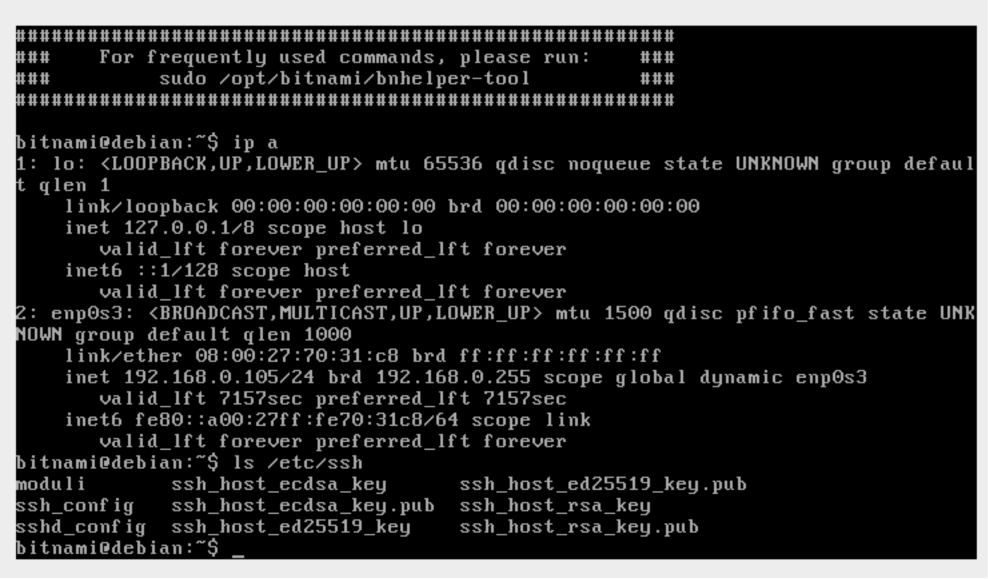
WEEK 6 ASSIGNMENT 2

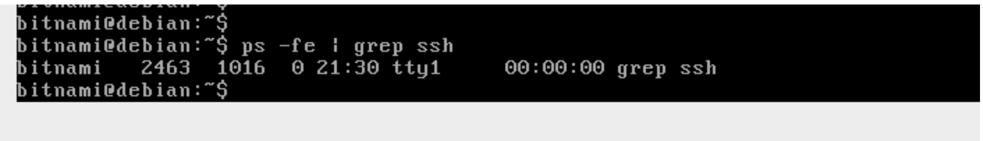
Large-Scale Data Storage Systems – DATA-5400 | Spring 2020

Christina Morgenstern

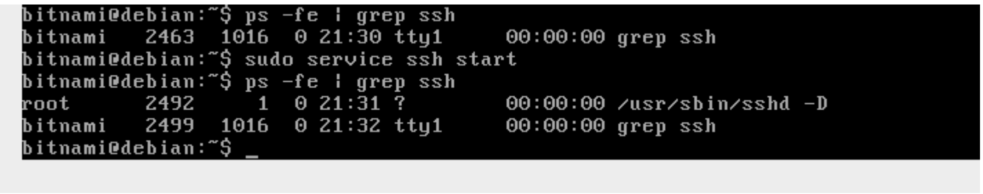
Creating an HDInsight Cluster didn´t work for me due to subscription issues. That´s why, I am using Bitnami Hadoop on my previously generated Virtual Machine.
Start the VM, log into Bitnami Hadoop and enable SSH so that I can connect to the VM from my Mac Terminal.



```
########################################################
###    For frequently used commands, please run:     ###
###         sudo /opt/bitnami/bnhelper-tool          ###
########################################################

bitnami@debian:~$ ip a
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group defaul
t qlen 1
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: enp0s3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UNK
NOWN group default qlen 1000
    link/ether 08:00:27:70:31:c8 brd ff:ff:ff:ff:ff:ff
    inet 192.168.0.105/24 brd 192.168.0.255 scope global dynamic enp0s3
       valid_lft 7157sec preferred_lft 7157sec
    inet6 fe80::a00:27ff:fe70:31c8/64 scope link
       valid_lft forever preferred_lft forever
bitnami@debian:~$ ls /etc/ssh
moduli       ssh_host_ecdsa_key      ssh_host_ed25519_key.pub
ssh_config   ssh_host_ecdsa_key.pub  ssh_host_rsa_key
sshd_config  ssh_host_ed25519_key    ssh_host_rsa_key.pub
bitnami@debian:~$ _
```

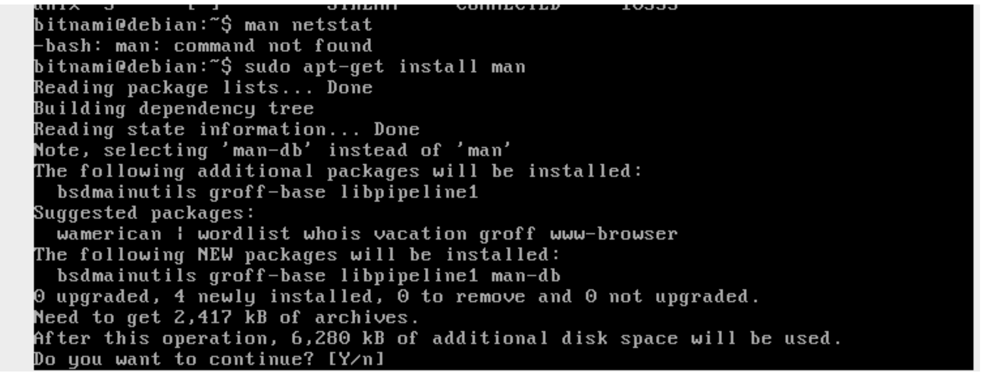


```
bitnami@debian:~$
bitnami@debian:~$ ps -fe | grep ssh
bitnami   2463  1016  0 21:30 tty1     00:00:00 grep ssh
bitnami@debian:~$
```



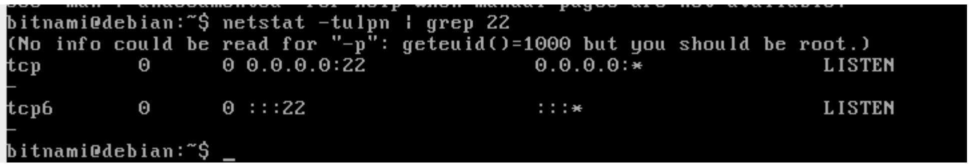
```
bitnami@debian:~$ ps -fe | grep ssh
bitnami   2463  1016  0 21:30 tty1     00:00:00 grep ssh
bitnami@debian:~$ sudo service ssh start
bitnami@debian:~$ ps -fe | grep ssh
root      2492     1  0 21:31 ?        00:00:00 /usr/sbin/sshd -D
bitnami   2499  1016  0 21:32 tty1     00:00:00 grep ssh
bitnami@debian:~$ _
```

Install documentation



```
bitnami@debian:~$ man netstat
-bash: man: command not found
bitnami@debian:~$ sudo apt-get install man
Reading package lists... Done
Building dependency tree
Reading state information... Done
Note, selecting 'man-db' instead of 'man'
The following additional packages will be installed:
  bsdmainutils groff-base libpipeline1
Suggested packages:
  wamerican | wordlist whois vacation groff www-browser
The following NEW packages will be installed:
  bsdmainutils groff-base libpipeline1 man-db
0 upgraded, 4 newly installed, 0 to remove and 0 not upgraded.
Need to get 2,417 kB of archives.
After this operation, 6,280 kB of additional disk space will be used.
Do you want to continue? [Y/n]
```

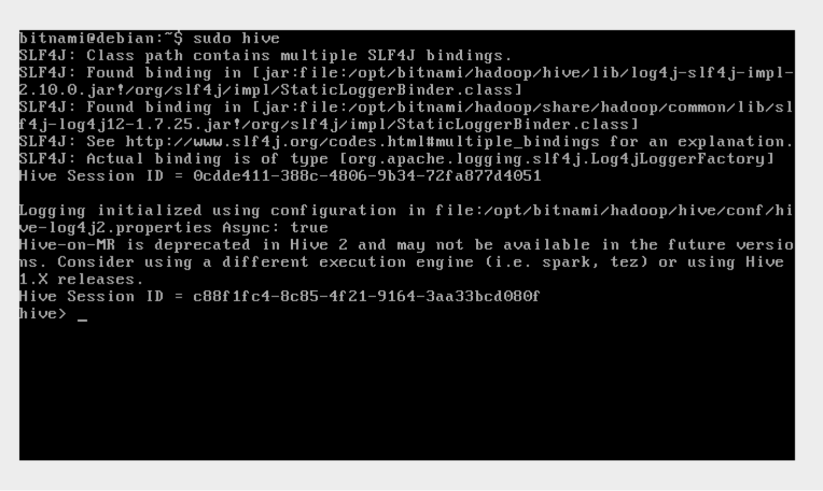
Check listening ports with `netstat.`



```
bitnami@debian:~$ netstat -tulpn | grep 22
(No info could be read for "-p": geteuid()=1000 but you should be root.)
tcp        0      0 0.0.0.0:22              0.0.0.0:*               LISTEN
-
tcp6       0      0 :::22                   :::*                    LISTEN
-
bitnami@debian:~$ _
```

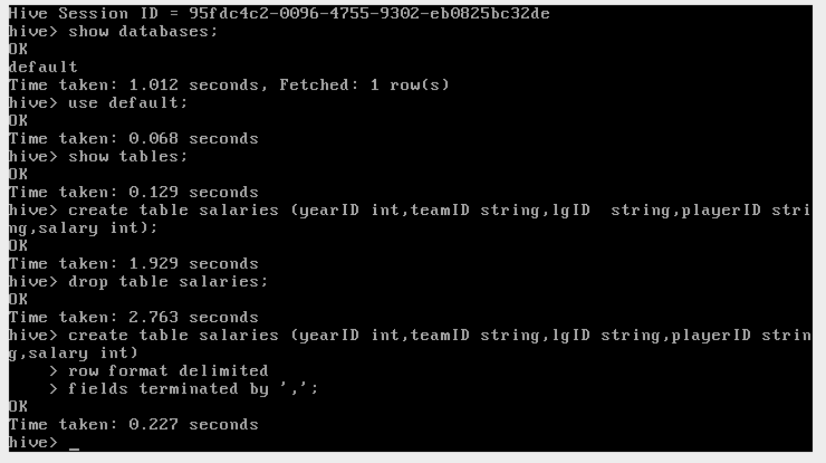
Test using Hive

Log into Hive by running command `sudo hive.`



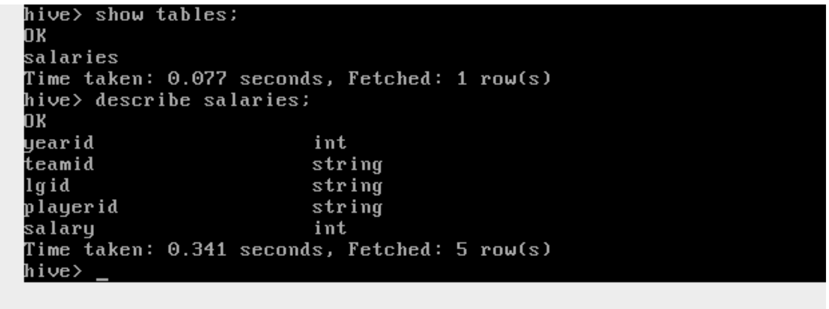
```
bitnami@debian:~$ sudo hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/bitnami/hadoop/hive/lib/log4j-slf4j-impl-
2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/bitnami/hadoop/share/hadoop/common/lib/sl
f4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
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 = 0cdde411-388c-4806-9b34-72fa877d4051

Logging initialized using configuration in file:/opt/bitnami/hadoop/hive/conf/hi
ve-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
Hive Session ID = c88f1fc4-8c85-4f21-9164-3aa33bcd080f
hive> _
```

Create a table in Hive.



