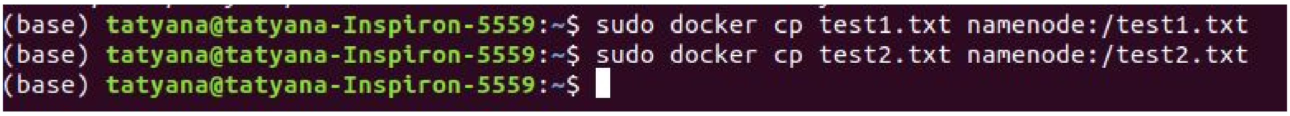
Лабораторная работа №4. Map-reduce: WordCount в Hadoop. Выполнила Бабушкина Татьяна

1. Создадим два текстовых файла для тестирования программы.
   a) test1.txt с текстом:
   First test

   b) test2.txt с текстом:
   This is the second test
   123
   Hello World!

2. Запускаем docker-compose: sudo docker-compose up

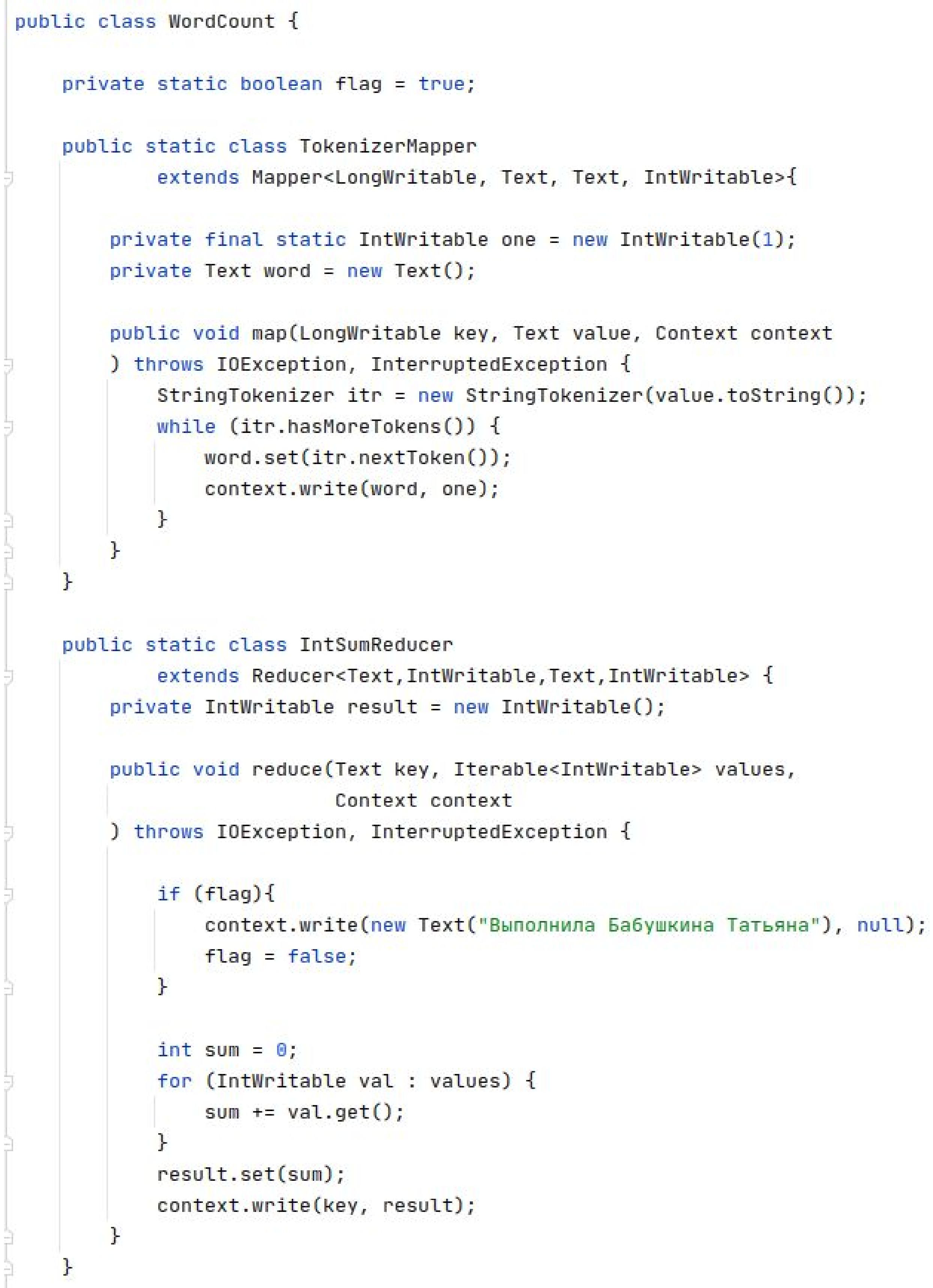
3. Копируем ранее созданные файлы test1.txt и test2.txt в ноду namenode, имеющую доступ к hdfs:



```
(base) tatyana@tatyana-Inspiron-5559:~$ sudo docker cp test1.txt namenode:/test1.txt
(base) tatyana@tatyana-Inspiron-5559:~$ sudo docker cp test2.txt namenode:/test2.txt
(base) tatyana@tatyana-Inspiron-5559:~$
```

