МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное   
образовательное учреждение высшего образования  
«Самарский национальный исследовательский университет   
имени академика С.П. Королева»

(Самарский университет)

Институт информатики, математики и электроники

Факультет информатики  
Кафедра технической кибернетики

**Отчет по лабораторной работе №1**

Дисциплина: «Большие Данные»

## Тема: «Введение в Apache Spark»

Выполнила: Аксенова Д.А.

Группа: 6133-010402D

Самара 2020

**ЗАДАНИЕ**

1. Изучить операции загрузки и выгрузки данных в HDFS;
2. ознакомиться с базовыми операциями Apache Spark в spark-shell;
3. создать проект по обработке данных в IDE;
4. отладить анализ данных велопарковок на локальном компьютере:
   1. Найти велосипед с максимальным пробегом.
   2. Найти наибольшее расстояние между станциями.
   3. Найти путь велосипеда с максимальным пробегом через станции.
   4. Найти количество велосипедов в системе.
   5. Найти пользователей потративших на поездки более 3 часов.
5. запустить анализ данных велопарковок на сервере.

**ХОД РАБОТЫ**

Скачаем и установим виртуальную машину Hortonworks с установленными Hadoop и Spark, подключимся к ней (рисунок 1).

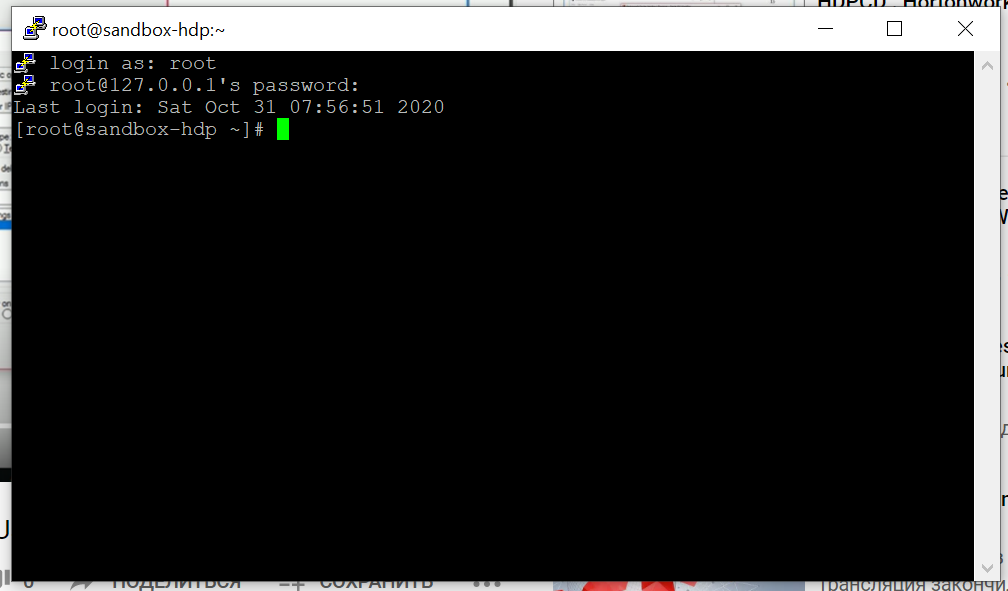


Рисунок 1 – Подключение к Hortonworks Sandbox

Переместим необходимые для работы наборы данных в РФС. Для этого сначала скопируем их в файловую систему узла кластера с помощью команды утилиты WinSCP, как показано на рисунке 2 и убедимся, что файлы перемещены (рисунок 3).

