1. Запустить один-два примера
Опробовать запуски map-reduce задач для кластера используя hadoop-mapreduce-examples.jar. Чтобы увидеть полный список нужно выполнить yarn jar $YARN_EXAMPLES/hadoop-mapreduce-examples.jar без параметров. (Там, например, wordcount тоже есть)

```
[student9_11@manager ~]$ YARN_EXAMPLES=/opt/cloudera/parcels/CDH-5.16.2-1.cdh5.1
6.2.p0.8/lib/hadoop-mapreduce/
[student9_11@manager ~]$ echo $YARN_EXAMPLES
/opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/lib/hadoop-mapreduce/
[student9_11@manager ~]$ yarn jar $YARN_EXAMPLES/hadoop-mapreduce-examples.jar
An example program must be given as the first argument.
Valid program names are:
  aggregatewordcount: An Aggregate based map/reduce program that counts the word
s in the input files.
  aggregatewordhist: An Aggregate based map/reduce program that computes the his
togram of the words in the input files.
  bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact di
gits of Pi.
  dbcount: An example job that count the pageview counts from a database.
  distbbp: A map/reduce program that uses a BBP-type formula to compute exact bi
ts of Pi.
  grep: A map/reduce program that counts the matches of a regex in the input.
  join: A job that effects a join over sorted, equally partitioned datasets
  multifilewc: A job that counts words from several files.
  pentomino: A map/reduce tile laying program to find solutions to pentomino pro
blems.
  pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.
  randomtextwriter: A map/reduce program that writes 10GB of random textual data


 of the length of the words in the input files.
[student9_11@manager ~]$ yarn jar $YARN_EXAMPLES/hadoop-mapreduce-examples.jar p
i 32 10000
Number of Maps  = 32
Samples per Map = 10000
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
Wrote input for Map #5
Wrote input for Map #6
Wrote input for Map #7
Wrote input for Map #8
Wrote input for Map #9
Wrote input for Map #10
Wrote input for Map #11
Wrote input for Map #12
Wrote input for Map #13
Wrote input for Map #14
Wrote input for Map #15
Wrote input for Map #16
Wrote input for Map #17
Wrote input for Map #18
```

```
                Failed Shuffles=0
                Merged Map outputs=32
                GC time elapsed (ms)=7745
                CPU time spent (ms)=34430
                Physical memory (bytes) snapshot=14952730624
                Virtual memory (bytes) snapshot=92206514176
                Total committed heap usage (bytes)=14779678720
        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=3776
        File Output Format Counters
                Bytes Written=97
Job Finished in 42.358 seconds
Estimated value of Pi is 3.14147500000000000000
```

