Topics in Computing Lab Assignment 5 : Hive

Team R : Vandita Goyal (2016ucp1004)
Nidheesh Panchal (2016ucp1008)
G. Jahnvi (2016ucp1332)

Objective:

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives a SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

It is used on top of Hadoop to perform more complicated map-reduce operations in a simpler manner

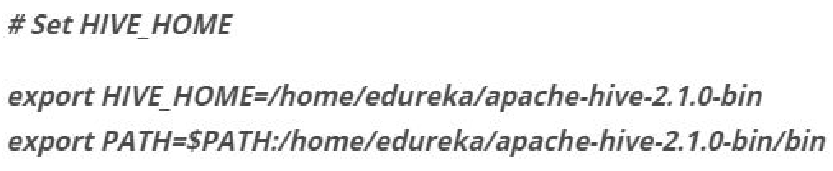
Implementation:

1. Installation of Hive:

The tutorial followed to achieve this:
https://www.edureka.co/blog/apache-hive-installation-on-ubuntu

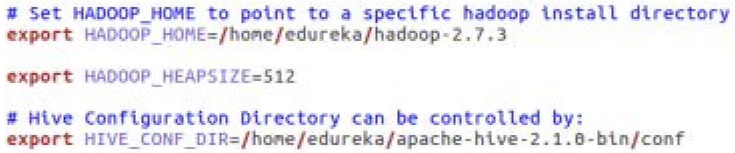
Steps followed:

1. Download Hive tar.
2. Extract the tar file.
3. Edit the ".bashrc" file to update the environment variables for user.



```
# Set HIVE_HOME

export HIVE_HOME=/home/edureka/apache-hive-2.1.0-bin
export PATH=$PATH:/home/edureka/apache-hive-2.1.0-bin/bin
```

4. Create Hive directories within HDFS. The directory 'warehouse' is the location to store the table or data related to hive.
5. Set read/write permissions for table.
6. Set Hadoop path in hive-env.sh

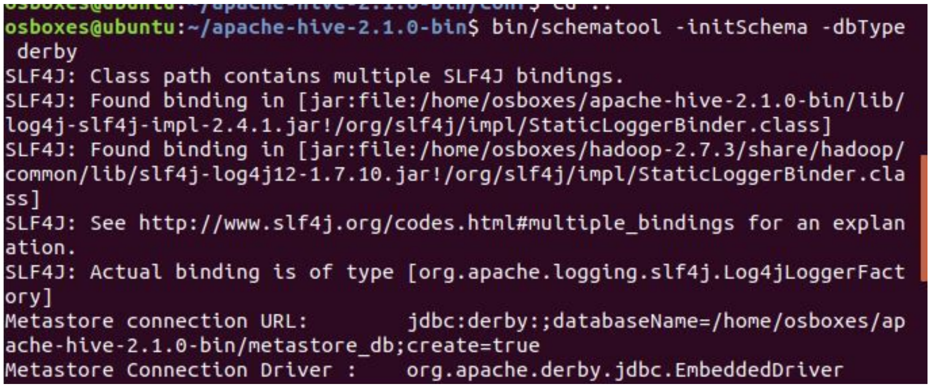


```
# Set HADOOP_HOME to point to a specific hadoop install directory
export HADOOP_HOME=/home/edureka/hadoop-2.7.3

export HADOOP_HEAPSIZE=512

# Hive Configuration Directory can be controlled by:
export HIVE_CONF_DIR=/home/edureka/apache-hive-2.1.0-bin/conf
```

