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

```
hadoop@akalashnikov-VirtualBox:~/lr2$ jps
3169 NodeManager
2643 SecondaryNameNode
2278 NameNode
2423 DataNode
4106 Jps
3022 ResourceManager
```

Рис. 1. Демонстрация запуска HDFS

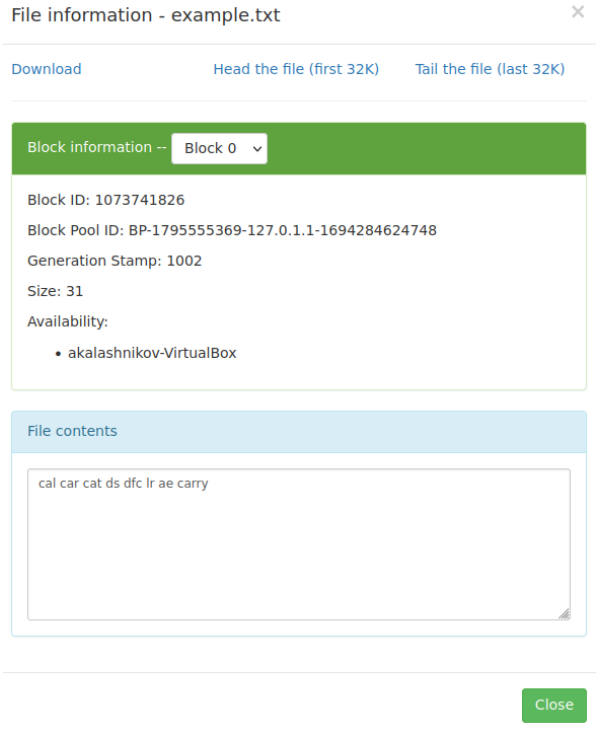

Рис. 2. Демонстрация данных для обработки в HDFS через веб-интерфейс

```
hadoop@akalashnikov-VirtualBox:~/lr2$ sh run.sh
Deleted /user/hduser/hadoop/lr2/output
2023-10-02 12:21:39,962 WARN streaming.StreamJob: -file option is deprecated, pl
ease use generic option -files instead.
packageJobJar: [./mapper.py, ./reducer.py] [] /tmp/streamjob10828136339241053137
.jar tmpDir=null
2023-10-02 12:21:41,426 INFO impl.MetricsConfig: Loaded properties from hadoop-m
etrics2.properties
2023-10-02 12:21:41,588 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot p
eriod at 10 second(s).
2023-10-02 12:21:41,588 INFO impl.MetricsSystemImpl: JobTracker metrics system s
tarted
2023-10-02 12:21:41,652 WARN impl.MetricsSystemImpl: JobTracker metrics system a
lready initialized!
2023-10-02 12:21:42,452 INFO mapred.FileInputFormat: Total input files to proces
s : 1
2023-10-02 12:21:42,571 INFO mapreduce.JobSubmitter: number of splits:1
2023-10-02 12:21:42,887 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local423298690_0001
2023-10-02 12:21:42,888 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-10-02 12:21:43,215 INFO mapred.LocalDistributedCacheManager: Localized file
:/home/hadoop/lr2/mapper.py as file:/tmp/hadoop-hadoop/mapred/local/job_local423
298690_0001_df58019d-380a-4a19-9631-a18985df643c/mapper.py
2023-10-02 12:21:43,252 INFO mapred.LocalDistributedCacheManager: Localized file
```