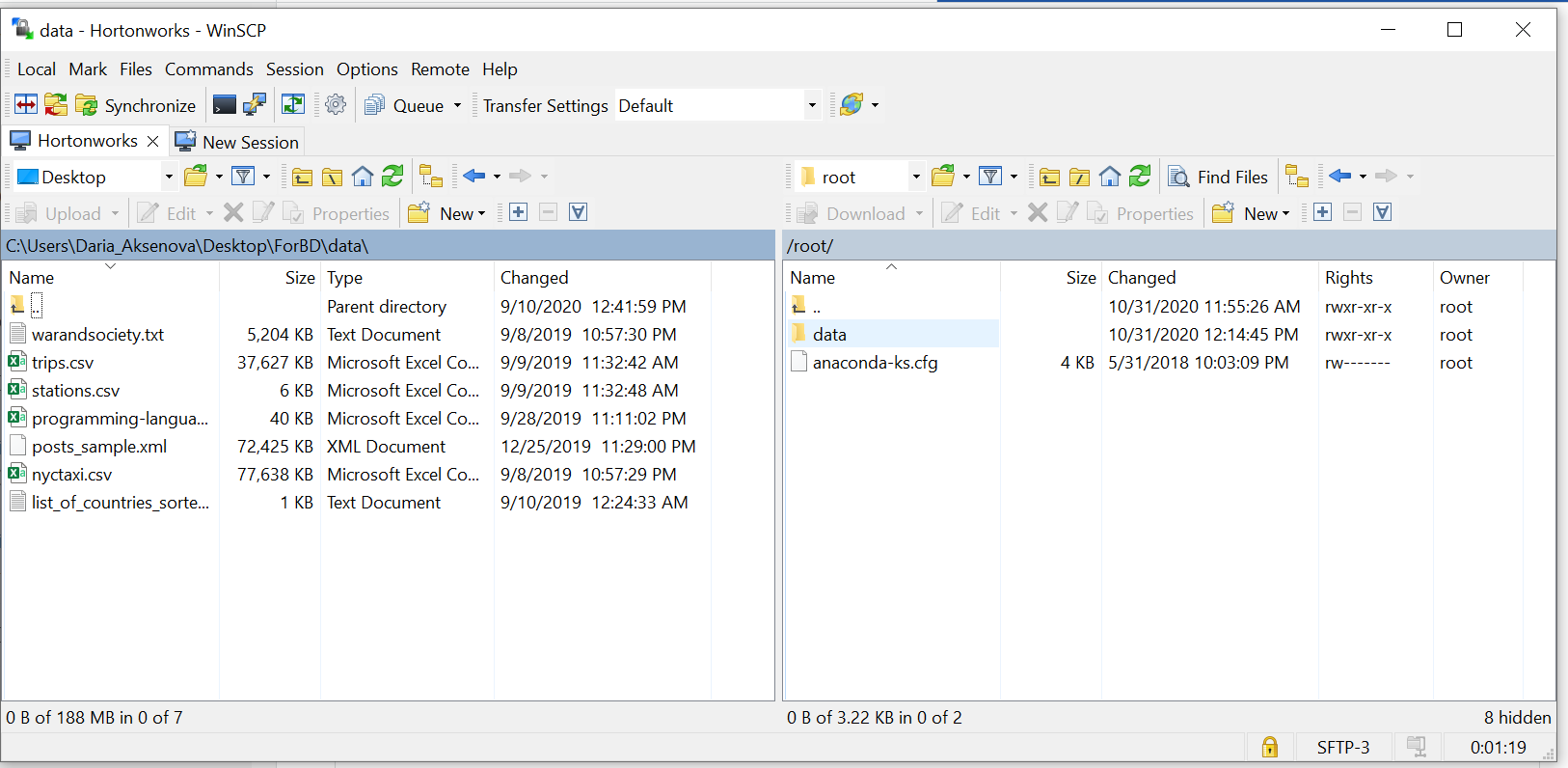


Рисунок 2 – Перемещение данных при помощи WinSCP

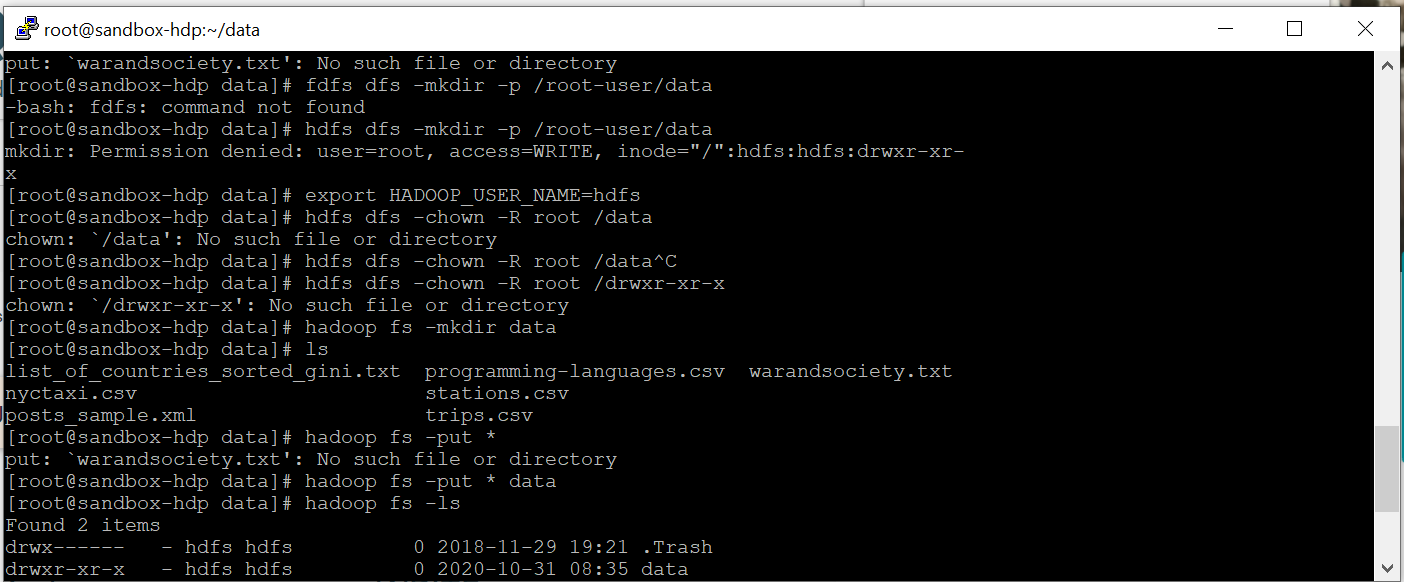


Рисунок 3 – Данные в директории

Затем запустим spark-shell и откроем веб интерфейс.

