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 №2

«MapReduce»

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 использования парадигмы MapReduce для обработки больших данных.

Задачи:

1. Изучить подход MapReduce.
2. Изучить принципы работы Hadoop MapReduce.
3. Получить практические навыки реализации MapReduce задач.
4. Уметь обрабатывать большие текстовые файлов с помощью MapReduce.

Задание:

Выполнить задание с помощью подхода MapReduce согласно варианту. В качестве входных текстовых файлов можно использовать книги в txt формате из библиотеки Project Gutenberg: https://www.gutenberg.org.

Список стоп-слов: http://xpo6.com/wp-content/uploads/2015/01/stopword-list.csv.

Вариант 7

Модифицировать программу подсчета слов WordCount. Результат должен содержать 100 самых часто встречающихся слов. Из результата должны быть удалены стоп-слова.

Листинг:

script.sh

```
#!/bin/bash
MAPPER="mapper.py"
REDUCER="reducer.py"
INPUT="/user/hduser/input"
OUTPUT="/user/hduser/output"
SW=`cat sw.csv`
DARG="SW=${SW}"
/usr/local/hadoop/bin/hdfs dfs -rm -R -f $OUTPUT
/usr/local/hadoop/bin/mapred streaming -D $DARG -input $INPUT -output $OUTPUT -
mapper $MAPPER -reducer $REDUCER
/usr/local/hadoop/bin/hdfs dfs -head "$OUTPUT/part-00000"
```

mapper.py

```
#!/usr/bin/python3.10
import sys
import os

try:
    stop_words = os.environ['SW'].split(',')
except:
    stop_words = []

for line in sys.stdin:
    line = line.strip()
    line = line.lower()
    words = line.split()
    for word in words:
        if word not in stop_words:
            print (word, 1)
```

reducer.py

```
#!/usr/bin/python3.10
import sys

words = {}
for line in sys.stdin:
    word = line.split()[0]
    if word in words:
        words[word] += 1
    else:
        words[word] = 1

words = dict(sorted(words.items(), key=lambda item: item[1], reverse=True))
i = 0
for word in words:
    i += 1
    print(f"{i} ({word}, {words [word]})")
    if i >= 100:
        break
```

Результат:

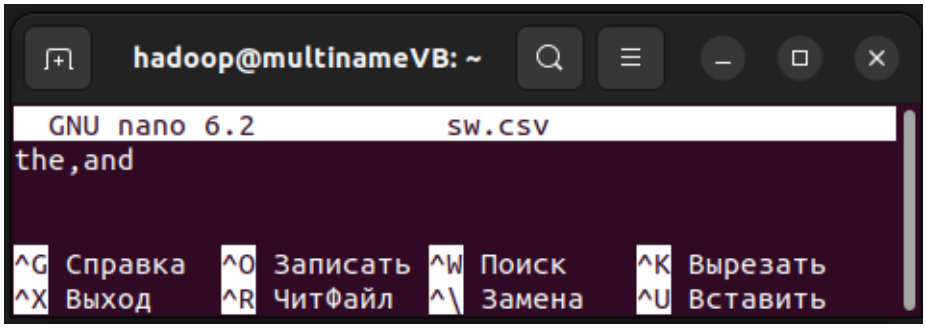

Рис. 1. Список стоп-слов

```
hadoop@multinameVB:~$ hdfs dfs -ls -R /user/hduser/input
-rw-r--r--   1 hadoop supergroup     465319 2023-09-27 14:45 /user/hduser/input/pg71725.txt
-rw-r--r--   1 hadoop supergroup     420882 2023-09-27 17:55 /user/hduser/input/pg71729.txt
```

