

Департамент образования и науки города Москвы
Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»
Институт цифрового образования
Департамент информатики, управления и технологий

ДИСЦИПЛИНА:
«Инструменты для хранения и обработки больших данных»
пр_06
Тема:
«Introduction to Yarn + Hive»

Выполнила: Николаева С.Г., АДЭУ-201
Преподаватель: Босенко Тимур Муртазович

Москва
2023

УСТАНОВКА DOCKER

```
Activities Terminal 20:58 st1@st1-vbox: ~
File Edit View Search Terminal Help
st1@st1-vbox:~$ sudo apt update
[sudo] password for st1:
Hit:1 http://ru.archive.ubuntu.com/ubuntu bionic InRelease
Get:2 http://ru.archive.ubuntu.com/ubuntu bionic-updates InRelease [88,7 kB]
Get:3 http://ru.archive.ubuntu.com/ubuntu bionic-backports InRelease [83,3 kB]
Hit:4 https://download.docker.com/linux/ubuntu focal InRelease
Get:5 https://dl.google.com/linux/chrome/deb stable InRelease [1 825 B]
Get:6 http://security.ubuntu.com/ubuntu bionic-security InRelease [88,7 kB]
Get:7 http://ru.archive.ubuntu.com/ubuntu bionic-updates/main i386 Packages [1 643 kB]
Get:8 http://ru.archive.ubuntu.com/ubuntu bionic-updates/main amd64 Packages [2 989 kB]
Get:9 http://ru.archive.ubuntu.com/ubuntu bionic-updates/main Translation-en [5 45 kB]
Get:10 http://ru.archive.ubuntu.com/ubuntu bionic-updates/main amd64 DEP-11 Metadata [297 kB]
Get:11 http://ru.archive.ubuntu.com/ubuntu bionic-updates/restricted amd64 Packages [1 269 kB]
Get:12 http://ru.archive.ubuntu.com/ubuntu bionic-updates/restricted Translation-en [176 kB]
Get:13 http://ru.archive.ubuntu.com/ubuntu bionic-updates/universe i386 Packages [1 659 kB]
Get:14 http://ru.archive.ubuntu.com/ubuntu bionic-updates/universe amd64 Packages [1 903 kB]
Get:15 http://ru.archive.ubuntu.com/ubuntu bionic-updates/universe amd64 DEP-11 Metadata [303 kB]
Get:16 http://ru.archive.ubuntu.com/ubuntu bionic-updates/multiverse amd64 DEP-11 Metadata [2 468 B]
Get:17 http://ru.archive.ubuntu.com/ubuntu bionic-backports/main amd64 DEP-11 M
33 packages can be upgraded now. Run 'apt list --upgradable' to see them.
st1@st1-vbox:~$ sudo apt install apt-transport-https ca-certificates curl software-properties-common
Reading package lists... Done
Building dependency tree
Reading state information... Done
ca-certificates is already the newest version (20211016ubuntu0.18.04.1).
curl is already the newest version (7.58.0-2ubuntu3.24).
The following packages were automatically installed and are no longer required:
  gir1.2-goa-1.0 gir1.2-snapd-1 linux-hwe-5.4-headers-5.4.0-84
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  python3-software-properties software-properties-gtk
  ubuntu-adantage-desktop-daemon ubuntu-adantage-tools
The following NEW packages will be installed:
  ubuntu-adantage-desktop-daemon
The following packages will be upgraded:
  apt-transport-https python3-software-properties software-properties-common
  software-properties-gtk ubuntu-adantage-tools
5 upgraded, 1 newly installed, 0 to remove and 28 not upgraded.
Need to get 294 kB of archives.
After this operation, 1 964 kB disk space will be freed.
Do you want to continue? [Y/n] 
Processing triggers for man-db (2.8.7.2-2ubuntu0.1) ...
st1@st1-vbox:~$ curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo apt-key add -
OK
st1@st1-vbox:~$ 
```

```
st1@st1-vbox:~$ sudo add-apt-repository "deb [arch=amd64] https://download.docker.com/linux/ubuntu focal stable"
Hit:1 http://ru.archive.ubuntu.com/ubuntu bionic InRelease
Hit:2 http://ru.archive.ubuntu.com/ubuntu bionic-updates InRelease
Hit:3 http://ru.archive.ubuntu.com/ubuntu bionic-backports InRelease
Hit:4 https://download.docker.com/linux/ubuntu focal InRelease
Hit:5 https://dl.google.com/linux/chrome/deb stable InRelease
Hit:6 http://security.ubuntu.com/ubuntu bionic-security InRelease
```

```
st1@st1-vbox:~$ sudo apt update
Hit:1 http://ru.archive.ubuntu.com/ubuntu bionic InRelease
Hit:2 http://ru.archive.ubuntu.com/ubuntu bionic-updates InRelease
Hit:3 http://ru.archive.ubuntu.com/ubuntu bionic-backports InRelease
Hit:4 https://dl.google.com/linux/chrome/deb stable InRelease
Hit:5 https://download.docker.com/linux/ubuntu focal InRelease
Hit:6 http://security.ubuntu.com/ubuntu bionic-security InRelease
Reading package lists... Done
Building dependency tree
Reading state information... Done
28 packages can be upgraded. Run 'apt list --upgradable' to see them.
```

```
st1@st1-vbox:~$ apt-cache policy docker-ce
docker-ce:
  Installed: 5:20.10.16~3-0~ubuntu-focal
  Candidate: 5:23.0.5-1~ubuntu.20.04-focal
  Version table:
    5:23.0.5-1~ubuntu.20.04~focal 500
      500 https://download.docker.com/linux/ubuntu focal/stable amd64 Package
    5:23.0.4-1~ubuntu.20.04~focal 500
      500 https://download.docker.com/linux/ubuntu focal/stable amd64 Package
    5:23.0.3-1~ubuntu.20.04~focal 500
      500 https://download.docker.com/linux/ubuntu focal/stable amd64 Package
    5:23.0.2-1~ubuntu.20.04~focal 500
      500 https://download.docker.com/linux/ubuntu focal/stable amd64 Package
    5:23.0.1-1~ubuntu.20.04~focal 500
      500 https://download.docker.com/linux/ubuntu focal/stable amd64 Package
    5:23.0.0-1~ubuntu.20.04~focal 500
      500 https://download.docker.com/linux/ubuntu focal/stable amd64 Package
    5:20.10.24~3-0~ubuntu-focal 500
      500 https://download.docker.com/linux/ubuntu focal/stable amd64 Package
    5:20.10.23~3-0~ubuntu-focal 500
      500 https://download.docker.com/linux/ubuntu focal/stable amd64 Package
```

```
st1@st1-vbox:~$ sudo apt install docker-ce
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer required:
  gir1.2-goa-1.0 gir1.2-snapd-1 linux-hwe-5.4-headers-5.4.0-84
Use 'sudo apt autoremove' to remove them.
Suggested packages:
  aufs-tools cgroupfs-mount | cgroup-lite
The following packages will be upgraded:
  docker-ce
1 upgraded, 0 newly installed, 0 to remove and 27 not upgraded.
Need to get 22,0 MB of archives.
After this operation, 1 584 kB disk space will be freed.
Get:1 https://download.docker.com/linux/ubuntu focal/stable amd64 docker-ce amd64 5:23.0.5-1~ubuntu.20.04~focal [22,0 MB]
Fetched 22,0 MB in 2s (8 909 kB/s)
(Reading database ... 75%
```

```
st1@st1-vbox:~$ sudo systemctl status docker
● docker.service - Docker Application Container Engine
   Loaded: loaded (/lib/systemd/system/docker.service; enabled; vendor preset:
     Active: active (running) since Thu 2023-05-04 21:04:21 MSK; 52s ago
       Docs: https://docs.docker.com
   Main PID: 7214 (dockerd)
     Tasks: 8
    CGroup: /system.slice/docker.service
            └─7214 /usr/bin/dockerd -H fd:// --containerd=/run/containerd/contai

мая 04 21:04:20 st1-vbox dockerd[7214]: time="2023-05-04T21:04:20.838374140+03:
мая 04 21:04:20 st1-vbox dockerd[7214]: time="2023-05-04T21:04:20.870200035+03:
мая 04 21:04:20 st1-vbox dockerd[7214]: time="2023-05-04T21:04:20.874058087+03:
мая 04 21:04:21 st1-vbox dockerd[7214]: time="2023-05-04T21:04:21.242196030+03:
мая 04 21:04:21 st1-vbox dockerd[7214]: time="2023-05-04T21:04:21.439072761+03:
мая 04 21:04:21 st1-vbox dockerd[7214]: time="2023-05-04T21:04:21.624174073+03:
мая 04 21:04:21 st1-vbox dockerd[7214]: time="2023-05-04T21:04:21.626728860+03:
мая 04 21:04:21 st1-vbox dockerd[7214]: time="2023-05-04T21:04:21.630481295+03:
мая 04 21:04:21 st1-vbox systemd[1]: Started Docker Application Container Engin
мая 04 21:04:21 st1-vbox dockerd[7214]: time="2023-05-04T21:04:21.688697418+03:
Lines 1-19/19 (END)
```

```
st1@st1-vbox:~$ docker
Usage: docker [OPTIONS] COMMAND
A self-sufficient runtime for containers

Options:
  --config string          Location of client config files (default
                           "/home/st1/.docker")
  -c, --context string     Name of the context to use to connect to the
                           daemon (overrides DOCKER_HOST env var and
                           default context set with "docker context use")
  -D, --debug               Enable debug mode
  -H, --host list           Daemon socket(s) to connect to
  -l, --log-level string   Set the logging level
                           ("debug"|"info"|"warn"|"error"|"fatal")
                           (default "info")
  --tls                     Use TLS; implied by --tlsverify
  --tlscacert string       Trust certs signed only by this CA (default
                           "/home/st1/.docker/ca.pem")
  --tlscert string          Path to TLS certificate file (default