7. Edit hive-site.xml

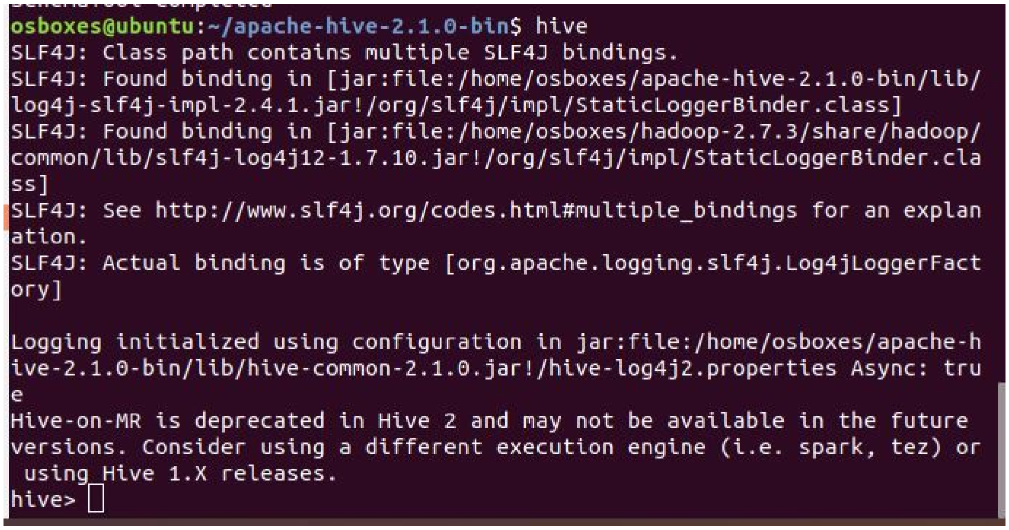
```
<configuration>
<property>
<name>javax.jdo.option.ConnectionURL</name>
<value>jdbc:derby:;databaseName=/home/edureka/apache-hive-2.1.0-bin/metastore_db;
create=true</value>
<description>
JDBC connect string for a JDBC metastore.
To use SSL to encrypt/authenticate the connection, provide database-specific SSL flag in
the connection URL.
For example, jdbc:postgresql://myhost/db?ssl=true for postgres database.
</description>
</property>
<property>
<name>hive.metastore.warehouse.dir</name>
<value>/user/hive/warehouse</value>
<description>location of default database for the warehouse</description>
</property>
<property>
<name>hive.metastore.uris</name>
<value/>
<description>Thrift URI for the remote metastore. Used by metastore client to connect to
remote metastore.</description>
</property>
<property>
<name>javax.jdo.option.ConnectionDriverName</name>
<value>org.apache.derby.jdbc.EmbeddedDriver</value>
<description>Driver class name for a JDBC metastore</description>
</property>
<property>
<name>javax.jdo.PersistenceManagerFactoryClass</name>
<value>org.datanucleus.api.jdo.JDOPersistenceManagerFactory</value>
<description>class implementing the jdo persistence</description>
</property>
</configuration>
```

8. Initialize Derby database.



```
osboxes@ubuntu:~/apache-hive-2.1.0-bin/conf$ cd ..
osboxes@ubuntu:~/apache-hive-2.1.0-bin$ bin/schematool -initSchema -dbType
 derby
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/osboxes/apache-hive-2.1.0-bin/lib/
log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/osboxes/hadoop-2.7.3/share/hadoop/
common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.cla
ss]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explan
ation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFact
ory]
Metastore connection URL:        jdbc:derby:;databaseName=/home/osboxes/ap
ache-hive-2.1.0-bin/metastore_db;create=true
Metastore Connection Driver :    org.apache.derby.jdbc.EmbeddedDriver
```

