



Final Project

Covid Use Case

Project By- Aniket Bhagwat Mohite.

Employee Id- 2261295

Domain- AIA-Azure

Cohort code- CDB23CSDDW006

I. Project Requirement:

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

II. Resource Provided:

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

III. Resource Used:

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

Process:

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

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

1. Copy Data from Source to Destination –

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

The screenshot shows the Microsoft Azure Resource Groups page. At the top, there are navigation links for Home, Resource groups, and Default Directory. Below the header, there are buttons for Create, Manage view, Refresh, Export to CSV, Open query, and Assign tags. A search bar is present with the placeholder "Search resources, services, and docs (G+)" and a dropdown menu for "Subscription equals all". There are also filters for Location (equals all) and Add filter. The main table displays one record: "covid-rg" under the Name column, with "Subscription" set to "Free Trial" and "Location" set to "Central India". The table includes sorting and grouping options like "Name ↑", "Subscription ↑", and "Location ↑". At the bottom, there are pagination controls ("Page 1 of 1") and a feedback link ("Give feedback"). The browser's address bar shows "portal.azure.com/#view/HubsExtension/BrowseResourceGroups". The taskbar at the bottom of the screen shows various pinned icons and the date/time as 07-06-2023 14:11.

- Create storage account using same resource group.

The screenshot shows the Microsoft Azure Storage Accounts page. The URL in the address bar is "portal.azure.com/#create/Microsoft.StorageAccount-ARM". The main heading is "Create a storage account" with a "..." button. Below it, there are tabs for Basics, Advanced, Networking, Data protection, Encryption, Tags, and Review. The Basics tab is selected. Under "Subscription", "Free Trial" is chosen. Under "Resource group", "covid-rg" is selected. In the "Instance details" section, the "Storage account name" is "covidads46", "Region" is "(Asia Pacific) Central India", and "Performance" is set to "Standard: Recommended for most scenarios (general-purpose v2 account)". The "Redundancy" is "Locally-redundant storage (LRS)". At the bottom, there are "Review" and "Next : Advanced >" buttons, along with a "Give feedback" link. The taskbar at the bottom shows various pinned icons and the date/time as 07-06-2023 14:14.

- create two containers named covid and transformedpath.

The screenshot shows the Microsoft Azure Storage container list for the storage account 'covidadls46'. The left sidebar shows navigation options like Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, and Storage browser. The main area displays a table of containers:

Name	Last modified	Public access level	Lease state
\$logs	6/7/2023, 2:15:52 PM	Private	Available
covid	6/7/2023, 2:17:01 PM	Private	Available
transformedpath	6/7/2023, 2:17:22 PM	Private	Available

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

The screenshot shows the Microsoft Azure Storage container details for the 'covid' container. The left sidebar shows navigation options like Overview, Diagnose and solve problems, Access Control (IAM), Settings (Shared access tokens, Manage ACL, Access policy, Properties, Metadata), and Authentication method (Access key). The main area displays a table of blobs:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
Destination					-	---
Source					-	---

- Upload all given files to source directory.

The screenshot shows the Microsoft Azure Storage Container 'covid'. The left sidebar has 'Overview' selected. The main area displays a table of blobs:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]	6/7/2023, 2:18:28 PM	Hot (Inferred)		Block blob	131.48 KiB	Available
case_deaths_uk_ind_only.csv	6/7/2023, 2:18:46 PM	Hot (Inferred)		Block blob	13.78 MiB	Available
cases_deaths.csv	6/7/2023, 2:18:27 PM	Hot (Inferred)		Block blob	46.2 KiB	Available
country_response.csv	6/7/2023, 2:18:29 PM	Hot (Inferred)		Block blob	1.01 MiB	Available
hospital_admissions.csv	6/7/2023, 2:18:27 PM	Hot (Inferred)		Block blob	83.87 KiB	Available
testing.csv	6/7/2023, 2:18:27 PM	Hot (Inferred)		Block blob	83.87 KiB	Available

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

The screenshot shows the 'Create Data Factory' page. The 'Basics' tab is selected. The 'Project details' section includes:

- Subscription: Free Trial
- Resource group: covid-rg
- Name: Covid-ADF-46
- Region: Central India
- Version: V2

At the bottom, there are navigation buttons: 'Review + create', '< Previous', 'Next : Git configuration >', and 'Give feedback'.

