Лабораторная работа № 21

Тема: Основы работы с Apache Spark. Обработка данных в формате json.

Цель: изучить основные возможности Apache Spark для работы с большими объемами данных. Научиться загружать, обрабатывать и агрегировать данные в формате JSON.

Задание:

1. Загрузите в кластер файл flights.json.

В файле flight.json содержатся данные о рейсах. Каждая строка представляет собой JSON-объект с информацией о конкретном рейсе. Вот описание каждого поля:

- FL_DATE: дата рейса в формате YYYY-MM-DD.
- DEP_DELAY: задержка вылета в минутах. Если значение положительное, это означает, что рейс вылетел позже запланированного времени. Если отрицательное — рейс вылетел раньше.
- ARR_DELAY: задержка прибытия в минутах. Положительное значение означает прибытие позже запланированного времени, отрицательное — прибытие раньше.
- AIR_TIME: время в пути в минутах (время, которое самолет провел в воздухе).
- DISTANCE: расстояние полета в милях.
- DEP_TIME: фактическое время вылета, представленное в виде десятичного числа, где целая часть — это час, а дробная — доля часа. Например, 9.083333 соответствует 9:05 утра.
- ARR_TIME: фактическое время прибытия, также в виде десятичного числа.

2. В Jupyter notebook создайте Spark-сессию. Прочитайте загруженный файл. Получите схему датасета и определите типы данных. Посчитайте общее количество записей в датасете. Выведите первые 5 строк.

3. Очистка данных. Удалите записи, в которых отсутствуют значения в ключевых полях (например задержка или расстояние у рейсов).

4. Фильтрация данных. Выведите только строки, относящиеся к определенной дате.

5. Выведите только определенный набор колонок (например: дата и задержка вылета).

6. Анализ данных. Вычислить среднюю задержку вылета и прибытия по каждому дню. Найти 5 самых длинных по расстоянию рейсов.

Отчет должен содержать (см. образец):
- номер и тему лабораторной работы;
- фамилию, номер группы студента и вариант задания;
- скриншоты подтверждающие выполнение заданий.

Отчеты в формате **pdf** отправлять на email:
colledge20education23@gmail.com