2. Изучить интерфейс Resource Manager

Выполнить любую задачу включенную в этот JAR

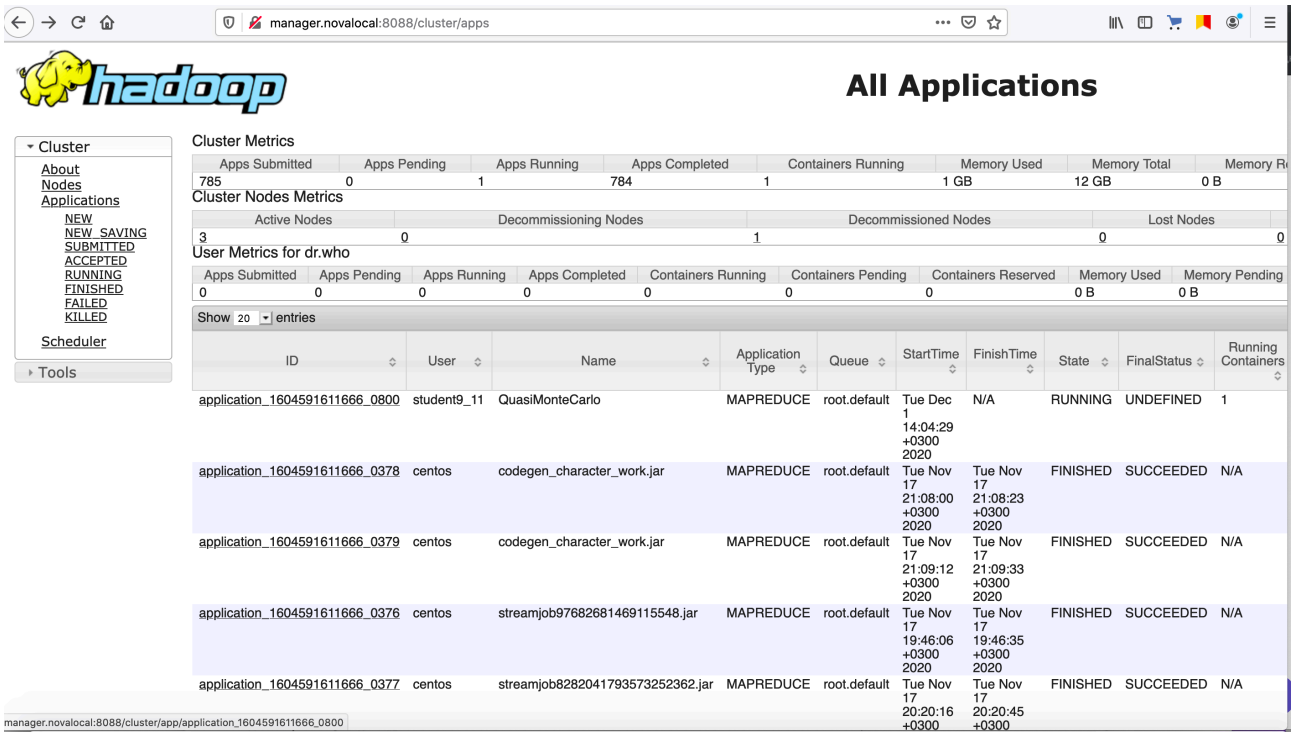
Найти свою задачи в интерфейсе Cloudera Manager

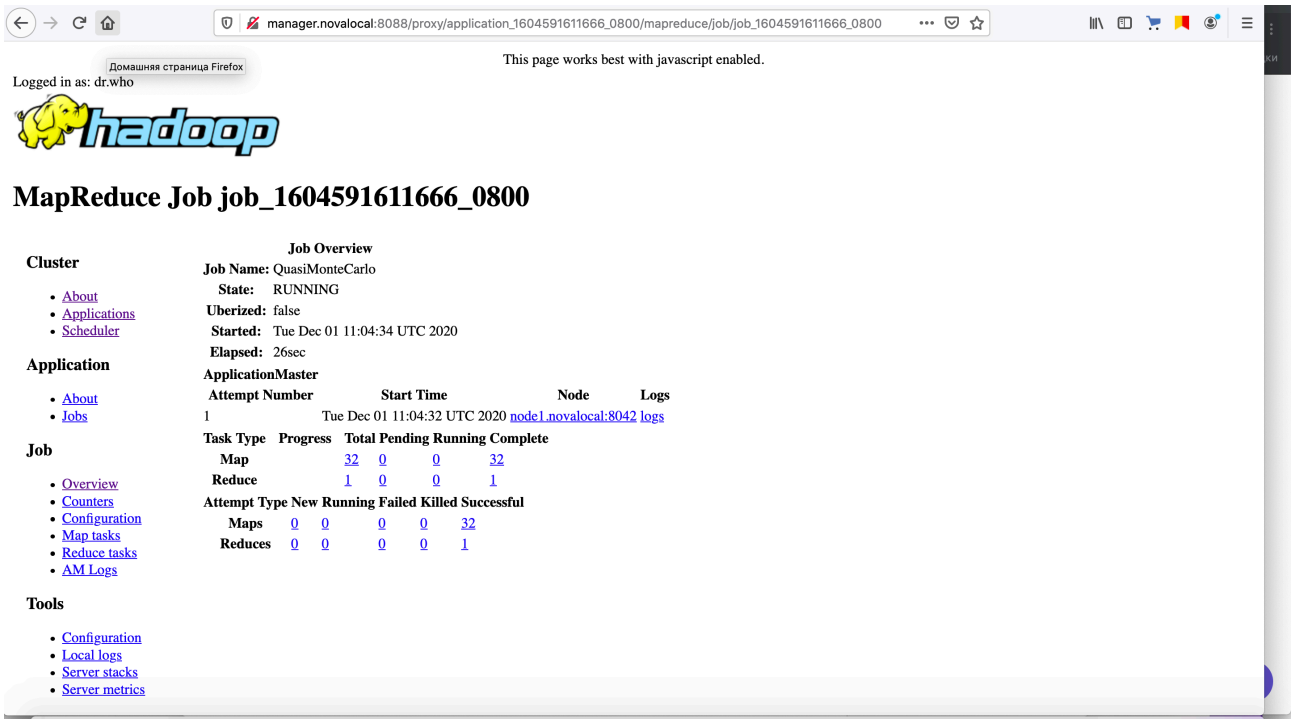
Пример:

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

yarn jar $YARN_EXAMPLES/hadoop-mapreduce-examples.jar pi 32 20000

```
[student9_11@manager ~]$ yarn jar $YARN_EXAMPLES/hadoop-mapreduce-examples.jar p
i 32 20000
Number of Maps  = 32
Samples per Map = 20000
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
Wrote input for Map #5
Wrote input for Map #6
Wrote input for Map #7
Wrote input for Map #8
Wrote input for Map #9
Wrote input for Map #10
Wrote input for Map #11
Wrote input for Map #12
Wrote input for Map #13
Wrote input for Map #14
Wrote input for Map #15
Wrote input for Map #16
Wrote input for Map #17
Wrote input for Map #18
Wrote input for Map #19
```


manager.novalocal:8088/cluster/apps
All Applications
Cluster
About
Nodes
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Cluster Metrics
Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Memory Used | Memory Total | Memory R
785 | 0 | 1 | 784 | 1 | 1 GB | 12 GB | 0 B
Cluster Nodes Metrics
Active Nodes | Decommissioning Nodes | Decommissioned Nodes | Lost Nodes
3 | 0 | 1 | 0 | 0
User Metrics for dr.who
Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Containers Pending | Containers Reserved | Memory Used | Memory Pending
0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 B | 0 B
Show 20 entries
ID | User | Name | Application Type | Queue | StartTime | FinishTime | State | FinalStatus | Running Containers
application_1604591611666_0800 | student9_11 | QuasiMonteCarlo | MAPREDUCE | root.default | Tue Dec 1 14:04:29 +0300 2020 | N/A | RUNNING | UNDEFINED | 1
application_1604591611666_0378 | centos | codegen_character_work.jar | MAPREDUCE | root.default | Tue Nov 17 21:08:00 +0300 2020 | Tue Nov 17 21:08:23 +0300 2020 | FINISHED | SUCCEEDED | N/A
application_1604591611666_0379 | centos | codegen_character_work.jar | MAPREDUCE | root.default | Tue Nov 17 21:09:12 +0300 2020 | Tue Nov 17 21:09:33 +0300 2020 | FINISHED | SUCCEEDED | N/A
application_1604591611666_0376 | centos | streamjob9768268146911554 8.jar | MAPREDUCE | root.default | Tue Nov 17 19:46:06 +0300 2020 | Tue Nov 17 19:46:35 +0300 2020 | FINISHED | SUCCEEDED | N/A
application_1604591611666_0377 | centos | streamjob8282041793573252362.jar | MAPREDUCE | root.default | Tue Nov 17 20:20:16 +0300 | Tue Nov 17 20:20:45 +0300
manager.novalocal:8088/cluster/app/application_1604591611666_0800