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

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the navigation menu includes General, Factory settings, Connections, Linked services (which is selected), Integration runtimes, Microsoft Purview, Source control, Git configuration, ARM template, Author, Triggers, Global parameters, Data flow libraries, Security, Credentials, Customer managed key, Outbound rules, and Managed private endpoints. The main area is titled "Linked services" and contains a sub-section "Linked service defines the connection information to a data store or compute." A "New" button is visible. On the right, a "New linked service" form is open, titled "Azure Data Lake Storage Gen2". The "Name" field is set to "LS_adls". The "Connect via integration runtime" dropdown is set to "AutoResolveIntegrationRuntime". The "Authentication type" is "Account key". Under "Account selection method", the "From Azure subscription" radio button is selected, and the "Azure subscription" dropdown shows "Free Trial (b578748c-dbc...-eef988ad24)". The "Storage account name" dropdown is set to "covidadls46". A "Test connection" section shows a successful connection. At the bottom, there are "Create", "Back", "Test connection", and "Cancel" buttons.

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

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the navigation menu includes Factory Resources (with a count of 1), Pipelines, Change Data Capture (preview), Datasets (with a count of 1, selected), Data flows, and Power Query. The main area shows a dataset named "ds_source" with a CSV icon. The "Properties" panel on the right shows the "General" tab selected, with the "Name" field set to "ds_source". The "Connection" tab shows "Linked service" set to "LS_adls". The "Schema" and "Parameters" tabs are also visible. At the bottom, there are "Create", "Back", "Test connection", and "Cancel" buttons.

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

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (0), 'Change Data Capture (preview)' (0), 'Datasets' (2), 'ds_destination' (selected), 'ds_source' (0), 'Data flows' (0), and 'Power Query' (0). In the main workspace, two datasets are shown: 'ds_source' (DelimitedText, CSV) and 'ds_destination'. The 'ds_destination' dataset properties are being edited. The 'General' tab is selected, showing the name 'ds_destination' and a description field. The 'Connection' tab shows a linked service 'LS_adls' and a file path configuration: 'covid' / 'Destination' / 'File name'. Other settings include compression type (None), column delimiter (Comma (,), Row delimiter (Default (\r\n, \n, or \r\n)), Encoding (Default(UTF-8)), and Quote character (Double quote ("))).

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

The screenshot shows the Microsoft Azure Data Factory interface with a pipeline named 'CopyCovidData-ADLS-Pipeline'. The pipeline consists of three main components: 'ds_source', 'ds_destination', and 'CopyCovidData-A...'. The 'ds_source' and 'ds_destination' datasets are the same as in the previous screenshot. The 'CopyCovidData-A...' component is a 'Copy data' activity. Its properties show it is configured to copy files from the 'ds_source' dataset to a 'Wildcard file path' in the 'ds_destination' dataset. The 'Wildcard paths' field is set to 'covid / [Source] / [*]'. The 'Source' tab of the 'Copy data' activity is selected, showing the source dataset 'ds_source' and the file path type 'Wildcard file path'.

The screenshot shows the Microsoft Azure Data Factory Pipeline Editor. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (CopyCovidData-ADLS-Pipeline), 'Datasets' (ds_destination, ds_source), and other components. The main workspace displays a pipeline step titled 'CopyCovidData-A...'. The 'Sink' tab is selected, showing configuration for 'ds_destination', 'Copy behavior' (None), 'Max concurrent connections', 'Block size (MB)', and 'Metadata'. The 'Properties' pane on the right shows the pipeline name as 'CopyCovidData-ADLS-Pipeline'. The system tray at the bottom indicates it's 34°C and sunny.

Output:

The screenshot shows the Microsoft Azure Storage Explorer interface. It displays a blob container named 'covid' under 'Storage accounts / covidadls46 / Containers'. The container contains several CSV files: 'case_deaths_uk_ind_only.csv', 'cases_deaths.csv', 'country_response.csv', 'hospital_admissions.csv', and 'testing.csv'. The 'Settings' sidebar on the left includes options like 'Shared access tokens', 'Manage ACL', 'Access policy', 'Properties', and 'Metadata'. The system tray at the bottom indicates it's 37°C and sunny.

