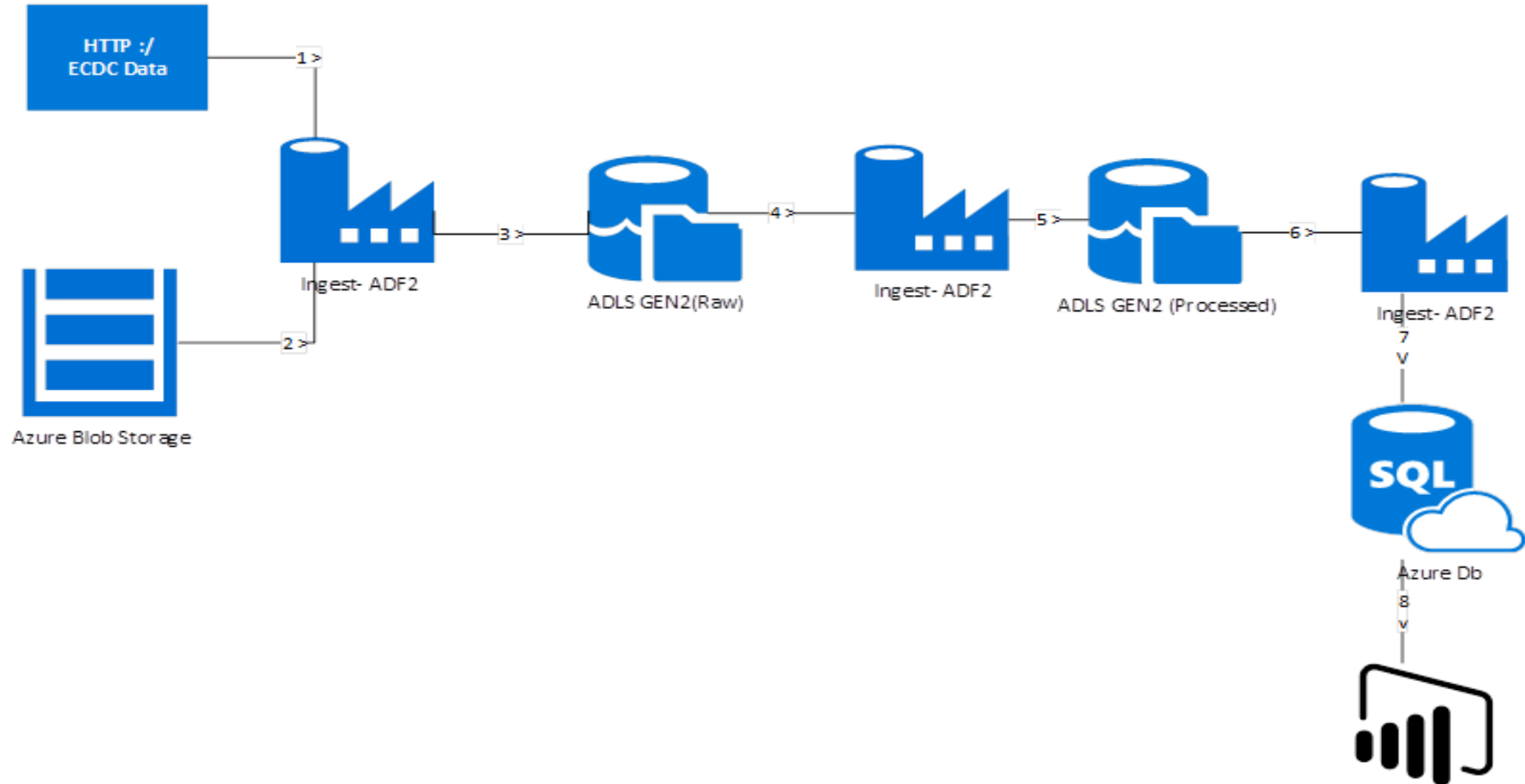


Overview of ECDC Data Processing using ADLS(Gen2) & Azure Data Factory 2

- Girish Kondeti

Architecture



ECDC (European Centre for Disease Control) Files from HTTP Source to ADLS Gen2 (Raw) - Pipeline (Steps 1,2 &3)

