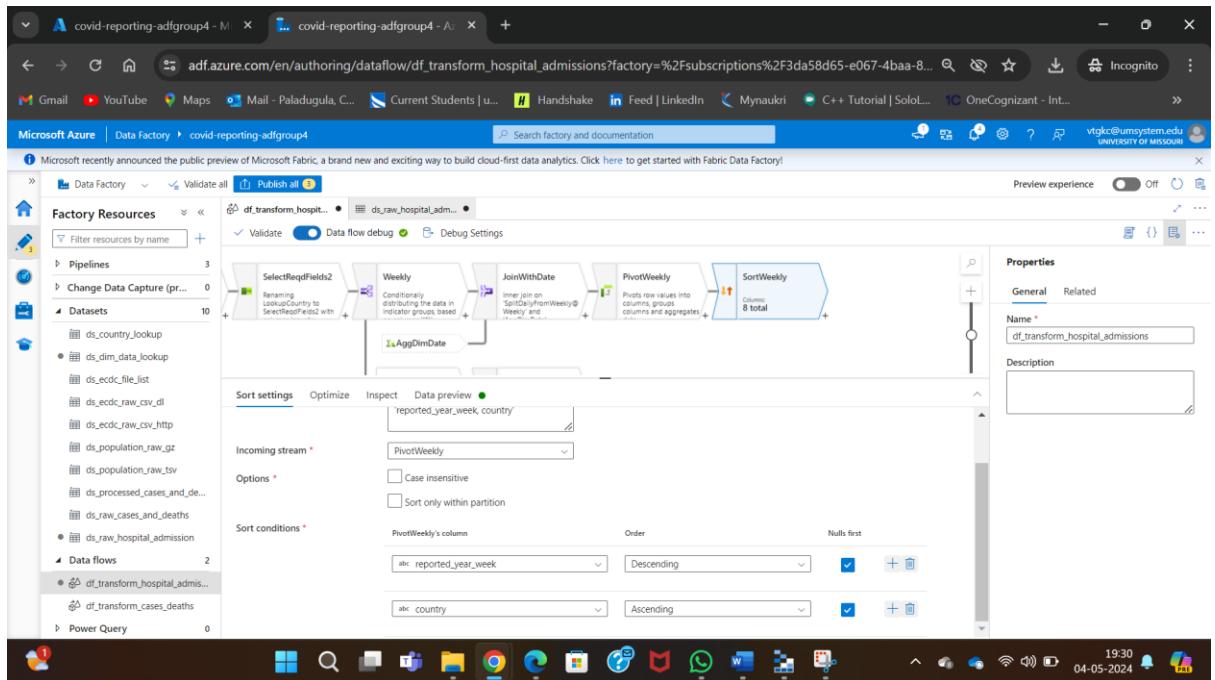


Here is the data preview,

The screenshot shows the Microsoft Azure Data Factory Data Flow blade with the same pipeline and configuration as the previous screenshot. The data preview pane now displays the full 100 rows of data from the "country" column:

country
Austria

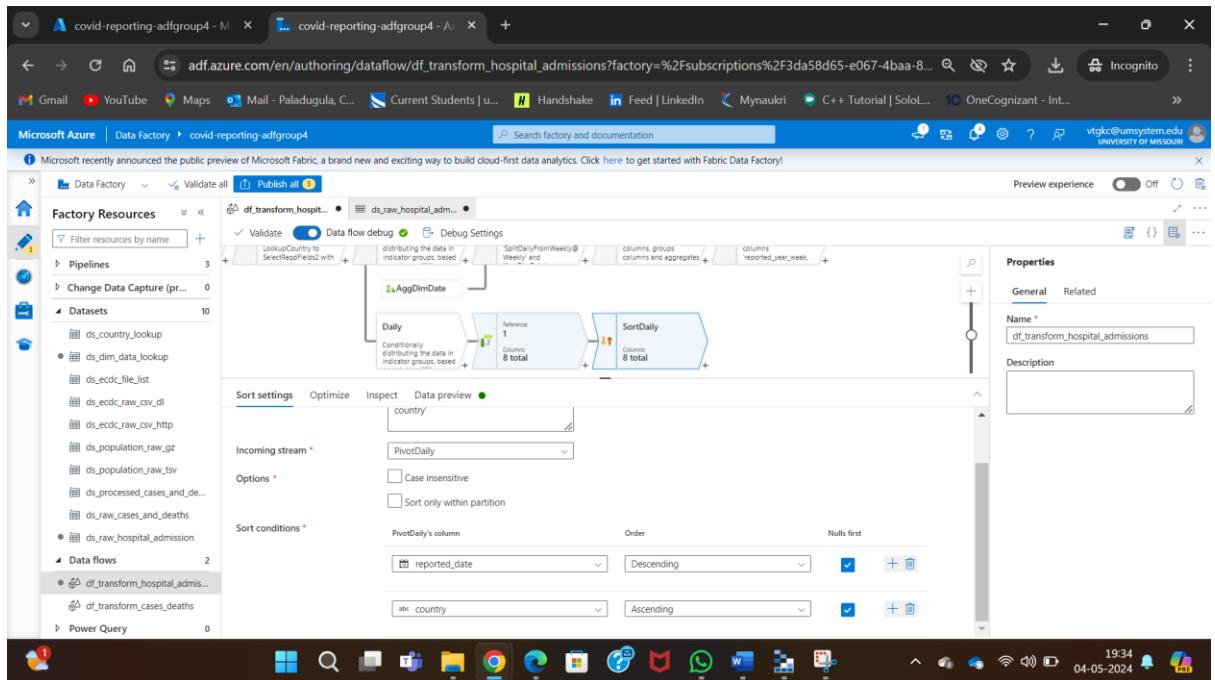
Now we create a sort transformation, To sort the records with country Ascending and reported year week descending, to recent date on the top.



Here is the data preview,

country	reported_year	source	population	reported_date	country_code
Belgium	BE	BEL	11455519	2020-W43	TESSy COVID-19
Belgium	BE	BEL	11455519	2020-W42	TESSy COVID-19
Belgium	BE	BEL	11455519	2020-W41	TESSy COVID-19
Belgium	BE	BEL	11455519	2020-W40	TESSy COVID-19
Belgium	BE	BEL	11455519	2020-W39	TESSy COVID-19
Belgium	BE	BEL	11455519	2020-W38	TESSy COVID-19

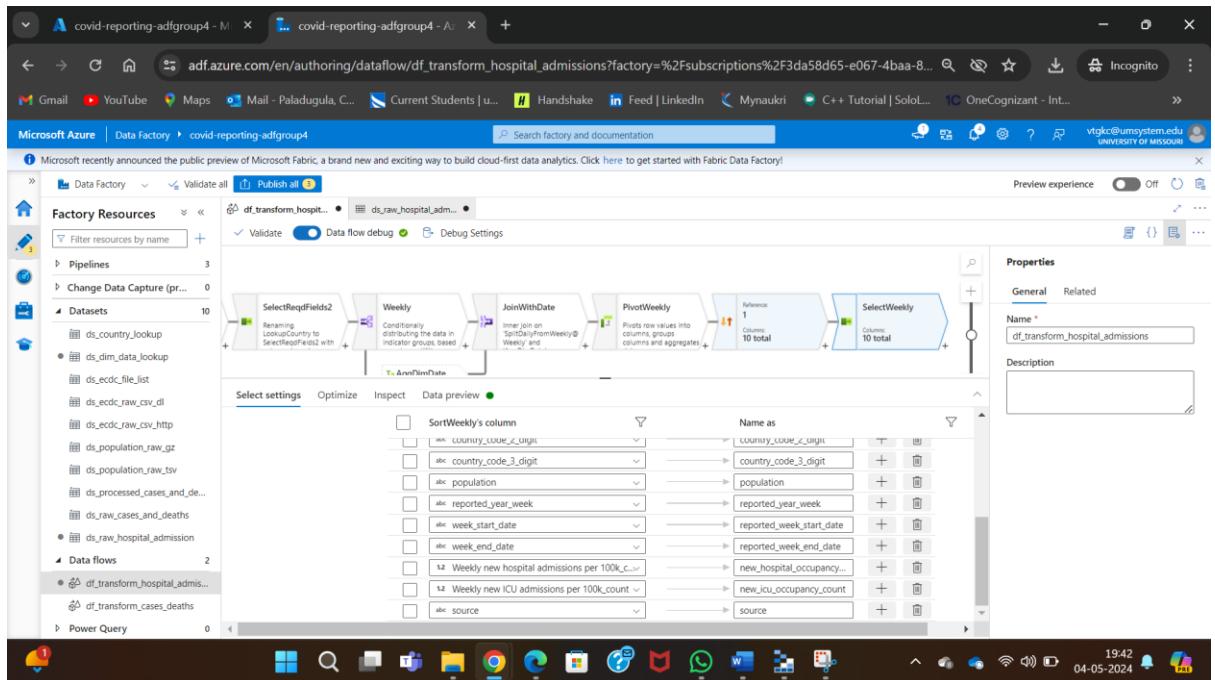
Now we perform the same for Daily,



Here is the data preview,

	country	country_code	population	reported_date	source
1	Austria	AT	8858775	2020-10-25	Country_Web...
2	Belgium	BE	11455519	2020-10-25	Country_Web...
3	Austria	AT	8858775	2020-10-24	Country_Web...
4	Belgium	BE	11455519	2020-10-24	Country_Web...
5	Austria	AT	8858775	2020-10-23	Country_Web...
6	Belgium	BE	11455519	2020-10-23	Country_Web...

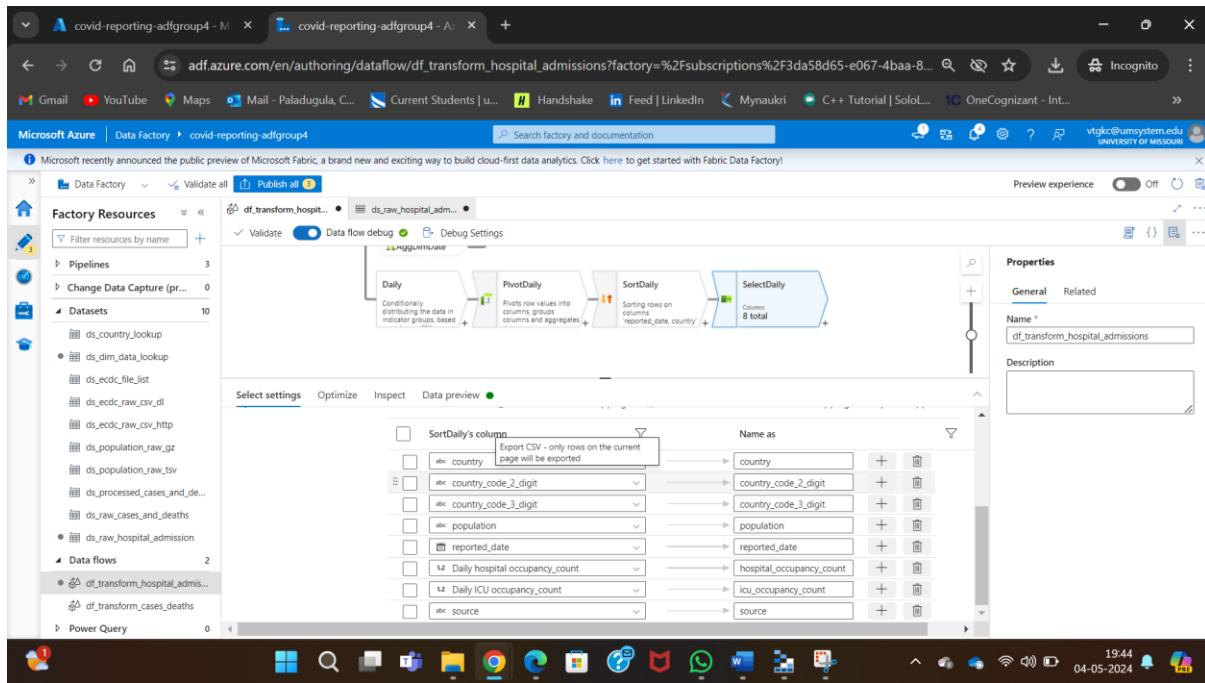
Now we perform Select and sink transformation,