9. Launch Hive

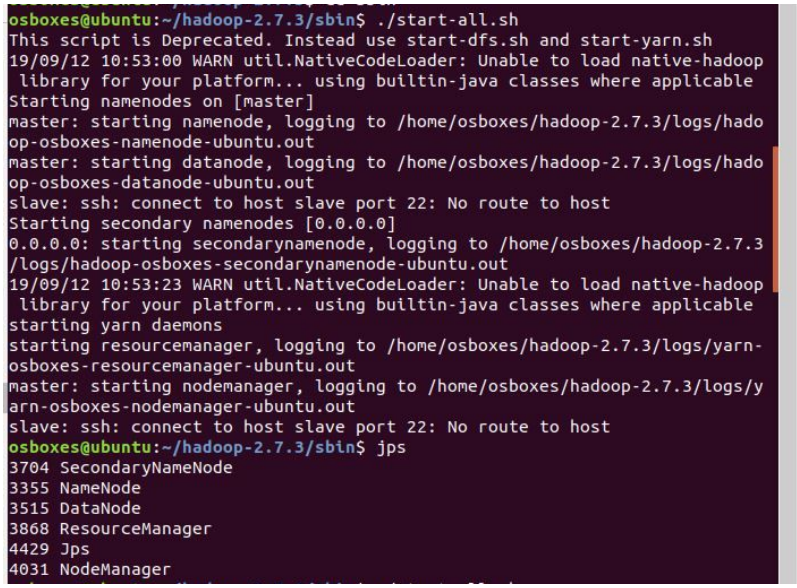


```
schemaTool completed
osboxes@ubuntu:~/apache-hive-2.1.0-bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/osboxes/apache-hive-2.1.0-bin/lib/
log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/osboxes/hadoop-2.7.3/share/hadoop/
common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.cla
ss]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explan
ation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFact
ory]

Logging initialized using configuration in jar:file:/home/osboxes/apache-h
ive-2.1.0-bin/lib/hive-common-2.1.0.jar!/hive-log4j2.properties Async: tru
e
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> 
```

2. Using hive perform query to find out the total sales done by each country

Steps followed:

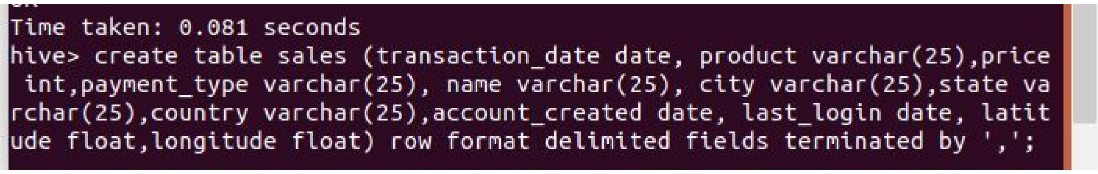
1. **Launch hadoop and hive**



```
osboxes@ubuntu:~/hadoop-2.7.3/sbin$ ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
19/09/12 10:53:00 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
Starting namenodes on [master]
master: starting namenode, logging to /home/osboxes/hadoop-2.7.3/logs/hado
op-osboxes-namenode-ubuntu.out
master: starting datanode, logging to /home/osboxes/hadoop-2.7.3/logs/hado
op-osboxes-datanode-ubuntu.out
slave: ssh: connect to host slave port 22: No route to host
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/osboxes/hadoop-2.7.3
/logs/hadoop-osboxes-secondarynamenode-ubuntu.out
19/09/12 10:53:23 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/osboxes/hadoop-2.7.3/logs/yarn-
osboxes-resourcemanager-ubuntu.out
master: starting nodemanager, logging to /home/osboxes/hadoop-2.7.3/logs/y
arn-osboxes-nodemanager-ubuntu.out
slave: ssh: connect to host slave port 22: No route to host
osboxes@ubuntu:~/hadoop-2.7.3/sbin$ jps
3704 SecondaryNameNode
3355 NameNode
3515 DataNode
3868 ResourceManager
4429 Jps
4031 NodeManager
```

