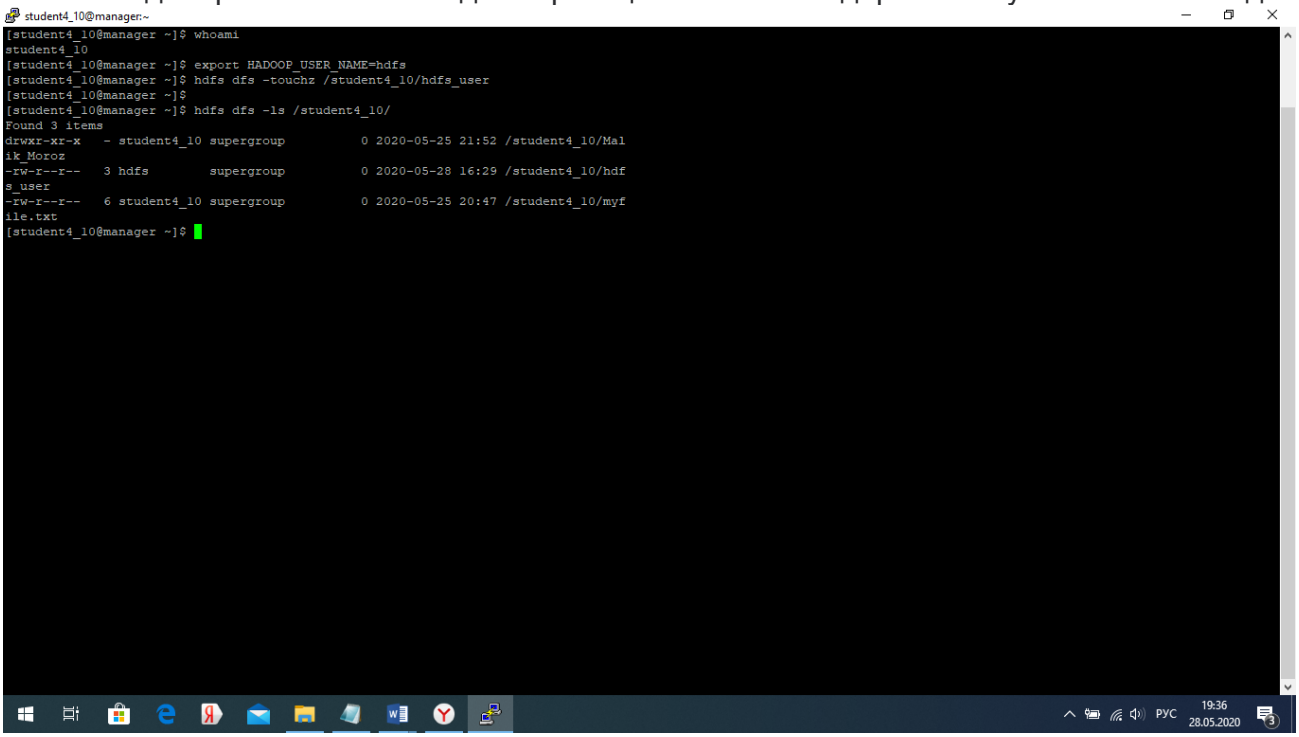
0. [Исследовательское задание]
Как получить доступ под пользователем hdfs в файловую систему не имея sudo?
Если до воскресенья никто не расскажет -- будет подсказка.

Создам переменную

- export HADOOP_USER_NAME = HDFS
- Вовремя создание нового файла в папке student4_10 с моей аутентификацией будет создан файл hdfs user под авторизацией hdfs благодаря выше указанной команде !

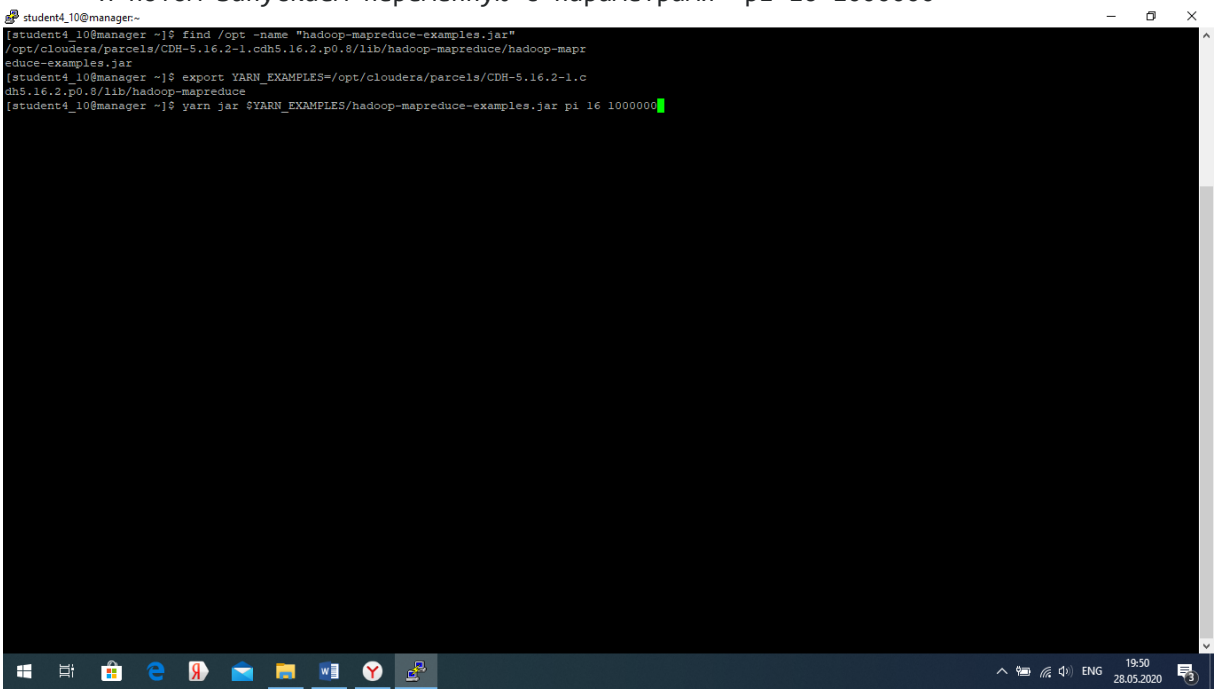


1. Опробовать запуски map-reduce задач для кластера используя hadoop-mapreduce-examples.jar.
а. Выполнить три любых задачи включенных в этот JAR.
б. Найти свои задачи в интерфейсе Cloudera Manager
в. Опробовать навигацию по интерфейсу YARN

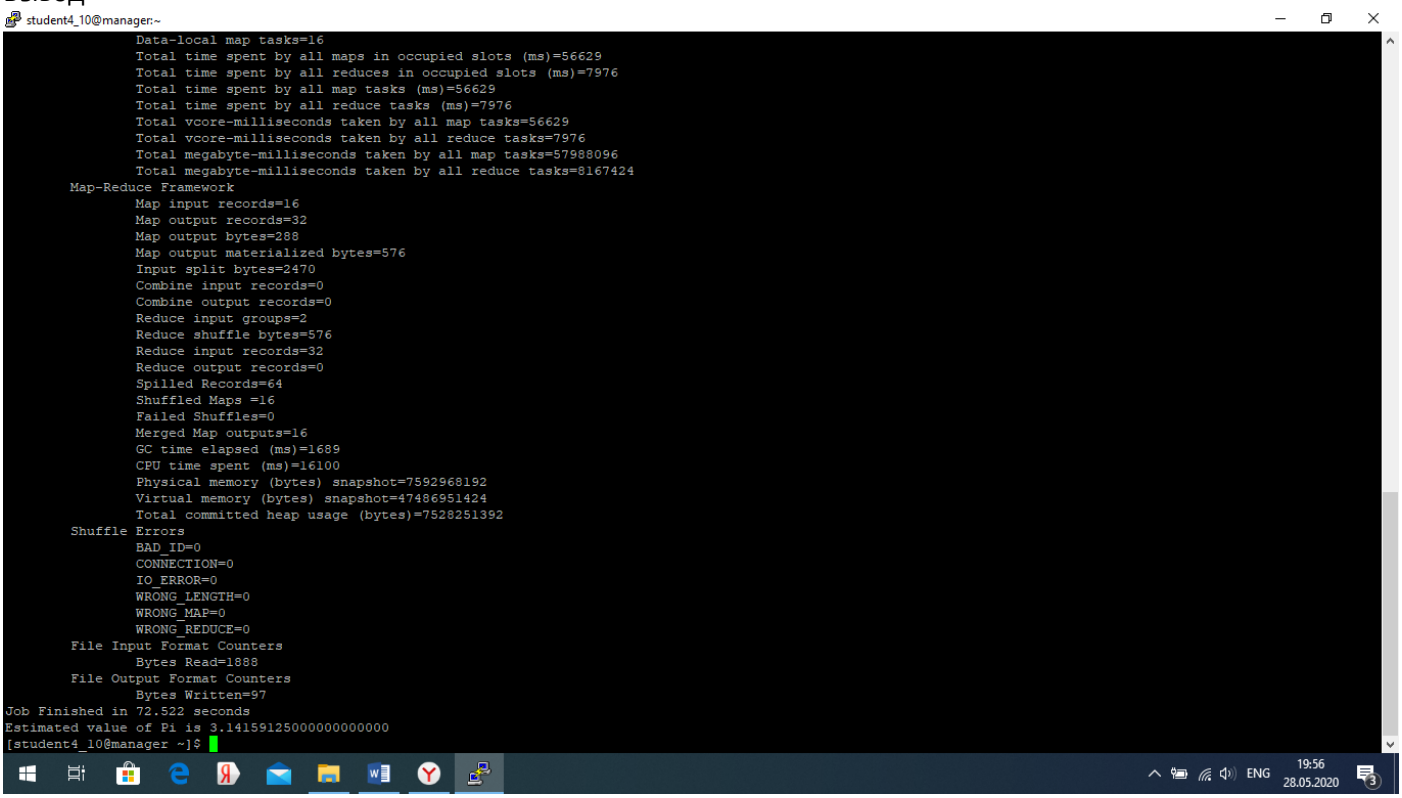
1. найдем jar file

```
find /some_directory -name some_file_name
```

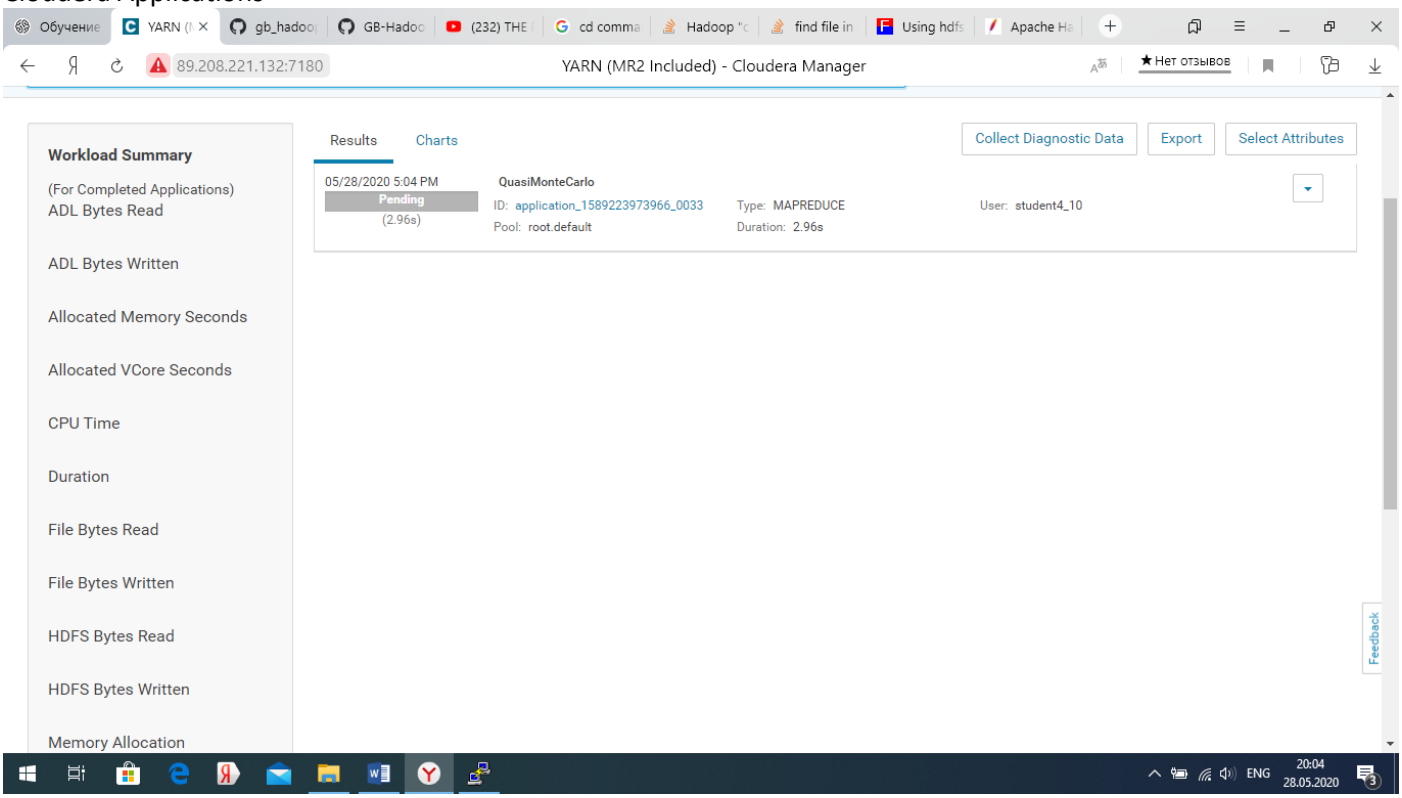
- Находим файк Hadoop-mapreduce через find
- Создаем переменную YARN_EXAMPLES
- И потом запускаем переменную с параметрами pi 16 1000000