Here is the data preview,

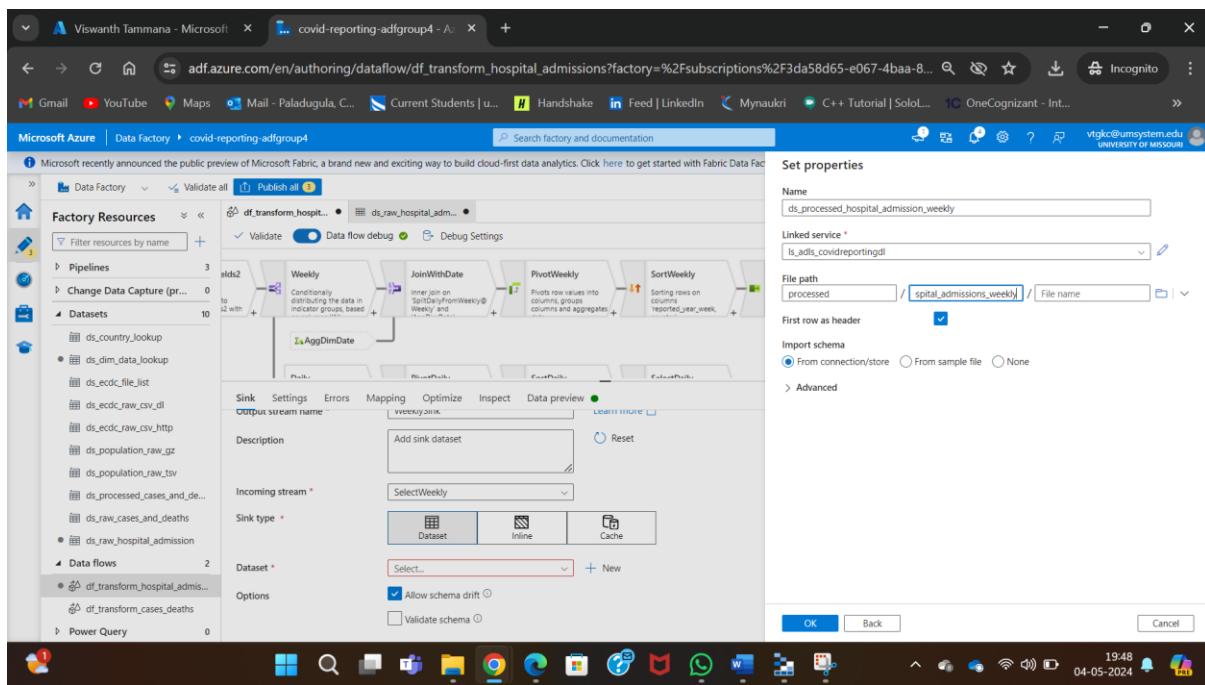
country	country_code_2_digit	country_code_3_digit	population	reported_year... week	reported_week_start_date	reported_week_end_date	new_hospital_occupancy... count	new_icu_occupancy... count	source
Belgium	BE	BEL	11455519	2020-W43	2020-10-18				
Belgium	BE	BEL	11455519	2020-W42	2020-10-11				
Belgium	BE	BEL	11455519	2020-W41	2020-10-04				
Belgium	BE	BEL	11455519	2020-W40	2020-09-27				
Belgium	BE	BEL	11455519	2020-W39	2020-09-20				
Belgium	BE	BEL	11455519	2020-W38	2020-09-13				

Now create the same select for daily records,

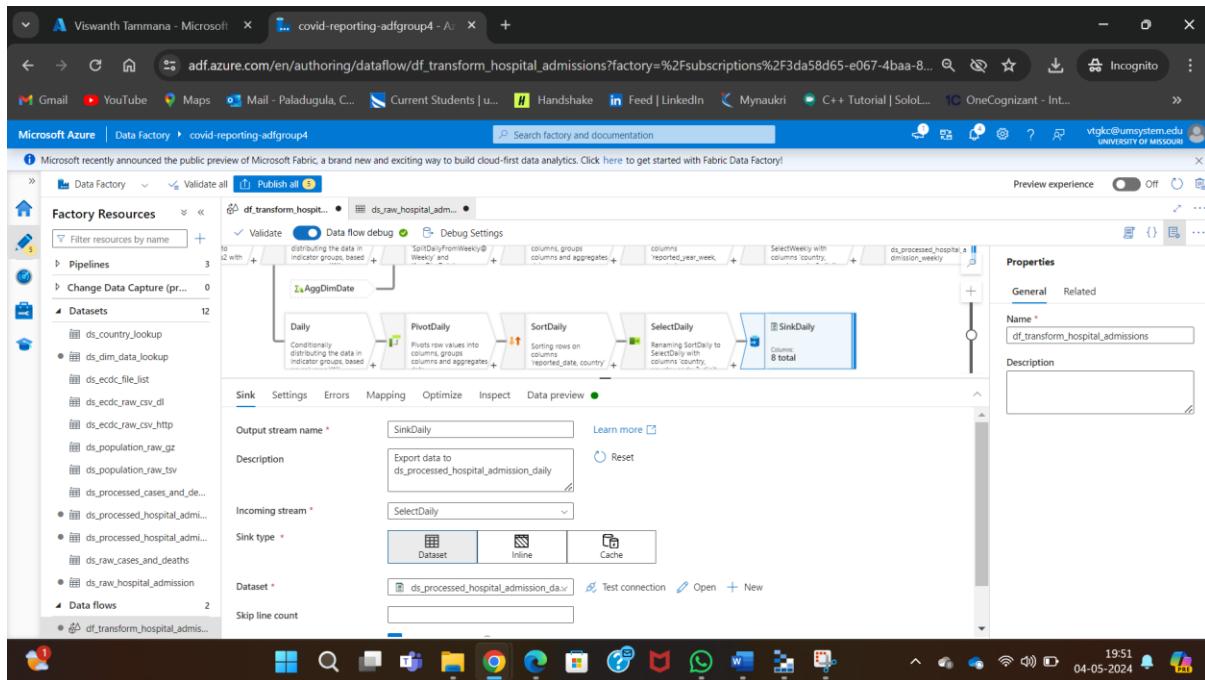


Now we create a new sink, along with a new dataset, dataset-

'ds\_processed\_hospital\_admission\_weekly'



Create the same for daily,



Here are the data preview for both weekly and daily,

Daily-

country	iso	country_code	reported_date	population	hospital_occ	icu_occ
Austria	AT	AUT	2020-10-25	8858775	1225.0	174.0
Belgium	BE	BEL	2020-10-25	11455519	4825.0	756.0
Austria	AT	AUT	2020-10-24	8858775	1177.0	175.0
Belgium	BE	BEL	2020-10-24	11455519	4408.0	694.0
Austria	AT	AUT	2020-10-23	8858775	1058.0	158.0
Belgium	BE	BEL	2020-10-23	11455519	4057.0	633.0

Weekly-

Microsoft Azure | Data Factory | covid-reporting-adfgroup4

df\_transform\_hospital\_admissions

Properties

Name: df\_transform\_hospital\_admissions

Description:

Data preview

country	country_code...	population	reported_year...	reported_week...
Belgium	BE	11455519	2020-W43	2020-10-18
Belgium	BE	11455519	2020-W42	2020-10-11
Belgium	BE	11455519	2020-W41	2020-10-04
Belgium	BE	11455519	2020-W40	2020-09-27
Belgium	BE	11455519	2020-W39	2020-09-20
Belgium	BE	11455519	2020-W38	2020-09-13

Now publish all the changes,

Now we have successfully created a dataflow we execute it from a pipeline,

New pipeline - pl\_processed\_hospital\_admissions\_data

Microsoft Azure | Data Factory | covid-reporting-adfgroup4

pl\_processed\_hospital\_admissions

Properties

Name: pl\_processed\_hospital\_admissions\_data

Description:

Annotations

Activities

- Move and transform
- Synapse
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDIInsight
- Iteration & conditionals
- Machine Learning
- Power Query

Use the existing dataflow,

The screenshot shows the Microsoft Azure Data Factory pipeline editor. On the left, the 'Factory Resources' sidebar lists Pipelines (4), Activities (12), Datasets (2), and Power Query (0). The main workspace displays a 'Data Flow' activity named 'Data flow1'. The 'Properties' panel on the right shows the pipeline name as 'pl\_processed\_hospital\_admissions\_data'. The 'Settings' tab is selected, showing the 'Data flow' dropdown set to 'df\_transform\_hospital\_admissions', 'Run on (Azure IR)' set to 'AutoResolveIntegrationRuntime', and 'Compute size' set to 'Custom'. The 'Advanced' section shows 'Compute type' as 'General purpose' and 'Core count' as '4 (+ 4 Driver cores)'. The status bar at the bottom indicates the date as 04-05-2024.

Now publish the pipeline,

Then we do the manual trigger by add trigger → trigger now.

Now navigate to monitor to see changes,

The screenshot shows the Microsoft Azure Data Factory Pipeline runs monitor. The left sidebar lists 'Runs' (1), 'Trigger runs', 'Change Data Capture (previ...', 'Runtimes & sessions', 'Integration runtimes', 'Data flow debug', and 'Notifications'. The main area displays a table of pipeline runs:

Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run	Parameters	Annotations	Ru
pl_processed_hospital_admiss...	5/4/2024, 8:06:09 PM	5/4/2024, 8:09:45 PM	3m 37s	Manual trigger	Succeeded	Original			ae
pl_processed_hospital_admiss...	5/4/2024, 8:01:02 PM	5/4/2024, 8:04:36 PM	3m 35s	Manual trigger	Succeeded	Original			3f
pl_ingroup_ecdc_data	5/4/2024, 7:25:00 PM	5/4/2024, 7:25:26 PM	26s	tr_ingest_ecdc_data	Succeeded	Original			12

The status bar at the bottom indicates the date as 04-05-2024.