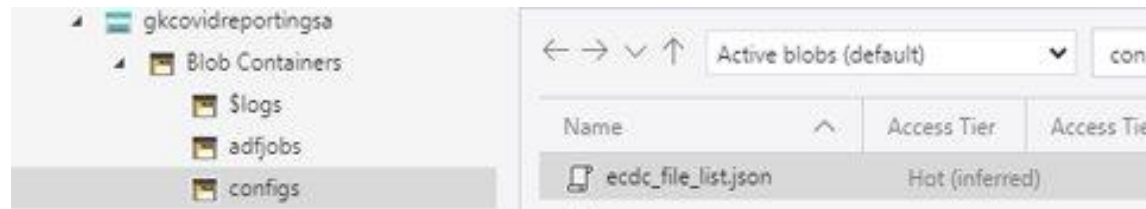
The screenshot displays the Microsoft Azure Data Factory portal interface. The browser address bar shows the URL: `adf.azure.com/en-us/authoring/pipeline/pl_ingest_ecdc_files?factory=%2Fsubscriptions%2F6d53f9f4-c0aa-476f-a768-52917162c8b5%2FresourceGroups%2FGKCovidReportin...`. The page title is "Microsoft Azure | GKCovidReportingADF".

The interface is divided into several sections:

- Factory Resources:** A sidebar on the left showing a tree view of resources. Under the "Ingest" folder, the pipeline "pl_ingest_ecdc_files" is selected.
- Activities:** A central panel showing a list of activities. The "Lookup" activity is selected, and its properties are displayed in the bottom panel.
- Lookup Activity Properties:** The bottom panel shows the "General" tab for the "Lookup ecdc file list" activity. The "Name" field is set to "Lookup ecdc file list", and the "Timeout" is set to "7.00:00:00".
- Pipeline Diagram:** The main canvas shows a pipeline diagram with two activities: "Lookup ecdc file list" and "ForEach1". A green arrow indicates the flow from the Lookup activity to the ForEach activity.

Pipeline - ECDC Files Ingest - Previous slide Explained

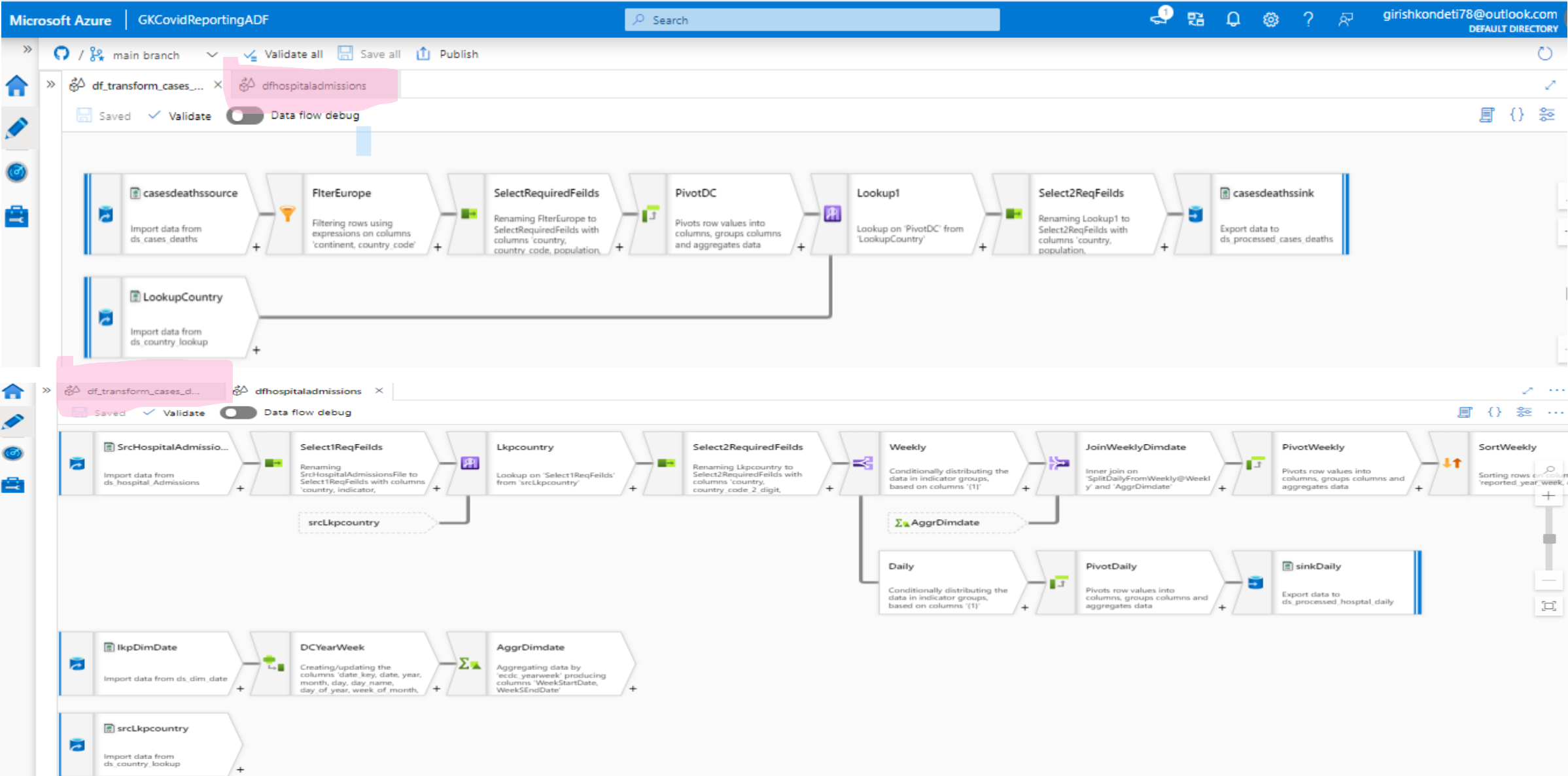
- The Pipeline copies all 4 raw files from source(https://) to destination ADLS GEN2 (Raw).
- The Lookup component is for JSON file configuration . The JSON will have details of file names and http location details.