4. Реализуем WordCount с помощью исходников на сайте: http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial

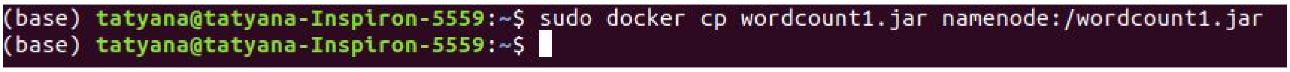
Допишем строчку для вывода фамилии и имени в файл ответа.



```
public class WordCount {

    private static boolean flag = true;

    public static class TokenizerMapper
            extends Mapper<LongWritable, Text, Text, IntWritable>{

        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(LongWritable key, Text value, Context context
        ) throws IOException, InterruptedException {
            StringTokenizer itr = new StringTokenizer(value.toString());
            while (itr.hasMoreTokens()) {
                word.set(itr.nextToken());
                context.write(word, one);
            }
        }
    }

    public static class IntSumReducer
            extends Reducer<Text,IntWritable,Text,IntWritable> {
        private IntWritable result = new IntWritable();

        public void reduce(Text key, Iterable<IntWritable> values,
                           Context context
        ) throws IOException, InterruptedException {

            if (flag){
                context.write(new Text("Выполнила Бабушкина Татьяна"), null);
                flag = false;
            }

            int sum = 0;
            for (IntWritable val : values) {
                sum += val.get();
            }
            result.set(sum);
            context.write(key, result);
        }
    }
```