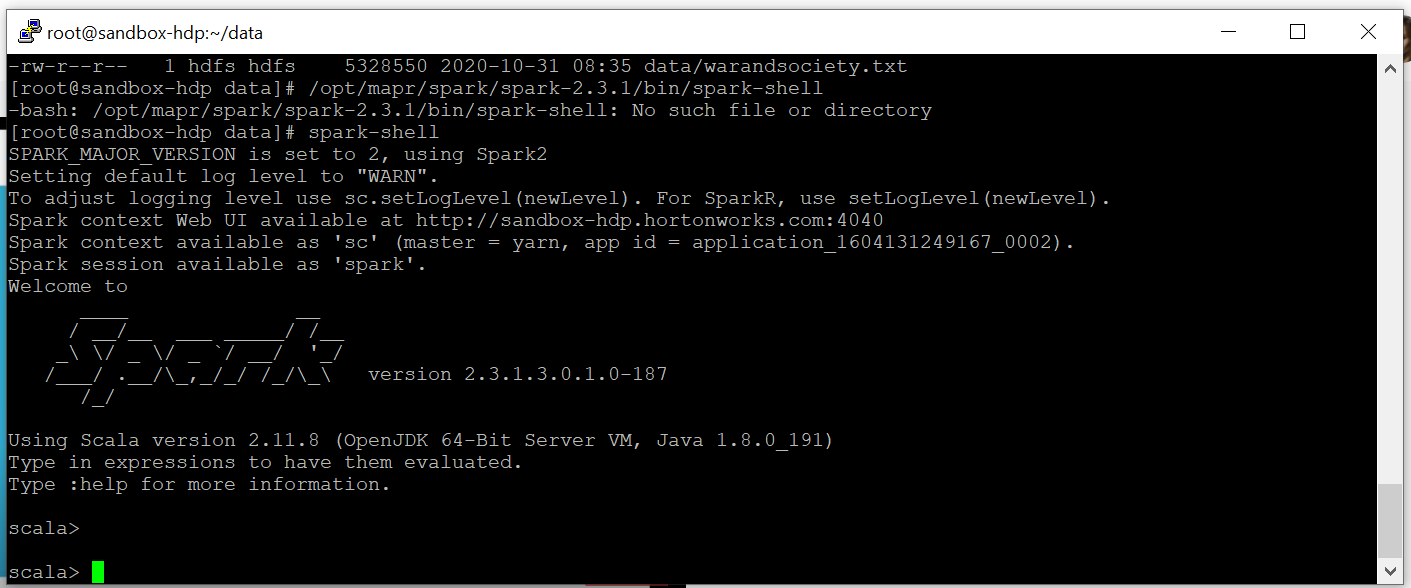


Рисунок 4 – Запуск spark-shell

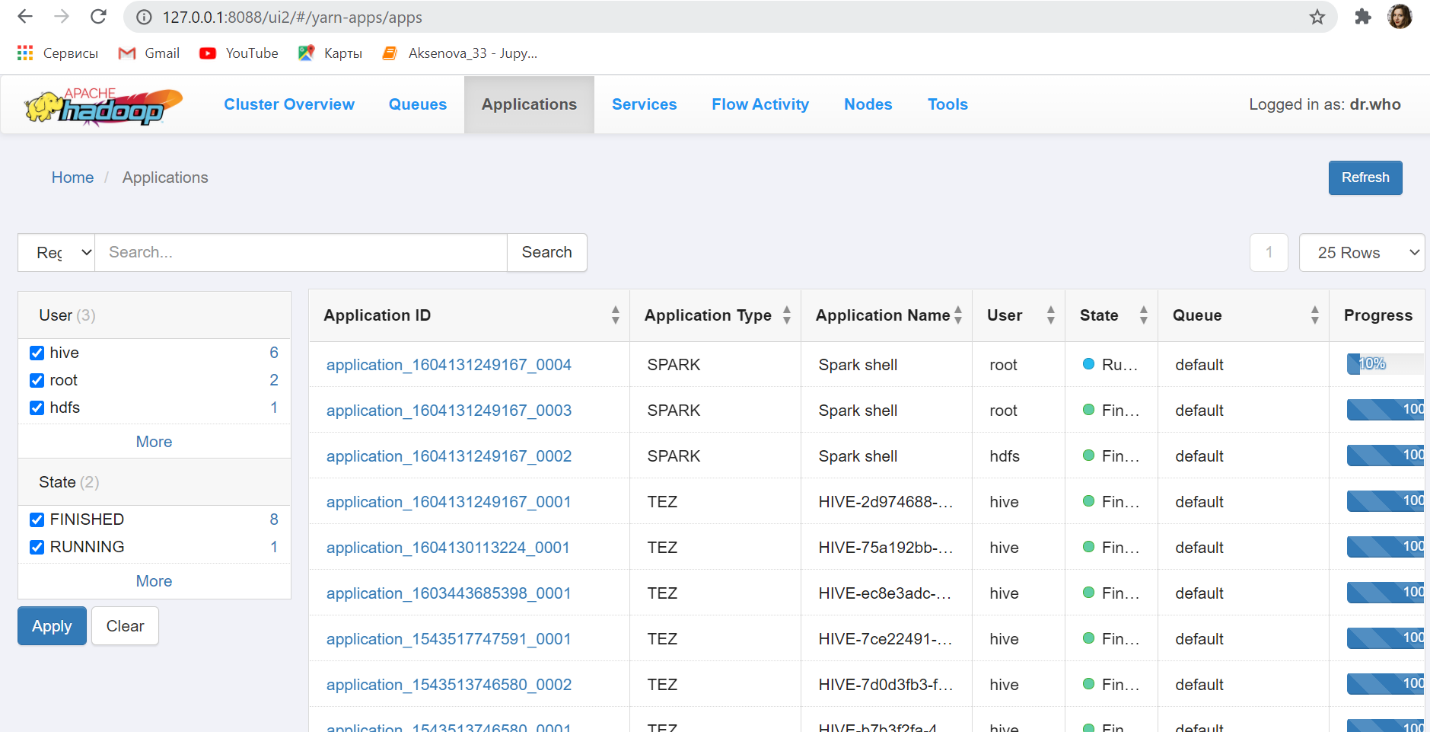


Рисунок 5 – Web-интерфейс spark-shell

Создадим RDD для текстового файла warandpeace.txt и считаем первые 10 строк.