All pipeline runs > pl\_processed\_hospital\_admissions\_data - Activity runs > df\_transform\_hospital\_admissions

**df\_transform\_hospital\_admissions**

Cluster startup time: 2m 29s Number of transformations: 17 Data flow status: Success

Refresh Auto refresh On Edit dataflow

Sinks All streams

Transform	Status	Time	Skew	Kurtosis
SelectReqFields	Succeeded	323ms	-	-
LookupCountry	Succeeded	-	13.7121	189.0156
SelectReqFields2	Succeeded	-	13.7121	189.0156
SplitDailyFromWeekly@Weekly	Succeeded	-	13.7121	189.0156
JoinWithDate	Succeeded	3s 723ms	13.7121	189.0156
PivotWeekly	Succeeded	114ms	-	-
SortWeekly	Succeeded	-	-	-

Now we use single partition in the sink to get the output file as a single file ,

Weekly sink file - hospital\_admissions\_weekly.csv

df\_transform\_hospital\_admissions

Validate all

Preview experience Off

Factory Resources

Pipelines

Data flows

Data sets

Power Query

df\_transform\_hospital\_admissions

Sink Settings Errors Mapping Optimize Inspect Data preview

Clear the folder

File name option Output to single file

Output to single file hospital\_admissions\_weekly.csv

Add dynamic content [Alt+Shift+D]

Quote All

Headers Enter expression...

Umask Owner R W X

Now we perform the same for daily sink as well,

Select single partition in both sortweekly and sortdaily,

Now we publish all the changes,

Then we manually trigger the pipeline,

Now we can see the files in the file explorer,

This is the daily records file,

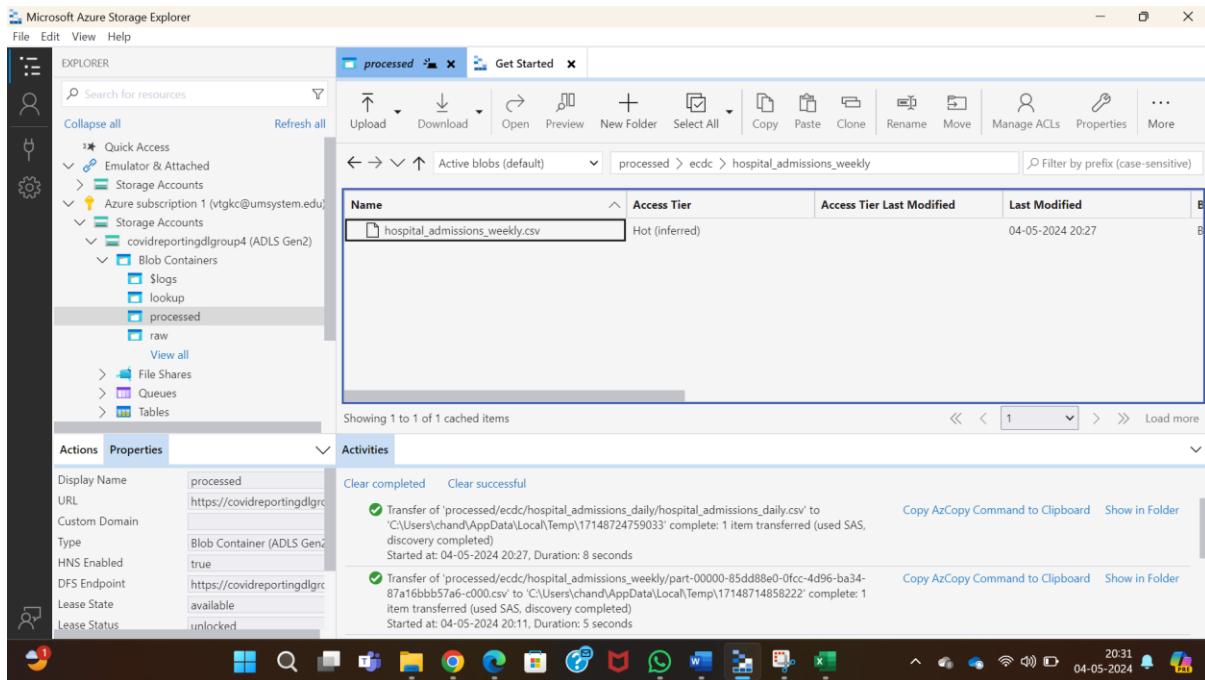
The screenshot shows the Microsoft Azure Storage Explorer interface. The left sidebar displays storage accounts and containers. The 'processed' container under 'ecdc' is selected. Inside 'processed', there is a blob named 'hospital\_admissions\_daily.csv'. The blob's properties show it has an inferred access tier and was last modified on 04-05-2024 20:27. The bottom pane shows activity logs for transfers involving this blob.

Here is the csv file,

The screenshot shows a Microsoft Excel spreadsheet titled 'hospital\_admissions\_daily.csv'. The data consists of 18 rows, each representing a country with its corresponding codes and counts. The columns are labeled: country, country\_code\_2\_digit, country\_code\_3\_digit, population, reported\_date, hospital\_occupancy\_count, icu\_occupancy\_count, and source. The data includes entries for Austria, Belgium, Bulgaria, Czechia, Denmark, Estonia, France, Hungary, Ireland, Italy, Latvia, Luxembourg, Netherlands, Portugal, Romania, Slovakia, and Slovenia.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	country	country_code_2_digit	country_code_3_digit	population	reported_date	hospital_occupancy_count	icu_occupancy_count	source					
2	Austria	AT	AUT	8858775	25-10-2020	1225	174	Country_Website					
3	Belgium	BE	BEL	11455519	25-10-2020	4825	756	Country_Website					
4	Bulgaria	BG	BGR	7000039	25-10-2020	1976	138	External_Github					
5	Czechia	CZ	CZE	10649800	25-10-2020	5613	828	Country_Website					
6	Denmark	DK	DNK	5806081	25-10-2020	127	18	Country_Website					
7	Estonia	EE	EST	1324820	25-10-2020	29	4	Country_API					
8	France	FR	FRA	67012883	25-10-2020	16454	2575	Country_Website					
9	Hungary	HU	HUN	9772756	25-10-2020	2449	221	JRC					
10	Ireland	IE	IRL	4904240	25-10-2020	319	39	Country_Website					
11	Italy	IT	ITA	60359546	25-10-2020	13214	1208	Country_Github					
12	Latvia	LV	LVA	1919968	25-10-2020	146	Country_Website						
13	Luxembourg	LU	LUX	613894	25-10-2020	114	12	Country_Website					
14	Netherlands	NL	NLD	17282163	25-10-2020		537	Other Website					
15	Portugal	PT	PRT	10276617	25-10-2020	1574	230	Country_Website					
16	Romania	RO	ROU	19414458	25-10-2020		828	Country_Website					
17	Slovakia	SK	SVK	5450421	25-10-2020	1117	JRC						
18	Slovenia	SI	SVN	2080908	25-10-2020	508	73	Country_Github					

This is the weekly records file,



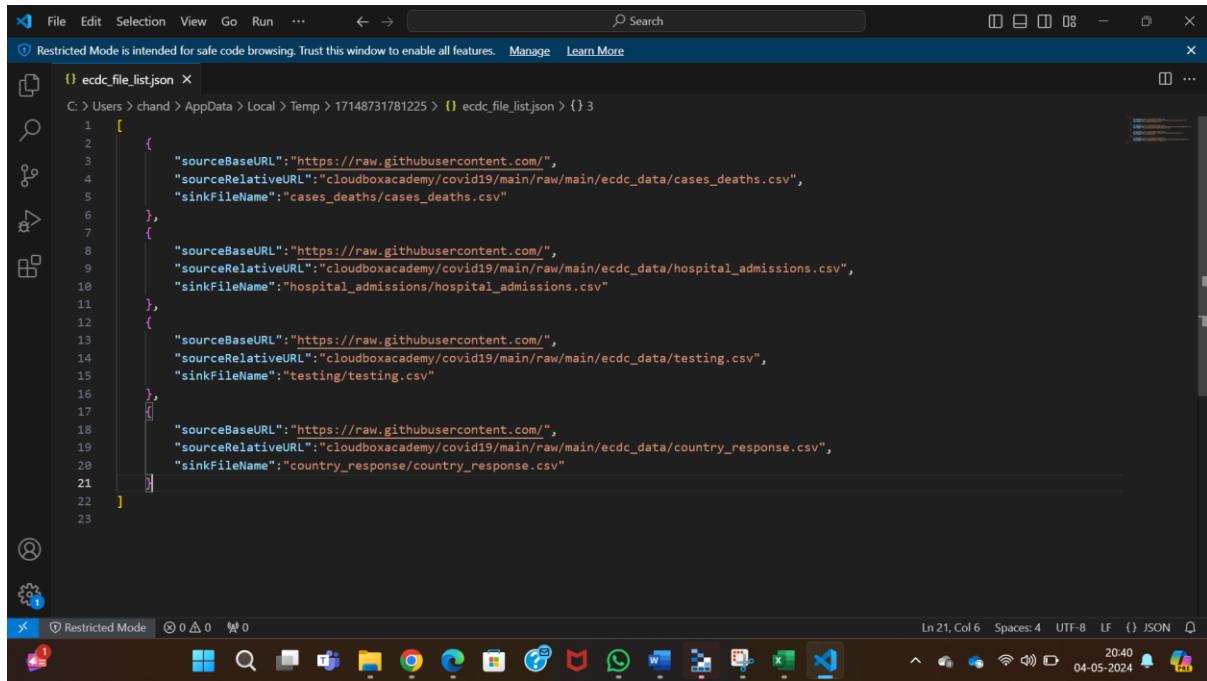
Here is the csv file,

	A	B	C	D	E	F	G	H	I
1	country	country_code_2_digit	country_code_3_digit	population	reported_year_week	reported_week_start_date	reported_week_end_date	new_hospital_occupancy_count	new_icu_occupancy_count
2	Belgium	BE	BEL	11455519	2020-W43	18-10-2020	24-10-2020	28.37933401	
3	Czechia	CZ	CZE	10649800	2020-W43	18-10-2020	24-10-2020	81.9545907	14.92985784
4	Denmark	DK	DNK	5806081	2020-W43	18-10-2020	24-10-2020	2.790178091	
5	Estonia	EE	EST	1324820	2020-W43	18-10-2020	24-10-2020		0.301927809
6	France	FR	FRA	67012883	2020-W43	18-10-2020	24-10-2020		2.70992668
7	Greece	EL	GRC	10724599	2020-W43	18-10-2020	24-10-2020		0.51283968
8	Iceland	IS	ISL	356991	2020-W43	18-10-2020	24-10-2020	14.84631265	
9	Ireland	IE	IRL	4904240	2020-W43	18-10-2020	24-10-2020	4.445133191	0.24468623
10	Latvia	LV	LVA	1919968	2020-W43	18-10-2020	24-10-2020	6.458440974	0.41667361
11	Netherlands	NL	NLD	17282163	2020-W43	18-10-2020	24-10-2020		1.868979016
12	Netherlands	NL	NLD	17282163	2020-W43	18-10-2020	24-10-2020	2.094645213	
13	Portugal	PT	PRT	10276617	2020-W43	18-10-2020	24-10-2020	2.38405304	
14	Romania	RO	ROU	19414458	2020-W43	18-10-2020	24-10-2020	136.0429428	
15	Slovenia	SI	SVN	2080908	2020-W43	18-10-2020	24-10-2020	7.59283928	
16	Spain	ES	ESP	46937060	2020-W43	18-10-2020	24-10-2020	6.406451533	0.32383792
17	Sweden	SE	SWE	10230185	2020-W43	18-10-2020	24-10-2020		0.27369984
18	United Kingdom	UK	GBR	66647112	2020-W43	18-10-2020	24-10-2020	10.25850903	

