

**Федеральное агентство связи  
Ордена трудового Красного Знамени  
Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«Московский технический университет связи и информатики»**



**Практическая работа № 2  
По дисциплине  
Введение в большие данные**

Группа: МБД2431  
ФИО: Киреев Артём Александрович

**Москва, 2025**

## Цель работы: получить навыки работы с MapReduce и YARN

### Первый запуск

```
yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar  
pi 5 123456789
```

```
-sh-4.2$ yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar pi 5 123456789  
Number of Maps = 5  
Samples per Map = 123456789  
Wrote input for Map #0  
Wrote input for Map #1  
Wrote input for Map #2  
Wrote input for Map #3  
Wrote input for Map #4  
Starting Job  
25/04/08 21:45:37 INFO client.RMProxy: Connecting to ResourceManager at node2.mtuci.cloud.ru/192.168.0.4:8050  
Job Finished in 21.211 seconds  
Estimated value of Pi is 3.14159321930849829571
```

Рисунок 1 - Запуск MapReduce для вычисления числа  $\pi$

```
yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar  
pi 5 12345678987
```

```
-sh-4.2$ yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar pi 5 1234567890  
Number of Maps = 5  
Samples per Map = 1234567890  
Wrote input for Map #0  
Wrote input for Map #1  
Wrote input for Map #2  
Wrote input for Map #3  
Wrote input for Map #4  
Starting Job  
25/04/06 14:22:08 INFO client.RMProxy: Connecting to ResourceManager at node2.mtuci.cloud.ru/192.168.0.4:8050  
25/04/06 14:22:09 INFO client.AHSProxy: Connecting to Application History server at node3.mtuci.cloud.ru/192.168.0.5:8050  
25/04/06 14:22:09 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/kireev_mbd_2431/stage/job_1743703651672_0035  
25/04/06 14:22:09 INFO input.FileInputFormat: Total input files to process : 5  
25/04/06 14:22:09 INFO mapreduce.JobSubmitter: number of splits:5  
25/04/06 14:22:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1743703651672_0035  
25/04/06 14:22:09 INFO mapreduce.JobSubmitter: Executing with tokens: []  
25/04/06 14:22:09 INFO conf.Configuration: found resource resource-types.xml at file:/etc/hadoop/3.1.4.0-315/0/resource-types.xml  
25/04/06 14:22:09 INFO impl.YarnClientImpl: Submitted application application_1743703651672_0035  
25/04/06 14:22:09 INFO mapreduce.Job: The url to track the job: http://node2.mtuci.cloud.ru:8088/proxy/application_1743703651672_0035/  
25/04/06 14:22:09 INFO mapreduce.Job: Running job: job_1743703651672_0035  
25/04/06 14:22:14 INFO mapreduce.Job: Job job_1743703651672_0035 running in uber mode : false  
25/04/06 14:22:14 INFO mapreduce.Job: map 0% reduce 0%  
Job Finished in 36.026 seconds  
Estimated value of Pi is 3.14159277542849425640
```

Рисунок 2 - Второй запуск (с увеличенным количеством точек)

Войди в Ambari под пользователем monitor/monitor и перейди в раздел Services  
-> YARN -> Quick Links -> ResourceManager UI

