|  |  |
| --- | --- |
| **Gerb-BMSTU_01** | **Министерство науки и высшего образования Российской Федерации**  Калужский филиал  федерального государственного бюджетного  образовательного учреждения высшего образования  ***«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»***  ***(КФ МГТУ им. Н.Э. Баумана)*** |

**ФАКУЛЬТЕТ** ***ИУК «Информатика и управление»***

**КАФЕДРА** \_\_***ИУК4 «Программное обеспечение ЭВМ, информационные технологии»***

**ЛАБОРАТОРНАЯ РАБОТА №4**

**«Язык Pig Latin»**

**ДИСЦИПЛИНА: «Технологии обработки больших данных»**

|  |  |  |
| --- | --- | --- |
| Выполнил: студент гр. ИУК4-72Б | | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ( Карельский М.К. )  (Подпись) |
| Проверил: | | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ( Голубева С.Е. )  (Подпись) |
| Дата сдачи (защиты):  Результаты сдачи (защиты): | | |
|  | - Балльная оценка:  - Оценка: | |

Калуга, 2023

**Цель:** формирование практических навыков реализации pig-скриптов для обработки больших данных.

**Задачи:**

1. Получить навыки обработки больших данных, использую Pig Latin.
2. Изучить принцип работы Pig Latin.
3. Изучить синтаксис Pig Latin.
4. Уметь писать запросы, комбинируя несколько источников данных

**Задание:**

1. Выполнить задание из лабораторной работы №2, используя язык Pig Latin
2. База данных твитов состоит из двух файлов. Выполить задание по варианту, используя Pig Latin.

Файл tweets.csv имеет формат:

tweet\_id, tweet, login

Файл users.csv имеет формат:

login, user\_name, state

**Вариант 7**

1. Модифицировать программу подсчета слов WordCount. Результат должен содержать 100 самых часто встречающихся слов. Из результата должны быть удалены стоп-слова.
2. Выбрать все твиты пользователей из штата NY. Вывести список 20 самых часто используемых слов в их твитах.

**Листинг:**

***LW4\_1.pig***

lines = LOAD '/user/hduser/input/task\_1' USING TextLoader() AS (line:chararray);

stopwords = LOAD '/user/hduser/input/stopwords.txt' AS (stopword:chararray);

words = FOREACH lines GENERATE FLATTEN(STRSPLIT(line, ' ')) AS word;

words = FOREACH words GENERATE REPLACE(word, '[^a-zA-z]', '') AS word;

words = FILTER words BY word != '';

words = FOREACH words GENERATE LOWER(word) AS word;

words\_filtered = JOIN words BY word LEFT OUTER, stopwords BY stopword USING 'replicated';

words\_filtered = FILTER words\_filtered BY stopword IS NULL;

words\_filtered = FOREACH words\_filtered GENERATE word;

word\_counts = FOREACH (GROUP words\_filtered BY word) GENERATE group AS word, COUNT(words\_filtered) as count;

word\_counts\_sorted = ORDER word\_counts BY count DESC;

word\_counts\_ranked = RANK word\_counts\_sorted;

word\_counts\_cutted = FILTER word\_counts\_ranked BY rank\_word\_counts\_sorted < 101;

%declare DT `date +%y%m%dT%H%M`;

STORE word\_counts\_cutted INTO '/user/hduser/output/task\_1/$DT' USING PigStorage(',');

***LW4\_2.pig***

tweets = LOAD '/user/hduser/input/task\_2/tweets.csv' USING PigStorage(',')

AS (tweet\_id:int, tweet:chararray, login:chararray);

users = LOAD '/user/hduser/input/task\_2/users.csv' USING PigStorage(',')

AS (login:chararray, user\_name:chararray, state:chararray);

joined = JOIN tweets BY login, users BY login;

joined = FILTER joined BY state == 'New York';

words = FOREACH joined GENERATE FLATTEN(TOKENIZE(tweet, ' ')) AS word;

words = FOREACH words GENERATE LOWER(word) AS word;

word\_counts = FOREACH (GROUP words BY word) GENERATE group AS word, COUNT(words) as count;

word\_counts\_sorted = ORDER word\_counts BY count DESC;

