**Uber Data Analytics | Modern Data Engineering GCP Project**

**Introduction**

This project showcases an end-to-end ETL (Extract, Transform, Load) pipeline for processing Uber trip records data. The pipeline retrieves data from the TLC Trip Record Data provided by New York City's Taxi and Limousine Commission (TLC) for yellow and green taxi trips. It includes various fields capturing pick-up and drop-off dates/times, locations, distances, fares, rate types, payment types, and passenger counts reported by drivers.

**Dataset**

The dataset used in this project is the TLC Trip Record Data for yellow and green taxi trips. It offers a comprehensive collection of information related to Uber trips in New York City. To access the dataset, please follow the link below:

Uber Data CSV - <https://storage.googleapis.com/uber-data-engineering-project-sameer/uber_data.csv>

For a deeper understanding of the dataset's structure and field descriptions, refer to the following resources:

1. Website - <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
2. Data Dictionary - <https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf>

**Architecture**

[A picture containing text, software, font, screenshot

Description automatically generated](https://github.com/megs1110/uber-etl-pipeline/blob/main/architecture.jpeg)

**Technology Used**

* Programming Language - Python
* Programming Language - Sql

Google Cloud Platform

1. Google Storage
2. Compute Instance
3. BigQuery
4. Looker Studio

**Project Steps**

This project follows a step-by-step process to build the Uber ETL pipeline:

**Step 1: Data Acquisition**

To start, obtain the TLC Trip Record Data for yellow and green taxi trips. You can download the dataset using the provided link or visit the official NYC TLC Trip Record Data website. This step ensures you have the necessary data to proceed with the pipeline.

**Step 2: Data Modeling**

In this step, you will learn about data modeling concepts such as fact and dimension tables. Fact tables contain core numerical data, while dimension tables provide descriptive attributes to provide context for the fact data. Understanding these concepts is crucial for designing an effective data model that enables efficient analysis and reporting.

[A picture containing text, diagram, parallel, font

Description automatically generated](https://github.com/megs1110/uber-etl-pipeline/blob/main/data_model.jpeg)

**Step 3: Data Table creation**

Using Jupyter Notebook, the data is read and transformed. This step involves performing various transformations such as data cleaning, dropping duplicates, reindexing, and restructuring the dataset.I learned how to apply these transformations to ensure data quality and consistency for downstream processing.

**Step 4: Data Loading,Transformation & Extraction**

The data is loaded into Google Storage, a scalable and reliable cloud storage solution. Additionally, a Compute Instance is set up to install and run MageAI, a tool used for data extraction, transformation, and loading. This step introduces me to the process of loading data into a cloud storage system, which is a crucial component of the ETL pipeline.

Modern Data Pipeine Tool - <https://www.mage.ai/>

**Step 5: BigQuery Analytics**

Leveraging the power of Google BigQuery, a fully-managed data warehouse, this step focuses on executing SQL queries to create the final analytical table.I learned how to use SQL queries to perform advanced analytics on the Uber trip data. BigQuery provides efficient data storage, retrieval, and querying capabilities, enabling me to gain valuable insights from the data.

**Step 6: Dashboard Visualization**

The project concludes by creating an interactive dashboard using Looker, a powerful data visualization and business intelligence platform. The dashboard visualizes the analytical results, making it easier to derive insights and present the findings to stakeholders. Through this step, I learned how to create visually appealing and informative dashboards that facilitate data exploration and interpretation.

[A screenshot of a computer

Description automatically generated with medium confidence](https://github.com/megs1110/uber-etl-pipeline/blob/main/Uber_Dashboard%20_page-0001.jpeg)

**Challenges Faced**

During the project, several challenges were encountered. One of the major difficulties arose from system limitations with MageAI. Due to system constraints, MageAI would frequently shut down and require restarts, resulting in time delays during the data transformation and loading process. These delays impacted the overall efficiency of the pipeline and required additional troubleshooting and optimization.

To mitigate these challenges, it is advisable to allocate sufficient system resources and optimize the MageAI environment to ensure smooth and uninterrupted processing. Additionally, implementing error handling mechanisms and monitoring systems can help detect and resolve issues promptly, minimizing the impact on the pipeline's performance.

Feel free to explore the project and leverage the provided resources to enhance your understanding of the Uber ETL pipeline and data engineering concepts. Happy coding!