|  |  |
| --- | --- |
| **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 №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:

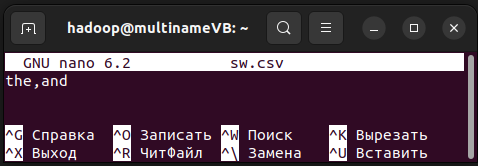
i += 1

print(f"{i} ({word}, {words [word]})")

if i >= 100:

break

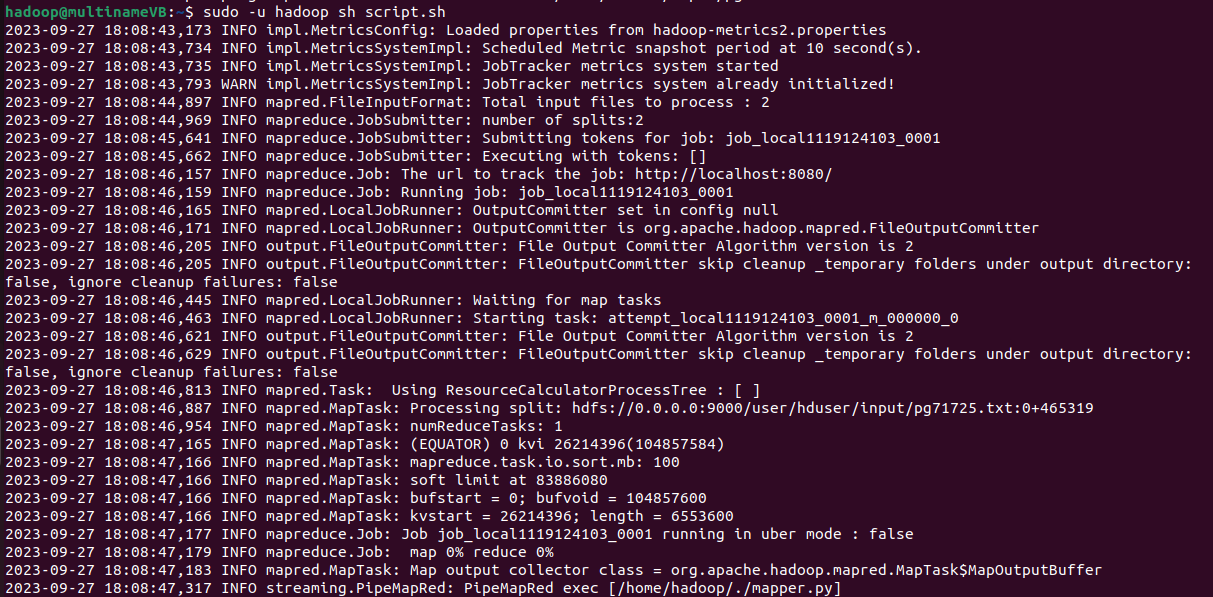
**Результат:**



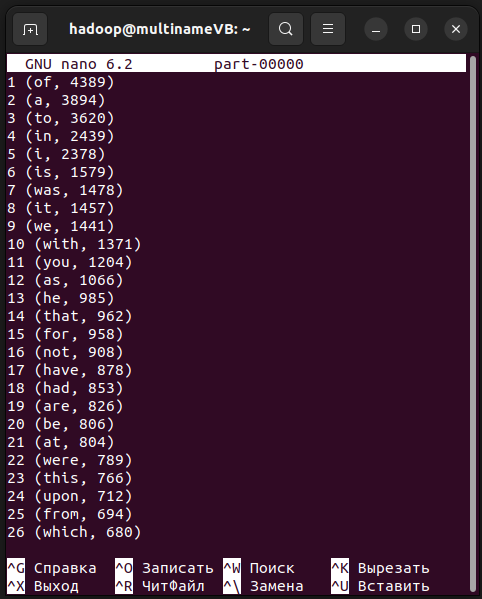
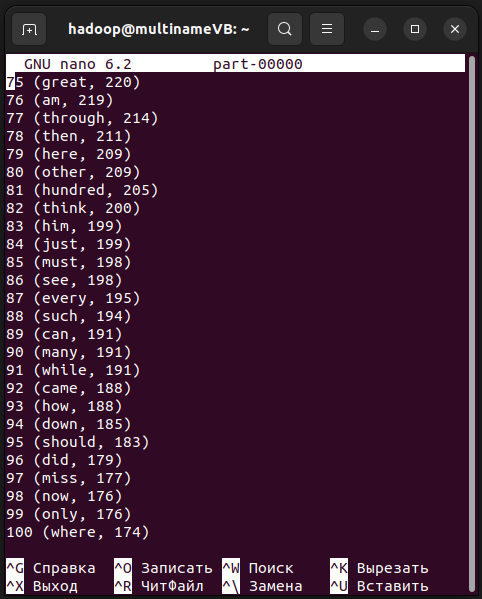
**Рис. 1.** Список стоп-слов



**Рис. 2.** Файлы для обработки



**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.