Prepare data for HD insights and Databricks

As we further do transformations in HDInsights and Spark we need the data to be in a folder, As they work on distributed computing. We need to make some changes in the JSON files,

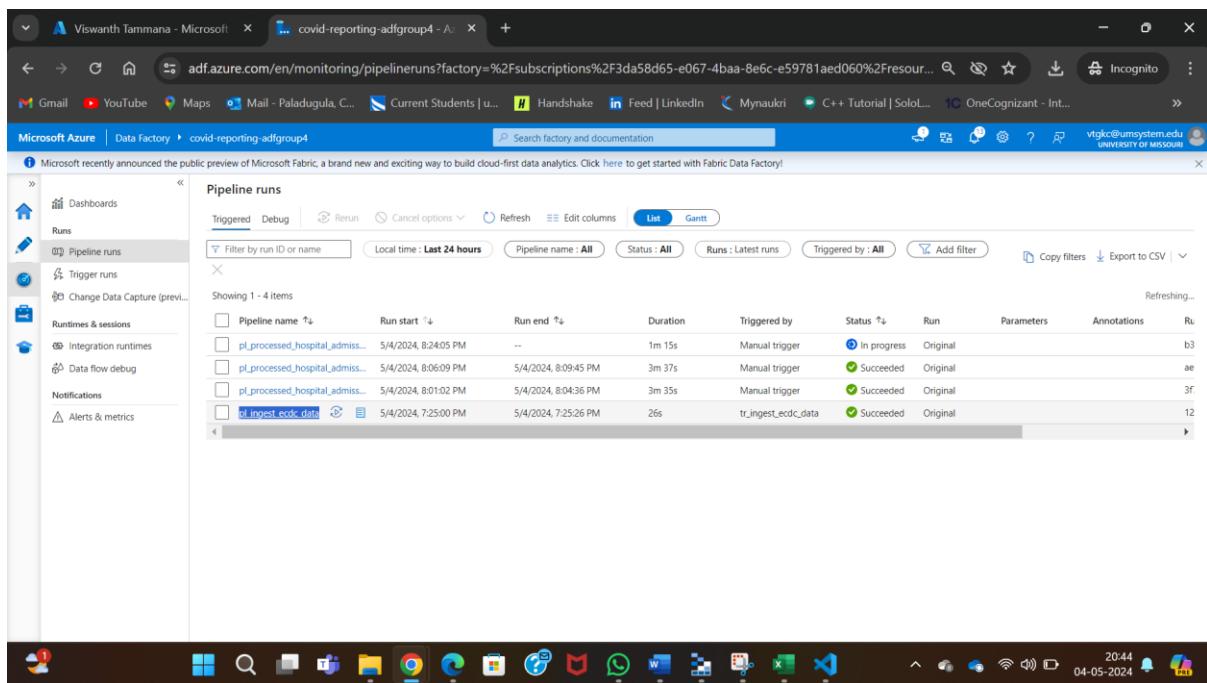
File path – covidreportingsagroup4 → blob containers → configs → ecdc\_file\_list.json



```
C:\> Users > chand > AppData > Local > Temp > 17148731781225 > {} ecdc_file_list.json > {} 3
[{"sourceBaseURL": "https://raw.githubusercontent.com/", "sourceRelativeURL": "cloubboxacademy/covid19/main/raw/main/ecdc_data/cases_deaths.csv", "sinkFileName": "cases_deaths/cases_deaths.csv"}, {"sourceBaseURL": "https://raw.githubusercontent.com/", "sourceRelativeURL": "cloubboxacademy/covid19/main/raw/main/ecdc_data/hospital_admissions.csv", "sinkFileName": "hospital_admissions/hospital_admissions.csv"}, {"sourceBaseURL": "https://raw.githubusercontent.com/", "sourceRelativeURL": "cloubboxacademy/covid19/main/raw/main/ecdc_data/testing.csv", "sinkFileName": "testing/testing.csv"}, {"sourceBaseURL": "https://raw.githubusercontent.com/", "sourceRelativeURL": "cloubboxacademy/covid19/main/raw/main/ecdc_data/country_response.csv", "sinkFileName": "country_response/country_response.csv"}]
```

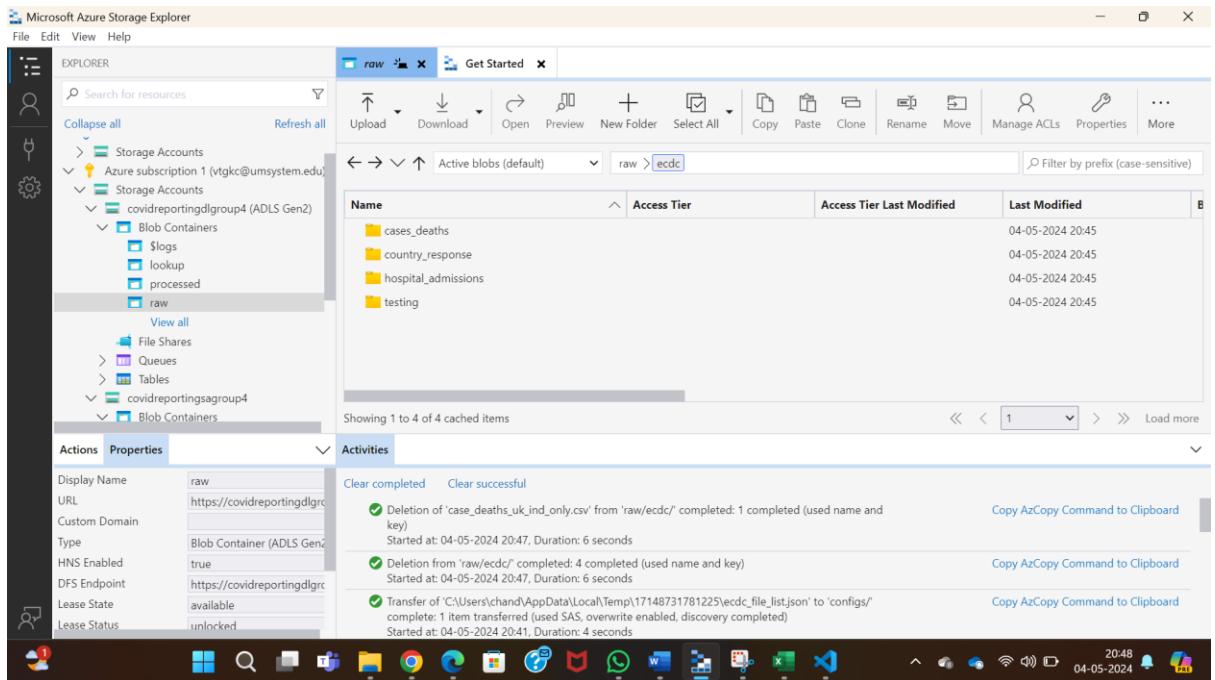
We change the file path here.

Then we rerun the pipeline from the monitor tab, name - pl\_ ingest\_ecdc\_data



Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run	Parameters	Annotations	Ru
pl_processed_hospital_admiss...	5/4/2024, 8:24:05 PM	--	1m 15s	Manual trigger	In progress	Original			b3
pl_processed_hospital_admiss...	5/4/2024, 8:06:09 PM	5/4/2024, 8:09:45 PM	3m 37s	Manual trigger	Succeeded	Original			ae
pl_processed_hospital_admiss...	5/4/2024, 8:01:02 PM	5/4/2024, 8:04:36 PM	3m 35s	Manual trigger	Succeeded	Original			3f
pl_ ingest_ecdc_data	5/4/2024, 7:25:00 PM	5/4/2024, 7:25:26 PM	26s	tr_ ingest_ecdc_data	Succeeded	Original			12

Now in explorer we see the changes in the ecdc folder,



Then we perform the location change in the dataset,

We do the same with all the datasets,

Then we publish all the changes,

Now trigger the two pipelines and monitor

The screenshot shows the Microsoft Azure Data Factory interface. The left sidebar has a 'Pipeline runs' section selected. The main area displays a table of pipeline runs with the following columns: Pipeline name, Run start, Run end, Duration, Triggered by, Status, Run, Parameters, Annotations, and Run ID. The table shows six runs, all triggered manually, with various statuses like In progress, Succeeded, and Rerun (Latest). The last refresh was 0 minutes ago.

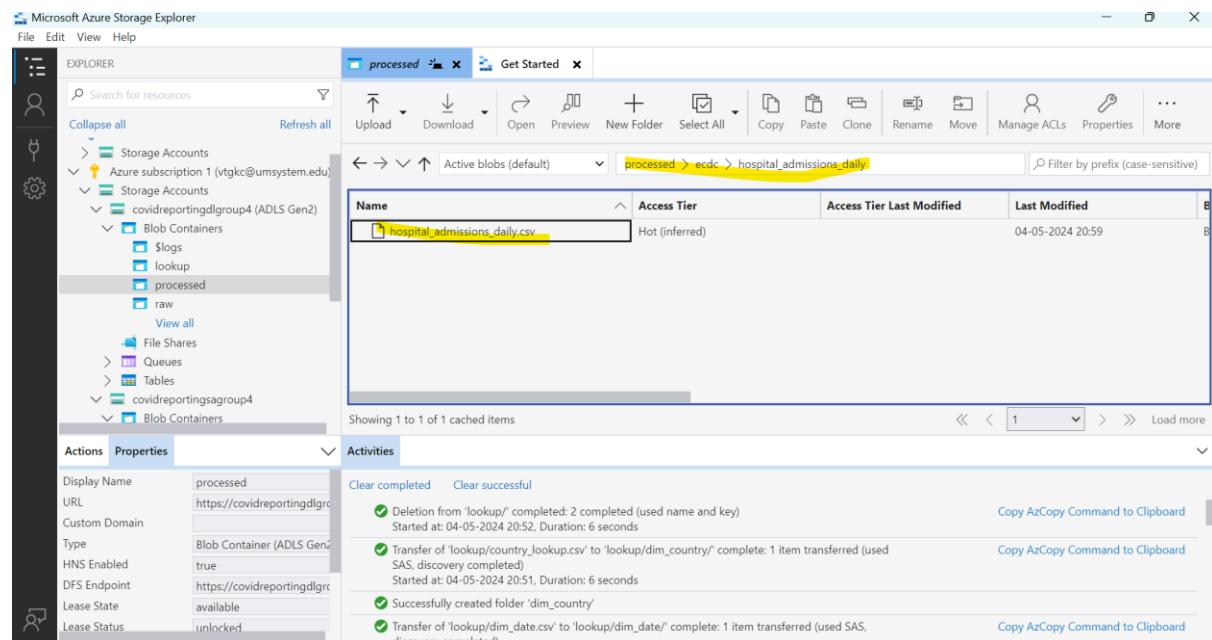
Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run	Parameters	Annotations	Run ID
pl_processed_hospital...	5/4/2024, 8:55:43 PM	--	9s	Manual trigger	In progress	Original			42
pl_process_cases_and...	5/4/2024, 8:55:32 PM	--	20s	Manual trigger	In progress	Original			60
> pl_ingest_ecdc_data	5/4/2024, 8:44:55 PM	5/4/2024, 8:45:18 PM	23s	Manual trigger	Succeeded	Rerun (Latest)			7e
pl_processed_hospital...	5/4/2024, 8:24:05 PM	5/4/2024, 8:27:35 PM	3m 31s	Manual trigger	Succeeded	Original			b3
pl_processed_hospital...	5/4/2024, 8:06:09 PM	5/4/2024, 8:09:45 PM	3m 37s	Manual trigger	Succeeded	Original			ae
pl_processed_hospital...	5/4/2024, 8:01:02 PM	5/4/2024, 8:04:36 PM	3m 35s	Manual trigger	Succeeded	Original			3f

Once done, make sure that the changes are seen in the explorer.

navigate to the path mentioned below and now you can see the updated file over there.