2. **Create table for sales in hive**

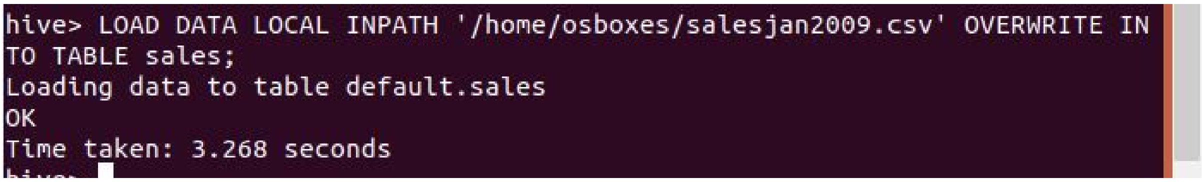
Note: The separator used is **‘,’**



```
Time taken: 0.081 seconds
hive> create table sales (transaction_date date, product varchar(25),price
 int,payment_type varchar(25), name varchar(25), city varchar(25),state va
rchar(25),country varchar(25),account_created date, last_login date, latit
ude float,longitude float) row format delimited fields terminated by ',';
```

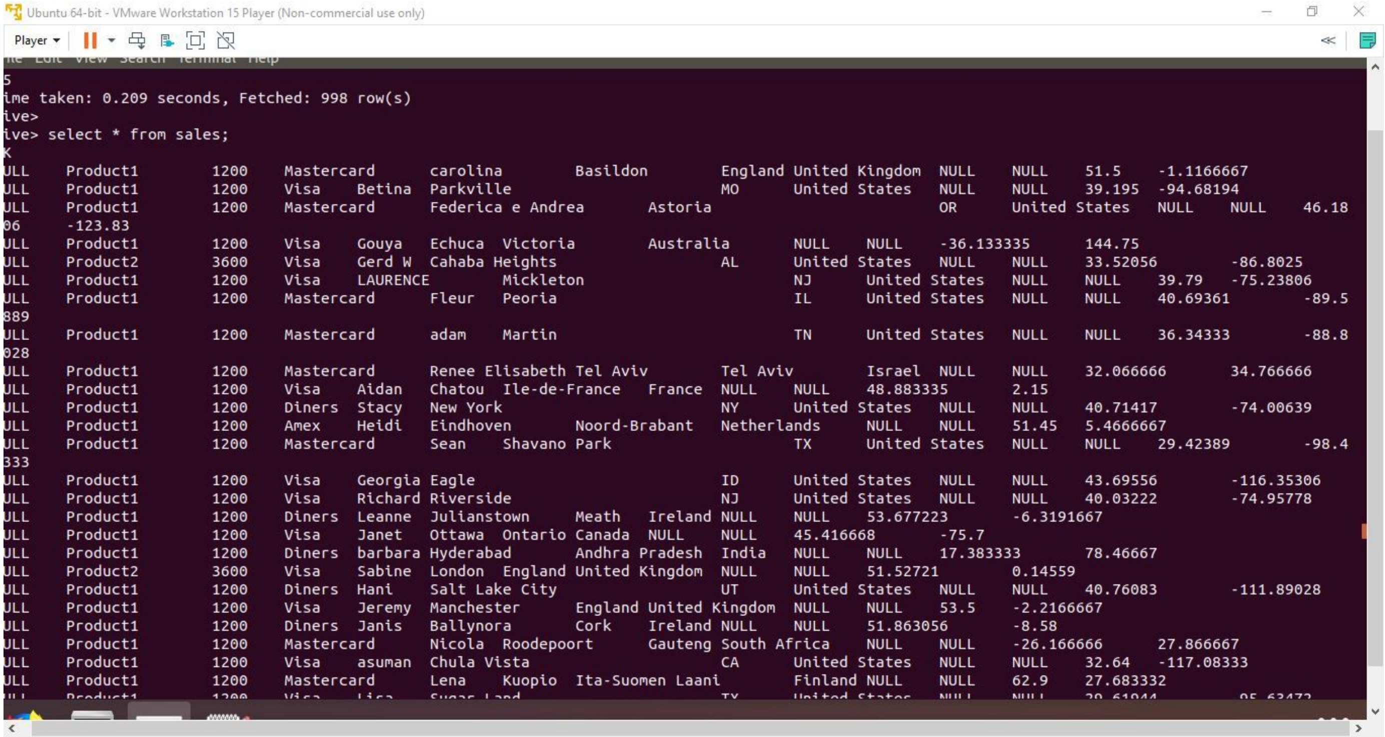
3. **Load the data in the table**

Note: we removed the top line of the data provided as that contained headings and not data



```
hive> LOAD DATA LOCAL INPATH '/home/osboxes/salesjan2009.csv' OVERWRITE IN
TO TABLE sales;
Loading data to table default.sales
OK
Time taken: 3.268 seconds
```