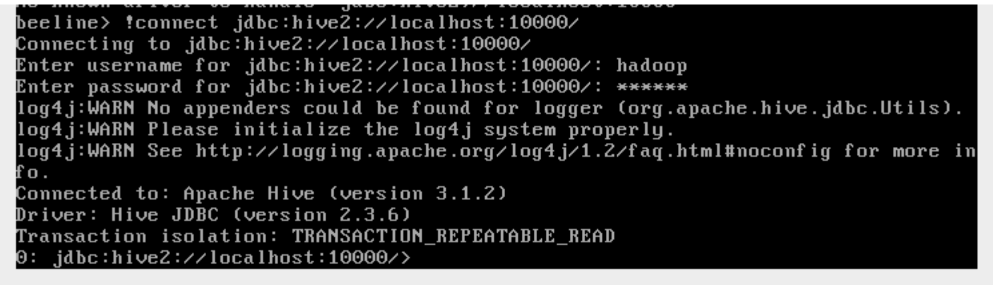
```
Hive Session ID = 95fdc4c2-0096-4755-9302-eb0825bc32de
hive> show databases;
OK
default
Time taken: 1.012 seconds, Fetched: 1 row(s)
hive> use default;
OK
Time taken: 0.068 seconds
hive> show tables;
OK
Time taken: 0.129 seconds
hive> create table salaries (yearID int,teamID string,lgID  string,playerID stri
ng,salary int);
OK
Time taken: 1.929 seconds
hive> drop table salaries;
OK
Time taken: 2.763 seconds
hive> create table salaries (yearID int,teamID string,lgID string,playerID strin
g,salary int)
    > row format delimited
    > fields terminated by ',';
OK
Time taken: 0.227 seconds
hive> _
```

Display tables and describe salaries table commands.



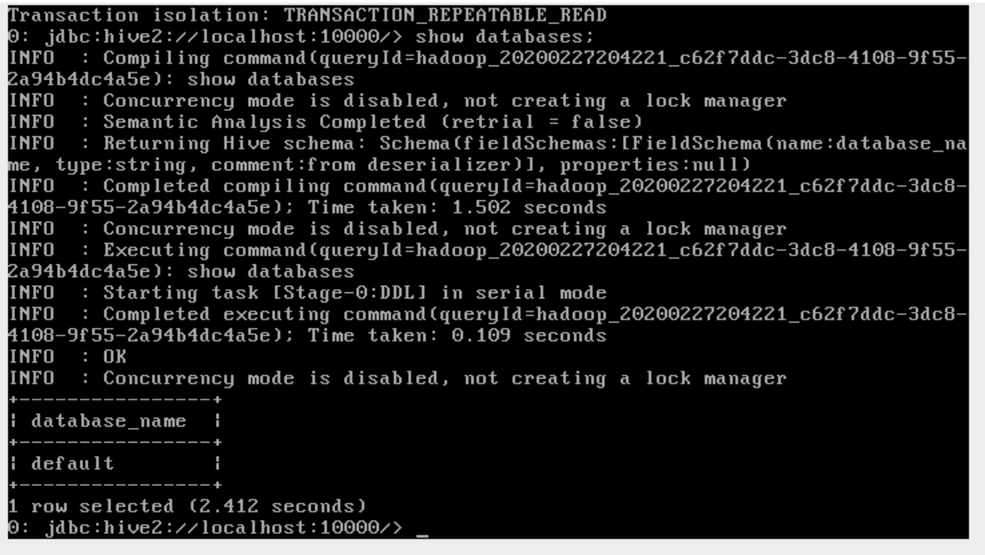
```
hive> show tables;
OK
salaries
Time taken: 0.077 seconds, Fetched: 1 row(s)
hive> describe salaries;
OK
yearid                  int
teamid                  string
lgid                    string
playerid                string
salary                  int
Time taken: 0.341 seconds, Fetched: 5 row(s)
hive> _
```

Login to and basic commands in Hive work, but preferred option is using beeline. Login to beeline using the `beeline` command and specifying username and password (both: hadoop). Connect to Hive afterwards using the following command: `!connect jdbc:hive2://localhost:10000/`



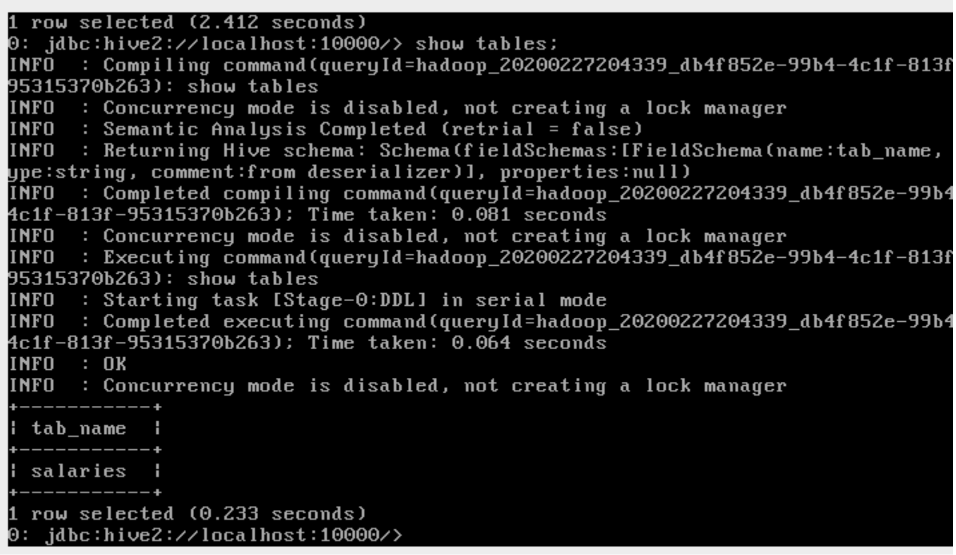
```
beeline> !connect jdbc:hive2://localhost:10000/
Connecting to jdbc:hive2://localhost:10000/
Enter username for jdbc:hive2://localhost:10000/: hadoop
Enter password for jdbc:hive2://localhost:10000/: ******
log4j:WARN No appenders could be found for logger (org.apache.hive.jdbc.Utils).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
Connected to: Apache Hive (version 3.1.2)
Driver: Hive JDBC (version 2.3.6)
Transaction isolation: TRANSACTION_REPEATABLE_READ
0: jdbc:hive2://localhost:10000/>
```

Run command `show databases;` to display available databases.



```
Transaction isolation: TRANSACTION_REPEATABLE_READ
0: jdbc:hive2://localhost:10000/> show databases;
INFO  : Compiling command(queryId=hadoop_20200227204221_c62f7ddc-3dc8-4108-9f55-
2a94b4dc4a5e): show databases
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:database_na
me, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200227204221_c62f7ddc-3dc8-
4108-9f55-2a94b4dc4a5e); Time taken: 1.502 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200227204221_c62f7ddc-3dc8-4108-9f55-
2a94b4dc4a5e): show databases
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hadoop_20200227204221_c62f7ddc-3dc8-
4108-9f55-2a94b4dc4a5e); Time taken: 0.109 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------+
| database_name  |
+----------------+
| default        |
+----------------+
1 row selected (2.412 seconds)
0: jdbc:hive2://localhost:10000/> _
```

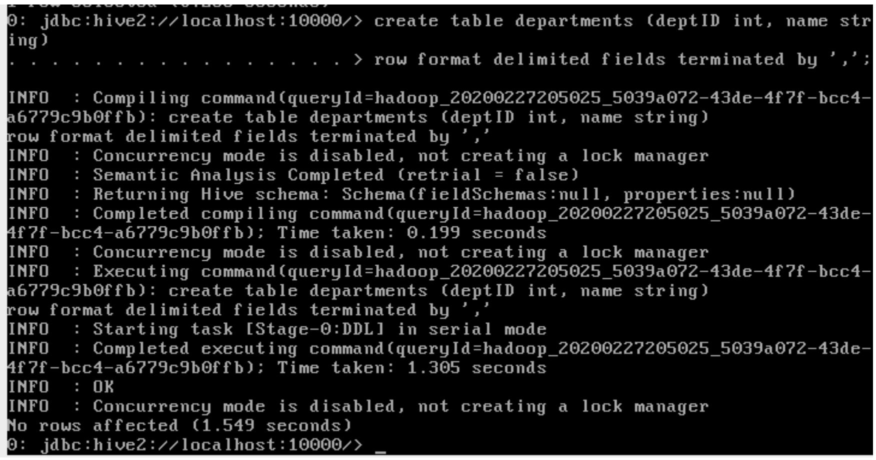
Run command `show tables;` to display tables.



```
1 row selected (2.412 seconds)
0: jdbc:hive2://localhost:10000/> show tables;
INFO  : Compiling command(queryId=hadoop_20200227204339_db4f852e-99b4-4c1f-813f
95315370b263): show tables
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name,
ype:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200227204339_db4f852e-99b4
4c1f-813f-95315370b263); Time taken: 0.081 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200227204339_db4f852e-99b4-4c1f-813f
95315370b263): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hadoop_20200227204339_db4f852e-99b4
4c1f-813f-95315370b263); Time taken: 0.064 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------+
| tab_name  |
+-----------+
| salaries  |
+-----------+
1 row selected (0.233 seconds)
0: jdbc:hive2://localhost:10000/>
```

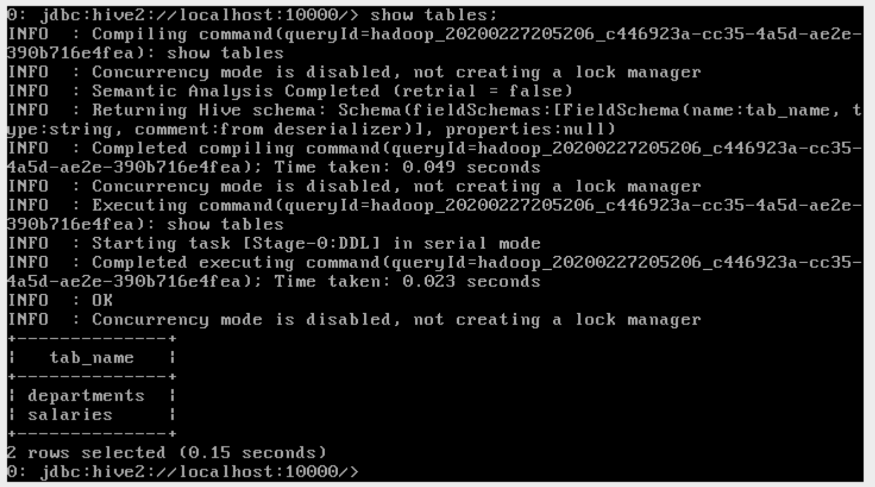
Create a table named departments with the columns deptID and name. The data types are integer for deptID and string for name.

```
create table departments (deptID int, name string)
row format delimited fields terminated by ',';
```



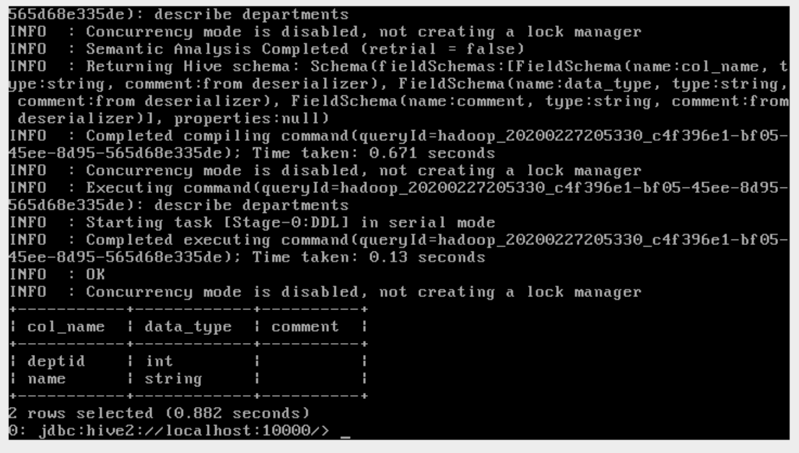
```
0: jdbc:hive2://localhost:10000/> create table departments (deptID int, name str
ing)
. . . . . . . . . . . . . . . . > row format delimited fields terminated by ',';

INFO  : Compiling command(queryId=hadoop_20200227205025_5039a072-43de-4f7f-bcc4-
a6779c9b0ffb): create table departments (deptID int, name string)
row format delimited fields terminated by ','
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200227205025_5039a072-43de-
4f7f-bcc4-a6779c9b0ffb); Time taken: 0.199 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200227205025_5039a072-43de-4f7f-bcc4-
a6779c9b0ffb): create table departments (deptID int, name string)
row format delimited fields terminated by ','
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hadoop_20200227205025_5039a072-43de-
4f7f-bcc4-a6779c9b0ffb); Time taken: 1.305 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (1.549 seconds)
0: jdbc:hive2://localhost:10000/> _
```

Check that table is created using the `show tables;` command.



```
0: jdbc:hive2://localhost:10000/> show tables;
INFO  : Compiling command(queryId=hadoop_20200227205206_c446923a-cc35-4a5d-ae2e-
390b716e4fea): show tables
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, t
ype:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200227205206_c446923a-cc35-
4a5d-ae2e-390b716e4fea); Time taken: 0.049 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200227205206_c446923a-cc35-4a5d-ae2e-
390b716e4fea): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hadoop_20200227205206_c446923a-cc35-
4a5d-ae2e-390b716e4fea); Time taken: 0.023 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------+
|   tab_name   |
+--------------+
| departments  |
| salaries     |
+--------------+
2 rows selected (0.15 seconds)
0: jdbc:hive2://localhost:10000/>
```

The `describe departments;` command lists the column names and data types.

