DE Case Study Q1

Chih-Chieh Lin

November 22, 2021

1 How should we store the data for each part of the process

For the sourcing, preprocessing, matching process, I would store the data in csv or the format can be easily utilized in Python script. It provides us more convenient way to process the data than doing it in SQL script.

2 What database technology should be used for delivery?

SQLite is utilized in my implementation. The reasons for choosing this technology are followed:

- SQLite is suitable for smaller project which does not require large scalability.
- Uses standard SQL syntax
- Easily be connected with python
- Only 5 data types which quite fits this project

As for the project which requires more scalability or demands security features, I would consider MySQL instead.

3 How would you model the database?

I come up with a relational database that is shown in Figure 1. For tripadvisor, outletid (restaurant id) and username are critical and related to each other, thus these two items are used to connect tables in database. For ubereats, only the outletid can link to each other.

4 How should the processes that we use to go from the raw data to the delivery database be run?

1. First, transform original json file into dataframe in Python
2. Do sourcing, preprocessing, matching in Python script before modeling database
3. Create relational database by SQL with processed data

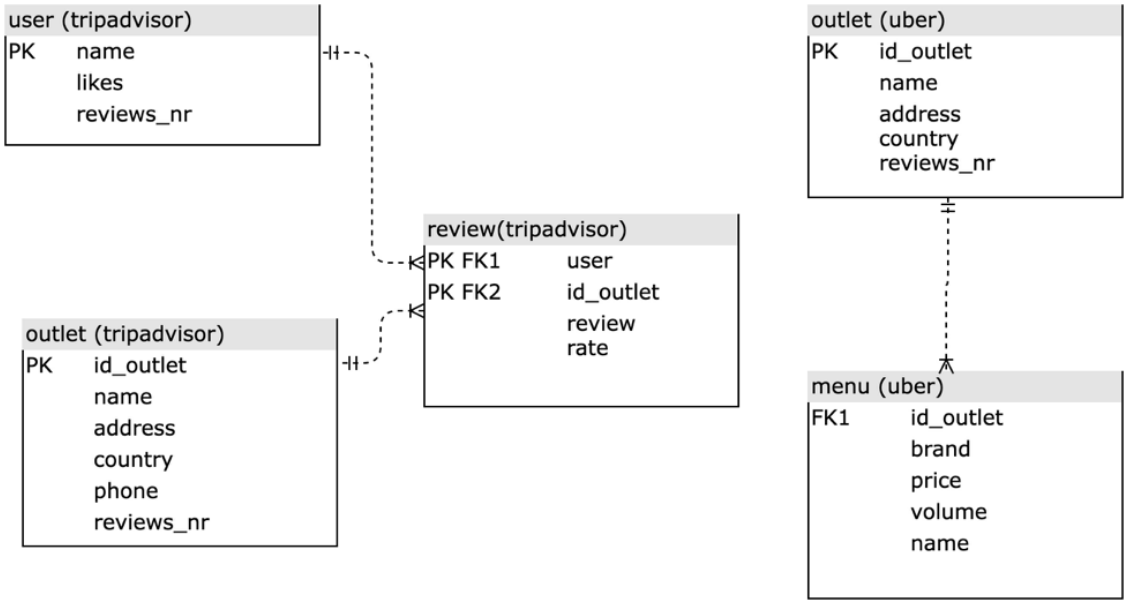

Figure 1: Relational database