Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
УРАЛЬСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ ПЕРВОГО ПРЕЗИДЕНТА РОССИИ Б.Н. ЕЛЬЦИНА
(УрФУ имени первого Президента России Б.Н. Ельцина)
Институт радиоэлектроники и информационных технологий — РТФ

ОТЧЁТ

по лабораторной работе №4

по дисциплине «Методы и инструменты анализа больших данных»

Преподаватель ____________ ____________ С.Г. Мирвода
(дата) (подпись)

Студент ____________ ____________ А.М. Белоусов
(дата) (подпись)

Студент ____________ ____________ А.В. Жиденко
(дата) (подпись)

Группа: РИМ-201211

Екатеринбург 2021

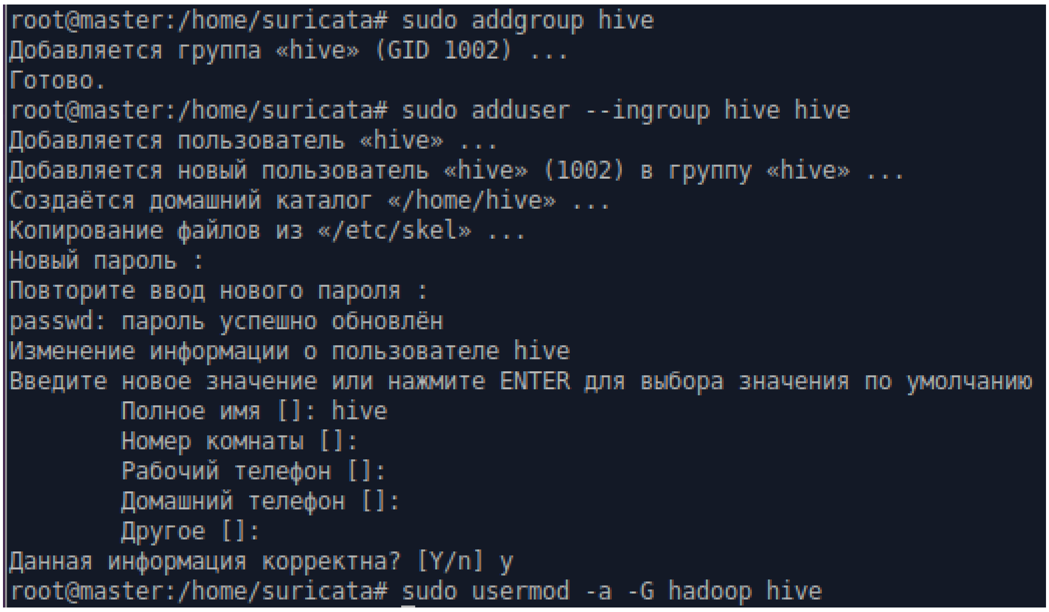
Цель работы: знакомство с базой данных HIVE.

Задание 0

Задача подготовить полигон

1. Установить на свой кластер hadoop 3.3 СУБД HIVE 3.1.2 согласно инструкции и примеру

Создание пользователя Hive



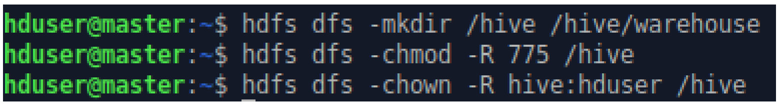
```
root@master:/home/suricata# sudo addgroup hive
Добавляется группа «hive» (GID 1002) ...
Готово.
root@master:/home/suricata# sudo adduser --ingroup hive hive
Добавляется пользователь «hive» ...
Добавляется новый пользователь «hive» (1002) в группу «hive» ...
Создаётся домашний каталог «/home/hive» ...
Копирование файлов из «/etc/skel» ...
Новый пароль :
Повторите ввод нового пароля :
passwd: пароль успешно обновлён
Изменение информации о пользователе hive
Введите новое значение или нажмите ENTER для выбора значения по умолчанию
        Полное имя []: hive
        Номер комнаты []:
        Рабочий телефон []:
        Домашний телефон []:
        Другое []:
Данная информация корректна? [Y/n] y
root@master:/home/suricata# sudo usermod -a -G hadoop hive
```

Установка Hive

```
wget https://downloads.apache.org/hive/hive-3.1.2/apache-hive-3.1.2-bin.tar.gz
tar -xf apache-hive-3.1.2-bin.tar.gz -C /usr/local/
chmod -R 755 /usr/local/apache-hive-3.1.2-bin
sudo chown -R hive:hive /usr/local/apache-hive-3.1.2-bin
```

Создание директории «warehouse» в HDFS

```
su - hduser
hdfs dfs -mkdir /hive /hive/warehouse
hdfs dfs -chmod -R 775 /hive
hdfs dfs -chown -R hive:hduser /hive
```

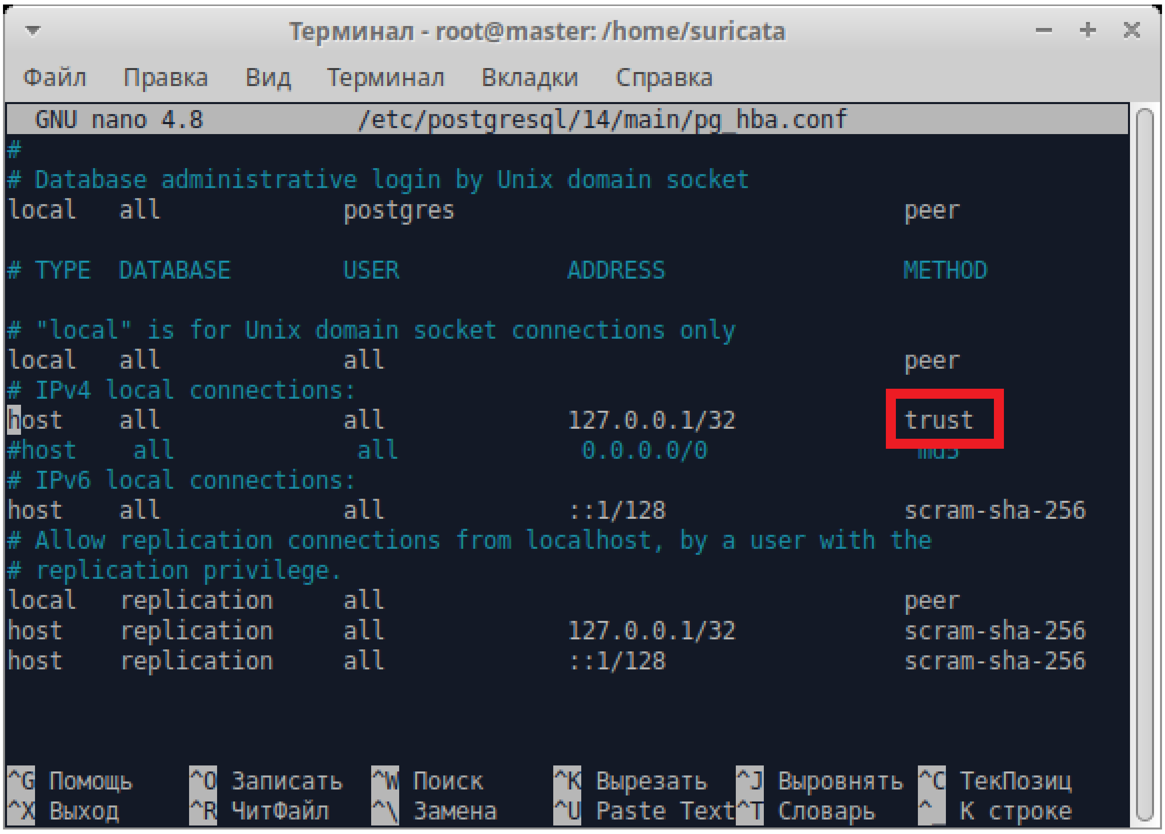


```
hduser@master:~$ hdfs dfs -mkdir /hive /hive/warehouse
hduser@master:~$ hdfs dfs -chmod -R 775 /hive
hduser@master:~$ hdfs dfs -chown -R hive:hduser /hive
```

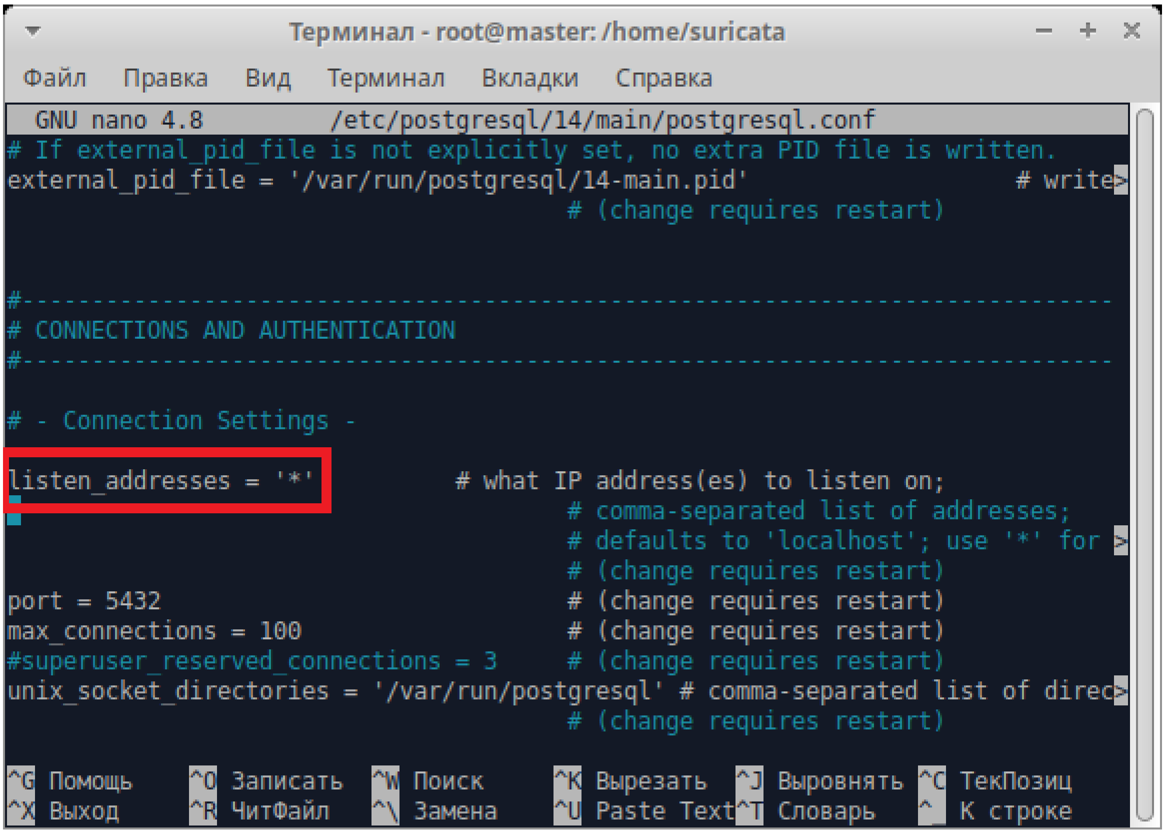
Установка БД PostgreSQL

```
echo "deb http://apt.postgresql.org/pub/repos/apt/ buster-pgdg main" >> /etc/apt/sources.list
wget --quiet -O - https://www.postgresql.org/media/keys/ACCC4CF8.asc | apt-key add -
apt-get update
apt-get install -y postgresql
service postgresql restart
```

Далее выполним редактирование конфигурационных файлов

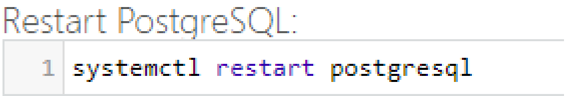


```
  GNU nano 4.8             /etc/postgresql/14/main/pg_hba.conf
#
# Database administrative login by Unix domain socket
local   all             postgres                                peer

# TYPE  DATABASE        USER            ADDRESS                 METHOD

# "local" is for Unix domain socket connections only
local   all             all                                     peer
# IPv4 local connections:
host    all             all             127.0.0.1/32            trust
#host    all             all             0.0.0.0/0               md5
# IPv6 local connections:
host    all             all             ::1/128                 scram-sha-256
# Allow replication connections from localhost, by a user with the
# replication privilege.
local   replication     all                                     peer
host    replication     all             127.0.0.1/32            scram-sha-256
host    replication     all             ::1/128                 scram-sha-256



^G Помощь     ^O Записать   ^W Поиск      ^K Вырезать   ^J Выровнять  ^C ТекПозиц
^X Выход      ^R ЧитФайл    ^\ Замена     ^U Paste Text^T Словарь     ^_ К строке
```



```
  GNU nano 4.8          /etc/postgresql/14/main/postgresql.conf
# If external_pid_file is not explicitly set, no extra PID file is written.
external_pid_file = '/var/run/postgresql/14-main.pid'                   # write>
                                        # (change requires restart)


#------------------------------------------------------------------------------
# CONNECTIONS AND AUTHENTICATION
#------------------------------------------------------------------------------

# - Connection Settings -

listen_addresses = '*'          # what IP address(es) to listen on;
                                        # comma-separated list of addresses;
                                        # defaults to 'localhost'; use '*' for >
                                        # (change requires restart)
port = 5432                             # (change requires restart)
max_connections = 100                   # (change requires restart)
#superuser_reserved_connections = 3     # (change requires restart)
unix_socket_directories = '/var/run/postgresql' # comma-separated list of direc>
                                        # (change requires restart)

^G Помощь     ^O Записать   ^W Поиск      ^K Вырезать   ^J Выровнять  ^C ТекПозиц
^X Выход      ^R ЧитФайл    ^\ Замена     ^U Paste Text^T Словарь     ^_ К строке
```

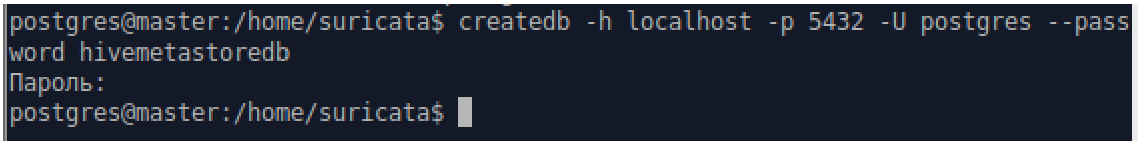
Выполним рестарт PostgreSQL


Restart PostgreSQL:

```
systemctl restart postgresql
```

Создание БД Hive metastore database (PostgreSQL)

su - postgres

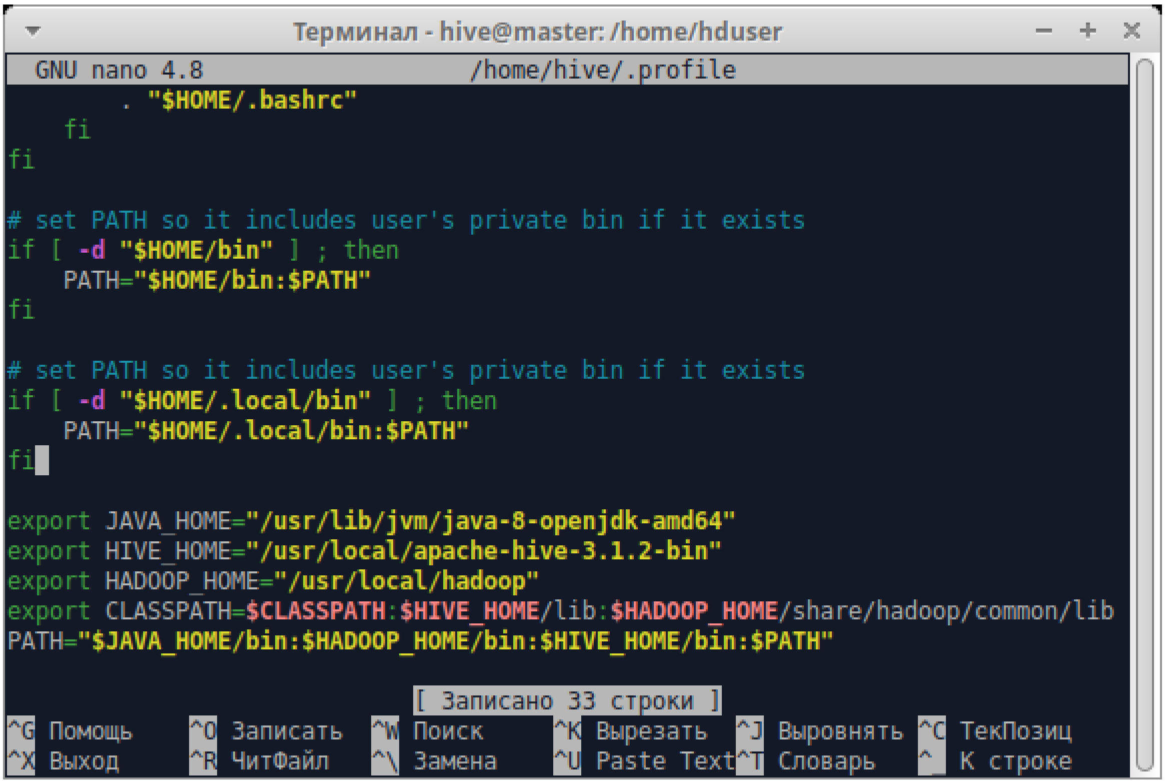


```
postgres@master:/home/suricata$ createdb -h localhost -p 5432 -U postgres --pass
word hivemetastoredb
Пароль:
postgres@master:/home/suricata$ 
```

Обновление файла «~/.profile»

su hive

nano ~/.profile

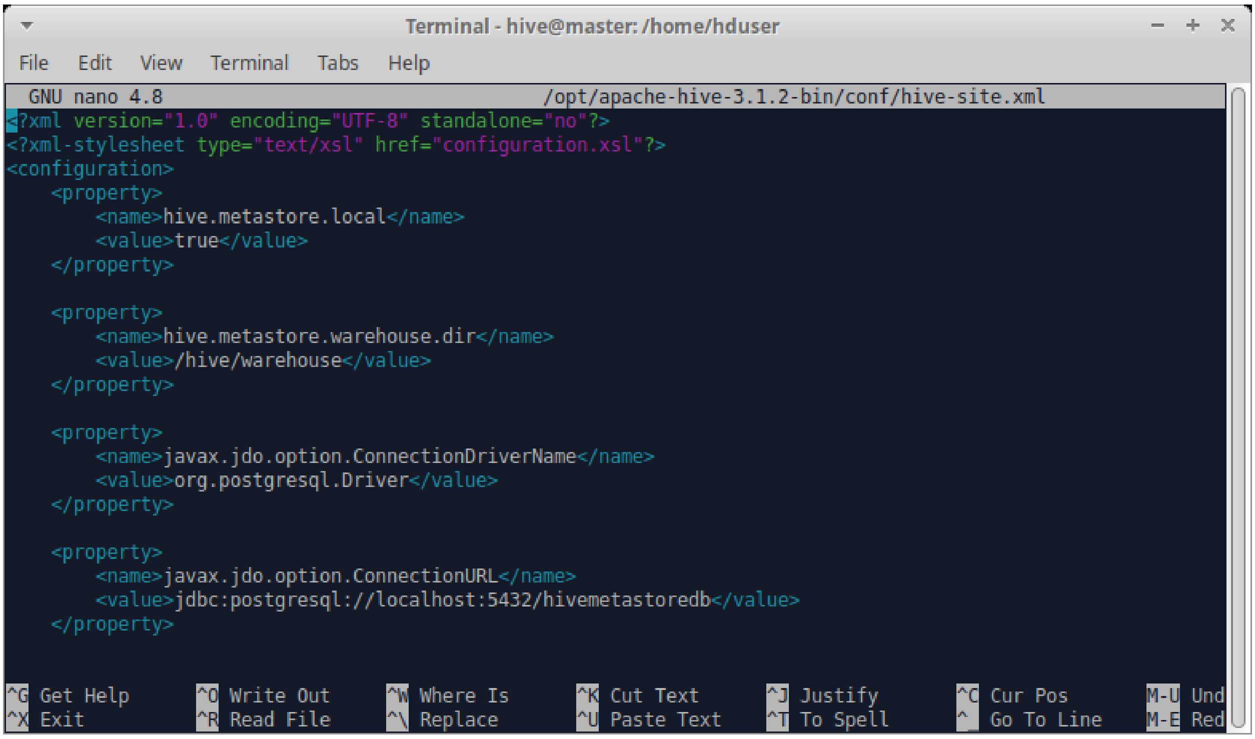


```
  GNU nano 4.8                     /home/hive/.profile
        . "$HOME/.bashrc"
    fi
fi

# set PATH so it includes user's private bin if it exists
if [ -d "$HOME/bin" ] ; then
    PATH="$HOME/bin:$PATH"
fi

# set PATH so it includes user's private bin if it exists
if [ -d "$HOME/.local/bin" ] ; then
    PATH="$HOME/.local/bin:$PATH"
fi

export JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64"
export HIVE_HOME="/usr/local/apache-hive-3.1.2-bin"
export HADOOP_HOME="/usr/local/hadoop"
export CLASSPATH=$CLASSPATH:$HIVE_HOME/lib:$HADOOP_HOME/share/hadoop/common/lib
PATH="$JAVA_HOME/bin:$HADOOP_HOME/bin:$HIVE_HOME/bin:$PATH"

                              [ Записано 33 строки ]
^G Помощь     ^O Записать   ^W Поиск      ^K Вырезать   ^J Выровнять  ^C ТекПозиц
^X Выход      ^R ЧитФайл    ^\ Замена     ^U Paste Text^T Словарь    ^_ К строке
```

source ~/.profile

Редактирование файла «${HIVE_HOME}/conf/hive-site.xml»