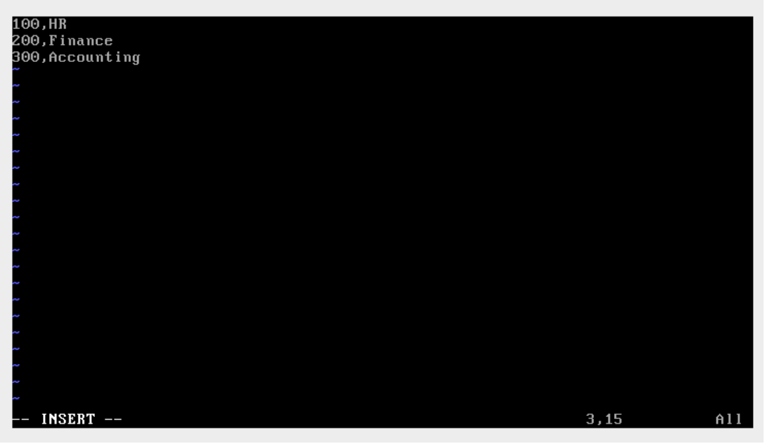


```
565d68e335de): describe departments
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, t
ype:string, comment:from deserializer), FieldSchema(name:data_type, type:string,
 comment:from deserializer), FieldSchema(name:comment, type:string, comment:from
 deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200227205330_c4f396e1-bf05-
45ee-8d95-565d68e335de); Time taken: 0.671 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200227205330_c4f396e1-bf05-45ee-8d95-
565d68e335de): describe departments
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hadoop_20200227205330_c4f396e1-bf05-
45ee-8d95-565d68e335de); Time taken: 0.13 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------+------------+----------+
| col_name  | data_type  | comment  |
+-----------+------------+----------+
| deptid    | int        |          |
| name      | string     |          |
+-----------+------------+----------+
2 rows selected (0.882 seconds)
0: jdbc:hive2://localhost:10000/> _
```

Quitting beeline `! quit`

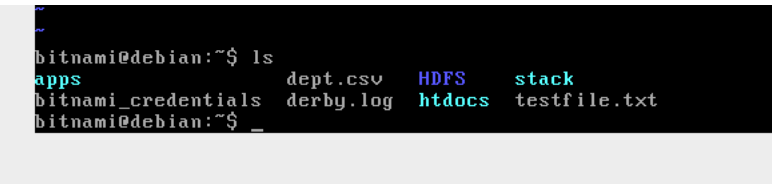
Create an internal table in Hive

Create a .csv file named dept.csv using the vim editor in the VM:
`vim dept.csv`
To edit the file press `i`.



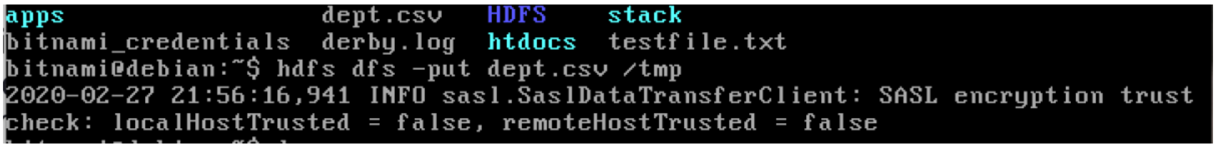
```
100,HR
200,Finance
300,Accounting
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
-- INSERT --                                          3,15          All
```

To save file hit `ESC`, type `:w` and `Enter` or `:x`



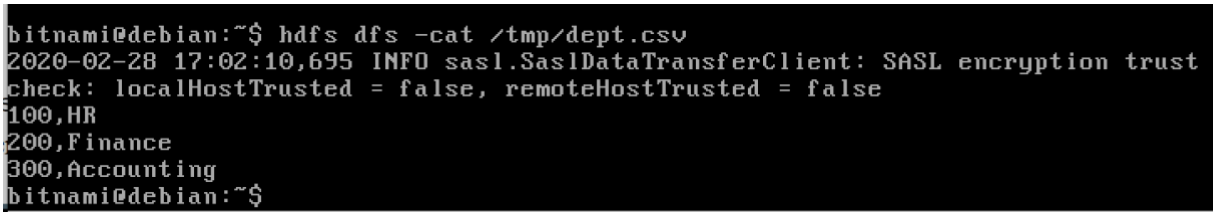
```
~
~
bitnami@debian:~$ ls
apps                  dept.csv   HDFS     stack
bitnami_credentials   derby.log  htdocs   testfile.txt
bitnami@debian:~$ _
```

Put dept.csv into /tmp directory.



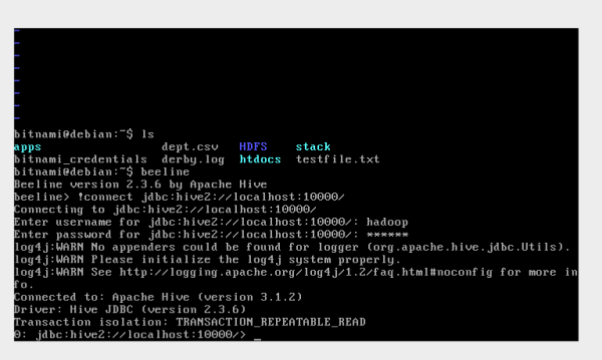
```
apps                  dept.csv   HDFS     stack
bitnami_credentials   derby.log  htdocs   testfile.txt
bitnami@debian:~$ hdfs dfs -put dept.csv /tmp
2020-02-27 21:56:16,941 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
```

Check for presence and content of dept.csv



```
bitnami@debian:~$ hdfs dfs -cat /tmp/dept.csv
2020-02-28 17:02:10,695 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
100,HR
200,Finance
300,Accounting
bitnami@debian:~$
```

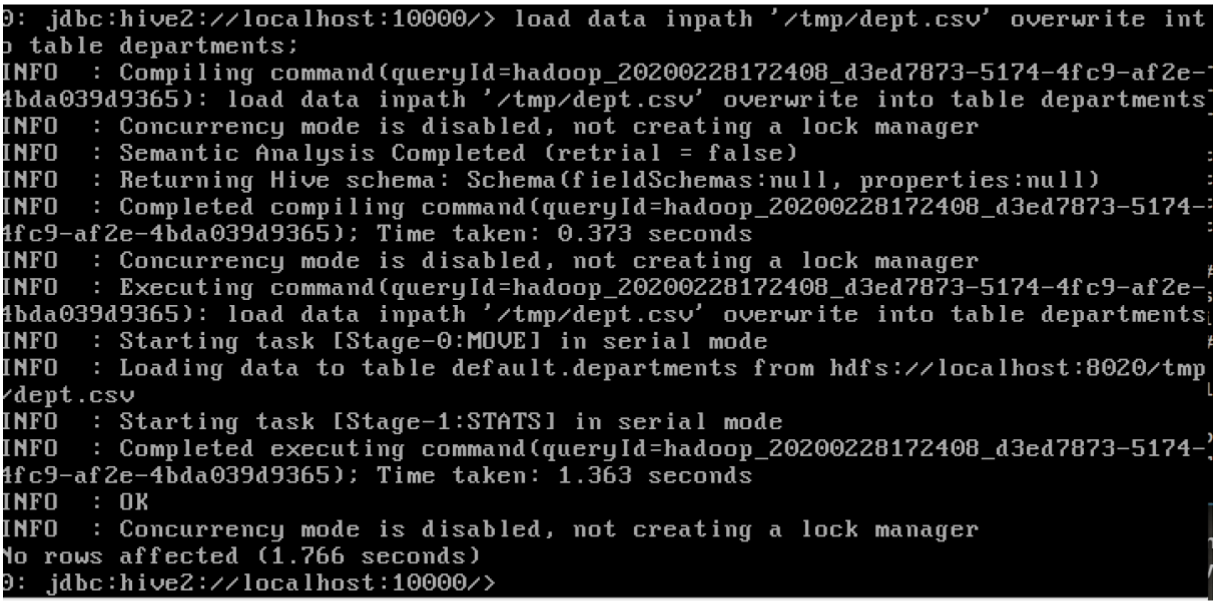
Connect to beeline again.



```
~
~
~
~
~
~
~
~
bitnami@debian:~$ ls
apps                  dept.csv   HDFS     stack
bitnami_credentials   derby.log  htdocs   testfile.txt
bitnami@debian:~$ beeline
Beeline version 2.3.6 by Apache Hive
beeline> !connect jdbc:hive2://localhost:10000/
Connecting to jdbc:hive2://localhost:10000/
Enter username for jdbc:hive2://localhost:10000/: hadoop
Enter password for jdbc:hive2://localhost:10000/: ******
log4j:WARN No appenders could be found for logger (org.apache.hive.jdbc.Utils).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
Connected to: Apache Hive (version 3.1.2)
Driver: Hive JDBC (version 2.3.6)
Transaction isolation: TRANSACTION_REPEATABLE_READ
0: jdbc:hive2://localhost:10000/> _
```

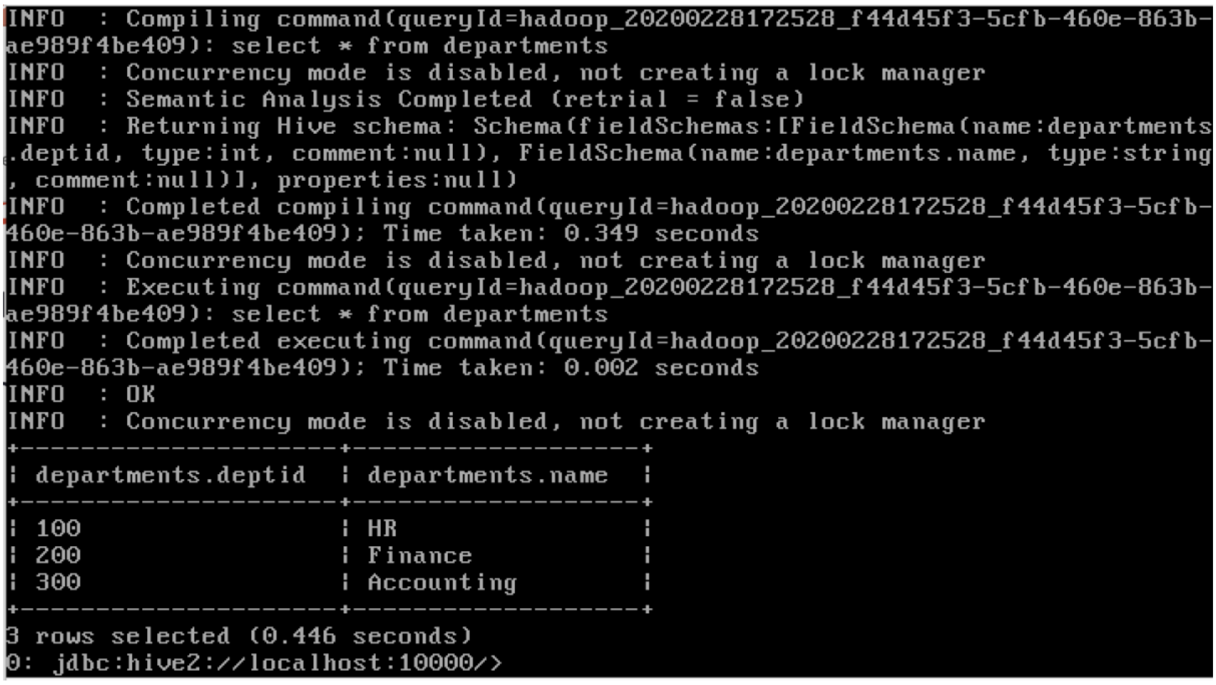
Load data from previously generated dept.csv and overwrite departments table. If we do not use local keyword, it assumes it as a HDFS Path.