Вывод



Cloudera Applications

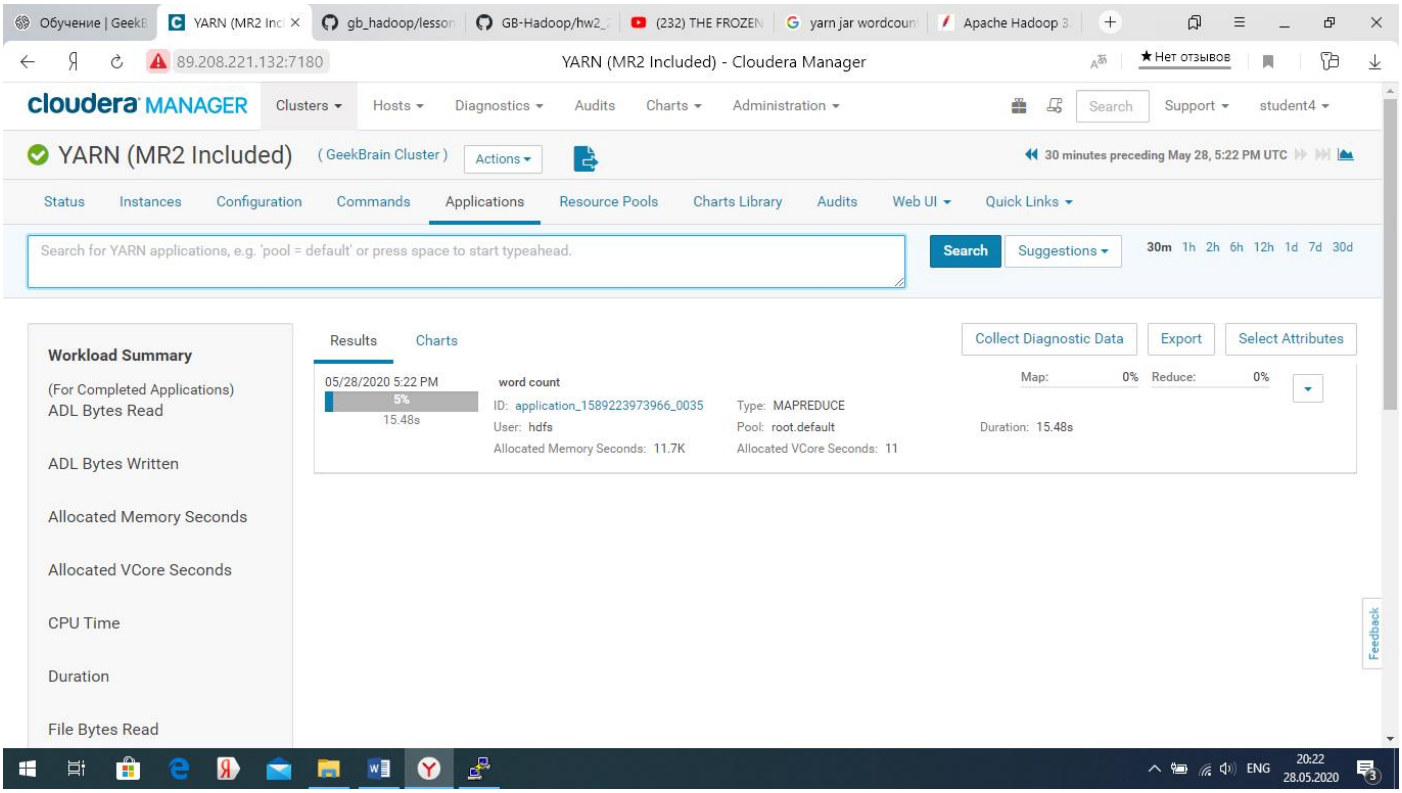


Работа в Cloudera

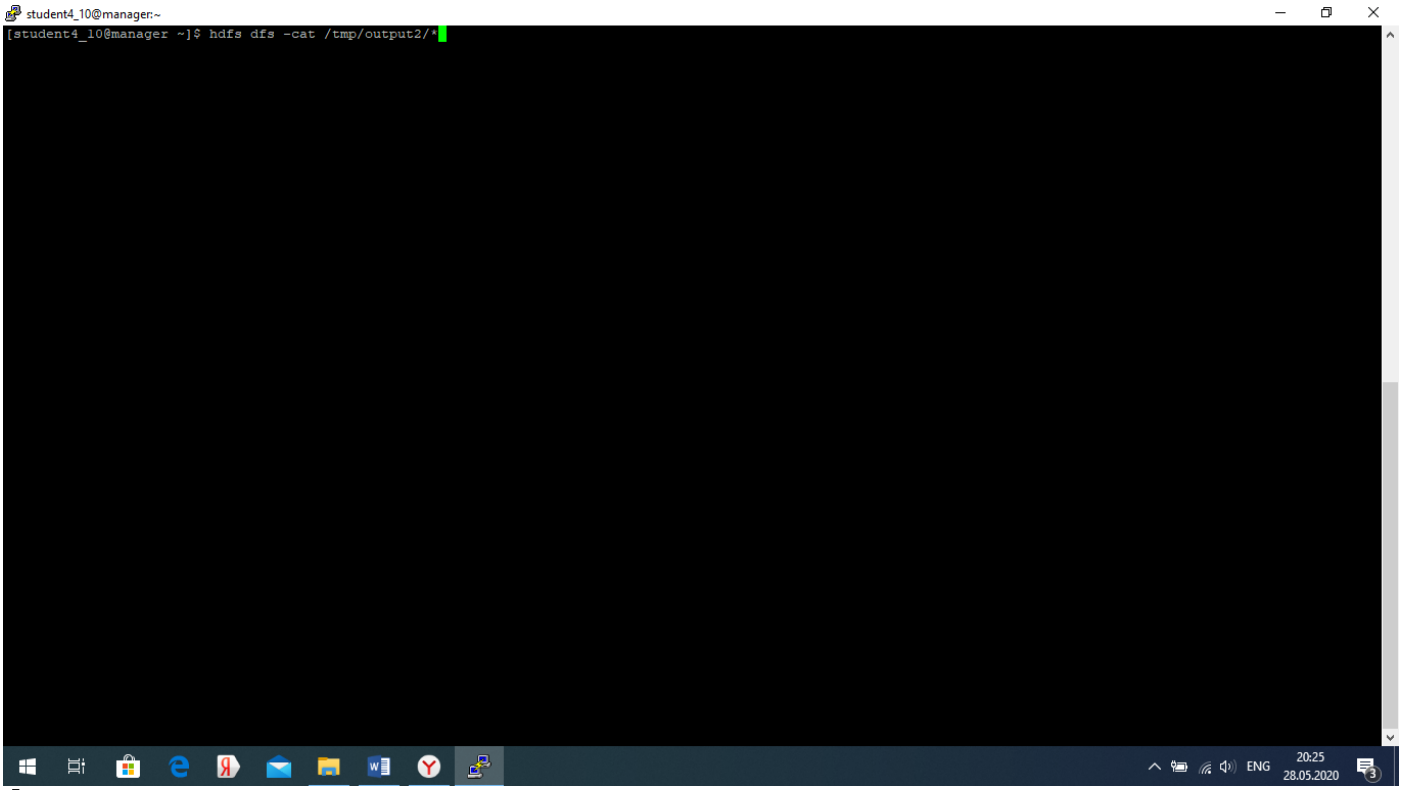
Второй пример wordcount, запускаем

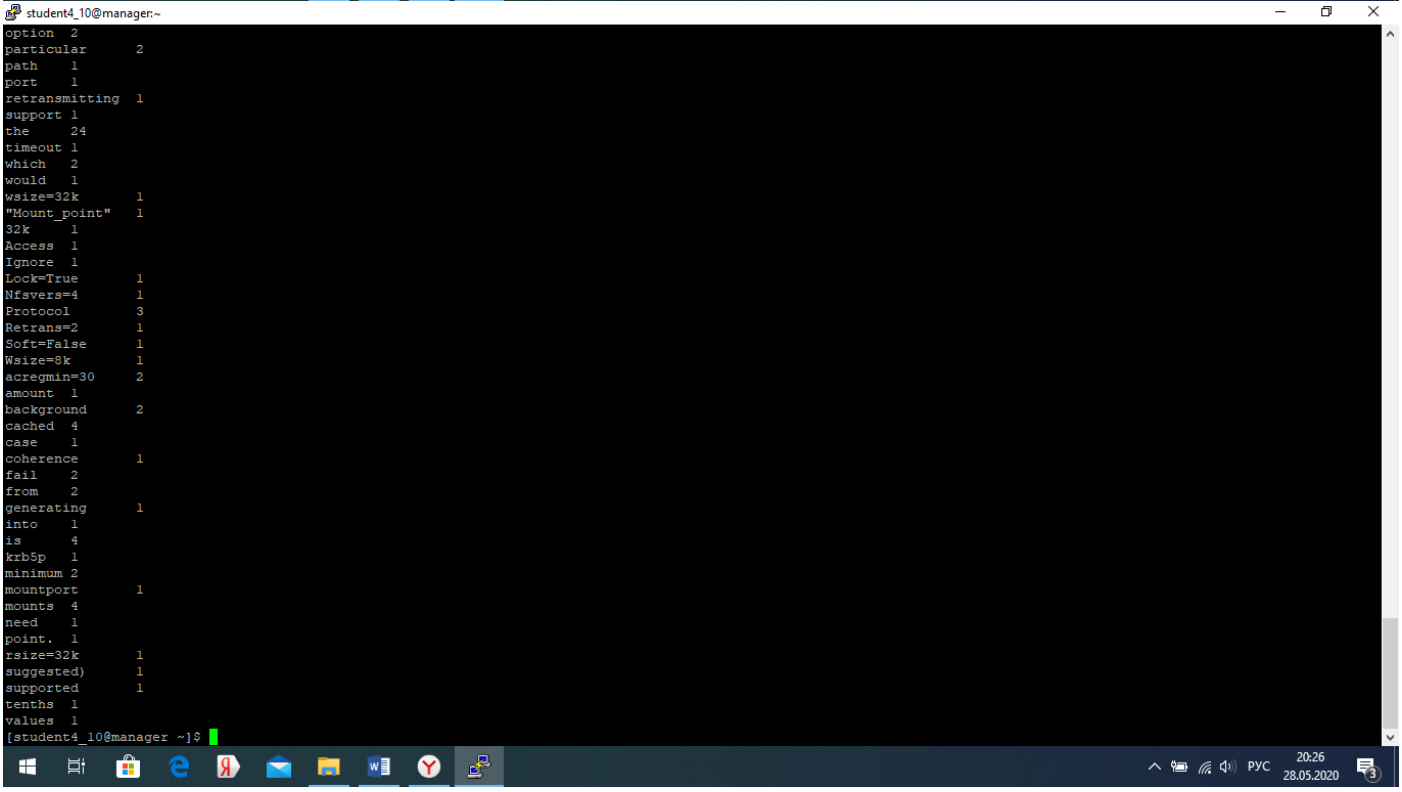
```
student4_10@manager:~
[student4_10@manager ~]$ yarn jar $YARN_EXAMPLES/hadoop-mapreduce-examples.jar wor
dcount /acldir/etc/nfsmount.conf /tmp/output2
20/05/28 17:13:44 INFO client.RMProxy: Connecting to ResourceManager at manager.no
valocal/89.208.221.132:8032
20/05/28 17:13:45 INFO input.FileInputFormat: Total input paths to process : 1
20/05/28 17:13:45 INFO mapreduce.JobSubmitter: number of splits:1
20/05/28 17:13:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1589
223973966_0034
20/05/28 17:13:45 INFO impl.YarnClientImpl: Submitted application application_1589
223973966_0034
20/05/28 17:13:45 INFO mapreduce.Job: The url to track the job: http://manager.nov
alocal:8088/proxy/application_1589223973966_0034/
20/05/28 17:13:45 INFO mapreduce.Job: Running job: job_1589223973966_0034
20/05/28 17:13:57 INFO mapreduce.Job: Job job_1589223973966_0034 running in uber m
ode : false
20/05/28 17:13:57 INFO mapreduce.Job:  map 0% reduce 0%
20/05/28 17:14:05 INFO mapreduce.Job:  map 100% reduce 0%
```

```
student4_10@manager:~
                Launched map tasks=1
                Launched reduce tasks=6
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4111
                Total time spent by all reduces in occupied slots (ms)=28702
                Total time spent by all map tasks (ms)=4111
                Total time spent by all reduce tasks (ms)=28702
                Total vcore-milliseconds taken by all map tasks=4111
                Total vcore-milliseconds taken by all reduce tasks=28702
                Total megabyte-milliseconds taken by all map tasks=4209664
                Total megabyte-milliseconds taken by all reduce tasks=29390848
        Map-Reduce Framework
                Map input records=135
                Map output records=593
                Map output bytes=5723
                Map output materialized bytes=2830
                Input split bytes=119
                Combine input records=593
                Combine output records=227
                Reduce input groups=227
                Reduce shuffle bytes=2830
                Reduce input records=227
                Reduce output records=227
                Spilled Records=454
                Shuffled Maps =6
                Failed Shuffles=0
                Merged Map outputs=6
                GC time elapsed (ms)=836
                CPU time spent (ms)=10100
                Physical memory (bytes) snapshot=1818824704
                Virtual memory (bytes) snapshot=19634774016
                Total committed heap usage (bytes)=1711800320
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=3390
        File Output Format Counters
                Bytes Written=2277
[student4_10@manager ~]$
```