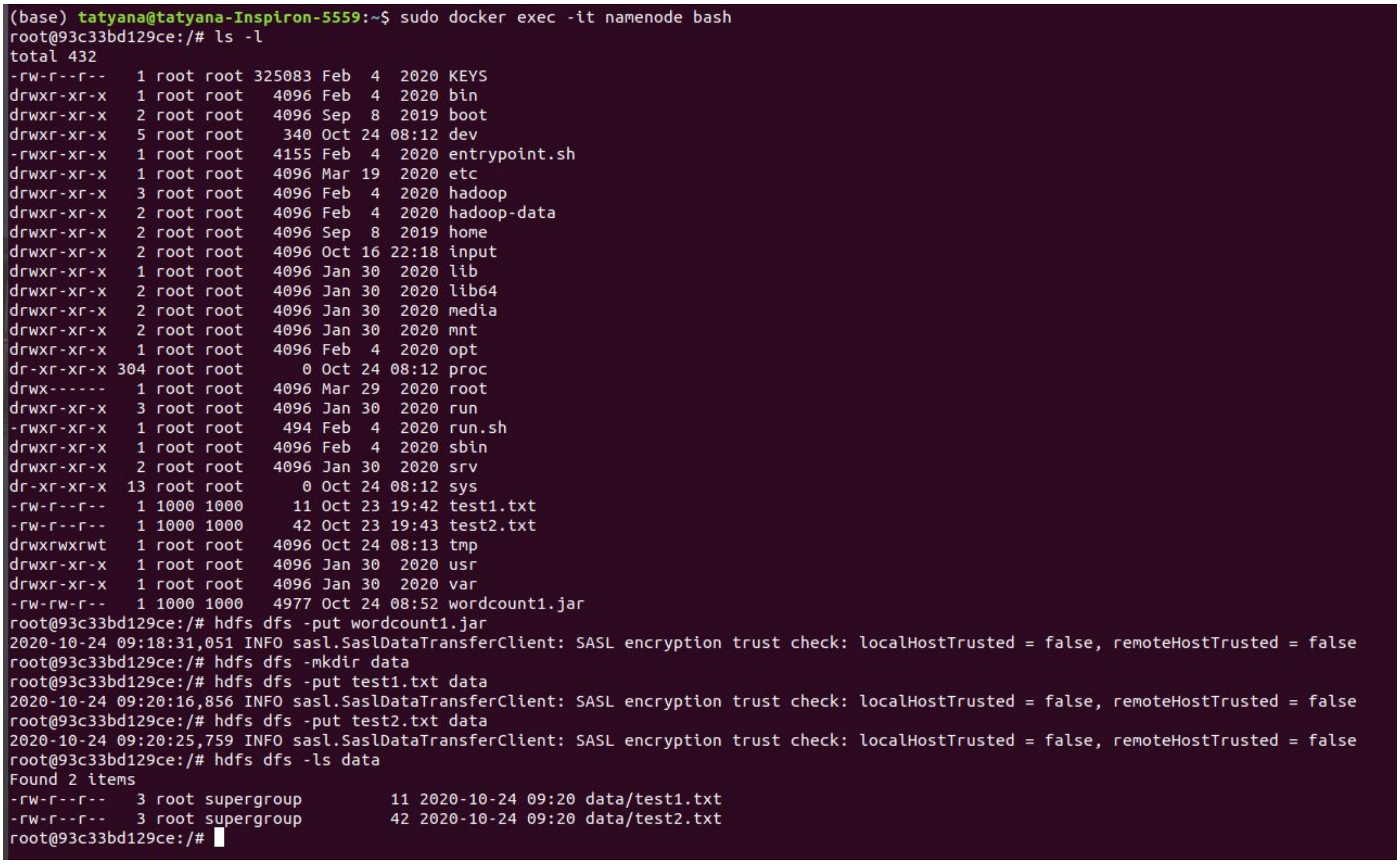
5. Собираем проект, получаем jar-артефакт.

6. Собранный jar-артефакт копируем в namenode:



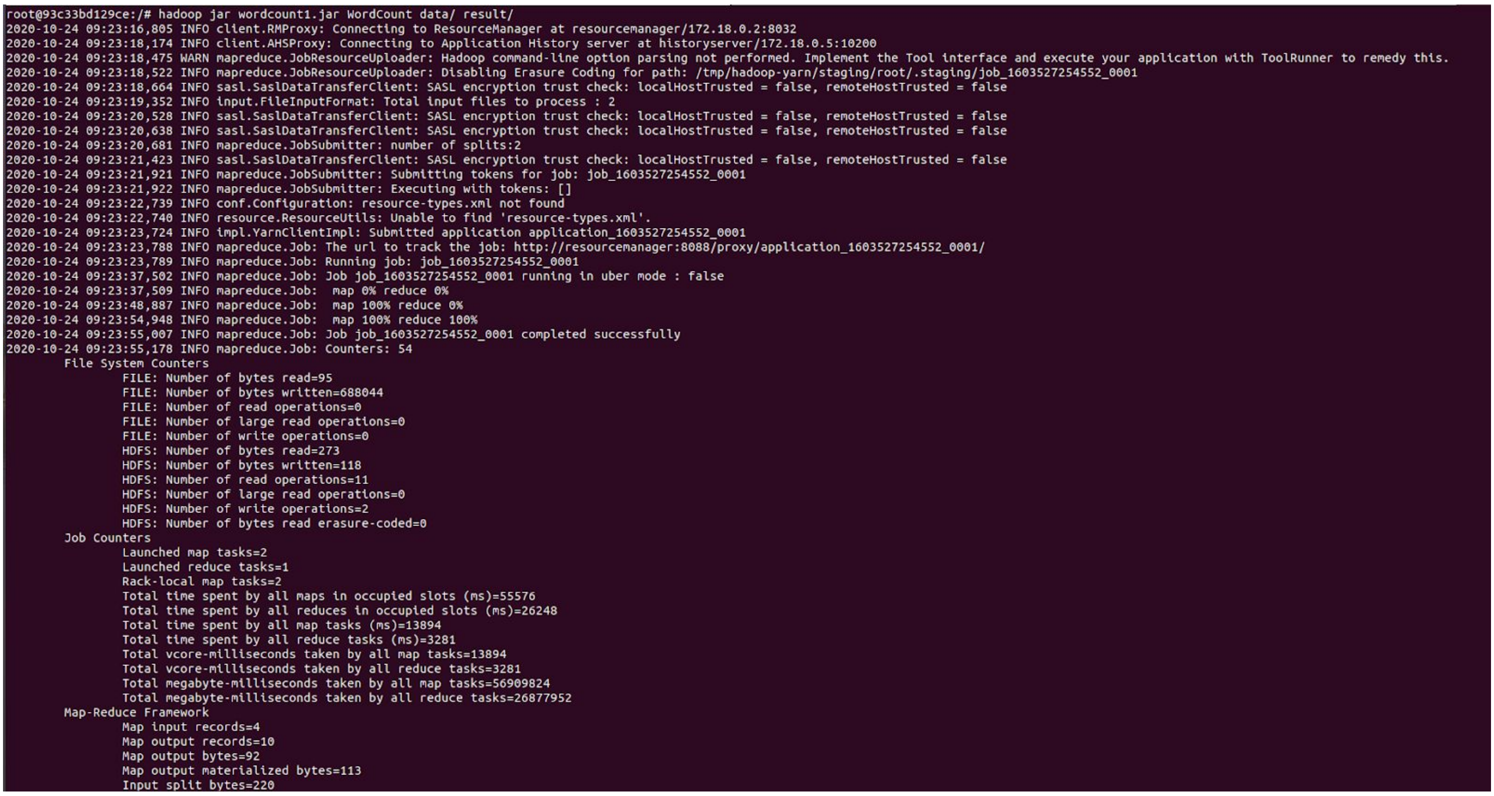
```
(base) tatyana@tatyana-Inspiron-5559:~$ sudo docker cp wordcount1.jar namenode:/wordcount1.jar
(base) tatyana@tatyana-Inspiron-5559:~$ 
```

7. Создаём в hdfs папку data и копируем туда файлы test1.txt и test2.txt для обработки программой.