The screenshot shows the Microsoft Azure Storage Explorer interface. The left sidebar shows storage accounts and containers. The main area shows a blob container named 'processed' with a single file 'case\_and\_deaths.csv'. The bottom pane shows activities log with several successful operations listed, such as deletion from 'lookup/' completed, transfer of 'dim\_country' folder, and transfer of 'dim\_date.csv' to 'dim\_dim\_date/'. The status bar at the bottom right shows the date as 04-05-2024 and time as 20:59.

we can see both the daily and weekly data is also updated



The screenshot shows the Microsoft Azure Storage Explorer interface. In the left sidebar, under 'EXPLORER', there are sections for 'Storage Accounts', 'Azure subscription 1 (vtgkc@umsystem.edu)', and 'covidreportingdlgroup4 (ADLS Gen2)'. Under 'covidreportingdlgroup4', there are 'Blob Containers' like '\$logs', 'lookup', 'processed' (which is selected), and 'raw'. Below these containers are 'File Shares', 'Queues', and 'Tables'. The 'processed' container is expanded, showing subfolders '\$logs', 'lookup', and 'raw'. The 'raw' folder contains a file named 'hospital\_admissions\_daily.csv'. The main pane shows a table with one item: 'Name' (hospital\_admissions\_daily.csv), 'Access Tier' (Hot (inferred)), 'Access Tier Last Modified' (04-05-2024 20:59), and 'Last Modified' (04-05-2024 20:59). At the bottom of the main pane, it says 'Showing 1 to 1 of 1 cached items'. The bottom section, 'Activities', lists four completed tasks: 'Deletion from 'lookup/' completed: 2 completed (used name and key)' (started at 04-05-2024 20:52, duration 6 seconds), 'Transfer of 'lookup/country\_lookup.csv' to 'lookup/dim\_country/' complete: 1 item transferred (used SAS, discovery completed)' (started at 04-05-2024 20:51, duration 6 seconds), 'Successfully created folder 'dim\_country'' (started at 04-05-2024 20:51, duration 1 second), and 'Transfer of 'lookup/dim\_date.csv' to 'lookup/dim\_date/' complete: 1 item transferred (used SAS, discovery completed)' (started at 04-05-2024 20:51, duration 1 second). Each task has a 'Copy AzCopy Command to Clipboard' link next to it.

Copy the data to Azure SQL:

Till now we have transformed the data using dataflow, HDInsight and databricks. The transformed data is written back to the dtaalake. No we need to copy this data back to the SQL data. Below 3 data are to be copied to azure sql.

- Cases and deaths
- Hospital admissions
- Testing

For that we need to create SQL

Navigate to covid database

Upload the sql file. (Click on open query)

The screenshot shows the Microsoft Azure portal interface for a SQL database named 'covid-dbgroup4'. In the left sidebar, under 'Intelligent Performance' and 'Query Performance Insight', there is a note: 'Showing limited object explorer here. For full capability please click here to open Azure Data Studio.' Below this are links for 'Tables', 'Views', and 'Stored Procedures'. The main area contains two tabs: 'Query 1' and 'Query 2'. The 'Query 1' tab is active and displays the following T-SQL code:

```
1 CREATE SCHEMA covid_reporting
2 GO
3
4 CREATE TABLE covid_reporting.cases_and_deaths
5 (
6     country           VARCHAR(100),
7     country_code_2_digit  VARCHAR(2),
8     country_code_3_digit  VARCHAR(3),
9     population        BIGINT,
10    cases_count       BIGINT,
```

Query succeeded, created tables

The screenshot shows the Microsoft Azure portal interface for a SQL database named 'covid-dbgroup4'. The left sidebar shows the newly created 'covid\_reporting.cases\_and\_deaths' table with columns: country (varchar, null), country\_code\_2\_digit (varchar, null), country\_code\_3\_digit (varchar, null), population (bigint, null), cases\_count (bigint, null), deaths\_count (bigint, null), reported\_date (date, null), and source (varchar, null). The main area contains two tabs: 'Query 1' and 'Query 2'. The 'Query 1' tab is active and displays the same T-SQL code as before. The 'Results' tab in the bottom right shows the message: 'Query succeeded: Affected rows: 0Affected rows: 0Affected rows: 0Affected rows: 0'. A small window in the bottom right corner shows a Word document titled 'CC\_Progress\_2.docx'.

Create a new pipeline pl\_sqlize\_cases\_and\_deaths\_data alongwith new linked service

## Pipeline created pl\_sqlize\_cases\_and\_deaths

Succesfully running commands on the table covid\_reporting.cases\_and\_deaths

The screenshot shows the Microsoft Azure SQL database Query editor (preview) interface. The left sidebar displays the database structure with tables like 'covid\_reporting.cases\_and\_deaths' and 'covid\_reporting.hospital\_admissions'. The main area shows a query result for 'Query 3' with the following data:

country	country_code_2_digit	country_code_3_digit	population	cases_count	deaths_count	reported_date	source
Bosnia and Herzego...	BA	BIH	3280815	9	0	2020-05-31	Epidemi...
Bosnia and Herzego...	BA	BIH	3280815	0	0	2020-10-12	Epidemi...
Armenia	AM	ARM	2963234	134	2	2020-09-07	Epidemi...
Bosnia and Herzego...	BA	BIH	3280815	69	1	2020-05-05	Epidemi...
Bosnia and Herzego...	BA	BIH	3280815	0	0	2020-02-10	Epidemi...

## COPY activity Hospital admissions data

Pipeline created for hospital admissions daily data.

The screenshot shows the Microsoft Azure Data Factory pipeline details for 'covid-reporting-adfgroup4'. It displays a successful copy activity run from 'Azure Data Lake Storage Gen2' to 'Azure SQL Database'. The activity summary shows:

- Activity run id: d27a9709-07d2-4d75-b9c9-dff31ec6c9ee
- Region: UK South
- Data read: 292,018 KB
- Files read: 1
- Rows read: 5,006
- Peak connections: 1
- Data written: 418.94 KB
- Rows written: 5,006
- Peak connections: 2
- Copy duration: 00:00:14
- Throughput: 58.404 KB/s

The properties panel on the right shows the pipeline name as 'pl\_sqlize\_hospital\_admissions\_daily\_data'.

Successfully fetched data from hospital admissions data

The screenshot shows the Microsoft Azure SQL database Query editor (preview). The query being run is:

```
1 SELECT * FROM covid_reporting.hospital_admissions_daily
```

The results table shows data from the hospital\_admissions\_daily table, with columns including country, country\_code\_2\_digit, country\_code\_3\_digit, population, reported\_date, hospital\_occupancy..., icu\_occupancy\_count, and source. The data includes entries for Austria, Belgium, Bulgaria, Czechia, and Denmark.

## Copy activity Testing Data

Created gen 2 dataset and sql dataset for testing ds\_processed\_testing and ds\_sql\_testing

Removing last 9,10 columns in the file

The screenshot shows the Microsoft Azure Data Factory pipeline configuration. A 'Copy data' activity is selected, showing its mapping settings. The mapping details how 12 columns from the source are mapped to specific columns in the sink, including country, country\_code\_2\_digit, country\_code\_3\_digit, year\_week, week\_start\_date, week\_end\_date, new\_cases, tests\_done, population, and testing\_data\_source. There are also notes about potential data truncation for columns 5 and 6.

Created pipeline for testing

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (with one named 'pl\_sqlize'), 'Datasets', and 'Change Data'. The main pane displays a 'Details' view for a pipeline run. The run summary indicates a 'Succeeded' status with an activity run ID of 7f6e4e80-ad09-4e80-9c06-c792ec67ac2. The flow shows data being copied from 'Azure Data Lake Storage Gen2' (Region: UK South) to 'Azure SQL Database' (Region: UK South). Key metrics include:

- Data read:** 116,837 KB
- Files read:** 1
- Rows read:** 1,108
- Peak connections:** 1
- Data written:** 104,83 KB
- Rows written:** 1,108
- Peak connections:** 2

The 'Copy duration' was 00:00:13, and the 'Throughput' was 38,946 KB/s. The 'Duration' section details the steps: Queue (00:00:07), Pre copy script (00:00:00), and Transfer (00:00:03). A note states 'Data consistency verification' is 'Not verified'.

SQL command successfully ran on testing data

The screenshot shows the Microsoft Azure Query editor (preview) interface. The top navigation bar includes 'Search resources, services, and docs (G+)', 'Home > covid-dbgroup4 (covid-srv4/covid-dbgroup4)', and user information 'vgkic@umsystem.edu UNIVERSITY OF MISSOURI (MAIL...)'. The main area shows a 'Query editor (preview)' tab for 'covid-dbgroup4 (admingroup4)'. The sidebar displays 'Intelligent Performance' and 'Query Performance Insight'. The query results pane shows a table with the following data:

country	country_code_2_digit	country_code_3_digit	year_week	week_start_date	week_e
Austria	AT	AUT	2020-W15	2020-04-05	2020-04-11
Austria	AT	AUT	2020-W16	2020-04-12	2020-04-18
Austria	AT	AUT	2020-W17	2020-04-19	2020-04-25
Austria	AT	AUT	2020-W18	2020-04-26	2020-04-30
Austria	AT	AUT	2020-W19	2020-05-03	2020-05-09

The message at the bottom of the results pane says 'Query succeeded 12s'.

## Making Pipeline Production Ready

We automate all the pipelines instead of manual triggering.

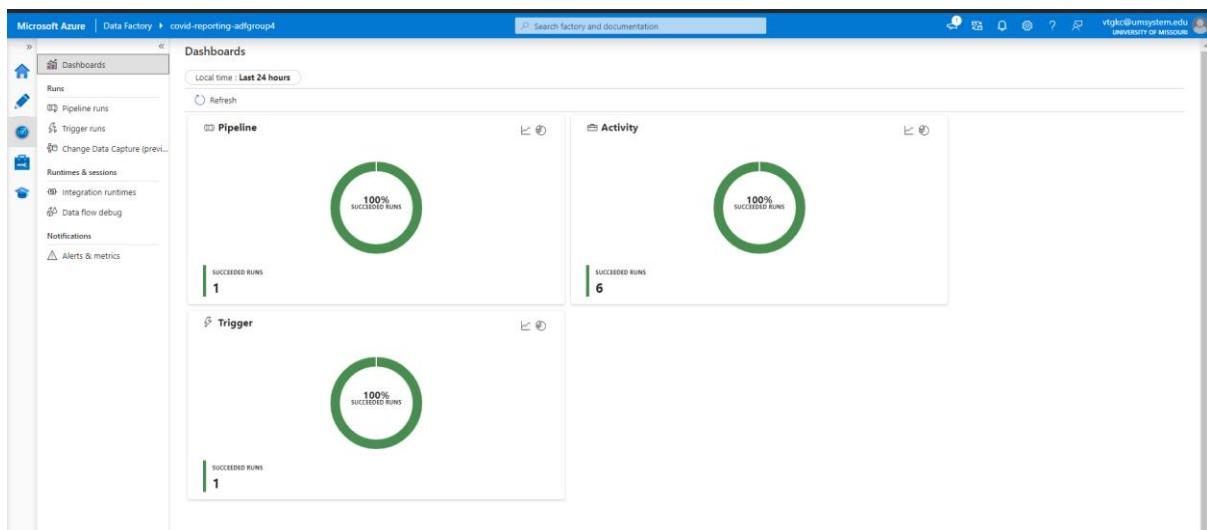
## Monitoring

As part of monitoring, we want to ensure that the pipelines are up and running and in a healthy state.

