|  |  |
| --- | --- |
| **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**

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

|  |  |  |
| --- | --- | --- |
| Выполнил: студент гр. ИУК4 -72Б | | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (\_\_\_\_\_Калашников А.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**: 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.

**Вариант 6**

Модифицировать программу подсчета слов WordCount для подсчета слов, начинающихся с заданной подстроки. Из результата должны быть удалены стоп-слова.

**Ход выполнения работы**

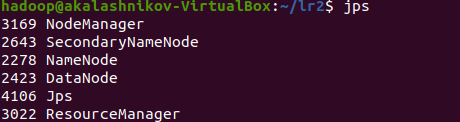
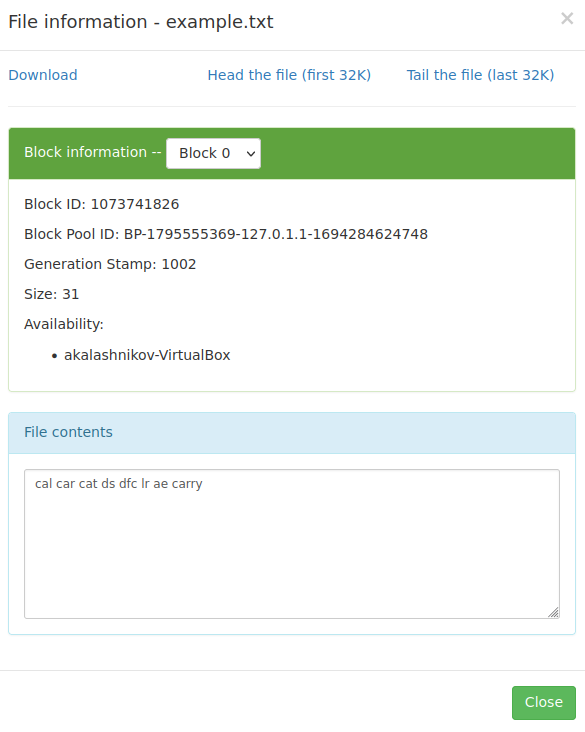
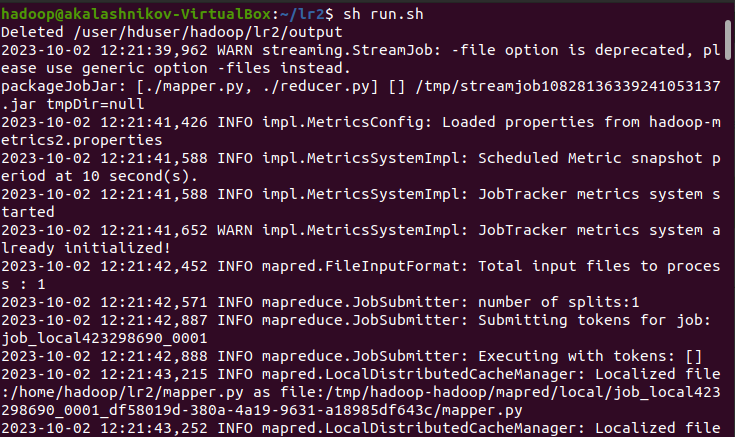


Рис. 1. Демонстрация запуска HDFS



**Рис. 2.** Демонстрация данных для обработки в HDFS

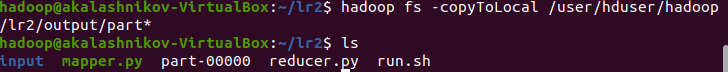
через веб-интерфейс



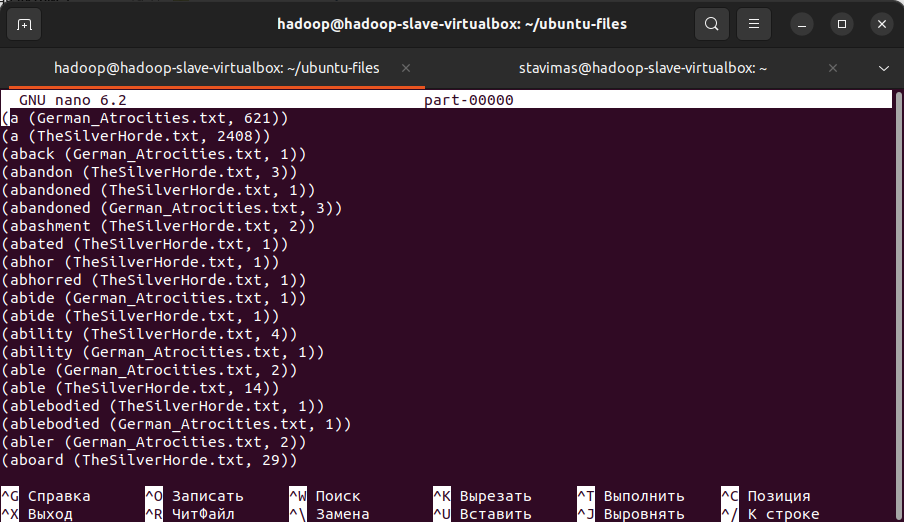
**Рис. 4.** Вызов программы в MapReduce и демонстрация успешного выполнения программы



**Рис. 5.** Демонстрация результата через веб-интерфейс



**Рис. 6.** Копирование результата на локальную файловую систему



**Рис. 7.** Демонстрация результата

**Листинг программы**

**run.sh**

hadoop fs -rm -r /user/hduser/hadoop/lr2/output

mapred streaming \

-input /user/hduser/hadoop/lr2/input \

-output /user/hduser/hadoop/lr2/output \

-file ./mapper.py -mapper 'python3 mapper.py' \

-file ./reducer.py -reducer 'python3 reducer.py'

**mapper.py**

#!/usr/bin/python3.10

import sys

import os

import re

try:

input\_file = os.environ['mapreduce\_map\_input\_file']

except KeyError:

try:

input\_file = os.environ['map\_input\_file']

except KeyError:

input\_file = "NOT\_FILE"

word\_count = 0

stop\_words = [""]

substr = ""

# Открываем файл для чтения

for line in sys.stdin:

# Разделяем строку на отдельные слова

words = line.split()

# Перебираем слова в строке

for word in words:

# Удаляем знаки препинания и приводим слово к нижнему регистру

cleaned\_word = word.strip('.,:;!?').lower()

print(f"{word},1")

**reducer.py**

#!/usr/bin/python3.10

import sys

import os

current\_count = 0

total\_count = 0

substr = "ca"

stop\_words = ["cal", "ds"]

for line in sys.stdin:

word, count = line.strip().split(",")

count = int(count)

if word.startswith(substr) and word not in stop\_words:

current\_count += 1

print(f"Количество слов, начинающихся с \"{substr}\", без стоп-слов: {current\_count}")

**Вывод:** в ходе выполнения данной лабораторной работы были сформированы практические навыки по установке и настройке кластера Hadoop и работе с файловой системой HDFS.