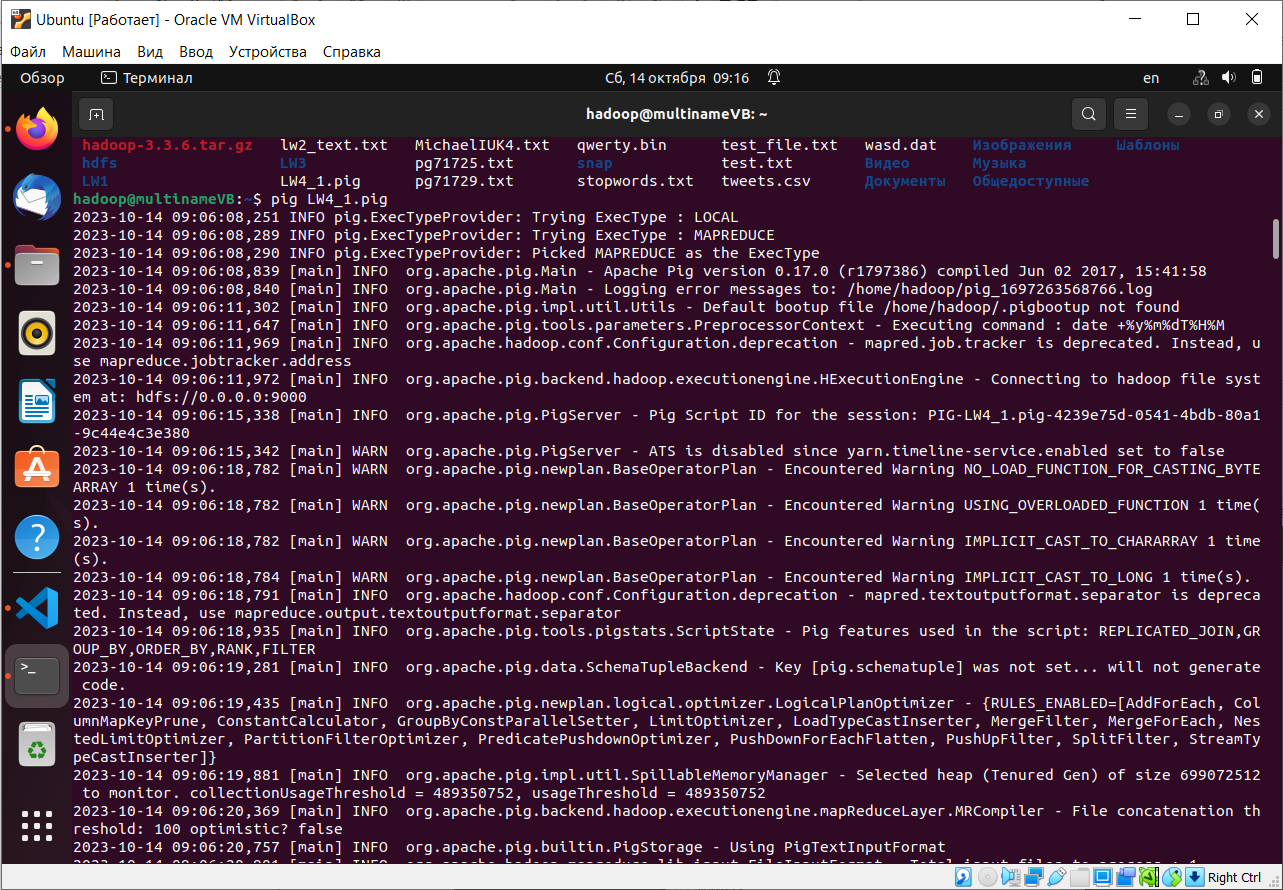
word\_counts\_ranked = RANK word\_counts\_sorted;