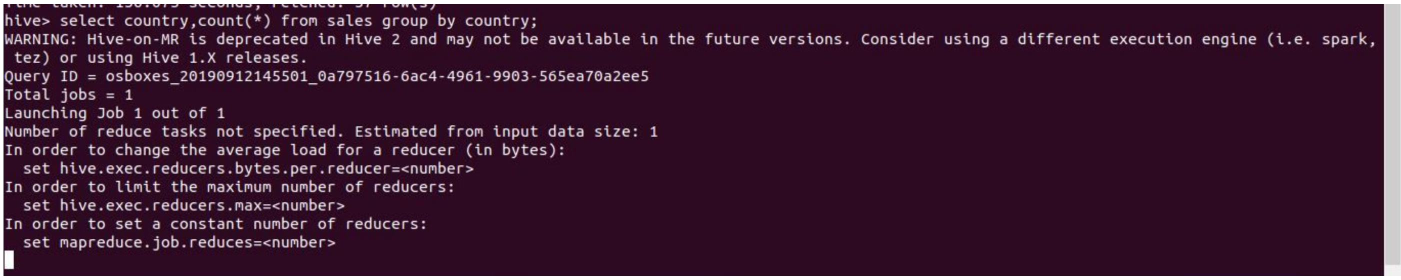
Loaded data:


Ubuntu 64-bit - VMware Workstation 15 Player (Non-commercial use only)
Player

```
5
ime taken: 0.209 seconds, Fetched: 998 row(s)
ive>
ive> select * from sales;
K
ULL    Product1        1200    Mastercard      carolina        Basildon        England United Kingdom  NULL    NULL    51.5    -1.1166667
ULL    Product1        1200    Visa    Betina  Parkville                       MO      United States   NULL    NULL    39.195  -94.68194
ULL    Product1        1200    Mastercard      Federica e Andrea       Astoria                         OR      United States   NULL    NULL    46.18
06     -123.83
ULL    Product1        1200    Visa    Gouya   Echuca  Victoria        Australia       NULL    NULL    -36.133335      144.75
ULL    Product2        3600    Visa    Gerd W  Cahaba Heights                  AL      United States   NULL    NULL    33.52056        -86.8025
ULL    Product1        1200    Visa    LAURENCE        Mickleton                       NJ      United States   NULL    NULL    39.79   -75.23806
ULL    Product1        1200    Mastercard      Fleur   Peoria                          IL      United States   NULL    NULL    40.69361        -89.5
889
ULL    Product1        1200    Mastercard      adam    Martin                          TN      United States   NULL    NULL    36.34333        -88.8
028
ULL    Product1        1200    Mastercard      Renee Elisabeth Tel Aviv        Tel Aviv        Israel  NULL    NULL    32.066666       34.766666
ULL    Product1        1200    Visa    Aidan   Chatou  Ile-de-France   France  NULL    NULL    48.883335       2.15
ULL    Product1        1200    Diners  Stacy   New York                        NY      United States   NULL    NULL    40.71417        -74.00639
ULL    Product1        1200    Amex    Heidi   Eindhoven       Noord-Brabant   Netherlands     NULL    NULL    51.45   5.4666667
ULL    Product1        1200    Mastercard      Sean    Shavano Park                    TX      United States   NULL    NULL    29.42389        -98.4
333
ULL    Product1        1200    Visa    Georgia Eagle                           ID      United States   NULL    NULL    43.69556        -116.35306
ULL    Product1        1200    Visa    Richard Riverside                       NJ      United States   NULL    NULL    40.03222        -74.95778
ULL    Product1        1200    Diners  Leanne  Julianstown     Meath   Ireland NULL    NULL    53.677223       -6.3191667
ULL    Product1        1200    Visa    Janet   Ottawa  Ontario Canada  NULL    NULL    45.416668       -75.7
ULL    Product1        1200    Diners  barbara Hyderabad       Andhra Pradesh  India   NULL    NULL    17.383333       78.46667
ULL    Product2        3600    Visa    Sabine  London  England United Kingdom  NULL    NULL    51.52721        0.14559
ULL    Product1        1200    Diners  Hani    Salt Lake City                  UT      United States   NULL    NULL    40.76083        -111.89028
ULL    Product1        1200    Visa    Jeremy  Manchester      England United Kingdom  NULL    NULL    53.5    -2.2166667
ULL    Product1        1200    Diners  Janis   Ballynora       Cork    Ireland NULL    NULL    51.863056       -8.58
ULL    Product1        1200    Mastercard      Nicola  Roodepoort      Gauteng South Africa    NULL    NULL    -26.166666      27.866667
ULL    Product1        1200    Visa    asuman  Chula Vista                     CA      United States   NULL    NULL    32.64   -117.08333
ULL    Product1        1200    Mastercard      Lena    Kuopio  Ita-Suomen Laani        Finland NULL    NULL    62.9    27.683332
```