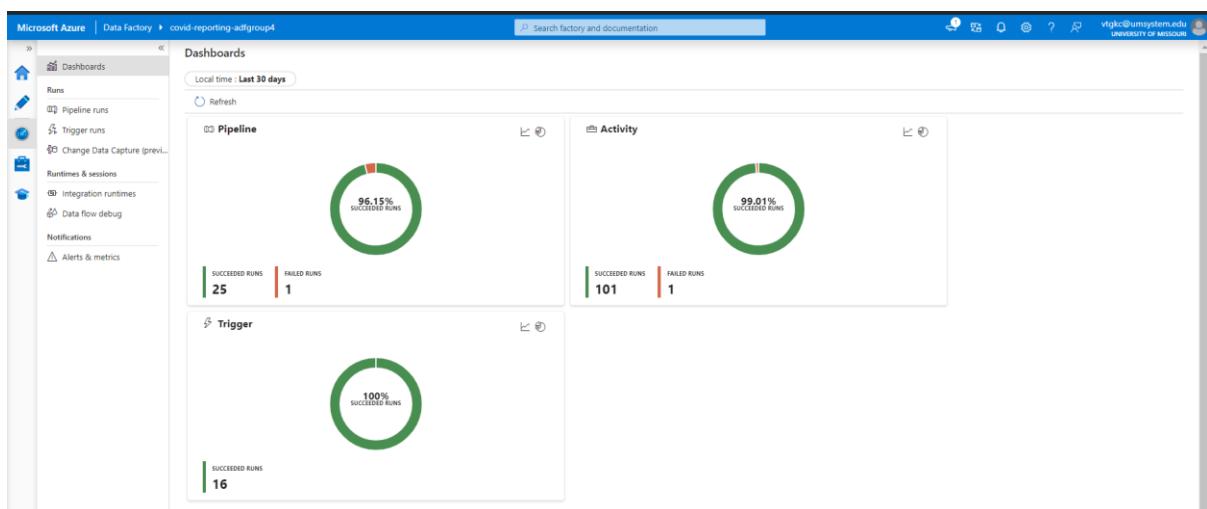
## Data Factory Monitor

Here we look at the status of triggers and pipelines.

Monitoring results in last 24 hours



Monitoring results in last 30 days



We can see all the trigger runs in the figure below. We have set the option to view the trigger runs from last 30 days. If you run a failed one or if you're rerunning a trigger you will see them here as well.

The screenshot shows the Microsoft Azure Data Factory interface for the 'covid-reporting-adfgroup4' factory. The left sidebar includes options like Dashboards, Runs, Pipeline runs, Trigger runs, Change Data Capture, Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, and Alerts & metrics. The main content area is titled 'Trigger runs' and displays a table of trigger runs. The table columns are: Trigger name, Trigger type, Trigger time, Status, Pipelines, Run, Message, Properties, and Run ID. The status column shows 'Succeeded' for all entries. The table lists 16 items, each corresponding to a different trigger name and type, such as 'tr\_ingest\_ecdc\_data' or 'tr\_process\_hospital\_admissions'. The Run ID column contains unique identifiers for each run.

Here we will see the start and end, and then the duration, and then whatever trigger it's triggered. So we'll see any parameters, we don't have any parameters on those ones.

And if we put any annotations, we can see them as well, If they're errors, we'll see them

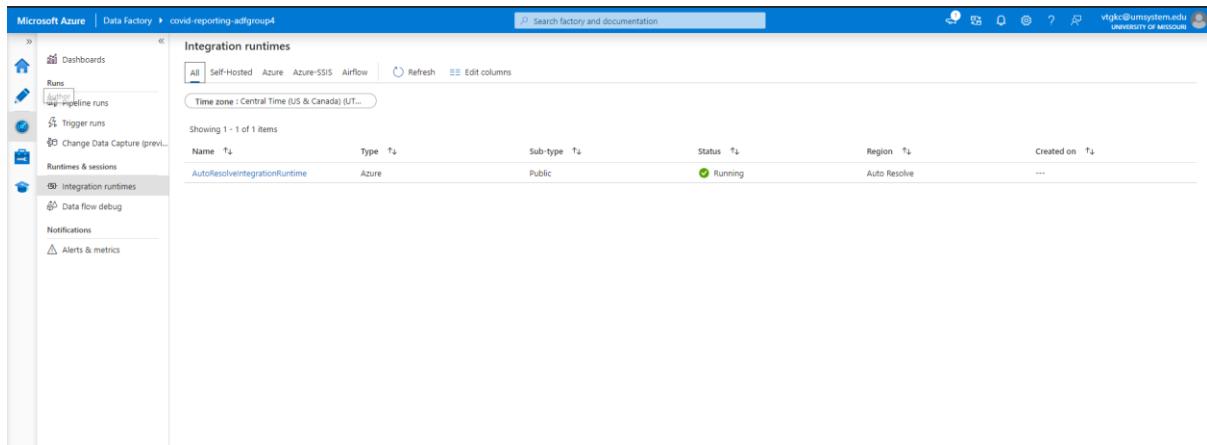
and we can add further filters as well.

So we've got the filters on pipeline names, status and runs, and we can add annotations.

filter on the annotation for example. So that is what we do there.

The screenshot shows the Microsoft Azure Data Factory interface for the 'covid-reporting-adfgroup4' factory. The left sidebar includes options like Dashboards, Runs, Pipeline runs, Trigger runs, Change Data Capture, Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, and Alerts & metrics. The main content area is titled 'Pipeline runs' and displays a table of pipeline runs. The table columns are: Pipeline name, Run start, Run end, Duration, Triggered by, Status, Run, Parameters, Annotations, and Run ID. The status column shows 'Succeeded' for most entries, with one entry labeled 'Run (Latest)'. The table lists 25 items, each corresponding to a different pipeline name and run start/end times. The Run ID column contains unique identifiers for each run.

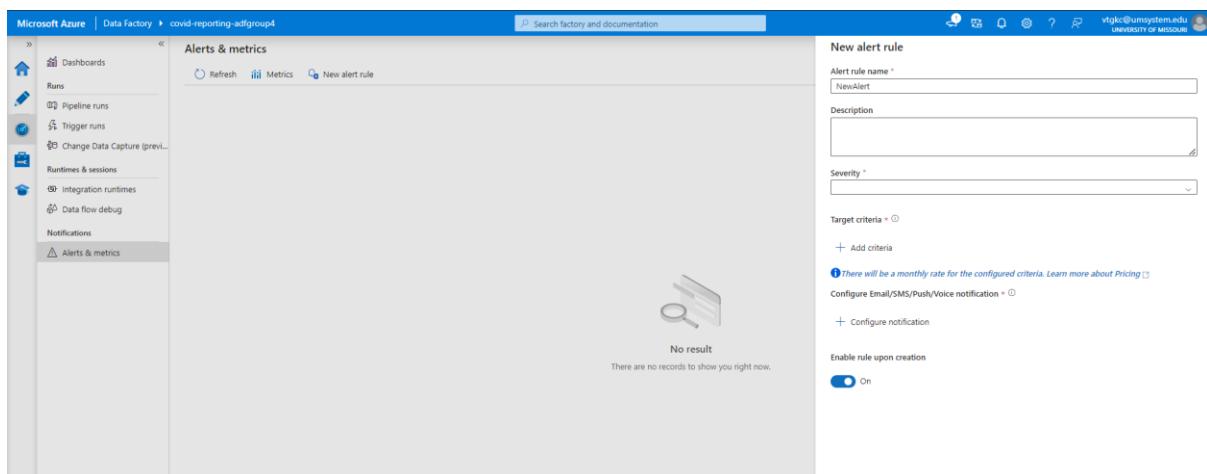
Next we've got the integration runtimes, so we can report on the status of the integration runtimes from here. At the moment, we only have the default order to resolve integration runtime.



The screenshot shows the 'Integration runtimes' page in the Microsoft Azure Data Factory interface. The left sidebar includes options like Dashboards, Runs, Pipeline runs, Trigger runs, Change Data Capture, Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, and Alerts & metrics. The main area displays a table with one item: 'AutoResolveIntegrationRuntime' (Name), 'Azure' (Type), 'Public' (Sub-type), 'Running' (Status), 'Auto Resolve' (Region), and a timestamp for Created on. There are filters for All, Self-Hosted, Azure, Azure-SSIS, and Airflow, along with a Refresh button and an 'Edit columns' link. A search bar at the top right says 'Search factory and documentation'.

If we want a report on any failures, or successes of triggers, or activities, or pipelines, we can create a new alert.

So we set your target criteria and tell what to do, whether to send an email, or send an SMS, or a voice notification for example.



The screenshot shows the 'Alerts & metrics' page in the Microsoft Azure Data Factory interface. The left sidebar includes options like Dashboards, Runs, Pipeline runs, Trigger runs, Change Data Capture, Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, and Alerts & metrics. The main area displays a table with no results, showing a magnifying glass icon and the text 'No result'. It also includes a note: 'There are no records to show you right now.' On the right, there's a form for creating a new alert rule. It asks for 'Alert rule name' (set to 'Newsletter'), 'Description' (empty), 'Severity' (empty), and 'Target criteria' (with a note: 'There will be a monthly rate for the configured criteria. Learn more about Pricing'). It also includes sections for 'Configure Email/SMS/Push/Voice notification' (with a note: '+ Add criteria'), 'Configure notification', and 'Enable rule upon creation' (with a toggle switch set to 'On').

Give the details to add the criteria for which the alert shall be created.



1

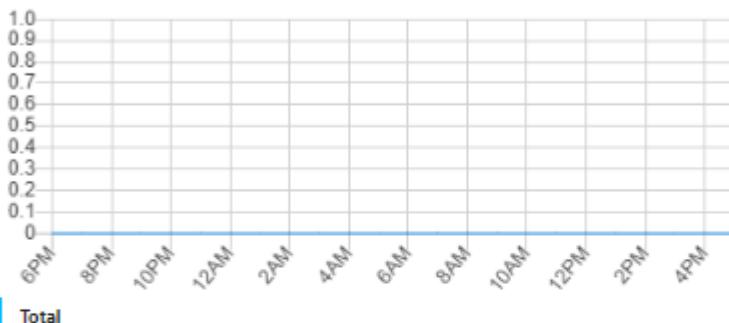
vtgkc@umsystem.edu  
UNIVERSITY OF MISSOURI

## Configure alert logic

Failed trigger runs metrics

Show history

Over the last 24 hours



Selecting the dimension values will help you filter to the right time series.