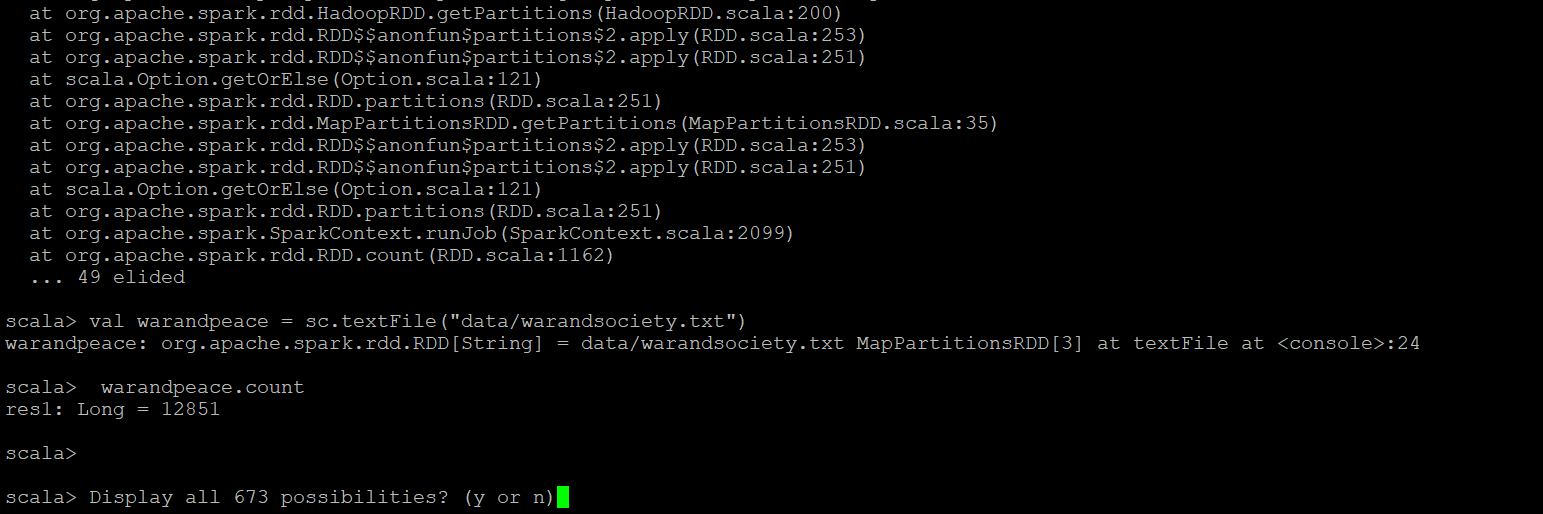


Рисунок 6 – Создание RDD

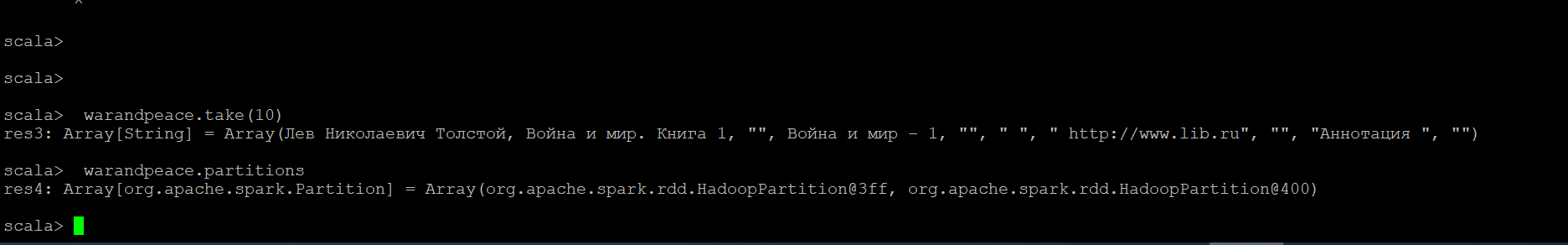


Рисунок 7 – Первые 10 строк RDD

Найдем строки, в которых содержится слово "война". Будем считать количество строк несколько раз подряд, замеряя при этом время работы программы. При выполнении команды count во второй и последующие разы можно заметить ускорение работы прогаммы за счет кэширования.

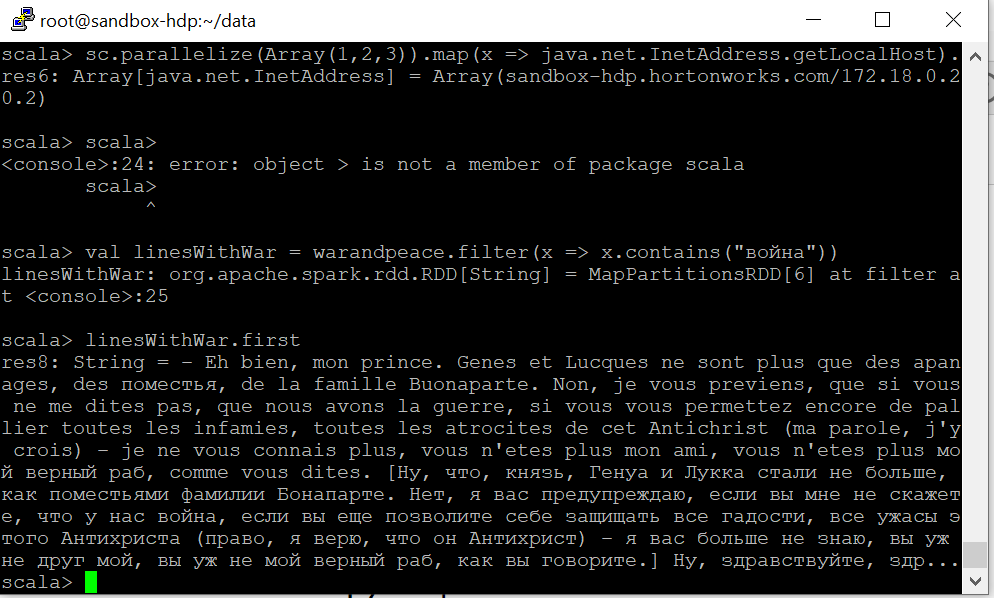


Рисунок 8 – Строка со словом «война»

