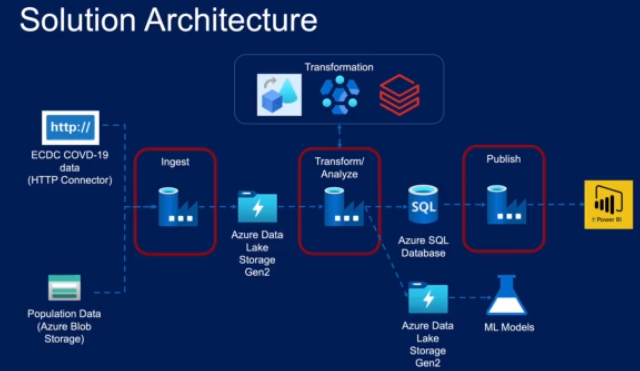
**Covid Project Requirements**

We want to create a data platform from which our data analysts can easily report on the COVID-19 trends using a reporting tool. The solution that we are building will be limited to reporting on the data related to EU countries and the UK only. The data lake that we build will be populated with details about the confirmed COVID-19 cases daily, the unfortunate mortalities as a result of COVID on a daily basis, the hospital admissions, and ICU cases.

We will have both new numbers per week as well as people in hospital at the end of the day, testing details such as tests being carried out per week, and any new COVID cases from the tests. And finally, we'll also get the statistics about the population in every country by age group. This data could then be used by our data scientists to predict the spread of the virus. We'll then populate the data warehouse with the subset of the data so that it can be used for reporting on trends. The data warehouse will include details about confirmed cases, unfortunate mortality rates, hospitalization, and ICU cases from our weekly counts in the data lake, as well as the testing numbers. We'll then build a Power BI report from this data.We'll use the European Center for Disease Prevention and Control website as the source for our confirmed cases, mortality, hospitalization cases, ICU cases, and testing numbers. And we'll use the Eurostat website for population by age data.

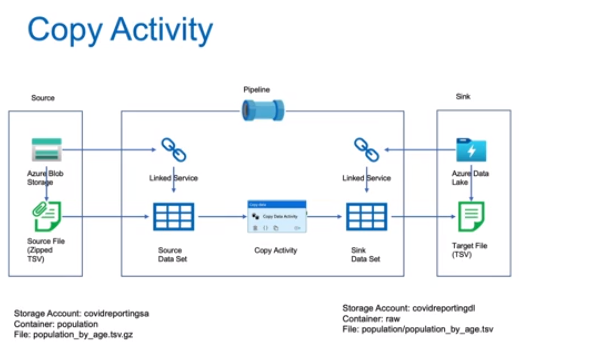
**Solution Architecture**



**Assignment 1:-**

**Data ingestion from Azure BLOB**

Copy activity from azure blob to azure data lake



**Steps followed:-**

Create a linked service for source Azure blob storage .

Create a linked service for target Azure data lake gen2 storage.

Create source dataset.

Create Sink dataset.

Create a pipeline with copy activity and add source and sink datasets created in the above step.

In the pipeline add control activity –validation activity to execute copy activity once file becomes available in source.

In the pipeline add control activity -metadata validation like checking number of columns and if file length is more than certain bytes, execute copy activity only if file contents are as expected.

In the pipeline add control activity - delete activity to delete file at source after copy activity is done

Schedule pipeline execution by creating event trigger.

**Please refer links from Microsoft documentation that you will need for this assignment:-**

**Data Ingestion From Blob**

https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-overview

**Linked Services & Datasets**

https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services

https://docs.microsoft.com/en-us/azure/data-factory/concepts-datasets-linked-services