Посмотрим что внутри файла



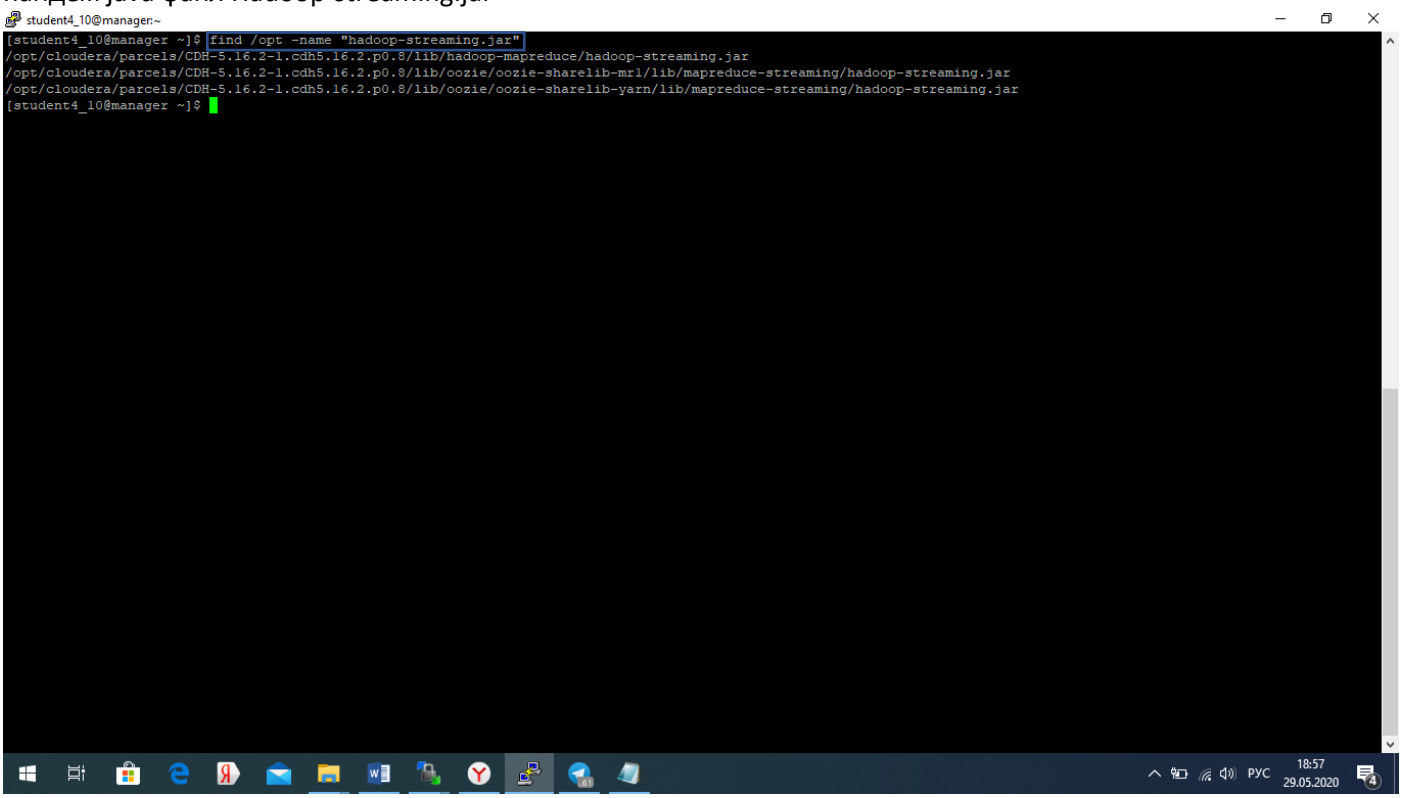


3. [Факультативное, для тех кто знает JAVA] Собрать программу для MR на Java и запустить ее. Wordcount будет вполне достаточен.

С java не знаком, по этому не получилось.

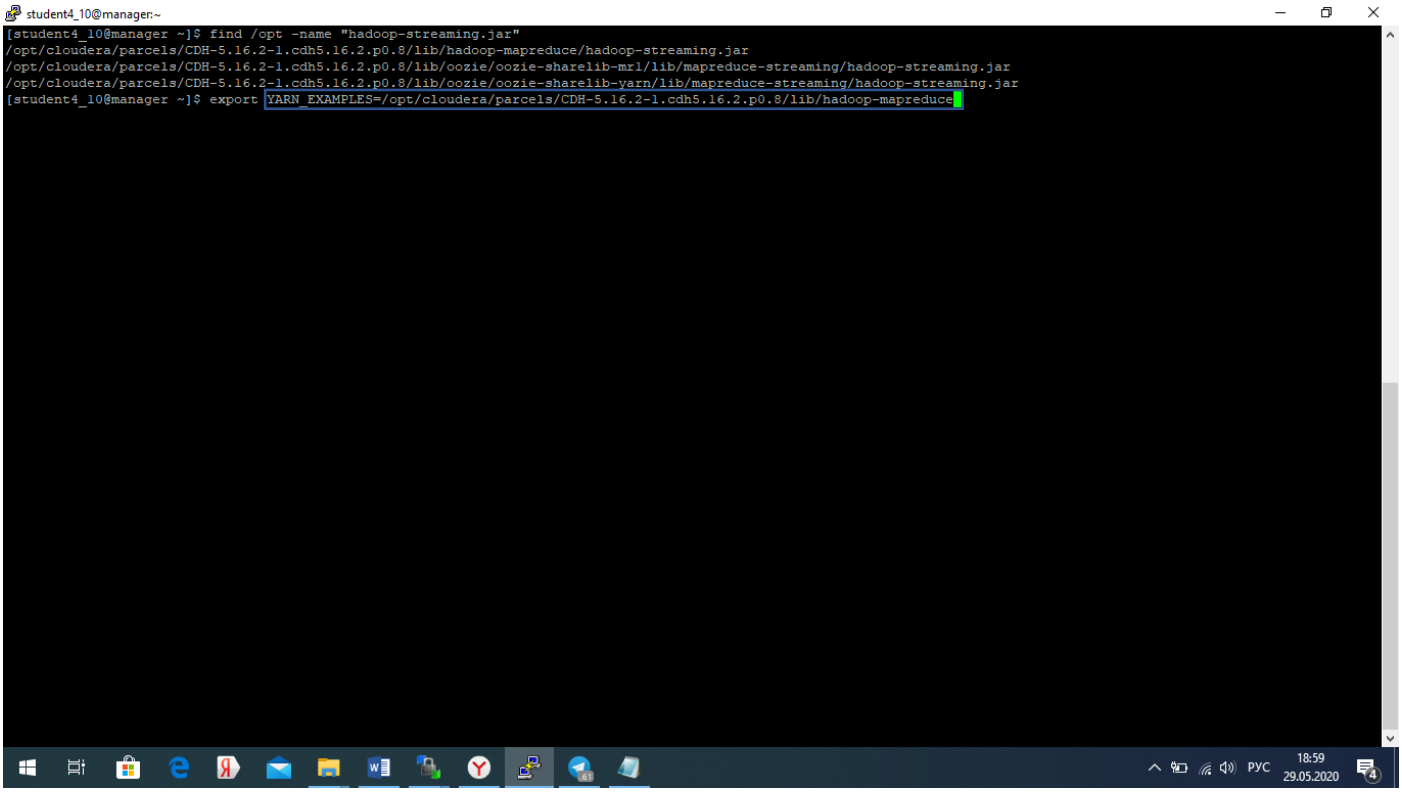
5++

найдем java файл Hadoop-streaming.jar



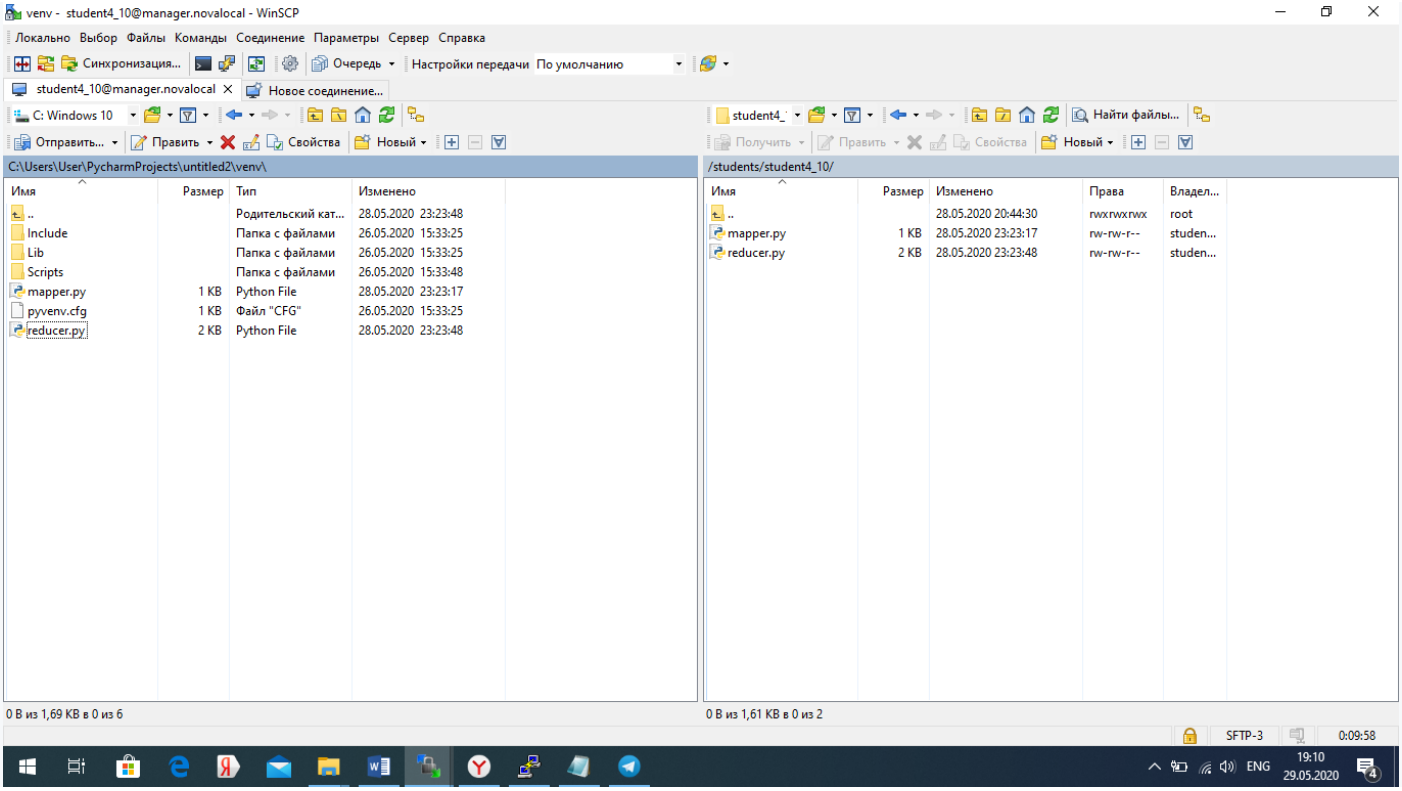
создаём переменную

0

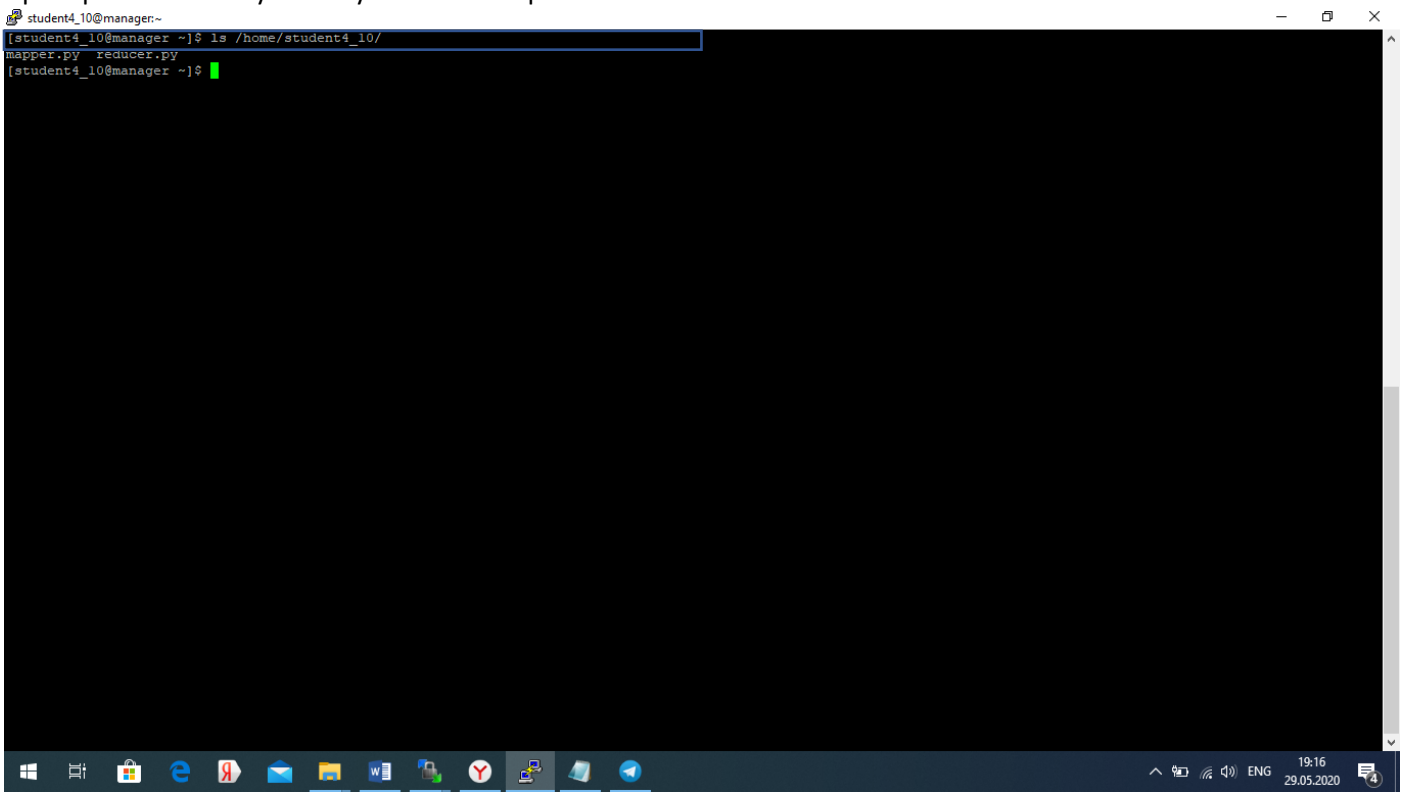


export YARN_EXAMPLES=/opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/lib/hadoop-mapreduce