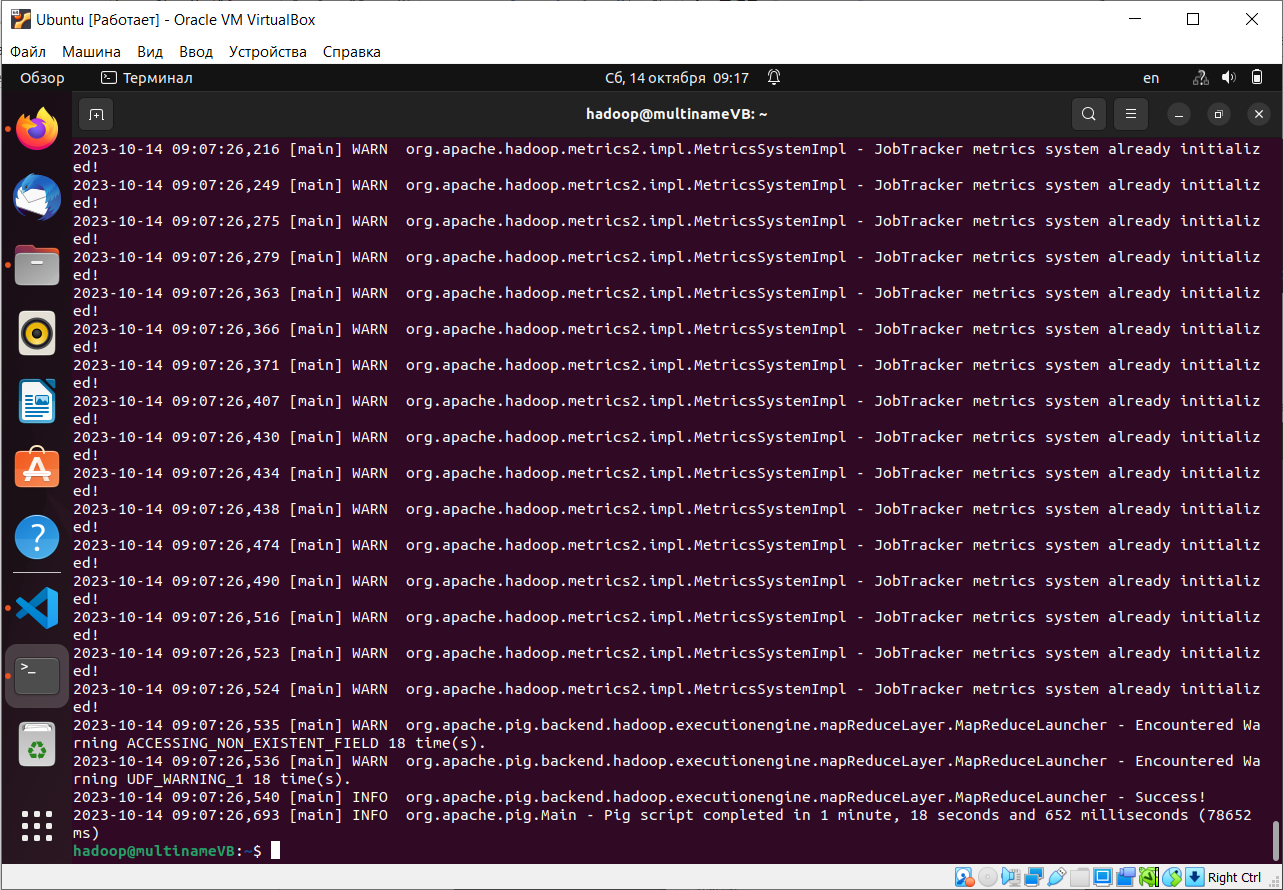
word\_counts\_cutted = FILTER word\_counts\_ranked BY rank\_word\_counts\_sorted < 21;

%declare DT `date +%y%m%dT%H%M`;