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

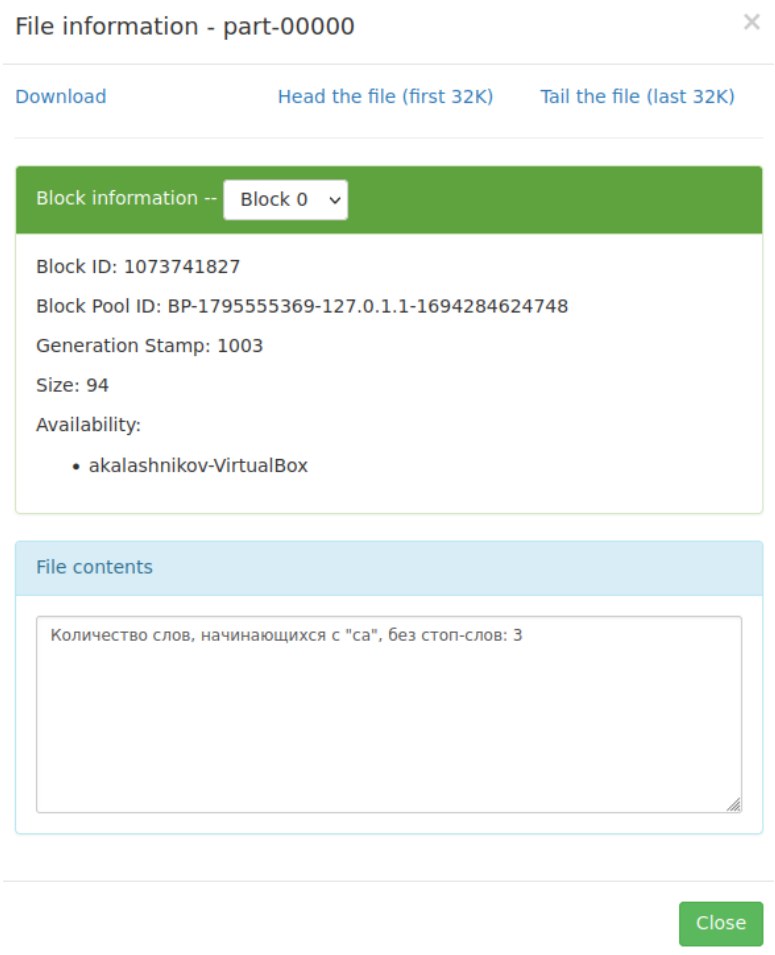

Рис. 5. Демонстрация результата через веб-интерфейс

```
hadoop@akalashnikov-VirtualBox:~/lr2$ hadoop fs -copyToLocal /user/hduser/hadoop
/lr2/output/part*
hadoop@akalashnikov-VirtualBox:~/lr2$ ls
input  mapper.py  part-00000  reducer.py  run.sh
```

Рис. 6. Копирование результата на локальную файловую систему

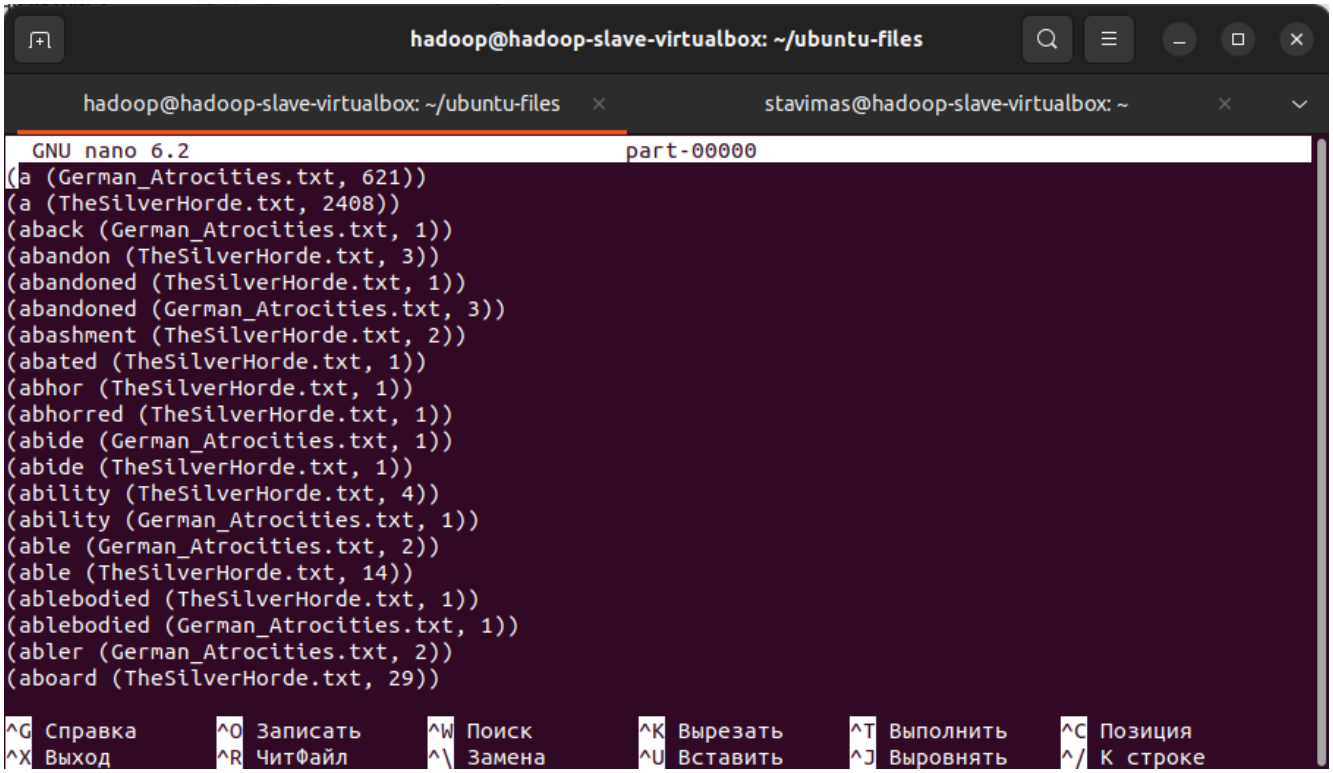

Рис. 7. Демонстрация результата

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.