Переносим с П.К. файлы mapper.py и reducer.py в students/student4_10

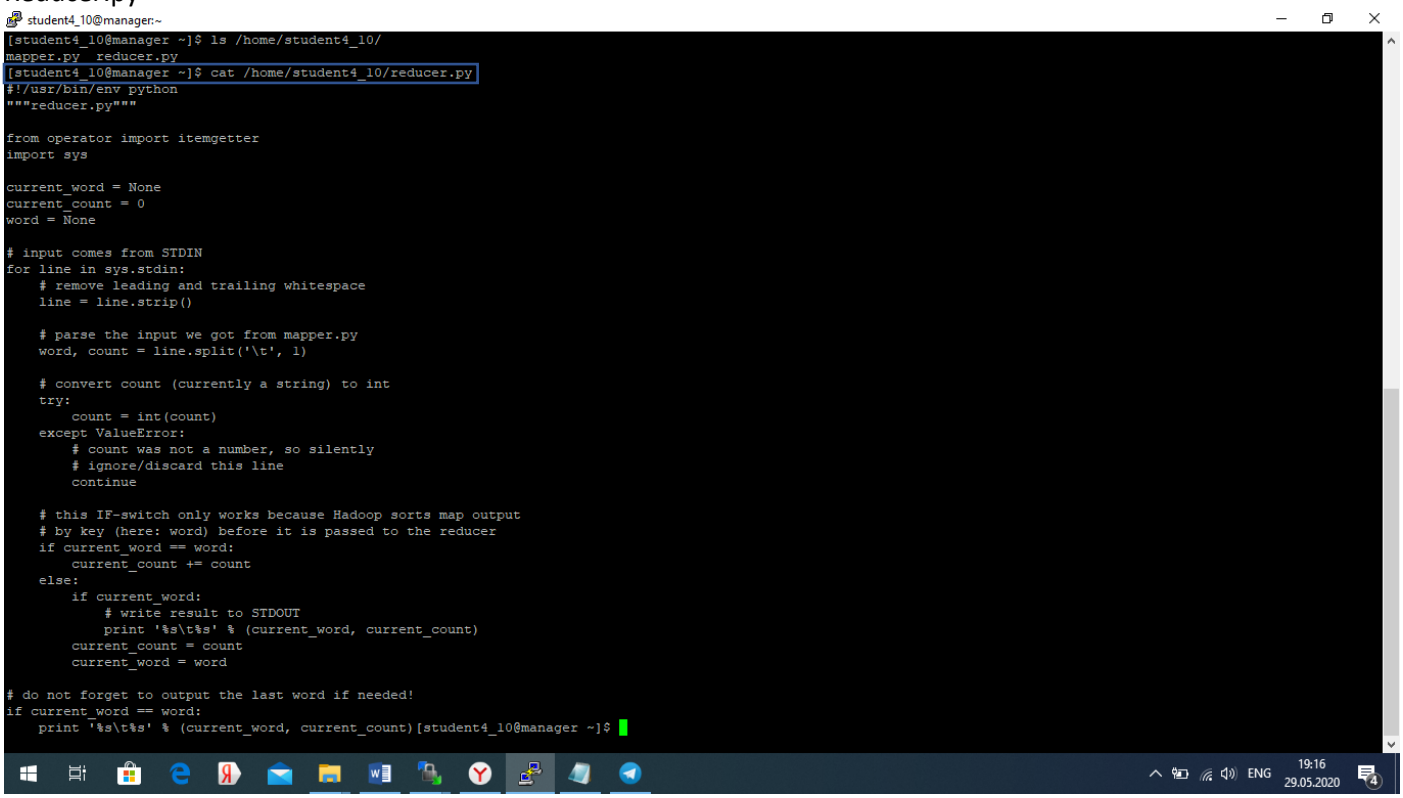


Проверяем локальную папку на наличие файлов

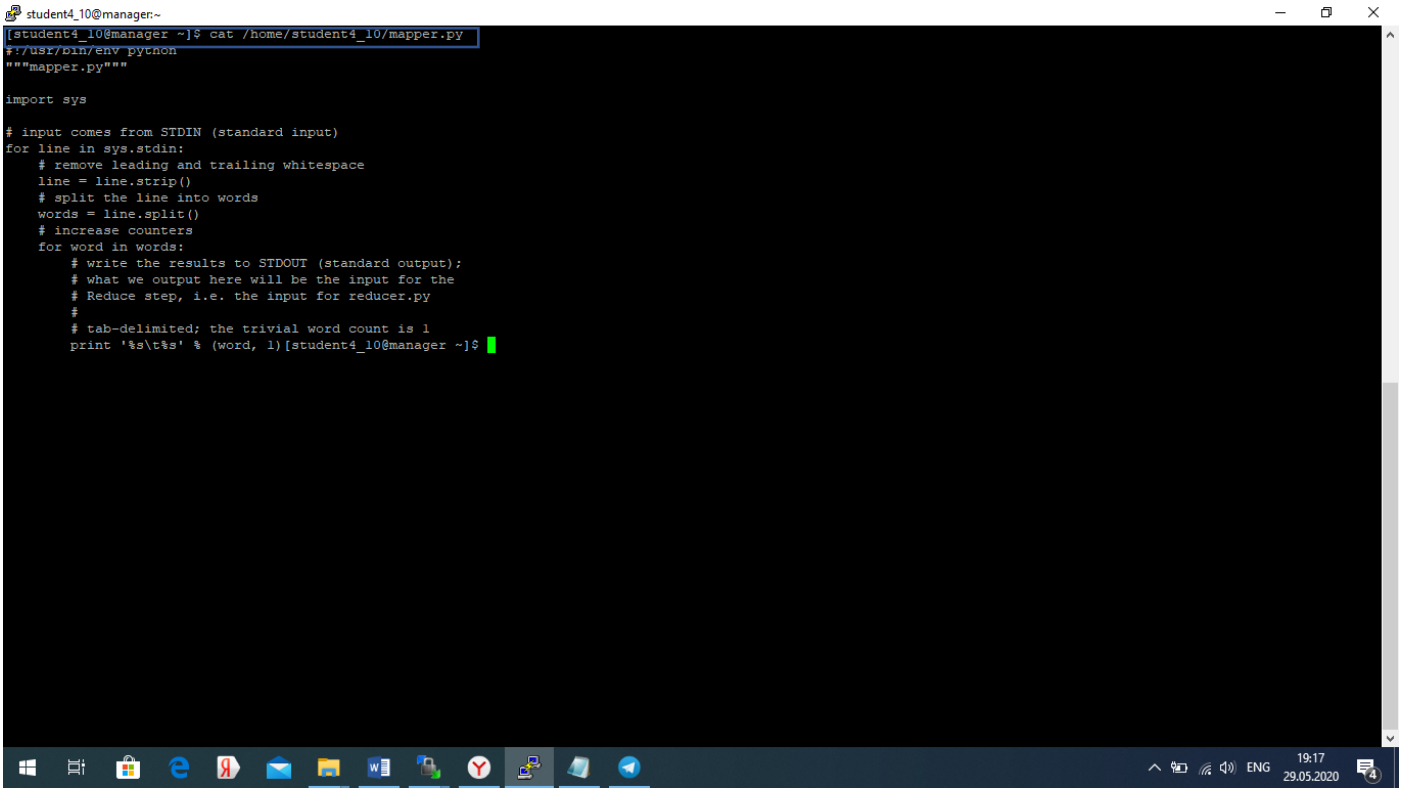


Проверяем внутренность файлов

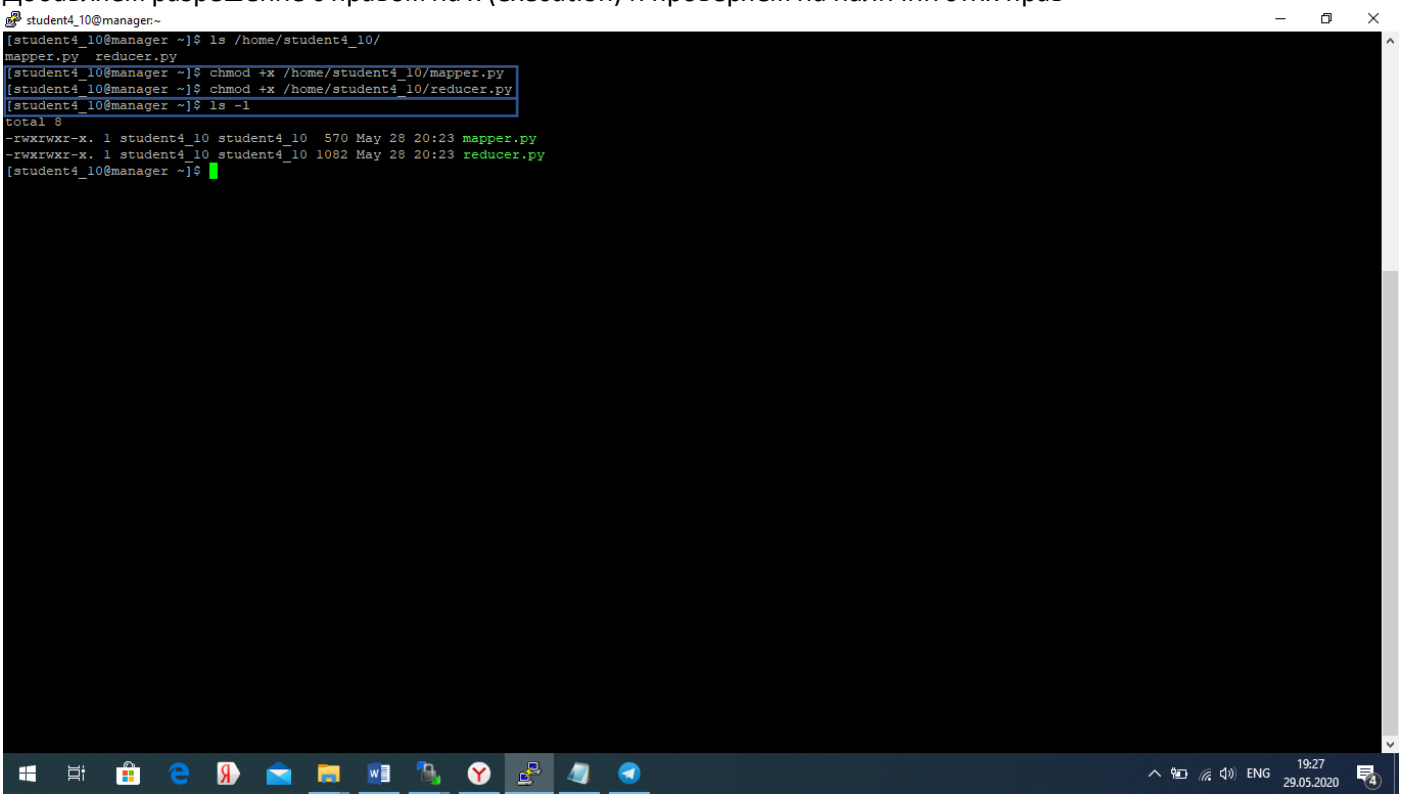
Reducer.py



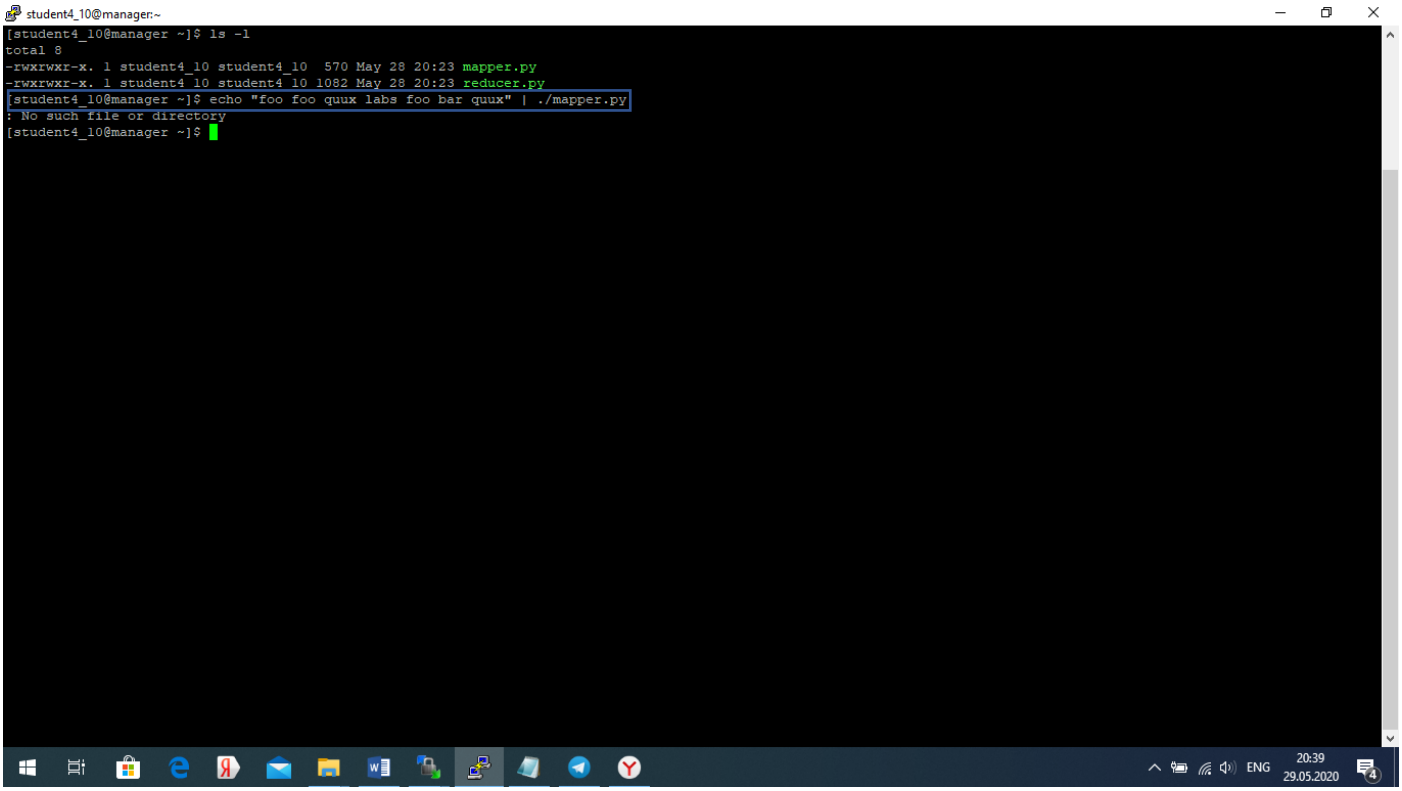
Mapper.py



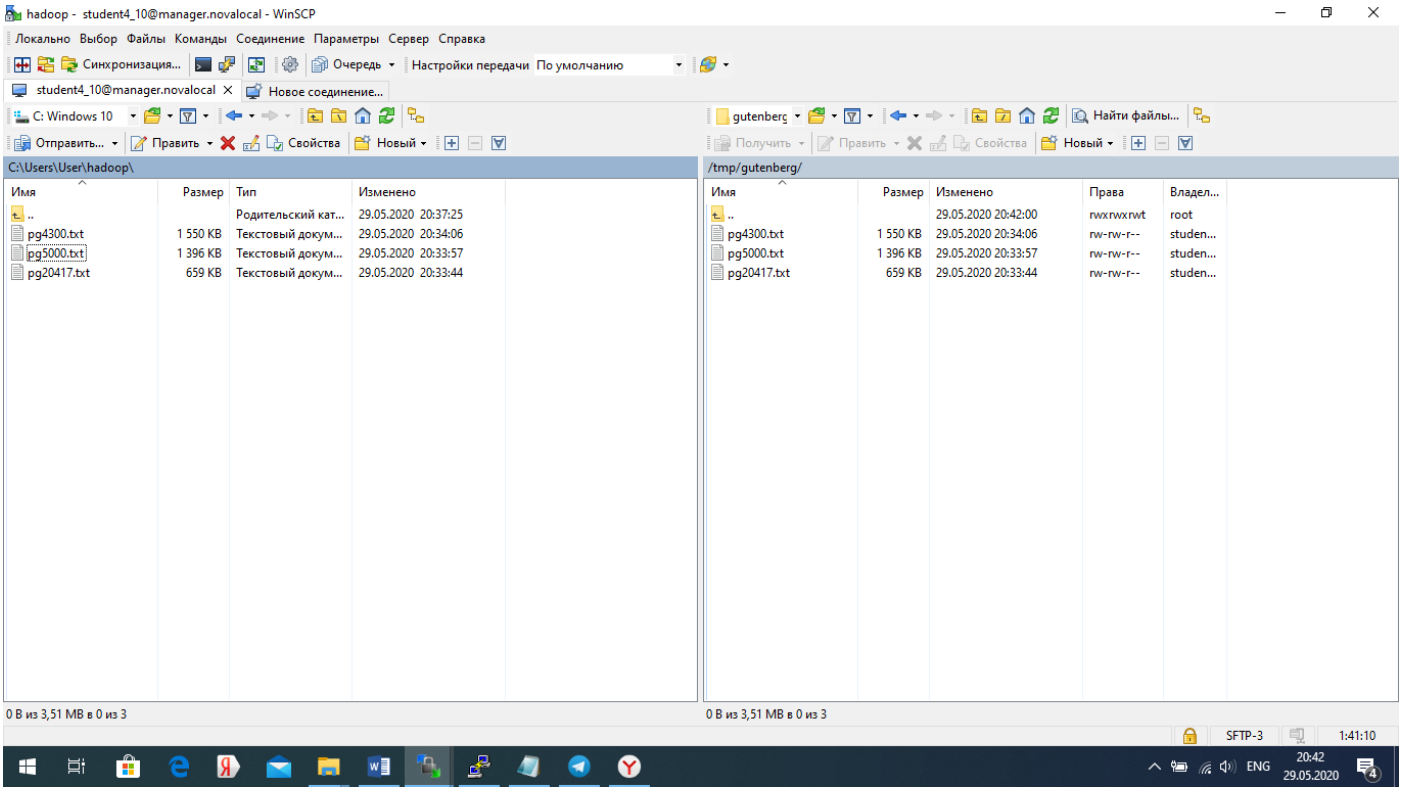