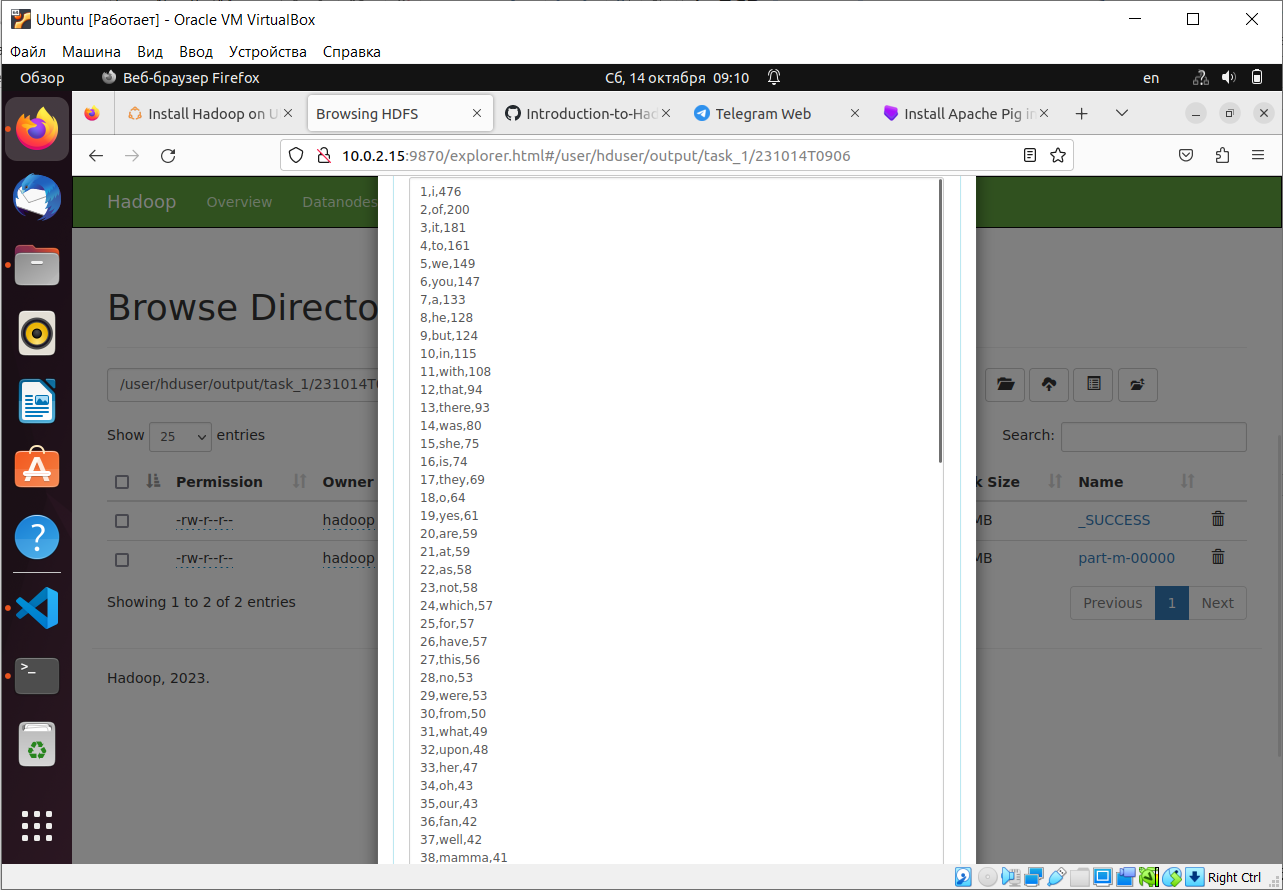
STORE word\_counts\_cutted INTO '/user/hduser/output/task\_2/$DT' USING PigStorage(',');

