Домашнее задание
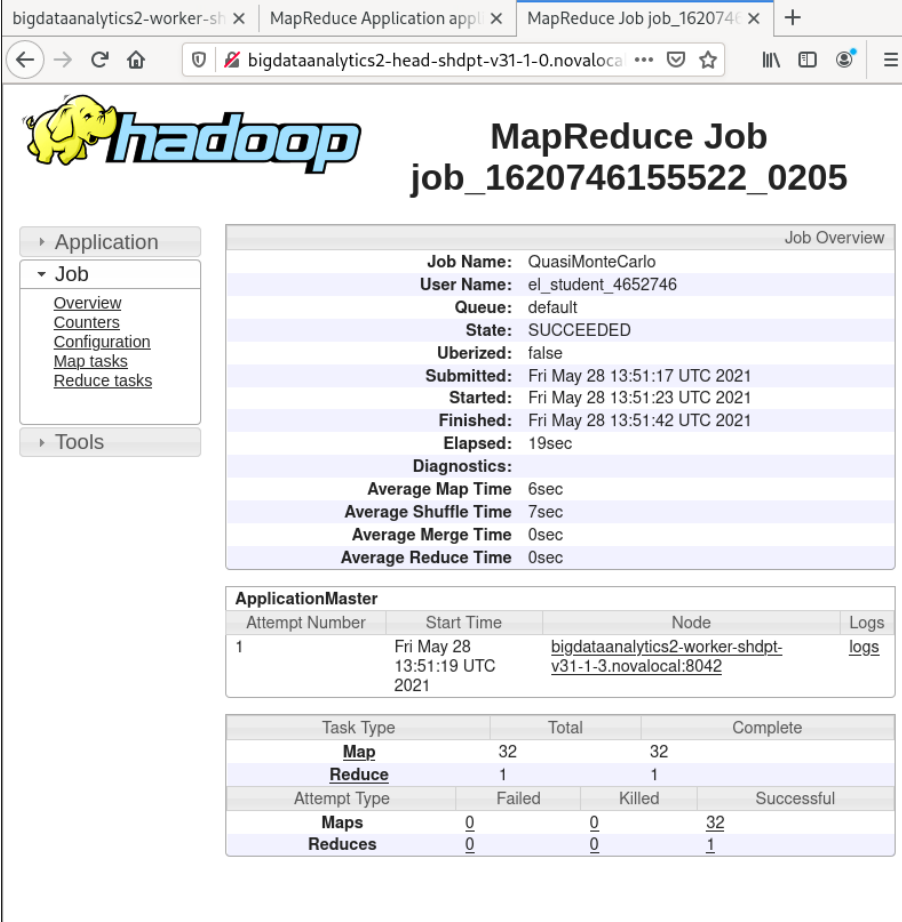
1. Запустить задачу из примеров, например, вычисление pi методом Монте-Карло

```
$ ssh -i .ssh/id_rsa_el_student_4652746 el_student_4652746@37.139.32.56 -D localhost:8080
```

```
$ yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar pi 32 10000
```

```
21/05/28 13:51:18 INFO mapreduce.Job: The url to track the job: http://bigdataanalytics2-head-shdpt-v31-1-0.nova
local:8088/proxy/application_1620746155522_0205/
21/05/28 13:51:18 INFO mapreduce.Job: Running job: job_1620746155522_0205
21/05/28 13:51:25 INFO mapreduce.Job: Job job_1620746155522_0205 running in uber mode : false
21/05/28 13:51:25 INFO mapreduce.Job:  map 0% reduce 0%
21/05/28 13:51:33 INFO mapreduce.Job:  map 22% reduce 0%
21/05/28 13:51:34 INFO mapreduce.Job:  map 25% reduce 0%
21/05/28 13:51:35 INFO mapreduce.Job:  map 38% reduce 0%
21/05/28 13:51:38 INFO mapreduce.Job:  map 63% reduce 0%
21/05/28 13:51:39 INFO mapreduce.Job:  map 75% reduce 0%
21/05/28 13:51:40 INFO mapreduce.Job:  map 78% reduce 0%
21/05/28 13:51:42 INFO mapreduce.Job:  map 88% reduce 0%
21/05/28 13:51:43 INFO mapreduce.Job:  map 100% reduce 0%
21/05/28 13:51:44 INFO mapreduce.Job:  map 100% reduce 100%
21/05/28 13:56:46 INFO mapreduce.Job: Job job_1620746155522_0205 completed successfully
21/05/28 13:56:46 INFO mapreduce.Job: Counters: 54
```

зайти на ResourceManager http://manager.novalocal:8088 и найти свою задачу.

2. Запустить WordCount и доработать скрипт из примера, чтобы удалялись знаки препинания и слова считались в нижнем регистре

./WordCountMR/mapper.py, ./WordCountMR/reducer.py

```
$ yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar -input
./krivosheenkova/test.txt -output ./krivosheenkova/result00 -mapper "python mapper.py" -reducer
"python reducer.py" -file mapper.py  -file reducer.py
```

3. *реализовать алгоритм join на MapReduce

./JoinMapReduce/joinmapper01.py, ./JoinMapReduce/joinreducer01.py

test:

```
$ cat ./country_* | python joinmapper01.py | sort | python joinreducer01.py >> test_result.txt
```

hdfs:

```
$ yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar -D
stream.num.map.output.key.fields=2 -input ./krivosheenkova/JoinMapReduce/country_indexes.csv
-input ./krivosheenkova/JoinMapReduce/country_score.csv -output
./krivosheenkova/JoinMapReduce/hadoop_result00 -mapper "python joinmapper.py" -reducer "python
joinreducer.py" -file joinmapper01.py  -file joinreducer01.py
```