Add permissions x (execution) and check permissions
Добавляем разрешение с правом на x (execution) и проверяем на наличии этих прав



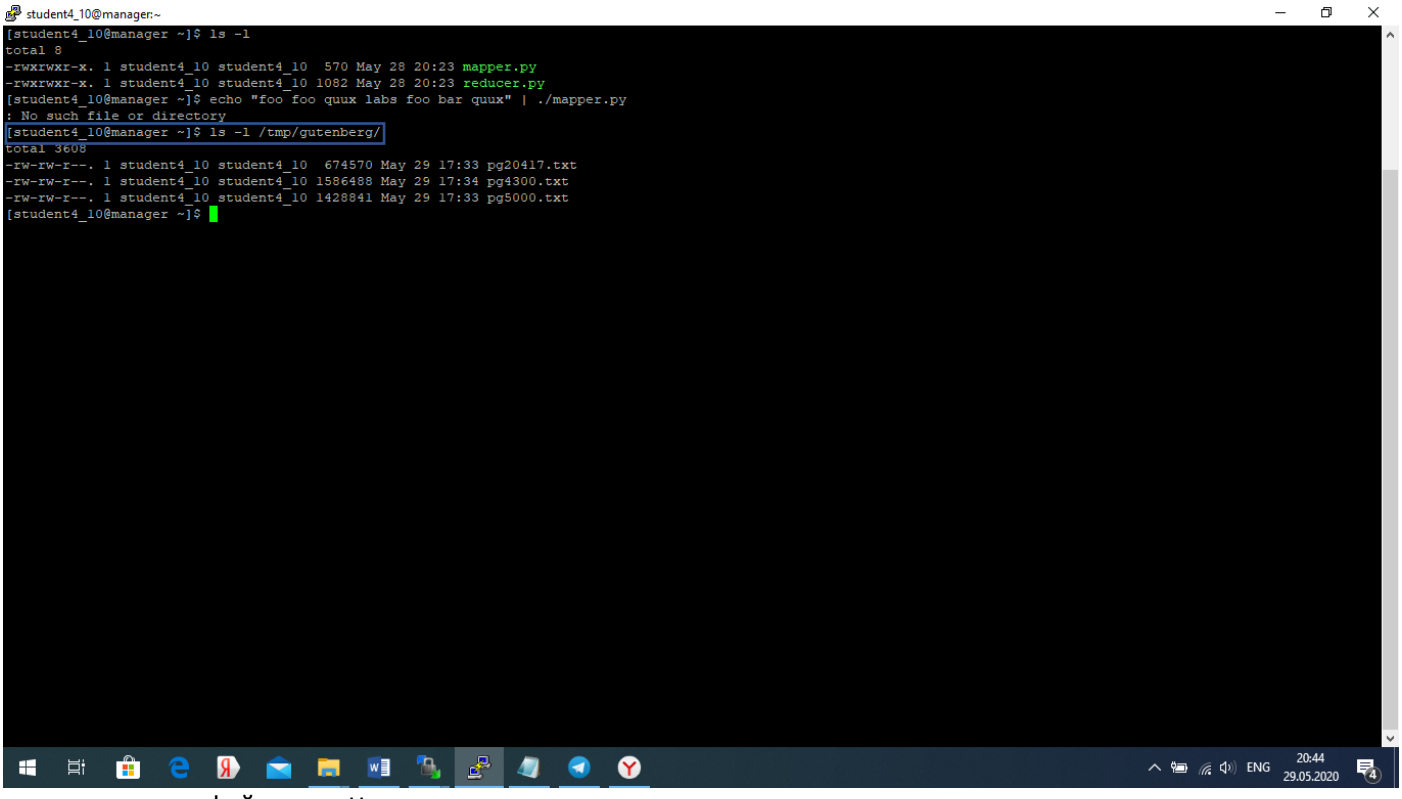
Вопрос
Почему мы надо запускать mapper и reducer с LOCAL а не с HDFS как например текстовые файлы e-books



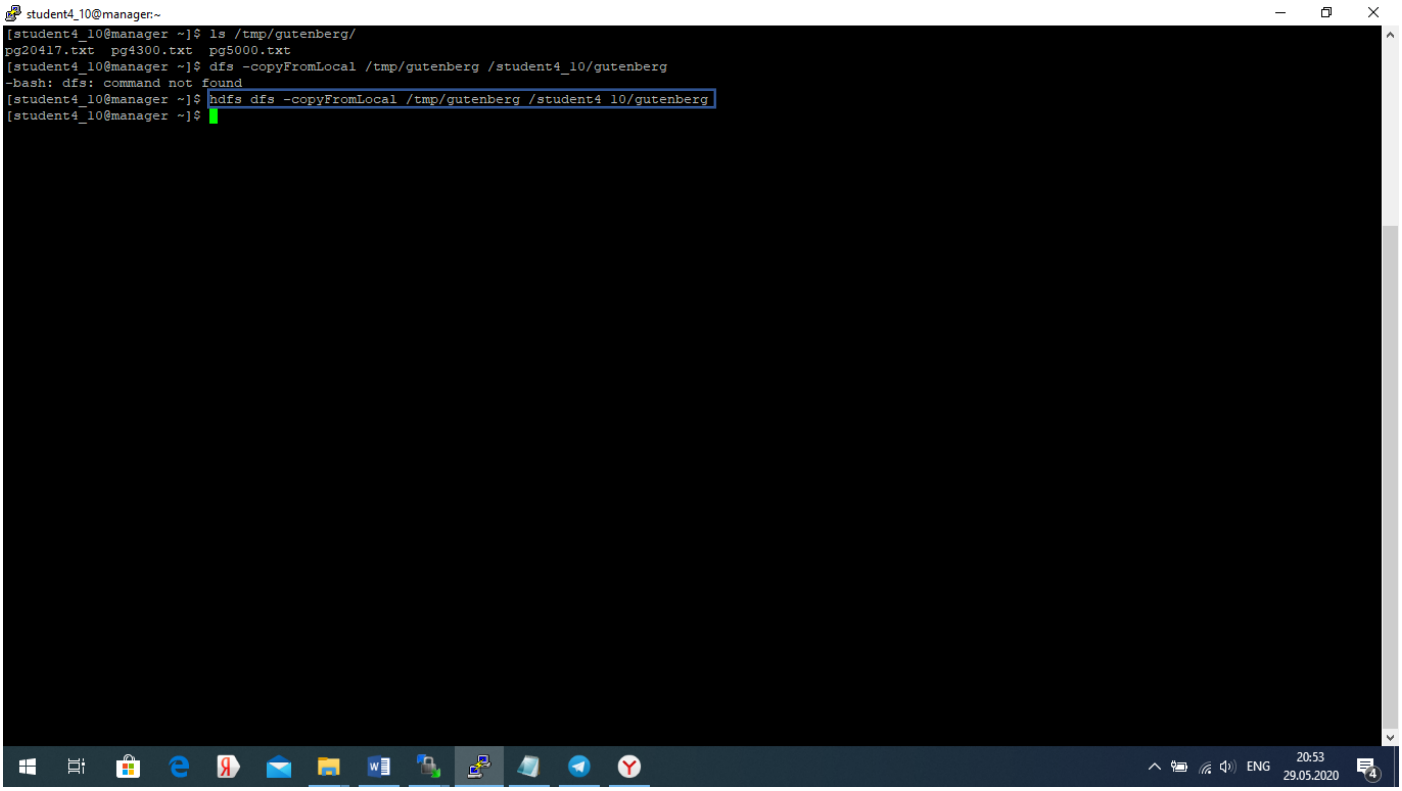
Вопрос
Не работает ECHO ?