### Dimension

### Values

Name

tr\_ingest\_ecdc\_data

FailureType

3 selected

## Alert logic

### Condition \* ⓘ

Greater than

### Time aggregation \* ⓘ

Total

### Threshold count \* ⓘ

0

## Evaluate based on

### Period \* ⓘ

Over the last 1 minutes

### Frequency \* ⓘ

Add criteria

Back

Cancel

Now, create a notification for the alert.

## Configure notification

Notify your team via email and text messages or automate actions using webhooks, runbooks, functions logic apps or integrating with external ITSM solutions.

Create new  Use existing

Action group name \*

covid-support-project

Short name \*

covidsupport

Notifications	Action type	Actions
EmailSend	Email/SMS/Push/Voice	 

[+ Add notification](#)

## Edit notification

Learn more about [Pricing](#) and [Privacy statement](#).

Action name \*

EmailSend

### Select which notifications you'd like to receive

Email

vtgkc@umsystem.edu

SMS

Country code

Phone number \*

1



1234567890

Carrier charges may apply.

Azure app push notifications

Enter your email used to log into your Azure account. [Learn about connecting to your Azure resources using the Azure app](#).

email@example.com

Voice

Country code

Phone number \*

1



1234567890

Error faced due to no subscription to Microsoft.insights.

✖ Failed

```
{"error": {"code": "MissingSubscriptionRegistration", "message": "The subscription is not registered to use namespace 'microsoft.insights'. See https://aka.ms/rps-not-found for how to register subscriptions.", "details": [{"code": "MissingSubscriptionRegistration", "target": "microsoft.insights", "message": "The subscription is not registered to use namespace 'microsoft.insights'. See https://aka.ms/rps-not-found for how to register subscriptions."}]}}
```

Now, we can see the alert is created and we can edit the alert as shown below.

The screenshot shows the 'Edit alert rule' interface. At the top, there are several icons: a bell with 1 notification, a gear, a user icon with 3 notifications, a question mark, and a person icon. To the right of these is the email address [vtgkc@umsystem.edu](mailto:vtgkc@umsystem.edu) and the text 'UNIVERSITY OF MISSOURI'. Below the header, the title 'Edit alert rule' is displayed. The main form fields include:

- Alert rule name \***: al\_trigger\_failure
- Description**: A large text area containing the placeholder text 'Add description'.
- Severity \***: Sev0

Below these fields is a table for 'Target criteria':

Target criteria	Actions
Whenever Trigger Failed Runs metric is Greater Than 1	<a href="#"></a> <a href="#"></a>

A link '+ Add criteria' is located below the table. A note indicates: **i** There will be a monthly rate for the configured criteria. Learn more about Pricing [»](#).

Below the target criteria is another table for 'Notifications':

Notifications	Action group type	Actions
covid-support-project	1 Email	<a href="#"></a>

A link '+ Configure notification' is located below the notifications table. At the bottom, there is a section for 'Enable rule upon creation' with a toggle switch set to 'On'.

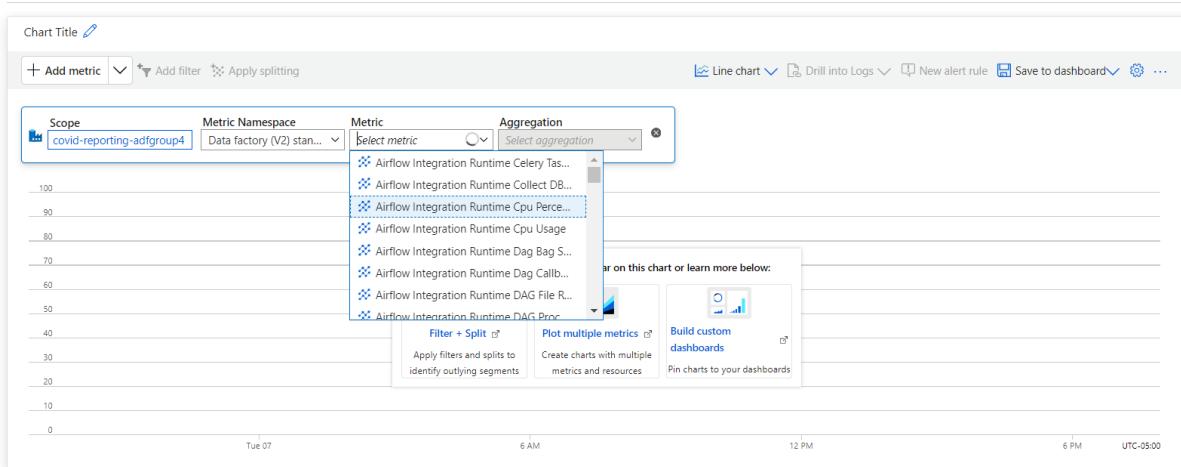
Now to monitor the metrics, navigate to alerts and metrics and click on metrics. Now we can see that a new tab is opened for monitoring. Now navigate to metrics to see all the resources available in our subscription.

Scope	Resource type	Location
Azure subscription 1	Subscription	-
covid-reporting-rg4	Resource group	-
covid-dbgroup4	SQL database	UK South
covid-reporting-adfgroup4	Data factory (V2)	UK South
covidreportingsagroup4	Storage account	UK South
covidreportingsagroup4-5951e67f-a130-4396-9c37...	Storage account	UK South
covidreportingsagroup4-5951e67f-a130-4396-9c37...	Event Grid System Topic	UK South
master	SQL database	UK South

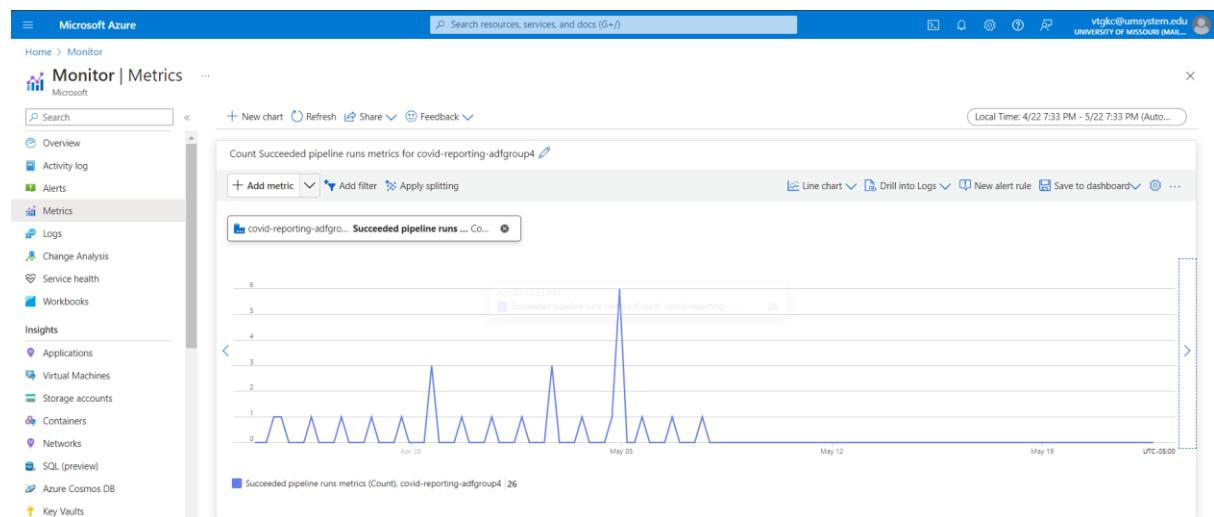
Now select data factory resource.

**Note:** Why can't I select multiple resources? Data factory (V2) resources have not enabled multi-selection with metrics. You can let the Data factory (V2) team know this capability is important and update this request.

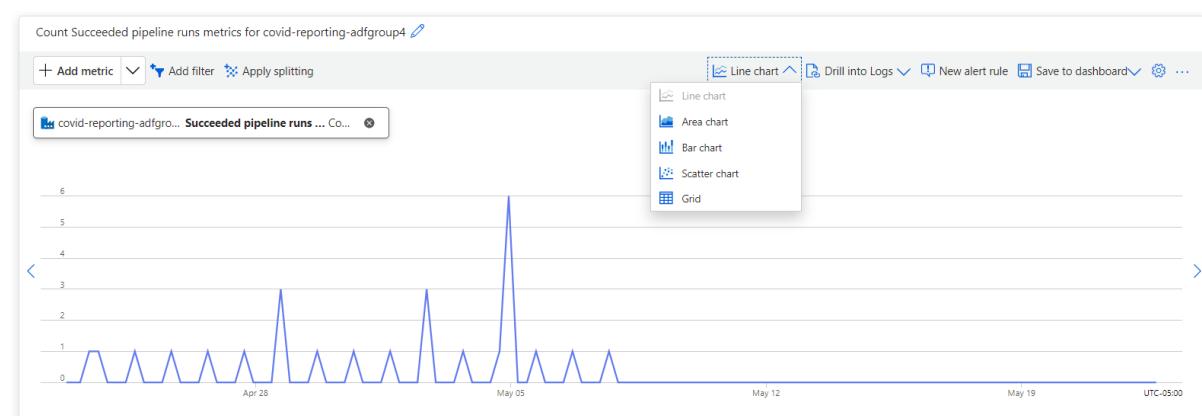
These are all the metrics available, we can select the metrics based upon the requirements.



So now we have selected succeeded pipeline run metrics with aggregation as count, below we can see metrics for last 30 days.



We can use different types of plots to visualize the metrics.



We can store the logs and metrics in the storage account by adding a diagnostic.

Subscription: Azure subscription 1  
Resource group: covid-reporting-rg4  
Resource type: Data factories (V2)  
Resource: covid-reporting-adgroup4

Name	Storage account	Event hub	Log Analytics workspace	Partner solution	Edit setting
No diagnostic settings defined					

+ Add diagnostic setting

Click 'Add Diagnostic setting' above to configure the collection of the following data:

- Pipeline activity runs log
- Pipeline runs log
- Trigger runs log
- Sandbox Pipeline runs log
- Sandbox Activity runs log

Deployment is completed for creating log analytics.

Deployment name: Microsoft.LogAnalyticsOMS  
Subscription: Azure subscription 1  
Resource group: covid-reporting-rg4

Your deployment is complete

Start time: 5/7/2024, 7:43:40 PM  
Correlation ID: 21e6661c-f8a1-4ff4-bce4-985727b23f01

Deployment details  
Next steps  
Go to resource

Give feedback  
Tell us about your experience with deployment

Cost management  
Get notified to stay within your budget and prevent unexpected charges on your bill.  
Set up cost alerts >

Microsoft Defender for Cloud  
Secure your apps and infrastructure  
Go to Microsoft Defender for Cloud >

Free Microsoft tutorials  
Start learning today >

In the diagnostic settings, we are enabling option to store the metrics and logs in storage account. Selecting ‘Resource specific’ option to create separate tables for each of the metrics and logs.

The screenshot shows the 'Diagnostic setting' configuration page in Microsoft Azure. On the left, there's a list of logs and categories. On the right, 'Destination details' are set up for 'Send to Log Analytics workspace'. It shows 'Subscription: Azure subscription 1' and 'Log Analytics workspace: covid-reporting-ws (aksouth)'. Below this, there are sections for 'Archive to a storage account' and 'Stream to an event hub'. A note about sending diagnostics to a storage account is visible.

