Tiny Yellow Cabs

Tiny Yellow Cabs is a Python program developed for the Data Engineer interview process at Tinybird. It fetches data from the NYC Yellow Taxi Trips Dataset and allows users to extract a subset based on specific criteria, saving the results in `.parquet` format. For 2024, data from January to June is used.

# Table of Contents

1. Installation  
2. Cloning the Repository  
3. Running the Program  
4. Project Approach  
5. Dataset Considerations  
6. Helpful Resources

## Scope

# Project Approach

## Ingestion Process

- \*\*Ingestion.py\*\* loads the data into a PostgreSQL database and processes records from `.parquet` files.  
- The script validates and inserts the data into a table named `yellow\_tripdata\_total\_records`.

## Data Retrieval

1. **Fetching total record count**: It retrieves the total number of records from a table named yellow\_tripdata\_total\_records using the SQL query SELECT count FROM yellow\_tripdata\_total\_records;. This value represents the total number of trip records.
2. **Calculating the 10% threshold**: After fetching the total count, the script calculates 10% of this total using math.ceil(total\_count \* 0.1). This ensures that the result is rounded up to the nearest integer, which represents the number of records that correspond to the top 10% of trips based on distance.
3. **Extracting the top 10% of trips**: It then fetches the top 10% of trips from the mv\_yellow\_tripdata\_desc\_trip\_distance materialized view. The query SELECT \* FROM mv\_yellow\_tripdata\_desc\_trip\_distance LIMIT {limit}; limits the number of rows retrieved based on the 10% threshold previously calculated.
4. **Exporting to CSV**: The selected rows are converted into a Pandas DataFrame with columns 'VendorID', 'pickup\_datetime', 'pep\_dropoff\_datetime', and 'trip\_distance'. This DataFrame is then exported to a CSV file named top\_10\_percentile\_trips.csv.  
   - \*\*db\_utils.py\*\*: This script is responsible for connecting to the PostgreSQL database.

## Additional Considerations

For a real-world solution, the system could be expanded to an event-driven design where file ingestion is triggered automatically as files arrive in the ingestion folder.

## Client Feedback

- \*\*Data Range\*\*: The solution currently processes 2024 data (January to June). Expanding this to historical data would be ideal.  
- \*\*Tools\*\*: Python with Pandas and PostgreSQL is utilized as per client requirements, avoiding Jupyter notebooks.  
- \*\*Data Quality\*\*: Data quality checks are in place.  
- \*\*Reporting\*\*: The top 10% of trips (by distance) are captured in the CSV file `top\_10\_percentile\_trips.csv`.

# Installation

Ensure you have Python 3.12.1 installed. You can download it from the [official Python website](https://www.python.org/downloads/).  
To verify the correct version, run:

$ python3 --version  
Python 3.12.1

# Cloning the Repository

Clone the repository and navigate to the project directory:

git clone https://github.com/rscottlundgren/tiny-yellow-cabs.git  
cd tiny-yellow-cabs

# Running the Program

1. Install required dependencies:  
pip3 install -r requirements.txt

2. To run the ingestion script, execute the following command with the necessary arguments (e.g., ingestion):  
python Ingestion.py yellow\_tripdata\_2024-01.parquet yellow\_tripdata\_2024-02.parquet yellow\_tripdata\_2024-03.parquet yellow\_tripdata\_2024-04.parquet yellow\_tripdata\_2024-05.parquet yellow\_tripdata\_2024-06.parquet

3.To run the retrieval script

python Retrieval.py