скачал e-books files и переносим эти text files в tmp/Gutenberg



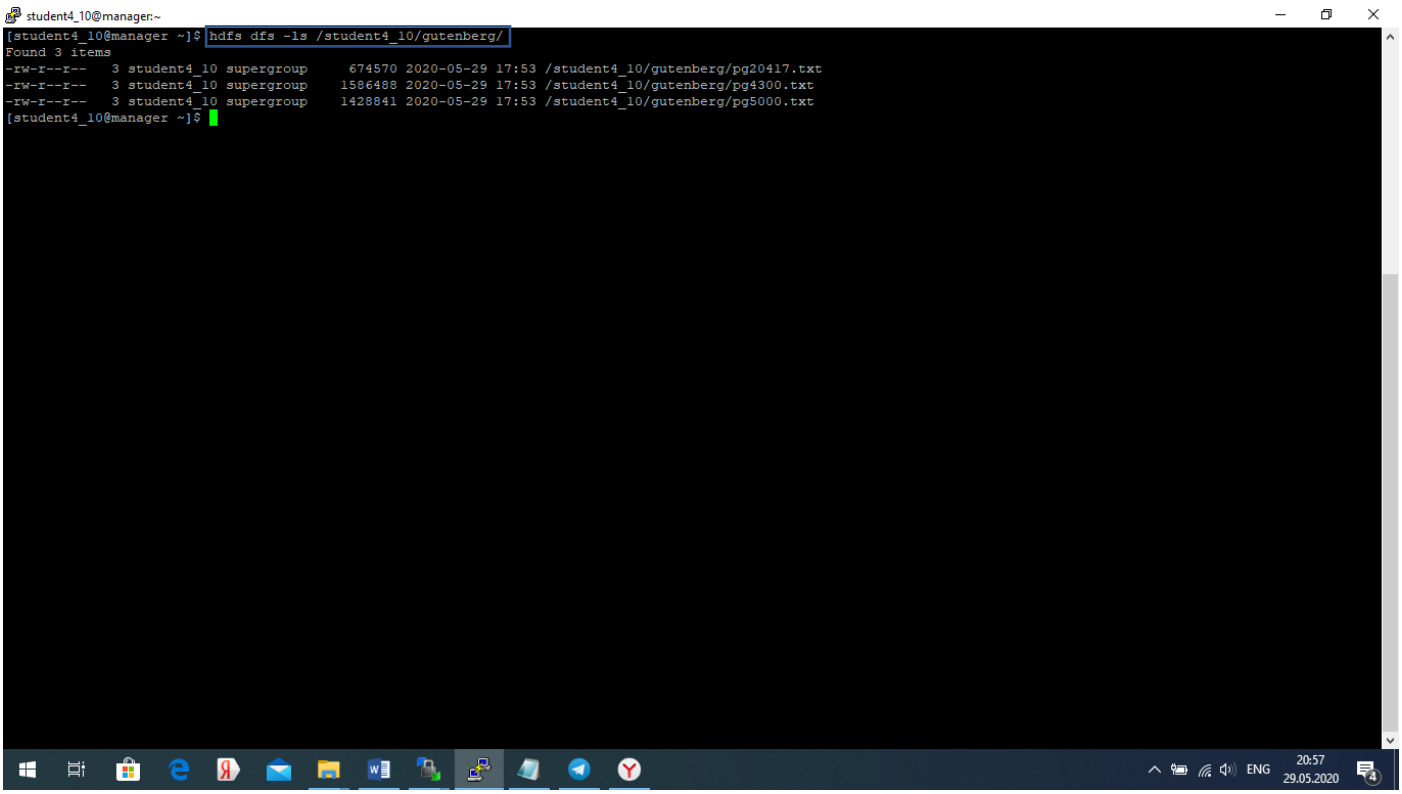
проверяем текс файлы в putty



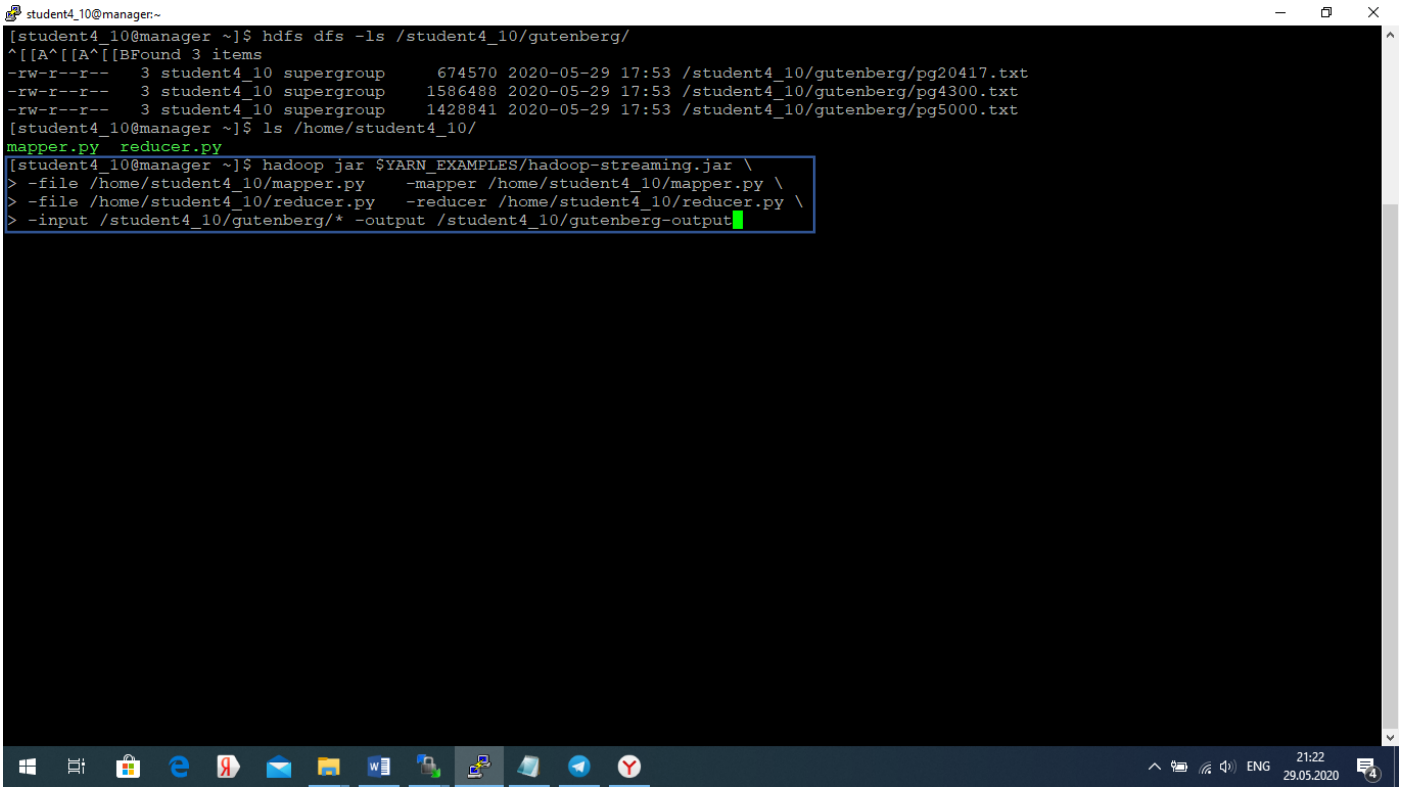
CopyFromLocal Text files into HDFS

Question

- Почему в примере Американцы не пишут HDFS
- Почему mapper.py и reducer.py мы оставили в локальной папке а вот e-books files перенесли в HDFS, почему не перенести все в HDFS или же оставить на локальном уровне.