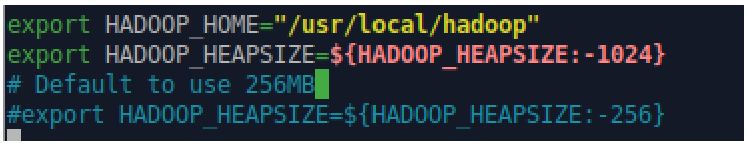
nano ${HIVE_HOME}/conf/hive-site.xml



```
  GNU nano 4.8                        /opt/apache-hive-3.1.2-bin/conf/hive-site.xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
    <property>
        <name>hive.metastore.local</name>
        <value>true</value>
    </property>

    <property>
        <name>hive.metastore.warehouse.dir</name>
        <value>/hive/warehouse</value>
    </property>

    <property>
        <name>javax.jdo.option.ConnectionDriverName</name>
        <value>org.postgresql.Driver</value>
    </property>

    <property>
        <name>javax.jdo.option.ConnectionURL</name>
        <value>jdbc:postgresql://localhost:5432/hivemetastoredb</value>
    </property>

^G Get Help     ^O Write Out    ^W Where Is     ^K Cut Text     ^J Justify      ^C Cur Pos      M-U Und
^X Exit         ^R Read File    ^\ Replace      ^U Paste Text   ^T To Spell     ^_ Go To Line   M-E Red
```

Редактирование файла «${HIVE_HOME}/bin/hive-config.sh»

nano ${HIVE_HOME}/bin/hive-config.sh

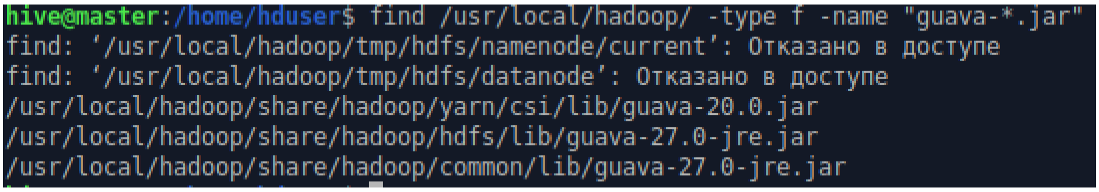


```
export HADOOP_HOME="/usr/local/hadoop"
export HADOOP_HEAPSIZE=${HADOOP_HEAPSIZE:-1024}
# Default to use 256MB
#export HADOOP_HEAPSIZE=${HADOOP_HEAPSIZE:-256}
```

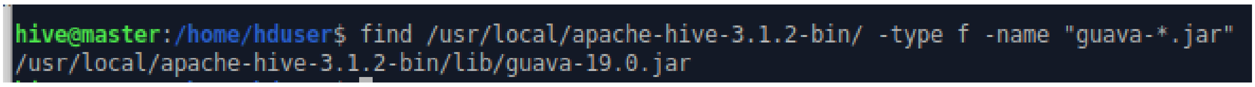
Создание схемы Hive (PostgreSQL)

${HIVE_HOME}/bin/schematool -initSchema -dbType postgres

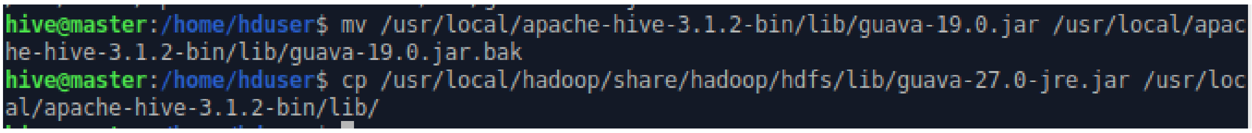
При выполнении команды может появиться ошибка, поэтому заранее выполним исправление согласно инструкции.



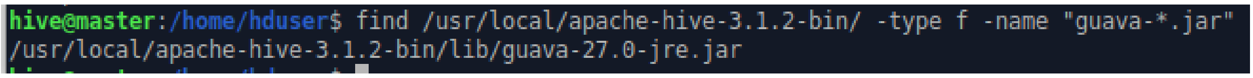
```
hive@master:/home/hduser$ find /usr/local/hadoop/ -type f -name "guava-*.jar"
find: '/usr/local/hadoop/tmp/hdfs/namenode/current': Отказано в доступе
find: '/usr/local/hadoop/tmp/hdfs/datanode': Отказано в доступе
/usr/local/hadoop/share/hadoop/yarn/csi/lib/guava-20.0.jar
/usr/local/hadoop/share/hadoop/hdfs/lib/guava-27.0-jre.jar
/usr/local/hadoop/share/hadoop/common/lib/guava-27.0-jre.jar
```



```
hive@master:/home/hduser$ find /usr/local/apache-hive-3.1.2-bin/ -type f -name "guava-*.jar"
/usr/local/apache-hive-3.1.2-bin/lib/guava-19.0.jar
```

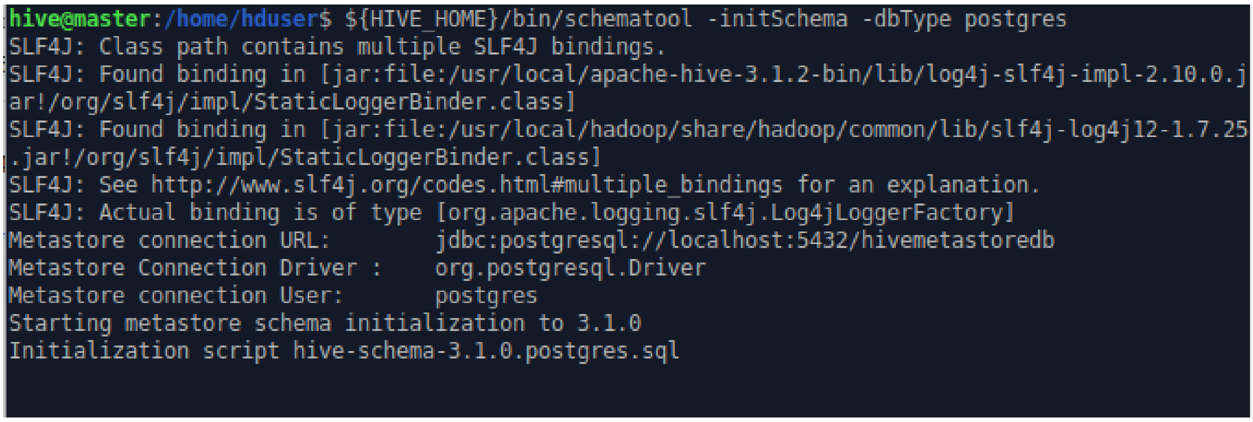


```
hive@master:/home/hduser$ mv /usr/local/apache-hive-3.1.2-bin/lib/guava-19.0.jar /usr/local/apac
he-hive-3.1.2-bin/lib/guava-19.0.jar.bak
hive@master:/home/hduser$ cp /usr/local/hadoop/share/hadoop/hdfs/lib/guava-27.0-jre.jar /usr/loc
al/apache-hive-3.1.2-bin/lib/
```

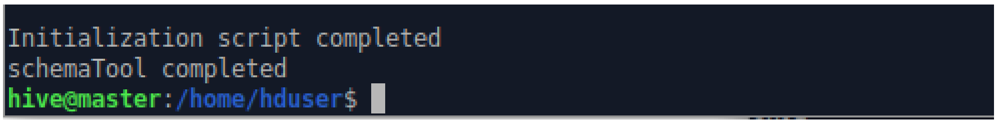


```
hive@master:/home/hduser$ find /usr/local/apache-hive-3.1.2-bin/ -type f -name "guava-*.jar"
/usr/local/apache-hive-3.1.2-bin/lib/guava-27.0-jre.jar
```

Создаем схему

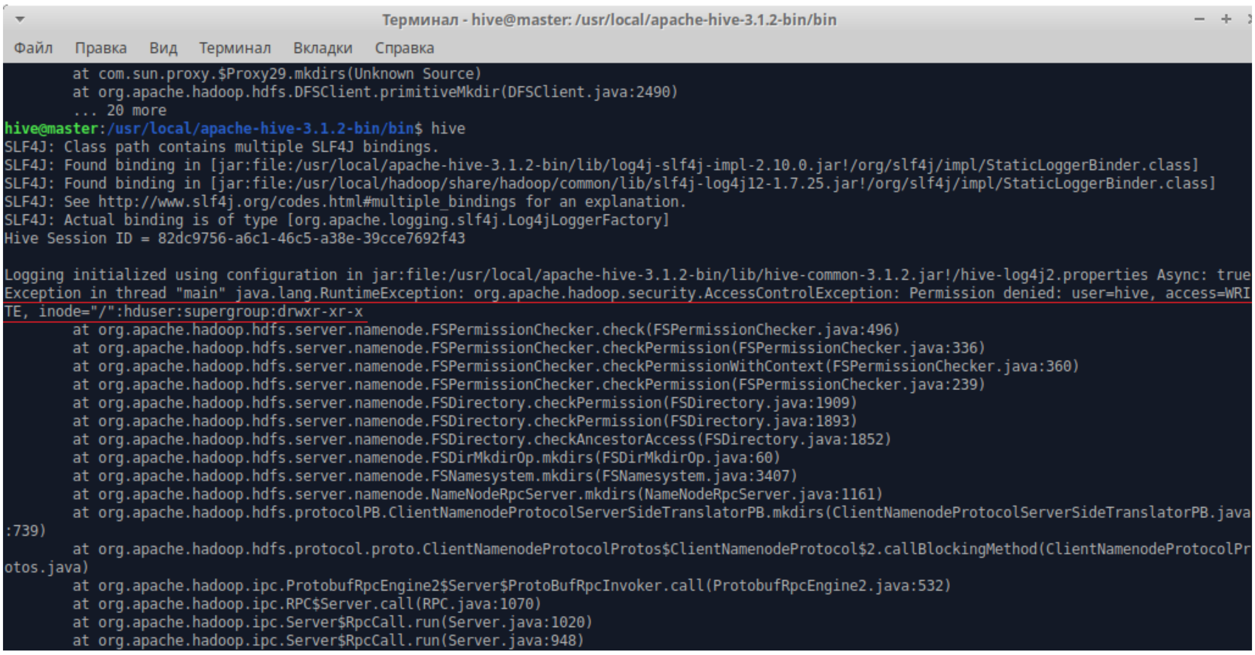


```
hive@master:/home/hduser$ ${HIVE_HOME}/bin/schematool -initSchema -dbType postgres
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.j
ar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-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]
Metastore connection URL:        jdbc:postgresql://localhost:5432/hivemetastoredb
Metastore Connection Driver :    org.postgresql.Driver
Metastore connection User:       postgres
Starting metastore schema initialization to 3.1.0
Initialization script hive-schema-3.1.0.postgres.sql
```



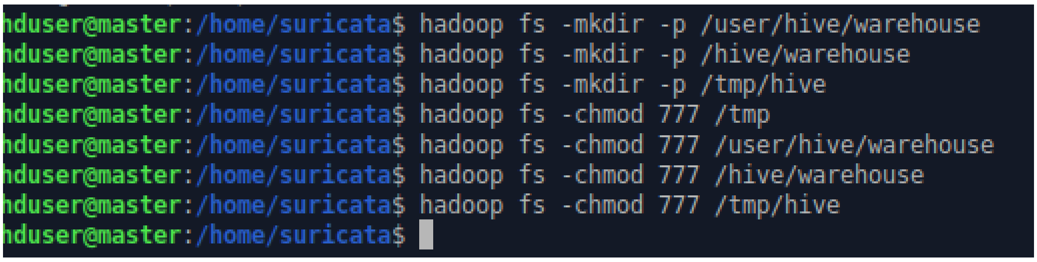
```
Initialization script completed
schemaTool completed
hive@master:/home/hduser$
```

При запуске Hive возникала ошибка, связанная с ограничением доступа пользователю hive.

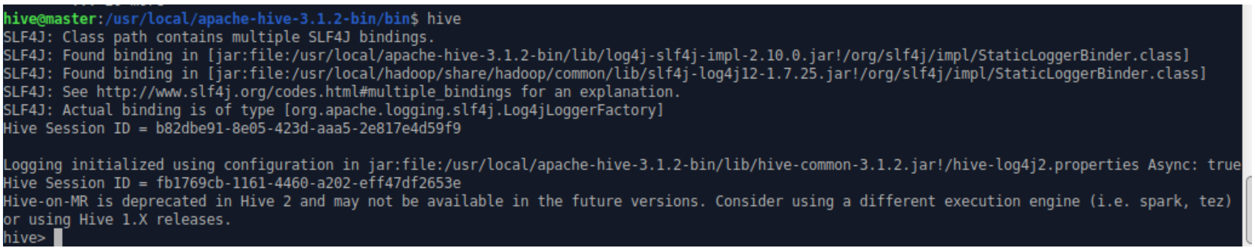

Терминал - hive@master: /usr/local/apache-hive-3.1.2-bin/bin
Файл Правка Вид Терминал Вкладки Справка

```
        at com.sun.proxy.$Proxy29.mkdirs(Unknown Source)
        at org.apache.hadoop.hdfs.DFSClient.primitiveMkdir(DFSClient.java:2490)
        ... 20 more
hive@master:/usr/local/apache-hive-3.1.2-bin/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-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 = 82dc9756-a6c1-46c5-a38e-39cce7692f43

Logging initialized using configuration in jar:file:/usr/local/apache-hive-3.1.2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
Exception in thread "main" java.lang.RuntimeException: org.apache.hadoop.security.AccessControlException: Permission denied: user=hive, access=WRI
TE, inode="/":hduser:supergroup:drwxr-xr-x
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.check(FSPermissionChecker.java:496)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:336)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermissionWithContext(FSPermissionChecker.java:360)
        at org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:239)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1909)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkPermission(FSDirectory.java:1893)
        at org.apache.hadoop.hdfs.server.namenode.FSDirectory.checkAncestorAccess(FSDirectory.java:1852)
        at org.apache.hadoop.hdfs.server.namenode.FSDirMkdirOp.mkdirs(FSDirMkdirOp.java:60)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.mkdirs(FSNamesystem.java:3407)
        at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.mkdirs(NameNodeRpcServer.java:1161)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.mkdirs(ClientNamenodeProtocolServerSideTranslatorPB.java
:739)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingMethod(ClientNamenodeProtocolPr
otos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine2.java:532)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:1070)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:1020)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:948)
```

Решение проблемы с запуском Hive

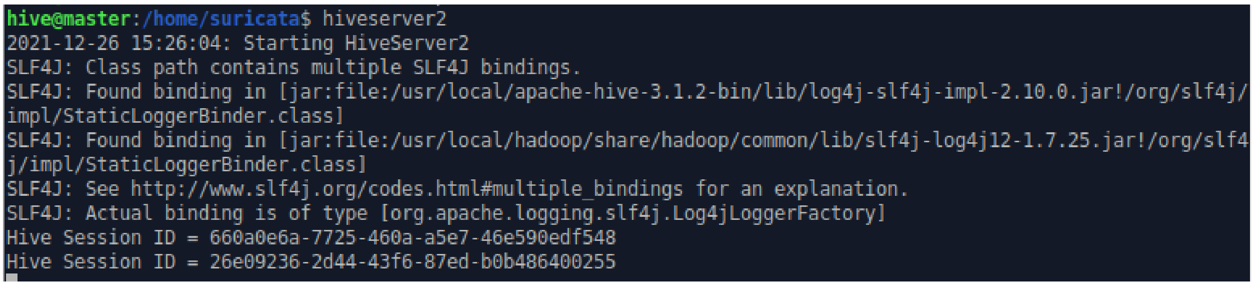


```
hduser@master:/home/suricata$ hadoop fs -mkdir -p /user/hive/warehouse
hduser@master:/home/suricata$ hadoop fs -mkdir -p /hive/warehouse
hduser@master:/home/suricata$ hadoop fs -mkdir -p /tmp/hive
hduser@master:/home/suricata$ hadoop fs -chmod 777 /tmp
hduser@master:/home/suricata$ hadoop fs -chmod 777 /user/hive/warehouse
hduser@master:/home/suricata$ hadoop fs -chmod 777 /hive/warehouse
hduser@master:/home/suricata$ hadoop fs -chmod 777 /tmp/hive
hduser@master:/home/suricata$
```

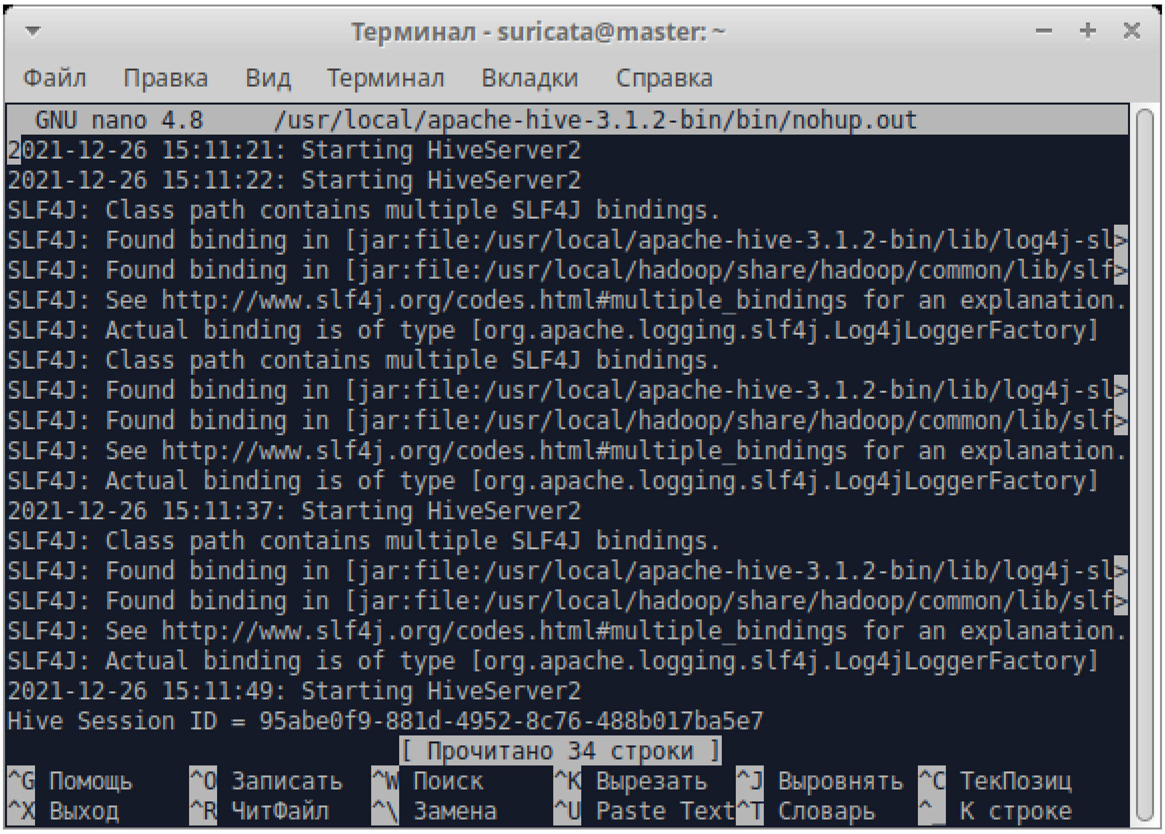


```
hive@master:/usr/local/apache-hive-3.1.2-bin/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/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 = b82dbe91-8e05-423d-aaa5-2e817e4d59f9

Logging initialized using configuration in jar:file:/usr/local/apache-hive-3.1.2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = fb1769cb-1161-4460-a202-eff47df2653e
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez)
or using Hive 1.X releases.
hive>
```

