**Capstone Project**

Case Study -

U F H

Its an ecommerce company dealing in organic groceries.

They are still an early stage startup. They operate in 20 Cities in India at this moment accross 6 States. They promise to deliver the order within 48 hours after order placement. They offer 50 different type of products in different SKUs.

At this moment they recieve between 500-1000 orders per day across India.

They soon plan to expand their capabilities to serve between 10000 to 20000 orders per day.

Work Flow-

Diagram

Description automatically generated

1. Create MySQL cloud SQL instance on GCP.
2. Create a Database in this instance and add a network under Connections section using the VM’s public IP.
3. Create an OLTP schema using MySQL
   1. Create an SQL script containing the queries for table creation and import it into your SQL instance.
   2. Customer\_Master –
      1. Customerid – Primary key
   3. Product\_Master –
      1. Productid – Primary key
   4. Order\_Details –
      1. Customerid – Foreign key
      2. (Orderid, Order\_status) – Primary key
   5. Order\_Items –
      1. (Orderid, Productid) – Primary key
      2. Orderid – Foreign key
      3. Productid – Foreign key
4. Create engine from sqlalchemy library to establish connection between python script and database.
5. Generate fake data using faker library and store in dataframe.
6. Insert the fake data generated into OLTP schema using dataframe.to\_sql method.
7. Create tables defined in Star Schema in Big Query(Create an SQL script containing the queries for table creation and run it in bigquery editor)
8. Build the ETL pipeline
   1. Create a service account giving Bigquery admin rights and generate a key for the same. Use this key to establish connection between python script and bigquery
   2. Create engine from sqlalchemy library to establish connection between python script and database.
   3. Extract data from cloud SQL tables into dataframes.
   4. Transform this data as per the star schema tables using pandas dataframe.
   5. Using pandas\_gbq library push the data from dataframes into bigquery tables.
9. Run incremental ETL to load 5000 more records and populate them into oltp and then in bigquery
   1. Create engine from sqlalchemy library to establish connection between python script and database.
   2. Extract data from cloud SQL tables into dataframes and get the last updated data.
   3. Generate fake data using faker library and store in dataframe.
   4. Append the fake data generated into OLTP schema using dataframe.to\_sql method.
10. Perform analysis on the transferred data using BigQuery.