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 _____ (Сафронов Н.С.)
(подпись) (Ф.И.О.)

Проверил: _____ (Голубева С.Е.)
(подпись) (Ф.И.О.)

Дата сдачи (защиты):

Результаты сдачи (защиты):

- Балльная оценка:
- Оценка:

Калуга, 2023

Цель работы: формирование практических навыков использования парадигмы MapReduce для обработки больших данных.

Постановка задачи

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.

Вариант 4

Построить индекс файла. Для каждого слова в файле результат должен содержать номера всех строка, в которых появляется данное слово. Индекс должен быть регистро-независимым. Результат должен быть сохранен в файле в виде: ((word1 (1 42 58)), (word2 (34, 55, 776, 3456), …).

Ход выполнения работы

Листинг программы

mapper.py:

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

import sys


print("(")
line_number = 1
for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        print(f"({word} ({line_number}))", end=",\n")
    line_number += 1

print(")")
```

reducer.py:

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

import re
import sys

current_word = None
```

```
current_locations = []
word = None

print(")", end="")
for line in sys.stdin:
    regex = re.compile(r"\(\S+\s\(\d+(?:\s+\d+)*\)\)")
    matches = regex.findall(line)
    matches.sort()
    for word_with_info in matches:
        word_with_info = word_with_info.strip("(),")
        word, appearances = word_with_info.split(" ")
        word = word.lower()
        locations = appearances.strip("()\n,")
        locations = locations.split(" ")
        if current_word == word:
            current_locations.append(*locations)
            current_locations = list(set(current_locations))
        else:
            if current_word:
                print(
                    f'({current_word} ({" ".join(current_locations)}))',
                    end=",\n"
                )
            current_locations = locations
            current_word = word
if current_word == word:
    print(f'({current_word} ({" ".join(current_locations)}))', end="")

print(")", end="")
```

Результаты выполнения программы

```
hadoop@keyone-laptop:~/lab2$ hdfs dfs -head /user/hduser/test.txt
The Project Gutenberg eBook of Less than kin

This ebook is for the use of anyone anywhere in the United States and
most other parts of the world at no cost and with almost no restrictions
whatsoever. You may copy it, give it away or re-use it under the terms
of the Project Gutenberg License included with this ebook or online
at www.gutenberg.org. If you are not located in the United States,
you will have to check the laws of the country where you are located
before using this eBook.

Title: Less than kin


Author: Alice Duer Miller

Release date: September 18, 2023 [eBook #71674]
```

Рисунок 1. Входной файл

```
doop@keyone-laptop:~/lab2$ /usr/local/hadoop/bin/mapred streaming -input /user/hduser/test.txt -output /user/hduser/lab2_output -mapper mapper.py -reducer reducer.py
```

Рисунок 2. Запуск MapReduce

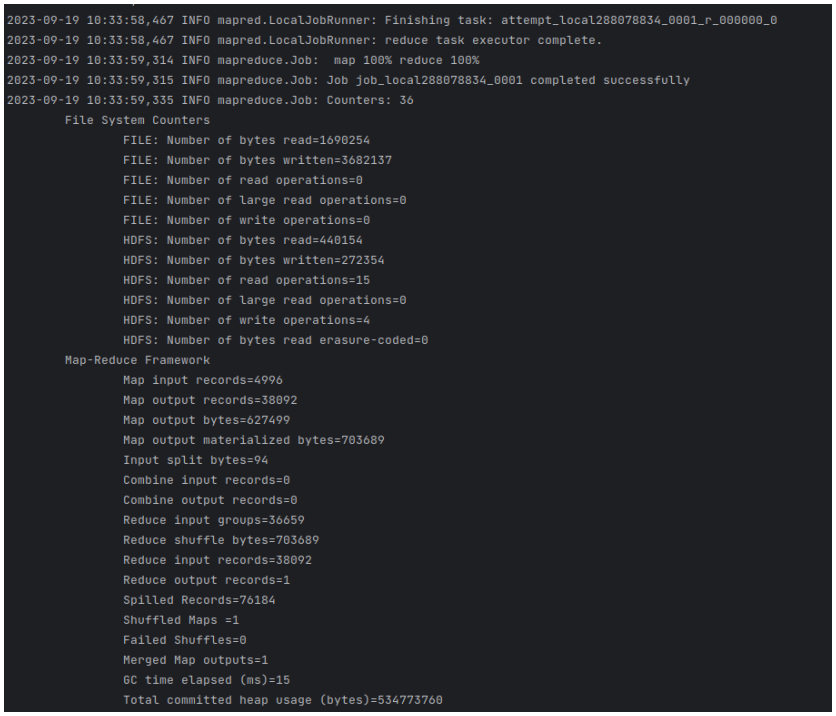

Рисунок 3. Результаты выполнения MapReduce

Рисунок 4. Демонстрация файла с результатом

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