2. Merge All files of Destination Folder

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

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

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists 'Pipelines', 'Datasets', and 'Data flows'. Under 'Datasets', there is a 'Merging_Datasets' folder containing five datasets: 'ds_file_1', 'ds_file_2', 'ds_file_3', 'ds_file_4', and 'ds_file_5'. The 'ds_file_1' dataset is selected. The main panel displays its properties. The 'Connection' tab shows 'Linked service' set to 'LS_adls', 'File path' set to 'covid / Destination / case_deaths_uk_ind...', and other settings like 'Compression type' (None), 'Column delimiter' (Comma), 'Row delimiter' (Default (\r\n, or \n)), 'Encoding' (Default(UTF-8)), and 'Quote character' (Double quote ('')). The 'Properties' pane on the right shows the name 'ds_file_1' and a description field. The bottom status bar shows the date as 07-06-2023 and time as 15:15.

2) Create Dataflow

1) Select each file Dataset as Source

The screenshot shows the Microsoft Azure Data Factory interface for creating a Dataflow. The 'Validate' button is checked. The main panel displays a dataflow pipeline with five parallel import steps from datasets 'ds_file_1' through 'ds_file_5', followed by a 'UnionOfFiles' step, then 'SelectColumns' (with 'Columns: 4 total'), and finally 'FilterForDeaths' (with 'Filtering rows using expressions on columns "Indicator"'). The output is 'SinkMergedFile' which exports data to 'ds_destination'. The 'Properties' pane on the right shows the name 'dataflow1' and a description field. The bottom status bar shows the date as 07-06-2023 and time as 16:03.

2) Use Union Activity to combine each file.

The screenshot shows the Microsoft Azure Data Factory Data Flow designer. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, and Data flows. Under 'Data flows', 'Merging_Data_Flow' is selected. The main workspace displays a data flow diagram with five input datasets: 'ds_file_1', 'ds_file_2', 'ds_file_3', 'ds_file_4', and 'ds_file_5'. These datasets are combined using a 'Union' activity, which then feeds into a 'select1' activity. The 'Properties' pane on the right shows the data flow is named 'Merging_Data_Flow'. The 'Union settings' tab is selected, showing the incoming stream is 'File1' and the union is by 'Name'. Below this, there are fields for 'Union with' containing 'File2', 'File3', 'File4', and 'File5'. The status bar at the bottom indicates it's 35°C and sunny.

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

The screenshot shows the Microsoft Azure Data Factory Data Flow designer. The 'Factory Resources' sidebar shows 'dataflow1' is selected under 'Data flows'. The main workspace displays a data flow diagram starting with 'File1' (import data from 'ds_file_1'), followed by a 'UnionOfFiles' activity (Combining rows from transformation 'File1, File2, File3, File4, File5'), and finally a 'select1' activity (Columns: 4 total). The 'Properties' pane on the right shows the data flow is named 'dataflow1'. The 'Select settings' tab is selected, showing options for skipping duplicate input and output columns. Below this, the 'Input columns' section shows four mappings: 'abc continent' to 'continent', 'abc indicator' to 'indicator', '123 daily_count' to 'daily_count', and 'date' to 'date'. The status bar at the bottom indicates it's 15:55 and 07-06-2023.

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

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

Output File -

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]	6/7/2023, 3:26:34 PM	Hot (Inferred)		Block blob	131.48 KIB	Available
case_deaths_uk_ind_only.csv	6/7/2023, 3:26:34 PM	Hot (Inferred)		Block blob	13.78 MiB	Available
cases_deaths.csv	6/7/2023, 3:26:34 PM	Hot (Inferred)		Block blob	46.2 KIB	Available
country_response.csv	6/7/2023, 3:26:34 PM	Hot (Inferred)		Block blob	1.01 MiB	Available
hospital_admissions.csv	6/7/2023, 3:26:34 PM	Hot (Inferred)		Block blob	1.69 MiB	Available
Merged_File.json	6/7/2023, 4:08:23 PM	Hot (Inferred)		Block blob	83.87 KIB	Available
testing.csv	6/7/2023, 3:26:34 PM	Hot (Inferred)		Block blob	83.87 KIB	Available

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

1) create Dataset for merged file

The screenshot shows the Microsoft Azure Data Factory studio interface. On the left, the 'Factory Resources' sidebar lists various components: Pipelines, Datasets, Data flows, and Power Query. In the center, a dataset named 'ds_merged_file' is being configured. The 'Properties' pane on the right shows the dataset's name as 'ds_merged_file' and its description. The 'Connection' tab is selected, showing the linked service 'LS_adls' and the file path '/covid/Merged_File.json'. The 'Schema' and 'Parameters' tabs are also visible. A battery saver notification is present at the bottom right.

