



Министерство науки и высшего образования Российской Федерации
Калужский филиал федерального государственного автономного
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

«MAPREDUCE»

по дисциплине: «Технологии обработки больших данных»

Выполнил: студент группы ИУК4-72Б

(Подпись)

Губин Е.В.

(И.О. Фамилия)

Проверил:

(Подпись)

Голубева С.Е.

(И.О. Фамилия)

Дата сдачи (защиты):

Результаты сдачи (защиты):

- Балльная оценка:

- Оценка:

Калуга, 2025

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

Задачи:

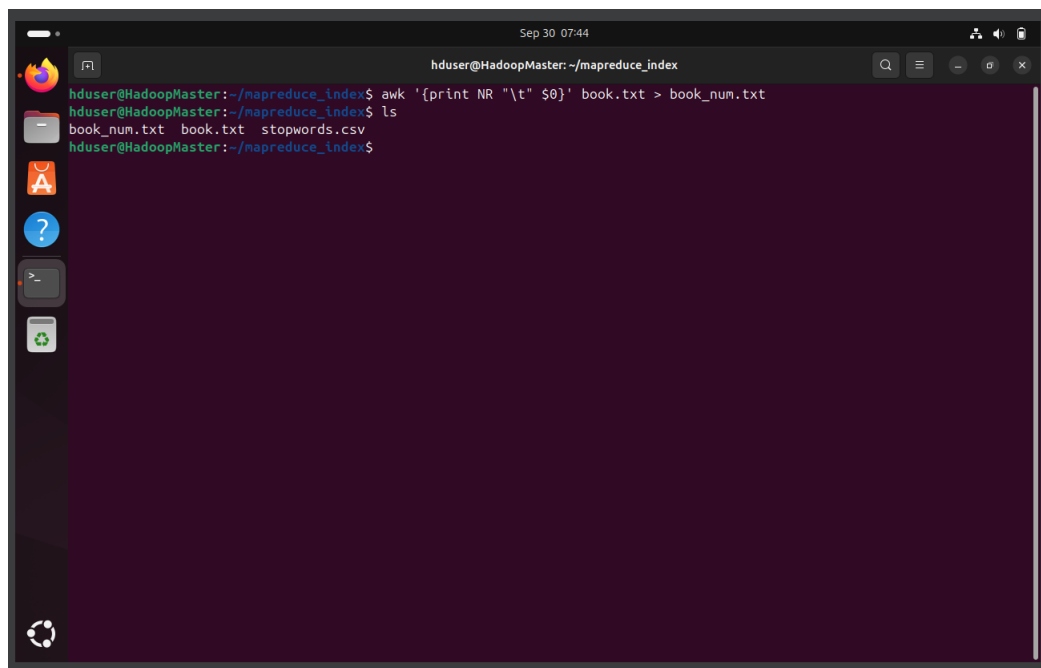
1. Изучить подход MapReduce.
2. Изучить принципы работы Hadoop MapReduce.
3. Получить практические навыки реализации MapReduce задач.
4. Уметь обрабатывать большие текстовые файлы с помощью MapReduce.

Формулировка задания (9 вариант):

Построить индекс файла. Для каждого слова в файле результат должен содержать номера всех строк, 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:



```
hduser@HadoopMaster: ~/mapreduce_index
hduser@HadoopMaster:~/mapreduce_index$ awk '{print NR "\t" $0}' book.txt > book_num.txt
hduser@HadoopMaster:~/mapreduce_index$ ls
book_num.txt  book.txt  stopwords.csv
hduser@HadoopMaster:~/mapreduce_index$
```

Рисунок 1 Нумерация строк в файле

```
Sep 30 07:44
hduser@HadoopMaster: ~/mapreduce_index

1 <feff>The Project Gutenberg eBook of Nora's twin sister
2
3 This ebook is for the use of anyone anywhere in the United States and
4 most other parts of the world at no cost and with almost no restrictions
5 whatsoever. You may copy it, give it away or re-use it under the terms
6 of the Project Gutenberg License included with this ebook or online
7 at www.gutenberg.org. If you are not located in the United States,
8 you will have to check the laws of the country where you are located
9 before using this eBook.
10
11 Title: Nora's twin sister
12
13 Author: Nina Rhoades
14
15 Illustrator: Nana French Bickford
16
17 Release date: September 30, 2025 [eBook #76955]
18
19 Language: English
20
21 Original publication: Boston: Lothrop, Lee & Shepard Co, 1919
22
23 Credits: Susan E., David E. Brown, and the Online Distributed Proofreading Team at https://www.pgdp.net (This fi
24 le was produced from images generously made available by The Internet Archive)
25
26 *** START OF THE PROJECT GUTENBERG EBOOK NORA'S TWIN SISTER ***
27
28
29
30
1,1 Top
```