**Create Pipeline**

https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities

**Control Flow Activities**

**Validation Activity -** https://docs.microsoft.com/en-us/azure/data-factory/control-flow-validation-activity

**Get Metadata Activity -** https://docs.microsoft.com/en-us/azure/data-factory/control-flow-get-metadata-activity

**If Condition Activity -** https://docs.microsoft.com/en-us/azure/data-factory/control-flow-if-condition-activity

**Web Activity -** https://docs.microsoft.com/en-us/azure/data-factory/control-flow-web-activity

**Delete Activity** - https://docs.microsoft.com/en-us/azure/data-factory/delete-activity

**Fail Activity** - https://learn.microsoft.com/en-us/azure/data-factory/control-flow-fail-activity

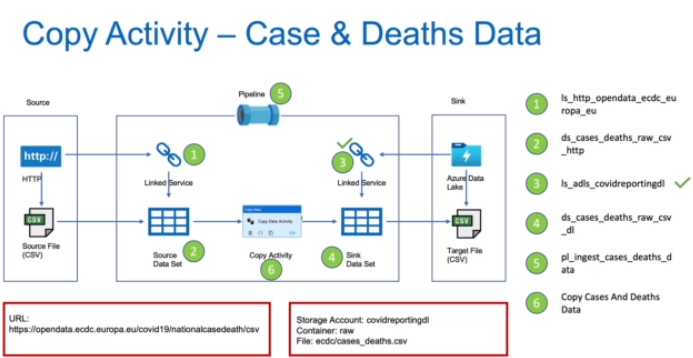
**Triggers -**<https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers>

**Assignment 2** ➖ **Data ingestion from Http**

**Assignment Requirement: -**

Copy the above csv files from http location to azure data lake gen2 storage.

First just copy the cases\_deaths.csv file from http location to Azure data lake Gen2 storage.



Steps to follow:-

Create a linked service for http connection by providing the Base url and Relative URL for only cases\_death.csv file.

**Example of base url and relative url:-**

BaseUrl - https://raw.githubusercontent.com/

RelativeUrl:- for each of the file will be

akjainade/akjain-adf/**cases\_deaths**.csv( change the file name)

Create a linked service for target Azure data lake gen2 storage.

Create source dataset.

Create Sink dataset.

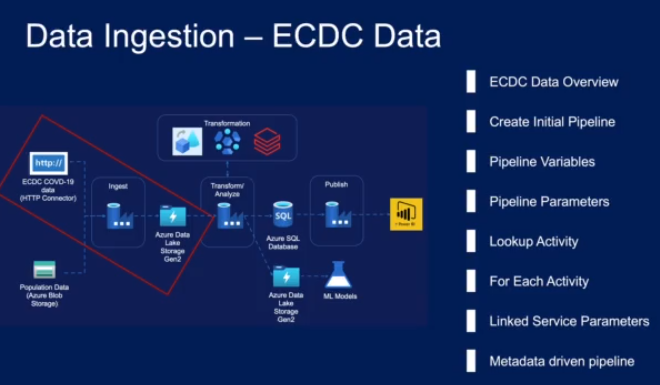
Create a pipeline with copy activity and add source and sink datasets created in the above step.

Debug the pipeline.

**Assignment 3** -

**Data ingestion from http using parameters and variables in the pipeline.**

Automate the process of copying data for multiple files from http location so that we can copy all files using the one single pipeline.



**There files that we will ingest from http link**

Cases\_deaths.csv

Hospital\_admissions,csv

Testing.csv

Json file for metadata driven pipeline attached in different folder.

Steps needed:-

1. Copy hospitalization data , use the same pipeline you created for cases and deaths file by passing file names using parameters and variables.A pipeline can receive a parameter, send that to a data set and the data set can then send that to a linked service. So all three components, pipelines, data sets, and link services can be parameterized.
2. At source dataset -parameterize the relative URL so that you can pass in different values depending on which file you want to copy
3. At sink dataset --parameterize the file name so that the data set can deal with a different file every time it's being called with one
4. At pipleline --create two variables -sourcerelativeurl and sinkfilename , map these variables to parameters in source dataset and sink dataset -this is to test the datasets.Parameterize the pipeline -We need the parameter sink file name for sink and the URL parameter for source.
5. Update copy activity to use parameters
6. Pass parameters from trigger ,create a schedule trigger, attach it to pipeline.

For each file we will have to create a trigger ,so there is a better way to do this.

1. Create a look up activity and source would be the json file which has a list of source url ,relative url and sink file name.
2. Add for each activity after this lookup so that we can execute the pipeline for all the files one by one or in parallel.
3. Add the copy activity inside for each activity to execute for each file. Update the copy activity to get input from items in for each and remove all the parameters that were created earlier.
4. Create a trigger and attach it to the pipeline.

