**Cohort 5.1:**

Requirements:

Source:

https://www.stats.govt.nz/assets/Uploads/Business-operations-survey/Business-operations-survey-2022/Download-data/business-operations-survey-2022-information-and-communications-technology.csv

https://www.stats.govt.nz/assets/Uploads/Business-operations-survey/Business-operations-survey-2022/Download-data/business-operations-survey-2022-business-finance.csv

Target: ADLS/Refined/<file>. 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): -

* Fetch data from given URL 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 ADF/ADB):

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

* show maximum & Mnium values for each industry.
* show what are difference between Wood & paper product or Agriculture with respective to value.
* Remove '\_' from header.

Setup email configuration on failure & success of pipeline to get notify.

**Deliverables**:

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