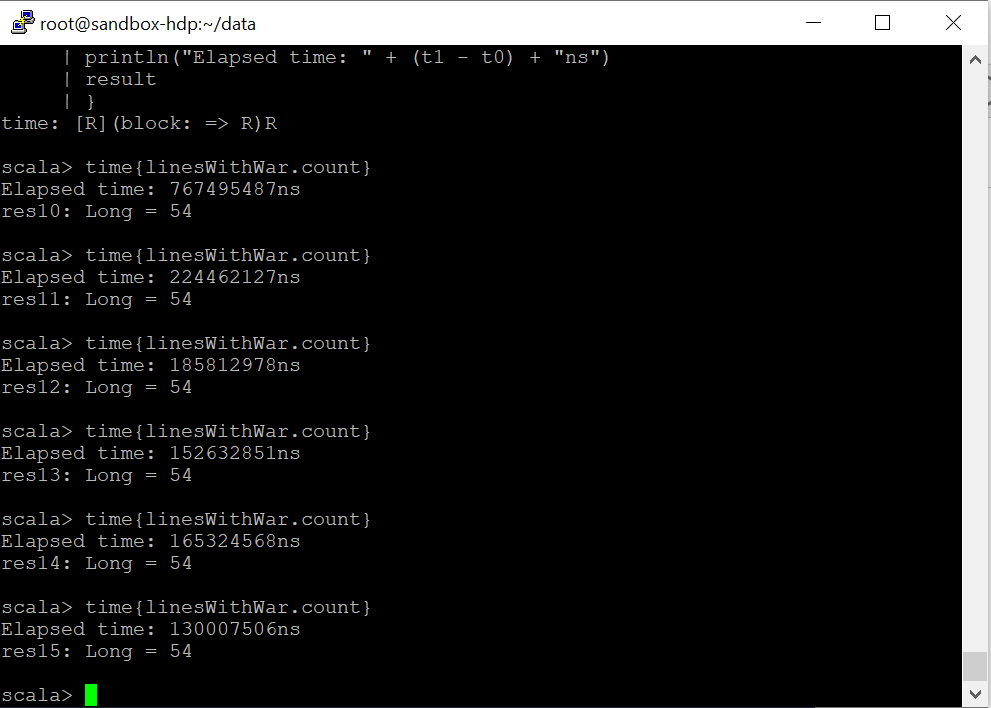


Рисунок 9 – Время подсчета строк

Найдем гистограмму слов.

Затем улучшим процедуру, убирая из слов лишние символы и трансформируя все слова в нижний регистр при помощи регулярных выражений.