**Please refer links from Microsoft documentation that you will need for this assignment:-**

#### Data Ingestion from HTTP

#### 

#### ECDC Data Overview

Link to ECDC Website - <https://www.ecdc.europa.eu/en/covid-19/data>

#### Control Flow Activities

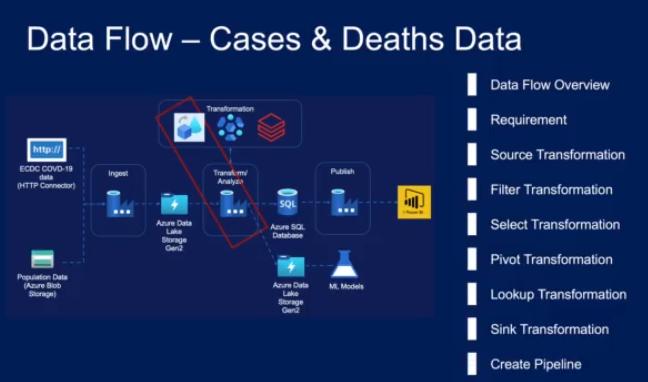
<https://docs.microsoft.com/en-us/azure/data-factory/control-flow-lookup-activity>

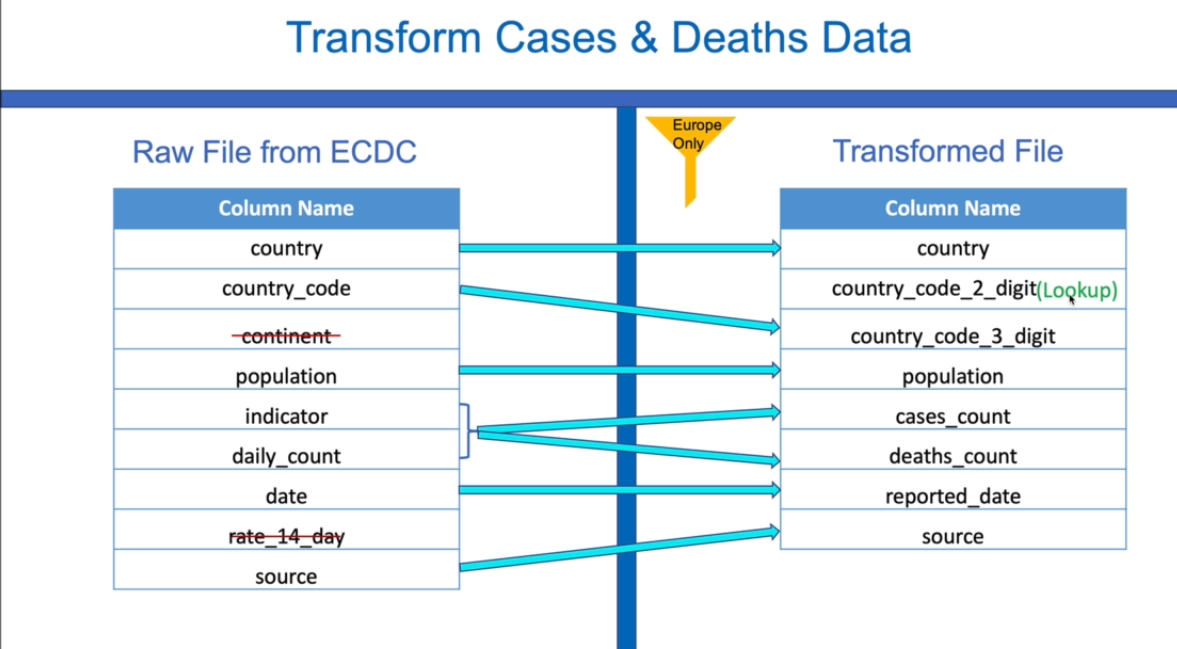
#### Linked Service Parameters

<https://docs.microsoft.com/en-us/azure/data-factory/parameterize-linked-services>

**Assignment 4:**- **Applying transformations using Data flows**

We have to apply transformations on cases\_deaths.csv file that we copied in Assignment2 and create a pipeline to execute the data flow.

****

****

f

**Transformation Required**

1. Get the 2-digit country code and create a new column country\_code\_2\_digit
2. Using indicator and daily count get cases\_count and deaths\_count column
3. Rename date to reported\_date
4. Remove rate\_14\_day, and continent
5. Take all the other columns.