                           "/home/st1/.docker/cert.pem")
  --tlskey string           Path to TLS key file (default
                           "/home/st1/.docker/key.pem")
  --tlsverify               Use TLS and verify the remote
  -v, --version              Print version information and quit
```

```
st1@st1-vbox:~$ sudo usermod -aG docker $USER
[sudo] password for st1:
st1@st1-vbox:~$
```

```
Activities Terminal 4T 21:40 ● st1@st1-vbox: ~
File Edit View Search Terminal Help
st1@st1-vbox:~$ docker pull marcelmittelstaedt/hive_base:latest
latest: Pulling from marcelmittelstaedt/hive_base
d519e2592276: Extracting 2.064MB/26.71MB
d22d2dfcf9c: Download complete
b3afe92c540b: Download complete
3ed0c27de97e: Downloading 29.56MB/212.3MB
4b2ad2c564d1: Download complete
badc5288d926: Download complete
14bcce92a89e: Download complete
3846a2a4c91d: Download complete
4af5e4a42180: Download complete
6673cbcddcc0: Waiting
8099d2fb2234: Waiting
babec1283197: Waiting
673052497f18: Waiting
a815e1d7f95c: Waiting
cc0f0cb32878: Waiting
2b09721e629b: Waiting
f119db364065: Waiting
1a8ca10727f4: Waiting
345cbcf50b54: Waiting
375923500aa4: Waiting
eb5f5cf68bcd: Waiting
dc0ee2623373: Waiting
e47c350c13e8: Waiting
6360d9a6e10a: Waiting
bf8513c3486c: Waiting
```

```
Activities Terminal 4T 21:43 ● st1@st1-vbox: ~
File Edit View Search Terminal Help
st1@st1-vbox:~$ docker pull marcelmittelstaedt/hive_base:latest
latest: Pulling from marcelmittelstaedt/hive_base
d519e2592276: Pull complete
d22d2dfcf9c: Pull complete
b3afe92c540b: Pull complete
3ed0c27de97e: Pull complete
4b2ad2c564d1: Pull complete
badc5288d926: Pull complete
14bcce92a89e: Pull complete
3846a2a4c91d: Pull complete
4af5e4a42180: Pull complete
6673cbcddcc0: Pull complete
8099d2fb2234: Pull complete
babec1283197: Pull complete
673052497f18: Pull complete
a815e1d7f95c: Pull complete
cc0f0cb32878: Pull complete
2b09721e629b: Pull complete
f119db364065: Pull complete
1a8ca10727f4: Pull complete
345cbcf50b54: Extracting 231.7MB/278.9MB
375923500aa4: Download complete
eb5f5cf68bcd: Download complete
dc0ee2623373: Download complete
e47c350c13e8: Download complete
6360d9a6e10a: Download complete
bf8513c3486c: Download complete
```

```
st1@st1-vbox:~$ docker run -dit --name hive_base_container -p 8088:8088 -p 9870 :9870 -p 9864:9864 marcelmittelstaedt/hive_base:latest
3e08b75870b9cd7551102170dbc6e8a0e4929fc12e64190cb0c7845cd0999f59
st1@st1-vbox:~$
```

```
st1@st1-vbox:~$ docker ps -a
CONTAINER ID        IMAGE               COMMAND             CREATED          NAMES
3e08b75870b9        marcelmittelstaedt/hive_base:latest   "/startup.sh"      53 seconds ago
                  Up 50 seconds   0.0.0.0:8088->8088/tcp, :::8088->8088/tcp, 0.0.0.0:9864->9864/tcp, :::9864->9864/tcp, 0.0.0.0:9870->9870/tcp, :::9870->9870/tcp   hive_base_container
st1@st1-vbox:~$
```

```
st1@st1-vbox:~$ docker logs hive_base_container
* Restarting OpenBSD Secure Shell server sshd                                         [ OK ]
First start of container.
Format HDFS.
Switched to User:
hadoop
WARNING: /home/hadoop/hadoop/logs does not exist. Creating.
2023-05-04 18:44:40,992 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = 3e08b75870b9/172.17.0.2
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.1.2
STARTUP_MSG: classpath = /home/hadoop/hadoop/etc/hadoop:/home/hadoop/hadoop/share/hadoop/common/lib/snappy-java-1.0.5.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jsr311-api-1.1.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jetty-server-9.3.24.v20180605.jar:/home/hadoop/hadoop/share/hadoop/common/lib/token-provider-1.0.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/htrace-core4-4.1.0-incubating.jar:/home/hadoop/hadoop/share/hadoop/common/lib/kerb-simplekdc-1.0.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jetty-servlet-9.3.24.v20180605.jar:/home/hadoop/hadoop/share/hadoop/common/lib/re2j-1.1.jar:/home/hadoop/hadoop/share/hadoop/common/lib/jackson-annotations-2.7.8.jar:/home/hadoop/hadoop/share/hadoop/common/lib/zookeeper-3.4.13.jar:/home/hadoop/hadoop/share/hadoop/common/lib/commons-io-2.5.jar:/home/hadoop/hadoop/share/hadoop/common/lib/woodstock-core-5.0.3.jar:/home/hadoop/hadoop/share/hadoop/common/lib/kerb-crypto-1.0.1.
```

```
st1@st1-vbox:~$ docker exec -it hive_base_container bash
root@3e08b75870b9:/#
```

```
st1@st1-vbox:~$ docker exec -it hive_base_container bash
root@3e08b75870b9:/# sudo su hadoop
hadoop@3e08b75870b9:/$ cd hadoop@c821a0e1bdcf:~$
```

```
st1@st1-vbox:~$ docker exec -it hive_base_container bash
root@3e08b75870b9:/# sudo su hadoop
hadoop@3e08b75870b9:/$ cd hadoop@c821a0e1bdcf:~$ 
bash: cd: hadoop@c821a0e1bdcf:~$: No such file or directory
hadoop@3e08b75870b9:/$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
```

```
hadoop@3e08b75870b9:/ $ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/hive/lib/log4j-slf4j-impl-2.10.0
.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hadoop/share/hadoop/common/lib/s
lf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation
.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = db222454-f9fd-430b-88c0-e185779ef1c3

Logging initialized using configuration in jar:file:/home/hadoop/hive/lib/hive-
common-3.1.2.jar!/hive-log4j2.properties Async: true
Thu May 04 18:51:41 GMT 2023 Thread[db222454-f9fd-430b-88c0-e185779ef1c3 main,5
,main] java.io.FileNotFoundException: derby.log (Permission denied)
Thu May 04 18:51:41 GMT 2023 Thread[db222454-f9fd-430b-88c0-e185779ef1c3 main,5
,main] Ignored duplicate property derby.module.replication.master in jar:file:/_
home/hadoop/hive/lib/hive-druid-handler-3.1.2.jar!/org/apache/derby/modules.pro
perties
Thu May 04 18:51:41 GMT 2023 Thread[db222454-f9fd-430b-88c0-e185779ef1c3 main,5
,main] Ignored duplicate property derby.module.resultSetStatisticsFactory in ja
r:file:/home/hadoop/hive/lib/hive-druid-handler-3.1.2.jar!/org/apache/derby/mod
ules.properties
Thu May 04 18:51:41 GMT 2023 Thread[db222454-f9fd-430b-88c0-e185779ef1c3 main,5
,main] Ignored duplicate property derby.module.NoneAuthentication in jar:file:/_
home/hadoop/hive/lib/hive-druid-handler-3.1.2.jar!/org/apache/derby/modules.pro
perties
```

```
hive> show databases;  
OK  
default  
Time taken: 1.44 seconds, Fetched: 1 row(s)  
hive>
```

```
hive> exit
      >
      > hadoop@3e08b75870b9:/$ ;
bash: syntax error near unexpected token `;'
hadoop@3e08b75870b9:/$ exit
exit
root@3e08b75870b9:# exit
exit
st1@st1-vbox:~$ wget https://datasets.imdbws.com/title.basics.tsv.gz
--2023-05-04 21:56:13--  https://datasets.imdbws.com/title.basics.tsv.gz
Resolving datasets.imdbws.com (datasets.imdbws.com)... 18.165.122.124, 18.165.1
22.50, 18.165.122.39, ...
Connecting to datasets.imdbws.com (datasets.imdbws.com)|18.165.122.124|:443...
connected.
HTTP request sent, awaiting response... 200 OK
Length: 171570495 (164M) [binary/octet-stream]
Saving to: 'title.basics.tsv.gz'

title.basics.tsv.gz  28%[====>]  46,37M  10,2MB/s    eta 12s
```

```
st1@st1-vbox:~$ wget https://datasets.imdbws.com/title.ratings.tsv.gz
--2023-05-04 21:57:48-- https://datasets.imdbws.com/title.ratings.tsv.gz
Resolving datasets.imdbws.com (datasets.imdbws.com)... 18.165.122.50, 18.165.12
2.47, 18.165.122.39, ...
Connecting to datasets.imdbws.com (datasets.imdbws.com)|18.165.122.50|:443... c
onnected.
HTTP request sent, awaiting response... 200 OK
Length: 6562168 (6,3M) [binary/octet-stream]
Saving to: 'title.ratings.tsv.gz'

title.ratings.tsv.g 100%[=====] 6,26M 8,66MB/s in 0,7s

2023-05-04 21:57:49 (8,66 MB/s) - 'title.ratings.tsv.gz' saved [6562168/6562168]
]

st1@st1-vbox:~$
```

```
st1@st1-vbox:~$ gunzip title.basics.tsv.gz
st1@st1-vbox:~$ gunzip title.ratings.tsv.gz
st1@st1-vbox:~$
```

```
st1@st1-vbox:~$ $HADOOP_HOME/bin/hadoop fs -mkdir /user/hadoop/imdb
mkdir: `hdfs://localhost:9000/user/hadoop': No such file or directory
st1@st1-vbox:~$ $HADOOP_HOME/bin/hadoop fs -mkdir -p /user/hadoop/imdb
st1@st1-vbox:~$ $HADOOP_HOME/bin/hadoop fs -mkdir -p /user/hadoop/imdb/title_ba
sics
st1@st1-vbox:~$ $HADOOP_HOME/bin/hadoop fs -mkdir -p /user/hadoop/imdb/title_ra
tings
st1@st1-vbox:~$
```

```
st1@st1-vbox:~$ $HADOOP_HOME/bin/hadoop fs -put title.basics.tsv /user/hadoop/i
mdb/title_basics/title.basics.tsv
st1@st1-vbox:~$ $HADOOP_HOME/bin/hadoop fs -put title.ratings.tsv /user/hadoop/
imdb/title_ratings/title.ratings.tsv
st1@st1-vbox:~$
```