`Load data inpath `/tmp/dept.csv`overwrite into table departments;`



```
0: jdbc:hive2://localhost:10000/> load data inpath '/tmp/dept.csv' overwrite int
o table departments;
INFO  : Compiling command(queryId=hadoop_20200228172408_d3ed7873-5174-4fc9-af2e-
4bda039d9365): load data inpath '/tmp/dept.csv' overwrite into table departments
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200228172408_d3ed7873-5174-
4fc9-af2e-4bda039d9365); Time taken: 0.373 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200228172408_d3ed7873-5174-4fc9-af2e-
4bda039d9365): load data inpath '/tmp/dept.csv' overwrite into table departments
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table default.departments from hdfs://localhost:8020/tmp
/dept.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hadoop_20200228172408_d3ed7873-5174-
4fc9-af2e-4bda039d9365); Time taken: 1.363 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (1.766 seconds)
0: jdbc:hive2://localhost:10000/>
```

Check contents of departments table using `select * from departments;` HiveQL command.



```
INFO  : Compiling command(queryId=hadoop_20200228172528_f44d45f3-5cfb-460e-863b-
ae989f4be409): select * from departments
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:departments
.deptid, type:int, comment:null), FieldSchema(name:departments.name, type:string
, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200228172528_f44d45f3-5cfb-
460e-863b-ae989f4be409); Time taken: 0.349 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200228172528_f44d45f3-5cfb-460e-863b-
ae989f4be409): select * from departments
INFO  : Completed executing command(queryId=hadoop_20200228172528_f44d45f3-5cfb-
460e-863b-ae989f4be409); Time taken: 0.002 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------------------+--------------------+
| departments.deptid  | departments.name   |
+---------------------+--------------------+
| 100                 | HR                 |
| 200                 | Finance            |
| 300                 | Accounting         |
+---------------------+--------------------+
3 rows selected (0.446 seconds)
0: jdbc:hive2://localhost:10000/>
```

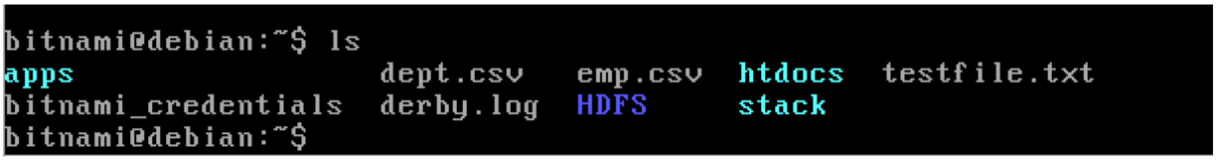
Creating an internal table in Hive worked.

Create an external table in Hive

Create a folder /tmp/emp

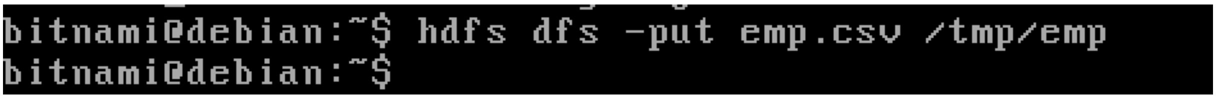
```
dfs dfs -mkdir /tmp/emp
```

Create csv file emp.csv using the vim-Editor.



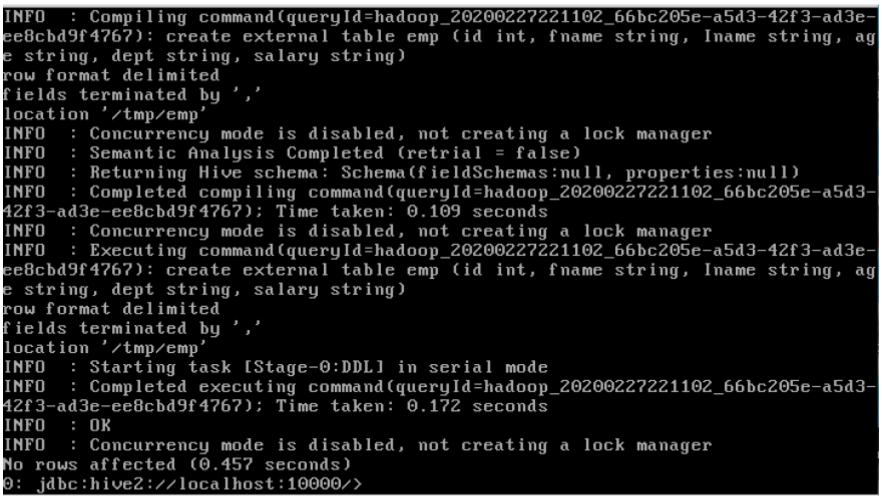
```
bitnami@debian:~$ ls
apps                  dept.csv   emp.csv   htdocs   testfile.txt
bitnami_credentials   derby.log  HDFS      stack
bitnami@debian:~$
```

Put emp.csv file into /tmp/emp



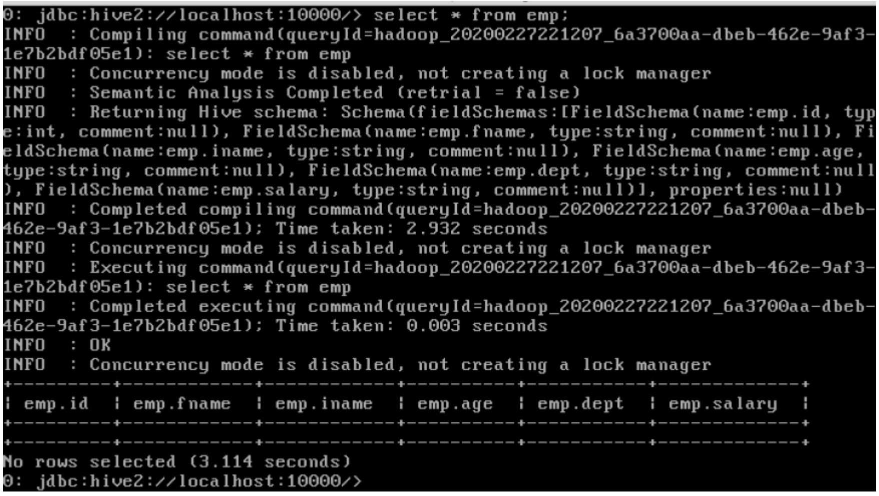
```
bitnami@debian:~$ hdfs dfs -put emp.csv /tmp/emp
bitnami@debian:~$
```

Connect to beeline and create an external table



```
INFO  : Compiling command(queryId=hadoop_20200227221102_66bc205e-a5d3-42f3-ad3e-
ee8cbd9f4767): create external table emp (id int, fname string, Iname string, ag
e string, dept string, salary string)
row format delimited
fields terminated by ','
location '/tmp/emp'
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200227221102_66bc205e-a5d3-
42f3-ad3e-ee8cbd9f4767); Time taken: 0.109 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200227221102_66bc205e-a5d3-42f3-ad3e-
ee8cbd9f4767): create external table emp (id int, fname string, Iname string, ag
e string, dept string, salary string)
row format delimited
fields terminated by ','
location '/tmp/emp'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hadoop_20200227221102_66bc205e-a5d3-
42f3-ad3e-ee8cbd9f4767); Time taken: 0.172 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.457 seconds)
0: jdbc:hive2://localhost:10000/>
```

Verify contents of table

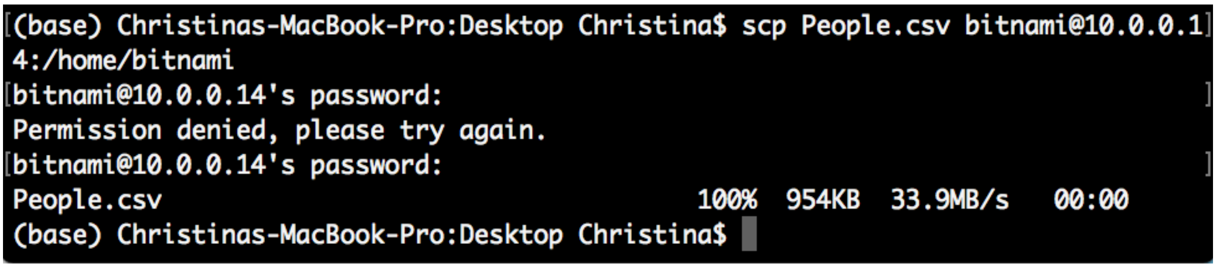


```
0: jdbc:hive2://localhost:10000/> select * from emp;
INFO  : Compiling command(queryId=hadoop_20200227221207_6a3700aa-dbeb-462e-9af3-
1e7b2bdf05e1): select * from emp
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:emp.id, typ
e:int, comment:null), FieldSchema(name:emp.fname, type:string, comment:null), Fi
eldSchema(name:emp.iname, type:string, comment:null), FieldSchema(name:emp.age,
type:string, comment:null), FieldSchema(name:emp.dept, type:string, comment:null
), FieldSchema(name:emp.salary, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200227221207_6a3700aa-dbeb-
462e-9af3-1e7b2bdf05e1); Time taken: 2.932 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200227221207_6a3700aa-dbeb-462e-9af3-
1e7b2bdf05e1): select * from emp
INFO  : Completed executing command(queryId=hadoop_20200227221207_6a3700aa-dbeb-
462e-9af3-1e7b2bdf05e1); Time taken: 0.003 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------+------------+------------+----------+-----------+-------------+
| emp.id  | emp.fname  | emp.iname  | emp.age  | emp.dept  | emp.salary  |
+---------+------------+------------+----------+-----------+-------------+
+---------+------------+------------+----------+-----------+-------------+
No rows selected (3.114 seconds)
0: jdbc:hive2://localhost:10000/>
```

Creating an external table in Hive worked.