Рисунок 2 Пронумерованные строки

```
Sep 30 09:24
hduser@HadoopMaster: ~/mapreduce_index_lw_2

1 You may obtain a copy of the License at
2
3 http://www.apache.org/licenses/LICENSE-2.0
4
5 Unless required by applicable law or agreed to in writing, software
6 distributed under the License is distributed on an "AS IS" BASIS,
7 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
8 See the License for the specific language governing permissions and
9 limitations under the License. See accompanying LICENSE file.
10
11 -->
12 <!-- Put site-specific property overrides in this file. -->
13
14 <configuration>
15   <property>
16     <name>mapreduce.framework.name</name>
17     <value>yarn</value>
18   </property>
19   <property>
20     <name>yarn.app.mapreduce.am.env</name>
21     <value>HADOOP_MAPRED_HOME=/home/hduser/hadoop</value>
22   </property>
23   <property>
24     <name>mapreduce.map.env</name>
25     <value>HADOOP_MAPRED_HOME=/home/hduser/hadoop</value>
26   </property>
27   <property>
28     <name>mapreduce.reduce.env</name>
29     <value>HADOOP_MAPRED_HOME=/home/hduser/hadoop</value>
30   </property>
31 </configuration>
32
33 "/home/hadoop/etc/hadoop/mapred-site.xml" 36L, 1221B
36,3 Bot
```