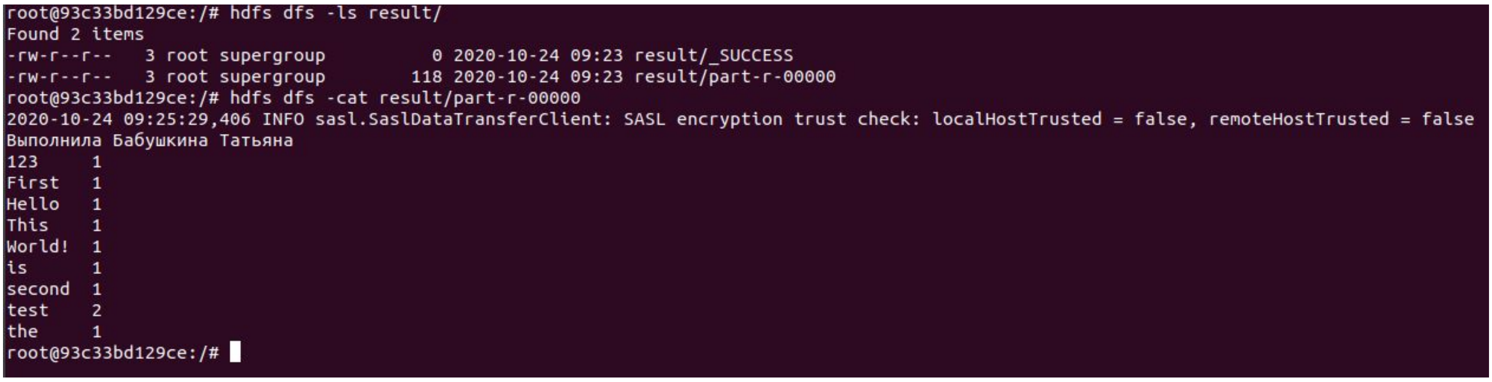
```
(base) tatyana@tatyana-Inspiron-5559:~$ sudo docker exec -it namenode bash
root@93c33bd129ce:/# ls -l
total 432
-rw-r--r--   1 root root 325083 Feb  4  2020 KEYS
drwxr-xr-x   1 root root   4096 Feb  4  2020 bin
drwxr-xr-x   2 root root   4096 Sep  8  2019 boot
drwxr-xr-x   5 root root    340 Oct 24 08:12 dev
-rwxr-xr-x   1 root root   4155 Feb  4  2020 entrypoint.sh
drwxr-xr-x   1 root root   4096 Mar 19  2020 etc
drwxr-xr-x   3 root root   4096 Feb  4  2020 hadoop
drwxr-xr-x   2 root root   4096 Feb  4  2020 hadoop-data
drwxr-xr-x   2 root root   4096 Sep  8  2019 home
drwxr-xr-x   2 root root   4096 Oct 16 22:18 input
drwxr-xr-x   1 root root   4096 Jan 30  2020 lib
drwxr-xr-x   2 root root   4096 Jan 30  2020 lib64
drwxr-xr-x   2 root root   4096 Jan 30  2020 media
drwxr-xr-x   2 root root   4096 Jan 30  2020 mnt
drwxr-xr-x   1 root root   4096 Feb  4  2020 opt
dr-xr-xr-x 304 root root      0 Oct 24 08:12 proc
drwx------   1 root root   4096 Mar 29  2020 root
drwxr-xr-x   3 root root   4096 Jan 30  2020 run
-rwxr-xr-x   1 root root    494 Feb  4  2020 run.sh
drwxr-xr-x   1 root root   4096 Feb  4  2020 sbin
drwxr-xr-x   2 root root   4096 Jan 30  2020 srv
dr-xr-xr-x  13 root root      0 Oct 24 08:12 sys
-rw-r--r--   1 1000 1000     11 Oct 23 19:42 test1.txt
-rw-r--r--   1 1000 1000     42 Oct 23 19:43 test2.txt
drwxrwxrwt   1 root root   4096 Oct 24 08:13 tmp
drwxr-xr-x   1 root root   4096 Jan 30  2020 usr
drwxr-xr-x   1 root root   4096 Jan 30  2020 var
-rw-rw-r--   1 1000 1000   4977 Oct 24 08:52 wordcount1.jar
root@93c33bd129ce:/# hdfs dfs -put wordcount1.jar
2020-10-24 09:18:31,051 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@93c33bd129ce:/# hdfs dfs -mkdir data
root@93c33bd129ce:/# hdfs dfs -put test1.txt data
2020-10-24 09:20:16,856 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@93c33bd129ce:/# hdfs dfs -put test2.txt data
2020-10-24 09:20:25,759 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@93c33bd129ce:/# hdfs dfs -ls data
Found 2 items
-rw-r--r--   3 root supergroup         11 2020-10-24 09:20 data/test1.txt
-rw-r--r--   3 root supergroup         42 2020-10-24 09:20 data/test2.txt
root@93c33bd129ce:/# 
```