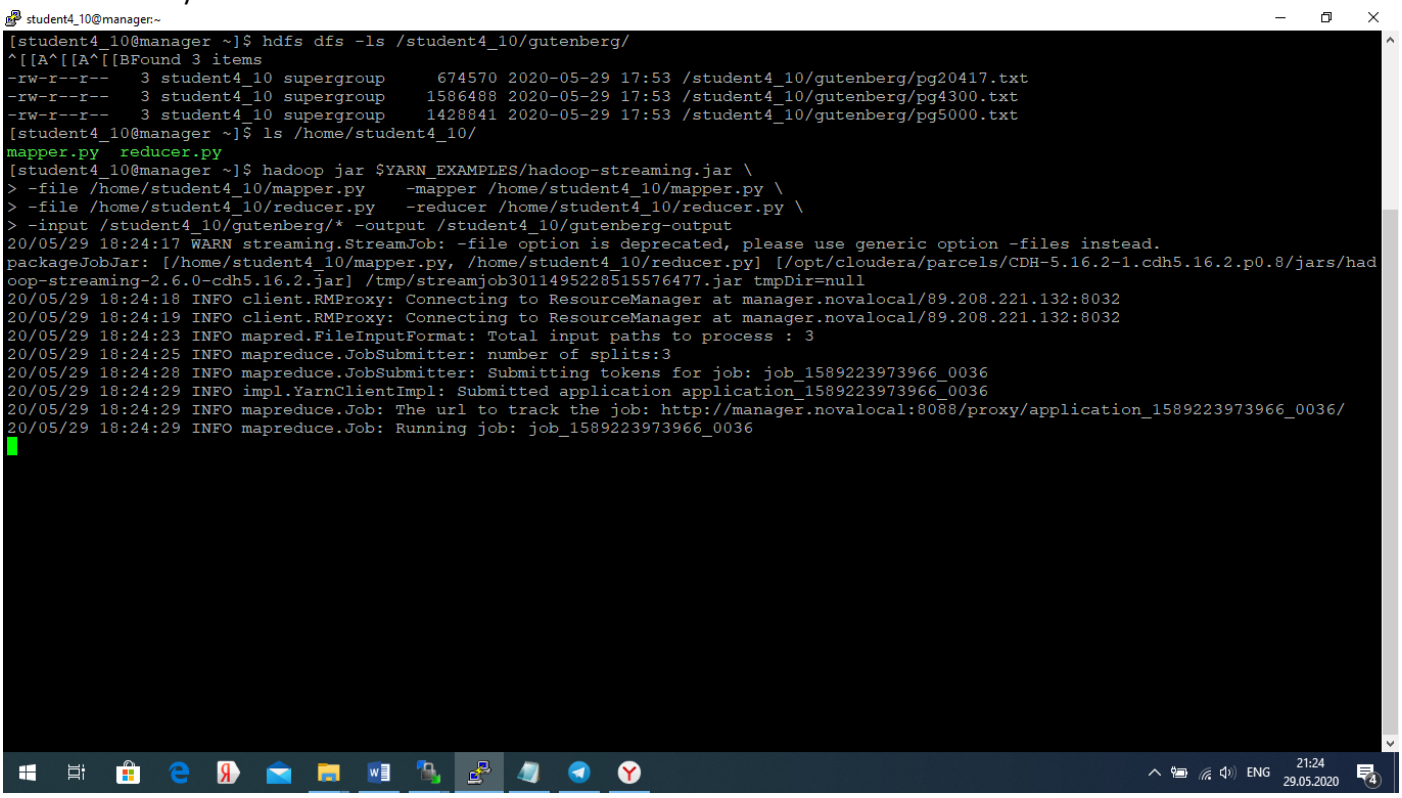
проверяем что все есть в HDFS

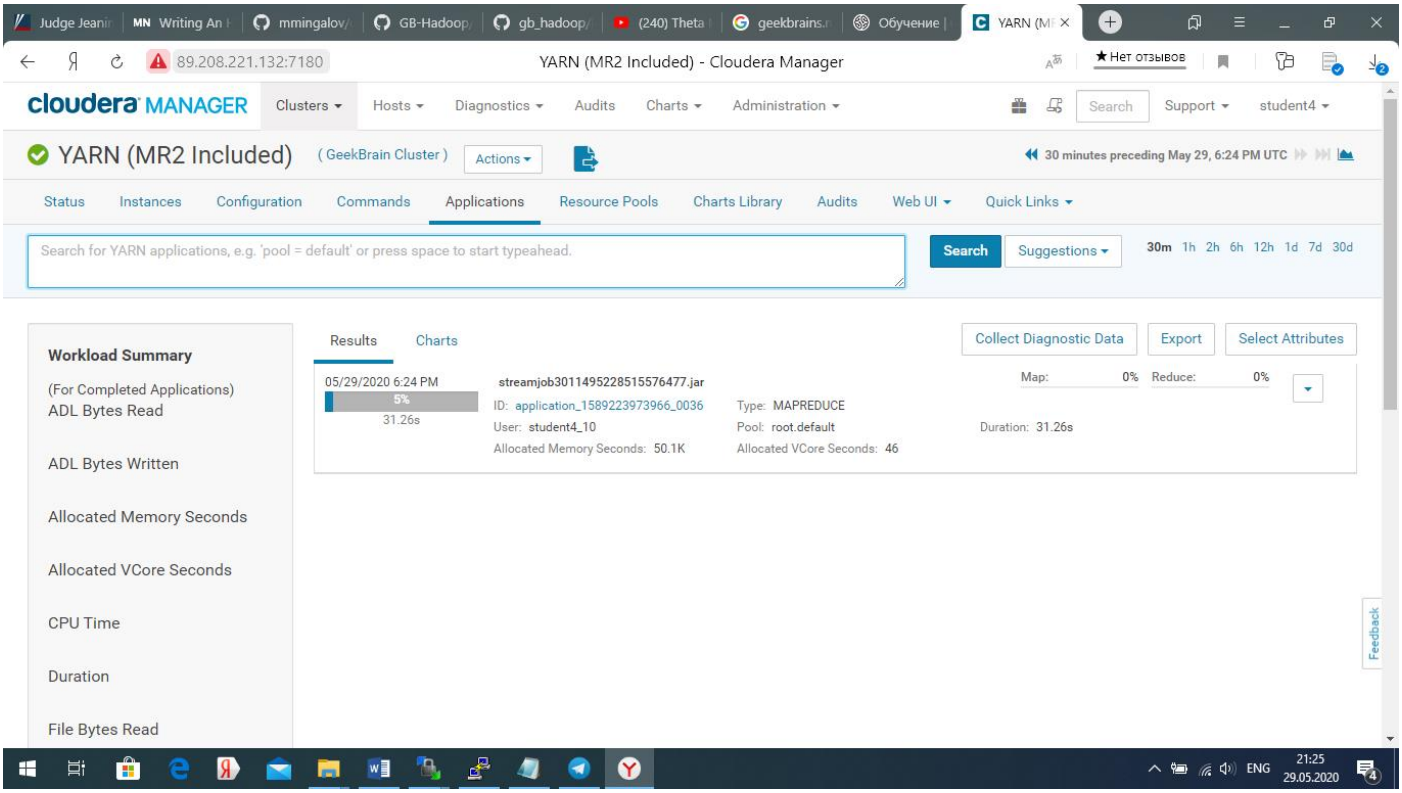


Запускаем jar файл

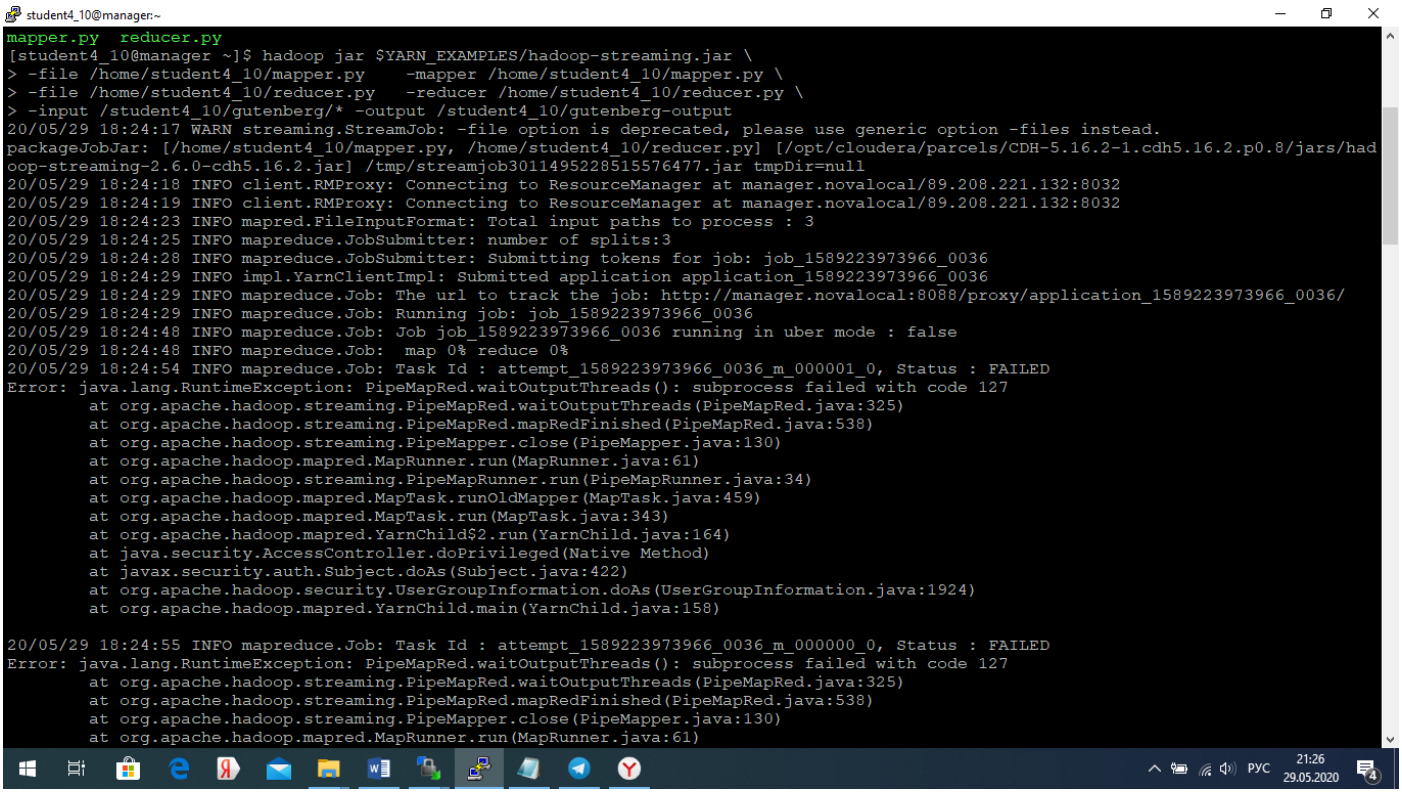
Question

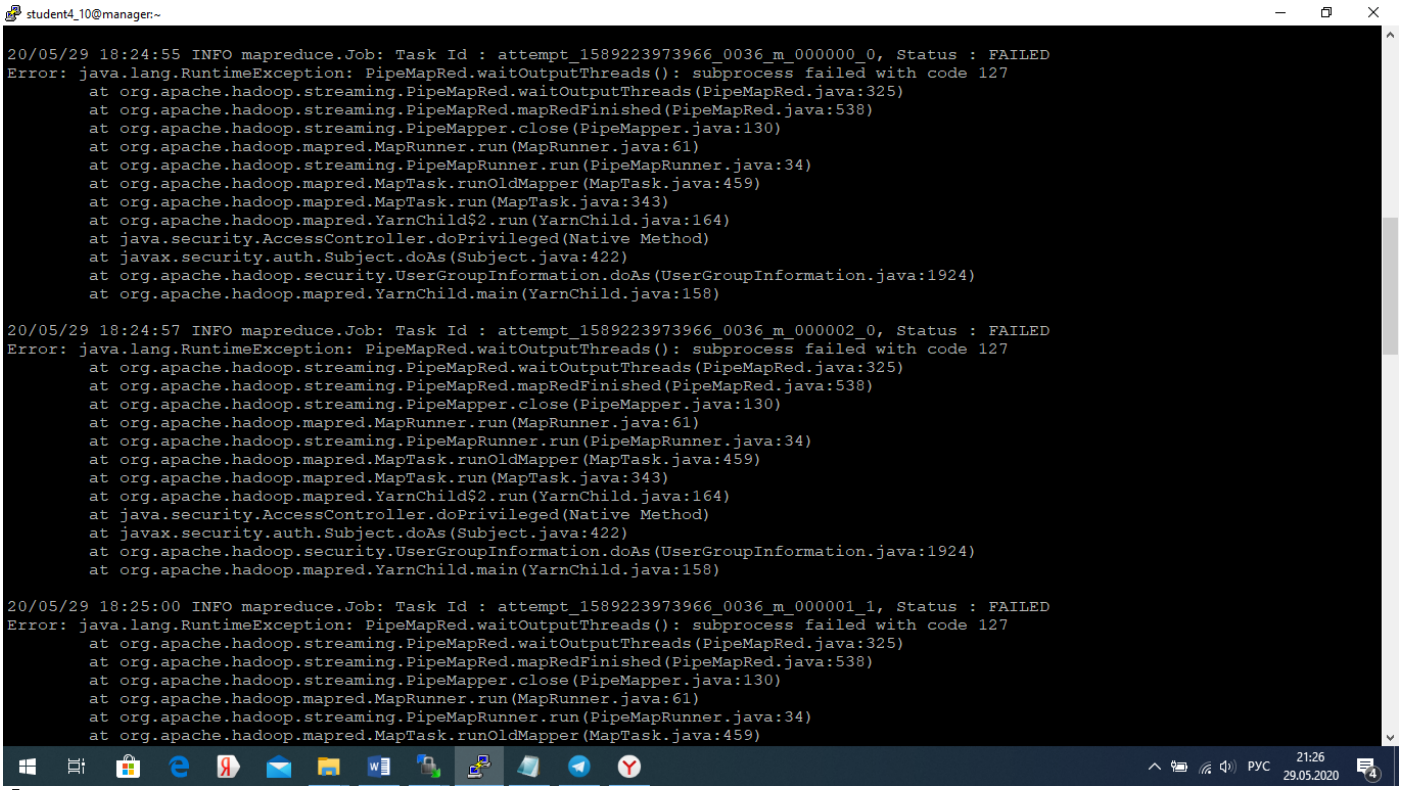
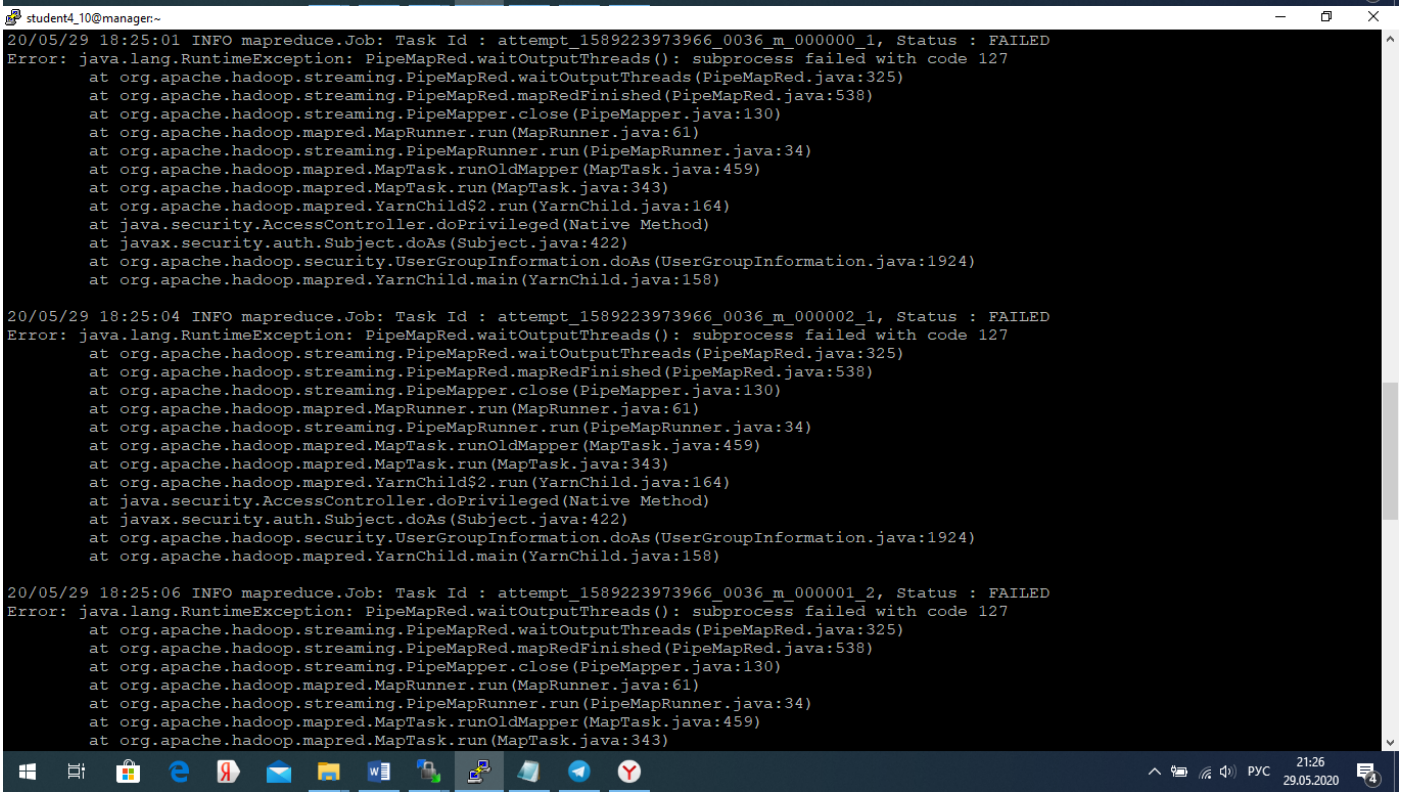
- Почему ошибка