**Результат:**

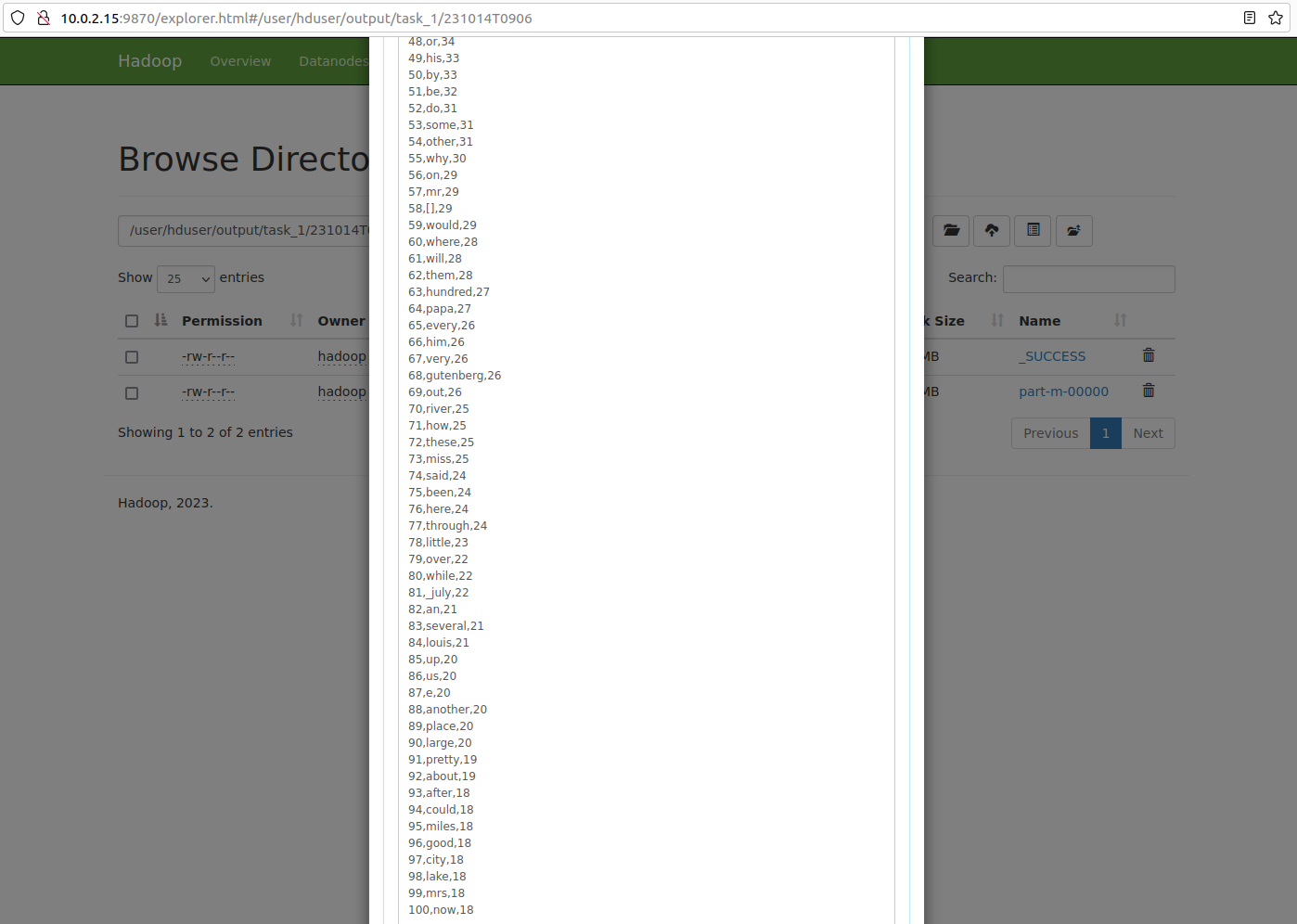




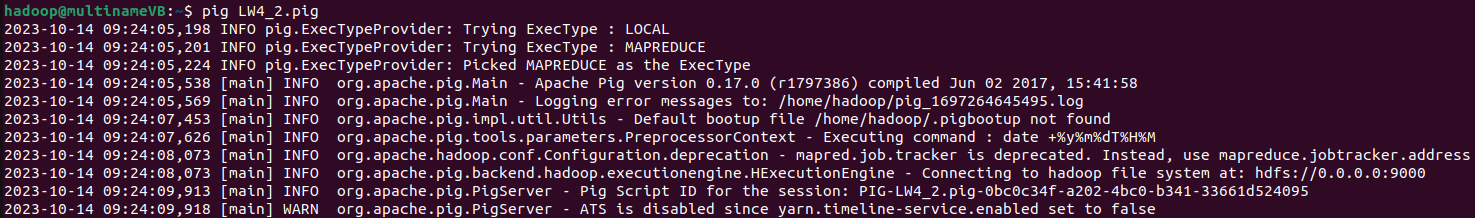
**Рис. 1.** Выполнение первой задачи

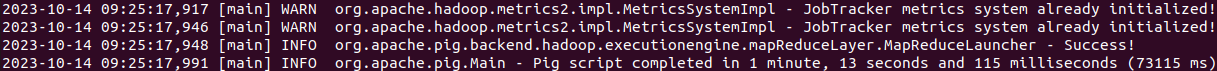


**Рис. 2.1.** Результат первой задачи

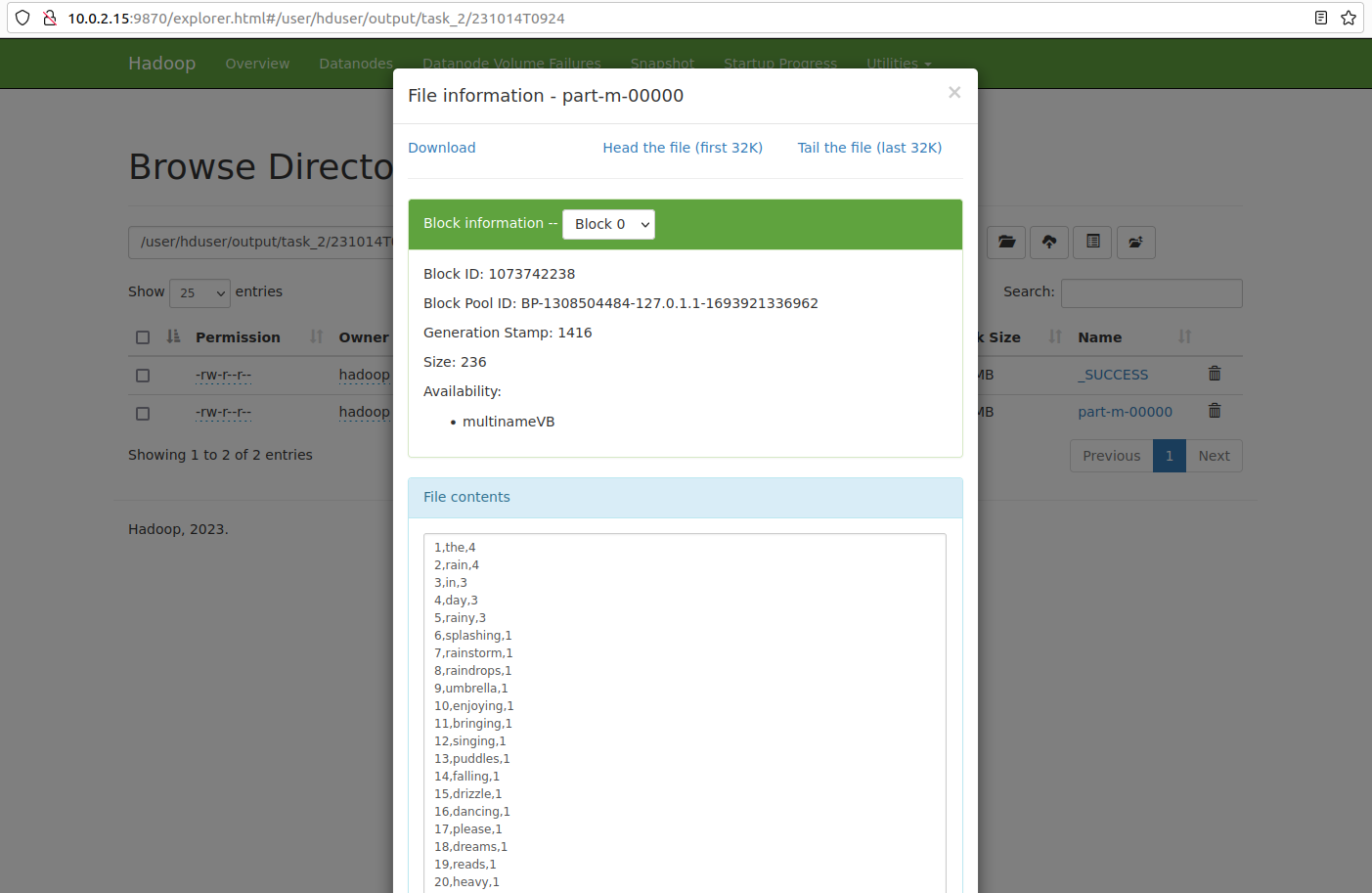


**Рис. 2.2.** Результат первой задачи





**Рис. 3.** Выполнение второй задачи



**Рис. 4.** Результат второй задачи

**Вывод:** в ходе выполнения лабораторной работы были получены практические навыки реализации pig-скриптов для обработки больших данных.