?????????????????????????????????????

```
st1@st1-vbox:~$ docker exec -it hive_base_container bash
root@a92ae4c95e66:/# sudo su hadoop
hadoop@a92ae4c95e66:/$ ls
CONTAINER_ALREADY_INITIALIZED dev lib mnt root srv tmp
bin etc lib64 opt run startup.sh usr
boot home media proc sbin sys var
hadoop@a92ae4c95e66:/$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [a92ae4c95e66]
Starting resourcemanager
Starting nodemanagers
hadoop@a92ae4c95e66:/$
```

```
root@a92ae4c95e66:/# wget https://datasets.imdbws.com/title.basics.tsv.gz
--2023-05-04 20:35:31-- https://datasets.imdbws.com/title.basics.tsv.gz
Resolving datasets.imdbws.com (datasets.imdbws.com)... 18.165.122.39, 18.165.122.124, 18.165.122.50, ...
Connecting to datasets.imdbws.com (datasets.imdbws.com)|18.165.122.39|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 171570495 (164M) [binary/octet-stream]
Saving to: 'title.basics.tsv.gz'

title.basics.tsv.gz 100%[=====] 163.62M 10.9MB/s    in 16s

2023-05-04 20:35:47 (10.2 MB/s) - 'title.basics.tsv.gz' saved [171570495/171570495]

root@a92ae4c95e66:/#
```

```
root@a92ae4c95e66:/# wget https://datasets.imdbws.com/title.ratings.tsv.gz
--2023-05-04 20:36:57-- https://datasets.imdbws.com/title.ratings.tsv.gz
Resolving datasets.imdbws.com (datasets.imdbws.com)... 18.165.122.39, 18.165.122.124, 18.165.122.50, ...
Connecting to datasets.imdbws.com (datasets.imdbws.com)|18.165.122.39|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 6562168 (6.3M) [binary/octet-stream]
Saving to: 'title.ratings.tsv.gz'

title.ratings.tsv.gz 100%[=====] 6.26M 8.56MB/s    in 0.7s

2023-05-04 20:36:58 (8.56 MB/s) - 'title.ratings.tsv.gz' saved [6562168/6562168]
```

```
root@a92ae4c95e66:/# gunzip title.basics.tsv.gz
root@a92ae4c95e66:/# gunzip title.ratings.tsv.gz
root@a92ae4c95e66:/#
```

```
root@a92ae4c95e66:/# sudo su hadoop
hadoop@a92ae4c95e66:$ hadoop fs -mkdir /user/hadoop/imdb
hadoop@a92ae4c95e66:$ hadoop fs -mkdir /user/hadoop/imdb/title_basics
hadoop@a92ae4c95e66:$ hadoop fs -mkdir /user/hadoop/imdb/title_ratings
hadoop@a92ae4c95e66:$
```

```
hadoop@a92ae4c95e66:$ hadoop fs -put title.basics.tsv /user/hadoop/imdb/title_basics/title.basics.tsv
hadoop@a92ae4c95e66:$ hadoop fs -put title.ratings.tsv /user/hadoop/imdb/title_ratings/title.ratings.tsv
hadoop@a92ae4c95e66:$
```

