**Uber Data Engineering Project:**

The task here is to create a data model(lucid/draw.io). Replicate the data model diagram in actual extraction through API, transformation/cleansing and loading back to Bigquery. The data from Bigquery can further be used for BI or building ML models.

Summary

Transformation logic in Python

Data storage in object storage

Mage – for deploying, scheduling, and triggering pipeline.

Load data on Bigquery

Looker for visualization

------------------------------------------------------------------------------------------------------------------------

GCP Services

Google cloud storage (GCS)– object storage

S3- object storage for AWS

Blob – Object storage for Azure

Compute engine- EC2 on AWS, Google compute engine for Google

Bigquery – Google cloud based

Looker and Data Studio is combined to create Looker Studio.

Mage is similar to Airflow

Fact and Dimension Table

Fact Tables

1. Contains quantitative metrics that are used for analysis.
2. Typically contains foreign keys that link to dimension tables.
3. Contains columns that have high cardinality and change frequently.
4. Columns are useful for analysis.

Dimension Tables

1. Contains columns that are attributes of the data being analyzed
2. Contains primary keys
3. Columns have low cardinality and don’t change frequently
4. Contains columns that can be used for grouping or filtering

# create a project

# Create google cloud storage account – upload data

# create a compute engine and click on ssh. Once the machine is up we can start working

# install pip, pandas and mage and start mage

# go to vm on google vm and get external ip with port available on vm itself

# it will not allow us to access as instance doent not accept any request from this port.

# We loaded the data using python data loader via api

# transformed the data using transformation python and returned a dictionary

# Now export the data using exporter to bigquery

# Service and API on console will contain all the details related to connections

# Creating a service account from API and services we give permission to connect VM with other GCP services. Generate key for big query.

# Add project details to config.io using json key generated in above step.