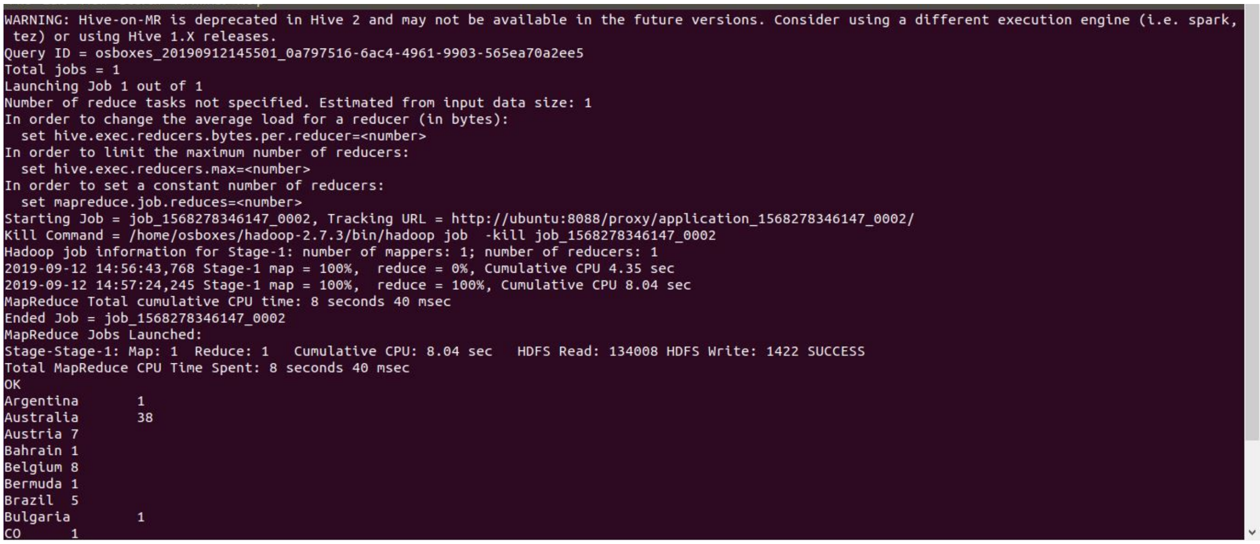
4. **Run the query**

   Select country, count(*) from sales group by country;