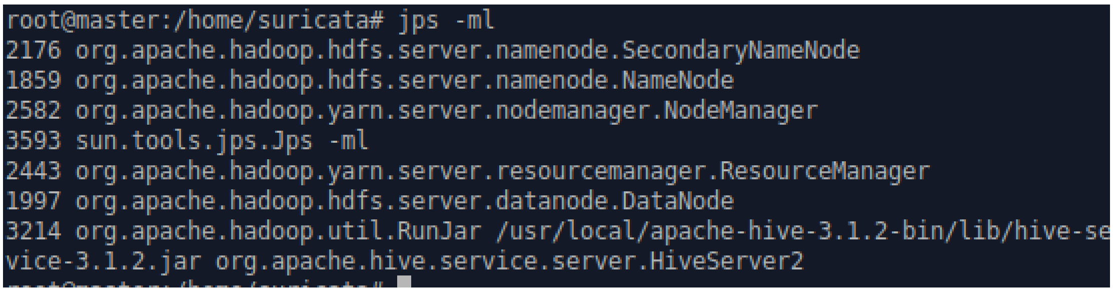
Start HiveServer2



```
hive@master:/home/suricata$ hiveserver2
2021-12-26 15:26:04: Starting HiveServer2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/
impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4
j/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 = 660a0e6a-7725-460a-a5e7-46e590edf548
Hive Session ID = 26e09236-2d44-43f6-87ed-b0b486400255
```

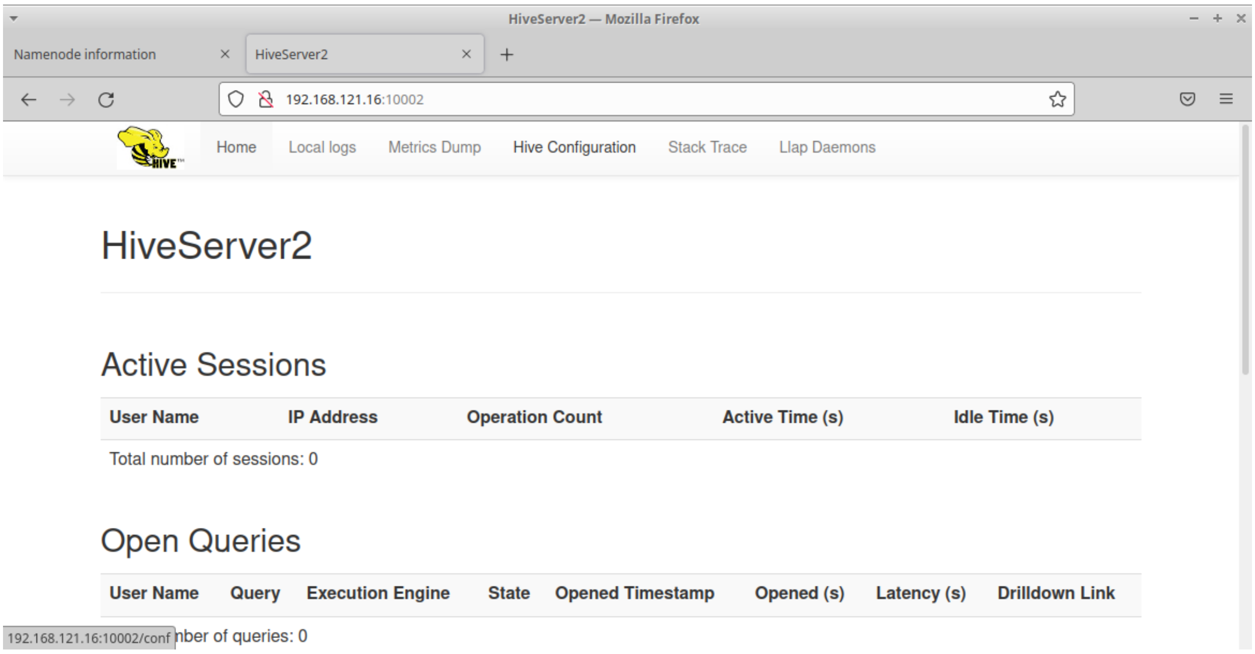

Терминал - suricata@master: ~
Файл Правка Вид Терминал Вкладки Справка

```
  GNU nano 4.8      /usr/local/apache-hive-3.1.2-bin/bin/nohup.out
2021-12-26 15:11:21: Starting HiveServer2
2021-12-26 15:11:22: Starting HiveServer2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-sl>
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf>
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-sl>
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf>
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
2021-12-26 15:11:37: Starting HiveServer2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-sl>
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf>
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
2021-12-26 15:11:49: Starting HiveServer2
Hive Session ID = 95abe0f9-881d-4952-8c76-488b017ba5e7
                            [ Прочитано 34 строки ]
^G Помощь      ^O Записать    ^W Поиск       ^K Вырезать    ^J Выровнять   ^C ТекПозиц
^X Выход       ^R ЧитФайл     ^\ Замена      ^U Paste Text^T Словарь     ^_ К строке
```



```
root@master:/home/suricata# jps -ml
2176 org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode
1859 org.apache.hadoop.hdfs.server.namenode.NameNode
2582 org.apache.hadoop.yarn.server.nodemanager.NodeManager
3593 sun.tools.jps.Jps -ml
2443 org.apache.hadoop.yarn.server.resourcemanager.ResourceManager
1997 org.apache.hadoop.hdfs.server.datanode.DataNode
3214 org.apache.hadoop.util.RunJar /usr/local/apache-hive-3.1.2-bin/lib/hive-se
vice-3.1.2.jar org.apache.hive.service.server.HiveServer2
```

http://192.168.121.16:10002/



HiveServer2

Active Sessions

User Name	IP Address	Operation Count	Active Time (s)	Idle Time (s)

Total number of sessions: 0

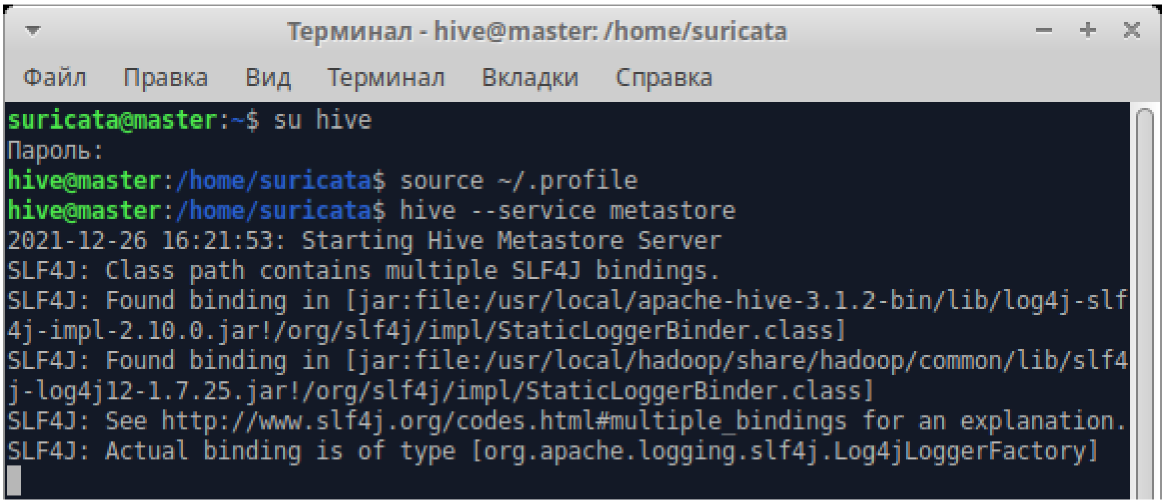
Open Queries

User Name	Query	Execution Engine	State	Opened Timestamp	Opened (s)	Latency (s)	Drilldown Link

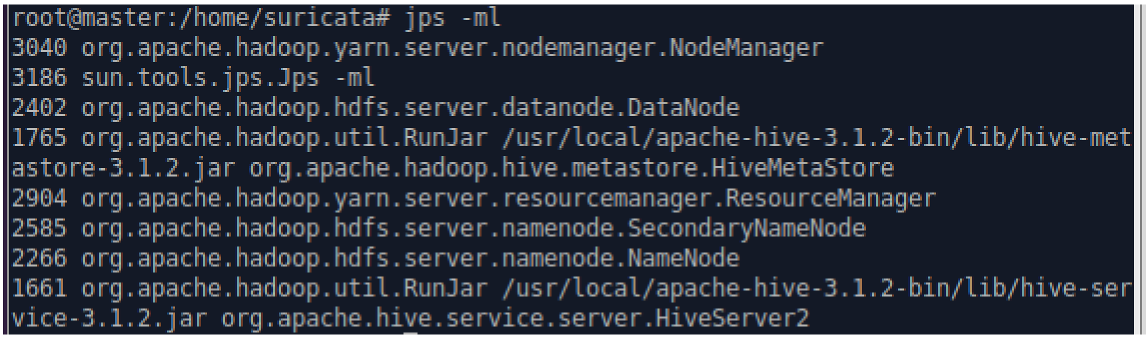
nber of queries: 0

Start Hive MetaStore

hive --service metastore

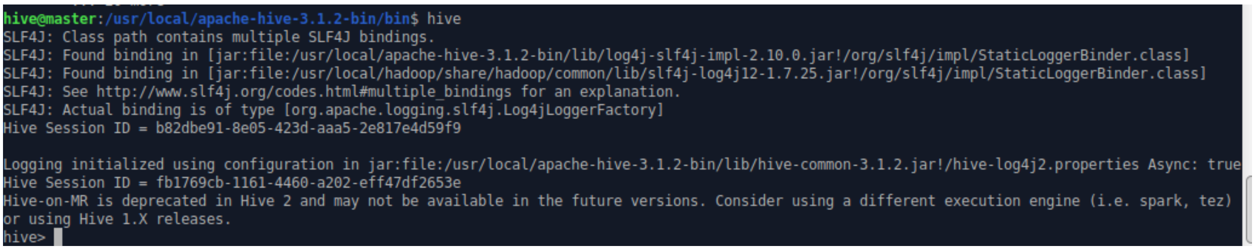


```
suricata@master:~$ su hive
Пароль:
hive@master:/home/suricata$ source ~/.profile
hive@master:/home/suricata$ hive --service metastore
2021-12-26 16:21:53: Starting Hive Metastore Server
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-slf
4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4
j-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]
```

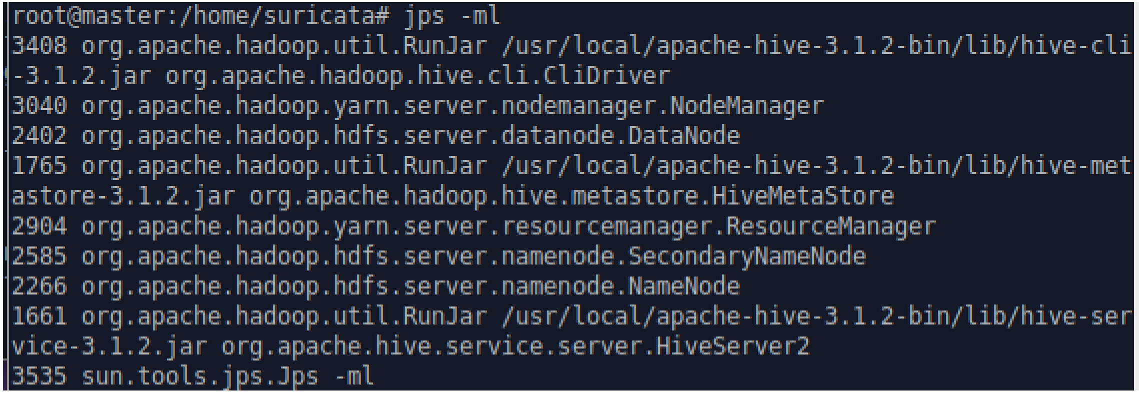


```
root@master:/home/suricata# jps -ml
3040 org.apache.hadoop.yarn.server.nodemanager.NodeManager
3186 sun.tools.jps.Jps -ml
2402 org.apache.hadoop.hdfs.server.datanode.DataNode
1765 org.apache.hadoop.util.RunJar /usr/local/apache-hive-3.1.2-bin/lib/hive-met
astore-3.1.2.jar org.apache.hadoop.hive.metastore.HiveMetaStore
2904 org.apache.hadoop.yarn.server.resourcemanager.ResourceManager
2585 org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode
2266 org.apache.hadoop.hdfs.server.namenode.NameNode
1661 org.apache.hadoop.util.RunJar /usr/local/apache-hive-3.1.2-bin/lib/hive-ser
vice-3.1.2.jar org.apache.hive.service.server.HiveServer2
```

2. Войти под пользователем hive и запустить консольную утилиту hive



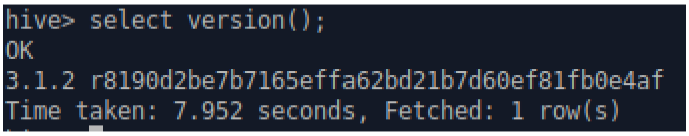
```
hive@master:/usr/local/apache-hive-3.1.2-bin/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/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 = b82dbe91-8e05-423d-aaa5-2e817e4d59f9

Logging initialized using configuration in jar:file:/usr/local/apache-hive-3.1.2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = fb1769cb-1161-4460-a202-eff47df2653e
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez)
or using Hive 1.X releases.
hive>
```



```
root@master:/home/suricata# jps -ml
3408 org.apache.hadoop.util.RunJar /usr/local/apache-hive-3.1.2-bin/lib/hive-cli
-3.1.2.jar org.apache.hadoop.hive.cli.CliDriver
3040 org.apache.hadoop.yarn.server.nodemanager.NodeManager
2402 org.apache.hadoop.hdfs.server.datanode.DataNode
1765 org.apache.hadoop.util.RunJar /usr/local/apache-hive-3.1.2-bin/lib/hive-met
astore-3.1.2.jar org.apache.hadoop.hive.metastore.HiveMetaStore
2904 org.apache.hadoop.yarn.server.resourcemanager.ResourceManager
2585 org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode
2266 org.apache.hadoop.hdfs.server.namenode.NameNode
1661 org.apache.hadoop.util.RunJar /usr/local/apache-hive-3.1.2-bin/lib/hive-ser
vice-3.1.2.jar org.apache.hive.service.server.HiveServer2
3535 sun.tools.jps.Jps -ml
```

3. Выполнить команду select version();

4. Записать в отчёт полученный ответ

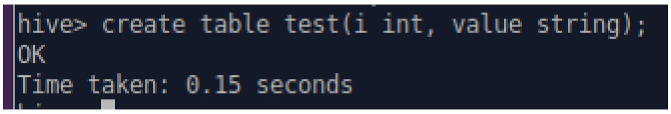


```
hive> select version();
OK
3.1.2 r8190d2be7b7165effa62bd21b7d60ef81fb0e4af
Time taken: 7.952 seconds, Fetched: 1 row(s)
```

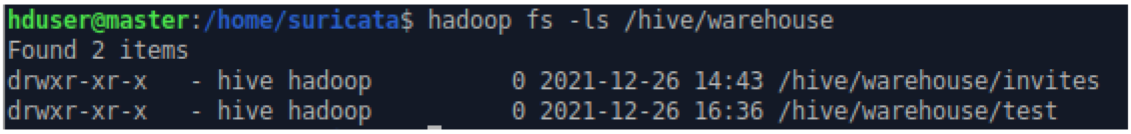
Задание 1

Задача познакомиться с базовыми командами HIVE.

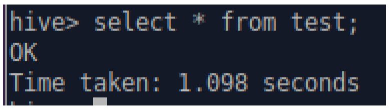
1. Воспроизведите примеры из лекции и сохраните скрипт в свой репозиторий



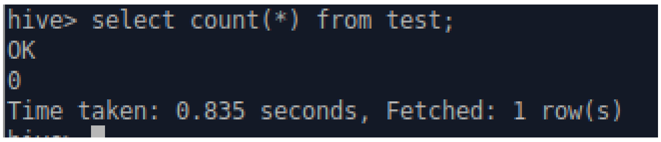
```
hive> create table test(i int, value string);
OK
Time taken: 0.15 seconds
```



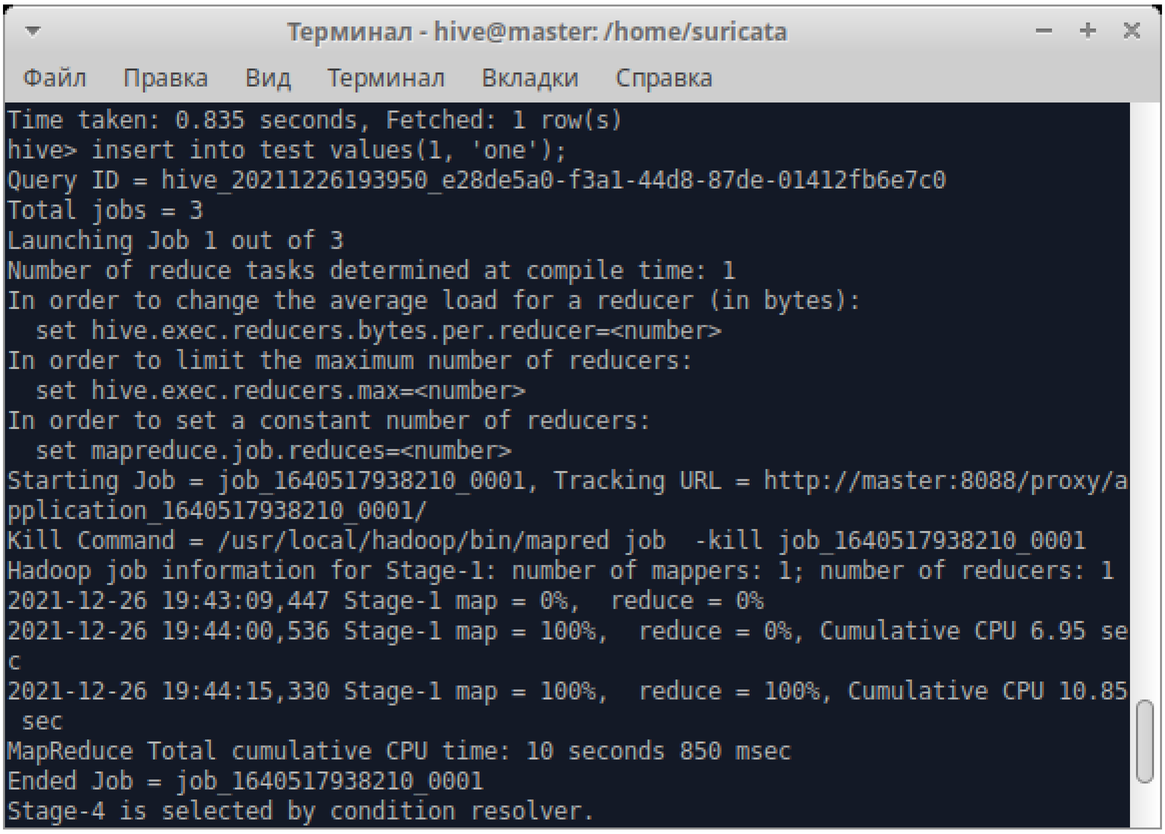
```
hduser@master:/home/suricata$ hadoop fs -ls /hive/warehouse
Found 2 items
drwxr-xr-x   - hive hadoop          0 2021-12-26 14:43 /hive/warehouse/invites
drwxr-xr-x   - hive hadoop          0 2021-12-26 16:36 /hive/warehouse/test
```



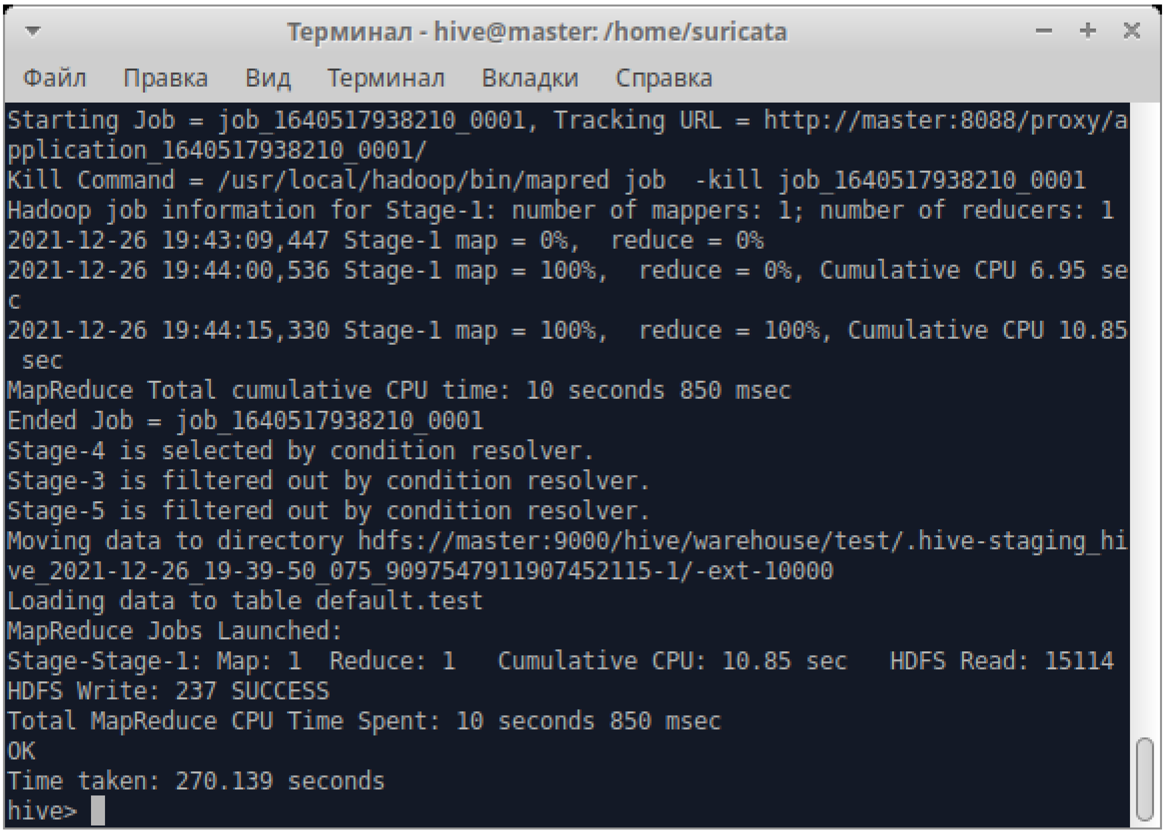
```
hive> select * from test;
OK
Time taken: 1.098 seconds
```