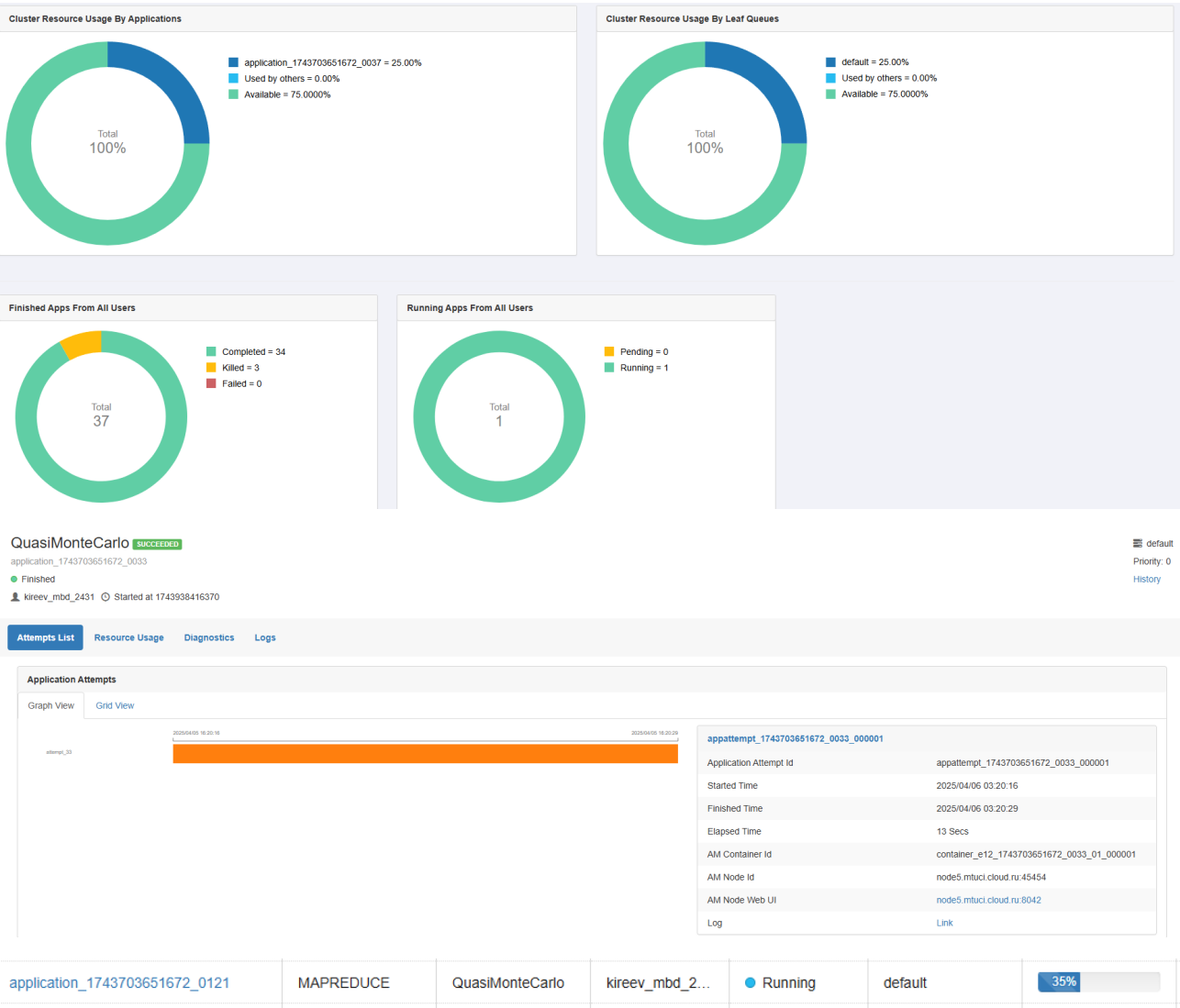
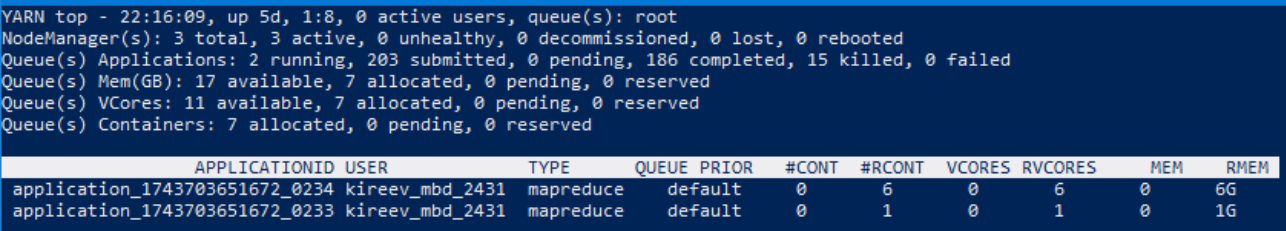


