Итоговая аттестация по курсу «Инженер данных» Анализ поездок такси в Нью-Йорке

Валерий Петрунин
27 декабря 2022 г.

Содержание

1. **Описание проекта.**
2. **Цели проекта с описание бизнес-задачи.**
3. **Требования.**
4. **Входные данные.**
5. **План реализации.**
6. **Используемые технологии с обоснованием.**
7. **Схемы/архитектуры с обоснованием.**
8. **Выводы.**

1. Описание проекта

Сегодня желтое нью-йоркское такси - это широко узнаваемый символ города.
На данный момент в Нью Йорке работает более 13 000 желтых такси и 50 000 водителей.
Пассажиропоток составляет 600 000 человек в день и 236 миллионов в год.

2. Цели проекта с описание бизнес-задачи

Необходимо, используя таблицу поездок для каждого дня рассчитать процент поездок по количеству человек в машине (без пассажиров, 1, 2,3,4 и более пассажиров). По итогу должна получиться таблица с колонками date, percentage_zero, percentage_1p, percentage_2p, percentage_3p, percentage_4p_plus.

Добавить столбцы к предыдущим результатам с самой дорогой и самой дешевой поездкой для каждой группы.

3. Требования

Все операции должны считаться локально.

Технологический стек – sql,scala (что-то одно).

Подготовить мини-отчет по качеству входных данных.

4. Входные данные.

Таблица, состоящая из поездок такси в Нью-Йорке(в csv файле).

Поле	Описание
VendorId	ИД компании
Trep_pickup_datetime	Время и дата, когда пассажир сел в такси
Trep_dropoff_datetime	Время и дата, когда пассажир вышел из такси
Passanger_count	Количество пассажиров
Trip_distance	Пройденное расстояние
Ratecodeid	Код скорости
Store_and_fwd_flag	Флаг, отвечающий за сохранение записи поездки перед ее отправкой поставщику
PulocationId	Широта, где была начата поездка
Dolocationid	Долгота, где была начата поездка
Payment_type	Тип оплаты
Fare_amount	Стоимость поездки
Mta_tax	Комиссия автопарка
Tip_amount	Чаевые
Tools_amount	Оплата за платные дороги
Improvement_surchange	Доплата за страховку
Total_amount	Полная стоимость поездки
Congestion_surchange	Дополнительный сбор

5. План реализации

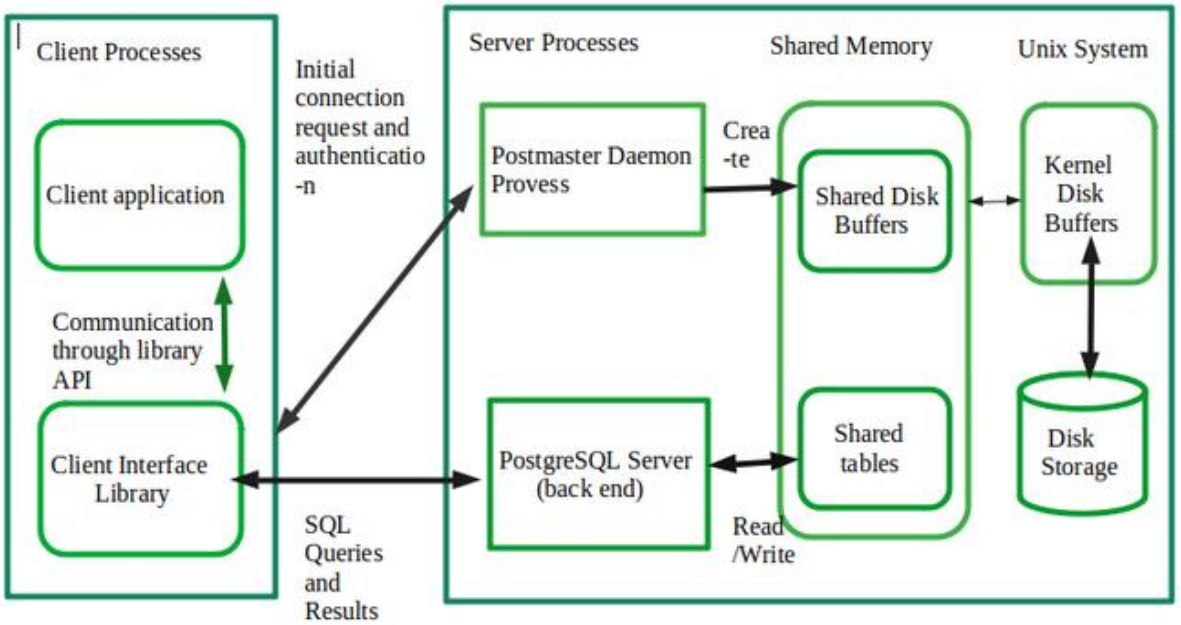
1. Загрузка данных.
2. Предобработка.
3. Анализ данных.
4. Выводы.

6. Используемые технологии с обоснованием.

Для решения этой задачи идеально подходит библиотека pandas на языке Python. И при этом по условиям выполнения проекта в нем нельзя пользоваться Python. Поэтому данный проект я выполнил на языке SQL в PostgreSQL.

7. Схемы/архитектуры с обоснованием.

В проекте я использовал стандартную БД PostgreSQL. В качестве клиента использовалась программа pgAdmin4.



8. Выводы

Была получена таблица поездок для каждого дня рассчитать процент поездок по количеству человек в машине.

Добавлены столбцы с самой дорогой и самой дешевой поездкой для каждой группы;

Наибольшее количество поездок с одним пассажиром.