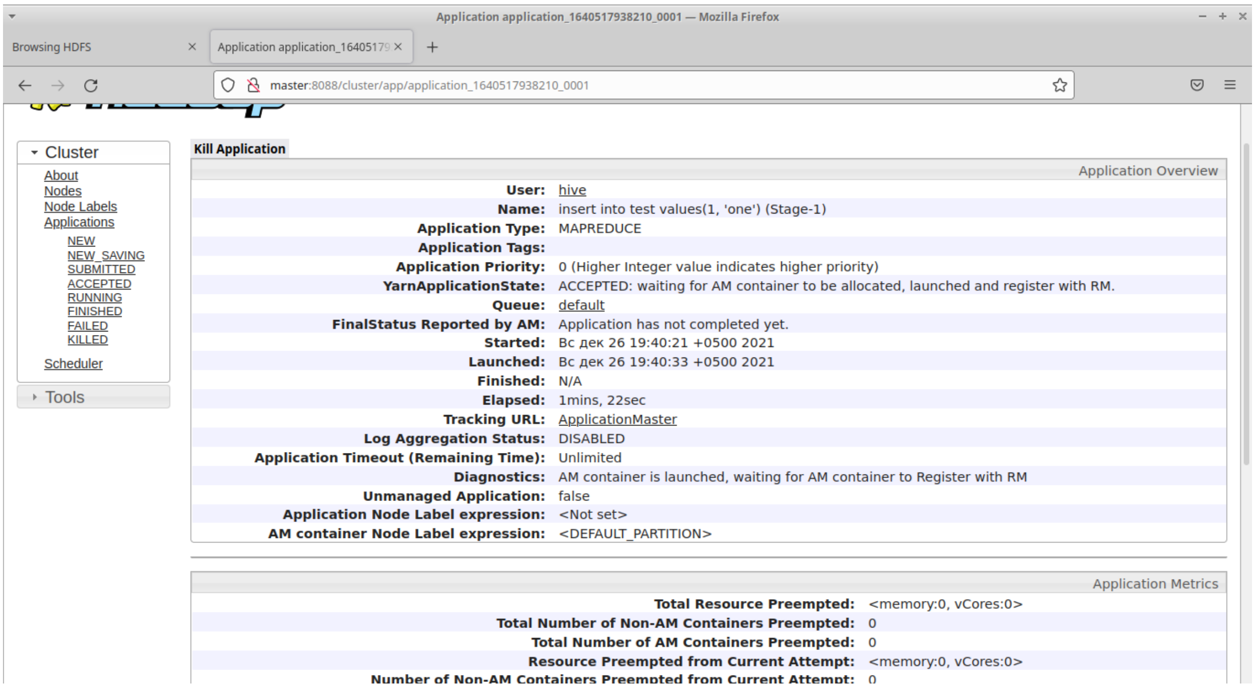
```
hive> select count(*) from test;
OK
0
Time taken: 0.835 seconds, Fetched: 1 row(s)
```



```
Time taken: 0.835 seconds, Fetched: 1 row(s)
hive> insert into test values(1, 'one');
Query ID = hive_20211226193950_e28de5a0-f3a1-44d8-87de-01412fb6e7c0
Total jobs = 3
Launching Job 1 out of 3
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>
Starting Job = job_1640517938210_0001, Tracking URL = http://master:8088/proxy/a
pplication_1640517938210_0001/
Kill Command = /usr/local/hadoop/bin/mapred job  -kill job_1640517938210_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-12-26 19:43:09,447 Stage-1 map = 0%,  reduce = 0%
2021-12-26 19:44:00,536 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.95 se
c
2021-12-26 19:44:15,330 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.85
 sec
MapReduce Total cumulative CPU time: 10 seconds 850 msec
Ended Job = job_1640517938210_0001
Stage-4 is selected by condition resolver.
```



```
Starting Job = job_1640517938210_0001, Tracking URL = http://master:8088/proxy/a
pplication_1640517938210_0001/
Kill Command = /usr/local/hadoop/bin/mapred job  -kill job_1640517938210_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-12-26 19:43:09,447 Stage-1 map = 0%,  reduce = 0%
2021-12-26 19:44:00,536 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.95 se
c
2021-12-26 19:44:15,330 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.85
 sec
MapReduce Total cumulative CPU time: 10 seconds 850 msec
Ended Job = job_1640517938210_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://master:9000/hive/warehouse/test/.hive-staging_hi
ve_2021-12-26_19-39-50_075_9097547911907452115-1/-ext-10000
Loading data to table default.test
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.85 sec   HDFS Read: 15114
HDFS Write: 237 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 850 msec
OK
Time taken: 270.139 seconds
hive>
```

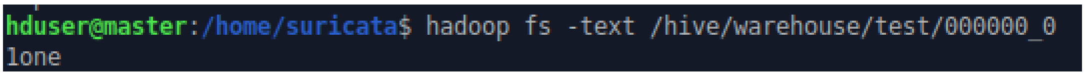

Application application_1640517938210_0001 — Mozilla Firefox
Browsing HDFS
Application application_1640517938
master:8088/cluster/app/application_1640517938210_0001
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Kill Application

Application Overview

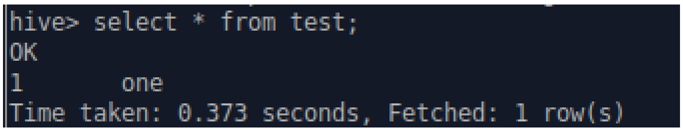
User: hive
Name: insert into test values(1, 'one') (Stage-1)
Application Type: MAPREDUCE
Application Tags:
Application Priority: 0 (Higher Integer value indicates higher priority)
YarnApplicationState: ACCEPTED: waiting for AM container to be allocated, launched and register with RM.
Queue: default
FinalStatus Reported by AM: Application has not completed yet.
Started: Вс дек 26 19:40:21 +0500 2021
Launched: Вс дек 26 19:40:33 +0500 2021
Finished: N/A
Elapsed: 1mins, 22sec
Tracking URL: ApplicationMaster
Log Aggregation Status: DISABLED
Application Timeout (Remaining Time): Unlimited
Diagnostics: AM container is launched, waiting for AM container to Register with RM
Unmanaged Application: false
Application Node Label expression: <Not set>
AM container Node Label expression: <DEFAULT_PARTITION>

Application Metrics

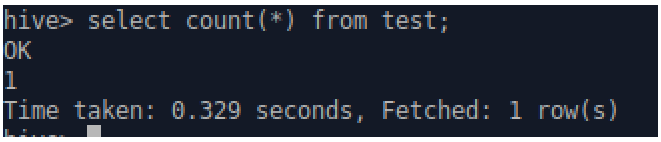
Total Resource Preempted: <memory:0, vCores:0>
Total Number of Non-AM Containers Preempted: 0
Total Number of AM Containers Preempted: 0
Resource Preempted from Current Attempt: <memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt: 0



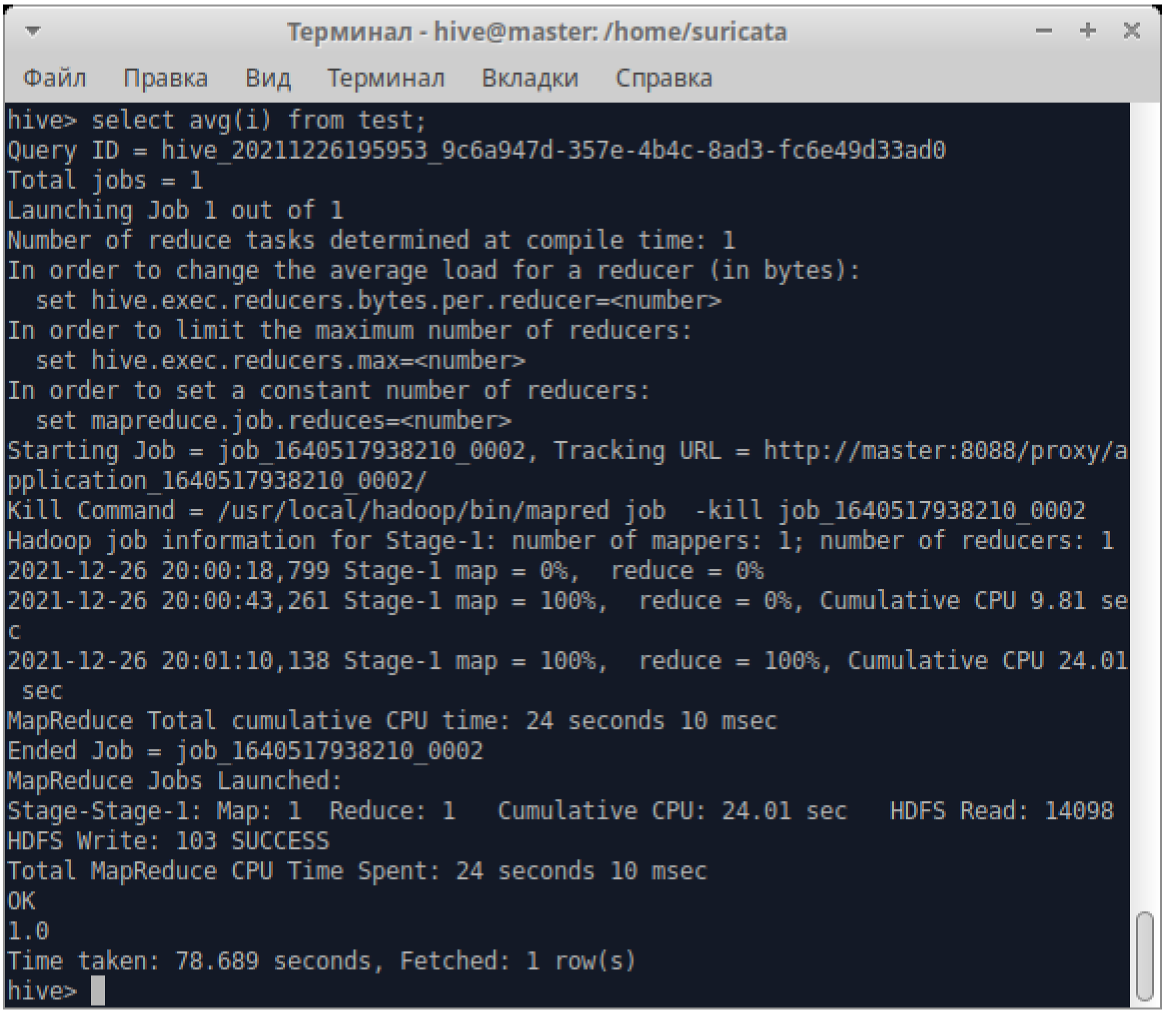
```
hduser@master:/home/suricata$ hadoop fs -text /hive/warehouse/test/000000_0
1one
```



```
hive> select * from test;
OK
1       one
Time taken: 0.373 seconds, Fetched: 1 row(s)
```



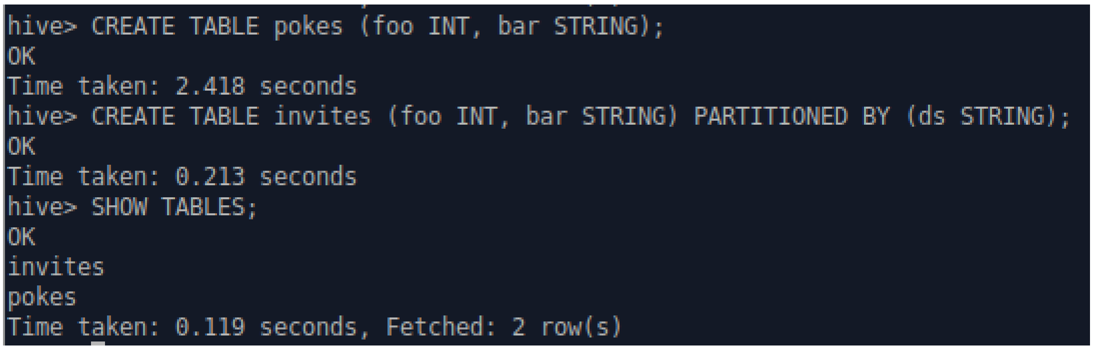
```
hive> select count(*) from test;
OK
1
Time taken: 0.329 seconds, Fetched: 1 row(s)
```


Терминал - hive@master: /home/suricata
Файл Правка Вид Терминал Вкладки Справка

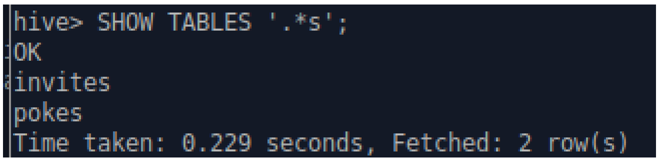
```
hive> select avg(i) from test;
Query ID = hive_20211226195953_9c6a947d-357e-4b4c-8ad3-fc6e49d33ad0
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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>
Starting Job = job_1640517938210_0002, Tracking URL = http://master:8088/proxy/a
pplication_1640517938210_0002/
Kill Command = /usr/local/hadoop/bin/mapred job  -kill job_1640517938210_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-12-26 20:00:18,799 Stage-1 map = 0%,  reduce = 0%
2021-12-26 20:00:43,261 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.81 se
c
2021-12-26 20:01:10,138 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 24.01
 sec
MapReduce Total cumulative CPU time: 24 seconds 10 msec
Ended Job = job_1640517938210_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 24.01 sec   HDFS Read: 14098
HDFS Write: 103 SUCCESS
Total MapReduce CPU Time Spent: 24 seconds 10 msec
OK
1.0
Time taken: 78.689 seconds, Fetched: 1 row(s)
hive>
```

2. Воспроизведите примеры из справки раздел DDL Operations и сохраните скрипты в свой репозиторий

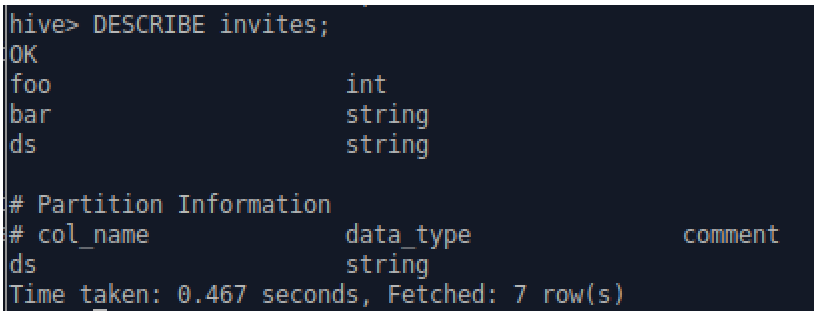
```
hive> CREATE TABLE pokes (foo INT, bar STRING);
hive> CREATE TABLE invites (foo INT, bar STRING) PARTITIONED BY (ds STRING);
hive> SHOW TABLES;
hive> SHOW TABLES '.*s';
hive> DESCRIBE invites;
```



```
hive> CREATE TABLE pokes (foo INT, bar STRING);
OK
Time taken: 2.418 seconds
hive> CREATE TABLE invites (foo INT, bar STRING) PARTITIONED BY (ds STRING);
OK
Time taken: 0.213 seconds
hive> SHOW TABLES;
OK
invites
pokes
Time taken: 0.119 seconds, Fetched: 2 row(s)
```

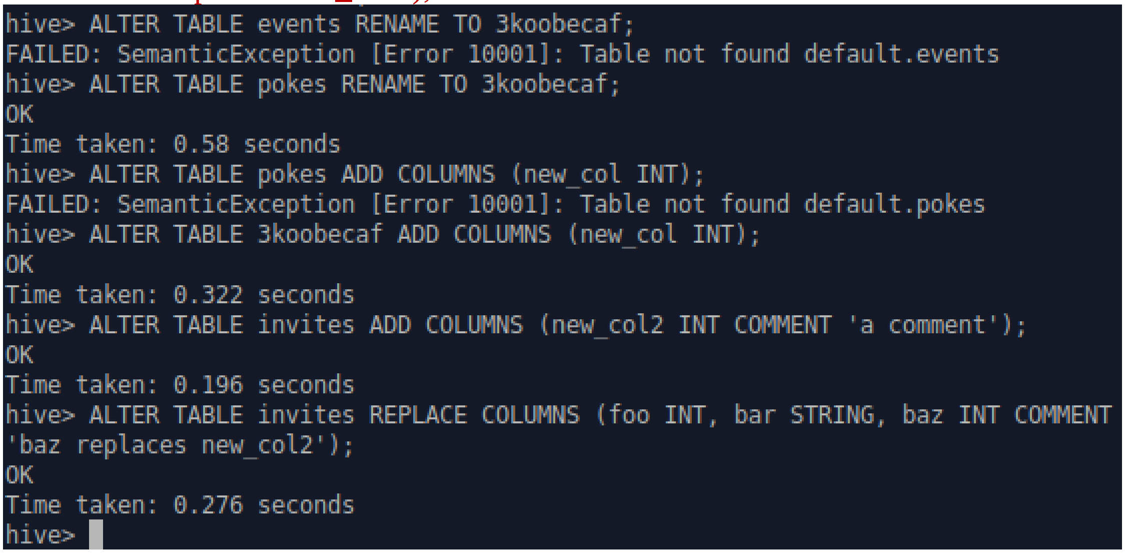


```
hive> SHOW TABLES '.*s';
OK
invites
pokes
Time taken: 0.229 seconds, Fetched: 2 row(s)
```



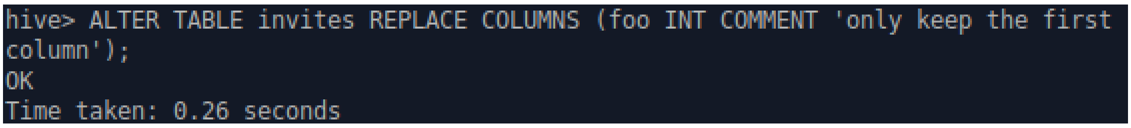
```
hive> DESCRIBE invites;
OK
foo                     int
bar                     string
ds                      string

# Partition Information
# col_name              data_type               comment
ds                      string
Time taken: 0.467 seconds, Fetched: 7 row(s)
```

hive> ALTER TABLE pokes RENAME TO 3koobecaf;
hive> ALTER TABLE 3koobecaf ADD COLUMNS (new_col INT);
hive> ALTER TABLE invites ADD COLUMNS (new_col2 INT COMMENT 'a comment');
hive> ALTER TABLE invites REPLACE COLUMNS (foo INT, bar STRING, baz INT COMMENT 'baz replaces new_col2');



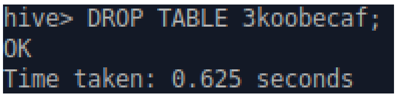
```
hive> ALTER TABLE events RENAME TO 3koobecaf;
FAILED: SemanticException [Error 10001]: Table not found default.events
hive> ALTER TABLE pokes RENAME TO 3koobecaf;
OK
Time taken: 0.58 seconds
hive> ALTER TABLE pokes ADD COLUMNS (new_col INT);
FAILED: SemanticException [Error 10001]: Table not found default.pokes
hive> ALTER TABLE 3koobecaf ADD COLUMNS (new_col INT);
OK
Time taken: 0.322 seconds
hive> ALTER TABLE invites ADD COLUMNS (new_col2 INT COMMENT 'a comment');
OK
Time taken: 0.196 seconds
hive> ALTER TABLE invites REPLACE COLUMNS (foo INT, bar STRING, baz INT COMMENT
'baz replaces new_col2');
OK
Time taken: 0.276 seconds
hive>
```

hive> ALTER TABLE invites REPLACE COLUMNS (foo INT COMMENT 'only keep the first column');



```
hive> ALTER TABLE invites REPLACE COLUMNS (foo INT COMMENT 'only keep the first
column');
OK
Time taken: 0.26 seconds
```

hive> DROP TABLE 3koobecaf;