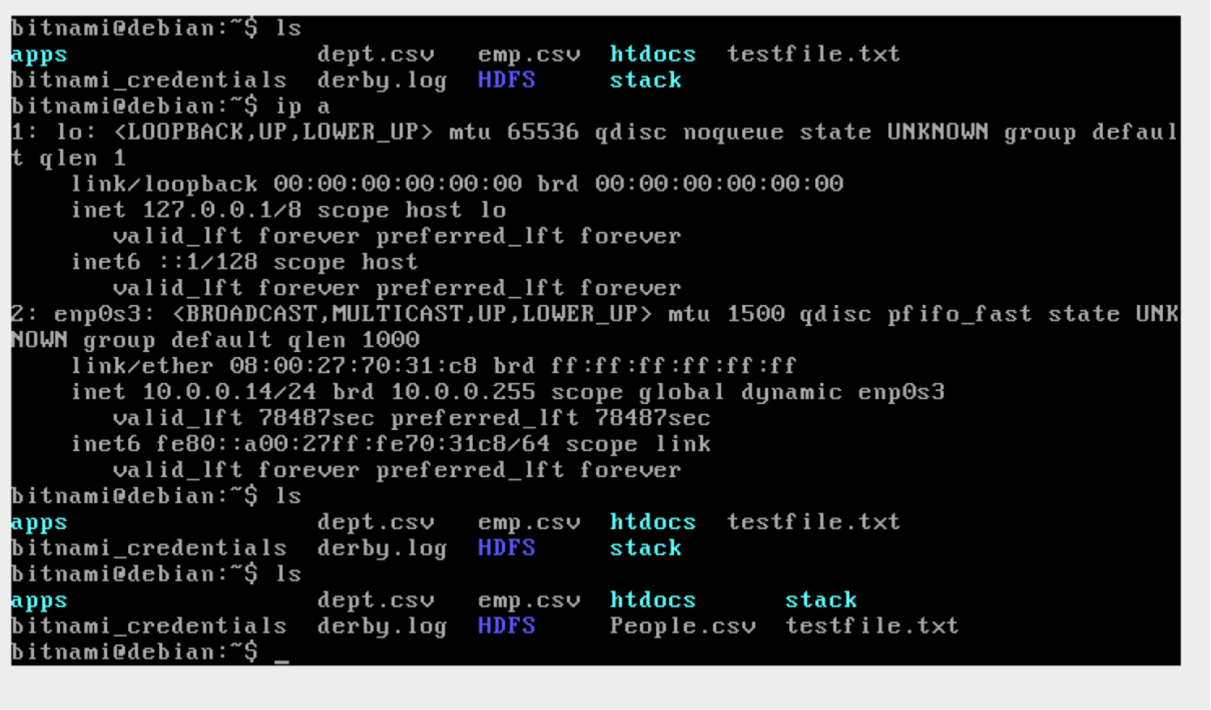
Create tables for baseball queries

Delete the columns not of use for the exercise and remove the header. Save the file as People.csv on my desktop.
Transfer the file to HDFS using SCP.



```
[(base) Christinas-MacBook-Pro:Desktop Christina$ scp People.csv bitnami@10.0.0.1]
4:/home/bitnami
[bitnami@10.0.0.14's password:                                                   ]
Permission denied, please try again.
[bitnami@10.0.0.14's password:                                                   ]
People.csv                                 100%  954KB  33.9MB/s   00:00
(base) Christinas-MacBook-Pro:Desktop Christina$ 
```

Verify transfer of file to HDFS



```
bitnami@debian:~$ ls
apps                 dept.csv   emp.csv  htdocs  testfile.txt
bitnami_credentials  derby.log  HDFS     stack
bitnami@debian:~$ ip a
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group defaul
t qlen 1
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: enp0s3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UNK
NOWN group default qlen 1000
    link/ether 08:00:27:70:31:c8 brd ff:ff:ff:ff:ff:ff
    inet 10.0.0.14/24 brd 10.0.0.255 scope global dynamic enp0s3
       valid_lft 78487sec preferred_lft 78487sec
    inet6 fe80::a00:27ff:fe70:31c8/64 scope link
       valid_lft forever preferred_lft forever
bitnami@debian:~$ ls
apps                 dept.csv   emp.csv  htdocs  testfile.txt
bitnami_credentials  derby.log  HDFS     stack
bitnami@debian:~$ ls
apps                 dept.csv   emp.csv  htdocs      stack
bitnami_credentials  derby.log  HDFS     People.csv  testfile.txt
bitnami@debian:~$ _
```

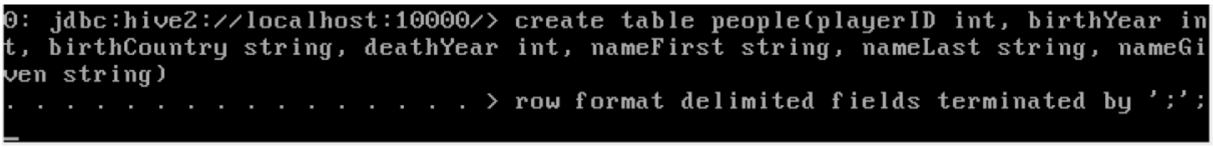
Transfer People.csv to tmp directory

```
hdfs dfs -put People.csv /tmp
```

Check contents of People.csv

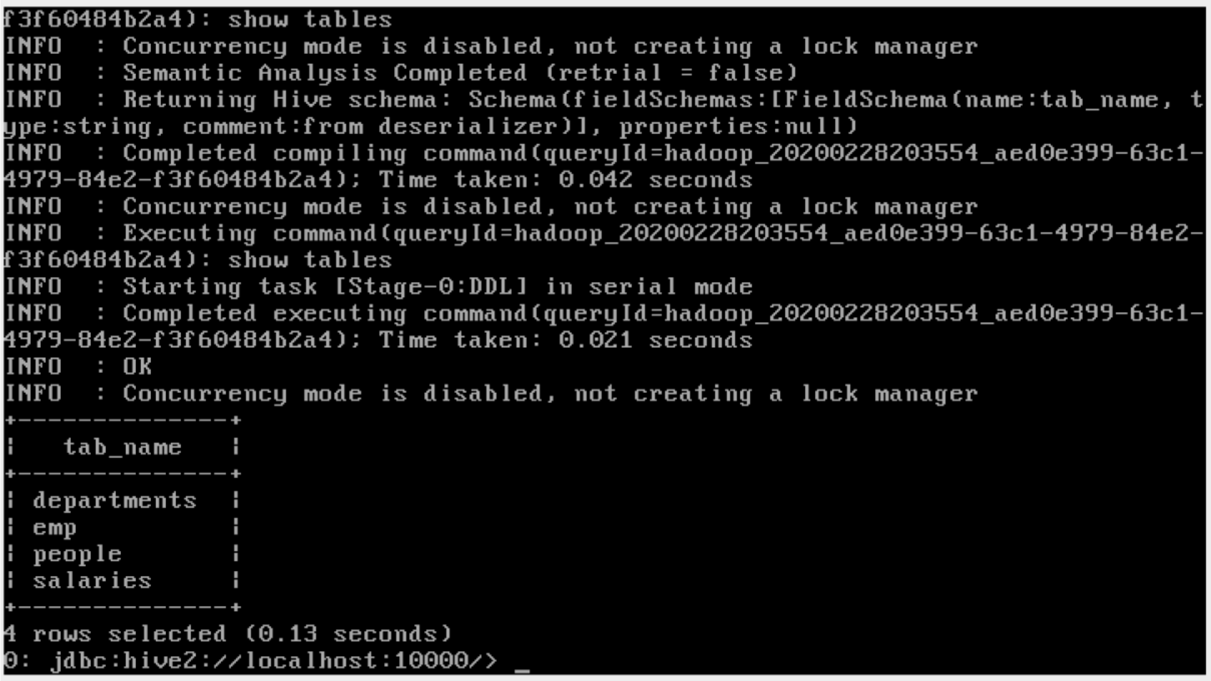
```
hdfs dfs -cat /tmp/People.csv
```

Connect to beeline and create a table named people with columns playerID, birthYear, birthCountry, deathYear, nameFirst, nameLast, nameGiven.



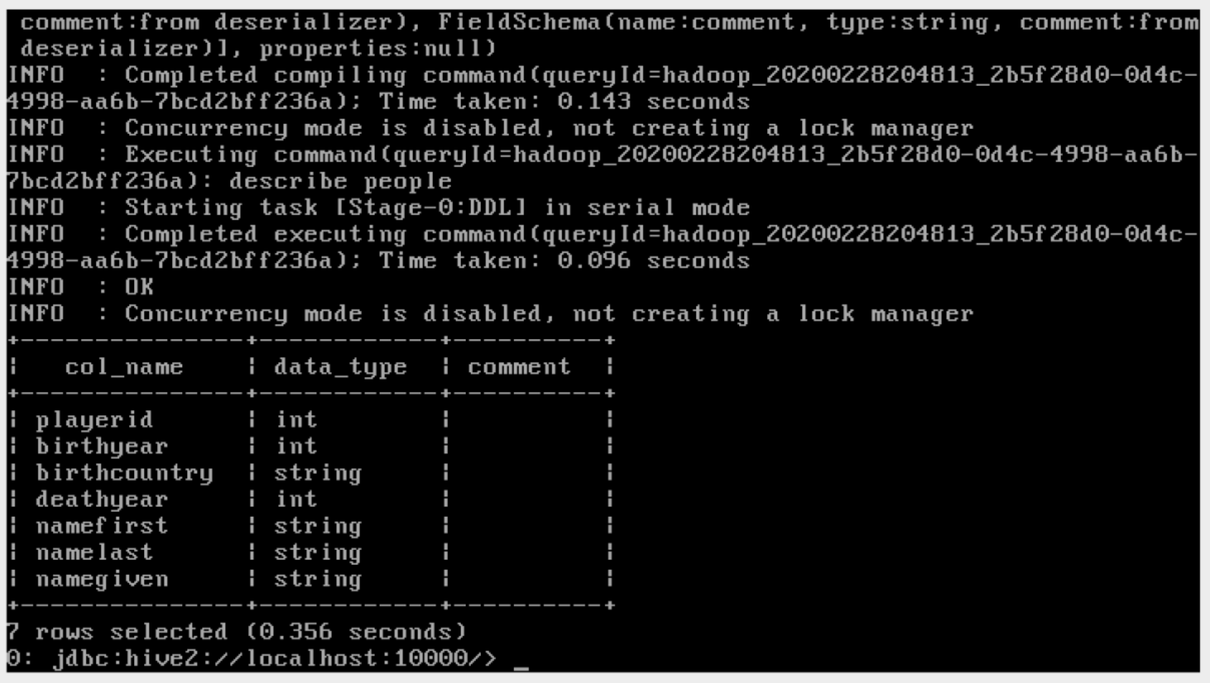
```
0: jdbc:hive2://localhost:10000/> create table people(playerID int, birthYear in
t, birthCountry string, deathYear int, nameFirst string, nameLast string, nameGi
ven string)
. . . . . . . . . . . . . . . . > row format delimited fields terminated by ';';
_
```

Verify creation of table



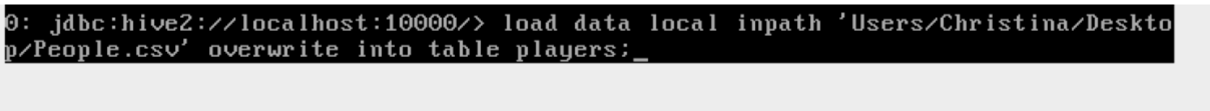
```
f3f60484b2a4): show tables
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, t
ype:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200228203554_aed0e399-63c1-
4979-84e2-f3f60484b2a4); Time taken: 0.042 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200228203554_aed0e399-63c1-4979-84e2-
f3f60484b2a4): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hadoop_20200228203554_aed0e399-63c1-
4979-84e2-f3f60484b2a4); Time taken: 0.021 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------+
¦   tab_name   ¦
+--------------+
¦ departments  ¦
¦ emp          ¦
¦ people       ¦
¦ salaries     ¦
+--------------+
4 rows selected (0.13 seconds)
0: jdbc:hive2://localhost:10000/> _
```