**Please refer links from Microsoft documentation that you will need for this assignment:-**

Data Flow – Cases & Deaths Data Transformation

#### Introduction to Data Flows

Overview - <https://docs.microsoft.com/en-us/azure/data-factory/concepts-data-flow-overview>

Available Regions - <https://docs.microsoft.com/en-us/azure/data-factory/concepts-data-flow-overview>

#### Filter Transformation

Overview - <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-filter>

#### Select Transformation

Overview - <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-select>

#### Pivot Transformation

Overview - <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-pivot>

#### Lookup Transformation

Overview - <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-lookup>

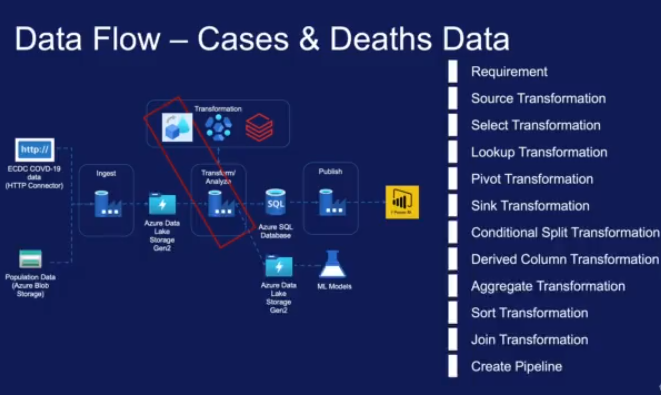
#### Sink Transformation

Overview - <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-sink>

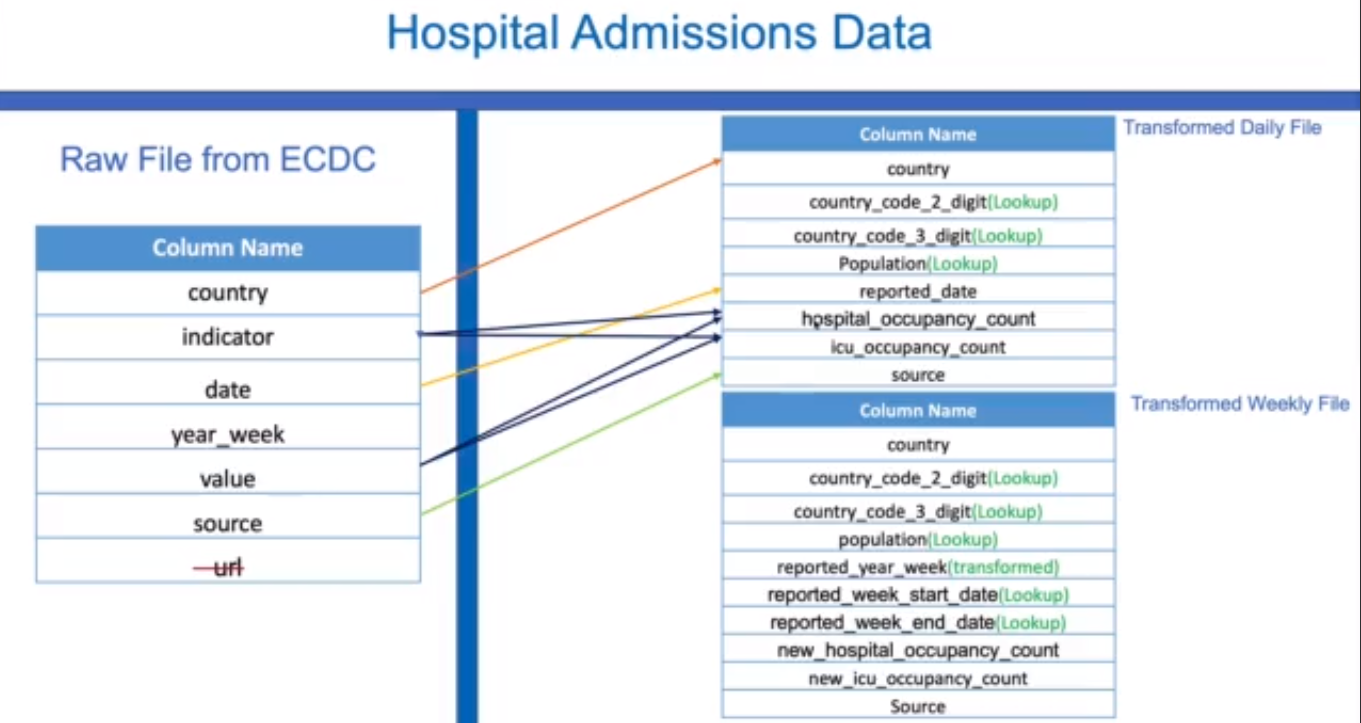
Supported Sink Types -https://docs.microsoft.com/en-us/azure/data-factory/data-flow-sink