- The copy component is wrapped with in the For each loop component to iterate for each file specified in JSON file. The files considered are:
 - cases_deaths.csv,
 - hospital_admissions.csv,
 - country_response.csv

Another pipe line for process (2) azure blob storage to ADLS GEN2(Raw) - Trigger when file arrives

Data processing from ADLS (Raw) to ADLS(Processed)- Steps 4 & 5



Data Flows - Processing - Previous Slide Explained

- Dataflow (df_transform_casesanddeaths) is created to transform the raw data – predominantly used to select columns, filter and lookup to country file saved in azure blob store, aggregate and to get country data and pivots the file data using an indicator column
- Dataflow (df_hospitaladmissions) is created to transform the raw data and works in similar way as above but separates the raw file to weekly and daily using conditional split.
- The dataflows can't be run directly so pipe lines are created with a data flow task. These pipe lines are scheduled or triggered

ADLS to Azure SQL Database (Copy activity) - Steps 6 & 7

The screenshot displays the Microsoft Azure portal interface for an Azure Data Factory (ADF) instance named 'GKCovidReportingADF'. The left sidebar shows navigation options: Home, Author, Monitor, and Manage. The 'Author' tab is active, showing a tree view of 'Factory Resources' on the left and a list of 'Activities' in the center. The 'Copy data' activity is selected, and its configuration is shown in the right pane.

Factory Resources:

- Pipeline: 8
 - Ingest: 2
 - Process: 3
 - Sqlize: 3
 - pl_sql_cases_and_deaths
 - pl_sql_hospital_admissions
 - pl_sql_testing
- Dataset: 16
- Data flows: 2
 - df_transform_cases_deaths
 - dfhospitaladmissions
- Power Query (Preview): 0
- Templates: 0

Activities:

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

Copy data activity configuration:

- General tab:** Sink dataset is set to 'ds_sql_cases_and_deaths'. Stored procedure name is 'Select...'. Table option is 'None'. Pre-copy script is 'truncate table covid_reporting.cases_and_deaths'.
- Source tab:** Source is 'coycasesanddeathssql'.
- Mapping tab:** Mapping is 'coycasesanddeathssql'.

Processed Data Loaded into Azure SQL Database (Step 7)

The screenshot displays the Microsoft SQL Server Management Studio interface. The title bar indicates the connection to 'SQLQuery1.sql - srv-covidsql.database.windows.net.covidrpt-db (Admin1 (91))'. The 'Object Explorer' on the left shows the database structure, including 'covidrpt-db' and its tables. The 'Query Editor' in the center contains a SQL query that selects the top 1000 rows from the 'cases_and_deaths' table in the 'covid_reporting' schema. The 'Results' pane at the bottom shows the output of the query, which is a table with 9 rows and 8 columns: country, country_code_2_digit, country_code_3_digit, population, cases_count, deaths_count, reported_date, and source. The first row is highlighted, showing data for Italy.