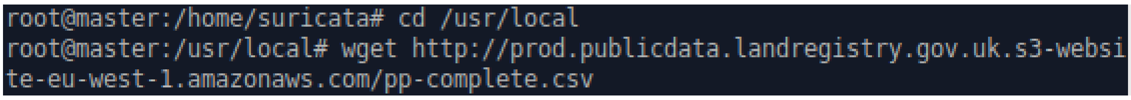
```
hive> DROP TABLE 3koobecaf;
OK
Time taken: 0.625 seconds
```

Задание 2

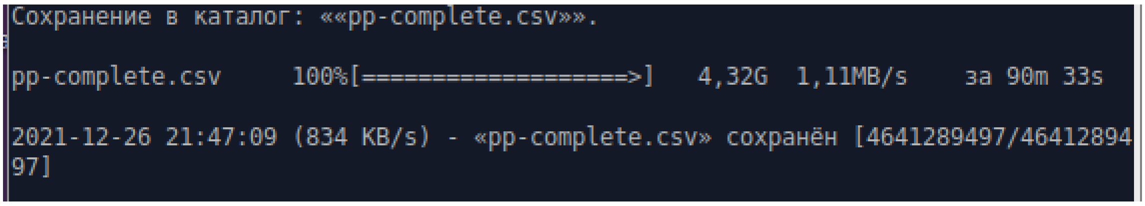
Загрузка данных в HIVE

1. Загрузите тестовый массив данных в текущую папку (файл большой и в облаке, может качаться долго).

wget http://prod.publicdata.landregistry.gov.uk.s3-website-eu-west-1.amazonaws.com/pp-complete.csv

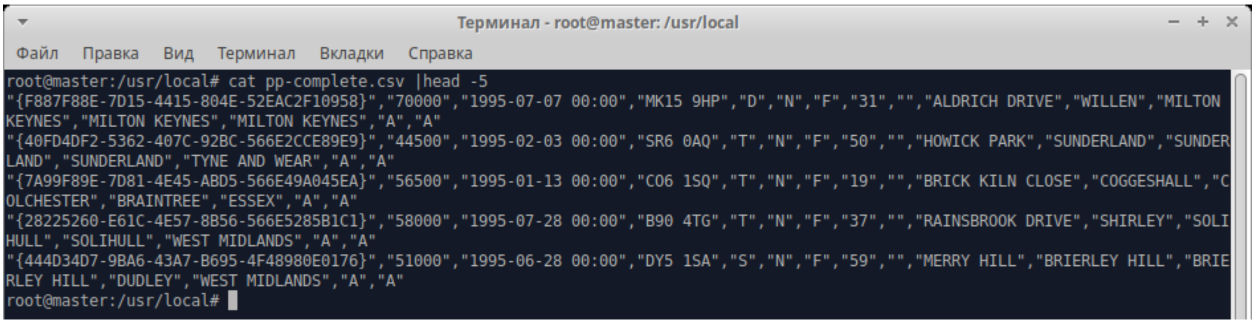


```
root@master:/home/suricata# cd /usr/local
root@master:/usr/local# wget http://prod.publicdata.landregistry.gov.uk.s3-websi
te-eu-west-1.amazonaws.com/pp-complete.csv
```



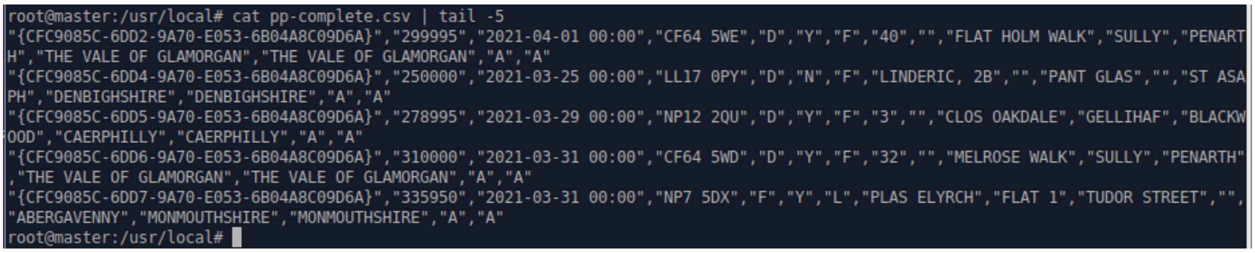
```
Сохранение в каталог: ««pp-complete.csv»».

pp-complete.csv     100%[===================>]   4,32G  1,11MB/s    за 90m 33s

2021-12-26 21:47:09 (834 KB/s) - «pp-complete.csv» сохранён [4641289497/46412894
97]
```

2. С помощью команд head и wc -l изучите его содержимое

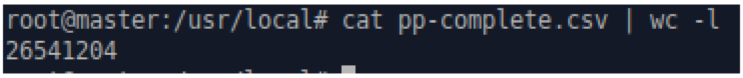


Терминал - root@master: /usr/local
Файл Правка Вид Терминал Вкладки Справка

```
root@master:/usr/local# cat pp-complete.csv |head -5
"{F887F88E-7D15-4415-804E-52EAC2F10958}","70000","1995-07-07 00:00","MK15 9HP","D","N","F","31","","ALDRICH DRIVE","WILLEN","MILTON
KEYNES","MILTON KEYNES","MILTON KEYNES","A","A"
"{40FD4DF2-5362-407C-92BC-566E2CCE89E9}","44500","1995-02-03 00:00","SR6 0AQ","T","N","F","50","","HOWICK PARK","SUNDERLAND","SUNDER
LAND","SUNDERLAND","TYNE AND WEAR","A","A"
"{7A99F89E-7D81-4E45-ABD5-566E49A045EA}","56500","1995-01-13 00:00","CO6 1SQ","T","N","F","19","","BRICK KILN CLOSE","COGGESHALL","C
OLCHESTER","BRAINTREE","ESSEX","A","A"
"{28225260-E61C-4E57-8B56-566E5285B1C1}","58000","1995-07-28 00:00","B90 4TG","T","N","F","37","","RAINSBROOK DRIVE","SHIRLEY","SOLI
HULL","SOLIHULL","WEST MIDLANDS","A","A"
"{444D34D7-9BA6-43A7-B695-4F48980E0176}","51000","1995-06-28 00:00","DY5 1SA","S","N","F","59","","MERRY HILL","BRIERLEY HILL","BRIE
RLEY HILL","DUDLEY","WEST MIDLANDS","A","A"
root@master:/usr/local# 
```



```
root@master:/usr/local# cat pp-complete.csv | tail -5
"{CFC9085C-6DD2-9A70-E053-6B04A8C09D6A}","299995","2021-04-01 00:00","CF64 5WE","D","Y","F","40","","FLAT HOLM WALK","SULLY","PENART
H","THE VALE OF GLAMORGAN","THE VALE OF GLAMORGAN","A","A"
"{CFC9085C-6DD4-9A70-E053-6B04A8C09D6A}","250000","2021-03-25 00:00","LL17 0PY","D","N","F","LINDERIC, 2B","","PANT GLAS","","ST ASA
PH","DENBIGHSHIRE","DENBIGHSHIRE","A","A"
"{CFC9085C-6DD5-9A70-E053-6B04A8C09D6A}","278995","2021-03-29 00:00","NP12 2QU","D","Y","F","3","","CLOS OAKDALE","GELLIHAF","BLACKW
OOD","CAERPHILLY","CAERPHILLY","A","A"
"{CFC9085C-6DD6-9A70-E053-6B04A8C09D6A}","310000","2021-03-31 00:00","CF64 5WD","D","Y","F","32","","MELROSE WALK","SULLY","PENARTH"
,"THE VALE OF GLAMORGAN","THE VALE OF GLAMORGAN","A","A"
"{CFC9085C-6DD7-9A70-E053-6B04A8C09D6A}","335950","2021-03-31 00:00","NP7 5DX","F","Y","L","PLAS ELYRCH","FLAT 1","TUDOR STREET","",
"ABERGAVENNY","MONMOUTHSHIRE","MONMOUTHSHIRE","A","A"
root@master:/usr/local# 
```



```
root@master:/usr/local# cat pp-complete.csv | wc -l
26541204
```

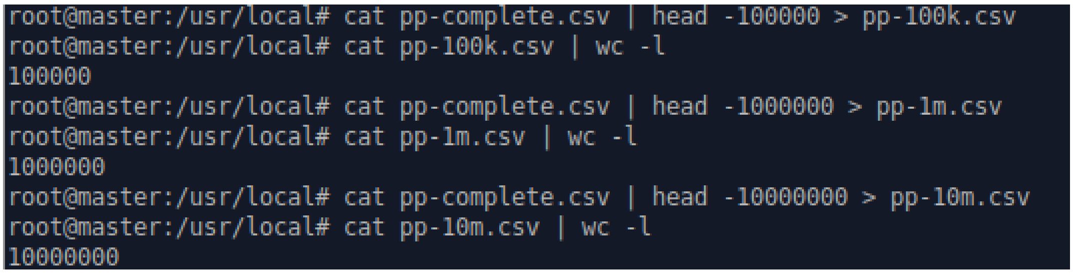
3. Сравните содержимое файла с описанием массива данных и подберите необходимые типы данных для колонок таблицы, перечень поддерживаемых типов данных приведён в справке

Data item	Explanation (where appropriate)	Data type BD
Transaction unique identifier	A reference number which is generated automatically recording each published sale. The number is unique and will change each time a sale is recorded.	String
Price	Sale price stated on the transfer deed.	Decimals / INT
Date of Transfer	Date when the sale was completed, as stated on the transfer deed.	TIMESTAMP
Postcode	This is the postcode used at the time of the original transaction. Note that postcodes can be reallocated and these changes are not reflected in the Price Paid Dataset.	String
Property Type	D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other Note that: - we only record the above categories to describe property type, we do not separately identify bungalows - end-of-terrace properties are included in the Terraced category above	String

	- ‘Other’ is only valid where the transaction relates to a property type that is not covered by existing values, for example where a property comprises more than one large parcel of land	
Old/New	Indicates the age of the property and applies to all price paid transactions, residential and non-residential. Y = a newly built property, N = an established residential building	String
Duration	Relates to the tenure: F = Freehold, L= Leasehold etc. Note that HM Land Registry does not record leases of 7 years or less in the Price Paid Dataset.	String
PAON	Primary Addressable Object Name. Typically the house number or name.	String
SAON	Secondary Addressable Object Name. Where a property has been divided into separate units (for example, flats), the PAON (above) will identify the building and a SAON will be specified that identifies the separate unit/flat.	String
Street		String
Locality		String
Town/City		String
District		String
County		String
PPD Category Type	Indicates the type of Price Paid transaction. A = Standard Price Paid entry, includes single residential property sold for value. B = Additional Price Paid entry including transfers under a power of sale/repossessions, buy-to-lets (where they can be identified by a Mortgage), transfers to non-private individuals and sales where the property type is classed as ‘Other’. Note that category B does not separately identify the transaction types stated. HM Land Registry has been collecting information on Category A transactions from January 1995. Category B transactions were identified from October 2013.	String

Record Status - monthly file only	Indicates additions, changes and deletions to the records.(see guide below).
	A = Addition
	C = Change
	D = Delete
	
	Note that where a transaction changes category type due to misallocation (as above) it will be deleted from the original category type and added to the correct category with a new transaction unique identifier.
		String

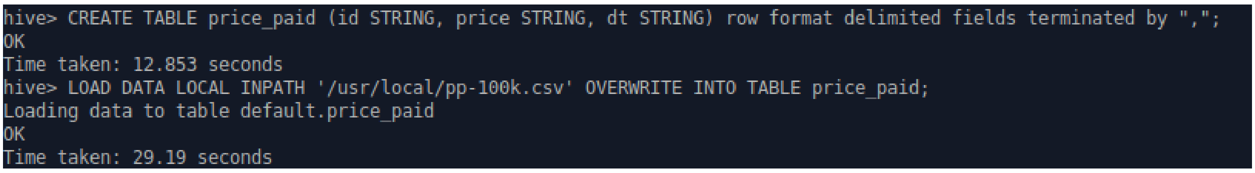
4. С помощью команды head -n сделайте 3 файла содержащие 100к, 1M и 10M строк.



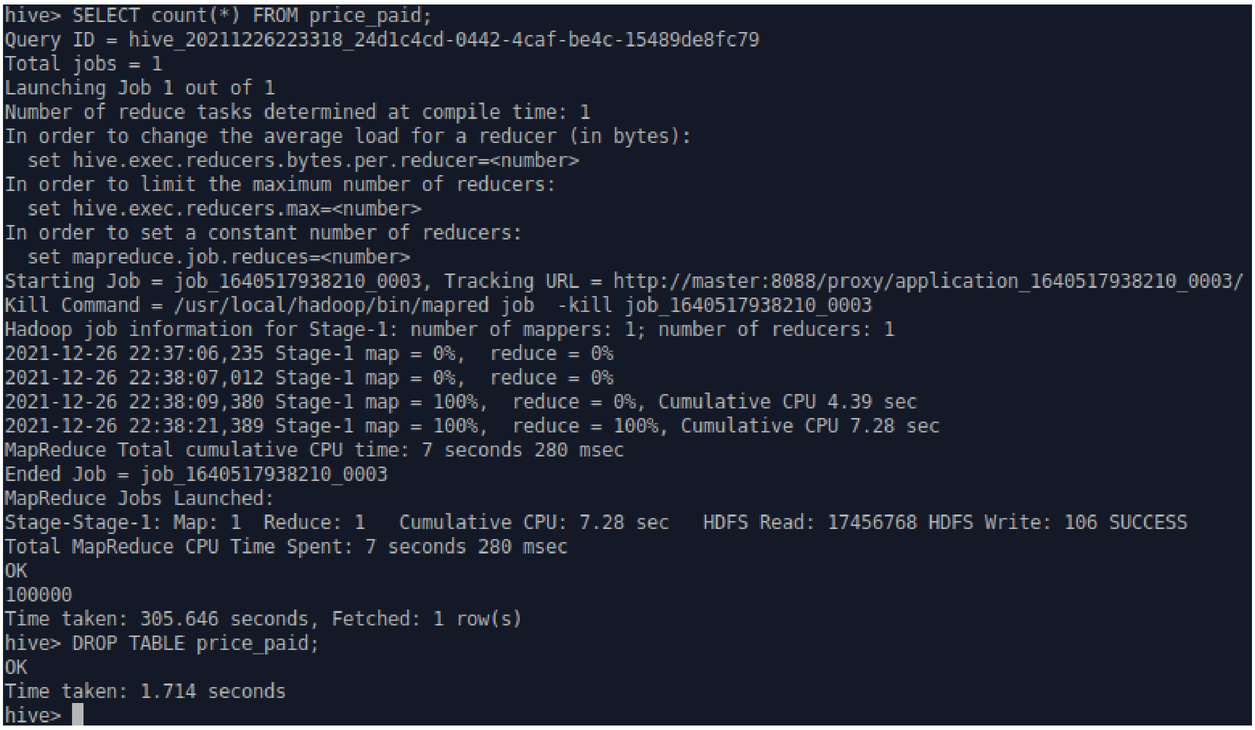
```
root@master:/usr/local# cat pp-complete.csv | head -100000 > pp-100k.csv
root@master:/usr/local# cat pp-100k.csv | wc -l
100000
root@master:/usr/local# cat pp-complete.csv | head -1000000 > pp-1m.csv
root@master:/usr/local# cat pp-1m.csv | wc -l
1000000
root@master:/usr/local# cat pp-complete.csv | head -10000000 > pp-10m.csv
root@master:/usr/local# cat pp-10m.csv | wc -l
10000000
```

5. Создайте тестовую таблицу при помощи кода в примере 1 и загрузите в неё данные записав в отчёт скорость записи каждого файла (для каждого следующего файла таблицу можно удалять или создавать новую с другим именем), количество строк и скорость выполнения запроса count(*).

100k

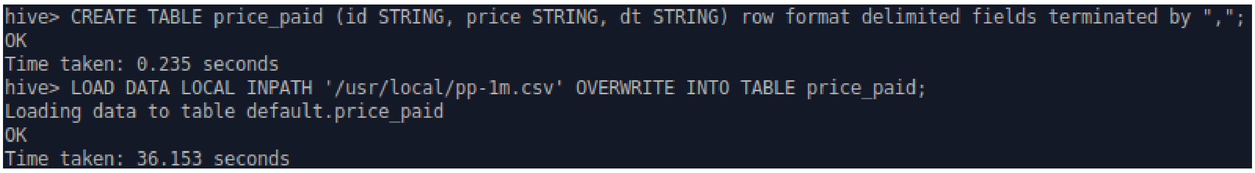


```
hive> CREATE TABLE price_paid (id STRING, price STRING, dt STRING) row format delimited fields terminated by ",";
OK
Time taken: 12.853 seconds
hive> LOAD DATA LOCAL INPATH '/usr/local/pp-100k.csv' OVERWRITE INTO TABLE price_paid;
Loading data to table default.price_paid
OK
Time taken: 29.19 seconds
```

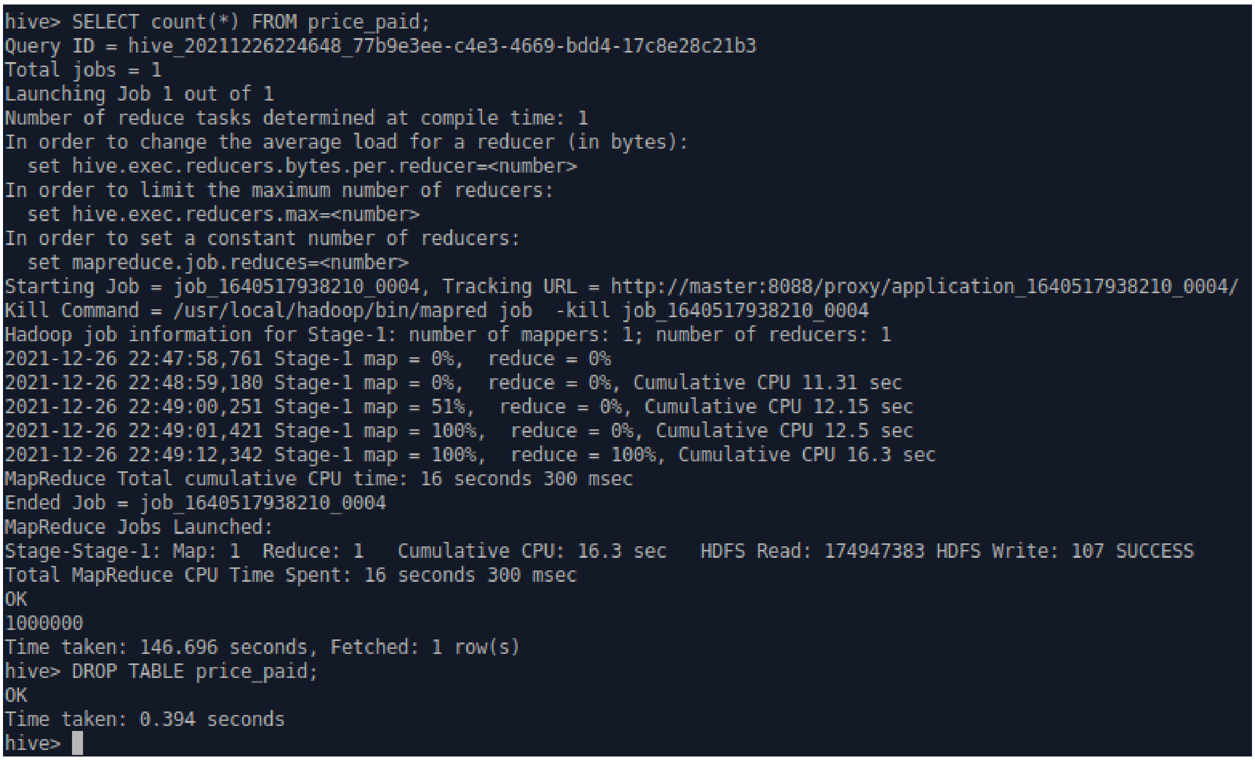


```
hive> SELECT count(*) FROM price_paid;
Query ID = hive_20211226223318_24d1c4cd-0442-4caf-be4c-15489de8fc79
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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>
Starting Job = job_1640517938210_0003, Tracking URL = http://master:8088/proxy/application_1640517938210_0003/
Kill Command = /usr/local/hadoop/bin/mapred job  -kill job_1640517938210_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-12-26 22:37:06,235 Stage-1 map = 0%,  reduce = 0%
2021-12-26 22:38:07,012 Stage-1 map = 0%,  reduce = 0%
2021-12-26 22:38:09,380 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.39 sec
2021-12-26 22:38:21,389 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.28 sec
MapReduce Total cumulative CPU time: 7 seconds 280 msec
Ended Job = job_1640517938210_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.28 sec   HDFS Read: 17456768 HDFS Write: 106 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 280 msec
OK
100000
Time taken: 305.646 seconds, Fetched: 1 row(s)
hive> DROP TABLE price_paid;
OK
Time taken: 1.714 seconds
hive>
```