Рисунок 3 Изменение конфигурации hadoop

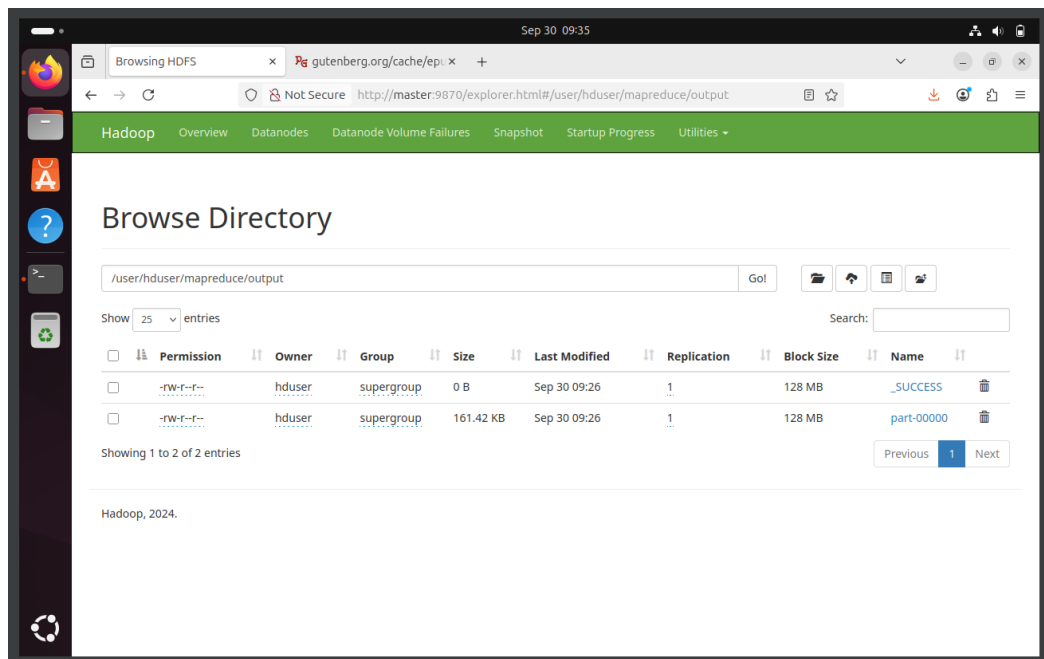


Рисунок 4 Результаты в веб-интерфейсе

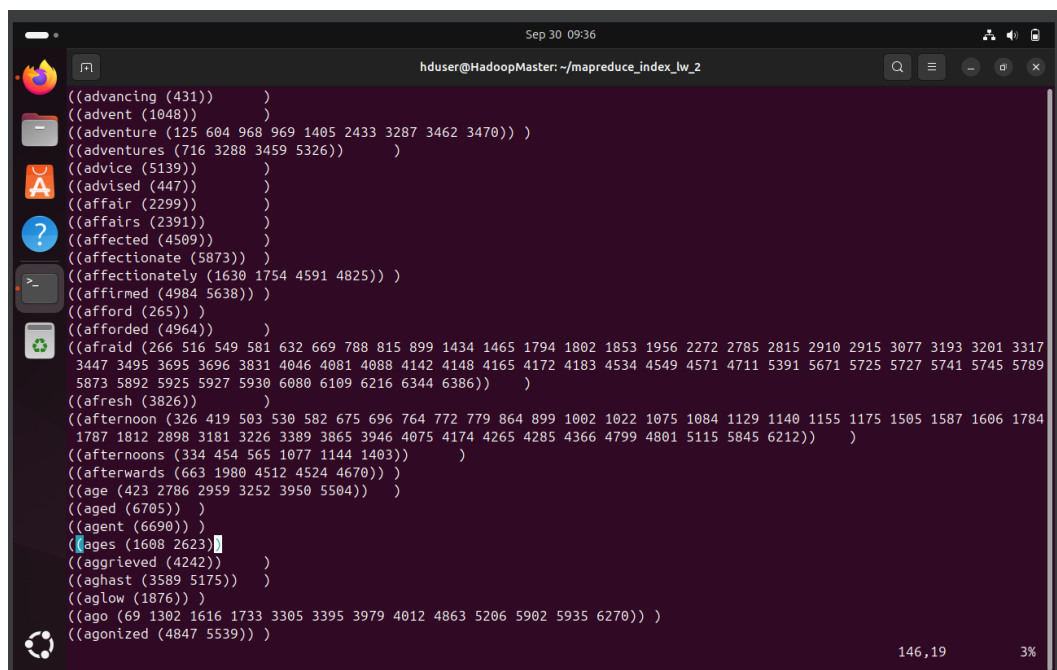


Рисунок 5 Результирующий файл

Листинги программ:

mapper.py:

```
#!/usr/bin/env python3

import sys, csv, re
stopwords = set()
try:
    with open('stopwords.csv', newline='', encoding='utf-8') as f:
        r = csv.reader(f)
        for row in r:
```

```

        if row:
            stopwords.add(row[0].strip().lower())
except Exception:
    stopwords = set()

word_re = re.compile(r"^[^W_]+", flags=re.UNICODE)

for raw in sys.stdin:
    raw = raw.rstrip('\n')
    if not raw:
        continue
    parts = raw.split('\t', 1)
    if len(parts) != 2:
        continue
    line_no, text = parts[0].strip(), parts[1]
    if not line_no.isdigit():
        continue
    for w in word_re.findall(text):
        w = w.lower()
        if not w or w in stopwords:
            continue
        # emit word<TAB>line_number
        print(f"{w}\t{line_no}")

```

reducer.py:

```

#!/usr/bin/env python3

import sys

current = None
lines_set = set()

def emit(word, s):
    if not word:
        return
    nums = sorted(int(x) for x in s)
    nums_str = " ".join(str(n) for n in nums)
    # печатаем кортеж для слова
    print(f"({word}) ({nums_str})")

for line in sys.stdin:
    line = line.strip()
    if not line:
        continue
    try:
        word, ln = line.split('\t', 1)
    except ValueError:
        continue
    if current is None:
        current = word
    if word != current:
        emit(current, lines_set)
        current = word
        lines_set = set()
    lines_set.add(ln)
if current is not None:
    emit(current, lines_set)

```

main.sh:

```
hdfs dfs -rm -r -f /user/hduser/mapreduce/output || true

STREAMING_JAR=$(ls $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming*.jar
| head -n1)
echo "Using streaming jar: $STREAMING_JAR"

hadoop jar "$STREAMING_JAR" \
  -files mapper.py, reducer.py, stopwords.csv \
  -input /user/hduser/mapreduce/input/book_num.txt \
  -output /user/hduser/mapreduce/output \
  -mapper "python3 mapper.py" \
  -reducer "python3 reducer.py" \
  -numReduceTasks 1
```

get_result.sh:

```
hdfs dfs -ls /user/hduser/mapreduce/output

hdfs dfs -cat /user/hduser/mapreduce/output/part-00000 | head -n 20

hdfs dfs -cat /user/hduser/mapreduce/output/part-00000 \
  | paste -sd', ' - \
  | sed 's/^/(/; s/$/)/' \
  > ~/mapreduce_index/index_result.txt

head -c 2000 ~/mapreduce_index/index_result.txt
```

Вывод: в ходе лабораторной работы были получены практические навыки по работе с MapReduce.