SQLQuery1.sql - srv-covidsql.database.windows.net.covidrpt-db (Admin1 (91)) - Microsoft SQL Server Management Studio

File Edit View Query Project Tools Window Help

Connect

srv-covidsql.database.windows.net (SQL Server 12.0.2)

- Databases
 - System Databases
 - covidrpt-db
 - Database Diagrams
 - Tables
 - System Tables
 - External Tables
 - GraphTables
 - covid_reporting.cases_and_deaths
 - covid_reporting.hospital_admissions_d
 - covid_reporting.testing
 - Views
 - External Resources
 - Synonyms
 - Programmability
 - Query Store
 - Extended Events
 - Storage
 - Security
 - Security
 - Integration Services Catalogs

SQLQuery1.sql - sr...pt-db (Admin1 (91))

```
/****** Script for SelectTopNRows command from SSMS *****/  
SELECT TOP (1000) [country]  
    , [country_code_2_digit]  
    , [country_code_3_digit]  
    , [population]  
    , [cases_count]  
    , [deaths_count]  
    , [reported_date]  
    , [source]  
FROM [covid_reporting].[cases_and_deaths]
```

100 %

Results Messages

	country	country_code_2_digit	country_code_3_digit	population	cases_count	deaths_count	reported_date	source
1	Italy	IT	ITA	60359546	NULL	0	2020-06-25	Epidemic intelligence, national daily data
2	Romania	RO	ROU	19414458	NULL	48	2020-08-29	Epidemic intelligence, national daily data
3	Slovakia	SK	SVK	5450421	NULL	0	2020-07-16	Epidemic intelligence, national daily data
4	Serbia	RS	SRB	6963764	NULL	0	2020-02-21	Epidemic intelligence, national daily data
5	Slovenia	SI	SVN	2080908	NULL	0	2020-02-22	Epidemic intelligence, national daily data
6	France	FR	FRA	67012883	NULL	18	2020-07-02	Epidemic intelligence, national daily data
7	Switzerland	CH	CHE	8544527	NULL	1	2020-08-23	Epidemic intelligence, national daily data
8	Luxembourg	LU	LUX	613894	NULL	3	2020-04-22	Epidemic intelligence, national daily data
9	Turkey	TR	TUR	82003882	NULL	47	2020-05-11	Epidemic intelligence, national daily data

Triggers - Tumbling Window runs at 5:30 AM IST : Remaining triggers are dependent on tr_ingest_ecdc_data.

Microsoft Azure | GKCovidReportingADF

main branch | Validate all | Save all | Publish

Connections

- Linked services
- Integration runtimes
- Azure Purview (Preview)
- Source control
- Git configuration
- ARM template
- Author
- Triggers
- Global parameters
- Security
- Customer managed key
- Managed private endpoints

Triggers

To execute a pipeline set the trigger. Triggers represent a unit of processing that determines when a pipeline execution needs to be kicked off.

+ New

Filter by name | Annotations : Any

Showing 1 - 6 of 6 items

Name ↑↓	Type ↑↓	Status ↑↓	Related ↑↓	Annotations ↑↓
tr_ingest_ecdc_data	Tumbling window	Started	3	
tr_process_cases_and_deaths	Tumbling window	Started	1	
tr_process_hospital_admissions	Tumbling window	Started	2	
tr_sqlize_cases_and_deaths	Tumbling window	Started	1	
tr_sqlize_hospital_admissions	Tumbling window	Started	1	
Tr_copyfile	Storage events	Started	1	