1m

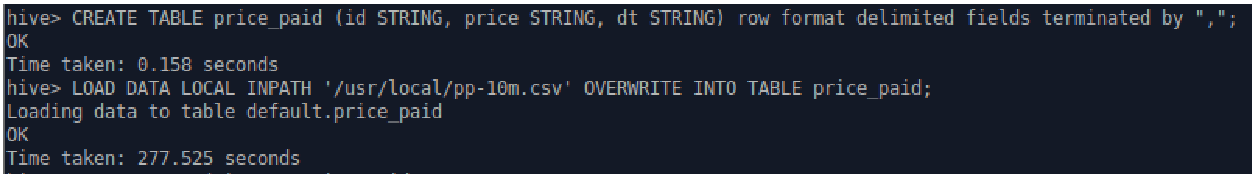


```
hive> CREATE TABLE price_paid (id STRING, price STRING, dt STRING) row format delimited fields terminated by ",";
OK
Time taken: 0.235 seconds
hive> LOAD DATA LOCAL INPATH '/usr/local/pp-1m.csv' OVERWRITE INTO TABLE price_paid;
Loading data to table default.price_paid
OK
Time taken: 36.153 seconds
```

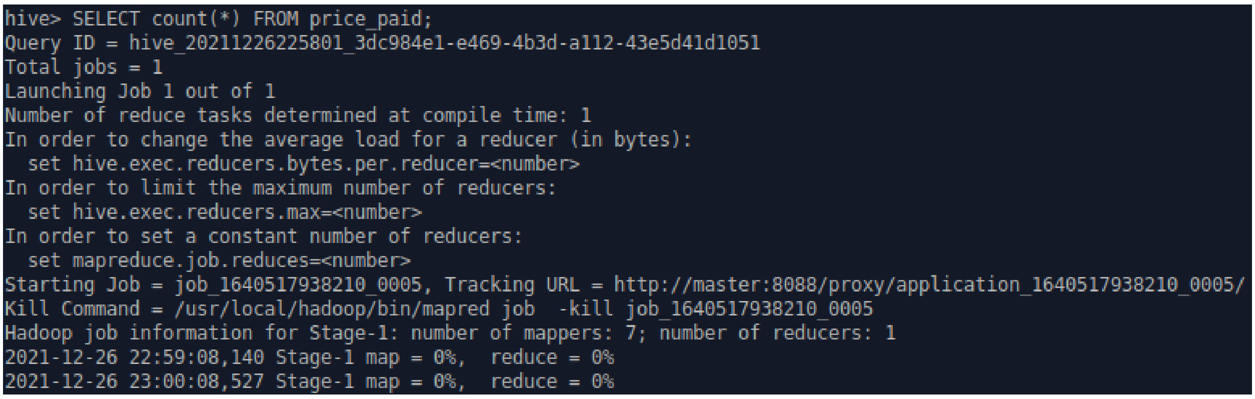


```
hive> SELECT count(*) FROM price_paid;
Query ID = hive_20211226224648_77b9e3ee-c4e3-4669-bdd4-17c8e28c21b3
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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>
Starting Job = job_1640517938210_0004, Tracking URL = http://master:8088/proxy/application_1640517938210_0004/
Kill Command = /usr/local/hadoop/bin/mapred job  -kill job_1640517938210_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-12-26 22:47:58,761 Stage-1 map = 0%,  reduce = 0%
2021-12-26 22:48:59,180 Stage-1 map = 0%,  reduce = 0%, Cumulative CPU 11.31 sec
2021-12-26 22:49:00,251 Stage-1 map = 51%,  reduce = 0%, Cumulative CPU 12.15 sec
2021-12-26 22:49:01,421 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 12.5 sec
2021-12-26 22:49:12,342 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 16.3 sec
MapReduce Total cumulative CPU time: 16 seconds 300 msec
Ended Job = job_1640517938210_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 16.3 sec   HDFS Read: 174947383 HDFS Write: 107 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 300 msec
OK
1000000
Time taken: 146.696 seconds, Fetched: 1 row(s)
hive> DROP TABLE price_paid;
OK
Time taken: 0.394 seconds
hive> █
```

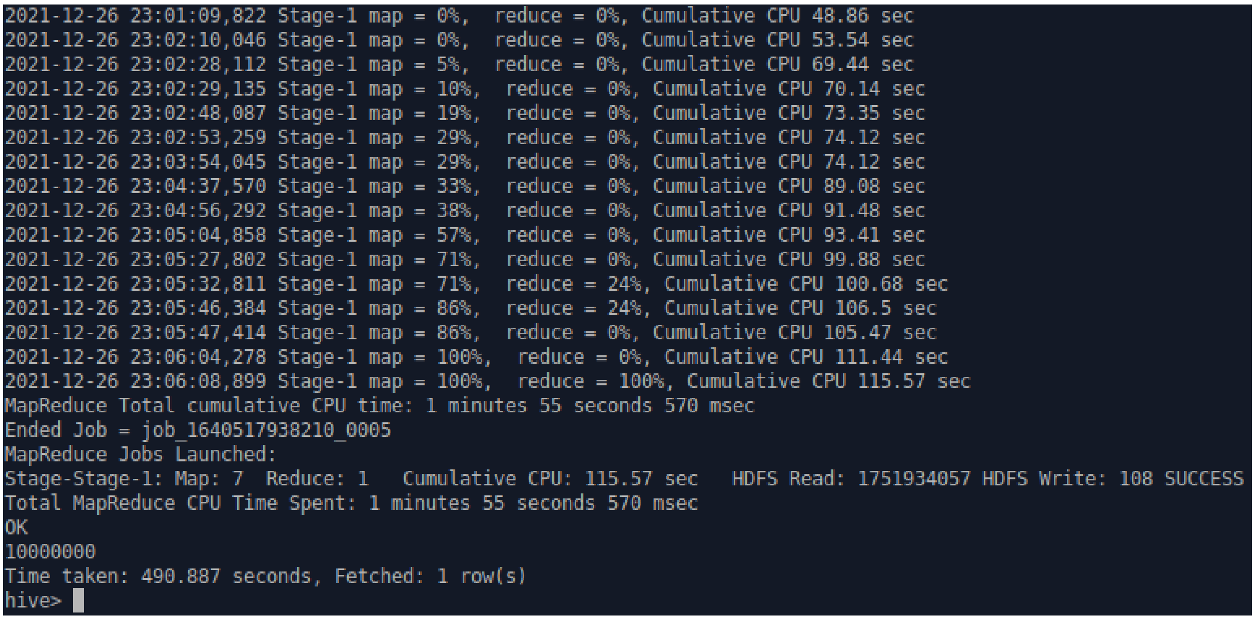
10m



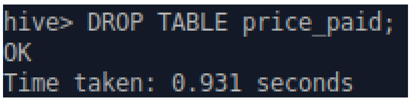
```
hive> CREATE TABLE price_paid (id STRING, price STRING, dt STRING) row format delimited fields terminated by ",";
OK
Time taken: 0.158 seconds
hive> LOAD DATA LOCAL INPATH '/usr/local/pp-10m.csv' OVERWRITE INTO TABLE price_paid;
Loading data to table default.price_paid
OK
Time taken: 277.525 seconds
```



```
hive> SELECT count(*) FROM price_paid;
Query ID = hive_20211226225801_3dc984e1-e469-4b3d-a112-43e5d41d1051
Total jobs = 1
Launching Job 1 out of 1
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>
Starting Job = job_1640517938210_0005, Tracking URL = http://master:8088/proxy/application_1640517938210_0005/
Kill Command = /usr/local/hadoop/bin/mapred job  -kill job_1640517938210_0005
Hadoop job information for Stage-1: number of mappers: 7; number of reducers: 1
2021-12-26 22:59:08,140 Stage-1 map = 0%,  reduce = 0%
2021-12-26 23:00:08,527 Stage-1 map = 0%,  reduce = 0%
```



```
2021-12-26 23:01:09,822 Stage-1 map = 0%,  reduce = 0%, Cumulative CPU 48.86 sec
2021-12-26 23:02:10,046 Stage-1 map = 0%,  reduce = 0%, Cumulative CPU 53.54 sec
2021-12-26 23:02:28,112 Stage-1 map = 5%,  reduce = 0%, Cumulative CPU 69.44 sec
2021-12-26 23:02:29,135 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 70.14 sec
2021-12-26 23:02:48,087 Stage-1 map = 19%,  reduce = 0%, Cumulative CPU 73.35 sec
2021-12-26 23:02:53,259 Stage-1 map = 29%,  reduce = 0%, Cumulative CPU 74.12 sec
2021-12-26 23:03:54,045 Stage-1 map = 29%,  reduce = 0%, Cumulative CPU 74.12 sec
2021-12-26 23:04:37,570 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 89.08 sec
2021-12-26 23:04:56,292 Stage-1 map = 38%,  reduce = 0%, Cumulative CPU 91.48 sec
2021-12-26 23:05:04,858 Stage-1 map = 57%,  reduce = 0%, Cumulative CPU 93.41 sec
2021-12-26 23:05:27,802 Stage-1 map = 71%,  reduce = 0%, Cumulative CPU 99.88 sec
2021-12-26 23:05:32,811 Stage-1 map = 71%,  reduce = 24%, Cumulative CPU 100.68 sec
2021-12-26 23:05:46,384 Stage-1 map = 86%,  reduce = 24%, Cumulative CPU 106.5 sec
2021-12-26 23:05:47,414 Stage-1 map = 86%,  reduce = 0%, Cumulative CPU 105.47 sec
2021-12-26 23:06:04,278 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 111.44 sec
2021-12-26 23:06:08,899 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 115.57 sec
MapReduce Total cumulative CPU time: 1 minutes 55 seconds 570 msec
Ended Job = job_1640517938210_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 7  Reduce: 1   Cumulative CPU: 115.57 sec   HDFS Read: 1751934057 HDFS Write: 108 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 55 seconds 570 msec
OK
10000000
Time taken: 490.887 seconds, Fetched: 1 row(s)
hive>
```

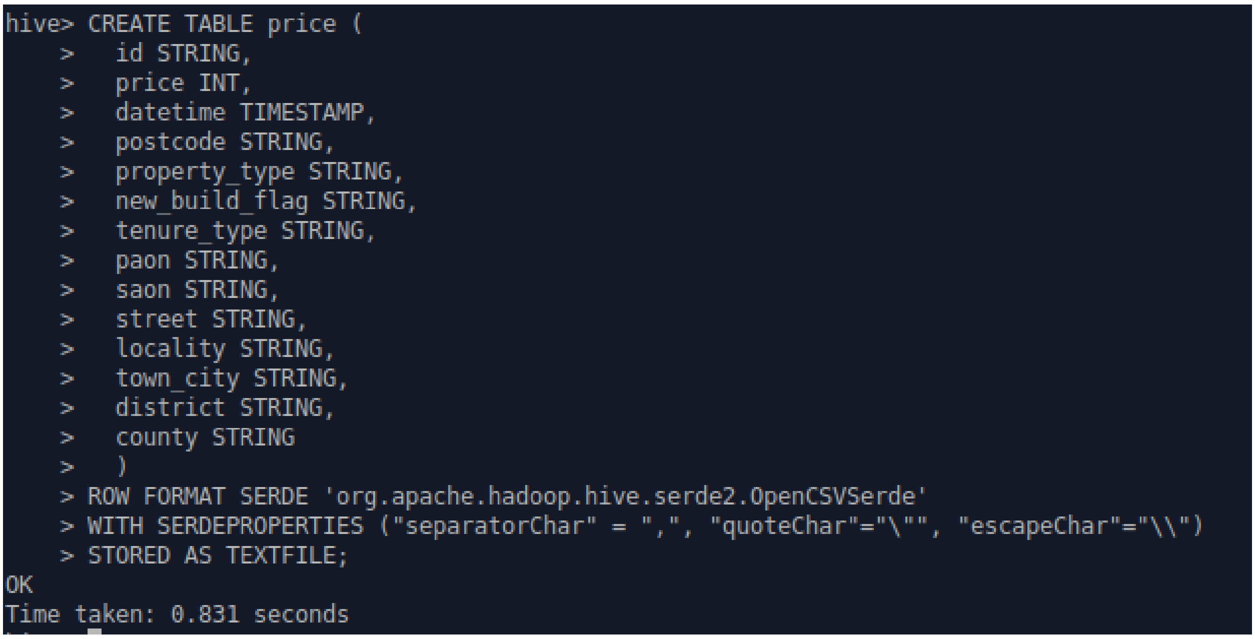


```
hive> DROP TABLE price_paid;
OK
Time taken: 0.931 seconds
```

Задание 3

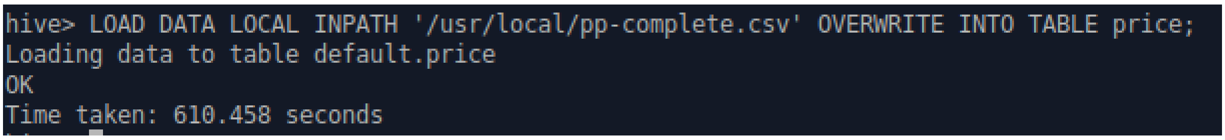
Типизация данных в HIVE.

1. Дополнив оставшимися колонками пример ниже загрузите данные в таблицы HIVE, замерьте время загрузки и запишите в отчёт



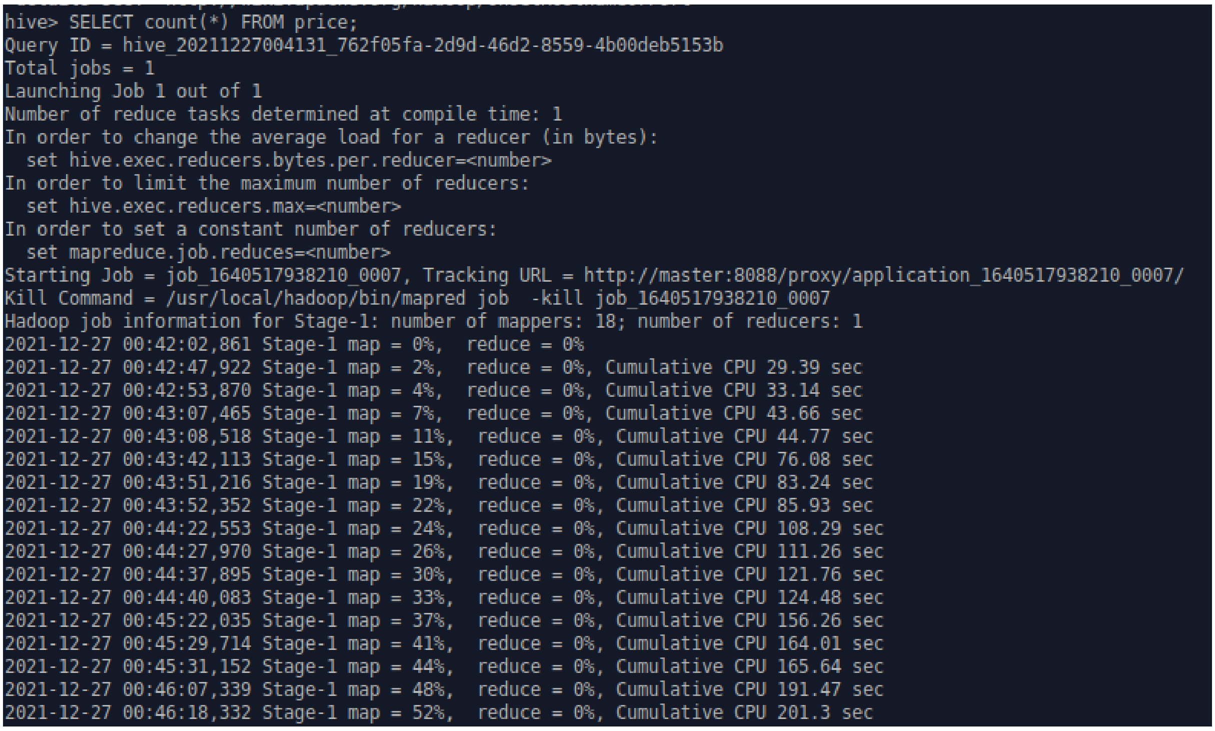
```
hive> CREATE TABLE price (
    >   id STRING,
    >   price INT,
    >   datetime TIMESTAMP,
    >   postcode STRING,
    >   property_type STRING,
    >   new_build_flag STRING,
    >   tenure_type STRING,
    >   paon STRING,
    >   saon STRING,
    >   street STRING,
    >   locality STRING,
    >   town_city STRING,
    >   district STRING,
    >   county STRING
    >   )
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES ("separatorChar" = ",", "quoteChar"="\"", "escapeChar"="\\")
    > STORED AS TEXTFILE;
OK
Time taken: 0.831 seconds
```

```
CREATE TABLE price (
  id STRING,
  price INT,
  datetime TIMESTAMP,
  postcode STRING,
  property_type STRING,
  new_build_flag STRING,
  tenure_type STRING,
  paon STRING,
  saon STRING,
  street STRING,
  locality STRING,
  town_city STRING,
  district STRING,
  county STRING,
  ppd STRING,
  rs STRING
  )
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES ("separatorChar" = ",", "quoteChar"="\"", "escapeChar"="\\")
STORED AS TEXTFILE;
```

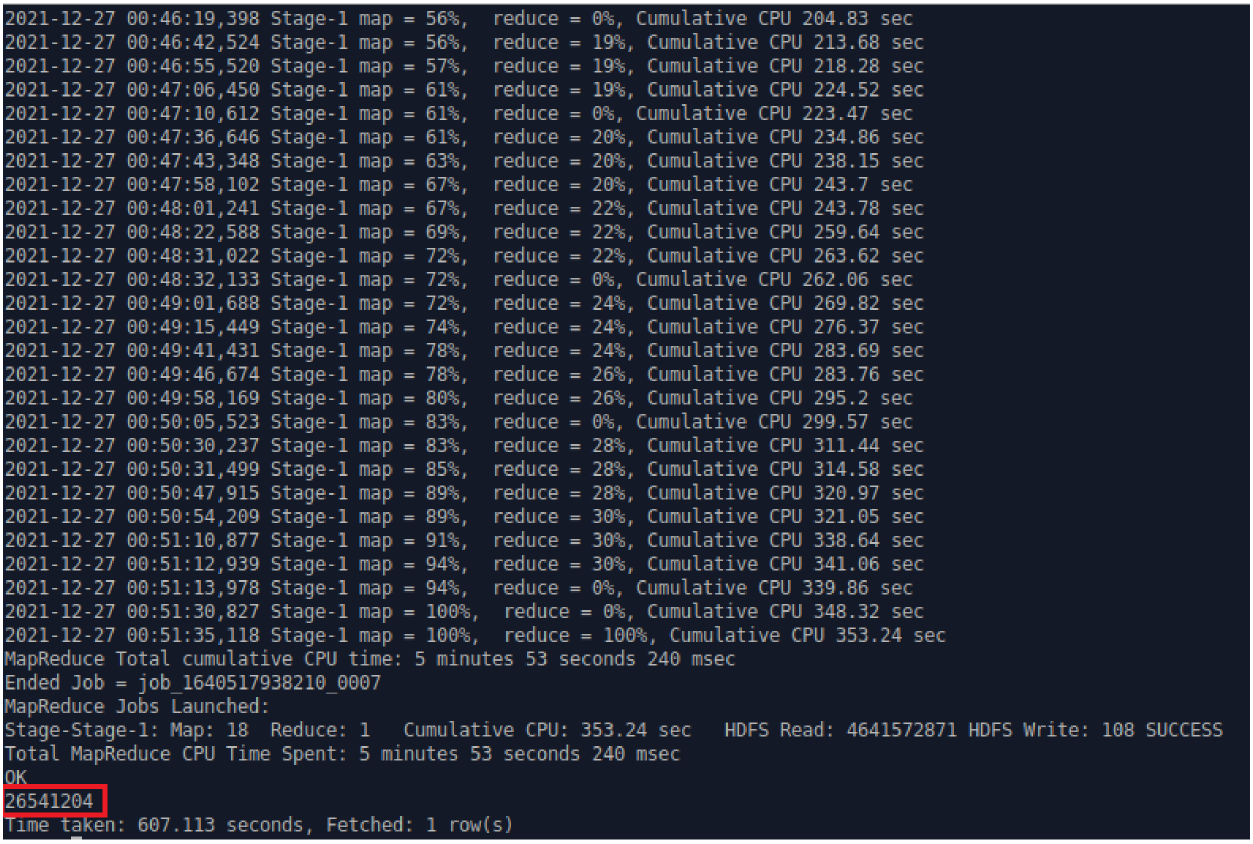


```
hive> LOAD DATA LOCAL INPATH '/usr/local/pp-complete.csv' OVERWRITE INTO TABLE price;
Loading data to table default.price
OK
Time taken: 610.458 seconds
```

2. В итоговой таблице должно содержаться 16 колонок и 26_541_204 строк.