Рис. 2. Файлы для обработки

```
hadoop@multinameVB:~$ sudo -u hadoop sh script.sh
2023-09-27 18:08:43,173 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-09-27 18:08:43,734 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-09-27 18:08:43,735 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-09-27 18:08:43,793 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2023-09-27 18:08:44,897 INFO mapred.FileInputFormat: Total input files to process : 2
2023-09-27 18:08:44,969 INFO mapreduce.JobSubmitter: number of splits:2
2023-09-27 18:08:45,641 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1119124103_0001
2023-09-27 18:08:45,662 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-09-27 18:08:46,157 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-09-27 18:08:46,159 INFO mapreduce.Job: Running job: job_local1119124103_0001
2023-09-27 18:08:46,165 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-09-27 18:08:46,171 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2023-09-27 18:08:46,205 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-09-27 18:08:46,205 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:
false, ignore cleanup failures: false
2023-09-27 18:08:46,445 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-09-27 18:08:46,463 INFO mapred.LocalJobRunner: Starting task: attempt_local1119124103_0001_m_000000_0
2023-09-27 18:08:46,621 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-09-27 18:08:46,629 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:
false, ignore cleanup failures: false
2023-09-27 18:08:46,813 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2023-09-27 18:08:46,887 INFO mapred.MapTask: Processing split: hdfs://0.0.0.0:9000/user/hduser/input/pg71725.txt:0+465319
2023-09-27 18:08:46,954 INFO mapred.MapTask: numReduceTasks: 1
2023-09-27 18:08:47,165 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2023-09-27 18:08:47,166 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2023-09-27 18:08:47,166 INFO mapred.MapTask: soft limit at 83886080
2023-09-27 18:08:47,166 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2023-09-27 18:08:47,166 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2023-09-27 18:08:47,177 INFO mapreduce.Job: Job job_local1119124103_0001 running in uber mode : false
2023-09-27 18:08:47,179 INFO mapreduce.Job:  map 0% reduce 0%
2023-09-27 18:08:47,183 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2023-09-27 18:08:47,317 INFO streaming.PipeMapRed: PipeMapRed exec [/home/hadoop/./mapper.py]
```

Рис. 3. Запуск программы

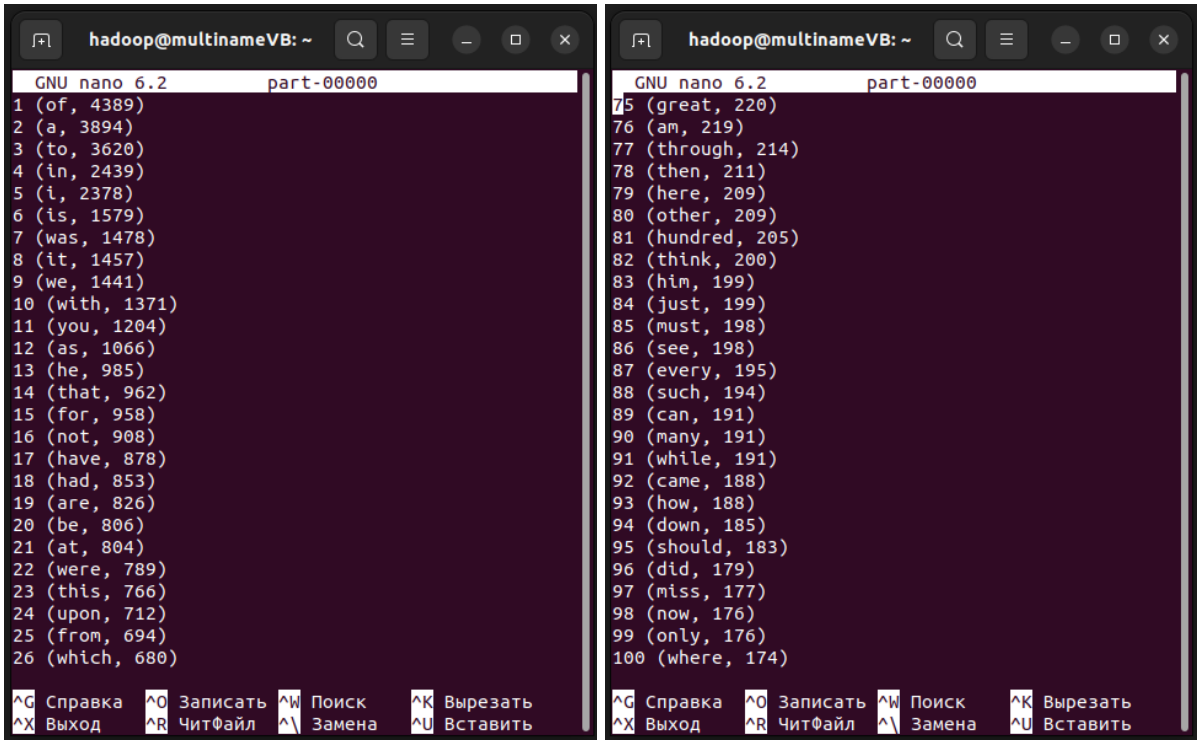

Рис. 4. Результат работы

Вывод: в ходе выполнения лабораторной работы были получены практические навыки использования парадигмы MapReduce для обработки больших данных.