Navigate to ‘Log Analytics Workspace’ to see the logs. We can see the tables under the log management.

Using Kusto Query language to query on these tables.

Querying on the trigger run table to count the number of records.

ADFTriggerRun – to see the details of the trigger run

The screenshot shows the 'Logs' section of the 'covid-reporting-ws' Log Analytics workspace. The 'ADFTriggerRun' table is selected. The results pane displays a single record with the following details:

TimeGenerated [UTC]	ResourceId	OperationName	Category	CorrelationId
5/9/2024, 12:25:02.254 AM	/SUBSCRIPTIONS/3DAS5B05-E067-4BA0-8E6C-E59781AED060/RESOURCEGROUPS/COVID-REPORTING-RG4/PROVIDERS/MICROSOFT.DATAPR	tr_ingest_eddc_data - Succeeded	TriggerRuns	afsefeed-31c2-449b-b1a1-309d4409b71d
	TenantId	e098df5f-6255-4264-875f-4f05f86f3e67		
	SourceSystem	Azure		
	TimeGenerated [UTC]	2024-05-09T00:25:02.254Z		
	ResourcedId	/SUBSCRIPTIONS/3DAS5B05-E067-4BA0-8E6C-E59781AED060/RESOURCEGROUPS/COVID-REPORTING-RG4/PROVIDERS/MICROSOFT.DATAPR		
	OperationName	tr_ingest_eddc_data - Succeeded		
	Category	TriggerRuns		
	CorrelationId	afsefeed-31c2-449b-b1a1-309d4409b71d		
	Level	Informational		
	Location	aksouth		
	Tags	0		
	Status	Succeeded		

To visualize the details of the trigger runs, the below code snippet in Kusto language is used.

We can find such details visually in the monitor section.

```
let latestADFTriggerRun = ADFTriggerRun
```

```
| summarize TimeGenerated = max(TimeGenerated) by TriggerId
```

```
| project TriggerId, TimeGenerated;
```

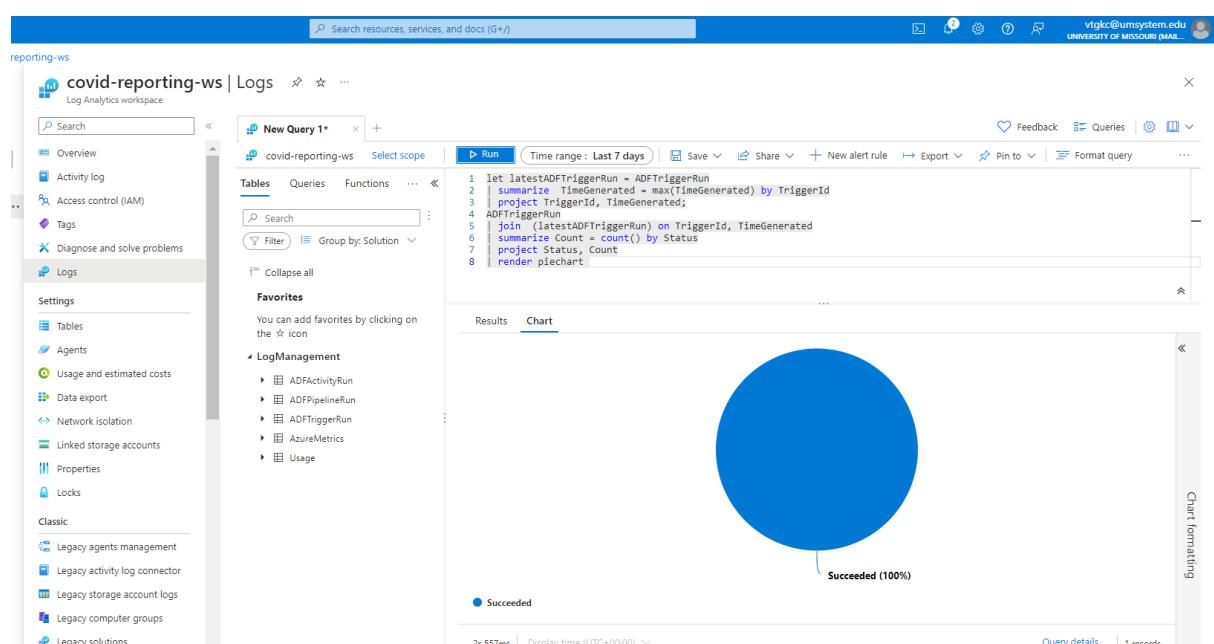
```
ADFTriggerRun
```

```
| join (latestADFTriggerRun) on TriggerId, TimeGenerated
```

```
| summarize Count = count() by Status
```

```
| project Status, Count
```

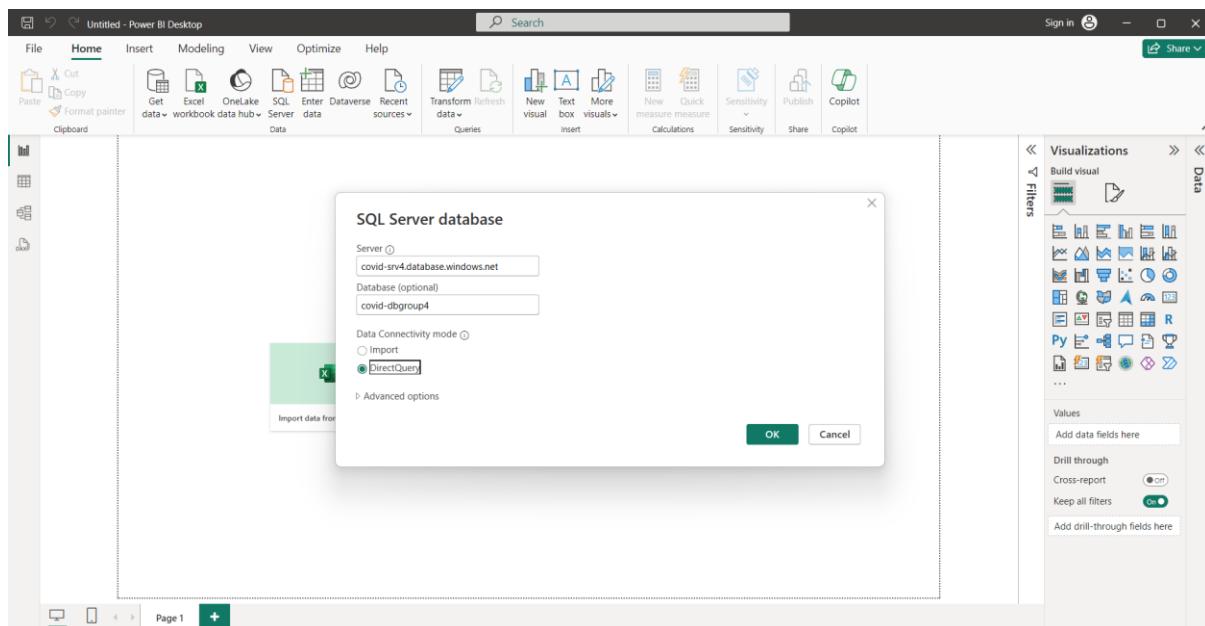
```
| render piechart
```



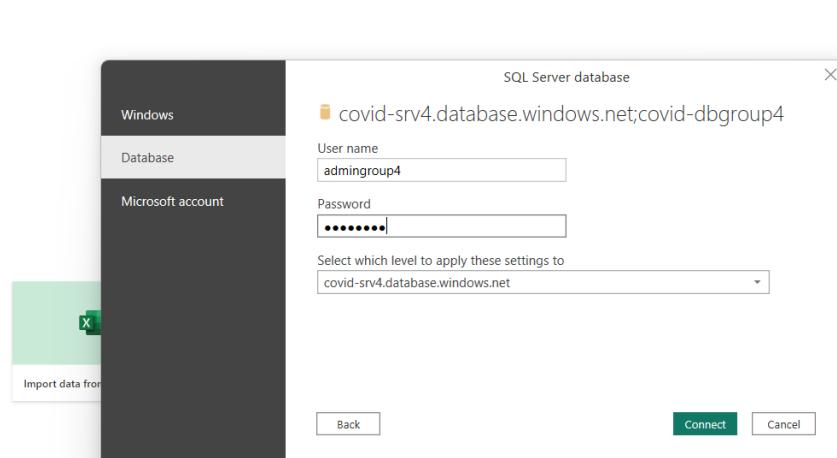
## Power BI dashboard

Installed Microsoft Power BI Desktop from the link <https://www.microsoft.com/en-us/download/details.aspx?id=58494>

Connected the Azure SQL database to Power BI desktop



Giving credentials of our database



The tables that are linked from database.

The screenshot shows the Power BI Desktop interface with the 'Data' tab selected. Three tables are listed in the data view:

- covid\_reporting cases\_and\_deaths**:
  - cases\_count
  - country
  - country\_code\_2\_digit
  - country\_code\_3\_digit
  - deaths\_count
  - population
  - reported\_date
  - source
- covid\_reporting hospital\_admissions\_daily**:
  - country
  - country\_code\_2\_digit
  - country\_code\_3\_digit
  - hospital\_occupancy\_count
  - icu\_occupancy\_count
  - population
  - reported\_date
  - source
- covid\_reporting testing**:
  - country
  - country\_code\_2\_digit
  - country\_code\_3\_digit
  - new\_cases
  - population
  - testing\_data\_source
  - tests\_done
  - week\_end\_date
  - week\_start\_date

The 'Properties' pane on the right shows settings for 'Cards':

- Show the database in the header when applicable: No (radio button)
- Show related fields when card is collapsed: Yes (radio button)
- Pin related fields to top of card: No (radio button)

The 'Data' pane on the right lists the three tables under the heading 'Search'.

## Conclusion

To sum up, our project effectively constructed an Azure data processing and ingestion system.

We successfully imported data from several sources, used pipelines to automate its flow, enhanced its quality through transformations, and utilized Azure SQL to enable perceptive analysis. The experience acquired taught me a lot about the value of monitoring and Azure's data management solutions. Future projects, including adding new data sources, executing sophisticated analytics, combining machine learning, or putting real-time processing in place, can make use of this foundation. By building on the accomplishments of this initiative, we may unleash even more possibilities for data-driven decision-making and additional innovation.

## **Challenges faced**

The project faced a number of difficulties, including insufficient CPU cores for Databricks and HDInsights cluster creation, which required a special request to support and could have resulted in significant costs and delays; IP allocation issues that initially prevented login to the SQL server but were resolved by assigning an available IP; difficulties reproducing PowerBI visualizations because of IP allocation issues for the SQL server; and difficulties creating alerts because of no subscription to Microsoft.insights, which were fixed by adding it to the subscription.

Video Link:

[https://drive.google.com/file/d/1jSFwxmFlDNdzbK\\_4VxyXTjeDZEz3cz5q/view?pli=1](https://drive.google.com/file/d/1jSFwxmFlDNdzbK_4VxyXTjeDZEz3cz5q/view?pli=1)

Github repo link:

[https://github.com/Viswanth13/CovidData\\_Azure4](https://github.com/Viswanth13/CovidData_Azure4)