Департамент образования и науки города Москвы

Государственное автономное образовательное учреждение

высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики управления и технологий

Инструменты хранения и анализа больших данных

Смоляков Руслан Игоревич БД-241м

**Практическая работа 1.2. Обработка данных с использованием Apache Spark и Python (PySpark)**

**Вариант 23**

Направление подготовки/специальность

38.04.05 - Бизнес-информатика

Бизнес-аналитика и большие данные

(очная форма обучения)

Руководитель дисциплины:

Босенко Т.М., доцент департамента

информатики, управления и технологий,

кандидат технических наук

Москва

2025

**Введение**

**Цель: освоение основ работы с Apache Spark и его интеграцией с Python через библиотеку PySpark. Студенты научатся обрабатывать большие объемы данных, используя распределенные вычисления, а также научатся применять базовые операции с RDD (Resilient Distributed Datasets) и DataFrame, работать с SQL-запросами в Spark SQL, а также визуализировать результаты обработки данных.**

**Задачи:**

1. **Установить Apache Spark и PySpark.  
   Настроить рабочую среду для использования PySpark в Python, установить необходимые зависимости и настроить Spark на локальном компьютере или через облачную платформу.**
2. **Загрузка данных и их предварительная обработка.  
   Скачать или подготовить исходные данные для анализа (например, текстовые файлы или CSV). Загружать данные в Spark через RDD или DataFrame, выполнить предварительную обработку: очистка данных, фильтрация, преобразования.**
3. **Применение операций с RDD и DataFrame.  
   Научиться работать с RDD и DataFrame, выполнять такие операции как map, filter, reduce, groupBy, join и другие стандартные операции для обработки данных в распределенной среде.**
4. **Применение SQL-запросов через Spark SQL.  
   Использование SQL-запросов в Spark для извлечения и агрегации данных, создание временных таблиц и выполнение сложных запросов для анализа данных.**
5. **Визуализация результатов анализа данных.  
   Визуализировать полученные результаты с помощью библиотеки Python для визуализации данных (например, matplotlib или seaborn). Построить графики для лучшего представления результатов.**
6. **Подготовка отчета.  
   Оформить отчет по выполненной практике, в котором будет описан процесс выполнения работы, анализ полученных результатов и выводы. Включить ссылки на репозиторий и прикрепить сам отчет в формате PDF или Markdown.**

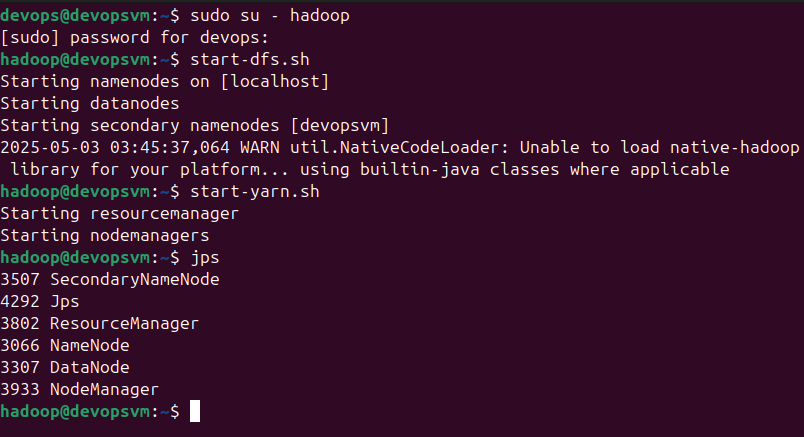
Вариант 23

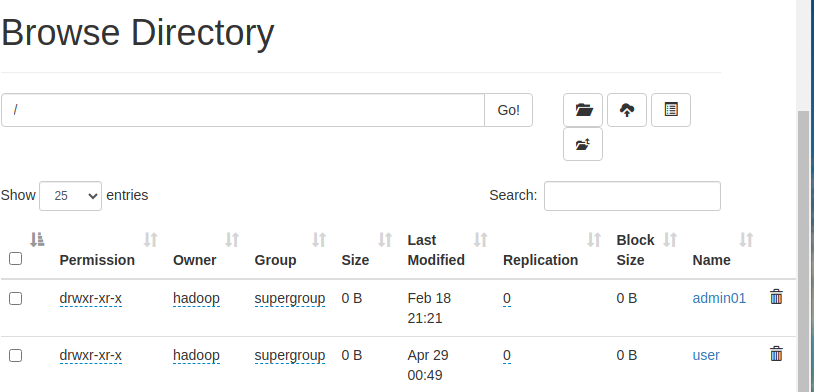
**Задание 1 (Spark + Hadoop): Анализ качества: загрузить quality.csv в HDFS, выявить проблемные товары**