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

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists various components: Pipelines, Datasets, and Data flows. Under 'Datasets', 'ds_transformation' is selected. The main panel displays the properties for this dataset. The 'Type' is set to 'Parquet'. The 'File path' field contains the value 'transformedpath' with a trailing slash. The 'Compression type' is set to 'snappy'. The 'Connection' dropdown is set to 'LS_adls'. The 'Properties' pane on the right shows the 'General' tab with 'Name' set to 'ds_transformation' and 'Description' left empty. The status bar at the bottom indicates it's 35°C and sunny.

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

This screenshot shows the same Microsoft Azure Data Factory interface as the previous one, but with a parameter added to the dataset. In the 'Parameters' section of the properties pane, there is a table with one row. The 'Name' column is 'FolderName', the 'Type' column is 'String', and the 'Default value' column is 'Value'. The rest of the interface and status bar are identical to the first screenshot.

Screenshot of the Microsoft Azure Data Factory interface showing the creation of a transformation dataset.

Left Sidebar: Factory Resources > Datasets > ds_transformation

Main Area:

- Name:** ds_transformation
- Type:** Parquet
- Connection:** LS_adls
- File path:** transformedpath / @dataset().FolderName
- Compression type:** snappy

Properties Panel:

- General:** Name: ds_transformation, Description:
- Related:** Preview experience, Off

4) Create Dataflow for Transformation.

1) Select Source for Dataflow.

Screenshot of the Microsoft Azure Data Factory interface showing the creation of a transformation dataflow.

Left Sidebar: Data Factory > Validate all

Main Area: Transformation Dataflow diagram showing a flow from MergedFile through derivedColumn, Aggregation, and a looped processing section (January to June) to sinkJan, sinkFeb, sinkMarch, sinkApril, sinkMay, and sinkJune.

Properties Panel:

- General:** Name: Transformation, Description:
- Related:** Preview experience, Off

Bottom Navigation: Parameters, Settings, New, 31°C Haze, Search, etc.

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

The screenshot shows the Microsoft Azure Data Factory Transformation Data Flow editor. On the left, the 'Factory Resources' sidebar lists various pipelines, datasets, and data flows. In the center, a data flow named 'Transformation_Data_flow' is displayed. The data flow consists of three main stages: 'MergedFile' (Import data from ds_merged_file), 'derivedColumn' (Creating/updating the columns 'continent, indicator, daily_count, date, month'), and 'aggregation' (Aggregating data by 'continent, month' producing columns 'monthly_deaths'). The 'derivedColumn' stage has a 'Columns' section where 'month' is mapped to 'month(date)'. On the right, the 'Properties' panel shows the general settings for the data flow, including its name ('Transformation_Data_flow') and description.

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

The screenshot shows the Microsoft Azure Data Factory Transformation Data Flow editor. The data flow 'Transformation_Data_flow' now includes an 'aggregation' stage after the 'derivedColumn' stage. The 'Aggregate settings' section shows 'Group by' selected, with 'abc continent' and '123 month' listed as group keys. The 'Name as' section shows 'continent' and 'month' respectively. The 'Properties' panel on the right remains the same, with the data flow named 'Transformation_Data_flow'.

4) Use Conditional Split to get data month wise.

The screenshot shows the Azure Data Factory Transformation Data Flow editor. On the left, the 'Factory Resources' sidebar lists various pipelines, datasets, and data flows. In the center, a data flow diagram is displayed. A 'MergedFile' dataset is connected to a 'derivedColumn' component, which then connects to an 'aggregation' component. The output of the aggregation component splits into three parallel paths labeled 'January', 'February', and 'March'. Below the diagram, the 'Conditional split settings' tab is selected, showing a table mapping months to conditions:

Month	Condition
January	month==1
February	month==2
March	month==3
April	month==4
May	month==5
June	month==6
July	month==7

The 'Properties' pane on the right shows the data flow is named 'Transformation_Data_flow'. The bottom status bar indicates the system is at 37°C and sunny.