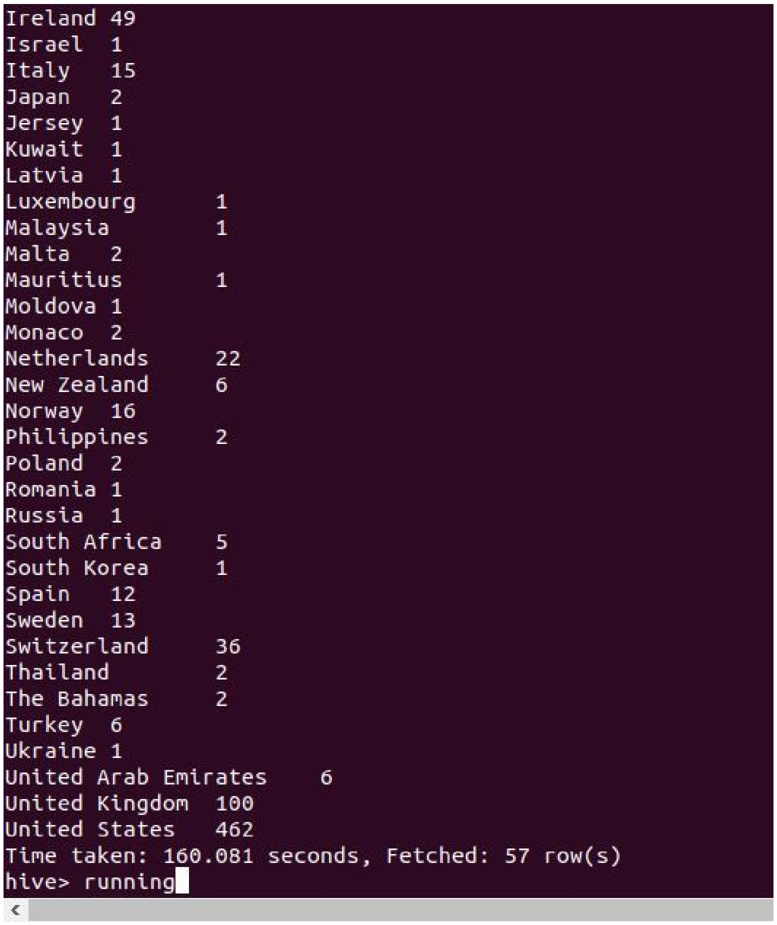


```
hive> select country,count(*) from sales group by country;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark,
 tez) or using Hive 1.X releases.
Query ID = osboxes_20190912145501_0a797516-6ac4-4961-9903-565ea70a2ee5
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 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>
```

Output:



```
WARNING: 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.
Query ID = osboxes_20190912145501_0a797516-6ac4-4961-9903-565ea70a2ee5
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 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_1568278346147_0002, Tracking URL = http://ubuntu:8088/proxy/application_1568278346147_0002/
Kill Command = /home/osboxes/hadoop-2.7.3/bin/hadoop job  -kill job_1568278346147_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-09-12 14:56:43,768 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.35 sec
2019-09-12 14:57:24,245 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.04 sec
MapReduce Total cumulative CPU time: 8 seconds 40 msec
Ended Job = job_1568278346147_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.04 sec   HDFS Read: 134008 HDFS Write: 1422 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 40 msec
OK
Argentina       1
Australia       38
Austria 7
Bahrain 1
Belgium 8
Bermuda 1
Brazil  5
Bulgaria        1
CO      1
```



```
Ireland 49
Israel  1
Italy   15
Japan   2
Jersey  1
Kuwait  1
Latvia  1
Luxembourg      1
Malaysia        1
Malta   2
Mauritius       1
Moldova 1
Monaco  2
Netherlands     22
New Zealand     6
Norway  16
Philippines     2
Poland  2
Romania 1
Russia  1
South Africa    5
South Korea     1
Spain   12
Sweden  13
Switzerland     36
Thailand        2
The Bahamas     2
Turkey  6
Ukraine 1
United Arab Emirates    6
United Kingdom  100
United States   462
Time taken: 160.081 seconds, Fetched: 57 row(s)
hive> running
```