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, используя 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:

```
hadoop@multinameVB:~$ pig LW4_1.pig
2023-10-14 09:06:08,251 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2023-10-14 09:06:08,289 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2023-10-14 09:06:08,290 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2023-10-14 09:06:08,839 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2023-10-14 09:06:08,840 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_1697263568766.log
2023-10-14 09:06:11,302 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not found
2023-10-14 09:06:11,647 [main] INFO  org.apache.pig.tools.parameters.PreprocessorContext - Executing command : date +%y%m%dT%H%M
2023-10-14 09:06:11,969 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2023-10-14 09:06:11,972 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file syst
em at: hdfs://0.0.0.0:9000
2023-10-14 09:07:26,535 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Wa
rning ACCESSING_NON_EXISTENT_FIELD 18 time(s).
2023-10-14 09:07:26,536 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Wa
rning UDF_WARNING_1 18 time(s).
2023-10-14 09:07:26,540 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-10-14 09:07:26,693 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 18 seconds and 652 milliseconds (78652
ms)
```

Рис. 1. Выполнение первой задачи

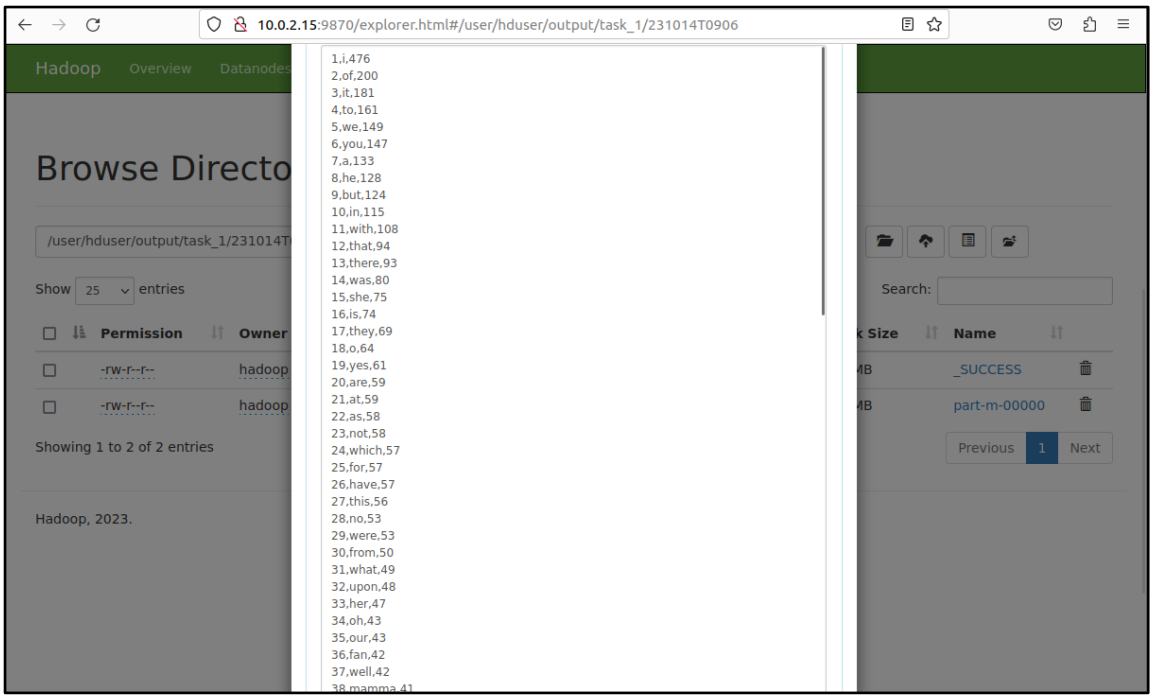

Рис. 2.1. Результат первой задачи

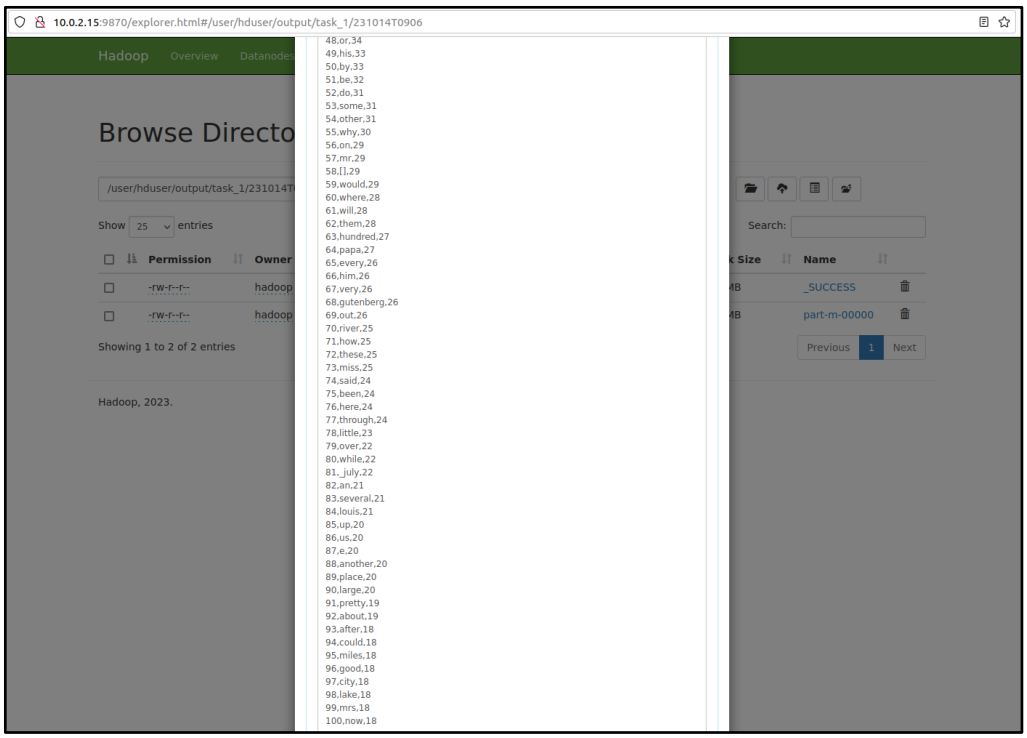

