Лабораторная работа № 20

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

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

Задание:

1. Загрузите в кластер файл cars.csv.

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. Фильтрация данных. Выведите только строки, относящиеся к определенной марке автомобиля (например Porsche).

4. Выведите только определенный набор колонок (например: марка, год выпуска, пробег, цена).

5. Анализ данных. Определите 5 марок автомобилей с наибольшей средней стоимостью. Найдите минимальную и максимальную цену для каждой марки.

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