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS title_ratings(
    > tconst STRING,
    > average_rating DECIMAL(2,1),
    > num_votes BIGINT
    > ) COMMENT 'IMDb Ratings'
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS
    > TEXTFILE LOCATION '/user/hadoop/imdb/title_ratings'
    > TBLPROPERTIES ('skip.header.line.count'=1');
OK
Time taken: 2.463 seconds
hive> █
```

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS title_basics (
  > tconst STRING,
  > title_type STRING,
  > primary_title STRING,
  > original_title STRING,
  > is_adult DECIMAL(1,0),
  > start_year DECIMAL(4,0),
  > end_year STRING,
  > runtime_minutes INT,
  > genres STRING
  > ) COMMENT 'IMDb Movies' ROW FORMAT DELIMITED FIELDS TERMINATED BY
  > '\t' STORED AS TEXTFILE LOCATION '/user/hadoop/imdb/title_basics'
  > TBLPROPERTIES ('skip.header.line.count'='1');
OK
Time taken: 0.298 seconds
hive>
```

```
hive> select * from title_basics limit 3;
OK
tt0000001      short    Carmencita      Carmencita      0      1894      NULL      1
Documentary,Short
tt0000002      short    Le clown et ses chiens  Le clown et ses chiens  0      1
892      NULL      5      Animation,Short
tt0000003      short    Pauvre Pierrot    Pauvre Pierrot    0      1892      NULL      4
Animation,Comedy,Romance
Time taken: 7.196 seconds, Fetched: 3 row(s)
hive> █
```

```

hive> select * from title_basics limit 3;
OK
tt0000001      short    Carmencita      Carmencita      0      1894      NULL  1
Documentary,Short
tt0000002      short    Le clown et ses chiens  Le clown et ses chiens  0      1
892      NULL  5      Animation,Short
tt0000003      short    Pauvre Pierrot  Pauvre Pierrot  0      1892      NULL  4
Animation,Comedy,Romance
Time taken: 7.196 seconds, Fetched: 3 row(s)
hive> select * from title_ratings limit 3;
OK
tt0000001      5.7      1967
tt0000002      5.8      264
tt0000003      6.5      1811
Time taken: 0.375 seconds, Fetched: 3 row(s)
hive> █

```

```

hive> SELECT * FROM title_basics b JOIN title_ratings r ON (b.tconst=r.tconst)
WHERE original_title = 'The Dark Knight' and title_type='movie';
Query ID = hadoop_20230504205745_92d40761-e1a1-410f-ab51-f50117588de3
Total jobs = 2
Stage-5 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
█

```

Пришлось удалить все контейнеры, поскольку он не дает зайти в них после выключения машины

```

st1@st1-vbox:~$ docker stop hive_base_container
hive_base_container
st1@st1-vbox:~$ docker rm $(docker ps -qa)
a92ae4c95e66
st1@st1-vbox:~$ docker ps -a
CONTAINER ID   IMAGE      COMMAND      CREATED      STATUS      PORTS      NAMES
st1@st1-vbox:~$ █

```

Создание и запуск заново

```

st1@st1-vbox:~$ docker run -dit --name hive_base_container -p 8088:8088 -p 9870
:9870 -p 9864:9864 marcelmittelstaedt/hive_base:latest
df81d624add6a46ab6da92c34b407d568f1fb76089ee2d8b995ee3a25e2d455
st1@st1-vbox:~$ docker ps -a
CONTAINER ID   IMAGE      COMMAND      CREATED      STATUS      PORTS      NAMES
          STATUS      PORTS
df81d624add6   marcelmittelstaedt/hive_base:latest  "/startup.sh"  20 seconds
ago  Up 16 seconds  0.0.0.0:8088->8088/tcp, :::8088->8088/tcp, 0.0.0.0:9864-
>9864/tcp, :::9864->9864/tcp, 0.0.0.0:9870->9870/tcp, :::9870->9870/tcp  hive_
base_container

```

```

hive> SELECT * FROM title_basics b JOIN title_ratings r ON (b.tconst=r.tconst)
WHERE original_title = 'The Dark Knight' and title_type = 'movie';

```

Google Chrome ▾ Пт 14:35 •

All Applications × Namenode information × +

localhost:8088/cluster Update :

 All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory
0	0	0	0	0	0 B

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Show 20 ▾ entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Runn	Contai
No data available											

Showing 0 to 0 of 0 entries

К сожалению, не хватает оперативной памяти для выполнения данной задачи, поэтому уже третий раз виртуальная машина зависает.

МАШИНА ОЖИЛА

Activities Google Chrome ▾ Пт 14:39 •

All Applications × Namenode information × +

localhost:8088/cluster Update :

 All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory
1	0	1	0	1	0 B

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Show 20 ▾ entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Runn	Contai
application_1683286243597_0001	hadoop	SELECT * FROM title_basics b JOIN ...movie' (Stage-1)	MAPREDUCE	default	0	Fri May 5 14:39:12 +0300 2023					

Showing 1 to 1 of 1 entries

```

root@df81d624add6: ~
File Edit View Search Terminal Help
2023-05-05 11:35:13      Processing rows:      1100000 Hashtable size: 1099999
Memory usage: 459198800      percentage: 0.096
Execution failed with exit status: 137
Obtaining error information

Task failed!
Task ID:
  Stage-5

Logs:

FAILED: Execution Error, return code 137 from org.apache.hadoop.hive.ql.exec.mr.MapredLocalTask
ATTEMPT: Execute BackupTask: org.apache.hadoop.hive.ql.exec.mr.MapRedTask
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 4
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1683286243597_0001, Tracking URL = http://df81d624add6:8088/proxy/application_1683286243597_0001
Kill Command = /home/hadoop/hadoop/bin/mapred job -kill job_1683286243597_0001
Hadoop job information for Stage-1: number of mappers: 5; number of reducers: 4
2023-05-05 11:40:10,406 Stage-1 map = 0%, reduce = 0%

```

На этом окончательно зависло

Часть 2

```
root@df81d624add6:/# wget https://datasets.imdbws.com/name.basics.tsv.gz
--2023-05-09 15:01:27-- https://datasets.imdbws.com/name.basics.tsv.gz
Resolving datasets.imdbws.com (datasets.imdbws.com)... 18.165.122.47, 18.165.12
2.39, 18.165.122.50, ...
Connecting to datasets.imdbws.com (datasets.imdbws.com)|18.165.122.47|:443... c
onnected.
HTTP request sent, awaiting response... 200 OK
Length: 245176378 (234M) [binary/octet-stream]
Saving to: 'name.basics.tsv.gz'

name.basics.tsv.gz 100%[=====] 233.82M 10.7MB/s    in 22s

2023-05-09 15:01:49 (10.6 MB/s) - 'name.basics.tsv.gz' saved [245176378/2451763
78]

root@df81d624add6:/# gunzip name.basics.tsv.gz
root@df81d624add6:/#
```

```
root@df81d624add6:/# sudo su hadoop
hadoop@df81d624add6:$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 223. Stop it first.
pdsh@df81d624add6: localhost: ssh exited with exit code 1
Starting datanodes
localhost: datanode is running as process 328. Stop it first.
pdsh@df81d624add6: localhost: ssh exited with exit code 1
Starting secondary namenodes [df81d624add6]
df81d624add6: secondarynamenode is running as process 502. Stop it first.
pdsh@df81d624add6: df81d624add6: ssh exited with exit code 1
Starting resourcemanager
resourcemanager is running as process 795. Stop it first.
Starting nodemanagers
localhost: nodemanager is running as process 906. Stop it first.
pdsh@df81d624add6: localhost: ssh exited with exit code 1
hadoop@df81d624add6:$ hadoop fs -mkdir /user/hadoop/imdb/name_basics
hadoop@df81d624add6:$
```

```
hadoop@df81d624add6:$ hadoop fs -put name.basics.tsv /user/hadoop/imdb/name_ba
sics/name.basics.tsv
```