Рис. 2.2. Результат первой задачи

```
hadoop@multinameVB:~$ pig LW4_2.pig
2023-10-14 09:24:05,198 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2023-10-14 09:24:05,201 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2023-10-14 09:24:05,224 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2023-10-14 09:24:05,538 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2023-10-14 09:24:05,569 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_1697264645495.log
2023-10-14 09:24:07,453 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not found
2023-10-14 09:24:07,626 [main] INFO  org.apache.pig.tools.parameters.PreprocessorContext - Executing command : date +%y%m%dT%H%M
2023-10-14 09:24:08,073 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-10-14 09:24:08,073 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://0.0.0.0:9000
2023-10-14 09:24:09,913 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-LW4_2.pig-0bc0c34f-a202-4bc0-b341-33661d524095
2023-10-14 09:24:09,918 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2023-10-14 09:25:17,917 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2023-10-14 09:25:17,946 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2023-10-14 09:25:17,948 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-10-14 09:25:17,991 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 13 seconds and 115 milliseconds (73115 ms)
```

Рис. 3. Выполнение второй задачи

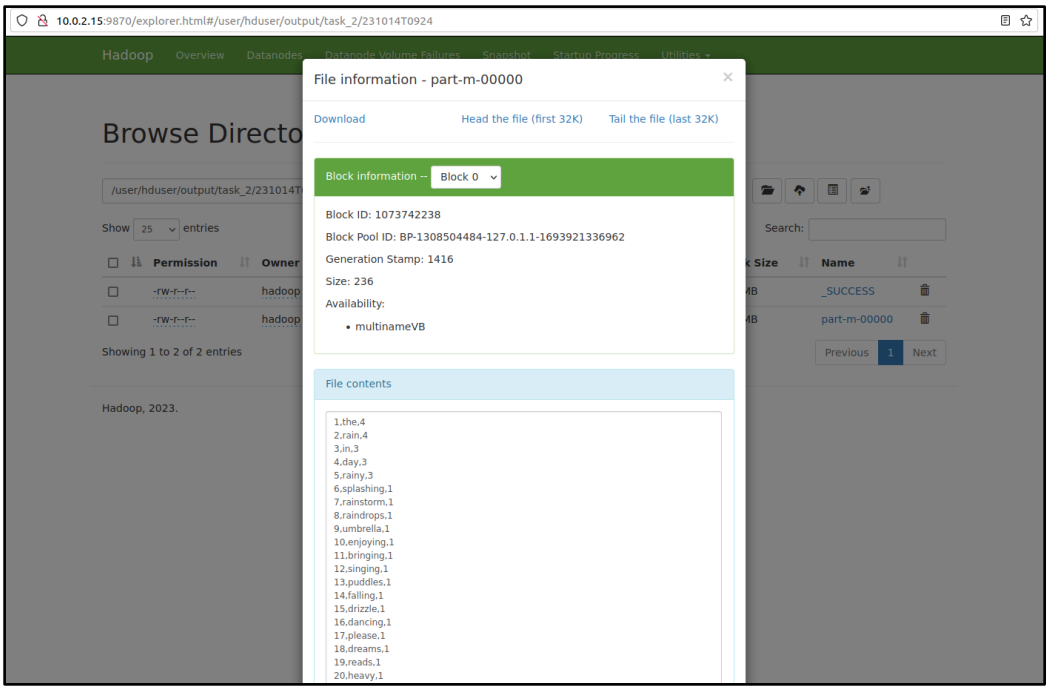

Рис. 4. Результат второй задачи

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