Microsoft Azure | GKCovidReportingADF

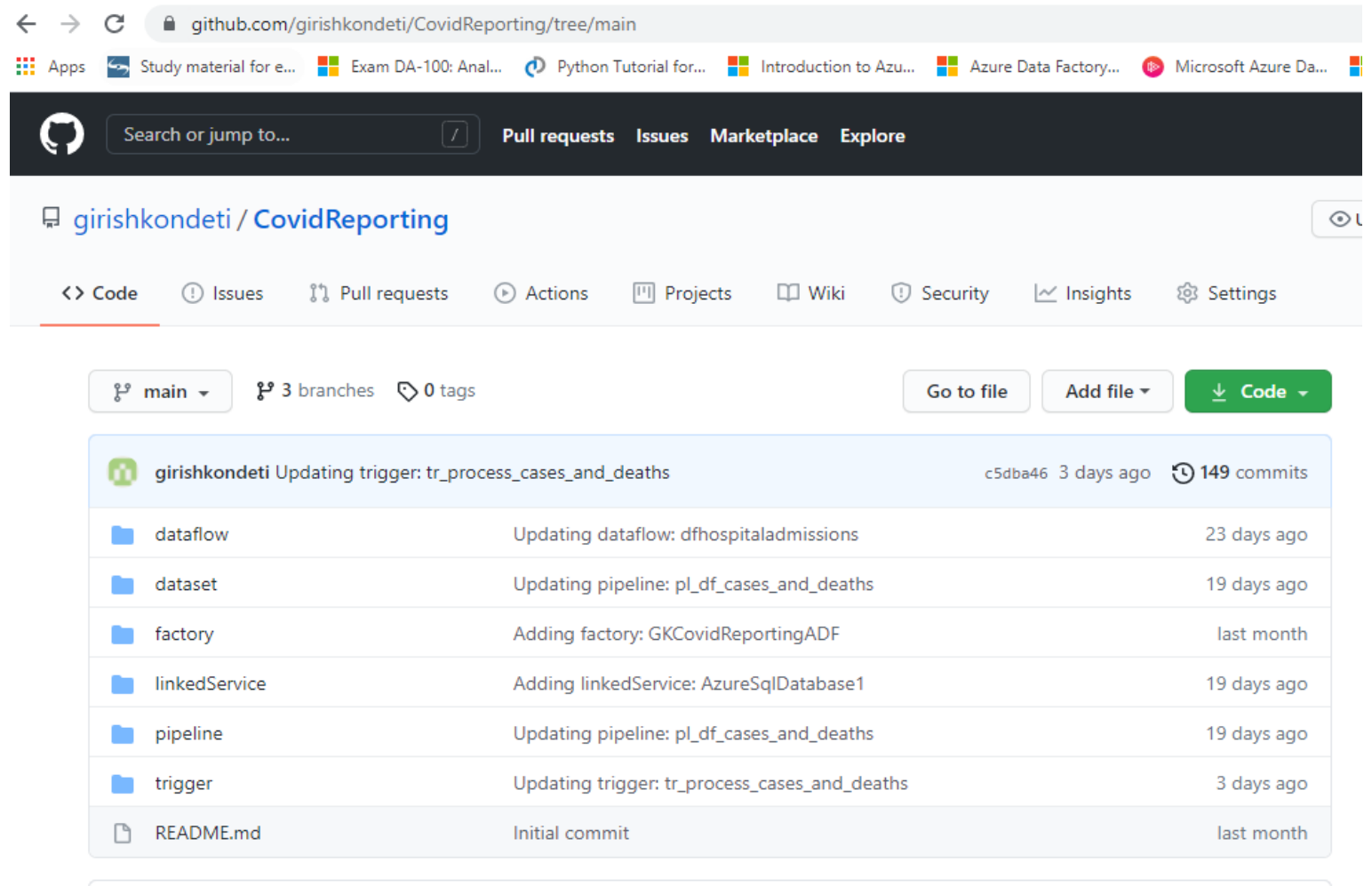
Triggers

Chennai, Kolkata, Mu... : Last 24 hours | Trigger name : All | Status : All | Runs : Latest runs

Showing 1 - 5 items

Trigger name	Trigger type	Trigger time ↑↓	Status	Pipelines	Run	Message	Properties	Run ID
tr_ingest_ecdc_data	Tumbling window trig...	5/28/21, 2:52:56 PM	Succeeded	1	Rerun (Latest)			08585794131088016245902715545CU68
tr_ingest_ecdc_data	Tumbling window trig...	5/28/21, 5:30:00 AM	Succeeded	1	Original			08585795332850820229895877894CU66
tr_process_hospital_admi...	Tumbling window trig...	5/28/21, 5:29:59 AM	Succeeded	1	Original			08585795332850738931321050757CU61
tr_sqlize_cases_and_deaths	Tumbling window trig...	5/28/21, 5:30:00 AM	Succeeded	1	Original			08585795332840160203419429697CU42
tr_sqlize_hospital_admissi...	Tumbling window trig...	5/28/21, 5:30:00 AM	Succeeded	1	Original			08585795332851646849908728436CU61
tr_process_cases_and_de...	Tumbling window trig...	5/28/21, 5:29:59 AM	Succeeded	1	Original			08585795332853855589424110449CU56

GitHub



github.com/girishkondeti/CovidReporting/tree/main

Search or jump to... Pull requests Issues Marketplace Explore

girishkondeti / CovidReporting

<> Code ! Issues 🔗 Pull requests ▶ Actions 📁 Projects 📖 Wiki 🛡 Security 📈 Insights ⚙ Settings

main 3 branches 0 tags Go to file Add file Code

	girishkondeti Updating trigger: tr_process_cases_and_deaths	c5dba46 3 days ago 149 commits
dataflow	Updating dataflow: dfhospitaladmissions	23 days ago
dataset	Updating pipeline: pl_df_cases_and_deaths	19 days ago
factory	Adding factory: GKCovidReportingADF	last month
linkedService	Adding linkedService: AzureSqlDatabase1	19 days ago
pipeline	Updating pipeline: pl_df_cases_and_deaths	19 days ago
trigger	Updating trigger: tr_process_cases_and_deaths	3 days ago
README.md	Initial commit	last month

Branches :

Master – used as Development branch

Main - Master merged to Main (ADF can only publish from Main)

ADFPipeline- Release maintained by ADF