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

The screenshot shows the Azure Data Factory Transformation Data Flow editor. The 'Factory Resources' sidebar is visible on the left. The main area displays a data flow diagram similar to the previous one, but the 'aggregation' component is followed by a 'rank1' activity, which then branches into '4 Columns'. The 'Rank settings' tab is selected in the properties pane, showing the following configuration:

- Incoming stream:** SplitForMonthwiseData@January
- Options:** Case insensitive, Dense (unchecked)
- Rank column:** Ranking
- Sort conditions:** SplitForMonthwiseData@January's column, Order: monthly_deaths, Descending

The 'Properties' pane shows the data flow is named 'Transformation_Data_flow'. The bottom status bar indicates the system is at 37°C and sunny.

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

The screenshot shows the Microsoft Azure Data Factory Transformation Data Flow editor. On the left, the 'Factory Resources' sidebar lists various pipelines, datasets, and data flows. In the center, a data flow named 'Transformation_Data_flow' is displayed. The data flow consists of several components: a 'derivedColumn' followed by an 'aggregation' step, which then branches into two parallel paths. The top path leads to a 'January' step, then a 'rank' step, and finally a '4 Columns' sink. The bottom path leads to a 'February' step, then a 'rank1' step, and finally a 'sort2' step. After 'sort2', the data splits into two parallel paths again. On the right, the 'Properties' panel is open for the 'Transformation_Data_flow' data flow, showing the name and description fields. Below the main editor, the Windows taskbar shows the date as 07-06-2023 and the time as 17:23.

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

The screenshot shows the Microsoft Azure Data Factory Transformation Data Flow editor. The setup is similar to the previous one, but now includes a 'sink2' component at the end of the bottom path. The 'Properties' panel on the right is configured for the 'Transformation_Data_flow' data flow, with the 'Sink' tab selected. It shows the 'Output stream name' as 'sink1', 'Description' as 'Export data to ds_transformation', 'Incoming stream' as 'sort1', 'Sink type' set to 'Dataset', and 'Dataset' set to 'ds_transformation'. The Windows taskbar at the bottom shows the date as 07-06-2023 and the time as 17:33.

Factory Resources

- Pipelines
 - CopyCovidData-ADLS-Pipeline
 - Merging_Files_Pipeline
 - Transformation_Pipeline**
- Datasets
 - ds_destination
 - ds_merged_file
 - ds_source
 - ds_transformation
- Data flows
 - Transformation_Data_flow
 - Merging_Data_Flow
- Power Query

Activities

- Merging_Files_Pipel...
- ds_merged_file
- ds_transformation
- Transformation_Data...
- Data flow**

Data flow

Properties

General

Name: Transformation_Pipeline

Description:

Annotations: + New

Name	Value	Type
sink1 parameters	FolderName: January	string
sink2 parameters	FolderName: February	string
sink3 parameters	FolderName: March	string
sink4 parameters	FolderName: April	string

Output of this project:

transformedpath - Microsoft Azure

Overview

Authentication method: Access key (Switch to Azure AD User Account)

Location: transformedpath

Search blobs by prefix (case-sensitive):

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
April					-	---
August					-	---
February					-	---
January					-	---
July					-	---
June					-	---
March					-	---
May					-	---
October					-	---
September					-	---

January Output In Parquet Format:

The screenshot shows the Microsoft Azure Storage Explorer interface. On the left, there's a sidebar with various settings and monitoring tools. The main area shows a container named 'transformedpath' containing a single file named 'January.parquet'. The file preview pane shows the first 12 lines of the data, which are highly compressed binary strings. The Azure interface includes a sidebar with various settings and monitoring tools.

January Output:

Parquet File Viewer For Web

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

Offset: 0

n rows: 5

[Prev Page](#) [Next Page](#) [Close File](#)

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

September output:

Parquet File Viewer For Web

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

Offset: 0

n rows: 5

[Prev Page](#) [Next Page](#) [Close File](#)

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

Conclusion:

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