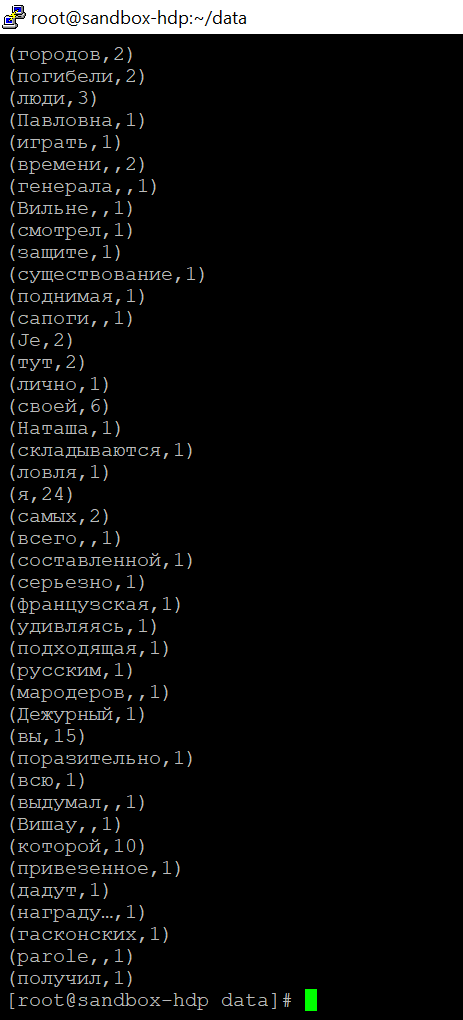


Рисунок 10 – Гистограмма слов

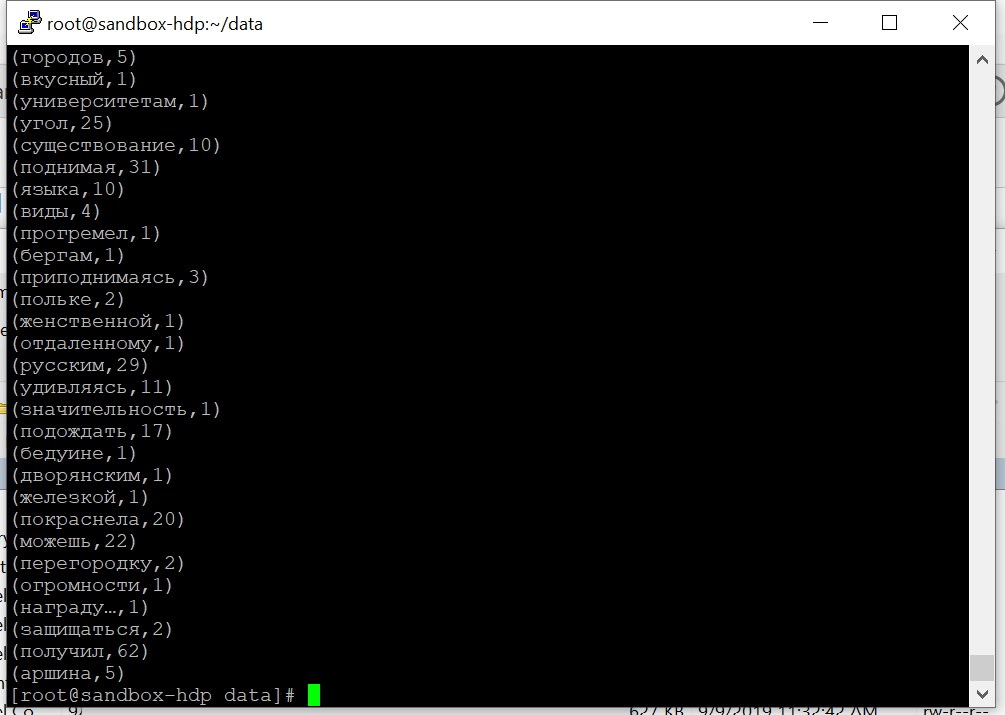


Рисунок 11 – Улучшенная гистограмма слов

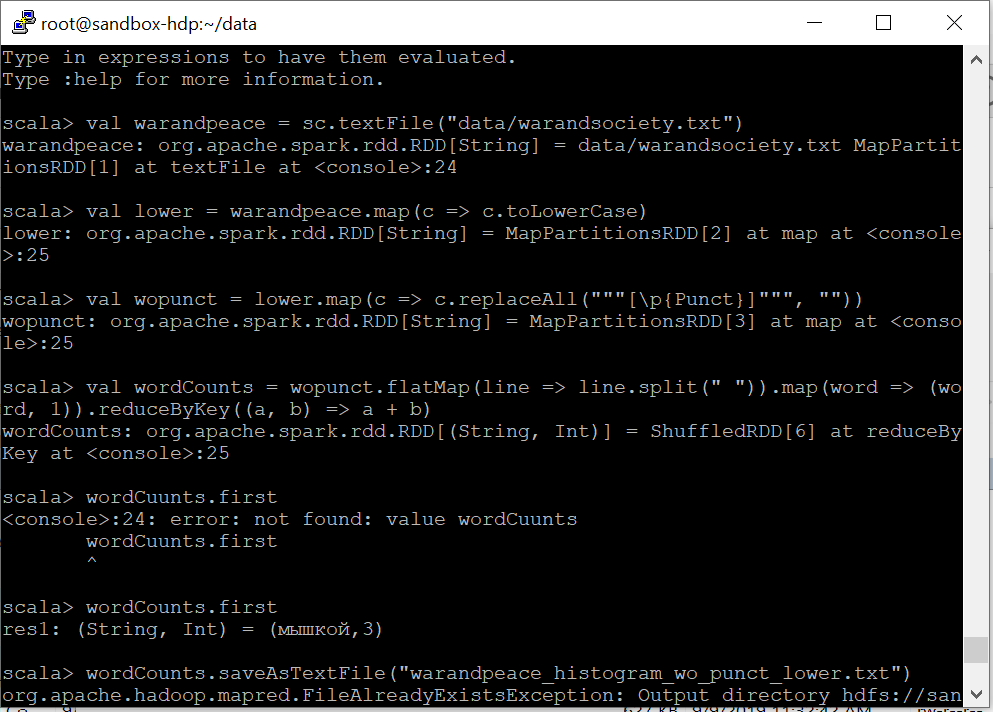


Рисунок 12 – Скрипт для улучшения гистограммы

Инициализируем два множества a и b. Найдем их объединение и уберем из него дубликаты при помощи функции distinct. Функция distinct вычисляет результакующий набор из уникальных значений при помощи HashSet. Найдем пересечение и разность множеств.

Рассмотрим работу с обзими и аккумулирующими переменными, а так же с парами «ключ-значение».