Задание 4

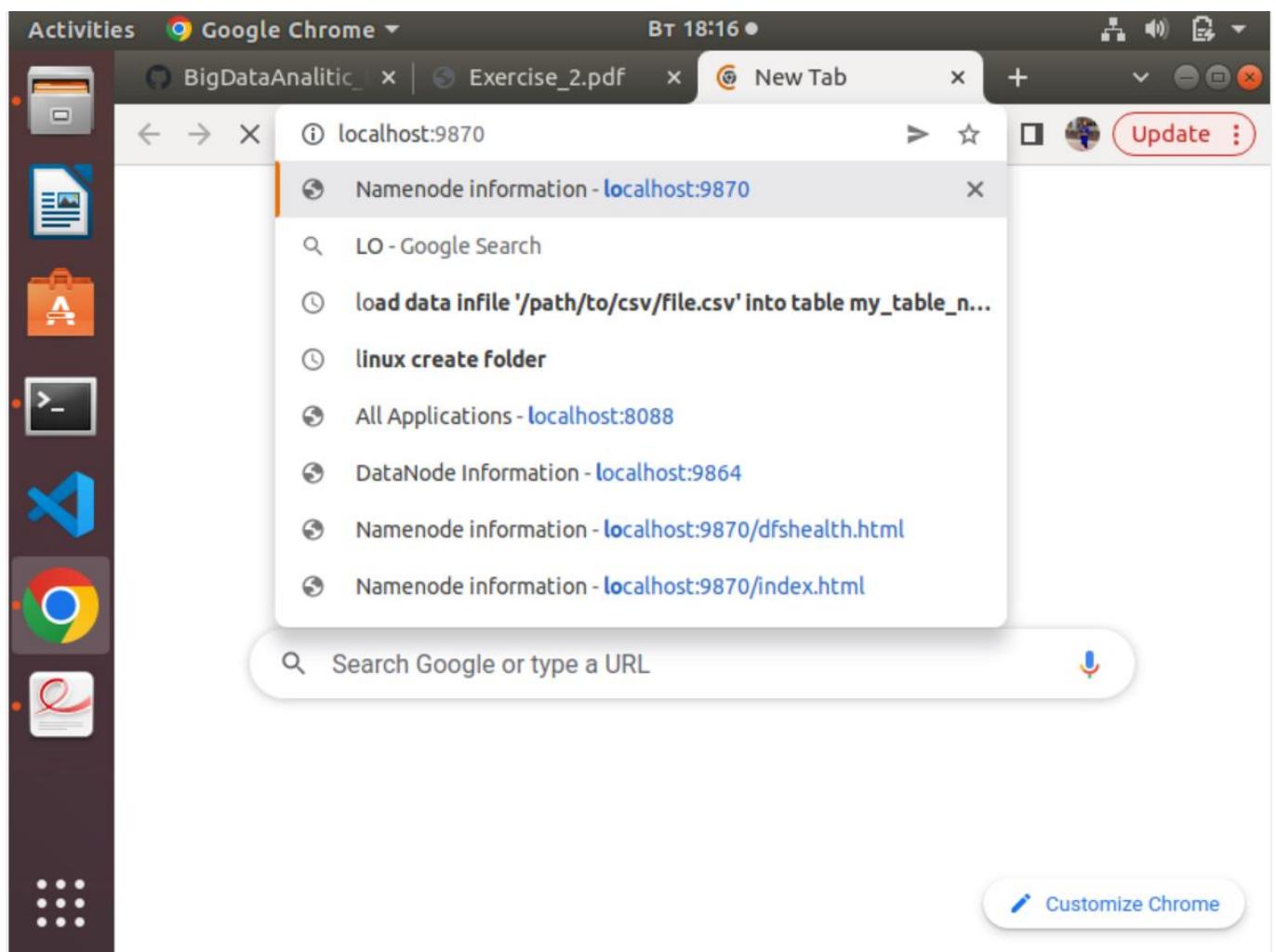
```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS name_basics(
    > nconst STRING,
    > primary_name STRING,
    > birth_year INT,
    > death_year STRING,
    > primary_profession STRING,
    > known_for_titles STRING
    >) COMMENT 'IMDb Actors' ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' ST
ORED AS TEXTFILE LOCATION '/user/hadoop/imdb/name_basics'
    > TBLPROPERTIES ('skip.header.line.count'='1');
OK
Time taken: 1.744 seconds
hive>
```

Задание 5

- а) Сколько фильмов и сериалов находится в наборе данных IMDB?

```
hive> SELECT m.title_type, count(*)
      > FROM title_basics m GROUP BY m.title_type;
Query ID = hadoop_20230509151423_1db2e3b9-d8fe-4789-a857-8134128927a7
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 4
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1683320683135_0001, Tracking URL = http://df81d624add6:8088/
proxy/application_1683320683135_0001/
Kill Command = /home/hadoop/hadoop/bin/mapred job -kill job_1683320683135_0001
■
```

Зависло



Перезапуск

```

s Terminal  BT 18:29 ●
root@df81d624add6:/
File Edit View Search Terminal Help
st1@st1-vbox:~$ docker ps -a
CONTAINER ID   IMAGE          COMMAND           CREATED          STATUS          PORTS
NAMES
df81d624add6   marcelmittelstaedt/hive_base:latest "/startup.sh"  4 days ago
    Exited (255) 2 minutes ago  0.0.0.0:8088->8088/tcp, :::8088->8088/tcp, 0.0.0.0:9864->9864/tcp, :::9864->9864/tcp, 0.0.0.0:9870->9870/tcp, :::9870->9870/tcp
hive_base_container
st1@st1-vbox:~$ docker start hive_base_container
hive_base_container
st1@st1-vbox:~$ docker exec -it hive_base_container bash
root@df81d624add6:/# sudo su hadoop
hadoop@df81d624add6:/$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [df81d624add6]
Starting resourcemanager
Starting nodemanagers
hadoop@df81d624add6:/$ █

```

Задача зарегистрирована, но опять зависло

The screenshot shows the Hadoop NameNode information page via a Google Chrome browser. The URL is `localhost:8088/cluster`. The page includes:

- Cluster Metrics:**

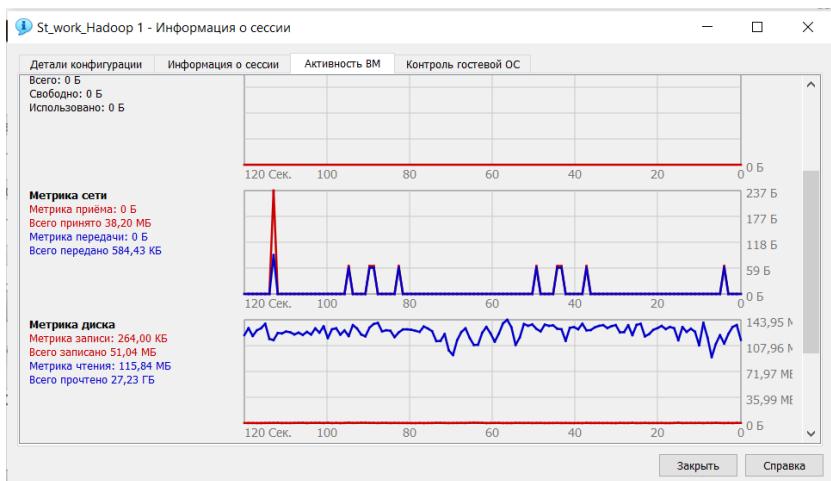
	Apps Submitted	Apps Pending	Apps Running	Apps Completed	
1	0	1	0	1	
- Cluster Nodes Metrics:**

Active Nodes	Decommissioning Nodes	Decom.
1	0	0
- Scheduler Metrics:**

Scheduler Type	Scheduling Resource Type
Capacity Scheduler	[memory-mb (unit=Mi), vcores]
- Applications:**

ID	User	Name	Application Type	Queue	Application Priority
application_1683646157447_0001	hadoop	SELECT m.title_type, count(*)...m.title_type (Stage-1)	MAPREDUCE	default	0