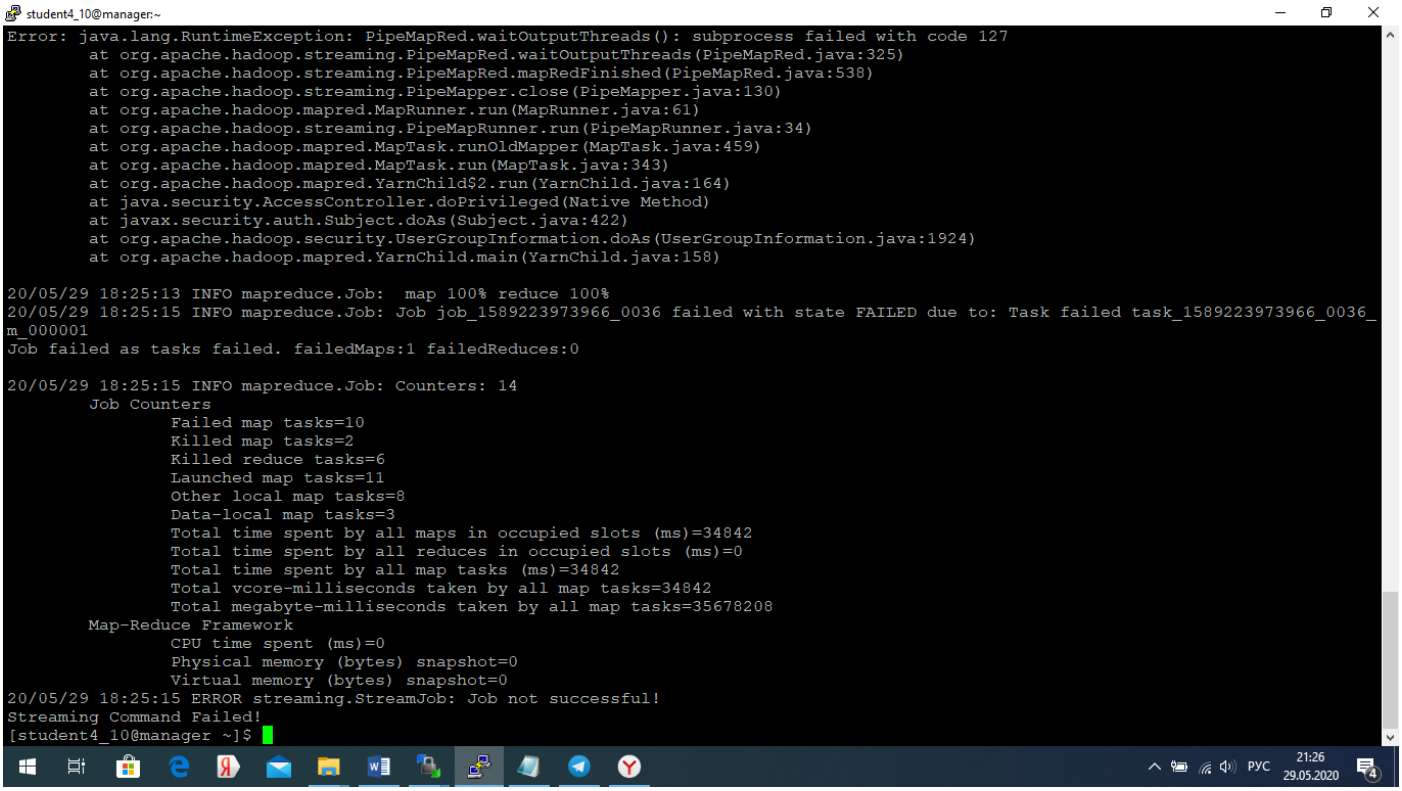




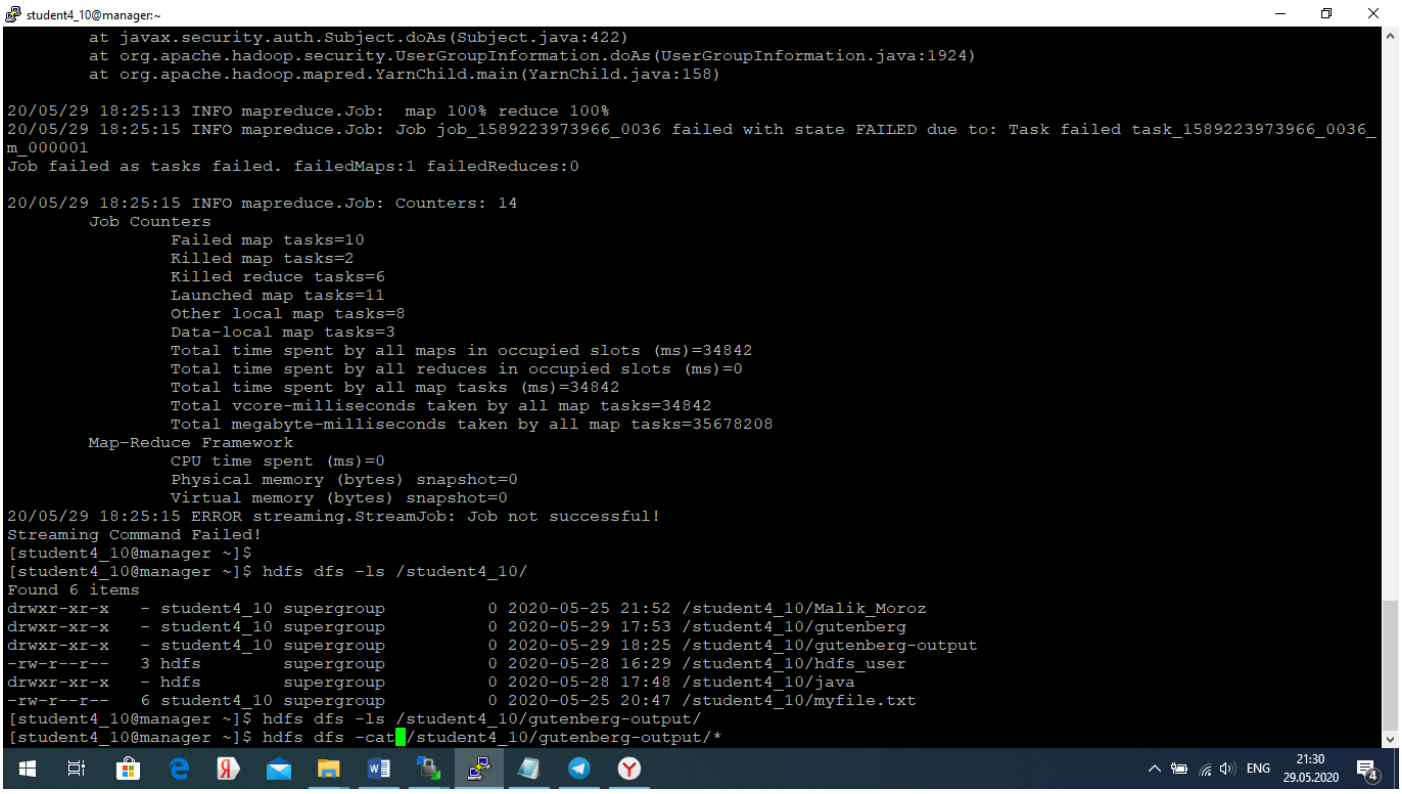
пишет что была ошибка



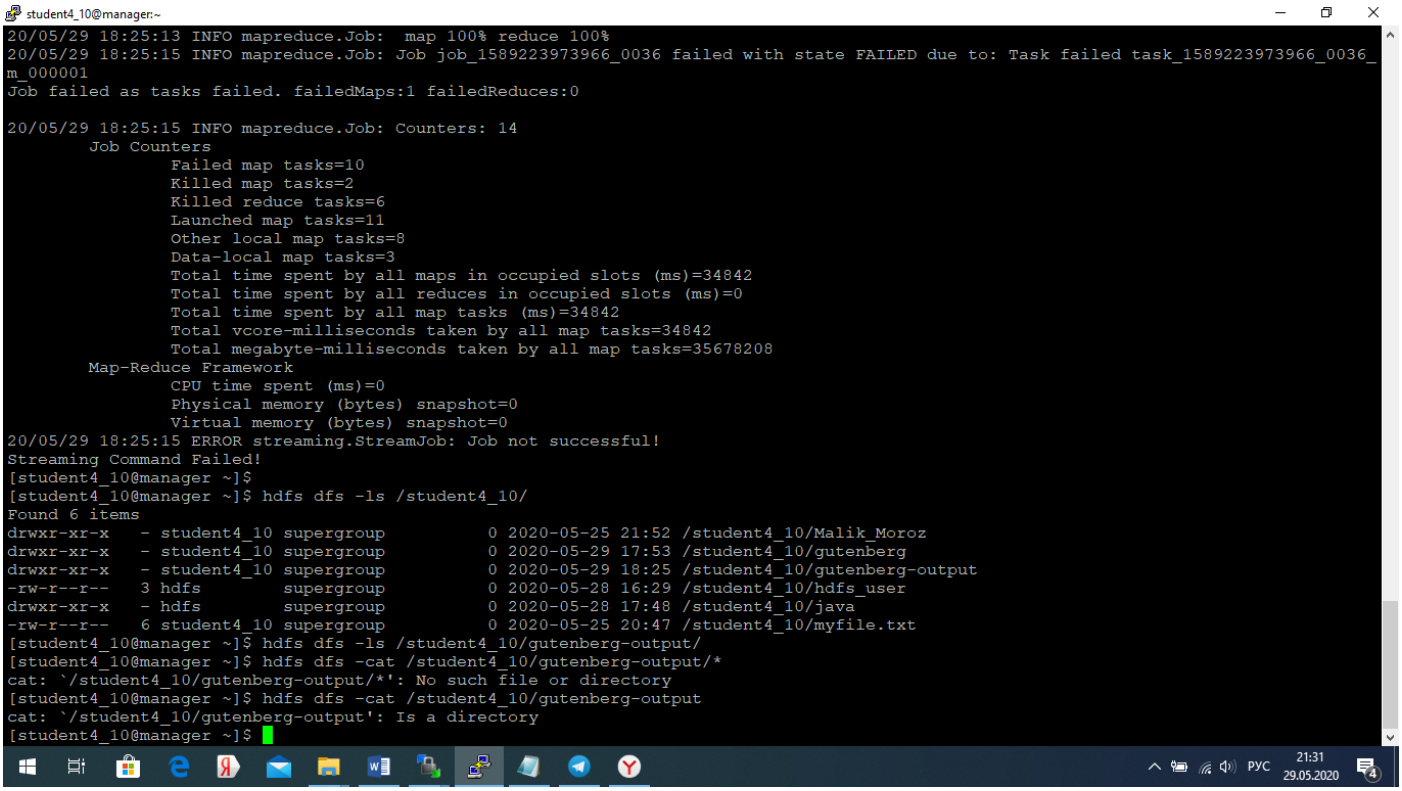

student4_10@manager:~
20/05/29 18:24:55 INFO mapreduce.Job: Task Id : attempt_1589223973966_0036_m_000000_0, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127
at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)
at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)
at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:459)
at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:164)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1924)
at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)
20/05/29 18:24:57 INFO mapreduce.Job: Task Id : attempt_1589223973966_0036_m_000002_0, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127
at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)
at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)
at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:459)
at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:164)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1924)
at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)
20/05/29 18:25:00 INFO mapreduce.Job: Task Id : attempt_1589223973966_0036_m_000001_1, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127
at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)
at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)
at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:459)
РУС
21:26
29.05.2020

student4_10@manager:~
20/05/29 18:25:01 INFO mapreduce.Job: Task Id : attempt_1589223973966_0036_m_000000_1, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127
at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)
at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)
at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:459)
at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:164)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1924)
at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)
20/05/29 18:25:04 INFO mapreduce.Job: Task Id : attempt_1589223973966_0036_m_000002_1, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127
at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)
at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)
at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:459)
at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:164)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1924)
at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)
20/05/29 18:25:06 INFO mapreduce.Job: Task Id : attempt_1589223973966_0036_m_000001_2, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127
at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)
at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)
at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:459)
at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
РУС
21:26
29.05.2020



Ничего не получилось ошибка !!!



проверяем что получили


student4_10@manager:~
20/05/29 18:25:13 INFO mapreduce.Job: map 100% reduce 100%
20/05/29 18:25:15 INFO mapreduce.Job: Job job_1589223973966_0036 failed with state FAILED due to: Task failed task_1589223973966_0036_m_000001
Job failed as tasks failed. failedMaps:1 failedReduces:0
20/05/29 18:25:15 INFO mapreduce.Job: Counters: 14
Job Counters
Failed map tasks=10
Killed map tasks=2
Killed reduce tasks=6
Launched map tasks=11
Other local map tasks=8
Data-local map tasks=3
Total time spent by all maps in occupied slots (ms)=34842
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=34842
Total vcore-milliseconds taken by all map tasks=34842
Total megabyte-milliseconds taken by all map tasks=35678208
Map-Reduce Framework
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
20/05/29 18:25:15 ERROR streaming.StreamJob: Job not successful!
Streaming Command Failed!
[student4_10@manager ~]$
[student4_10@manager ~]$ hdfs dfs -ls /student4_10/
Found 6 items
drwxr-xr-x - student4_10 supergroup 0 2020-05-25 21:52 /student4_10/Malik_Moroz
drwxr-xr-x - student4_10 supergroup 0 2020-05-29 17:53 /student4_10/gutenberg
drwxr-xr-x - student4_10 supergroup 0 2020-05-29 18:25 /student4_10/gutenberg-output
-rw-r--r-- 3 hdfs supergroup 0 2020-05-28 16:29 /student4_10/hdfs_user
drwxr-xr-x - hdfs supergroup 0 2020-05-28 17:48 /student4_10/java
-rw-r--r-- 6 student4_10 supergroup 0 2020-05-25 20:47 /student4_10/myfile.txt
[student4_10@manager ~]$ hdfs dfs -ls /student4_10/gutenberg-output/
[student4_10@manager ~]$ hdfs dfs -cat /student4_10/gutenberg-output/*
cat: `/student4_10/gutenberg-output/*': No such file or directory
[student4_10@manager ~]$ hdfs dfs -cat /student4_10/gutenberg-output
cat: `/student4_10/gutenberg-output': Is a directory
[student4_10@manager ~]$
РУС
21:31
29.05.2020

ничего нету