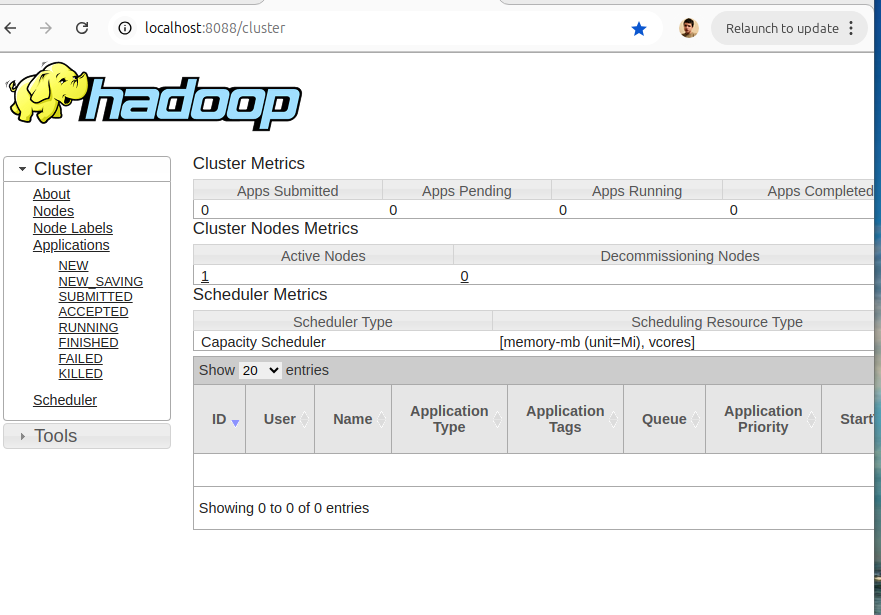
**Задание 2 (Spark Local + SQL): SQL-анализ: определить процент брака по производителям**

**Задание 3 (Визуализация): Построить диаграмму качества продукции**

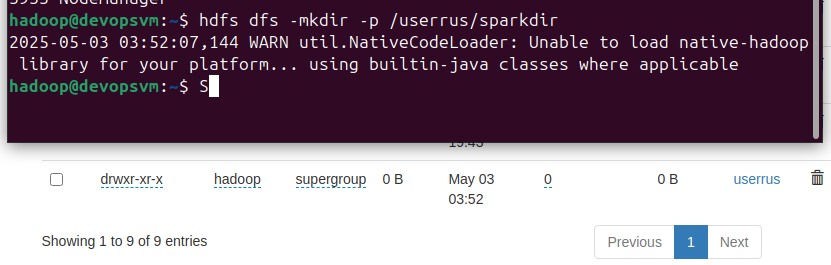
Перейдем на пользователя «hadoop» и запустим Hadoop:



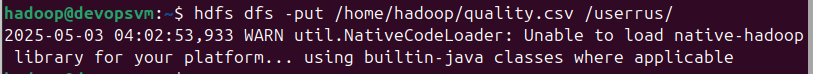




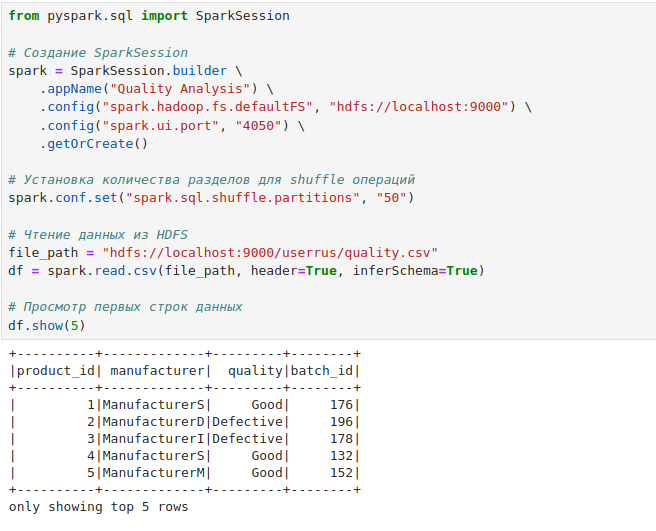
Создали директорию и сразу проверяем:



