Лабораторная работа № 22

Тема: Обработка данных с помощью Pandas на кластере PySpark.

Цель: Научиться работать с данными с помощью Pandas и PySpark, выполняя очистку, преобразование и анализ данных на кластере.

Задание:

1. Подготовка и загрузка данных. Загрузите предоставленный CSV или JSON файл (например, данные о продажах, статистика пользователей, информация о полетах) в кластер PySpark. Проверьте схему данных и определите типы данных для всех столбцов. Выполните предварительный анализ данных, включая проверку наличия пустых или некорректных значений.

2. Очистка данных. Удалите записи с пустыми значениями в ключевых полях, таких как цена, количество, расстояние или дата (в зависимости от набора данных). Преобразуйте значения типов данных в оптимальные типы для экономии памяти (например, преобразование строк в категориальные данные при использовании Pandas). Замените некорректные значения (например, отрицательные значения в полях расстояния или количества) на средние или медианные значения.

3. Обработка данных с использованием Pandas и PySpark. Разделите данные на несколько частей и обработайте их с помощью Pandas на кластере, используя pandas_on_spark. Используйте Pandas для выполнения операций агрегации, таких как подсчет количества записей, средних значений и медиан по ключевым полям, например, по регионам или продуктам. Найдите 5 записей с максимальными значениями в ключевых полях, таких как объем продаж, задержка или расстояние.

4. Анализ данных. Постройте сводные таблицы с использованием Pandas, сгруппировав данные по категориям (например, по месяцам или типам продуктов) и рассчитав средние или суммарные значения. Выполните сортировку данных, чтобы найти топ-5 элементов по выбранным метрикам (например, города с наибольшим объемом продаж или маршруты с самой высокой частотой задержек).

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