Showing 1 to 1 of 1 entries



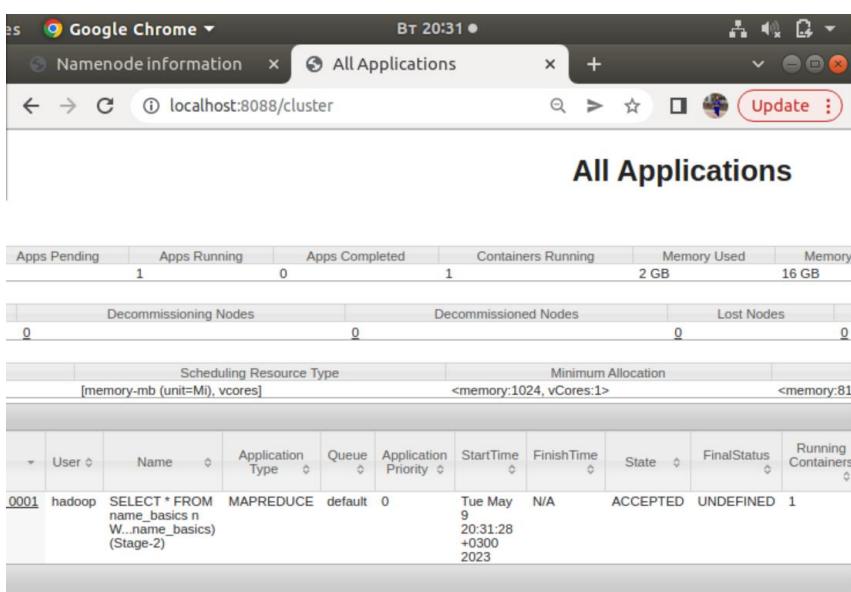
Спустя 25 минут не отвисло



b) Кто самый молодой актер/сценарист/... в наборе данных?

```
hive> SELECT * FROM name_basics n
  > WHERE n.birth_year = ( SELECT MAX(birth_year) FROM name_basics);■
```

Задача зарегистрирована



The screenshot shows the Apache Hadoop ResourceManager web interface. It displays various metrics and application logs.

Cluster Metrics:

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running
2	0	1	1	1

Cluster Nodes Metrics:

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

Scheduler Metrics:

Scheduler Type	Scheduling Resource Type	Minimum Allocat
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Application Log:

```

2023-05-09T20:38:20+0300
application_1683653177595_0002 hadoop SELECT * FROM title_basics m GROUP BY m.title_type;
2023-05-09T20:31:28+0300
application_1683653177595_0001 hadoop SELECT * FROM title_basics m GROUP BY m.title_type;
  
```

Showing 1 to 2 of 2 entries

Выполнение на другом компьютере

- a) Сколько фильмов и сериалов находится в наборе данных IMDB?

```

hive> SELECT m.title_type, count(*)
      > FROM title_basics m GROUP BY m.title_type;  
```

The screenshot shows the Apache Hadoop ResourceManager web interface. It displays various metrics and application logs.

Cluster Metrics:

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running
0	0	1	0	1

Cluster Nodes Metrics:

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
0	0	0

Scheduler Metrics:

Scheduler Type	Scheduling Resource Type	Minimum Allocat
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Application Log:

```

2023-05-09T22:16:22+0300
application_1683658545148_0001 hadoop SELECT m.title_type, count(*)...m.title_type
                                         (Stage-1)
  
```

Showing 1 to 1 of 1 entries

The screenshot shows the Apache Ambari web interface for the 'BigDataAnalytic_Practice' cluster. At the top, there are tabs for 'All Applications' and a search bar. The URL is 'localhost:8088/cluster'. A red 'Update' button is visible in the top right. The page displays various cluster metrics:

Total	Memory Reserved	Vcores Used	Vcores Total	Vcores Reserved
0 B		5	8	0

Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
0		0

Maximum Allocation	Maximum Cluster Application Priority
<vCores:4>	0

Below these tables is a search bar labeled 'Search:' and a table header for monitoring application details:

Allocated CPU Vcores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
----------------------	---------------------	---------------------	--------------------	------------	--------------	----------	-------------	-------------------

A single row of data is shown:

5	14336	0	0	87.5	87.5	<input type="checkbox"/>	ApplicationMaster	0
---	-------	---	---	------	------	--------------------------	-------------------	---

At the bottom are navigation buttons: First, Previous, 1, Next, Last.

Успешное выполнение задачи

The screenshot shows the 'All Applications' page for the 'BigDataAnalytic_Practice' cluster. The title 'All Applications' is prominently displayed. The URL is 'localhost:8088/cluster'. A red 'Update' button is visible in the top right. The page displays cluster metrics and a table of completed applications:

Running	Memory Used	Memory Total	Memory Reserved	Vcores Used	Vcores Total	Vcores Reserved
0 B	16 GB	0 B	0 B	0	8	0

Lost	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
0	0	0	0	0

Maximum Allocation	Maximum Allocation	Maximum Cluster Application Priority
<memory:8192, vCores:1>	<memory:8192, vCores:4>	0

Below these tables is a search bar labeled 'Search:' and a table header for monitoring application details:

FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	% of Queue	% of Cluster	Progress
------------	-------	-------------	--------------------	----------------------	---------------------	---------------------	--------------------	------------	--------------	----------

A single row of data is shown, representing a completed task:

Tue May 9 22:19:34 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<input type="checkbox"/>
-------------------------------	----------	-----------	-----	-----	-----	-----	-----	-----	-----	--------------------------

At the bottom are navigation buttons: First, Previous, 1, Next, Last.

```

Ended Job = job_1683658545148_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 4 Reduce: 4 Cumulative CPU: 74.35 sec HDFS Read: 84259
5831 HDFS Write: 643 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 14 seconds 350 msec
OK
movie    644471
short    929121
tvMiniSeries 48463
tvEpisode   7471167
tvSeries    243107
videoGame   34480
tvMovie    141421
tvPilot    1
tvShort    10093
tvSpecial   41497
video     273907
Time taken: 219.126 seconds, Fetched: 11 row(s)
hive> Show Applications
hive>

```

b) Кто самый молодой актер/сценарист/... в наборе данных?

```

hive> SELECT * FROM name_basics n
> WHERE n.birth_year = ( SELECT MAX(birth_year) FROM name_basics);
Query ID = hadoop_20230509192408_913ff0d1-8591-4ddf-9f94-f7a804a3cd0e
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>

```

Задача зарегистрирована

The screenshot shows the Apache Ambari web interface. At the top, there's a navigation bar with tabs for 'BigDataAnalytic_Practice' and 'All Applications'. Below the navigation is a search bar with the URL 'localhost:8088/cluster'. The main content area displays cluster statistics: Apps Pending (1), Apps Running (1), Apps Completed (2), Containers Running (2), Memory Used (5 GB), and Mem (16 GB). It also shows resource usage: Decommissioning Nodes (0), Decommissioned Nodes (0), and Lost Nodes (0). A detailed table below lists two running applications:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Run Conta
02	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-2)	MAPREDUCE	default	0	Tue May 9 22:24:19 +0300 2023	N/A	RUNNING	UNDEFINED	2
01	hadoop	SELECT m.title_type, count(*)...m.title_type (Stage-1)	MAPREDUCE	default	0	Tue May 9 22:16:22 +0300 2023	Tue May 9 22:19:34 +0300 2023	FINISHED	SUCCEEDED	N/A

Успешное выполнение задачи

Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total
	2 GB	16 GB	0 B	1	8

Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Sh
0	0	0	0	0

Minimum Allocation <vCores:1>	Maximum Allocation <memory:8192, vCores:4>	Maximum Cluster Application 0

FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	% of Queue	% of Cluster	Progress
Tue May 9 22:26:22 +0300 2023	FINISHED	SUCCEEDED	1	1	2048	0	0	12.5	12.5	
Tue May 9 22:25:09 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	
Tue May 9 22:19:34 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	


```

2023-05-09 19:47:36      Starting to launch local task to process map join;   m
maximum memory = 4772593664
2023-05-09 19:47:40      Uploaded 1 File to: file:/tmp/hadoop/c4501f99-87e7-4c7d
-9259-05781cc52c09/hive_2023-05-09_19-46-12_511_322076066008237280-1/-local-100
05/HashTable-Stage-3/MapJoin-mapfile30--.hashtable (260 bytes)
2023-05-09 19:47:40      End of local task; Time Taken: 4.344 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 2
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1683658545148_0010, Tracking URL = http://16c4596441db:8088/
proxy/application_1683658545148_0010/
Kill Command = /home/hadoop/hadoop/bin/mapred job -kill job_1683658545148_0010
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2023-05-09 19:48:00,526 Stage-3 map = 0%,  reduce = 0%
2023-05-09 19:48:17,404 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 9.41 s
ec
MapReduce Total cumulative CPU time: 9 seconds 410 msec
Ended Job = job_1683658545148_0010
MapReduce Jobs Launched:
Stage-Stage-2: Reduce: 1  Cumulative CPU: 10.41 sec  HDFS Read: 6061 HDFS Write:
96 SUCCESS
Stage-Stage-3: Map: 1  Cumulative CPU: 9.41 sec  HDFS Read: 7325 HDFS Write:
87 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 820 msec
OK
Time taken: 128.329 seconds
hive>

```

c) Создайте список (tconst, original_title, start_year, average_rating, num_votes), который состоит из: - фильм вышел в 2010 году или позднее; - фильм имеет средний рейтинг, равный или превышающий 8,1 - проголосовали более 100 000 раз

```

hive> SELECT m.tconst, m.original_title, m.start_year, r.average_rating, r.num_votes
    > FROM title_basics m JOIN title_ratings r ON (m.tconst = r.tconst)
    > WHERE r.average_rating >= 8.1 AND m.start_year >= 2010 AND m.title_type =
      'movie'
    > AND r.num_votes > 100000
    > ORDER BY r.average_rating DESC, r.num_votes DESC;

```

Задача зарегистрирована

localhost:8088/cluster						
ID	User	Query	Type	Stage	Start Time	End Time
5148_0006	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-2)	MAPREDUCE	default	0	Tue May 9 22:35:58 +0300 2023
5148_0005	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-3)	MAPREDUCE	default	0	Tue May 9 22:34:41 +0300 2023
5148_0004	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-2)	MAPREDUCE	default	0	Tue May 9 22:30:16 +0300 2023
5148_0003	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-3)	MAPREDUCE	default	0	Tue May 9 22:25:54 +0300 2023
5148_0002	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-2)	MAPREDUCE	default	0	Tue May 9 22:24:19 +0300 2023
5148_0001	hadoop	SELECT m.title_type, count(*)...m.title_type (Stage-1)	MAPREDUCE	default	0	Tue May 9 22:16:22 +0300 2023

Успешное выполнение задачи

localhost:8088/cluster						
ID	User	Query	Type	Stage	Start Time	End Time
5148_0006	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-2)	MAPREDUCE	default	N/A	N/A
5148_0005	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-3)	MAPREDUCE	default	N/A	N/A
5148_0004	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-2)	MAPREDUCE	default	N/A	N/A
5148_0003	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-3)	MAPREDUCE	default	N/A	N/A
5148_0002	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-2)	MAPREDUCE	default	N/A	N/A
5148_0001	hadoop	SELECT m.title_type, count(*)...m.title_type (Stage-1)	MAPREDUCE	default	N/A	N/A

```

File Edit View Search Terminal Help
tt10272386      The Father      2020     8.2      166592
tt4729430       Klaus          2019     8.2      165327
tt10366206      John Wick: Chapter 4 2023     8.2      141800
tt5813916       Dag II           2016     8.2      109109
tt4849438       Bahubali 2: The Conclusion 2017     8.2      106227
tt1392190       Mad Max: Fury Road    2015     8.1      1026676
tt2267998       Gone Girl        2014     8.1      1005602
tt1201607      Harry Potter and the Deathly Hallows: Part 2 2011     8.1      8
97670
tt2278388      The Grand Budapest Hotel 2014     8.1      835208
tt3315342       Logan            2017     8.1      784418
tt0892769      How to Train Your Dragon 2010     8.1      758759
tt1392214      Prisoners         2013     8.1      742112
tt2096673       Inside Out        2015     8.1      733470
tt2024544      12 Years a Slave   2013     8.1      714223
tt2119532       Hacksaw Ridge    2016     8.1      546845
tt5027774      Three Billboards Outside Ebbing, Missouri 2017     8.1      5
24269
tt1979320      Rush              2013     8.1      489789
tt1895587      Spotlight         2015     8.1      480871
tt1454029      The Help          2011     8.1      471123
tt3170832      Room              2015     8.1      430575
tt1950186      Ford v Ferrari 2019     8.1      415112
tt3011894      Relatos salvajes 2014     8.1      204458
tt2338151      PK                2014     8.1      192606
tt4016934      Ah-ga-ssi        2016     8.1      156050
Time taken: 236.925 seconds, Fetched: 56 row(s)
hive> ■

```

d) Сколько фильмов находится в списке с)?

```

hive> SELECT count(*)
   > FROM title_basics m JOIN title_ratings r on (m.tconst = r.tconst)
   > WHERE r.average_rating >= 8.1 and m.start_year >= 2010 and m.title_type =
'movie'
   > and r.num_votes > 100000;■

```

Задача зарегистрирована и выполнена

ID	User	Query	Type	Status	Start Time	End Time	Duration	Accepted	Undefined	Count
4	hadoop	SELECT count(*) FROM title_basics m...100000 (Stage-2)	MAPREDUCE	default	0	Tue May 9 22:59:01 +0300 2023	N/A	ACCEPTED	UNDEFINED	1
3	hadoop	SELECT count(*) FROM title_basics m...100000 (Stage-4)	MAPREDUCE	default	0	Tue May 9 22:57:30 +0300 2023	Tue May 9 22:58:54 +0300 2023	FINISHED	SUCCEEDED	N/A
2	hadoop	SELECT m.tconst, m.original_title, m....DESC (Stage-2)	MAPREDUCE	default	0	Tue May 9 22:52:02 +0300 2023	Tue May 9 22:53:42 +0300 2023	FINISHED	SUCCEEDED	N/A
1	hadoop	SELECT m.tconst, m.original_title, m....DESC (Stage-4)	MAPREDUCE	default	0	Tue May 9 22:50:32 +0300 2023	Tue May 9 22:51:47 +0300 2023	FINISHED	SUCCEEDED	N/A
0	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-3)	MAPREDUCE	default	0	Tue May 9 22:47:46 +0300 2023	Tue May 9 22:48:18 +0300 2023	FINISHED	SUCCEEDED	N/A
9	hadoop	SELECT * FROM name_basics n W...name_basics) (Stage-2)	MAPREDUCE	default	0	Tue May 9 22:46:17 +0300 2023	Tue May 9 22:47:03 +0300 2023	FINISHED	SUCCEEDED	N/A
8	hadoop	SELECT m.title_type	MAPREDUCE	default	0	Tue May 9 22:41:40	Tue May 9 22:41:40	FINISHED	SUCCEEDED	N/A

The screenshot shows a Linux desktop environment with several windows open. In the foreground, a terminal window displays the following output:

```

Ended Job = job_1683658545148_0014
MapReduce Jobs Launched:
Stage-Stage-4: Map: 4   Cumulative CPU: 54.4 sec   HDFS Read: 842586363 HDFS Write: 456 SUCCESS
Stage-Stage-2: Map: 1   Reduce: 1   Cumulative CPU: 13.3 sec   HDFS Read: 8104 HDFS Write: 102 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 7 seconds 700 msec
OK
56
Time taken: 238.195 seconds, Fetched: 1 row(s)
hive> 

```

Below the terminal, a browser window shows a table titled "BigDataAnalytic_Practice" with the URL "localhost:8088/cluster". The table lists six completed jobs, each with a timestamp, state, final status, and various performance metrics like CPU and memory usage.