Describe table

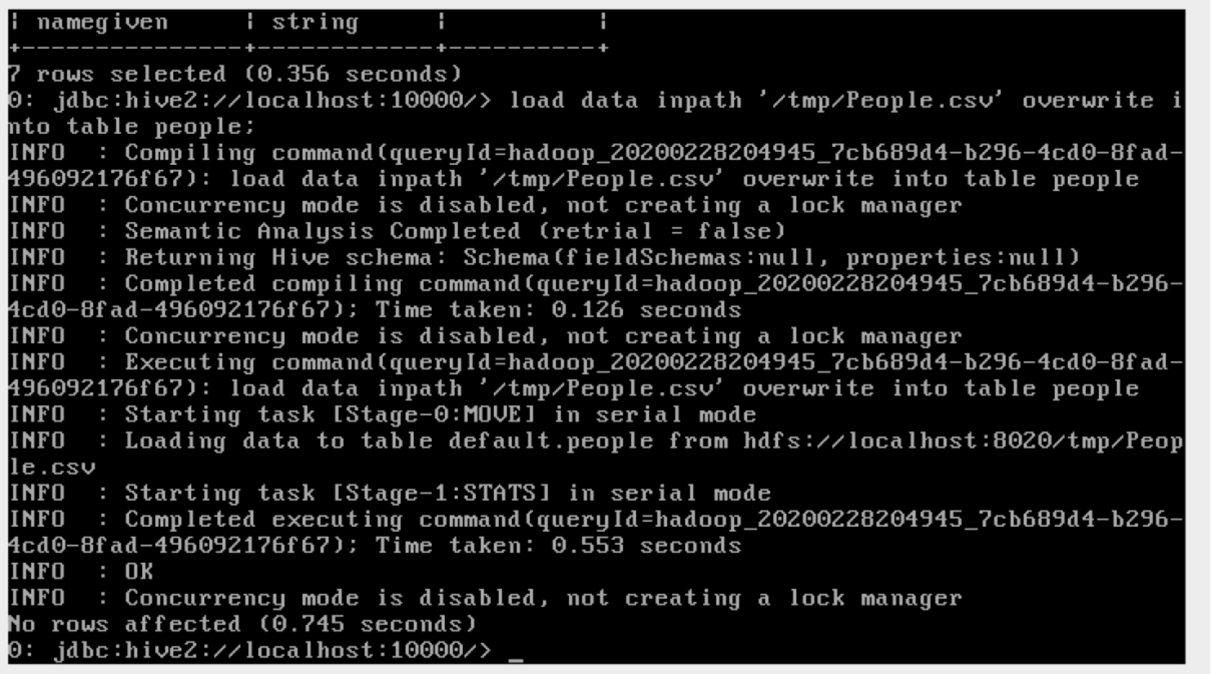


```
 comment:from deserializer), FieldSchema(name:comment, type:string, comment:from
 deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200228204813_2b5f28d0-0d4c-
4998-aa6b-7bcd2bff236a); Time taken: 0.143 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200228204813_2b5f28d0-0d4c-4998-aa6b-
7bcd2bff236a): describe people
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hadoop_20200228204813_2b5f28d0-0d4c-
4998-aa6b-7bcd2bff236a); Time taken: 0.096 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------------+------------+----------+
¦   col_name    ¦ data_type  ¦ comment  ¦
+---------------+------------+----------+
¦ playerid      ¦ int        ¦          ¦
¦ birthyear     ¦ int        ¦          ¦
¦ birthcountry  ¦ string     ¦          ¦
¦ deathyear     ¦ int        ¦          ¦
¦ namefirst     ¦ string     ¦          ¦
¦ namelast      ¦ string     ¦          ¦
¦ namegiven     ¦ string     ¦          ¦
+---------------+------------+----------+
7 rows selected (0.356 seconds)
0: jdbc:hive2://localhost:10000/> _
```

Load data into table



```
0: jdbc:hive2://localhost:10000/> load data local inpath 'Users/Christina/Deskto
p/People.csv' overwrite into table players;_
```

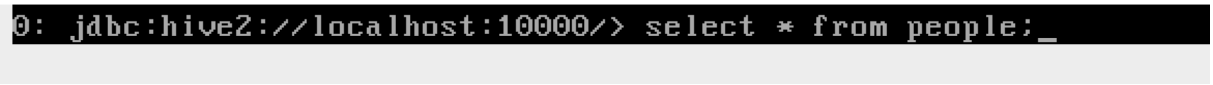


```
¦ namegiven      ¦ string      ¦           ¦
+---------------+------------+----------+
7 rows selected (0.356 seconds)
0: jdbc:hive2://localhost:10000/> load data inpath '/tmp/People.csv' overwrite i
nto table people;
INFO  : Compiling command(queryId=hadoop_20200228204945_7cb689d4-b296-4cd0-8fad-
496092176f67): load data inpath '/tmp/People.csv' overwrite into table people
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200228204945_7cb689d4-b296-
4cd0-8fad-496092176f67); Time taken: 0.126 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200228204945_7cb689d4-b296-4cd0-8fad-
496092176f67): load data inpath '/tmp/People.csv' overwrite into table people
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table default.people from hdfs://localhost:8020/tmp/Peop
le.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hadoop_20200228204945_7cb689d4-b296-
4cd0-8fad-496092176f67); Time taken: 0.553 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.745 seconds)
0: jdbc:hive2://localhost:10000/> _
```

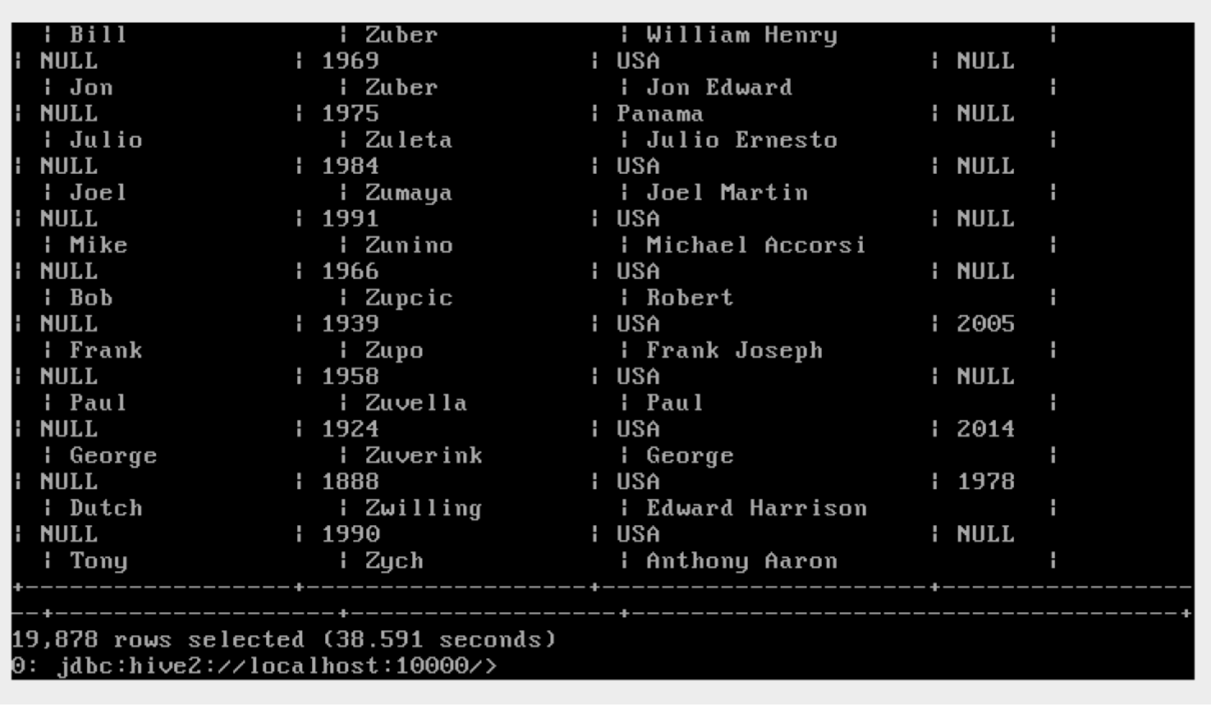
Answer the questions from assignment 1.

a. What is the total number of baseball players?

```
SELECT * FROM People;
```



```
0: jdbc:hive2://localhost:10000/> select * from people;_
```

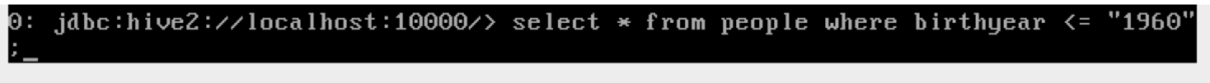


```
  ¦ Bill              ¦ Zuber              ¦ William Henry              ¦
¦ NULL              ¦ 1969               ¦ USA                ¦ NULL
  ¦ Jon               ¦ Zuber              ¦ Jon Edward                 ¦
¦ NULL              ¦ 1975               ¦ Panama             ¦ NULL
  ¦ Julio             ¦ Zuleta             ¦ Julio Ernesto              ¦
¦ NULL              ¦ 1984               ¦ USA                ¦ NULL
  ¦ Joel              ¦ Zumaya             ¦ Joel Martin                ¦
¦ NULL              ¦ 1991               ¦ USA                ¦ NULL
  ¦ Mike              ¦ Zunino             ¦ Michael Accorsi            ¦
¦ NULL              ¦ 1966               ¦ USA                ¦ NULL
  ¦ Bob               ¦ Zupcic             ¦ Robert                     ¦
¦ NULL              ¦ 1939               ¦ USA                ¦ 2005
  ¦ Frank             ¦ Zupo               ¦ Frank Joseph               ¦
¦ NULL              ¦ 1958               ¦ USA                ¦ NULL
  ¦ Paul              ¦ Zuvella            ¦ Paul                       ¦
¦ NULL              ¦ 1924               ¦ USA                ¦ 2014
  ¦ George            ¦ Zuverink           ¦ George                     ¦
¦ NULL              ¦ 1888               ¦ USA                ¦ 1978
  ¦ Dutch             ¦ Zwilling           ¦ Edward Harrison            ¦
¦ NULL              ¦ 1990               ¦ USA                ¦ NULL
  ¦ Tony              ¦ Zych               ¦ Anthony Aaron              ¦
+-------------------+--------------------+----------------------+-----------------
--+-------------------+--------------------+-------------------------------------+
19,878 rows selected (38.591 seconds)
0: jdbc:hive2://localhost:10000/>
```

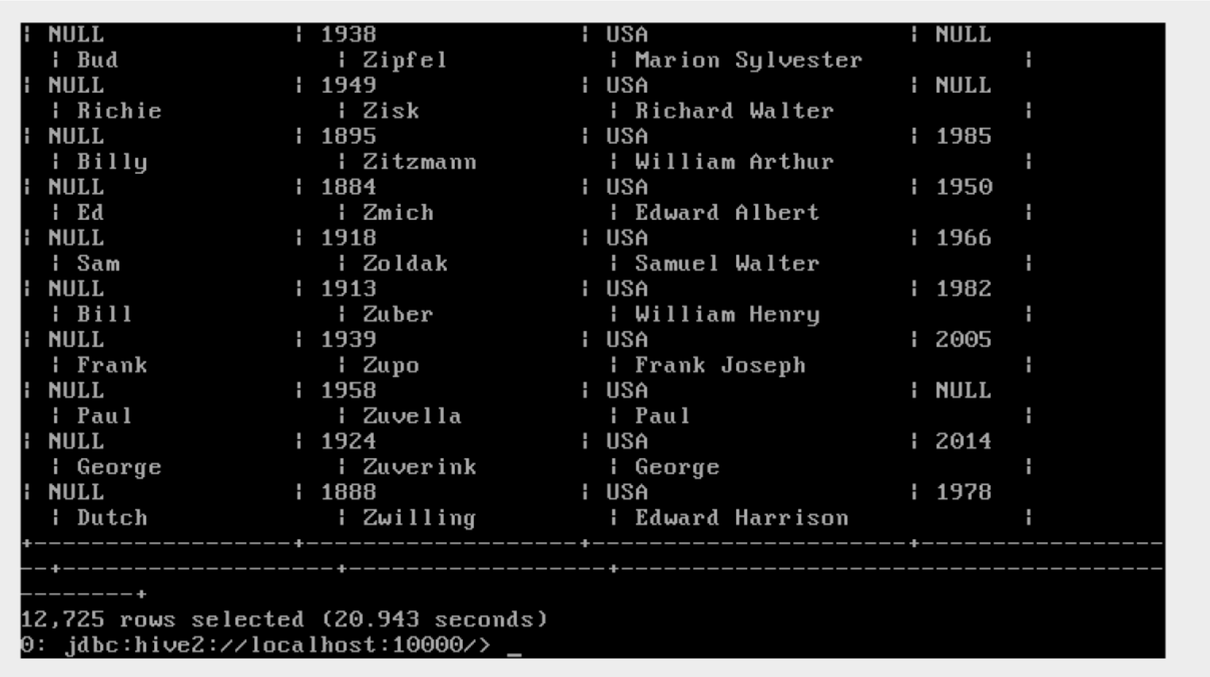
The total number of players in the database is 19,878.

