Домашнее задание. Практическая работа

Дисциплина

Системы хранения данных

Тема

Обработка данных в Data Lake

Форма проверки

Домашнее задание с самопроверкой.

Совет: выполняйте домашнее задание сразу, как только изучите тему.

Имя преподавателя

Вадим Заигрни

Время выполнения

120 минут.

Цель задания

Научиться обрабатывать данные на Spark

Инструменты для выполнения ДЗ

Apache Spark/Spark в Yandex Cloud

Правила приёма работы

- Прикрепите файл с кодом с результатами выполнения задания в LMS

Важно:

- убедитесь, что к файлу есть доступ;
- название файла должно содержать фамилию и имя студента, номер и название ДЗ.

Чек-лист самопроверки

Задание считается выполненным, если:

- Прикреплён файл с выполненным заданием;
- Файл содержит выполненные действия по заданию;
- доступ к материалам открыт.

Задание не выполнено, если:

- файл с заданием не прикреплён или отсутствует доступ к нему;
- файл не содержит выполненные действия по заданию.

Дедлайн

20.11.25

Описание задания

Для выполнения задания вам понадобится набор данных о поездках «жёлтого» такси (yellow taxi) в Нью-Йорке:
Набор данных
Описание набора данных.

Задание

1. Скачайте данные о поездках за первые полгода 2025 года в формате parquet:

```
- https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2025-01.parquet
- https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2025-02.parquet
- https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2025-03.parquet
- https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2025-04.parquet
- https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2025-05.parquet
- https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2025-06.parquet
```

2. Загрузите данные о поездках в один набор данных (dataframe)

3. Очистите набор данных от записей, в которых:
   - время посадки меньше 1 января 2025 года
   - время посадки больше 30 июня 2025 года
   - время высадки меньше 1 января 2025 года
   - время высадки больше 30 июня 2025 года
   - дистанция поездки меньше или равно нулю
   - количество пассажиров меньше или равно нулю

4. Добавьте в набор данных колонки с часом посадки и высадки.

5. Оставьте в наборе данных только колонки:
   - время посадки
   - время высадки
   - количество пассажиров
   - дистанция поездки
   - идентификатор зоны посадки

- идентификатор зоны высадки
- полная стоимость поездки
- час посадки
- час высадки

6. Скачайте данные о зонах в формате CSV:

7. Соедините данные о поездках с данными о зонах так, чтобы в результате в наборе оказались названия зон посадки и высадки вместо идентификаторов зон.

8. Соберите набор данных с часовой агрегацией количества заказов по зоне посадки.

9. Соберите набор данных со средним количеством заказов в каждой зоне.

- Набор должен содержать колонки: зона посадки и часы от 0 до 23, т.е. должно быть 25 колонок.
- Каждая строка набора — одна зона.
- На пересечении строк и столбов находится среднее количество заказов в зоне в час.

Подсказка: надо использовать функцию pivot()

10. Сохраните итоговый набор данных в формате parquet в режиме перезаписи.

Прикрепите файл с кодом с результатами выполнения задания в LMS