**Data Engineering Project: NYC Yellow Taxi Trip Data ETL Pipeline**

This project focuses on building a robust data pipeline to process and analyze NYC Yellow Taxi trip records using modern cloud-based tools and frameworks. The dataset was sourced from the [NYC Taxi & Limousine Commission's Trip Record Data](https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page), specifically the records for **January 2023** and **December 2024**, totaling over **5 million records**.

**Objectives**

1. **Data Ingestion**: Extract the data from the source files into the ETL framework using **Mage AI**.
2. **Data Transformation**: Cleanse, transform, and preprocess the dataset to ensure data quality, uniformity, and usability for downstream analysis.
3. **Data Storage**: Store the processed data in **Google Cloud Storage (GCS)** for scalability and durability.
4. **Data Loading**: Load the transformed data into **Google BigQuery**, enabling efficient querying and analysis.
5. **Insights Generation**: Use BigQuery's analytical capabilities to generate insights into taxi trip patterns, revenue trends, and peak demand times.

**Tools and Technologies**

* **Mage AI**: Used to orchestrate the ETL process with a pipeline to automate data extraction, transformation, and loading.
* **Google Cloud Storage**: Served as the intermediate storage solution for the transformed data files.
* **Google BigQuery**: A cloud-based data warehouse for storing and querying the final dataset.
* **Python**: For scripting and integrating Mage AI with the cloud services.

**Pipeline Workflow**

1. **Extract**:
   * Downloaded trip records in .csv format for January 2023 and December 2024.
   * Ingested these files into Mage AI's pipeline.
2. **Transform**:
   * Performed data cleaning (e.g., handling missing values, correcting data types).
   * Generated additional features, such as trip duration, fare per mile, and peak hour flags.
3. **Load**:
   * Transformed data was stored as .parquet files in Google Cloud Storage.
   * These files were subsequently loaded into BigQuery using optimized schemas for querying.

**Outcomes and Insights**

* Created a scalable pipeline capable of handling large datasets efficiently.
* Generated key insights, including:
  + **Revenue Analysis**: Monthly revenue patterns and peak income sources.
  + **Demand Trends**: Identification of peak demand hours and busy zones.
  + **Operational Metrics**: Average trip duration, fare per mile, and trip distances.

This project demonstrates the integration of modern data engineering practices and tools to solve real-world data challenges effectively.