b. How many players were born in the year 1960 and earlier?

```
SELECT * FROM People WHERE birthYear <= “1960”;
```



```
0: jdbc:hive2://localhost:10000/> select * from people where birthyear <= "1960"
;_
```

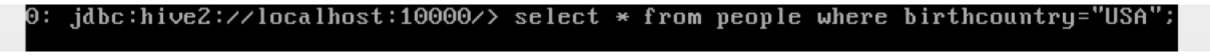


```
¦ NULL              ¦ 1938              ¦ USA                   ¦ NULL
  ¦ Bud               ¦ Zipfel            ¦ Marion Sylvester            ¦
¦ NULL              ¦ 1949              ¦ USA                   ¦ NULL
  ¦ Richie            ¦ Zisk              ¦ Richard Walter              ¦
¦ NULL              ¦ 1895              ¦ USA                   ¦ 1985
  ¦ Billy             ¦ Zitzmann          ¦ William Arthur              ¦
¦ NULL              ¦ 1884              ¦ USA                   ¦ 1950
  ¦ Ed                ¦ Zmich             ¦ Edward Albert               ¦
¦ NULL              ¦ 1918              ¦ USA                   ¦ 1966
  ¦ Sam               ¦ Zoldak            ¦ Samuel Walter               ¦
¦ NULL              ¦ 1913              ¦ USA                   ¦ 1982
  ¦ Bill              ¦ Zuber             ¦ William Henry               ¦
¦ NULL              ¦ 1939              ¦ USA                   ¦ 2005
  ¦ Frank             ¦ Zupo              ¦ Frank Joseph                ¦
¦ NULL              ¦ 1958              ¦ USA                   ¦ NULL
  ¦ Paul              ¦ Zuvella           ¦ Paul                        ¦
¦ NULL              ¦ 1924              ¦ USA                   ¦ 2014
  ¦ George            ¦ Zuverink          ¦ George                      ¦
¦ NULL              ¦ 1888              ¦ USA                   ¦ 1978
  ¦ Dutch             ¦ Zwilling          ¦ Edward Harrison             ¦
+-------------------+-------------------+-----------------------+-----------------
--+-------------------+-------------------+-----------------------------------------
--------+
12,725 rows selected (20.943 seconds)
0: jdbc:hive2://localhost:10000/> _
```

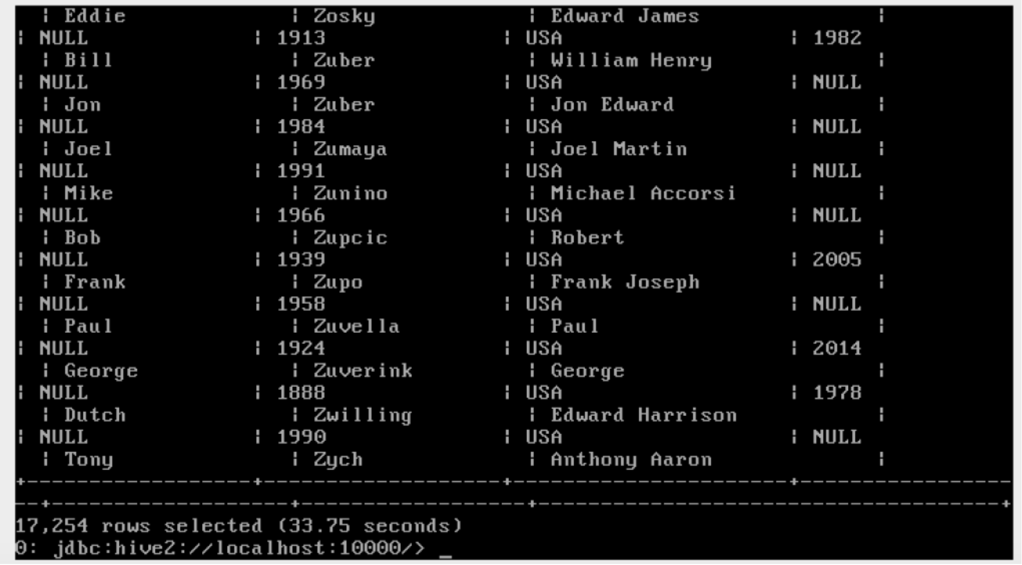
The number of players born in 1960 or before is 12,725.

c. How many players were born in the USA?

```
SELECT * FROM People WHERE birthCountry = “USA”;
```



```
0: jdbc:hive2://localhost:10000/> select * from people where birthcountry="USA";
```

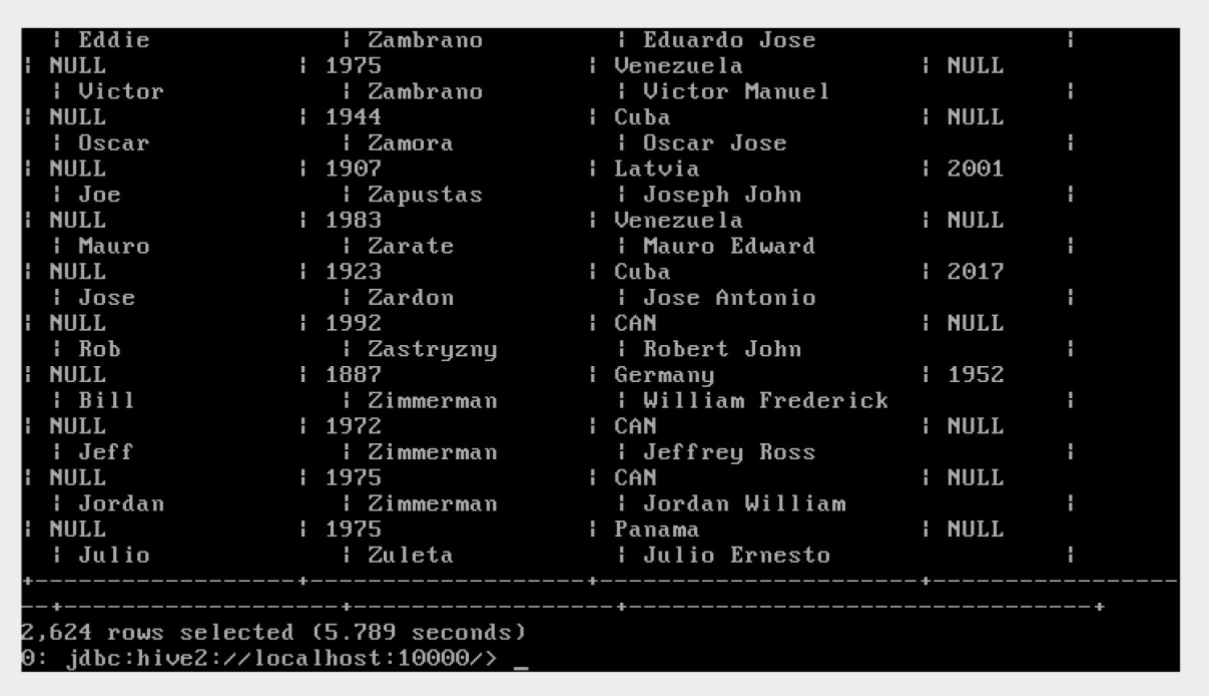


```
  ¦ Eddie             ¦ Zosky             ¦ Edward James              ¦
¦ NULL              ¦ 1913              ¦ USA                 ¦ 1982
  ¦ Bill              ¦ Zuber             ¦ William Henry             ¦
¦ NULL              ¦ 1969              ¦ USA                 ¦ NULL
  ¦ Jon               ¦ Zuber             ¦ Jon Edward                ¦
¦ NULL              ¦ 1984              ¦ USA                 ¦ NULL
  ¦ Joel              ¦ Zumaya            ¦ Joel Martin               ¦
¦ NULL              ¦ 1991              ¦ USA                 ¦ NULL
  ¦ Mike              ¦ Zunino            ¦ Michael Accorsi           ¦
¦ NULL              ¦ 1966              ¦ USA                 ¦ NULL
  ¦ Bob               ¦ Zupcic            ¦ Robert                    ¦
¦ NULL              ¦ 1939              ¦ USA                 ¦ 2005
  ¦ Frank             ¦ Zupo              ¦ Frank Joseph              ¦
¦ NULL              ¦ 1958              ¦ USA                 ¦ NULL
  ¦ Paul              ¦ Zuvella           ¦ Paul                      ¦
¦ NULL              ¦ 1924              ¦ USA                 ¦ 2014
  ¦ George            ¦ Zuverink          ¦ George                    ¦
¦ NULL              ¦ 1888              ¦ USA                 ¦ 1978
  ¦ Dutch             ¦ Zwilling          ¦ Edward Harrison           ¦
¦ NULL              ¦ 1990              ¦ USA                 ¦ NULL
  ¦ Tony              ¦ Zych              ¦ Anthony Aaron             ¦
+-------------------+-------------------+---------------------+-----------------
--+-------------------+-------------------+-----------------------------------+
17,254 rows selected (33.75 seconds)
0: jdbc:hive2://localhost:10000/> _
```

The number of players born in the USA is 17,254.

d. How many players were born outside the USA?

```
SELECT * FROM People WHERE birthCountry != “USA”;
```

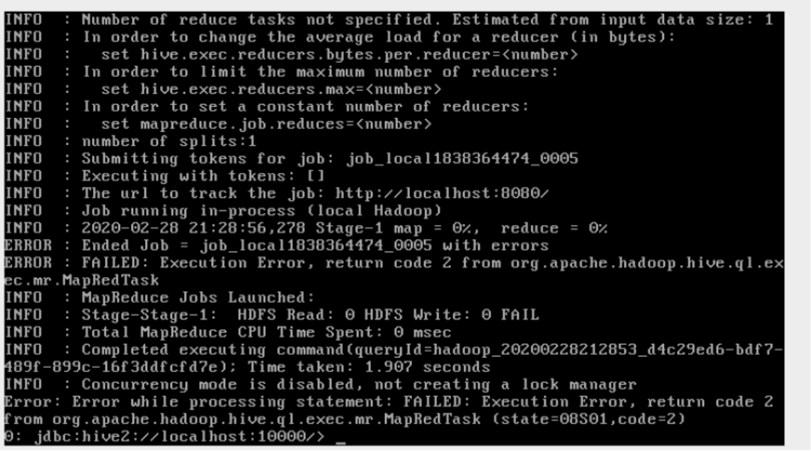


```
  ¦ Eddie               ¦ Zambrano          ¦ Eduardo Jose                ¦
¦ NULL              ¦ 1975                ¦ Venezuela             ¦ NULL
  ¦ Victor              ¦ Zambrano          ¦ Victor Manuel               ¦
¦ NULL              ¦ 1944                ¦ Cuba                  ¦ NULL
  ¦ Oscar               ¦ Zamora            ¦ Oscar Jose                  ¦
¦ NULL              ¦ 1907                ¦ Latvia                ¦ 2001
  ¦ Joe                 ¦ Zapustas          ¦ Joseph John                 ¦
¦ NULL              ¦ 1983                ¦ Venezuela             ¦ NULL
  ¦ Mauro               ¦ Zarate            ¦ Mauro Edward                ¦
¦ NULL              ¦ 1923                ¦ Cuba                  ¦ 2017
  ¦ Jose                ¦ Zardon            ¦ Jose Antonio                ¦
¦ NULL              ¦ 1992                ¦ CAN                   ¦ NULL
  ¦ Rob                 ¦ Zastryzny         ¦ Robert John                 ¦
¦ NULL              ¦ 1887                ¦ Germany               ¦ 1952
  ¦ Bill                ¦ Zimmerman         ¦ William Frederick           ¦
¦ NULL              ¦ 1972                ¦ CAN                   ¦ NULL
  ¦ Jeff                ¦ Zimmerman         ¦ Jeffrey Ross                ¦
¦ NULL              ¦ 1975                ¦ CAN                   ¦ NULL
  ¦ Jordan              ¦ Zimmerman         ¦ Jordan William              ¦
¦ NULL              ¦ 1975                ¦ Panama                ¦ NULL
  ¦ Julio               ¦ Zuleta            ¦ Julio Ernesto               ¦
+-------------------+---------------------+-----------------------+----------------
--+-------------------+-------------------+----------------------------+
2,624 rows selected (5.789 seconds)
0: jdbc:hive2://localhost:10000/> _
```