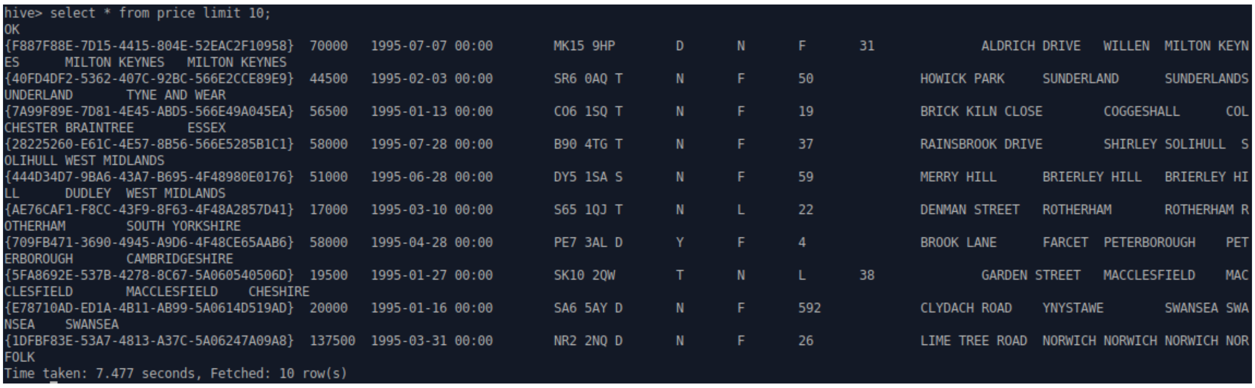


```
hive> SELECT count(*) FROM price;
Query ID = hive_20211227004131_762f05fa-2d9d-46d2-8559-4b00deb5153b
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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>
Starting Job = job_1640517938210_0007, Tracking URL = http://master:8088/proxy/application_1640517938210_0007/
Kill Command = /usr/local/hadoop/bin/mapred job  -kill job_1640517938210_0007
Hadoop job information for Stage-1: number of mappers: 18; number of reducers: 1
2021-12-27 00:42:02,861 Stage-1 map = 0%,  reduce = 0%
2021-12-27 00:42:47,922 Stage-1 map = 2%,  reduce = 0%, Cumulative CPU 29.39 sec
2021-12-27 00:42:53,870 Stage-1 map = 4%,  reduce = 0%, Cumulative CPU 33.14 sec
2021-12-27 00:43:07,465 Stage-1 map = 7%,  reduce = 0%, Cumulative CPU 43.66 sec
2021-12-27 00:43:08,518 Stage-1 map = 11%,  reduce = 0%, Cumulative CPU 44.77 sec
2021-12-27 00:43:42,113 Stage-1 map = 15%,  reduce = 0%, Cumulative CPU 76.08 sec
2021-12-27 00:43:51,216 Stage-1 map = 19%,  reduce = 0%, Cumulative CPU 83.24 sec
2021-12-27 00:43:52,352 Stage-1 map = 22%,  reduce = 0%, Cumulative CPU 85.93 sec
2021-12-27 00:44:22,553 Stage-1 map = 24%,  reduce = 0%, Cumulative CPU 108.29 sec
2021-12-27 00:44:27,970 Stage-1 map = 26%,  reduce = 0%, Cumulative CPU 111.26 sec
2021-12-27 00:44:37,895 Stage-1 map = 30%,  reduce = 0%, Cumulative CPU 121.76 sec
2021-12-27 00:44:40,083 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 124.48 sec
2021-12-27 00:45:22,035 Stage-1 map = 37%,  reduce = 0%, Cumulative CPU 156.26 sec
2021-12-27 00:45:29,714 Stage-1 map = 41%,  reduce = 0%, Cumulative CPU 164.01 sec
2021-12-27 00:45:31,152 Stage-1 map = 44%,  reduce = 0%, Cumulative CPU 165.64 sec
2021-12-27 00:46:07,339 Stage-1 map = 48%,  reduce = 0%, Cumulative CPU 191.47 sec
2021-12-27 00:46:18,332 Stage-1 map = 52%,  reduce = 0%, Cumulative CPU 201.3 sec
```



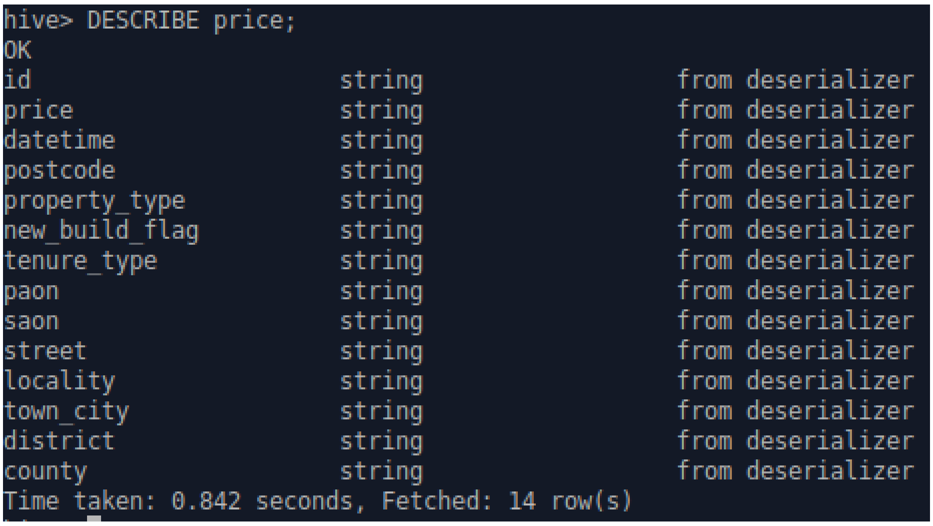
```
2021-12-27 00:46:19,398 Stage-1 map = 56%,  reduce = 0%, Cumulative CPU 204.83 sec
2021-12-27 00:46:42,524 Stage-1 map = 56%,  reduce = 19%, Cumulative CPU 213.68 sec
2021-12-27 00:46:55,520 Stage-1 map = 57%,  reduce = 19%, Cumulative CPU 218.28 sec
2021-12-27 00:47:06,450 Stage-1 map = 61%,  reduce = 19%, Cumulative CPU 224.52 sec
2021-12-27 00:47:10,612 Stage-1 map = 61%,  reduce = 0%, Cumulative CPU 223.47 sec
2021-12-27 00:47:36,646 Stage-1 map = 61%,  reduce = 20%, Cumulative CPU 234.86 sec
2021-12-27 00:47:43,348 Stage-1 map = 63%,  reduce = 20%, Cumulative CPU 238.15 sec
2021-12-27 00:47:58,102 Stage-1 map = 67%,  reduce = 20%, Cumulative CPU 243.7 sec
2021-12-27 00:48:01,241 Stage-1 map = 67%,  reduce = 22%, Cumulative CPU 243.78 sec
2021-12-27 00:48:22,588 Stage-1 map = 69%,  reduce = 22%, Cumulative CPU 259.64 sec
2021-12-27 00:48:31,022 Stage-1 map = 72%,  reduce = 22%, Cumulative CPU 263.62 sec
2021-12-27 00:48:32,133 Stage-1 map = 72%,  reduce = 0%, Cumulative CPU 262.06 sec
2021-12-27 00:49:01,688 Stage-1 map = 72%,  reduce = 24%, Cumulative CPU 269.82 sec
2021-12-27 00:49:15,449 Stage-1 map = 74%,  reduce = 24%, Cumulative CPU 276.37 sec
2021-12-27 00:49:41,431 Stage-1 map = 78%,  reduce = 24%, Cumulative CPU 283.69 sec
2021-12-27 00:49:46,674 Stage-1 map = 78%,  reduce = 26%, Cumulative CPU 283.76 sec
2021-12-27 00:49:58,169 Stage-1 map = 80%,  reduce = 26%, Cumulative CPU 295.2 sec
2021-12-27 00:50:05,523 Stage-1 map = 83%,  reduce = 0%, Cumulative CPU 299.57 sec
2021-12-27 00:50:30,237 Stage-1 map = 83%,  reduce = 28%, Cumulative CPU 311.44 sec
2021-12-27 00:50:31,499 Stage-1 map = 85%,  reduce = 28%, Cumulative CPU 314.58 sec
2021-12-27 00:50:47,915 Stage-1 map = 89%,  reduce = 28%, Cumulative CPU 320.97 sec
2021-12-27 00:50:54,209 Stage-1 map = 89%,  reduce = 30%, Cumulative CPU 321.05 sec
2021-12-27 00:51:10,877 Stage-1 map = 91%,  reduce = 30%, Cumulative CPU 338.64 sec
2021-12-27 00:51:12,939 Stage-1 map = 94%,  reduce = 30%, Cumulative CPU 341.06 sec
2021-12-27 00:51:13,978 Stage-1 map = 94%,  reduce = 0%, Cumulative CPU 339.86 sec
2021-12-27 00:51:30,827 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 348.32 sec
2021-12-27 00:51:35,118 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 353.24 sec
MapReduce Total cumulative CPU time: 5 minutes 53 seconds 240 msec
Ended Job = job_1640517938210_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 18  Reduce: 1   Cumulative CPU: 353.24 sec   HDFS Read: 4641572871 HDFS Write: 108 SUCCESS
Total MapReduce CPU Time Spent: 5 minutes 53 seconds 240 msec
OK
26541204
Time taken: 607.113 seconds, Fetched: 1 row(s)
```

Посмотрим первые 10 записей



```
hive> select * from price limit 10;
OK
{F887F88E-7D15-4415-804E-52EAC2F10958}  70000   1995-07-07 00:00        MK15 9HP        D       N       F       31              ALDRICH DRIVE   WILLEN  MILTON KEYN
ES      MILTON KEYNES   MILTON KEYNES
{40FD4DF2-5362-407C-92BC-566E2CCE89E9}  44500   1995-02-03 00:00        SR6 0AQ T       N       F       50              HOWICK PARK     SUNDERLAND      SUNDERLANDS
UNDERLAND       TYNE AND WEAR
{7A99F89E-7D81-4E45-ABD5-566E49A045EA}  56500   1995-01-13 00:00        CO6 1SQ T       N       F       19              BRICK KILN CLOSE        COGGESHALL      COL
CHESTER BRAINTREE       ESSEX
{28225260-E61C-4E57-8B56-566E5285B1C1}  58000   1995-07-28 00:00        B90 4TG T       N       F       37              RAINSBROOK DRIVE        SHIRLEY SOLIHULL  S
OLIHULL WEST MIDLANDS
{444D34D7-9BA6-43A7-B695-4F48980E0176}  51000   1995-06-28 00:00        DY5 1SA S       N       F       59              MERRY HILL      BRIERLEY HILL   BRIERLEY HI
LL      DUDLEY  WEST MIDLANDS
{AE76CAF1-F8CC-43F9-8F63-4F48A2857D41}  17000   1995-03-10 00:00        S65 1QJ T       N       L       22              DENMAN STREET   ROTHERHAM       ROTHERHAM R
OTHERHAM        SOUTH YORKSHIRE
{709FB471-3690-4945-A9D6-4F48CE65AAB6}  58000   1995-04-28 00:00        PE7 3AL D       Y       F       4               BROOK LANE      FARCET  PETERBOROUGH    PET
ERBOROUGH       CAMBRIDGESHIRE
{5FA8692E-537B-4278-8C67-5A060540506D}  19500   1995-01-27 00:00        SK10 2QW        T       N       L       38              GARDEN STREET   MACCLESFIELD    MAC
CLESFIELD       MACCLESFIELD    CHESHIRE
{E78710AD-ED1A-4B11-AB99-5A0614D519AD}  20000   1995-01-16 00:00        SA6 5AY D       N       F       592             CLYDACH ROAD    YNYSTAWE        SWANSEA SWA
NSEA    SWANSEA
{1DFBF83E-53A7-4813-A37C-5A06247A09A8}  137500  1995-03-31 00:00        NR2 2NQ D       N       F       26              LIME TREE ROAD  NORWICH NORWICH NORWICH NOR
FOLK
Time taken: 7.477 seconds, Fetched: 10 row(s)
```

Структура таблицы

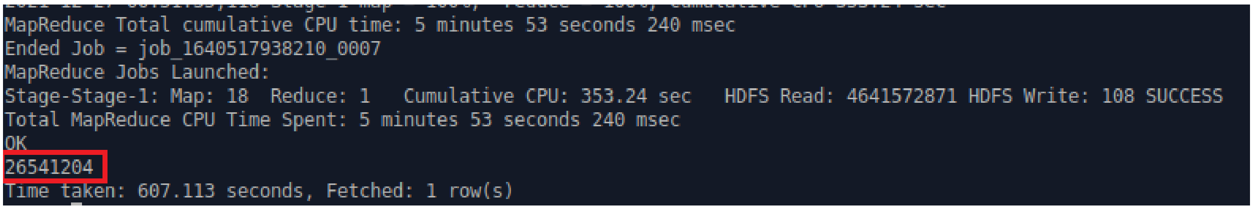


```
hive> DESCRIBE price;
OK
id                      string                  from deserializer
price                   string                  from deserializer
datetime                string                  from deserializer
postcode                string                  from deserializer
property_type           string                  from deserializer
new_build_flag          string                  from deserializer
tenure_type             string                  from deserializer
paon                    string                  from deserializer
saon                    string                  from deserializer
street                  string                  from deserializer
locality                string                  from deserializer
town_city               string                  from deserializer
district                string                  from deserializer
county                  string                  from deserializer
Time taken: 0.842 seconds, Fetched: 14 row(s)
```

3. Напишите запросы к загруженным данным, выполните их и запишите в отчёт: текст запроса, результат выполнения, время выполнения:

3.1. Количество загруженных строк данных

select count(*) from price;



```
MapReduce Total cumulative CPU time: 5 minutes 53 seconds 240 msec
Ended Job = job_1640517938210_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 18  Reduce: 1   Cumulative CPU: 353.24 sec   HDFS Read: 4641572871 HDFS Write: 108 SUCCESS
Total MapReduce CPU Time Spent: 5 minutes 53 seconds 240 msec
OK
26541204
Time taken: 607.113 seconds, Fetched: 1 row(s)
```

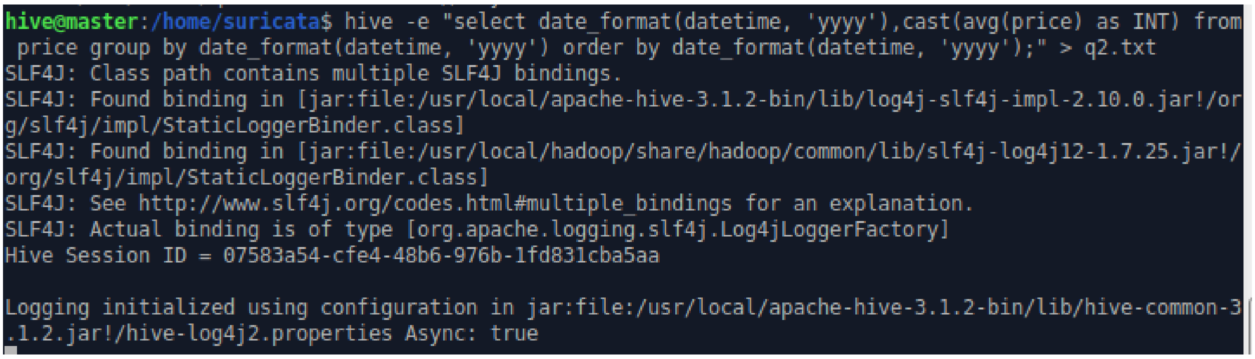
3.2. Средняя цена за год

select date_format(datetime, 'yyyy'),cast(avg(price) as INT)

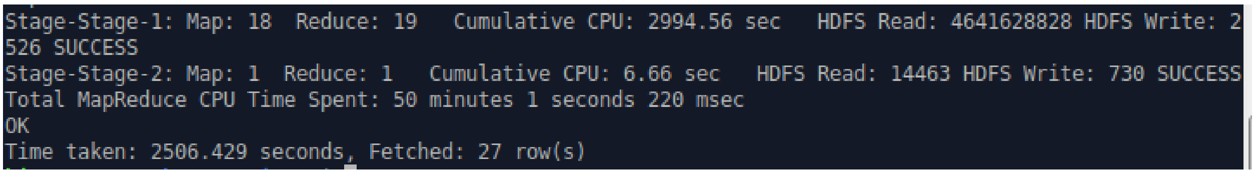
from price

group by date_format(datetime, 'yyyy')

order by date_format(datetime, 'yyyy');

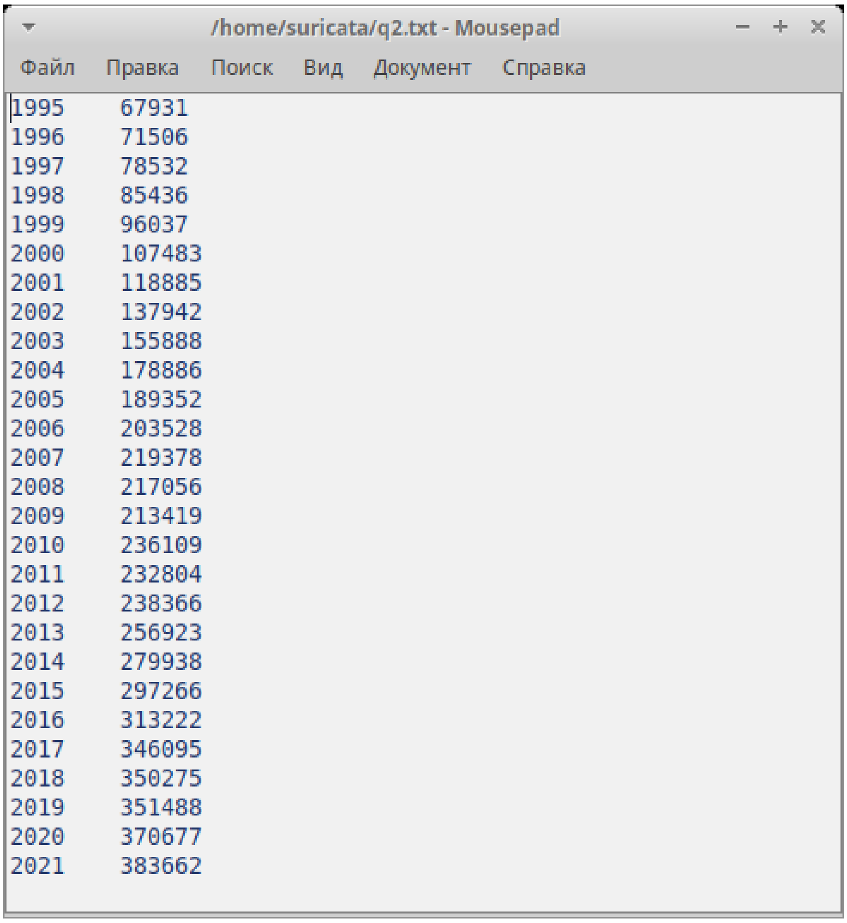


```
hive@master:/home/suricata$ hive -e "select date_format(datetime, 'yyyy'),cast(avg(price) as INT) from
 price group by date_format(datetime, 'yyyy') order by date_format(datetime, 'yyyy');" > q2.txt
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/or
g/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-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 = 07583a54-cfe4-48b6-976b-1fd831cba5aa

Logging initialized using configuration in jar:file:/usr/local/apache-hive-3.1.2-bin/lib/hive-common-3
.1.2.jar!/hive-log4j2.properties Async: true
```



```
Stage-Stage-1: Map: 18  Reduce: 19   Cumulative CPU: 2994.56 sec   HDFS Read: 4641628828 HDFS Write: 2
526 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 6.66 sec   HDFS Read: 14463 HDFS Write: 730 SUCCESS
Total MapReduce CPU Time Spent: 50 minutes 1 seconds 220 msec
OK
Time taken: 2506.429 seconds, Fetched: 27 row(s)
```

