Data Engineering Challenge

Description

This challenge will evaluate your proficiency in Data Engineering, and your knowledge in Software development as well.

Assignment

Your task is to build an automatic process to ingest data on an on-demand basis. The data represents trips taken by different vehicles, and include a city, a point of origin and a destination.

This CSV file gives you a small sample of the data your solution will have to handle. We would like to have some visual reports of this data, but in order to do that, we need the following features.

Mandatory Features

- There must be an automated process to ingest and store the data.
- Trips with similar origin, destination, and time of day should be grouped together.
- Develop a way to obtain the weekly average number of trips for an area, defined by a bounding box (given by coordinates) or by a region.
- Develop a way to inform the user about the status of the data ingestion without using a polling solution.
- The solution should be scalable to 100 million entries. It is encouraged to simplify the data by a data model. Please add proof that the solution is scalable.
- Use a SQL database.

Bonus features

- Containerize your solution.
- Sketch up how you would set up the application using any cloud provider (AWS, Google Cloud, etc).

Deliverables

Your project should be stored in a public GitHub repository. When you are done, send us the link to your repo.

It's not necessary to host this application anywhere (although you can if you like). Just make sure your repo has a README.md which contains any instructions we might need for running your project.