8. Запускаем WordCount с указанием папок для входных и выходных данных внутри hadoop.



```
root@93c33bd129ce:/# hadoop jar wordcount1.jar WordCount data/ result/
2020-10-24 09:23:16,805 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.18.0.2:8032
2020-10-24 09:23:18,174 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.18.0.5:10200
2020-10-24 09:23:18,475 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2020-10-24 09:23:18,522 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1603527254552_0001
2020-10-24 09:23:18,664 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-24 09:23:19,352 INFO input.FileInputFormat: Total input files to process : 2
2020-10-24 09:23:20,528 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-24 09:23:20,638 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-24 09:23:20,681 INFO mapreduce.JobSubmitter: number of splits:2
2020-10-24 09:23:21,423 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-24 09:23:21,921 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1603527254552_0001
2020-10-24 09:23:21,922 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-10-24 09:23:22,739 INFO conf.Configuration: resource-types.xml not found
2020-10-24 09:23:22,740 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2020-10-24 09:23:23,724 INFO impl.YarnClientImpl: Submitted application application_1603527254552_0001
2020-10-24 09:23:23,788 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1603527254552_0001/
2020-10-24 09:23:23,789 INFO mapreduce.Job: Running job: job_1603527254552_0001
2020-10-24 09:23:37,502 INFO mapreduce.Job: Job job_1603527254552_0001 running in uber mode : false
2020-10-24 09:23:37,509 INFO mapreduce.Job:  map 0% reduce 0%
2020-10-24 09:23:48,887 INFO mapreduce.Job:  map 100% reduce 0%
2020-10-24 09:23:54,948 INFO mapreduce.Job:  map 100% reduce 100%
2020-10-24 09:23:55,007 INFO mapreduce.Job: Job job_1603527254552_0001 completed successfully
2020-10-24 09:23:55,178 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=95
                FILE: Number of bytes written=688044
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=273
                HDFS: Number of bytes written=118
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=55576
                Total time spent by all reduces in occupied slots (ms)=26248
                Total time spent by all map tasks (ms)=13894
                Total time spent by all reduce tasks (ms)=3281
                Total vcore-milliseconds taken by all map tasks=13894
                Total vcore-milliseconds taken by all reduce tasks=3281
                Total megabyte-milliseconds taken by all map tasks=56909824
                Total megabyte-milliseconds taken by all reduce tasks=26877952
        Map-Reduce Framework
                Map input records=4
                Map output records=10
                Map output bytes=92
                Map output materialized bytes=113
                Input split bytes=220
```

9. После выполнения программы открываем папку result и проверяем количество посчитанных слов. Убеждаемся, что все слова посчитаны верно и фамилия-имя выводятся.



```
root@93c33bd129ce:/# hdfs dfs -ls result/
Found 2 items
-rw-r--r--   3 root supergroup          0 2020-10-24 09:23 result/_SUCCESS
-rw-r--r--   3 root supergroup        118 2020-10-24 09:23 result/part-r-00000
root@93c33bd129ce:/# hdfs dfs -cat result/part-r-00000
2020-10-24 09:25:29,406 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Выполнила Бабушкина Татьяна
123     1
First   1
Hello   1
This    1
World!  1
is      1
second  1
test    2
the     1
root@93c33bd129ce:/# 
```