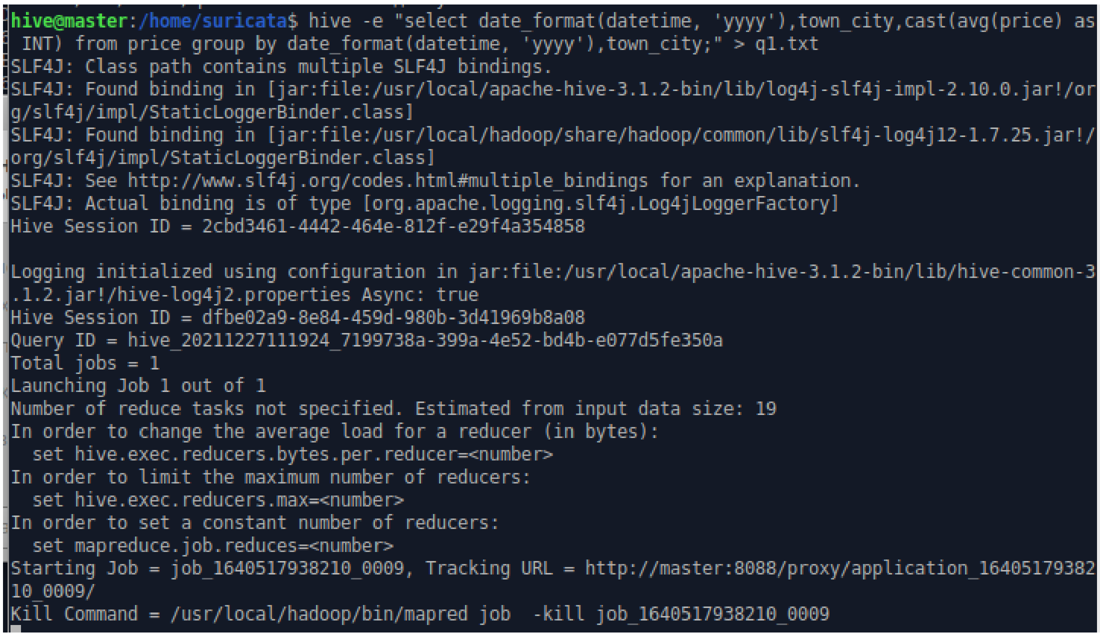
Результат в файле q2.txt



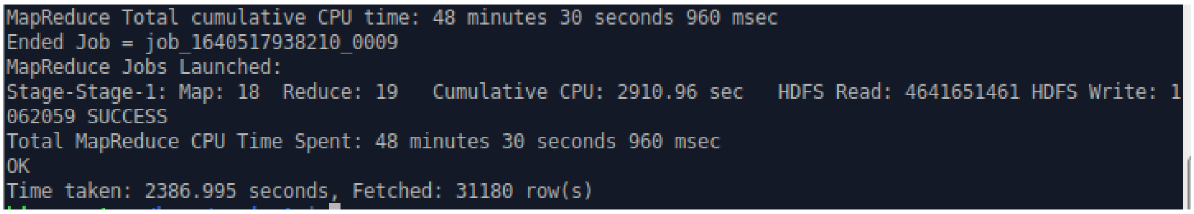
```
1995	67931
1996	71506
1997	78532
1998	85436
1999	96037
2000	107483
2001	118885
2002	137942
2003	155888
2004	178886
2005	189352
2006	203528
2007	219378
2008	217056
2009	213419
2010	236109
2011	232804
2012	238366
2013	256923
2014	279938
2015	297266
2016	313222
2017	346095
2018	350275
2019	351488
2020	370677
2021	383662
```

3.3 Средняя цена за год в Городе

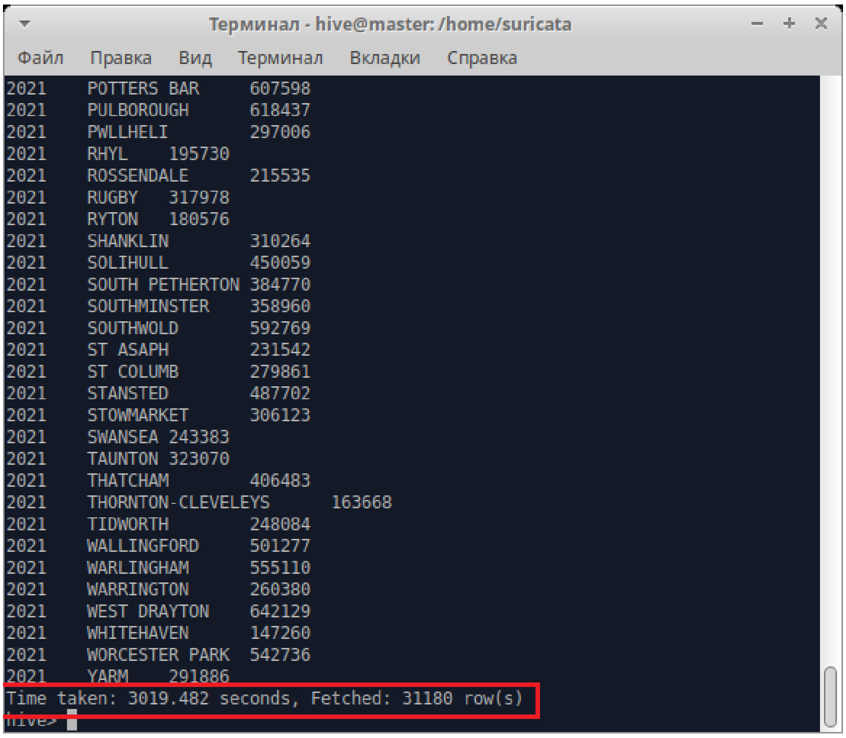
```
select date_format(datetime, 'yyyy'),town_city, cast(avg(price) as INT)
from price
group by date_format(datetime, 'yyyy'), town_city
order by date_format(datetime, 'yyyy');
```



```
hive@master:/home/suricata$ hive -e "select date_format(datetime, 'yyyy'),town_city,cast(avg(price) as
 INT) from price group by date_format(datetime, 'yyyy'),town_city;" > q1.txt
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/or
g/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-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 = 2cbd3461-4442-464e-812f-e29f4a354858

Logging initialized using configuration in jar:file:/usr/local/apache-hive-3.1.2-bin/lib/hive-common-3
.1.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = dfbe02a9-8e84-459d-980b-3d41969b8a08
Query ID = hive_20211227111924_7199738a-399a-4e52-bd4b-e077d5fe350a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 19
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_1640517938210_0009, Tracking URL = http://master:8088/proxy/application_16405179382
10_0009/
Kill Command = /usr/local/hadoop/bin/mapred job  -kill job_1640517938210_0009
```

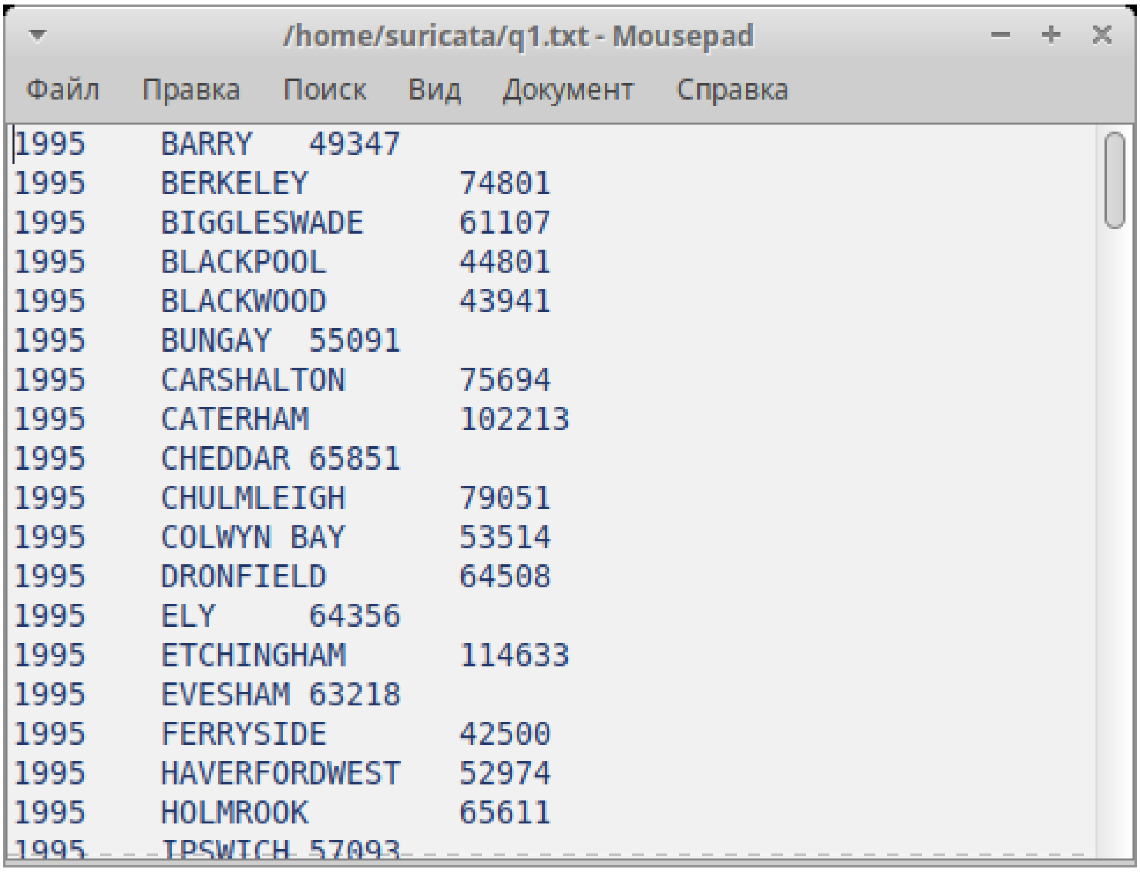


```
MapReduce Total cumulative CPU time: 48 minutes 30 seconds 960 msec
Ended Job = job_1640517938210_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 18  Reduce: 19   Cumulative CPU: 2910.96 sec   HDFS Read: 4641651461 HDFS Write: 1
062059 SUCCESS
Total MapReduce CPU Time Spent: 48 minutes 30 seconds 960 msec
OK
Time taken: 2386.995 seconds, Fetched: 31180 row(s)
```



```
2021    POTTERS BAR     607598
2021    PULBOROUGH      618437
2021    PWLLHELI        297006
2021    RHYL    195730
2021    ROSSENDALE      215535
2021    RUGBY   317978
2021    RYTON   180576
2021    SHANKLIN        310264
2021    SOLIHULL        450059
2021    SOUTH PETHERTON 384770
2021    SOUTHMINSTER    358960
2021    SOUTHWOLD       592769
2021    ST ASAPH        231542
2021    ST COLUMB       279861
2021    STANSTED        487702
2021    STOWMARKET      306123
2021    SWANSEA 243383
2021    TAUNTON 323070
2021    THATCHAM        406483
2021    THORNTON-CLEVELEYS      163668
2021    TIDWORTH        248084
2021    WALLINGFORD     501277
2021    WARLINGHAM      555110
2021    WARRINGTON      260380
2021    WEST DRAYTON    642129
2021    WHITEHAVEN      147260
2021    WORCESTER PARK  542736
2021    YARM    291886
Time taken: 3019.482 seconds, Fetched: 31180 row(s)
hive>
```

Результат в файле q1.txt



```
1995    BARRY   49347
1995    BERKELEY        74801
1995    BIGGLESWADE     61107
1995    BLACKPOOL       44801
1995    BLACKWOOD       43941
1995    BUNGAY  55091
1995    CARSHALTON      75694
1995    CATERHAM        102213
1995    CHEDDAR 65851
1995    CHULMLEIGH      79051
1995    COLWYN BAY      53514
1995    DRONFIELD       64508
1995    ELY     64356
1995    ETCHINGHAM      114633
1995    EVESHAM 63218
1995    FERRYSIDE       42500
1995    HAVERFORDWEST   52974
1995    HOLMROOK        65611
1995    IPSWICH 57093
```

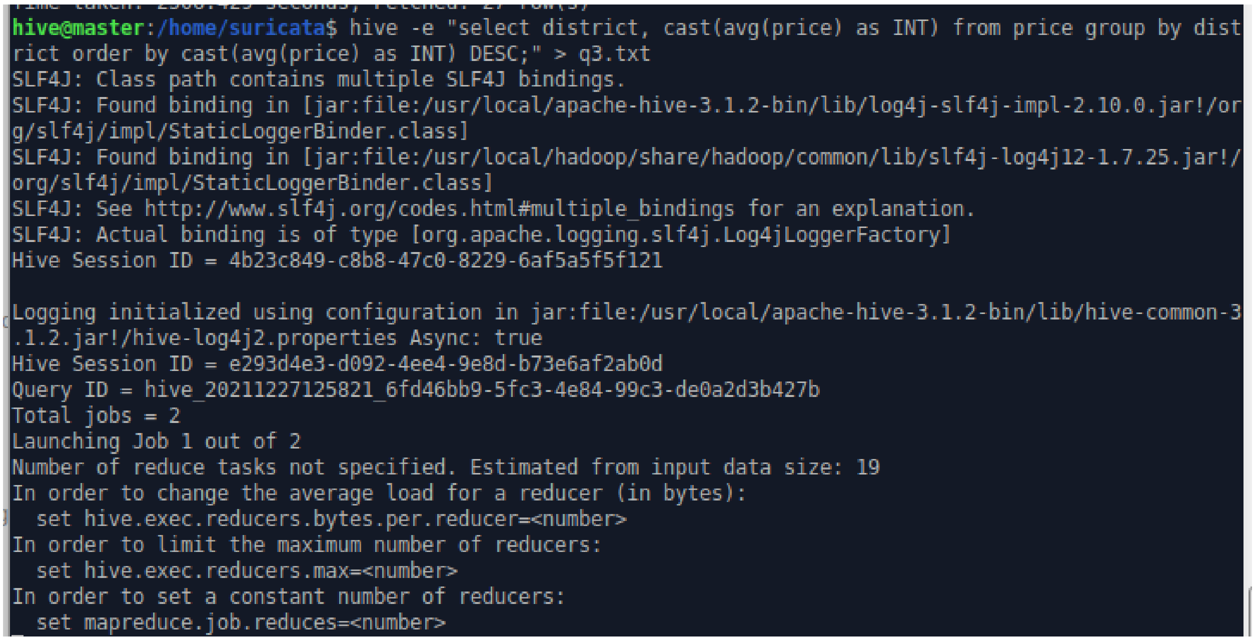
4.4 Самые дорогие районы

select district, cast(avg(price) as INT)

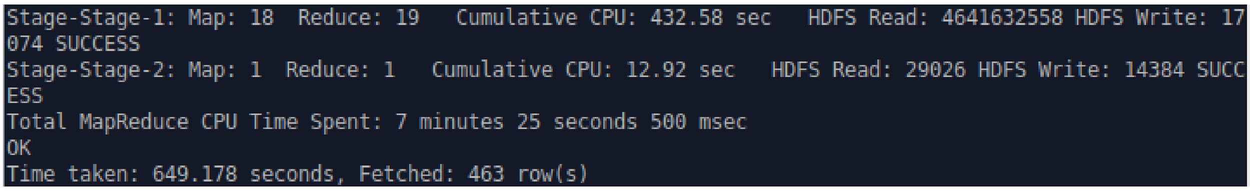
from price

group by district

order by cast(avg(price) as INT) DESC;

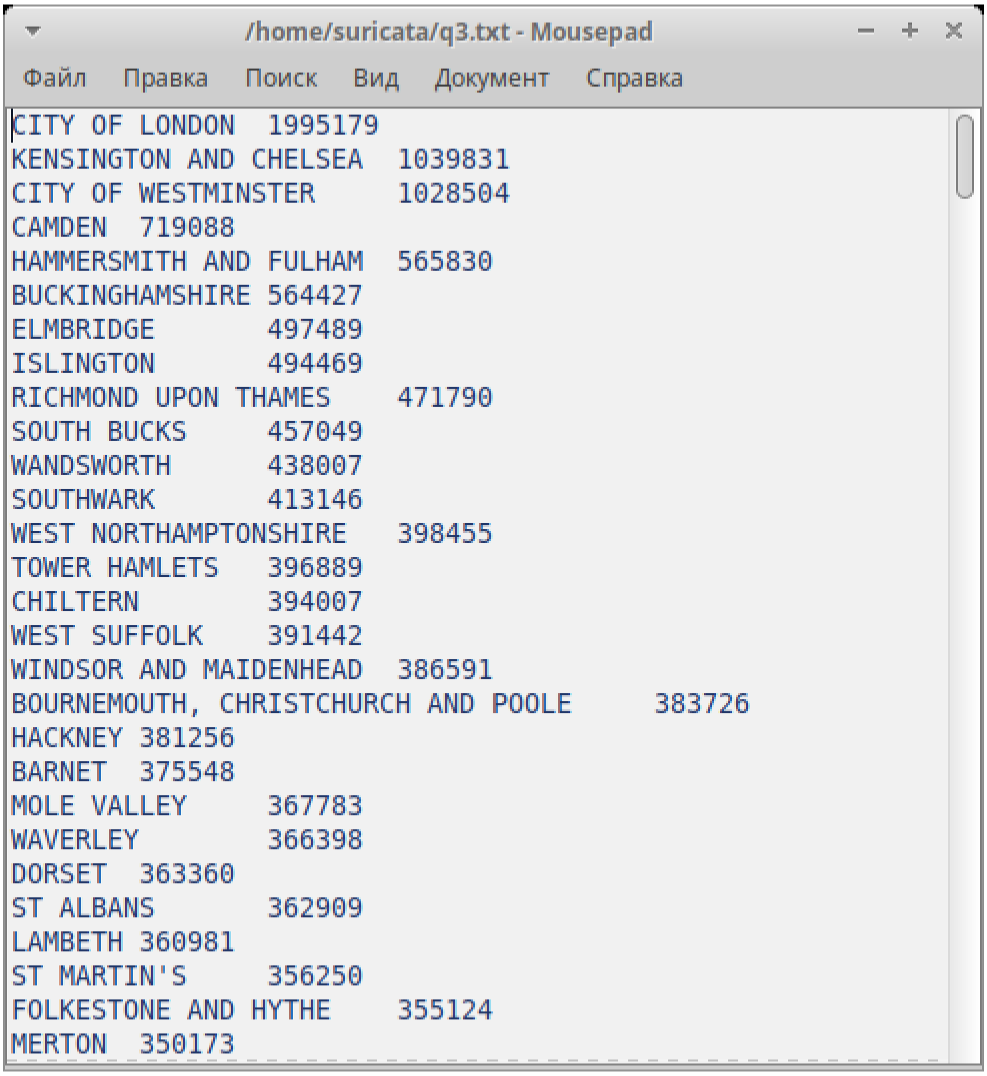


```
Time taken: 2300.425 seconds, Fetched: 27 row(s)
hive@master:/home/suricata$ hive -e "select district, cast(avg(price) as INT) from price group by dist
rict order by cast(avg(price) as INT) DESC;" > q3.txt
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/or
g/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-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 = 4b23c849-c8b8-47c0-8229-6af5a5f5f121

Logging initialized using configuration in jar:file:/usr/local/apache-hive-3.1.2-bin/lib/hive-common-3
.1.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = e293d4e3-d092-4ee4-9e8d-b73e6af2ab0d
Query ID = hive_20211227125821_6fd46bb9-5fc3-4e84-99c3-de0a2d3b427b
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 19
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>
```



```
Stage-Stage-1: Map: 18  Reduce: 19   Cumulative CPU: 432.58 sec   HDFS Read: 4641632558 HDFS Write: 17
074 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 12.92 sec   HDFS Read: 29026 HDFS Write: 14384 SUCC
ESS
Total MapReduce CPU Time Spent: 7 minutes 25 seconds 500 msec
OK
Time taken: 649.178 seconds, Fetched: 463 row(s)
```

Результат в файле q3.txt



```
CITY OF LONDON  1995179
KENSINGTON AND CHELSEA  1039831
CITY OF WESTMINSTER     1028504
CAMDEN  719088
HAMMERSMITH AND FULHAM  565830
BUCKINGHAMSHIRE 564427
ELMBRIDGE       497489
ISLINGTON       494469
RICHMOND UPON THAMES    471790
SOUTH BUCKS     457049
WANDSWORTH      438007
SOUTHWARK       413146
WEST NORTHAMPTONSHIRE   398455
TOWER HAMLETS   396889
CHILTERN        394007
WEST SUFFOLK    391442
WINDSOR AND MAIDENHEAD  386591
BOURNEMOUTH, CHRISTCHURCH AND POOLE     383726
HACKNEY 381256
BARNET  375548
MOLE VALLEY     367783
WAVERLEY        366398
DORSET  363360
ST ALBANS       362909
LAMBETH 360981
ST MARTIN'S     356250
FOLKESTONE AND HYTHE    355124
MERTON  350173
```