Data Engineer Assignment

In this task, you will have to use Yellow Taxi Trip dataset to perform simple data processing and deliver insights. We would prefer you to use Python for this assignment, but it is not mandatory. You should use the database of your choice (SQLite, Postgres, MySQL, etc.), hosted locally (on local OS or using developer tools like Docker) or in any public cloud. The code should include README.md file with documentation about how to run the code and any design, technology decisions that are important.

Please download the data from <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page> for 2020 January Yellow Taxi Trip Records (CSV). The exact link to the dataset is here: <https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2020-01.csv> .

You can find more information about what the data means in the "Data Dictionaries and MetaData" section.

Consider 4 hours as a rough estimate guidance on how long it should take to complete the assignment. There is no official time limit for this task. Please feel free to skip any number of tasks or only provide a pseudo-code of how you would approach the problem if you feel that the questions are outside of your knowledge.

Questions:

1. Imagine that you receive data in single day batches, based on the field “tpep\_dropoff\_datetime”. Split the .csv file into multiple files based on this field. You should end up in roughly 30 files. Consider using a more efficient file format than csv.

2. Your next task is to implement a process that loads all of the files into the database. It would be beneficial that the implementation is reusable, i.e. can be used for the future data files as well.

3. Test the quality of the data. What metrics would you use to monitor the quality of the data? Are there any anomalous or suspicious data points in the data? We would like to collect data quality metrics of processed data each time it is loaded into the database. Could you implement this functionality and store such metrics in a separate database table or any other place that you seem fit?

4. Model the dataset to represent a fact-dimension type of schema. Draw the entity relationship diagram of your model and transform the data to fit it.

5. Can you identify the weak points of the proposed solution? If you had more time, what parts of your solution would you improve?