Рисунок 3 - Мониторинг через Ambari

## Использование yarn top и принудительное завершение задачи

yarn top

yarn app -kill application\_1743703651672\_0234



```

^C-sh-4.2$ yarn app -kill application_1743703651672_0233
25/04/08 22:17:09 INFO client.RMProxy: Connecting to ResourceManager at node2.mtuci.cloud.ru/192.168.0.4:8050
25/04/08 22:17:10 INFO client.AHSProxy: Connecting to Application History server at node3.mtuci.cloud.ru/192.168.0.5:10200
Killing application application_1743703651672_0233
25/04/08 22:17:10 INFO impl.YarnClientImpl: Killed application application_1743703651672_0233
-sh-4.2$ yarn app -kill application_1743703651672_0234
25/04/08 22:17:19 INFO client.RMProxy: Connecting to ResourceManager at node2.mtuci.cloud.ru/192.168.0.4:8050
25/04/08 22:17:20 INFO client.AHSProxy: Connecting to Application History server at node3.mtuci.cloud.ru/192.168.0.5:10200
Killing application application_1743703651672_0234
25/04/08 22:17:20 INFO impl.YarnClientImpl: Killed application application_1743703651672_0234
-sh-4.2$ █

```

Рисунок 4 - Остановка задачи

## Nadoop Streaming с удалением стоп-слов

### Подготовка ВХОДНЫХ ДАННЫХ И СТОП-СЛОВ

#### mapper\_length.py

```

#!/usr/bin/env python3

import sys

import re

for line in sys.stdin:
    words = re.findall(r'\b\w+\b', line.lower())

    for word in words:
        if 6 <= len(word) <= 9:
            print(f"{word}\t1")

```

#### mapper\_stopwords.py

```

#!/usr/bin/env python3

import sys

import re

stopwords = set([
    'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you',
    'you're", "you've", "you'll", "you'd",
    'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself',
    'she', "she's", 'her', 'hers',
    'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs',
    'themselves', 'what', 'which',
    'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is',
    'are', 'was', 'were', 'be', 'been',

```

```

    'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a',
    'an', 'the', 'and', 'but', 'if',

    'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with',
    'about', 'against', 'between',

    'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to',
    'from', 'up', 'down', 'in', 'out',

    'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here',
    'there', 'when', 'where', 'why',

    'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some',
    'such', 'no', 'nor', 'not',

    'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will',
    'just', 'don', "don't",

    'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain',
    'aren', "aren't", 'couldn',

    "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn',
    "hasn't", 'haven', "haven't",

    'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn',
    "needn't", 'shan', "shan't",

    'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won',
    "won't", 'wouldn', "wouldn't"
])

for line in sys.stdin:
    words = re.findall(r'\b\w+\b', line.lower())

    filtered = [word for word in words if word not in stopwords]

    for word in filtered:
        print(word)

```

## reducer\_sum.py

```

#!/usr/bin/env python3

import sys

current_word = None
current_count = 0

for line in sys.stdin:
    word, count = line.strip().split("\t")

    count = int(count)

    if current_word == word:
        current_count += count
    else:
        if current_word:

```

```

        print(f"{current_word}\t{current_count}")

    current_word = word
    current_count = count

if current_word == word:
    print(f"{current_word}\t{current_count}")

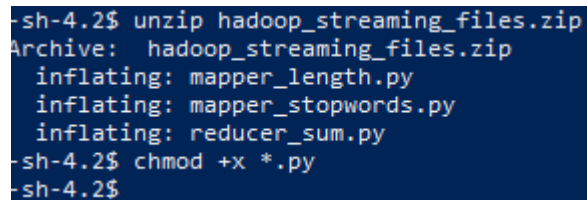
```

## Делаем файлы исполняемыми и закидываем их в HDFS

```

hdfs dfs -mkdir -p /data/kireev_mbd_2431/
hdfs dfs -chmod 777 /data/kireev_mbd_2431/
unzip hadoop_streaming_files.zip
chmod +x *.py

```



```

-sh-4.2$ unzip hadoop_streaming_files.zip
Archive:  hadoop_streaming_files.zip
  inflating: mapper_length.py
  inflating: mapper_stopwords.py
  inflating: reducer_sum.py
-sh-4.2$ chmod +x *.py
-sh-4.2$