**Assignment 5:-**

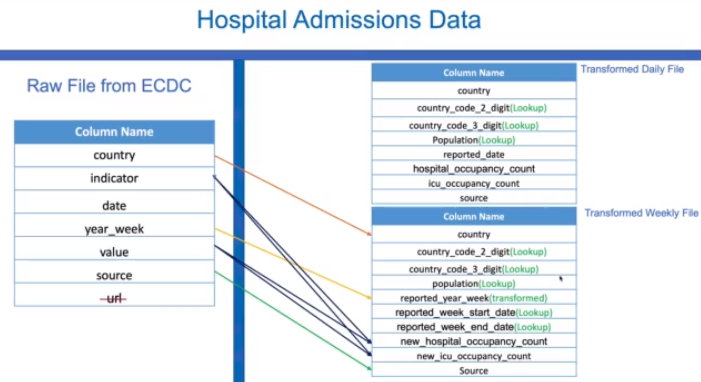
Transformation on hospital admissions file using data flows:-



Daily file requirements:-



Weekly File Requirements:-



1. Select required columns (remove the url column) and also rename the columns **date** to **reported\_date** and **year\_week** to **reported\_year\_week**
2. Add country source and then add lookup transformation to get country from country dataset. Lookup transformation gets input from step2 and country dataset.
3. Add select transformation to select required columns from step 3 and select duplicate columns only from one table and remove the columns that are not needed.
4. Add conditional split transformation to split Daily and weekly data based on Indicator column
5. Weekly stream we have to get the week start date and week end date ,so we will use the date dim table to get that. Add another source transformation to add the date dim table to further use it in derived transformation to transform some columns. Add new column ecdc\_year\_week as format in hospital admissions file is different and we need to transform column in date\_dim to be able to join. And derive it from concatenating existing columns year + “-W’ + lpad(year\_of\_week,2,’0’)
6. Add aggregate transformation to create one row per week Groupby ecdc\_year\_week and add new columns week\_start\_date = min(date) and week\_end\_date=max(date)
7. Add join transformation to join Weekly stream with date dim output in step7 on ecdc\_year\_week and reported\_year\_week columns.
8. We need to get hospital occupancy count and icu occupancy count based on indicator column and sum(value) column will be the count. So we need to add pivot transformation for this for both the weekly stream as well as daily stream.
9. Add sort transformation to get most recent data for both weekly and daily stream to be at the top.
10. Add select transformation to select columns that we need in final output and rename the columns as needed in output.
11. Add sink transformation and create the data file in processed in adls gen2 storage for weekly and daily both.
12. Add the data flow in a pipeline to be able to invoke it.

**Links from microsoft documentation:-**

#### Data Flow - Hospital Admissions Data Transformation

#### Conditional Split Transformation

Overview - <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split>

#### Derived Column Transformation

Overview - <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column>

#### Aggregate Transformation

Overview - <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-aggregate>

#### Join Transformation

Overview - <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-join>

#### Sort Transformation

Overview - <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-sort>

**Assignment 6:- Transforming the testing file**

1. We need to add two- and three-character country codes using country dimension.
2. We need to add week start date and week end date using date dim.

#### Create ADF Pipeline Databricks Notebook Activity

<https://docs.microsoft.com/en-us/azure/data-factory/transform-data-using-databricks-notebook>

**Assignment 7:- Copy data to Azure Sql from ADLS Gen2**

Copy the files created in processed folder in ADLS Gen2 to Azure SQL

**Links from Microsoft Documentation:-**

#### 

#### Copy Data to Azure SQL

Copy Activity - <https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-overview>

Azure Data Studio download - <https://docs.microsoft.com/en-us/sql/azure-data-studio/download-azure-data-studio?view=sql-server-ver15>