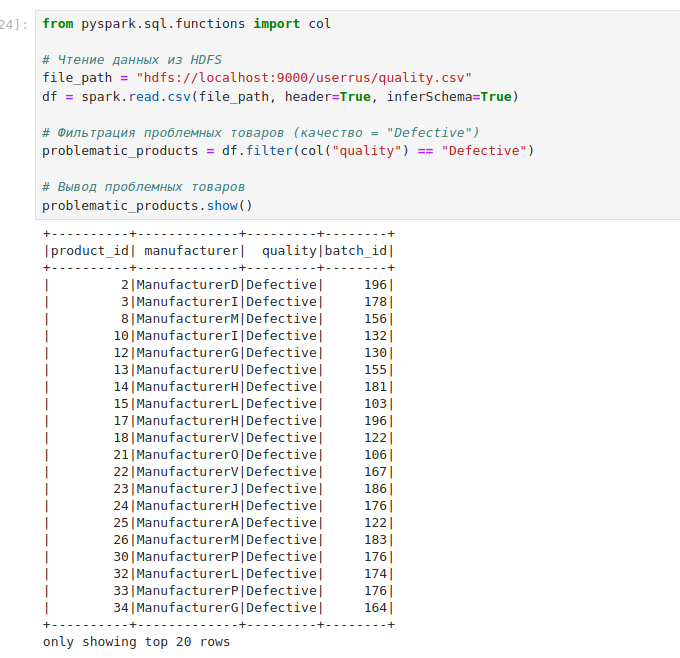
1. Задание 1 (Spark + Hadoop): Анализ качества: загрузить quality.csv в HDFS, выявить проблемные товары



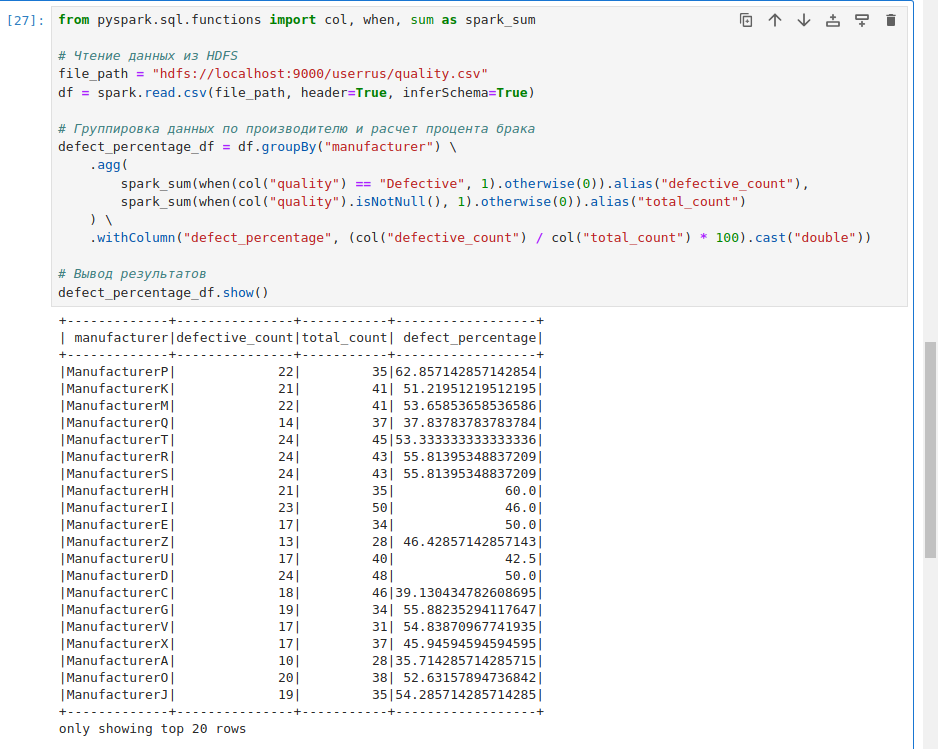
Откроем блокнот, проверим подключение и выведем первые 5 строк из датасета:



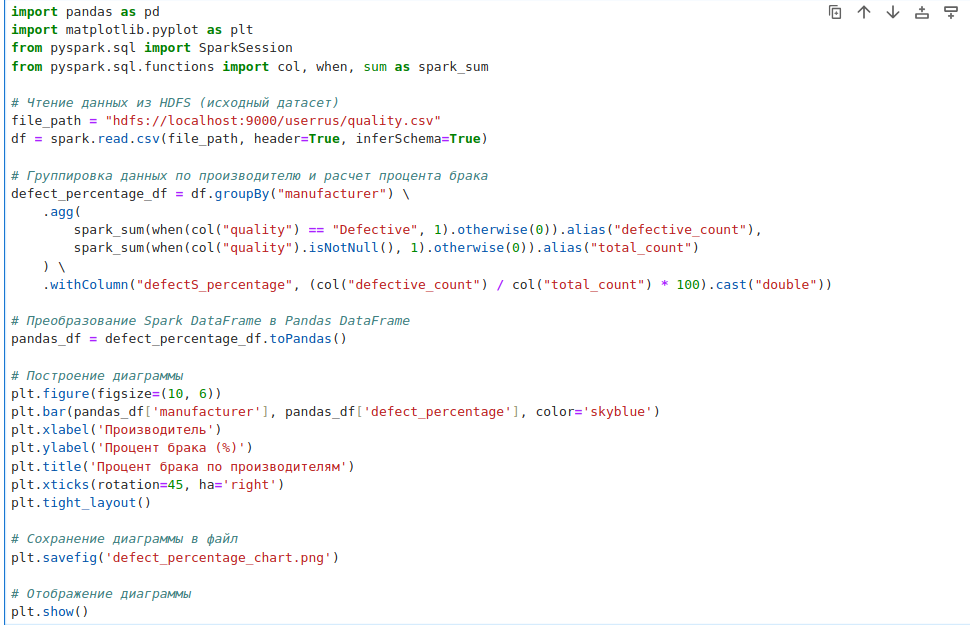
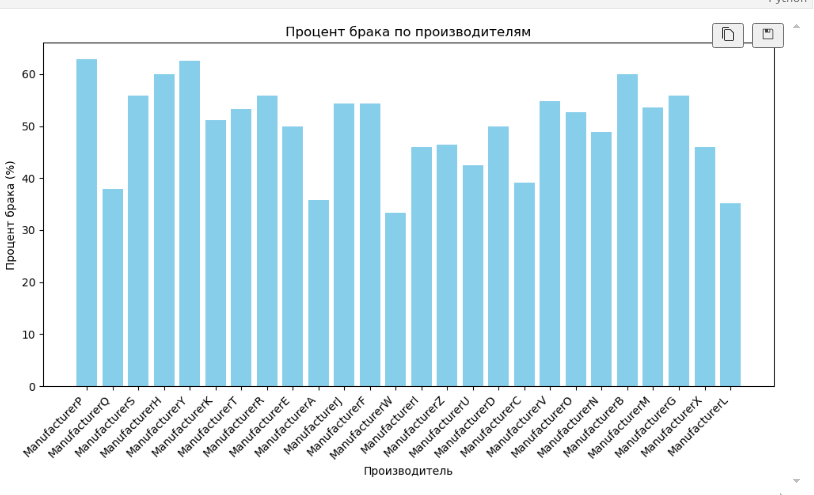
Результат выполнения задания:



1. Задание 2 (Spark Local + SQL): SQL-анализ: определить процент брака по производителям



1. Задание 3 (Визуализация): Построить диаграмму качества продукции

  
Вывод:  


Вывод:

В ходе выполнения практической работы были успешно освоены основы работы с Apache Spark и его интеграция с Python через библиотеку PySpark. В процессе выполнения заданий были приобретены навыки обработки больших данных с использованием распределенных вычислений, включая работу с RDD и DataFrame, выполнение SQL-запросов в Spark SQL, а также визуализацию результатов анализа данных.