```

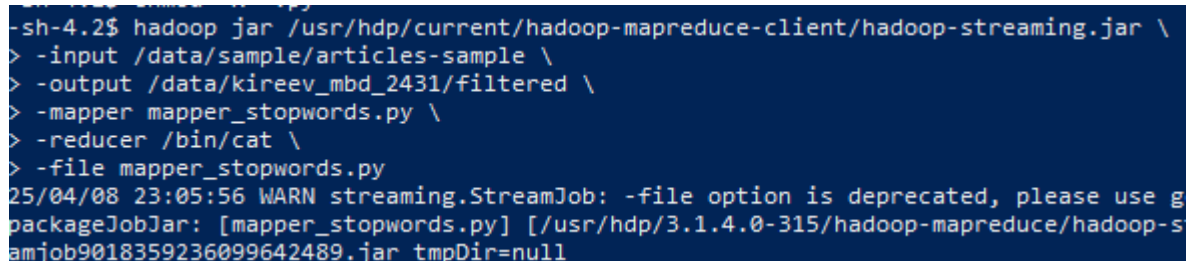
Рисунок 5 - Загрузка файлов

## Запуск Hadoop Streaming удаление стоп слов

```

hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar \
-input /data/sample/articles-sample \
-output /data/kireev_mbd_2431/filtered \
-mapper mapper_stopwords.py \
-reducer /bin/cat \
-file mapper_stopwords.py

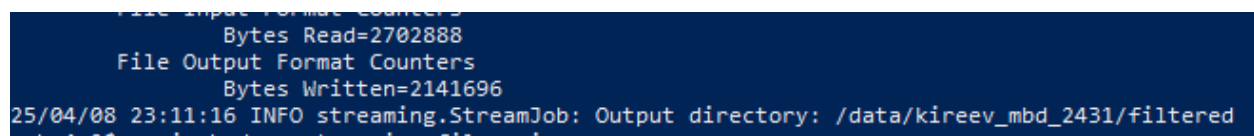
```



```

-sh-4.2$ hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar \
> -input /data/sample/articles-sample \
> -output /data/kireev_mbd_2431/filtered \
> -mapper mapper_stopwords.py \
> -reducer /bin/cat \
> -file mapper_stopwords.py
25/04/08 23:05:56 WARN streaming.StreamJob: -file option is deprecated, please use g
packageJobJar: [mapper_stopwords.py] [/usr/hdp/3.1.4.0-315/hadoop-mapreduce/hadoop-s
amjob9018359236099642489.jar tmpDir=null

```



```

File Input Format Counters
  Bytes Read=2702888
File Output Format Counters
  Bytes Written=2141696
25/04/08 23:11:16 INFO streaming.StreamJob: Output directory: /data/kireev_mbd_2431/filtered

```

Рисунок 6 - Запуск команды

```

hdfs dfs -ls /data/kireev_mbd_2431/filtered

```

```
hdfs dfs -cat /data/kireev_mbd_2431/filtered/part-00000
```

```
-sh-4.2$ hdfs dfs -ls /data/kireev_mbd_2431/filtered
Found 2 items
-rw-r--r--  3 kireev_mbd_2431 hdfs          0 2025-04-08 23:06 /data/kireev_mbd_2431/filtered/_SUCCESS
-rw-r--r--  3 kireev_mbd_2431 hdfs    2141696 2025-04-08 23:06 /data/kireev_mbd_2431/filtered/part-00000
-sh-4.2$
```

Рисунок 7 - Просмотр результатов

## Сортировка по убыванию количества слов

```
hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar \
-input /data/sample/articles-sample \
-output /data/kireev_mbd_2431/wordcount_6to9 \
-mapper mapper_length.py \
-reducer reducer_sum.py \
-file mapper_length.py \
-file reducer_sum.py
```

```
-sh-4.2$ hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar \
> -input /data/sample/articles-sample \
> -output /data/kireev_mbd_2431/wordcount_6to9 \
> -mapper mapper_length.py \
> -reducer reducer_sum.py \
> -file mapper_length.py \
> -file reducer_sum.py
25/04/08 23:20:35 WARN streaming.StreamJob: -file option is deprecated, please use gene
```

Рисунок 8 - Запуск команды

```
hdfs dfs -cat /data/kireev_mbd_2431/wordcount_6to9/* | sort -k2 -nr | head -n 10
```

```
25/04/08 23:25:54 INFO streaming.StreamJob: Output directory: /data/kireev_mbd_2431/wordcount_6to9
-sh-4.2$
25/04/08 23:25:54 INFO streaming.StreamJob: Output directory: /data/kireev_mbd_2431/wordcount_6to9
-sh-4.2$ hdfs dfs -cat /data/kireev_mbd_2431/wordcount_6to9/* | sort -k2 -nr | head -n 10
states 480
during 445
apollo 438
between 428
people 380
lincoln 369
century 369
united 348
american 318
through 315
-sh-4.2$
```

Рисунок 9 - Результат сортировки