Cohort 2.1:

Requirements:



Source: ADLS/RAW/Customer\_source.csv

Target: ADLS/Refined/deltatable\_customer.Parquet

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, Py spark etc.

* If date is NULL or blank, give default date as ‘2020-11-28’. Format of date column should be YYYY-MM-DD.
* In Gender column “F” & “M” has been given that needs to replace f-> Female & M-> Male.
* Remove ‘\_’ from column header.
* In some of the records, the names are mixed case (e.g., "Carol"), while in others, they are uppercase (e.g., "CAROL").
* The phone numbers aren't consistent, either. Some of them are hyphenated (e.g., "992-83-4829"), while others are missing hyphens ("992834829") & rename column phone with cellphone.

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