**Cohort 4.1:**

Requirements:

Source : ADLS/RAW/ecdc\_data.zip

Staging layer : ADLS/Refined/<Individual folder for each file>/<respectivefile.parquet>

Target: ADLS/Refined/Delta tables

Problem statement: Fetch data from source url & load in staging raw layer as zip file & processed or clean data that has been arrived in last 24 hrs from source & load into ADLS refined container in parquet file as delta table while following below instruction:

Step 1 - Data preparation (use ADF):-

* Unzip file with help of ADF/ADB & load under raw<participant\_name>/Unzippeddata/file.
* Fetch data that has been processed in last 24 hrs from source & load into ADLS refined container in parquet file with respective folder without using parameter table.
* Check if the file is available in the path. If it’s not available, there should be timeout after 1 minute.

Step 2 - Data transformation(use ADB):

Write a code to extract, transform and load the data from source to target. For ETL, use Python, Panda, Pyspark etc.

* If date is NULL or blank, give default date as ‘2020-11-28’. Format of date column should be YYYY-MM-DD.
* Display which country has more cases with respect to population.
* Remove the entries which has URL field value as ‘ERROR’.
* Transform the values of column country with their acronyms. For eg: Austria would be replaced by ‘AUST’, Belgium by ‘BELG’ etc. You can find data in below url:

[https://travel.state.gov/content/travel/en/us-visas/visa-information-resources/fees/countryhttps://travel.state.gov/content/travel/en/us-visas/visa-information-resources/fees/country-acronyms.htmlacronyms.html](https://travel.state.gov/content/travel/en/us-visas/visa-information-resources/fees/country-acronyms.html)

Note: Use only two resource : adls, ADF,ADB.

**Deliverables**:

1. Architecture diagram of flow
2. Best Practice in Pipeline
3. Best Practice for Database
4. Best Practice for ETL
5. Target Tables