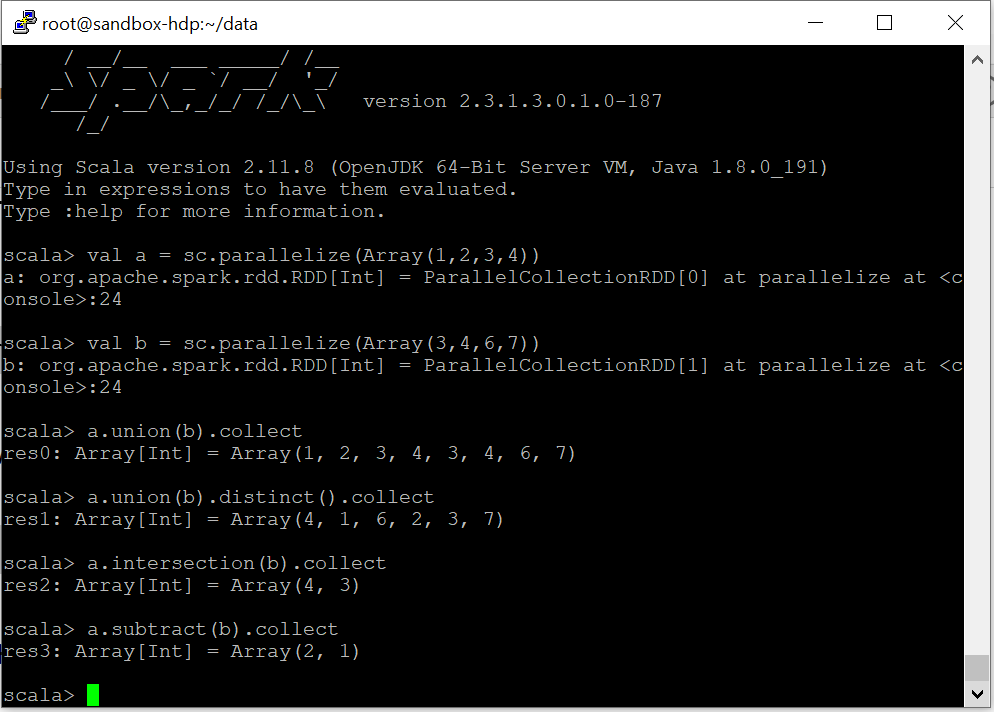


Рисунок 13 – Операции над множествами

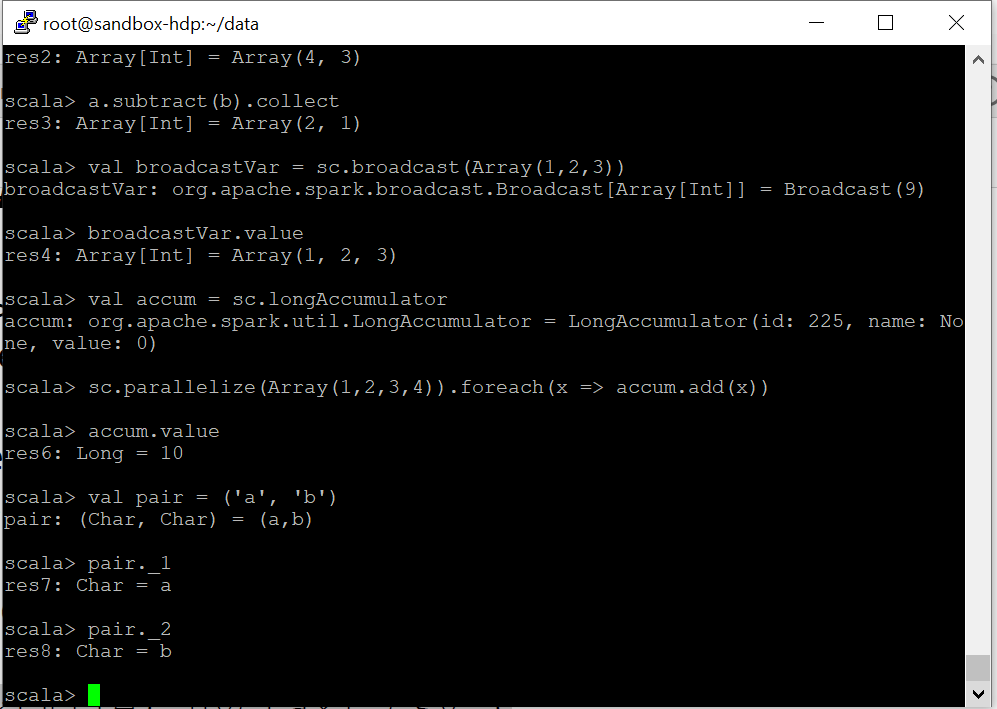


Рисунок 14 – Общие и аккумулирующие переменные, пары ключ-значение

Проанализируем данные о поездках такси в Нью-Йорке. Найдем общее количество номеров такси несколько раз, измеряя время выполнения операции. Как видно из рисунка 15, за счет кэширования время выполнения операции с каждым разом уменьшается.

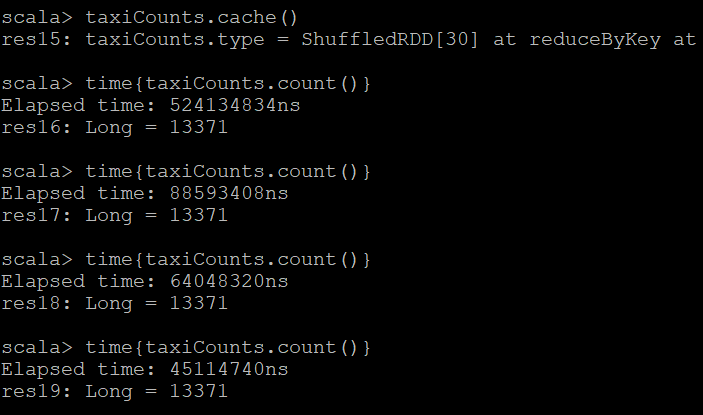


Рисунок 15 – Время подсчета общего числа такси

В ходе аботы так же был создан проект на локальном компьютере при помощи IntelliJ IDEA на языке Scala. И были проагализированы данные влопарковок и поездок на велосипедах.

Код написанной программы доступен в репозитории по ссылке <https://github.com/darinaksena/Big_Data_Labs> .

Запуск проекта на виртуальной машине Hortonworks так же прошел успешно.

**ЗАКЛЮЧЕНИЕ**

В результате лабораторной работы были изучены операции загрузки и выгрузки данных в HDFS, базовые операции Apache Spark в spark-shell, был создан проект по обработке данных в IDE на локальном компьютере, были проанализированы данные и выполнено задание, а так же ь=был запущен проект на виртуальной машине Hortonworks.