manager.novalocal:8088/proxy/application_1604591611666_0800/mapreduce/job/job_1604591611666_0800
Домашняя страница Firefox
This page works best with javascript enabled.
Logged in as: dr.who
MapReduce Job job_1604591611666_0800
Cluster
About
Applications
Scheduler
Application
About
Jobs
Job
Overview
Counters
Configuration
Map tasks
Reduce tasks
AM Logs
Tools
Configuration
Local logs
Server stacks
Server metrics
Job Overview
Job Name: QuasiMonteCarlo
State: RUNNING
Uberized: false
Started: Tue Dec 01 11:04:34 UTC 2020
Elapsed: 26sec
ApplicationMaster
Attempt Number | Start Time | Node | Logs
1 | Tue Dec 01 11:04:32 UTC 2020 | node1.novalocal:8042 | logs
Task Type | Progress | Total | Pending | Running | Complete
Map | 32 | 0 | 0 | 32
Reduce | 1 | 0 | 0 | 1
Attempt Type | New | Running | Failed | Killed | Successful
Maps | 0 | 0 | 0 | 0 | 32
Reduces | 0 | 0 | 0 | 0 | 1

3. Написать wordcount

Опираясь на лекцию написать wordcount, используя hadoop-streaming
или используя java. (За примерами по java пишите в telegram)

```
                a sequence file in the specified path
  [streamjob] <args> Runs streaming job with given arguments
[student9_11@manager ~]$ yarn jar /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p
0.8/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.16.2.jar -input /user/cento
s/text_example.txt -output result -file mapper.py -file reducer.py -mapper "pyth
on3 mapper.py" -reducer "python3 reducer.py"
20/12/01 10:52:57 WARN streaming.StreamJob: -file option is deprecated, please u
se generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [/opt/cloudera/parcels/CDH-5.16.2-1.cdh5.
16.2.p0.8/jars/hadoop-streaming-2.6.0-cdh5.16.2.jar] /tmp/streamjob7094625425660
745038.jar tmpDir=null
20/12/01 10:52:58 INFO client.RMProxy: Connecting to ResourceManager at manager.
novalocal/89.208.221.132:8032
20/12/01 10:52:58 INFO client.RMProxy: Connecting to ResourceManager at manager.
novalocal/89.208.221.132:8032
20/12/01 10:53:00 INFO mapred.FileInputFormat: Total input paths to process : 1
20/12/01 10:53:00 INFO mapreduce.JobSubmitter: number of splits:2
20/12/01 10:53:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
04591611666_0799
20/12/01 10:53:00 INFO impl.YarnClientImpl: Submitted application application_16
04591611666_0799
20/12/01 10:53:00 INFO mapreduce.Job: The url to track the job: http://manager.n
ovalocal:8088/proxy/application_1604591611666_0799/
20/12/01 10:53:00 INFO mapreduce.Job: Running job: job_1604591611666_0799
```

```
ovalocal:8088/proxy/application_1604591611666_0799/
20/12/01 10:53:00 INFO mapreduce.Job: Running job: job_1604591611666_0799
20/12/01 10:53:09 INFO mapreduce.Job: Job job_1604591611666_0799 running in uber
 mode : false
20/12/01 10:53:09 INFO mapreduce.Job:  map 0% reduce 0%
20/12/01 10:53:14 INFO mapreduce.Job: Task Id : attempt_1604591611666_0799_m_000
000_0, Status : FAILED
Error: java.lang.RuntimeException: Error in configuring object
        at org.apache.hadoop.util.ReflectionUtils.setJobConf(ReflectionUtils.jav
a:109)
        at org.apache.hadoop.util.ReflectionUtils.setConf(ReflectionUtils.java:7
5)
        at org.apache.hadoop.util.ReflectionUtils.newInstance(ReflectionUtils.ja
va:133)
        at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)
        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(UserGroupInforma
tion.java:1924)
        at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)
Caused by: java.lang.reflect.InvocationTargetException
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
```

К сожалению, у меня Hadoop-streaming падает с ошибкой, посмотрю, если какое-то решение

4. Ответить на вопросы

Почему Map Reduce долго выполняется?

Map Reduce долго выполняется потому что часто возникает проблема нехватки памяти из-за того, что существует много запросов на обработку, состоящих из нескольких шагов (можно сказать пайплайнов)

Почему Map Reduce не выполняется?

Возможно, из-за того, что некоторые стадии могут падать с ошибкой