	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress
1	Tue May 9 23:00:37 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%;"></div>
2	Tue May 9 22:58:54 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%;"></div>
3	Tue May 9 22:53:42 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%;"></div>
4	Tue May 9 22:51:47 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%;"></div>
5	Tue May 9 22:48:18 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%;"></div>
6	Tue May 9 22:47:03 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%;"></div>

Time taken: 238.195 seconds, Fetched: 1 row(s)

e) Мы хотим знать, какие годы были великими для кинематографа. Создайте список с одной строкой в год и соответствующим количеством фильмов, которые: - имеют средний рейтинг выше 8; - были проголосованы более 100 000 раз в порядке убывания количества фильмов.

```

hive> SELECT m.start_year, count(*)
> FROM title_basics m JOIN title_ratings r ON (m.tconst = r.tconst)
> WHERE r.average_rating > 8 AND m.title_type = 'movie'
> AND r.num_votes > 100000
> GROUP BY m.start_year
> ORDER BY count(*) DESC;
Query ID = hadoop_20230509200557_46f91165-505b-49bb-99c8-666073f9bc11
Total jobs = 4
Stage-7 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
SLF4J: Found binding in [jar:file:/home/hadoop/hive/lib/log4j-slf4j-impl-2.10.0.jar!/_org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hadoop/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/_org/slf4j/impl/StaticLoggerBinder.class]

```

Задача зарегистрирована

SELECT m.start_year, count(*) FROM ti...DESC (Stage-5)	MAPREDUCE	default	0	Tue May 9 23:06:49 +0300 2023	N/A	ACCEPTED	UNDEFINED	1	1
SELECT count(*) FROM title_basics m...100000 (Stage-2)	MAPREDUCE	default	0	Tue May 9 22:59:01 +0300 2023	Tue May 9 23:00:37 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A
SELECT count(*) FROM title_basics m...100000 (Stage-4)	MAPREDUCE	default	0	Tue May 9 22:57:30 +0300 2023	Tue May 9 22:58:54 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A
SELECT m.tconst, m.original_title, m....DESC (Stage-2)	MAPREDUCE	default	0	Tue May 9 22:52:02 +0300 2023	Tue May 9 22:53:42 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A
SELECT m.tconst, m.original_title, m....DESC (Stage-4)	MAPREDUCE	default	0	Tue May 9 22:50:32 +0300 2023	Tue May 9 22:51:47 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A
SELECT * FROM name_basics n W...name_basics) (Stage-3)	MAPREDUCE	default	0	Tue May 9 22:47:46 +0300 2023	Tue May 9 22:48:18 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A
SELECT * FROM	MAPREDUCE	default	0	Tue May 9 22:47:50 +0300 2023	Tue May 9 22:49:00 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A

Успешное выполнение задачи

Tue May 9 23:10:30 +0300 2023	FINISHED	SUCCEEDED	1	1	2048	0	0	12.5	12.5
Tue May 9 23:08:59 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
Tue May 9 23:00:37 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
Tue May 9 22:58:54 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
Tue May 9 22:53:42 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
Tue May 9 22:51:47	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0

St_work_Hadoop [Работает] - Oracle VM VirtualBox

Activities Terminal BT 23:11
root@16c4596441db: /

```
File Edit View Search Terminal Help
1931      2
1985      2
1992      2
2023      1
1990      1
1987      1
1978      1
1977      1
1972      1
1971      1
1967      1
1964      1
1963      1
1961      1
1958      1
1952      1
1949      1
1946      1
1944      1
1942      1
1941      1
1936      1
1934      1
1927      1
1925      1
1921      1
Time taken: 338.298 seconds, Fetched: 82 row(s)
hive> █
```

St_work_Hadoop [Работает] - Oracle VM VirtualBox

Activities Terminal BT 23:12
root@16c4596441db: /

```
File Edit View Search Terminal Help
OK
1995      8
2014      6
1957      6
2009      6
2004      6
2003      6
2001      6
2000      6
1999      6
1994      6
2019      6
2006      5
2010      5
1998      5
2007      5
2011      5
2016      5
2015      4
1984      4
1988      4
2017      4
1954      4
1959      4
2002      4
1975      4
2020      4
1997      4
1980      4
```

So 1995 seems to be a really good year for cinema, 8 really good movies have been releases, but which are they?

```

hive> SELECT
    > m.tconst, m.original_title, m.start_year, r.average_rating,
    > r.num_votes
    > FROM title_basics m JOIN title_ratings r ON (m.tconst = r.tconst)
    > WHERE
    > r.average_rating > 8 AND m.title_type = 'movie'
    > AND r.num_votes > 100000 AND m.start_year = 1995
    > ORDER BY r.average_rating DESC;
Query ID = hadoop_20230509201356_62117f35-e47b-40bb-a7f5-5b1256de0dac
Total jobs = 3
Stage-6 is selected by condition resolver.

```

Задача зарегистрирована

Show 20 ▾ entries								
ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	
application_1683658545148_0018	hadoop	SELECT m.tconst, m.original_title, m....DESC (Stage-4)	MAPREDUCE	default	0	Tue May 9 23:14:44 +0300 2023	N/A	
application_1683658545148_0017	hadoop	SELECT m.start_year, count(*) FROM ti...DESC (Stage-3)	MAPREDUCE	default	0	Tue May 9 23:10:34 +0300 2023	Tue May 9 23:11:34 +0300 2023	
application_1683658545148_0016	hadoop	SELECT m.start_year, count(*) FROM ti...DESC (Stage-2)	MAPREDUCE	default	0	Tue May 9 23:09:10 +0300 2023	Tue May 9 23:10:30 +0300 2023	
application_1683658545148_0015	hadoop	SELECT m.start_year, count(*) FROM ti...DESC (Stage-5)	MAPREDUCE	default	0	Tue May 9 23:06:49 +0300 2023	Tue May 9 23:08:59 +0300 2023	
application_1683658545148_0014	hadoop	SELECT count(*) FROM title_basics m...100000 (Stage-2)	MAPREDUCE	default	0	Tue May 9 22:59:01 +0300 2023	Tue May 9 23:00:37 +0300 2023	

Успешное выполнение задачи

FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	% of Queue	% of Cluster	Progress
Tue May 9 23:18:11 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%; background-color: #ccc;"></div>
Tue May 9 23:16:33 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%; background-color: #ccc;"></div>
Tue May 9 23:11:34 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%; background-color: #ccc;"></div>
Tue May 9 23:10:30 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%; background-color: #ccc;"></div>
Tue May 9 23:08:59 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%; background-color: #ccc;"></div>
Tue May 9 23:00:37	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	<div style="width: 100%; background-color: #ccc;"></div>

```

Ended Job = job_1683658545148_0019
MapReduce Jobs Launched:
Stage-Stage-4: Map: 4   Cumulative CPU: 87.06 sec   HDFS Read: 842584195 HDFS W
rite: 738 SUCCESS
Stage-Stage-2: Map: 1   Reduce: 1   Cumulative CPU: 19.21 sec   HDFS Read: 10310
HDFS Write: 477 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 46 seconds 270 msec
OK
tt0114369      Se7en    1995     8.6      1691579
tt0114814      The Usual Suspects    1995     8.5      1101917
tt0112573      Braveheart    1995     8.4      1053495
tt0114709      Toy Story     1995     8.3      1014798
tt0113277      Heat       1995     8.3      671467
tt0112641      Casino      1995     8.2      532882
tt0113247      La haine     1995     8.1      180616
tt0112471      Before Sunrise 1995     8.1      319139
Time taken: 256.133 seconds, Fetched: 8 row(s)

```

```

hive> SELECT * FROM title_basics b JOIN title_ratings r ON (b.tconst=r.tconst)
WHERE original_title = 'The Dark Knight' and title_type='movie';
Query ID = hadoop_20230509202109_ed613400-2b11-4b78-b6a5-ff4c5af5917a
Total jobs = 2
Stage-5 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.

```

Задача зарегистрирована

	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus
0020	hadoop	SELECT * FROM title_basics WHERE original_title = 'The Dark Knight' and title_type='movie' (Stage-3)	MAPREDUCE	default	0	Tue May 9 23:22:04 +0300 2023	N/A	ACCEPTED	UNDEFINE