2,624 players were born outside the USA.

e. Display the number of players born in each year starting from 1960 thru 2000. For example, the output should show: 1980 4 (where 4 is the number of players born in 1980)

```
SELECT birthYear, COUNT(*) FROM People.csv GROUP BY  birthyear;
```

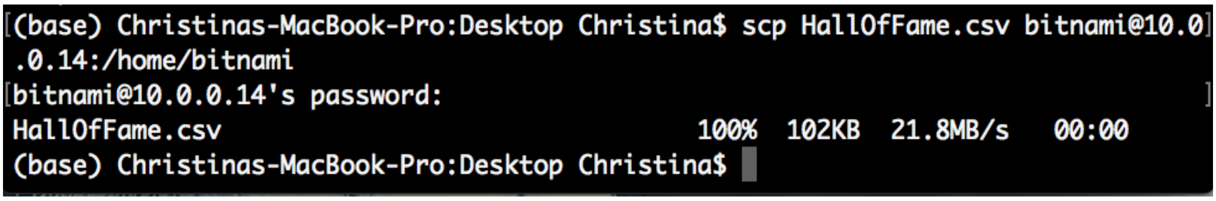


```
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_local1838364474_0005
INFO  : Executing with tokens: []
INFO  : The url to track the job: http://localhost:8080/
INFO  : Job running in-process (local Hadoop)
INFO  : 2020-02-28 21:28:56,278 Stage-1 map = 0%,  reduce = 0%
ERROR : Ended Job = job_local1838364474_0005 with errors
ERROR : FAILED: Execution Error, return code 2 from org.apache.hadoop.hive.ql.ex
ec.mr.MapRedTask
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1:  HDFS Read: 0 HDFS Write: 0 FAIL
INFO  : Total MapReduce CPU Time Spent: 0 msec
INFO  : Completed executing command(queryId=hadoop_20200228212853_d4c29ed6-bdf7-
489f-899c-16f3ddfcfd7e); Time taken: 1.907 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
Error: Error while processing statement: FAILED: Execution Error, return code 2
from org.apache.hadoop.hive.ql.exec.mr.MapRedTask (state=08S01,code=2)
0: jdbc:hive2://localhost:10000/> _
```

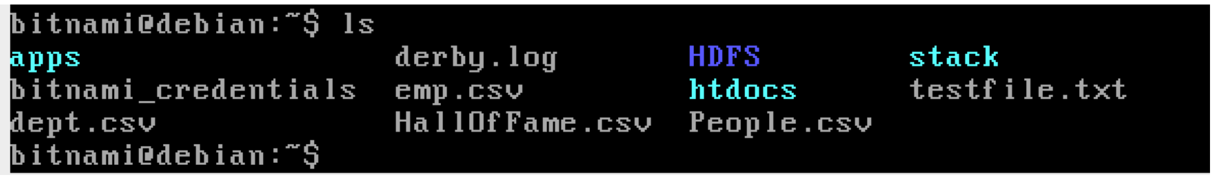
This command raised an error.

f. How many players and managers were inducted into the Hall of Fame?

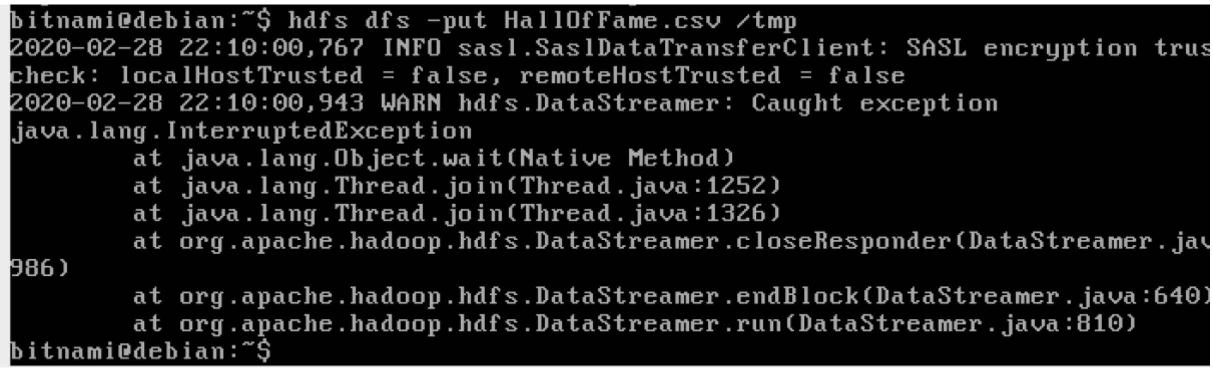
To answer this question, the HallOfFame.csv table was used. The csv file was amended to have only playerID, yearID, inducted and category was chosen. The other columns and the header were deleted. The file was transferred to the VM using SCP.



```
[(base) Christinas-MacBook-Pro:Desktop Christina$ scp HallOfFame.csv bitnami@10.0]
 .0.14:/home/bitnami
[bitnami@10.0.0.14's password:                                                    ]
 HallOfFame.csv                              100%  102KB  21.8MB/s   00:00
 (base) Christinas-MacBook-Pro:Desktop Christina$
```

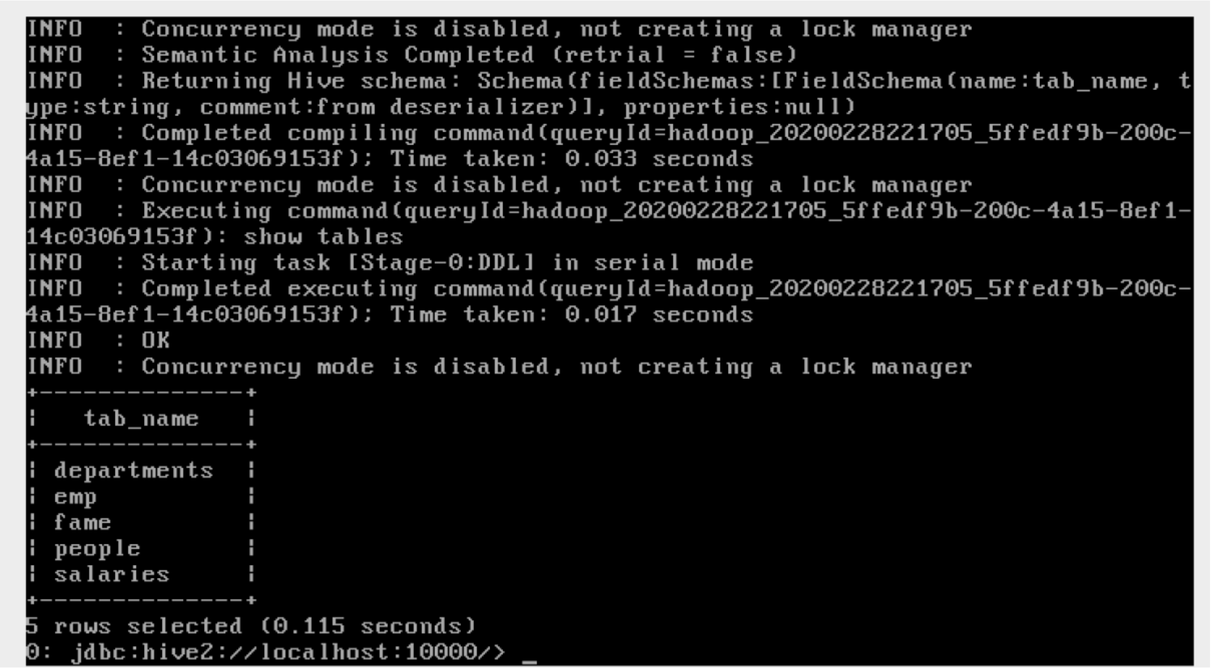


```
bitnami@debian:~$ ls
apps                  derby.log       HDFS        stack
bitnami_credentials   emp.csv         htdocs      testfile.txt
dept.csv              HallOfFame.csv  People.csv
bitnami@debian:~$
```



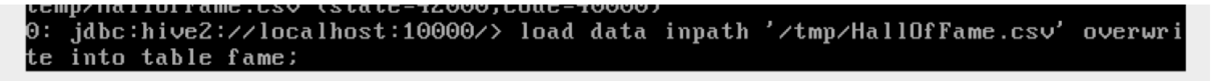
```
bitnami@debian:~$ hdfs dfs -put HallOfFame.csv /tmp
2020-02-28 22:10:00,767 INFO sasl.SaslDataTransferClient: SASL encryption trus
check: localHostTrusted = false, remoteHostTrusted = false
2020-02-28 22:10:00,943 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.ja
986)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:640)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:810)
bitnami@debian:~$
```

Create a table fame as previously.

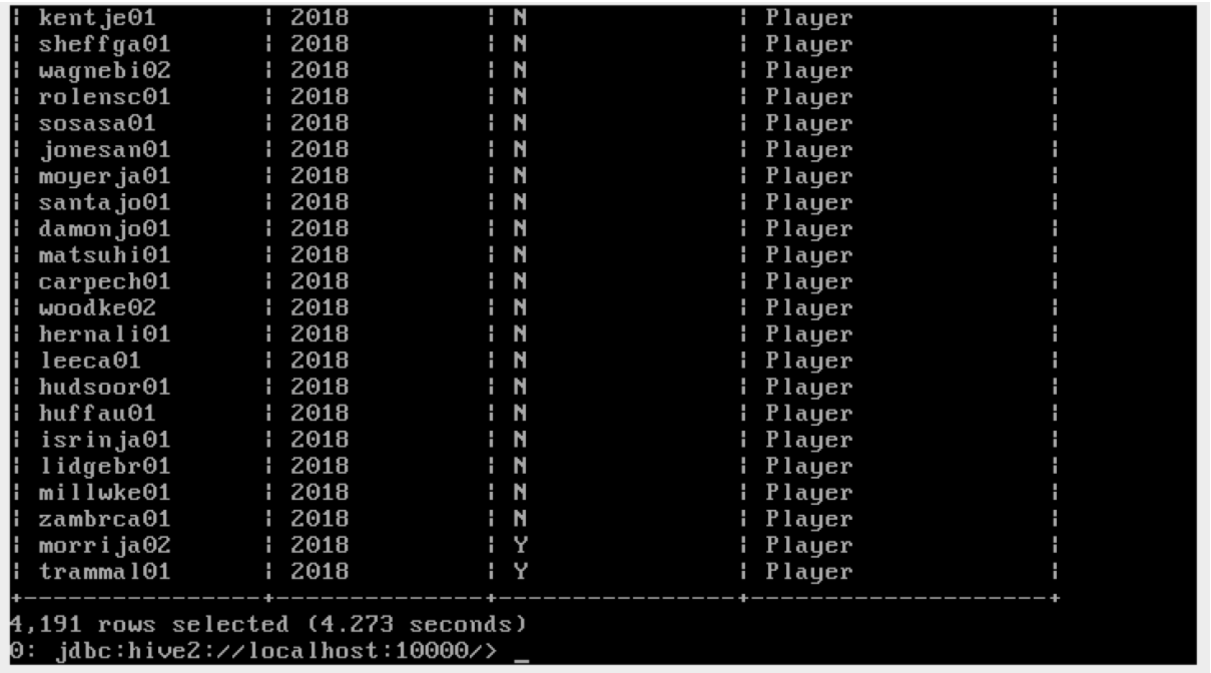


```
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, t
ype:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hadoop_20200228221705_5ffedf9b-200c-
4a15-8ef1-14c03069153f); Time taken: 0.033 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hadoop_20200228221705_5ffedf9b-200c-4a15-8ef1-
14c03069153f): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hadoop_20200228221705_5ffedf9b-200c-
4a15-8ef1-14c03069153f); Time taken: 0.017 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------+
|   tab_name   |
+--------------+
| departments  |
| emp          |
| fame         |
| people       |
| salaries     |
+--------------+
5 rows selected (0.115 seconds)
0: jdbc:hive2://localhost:10000/> _
```

Load data from HallOfFame.csv into table fame.



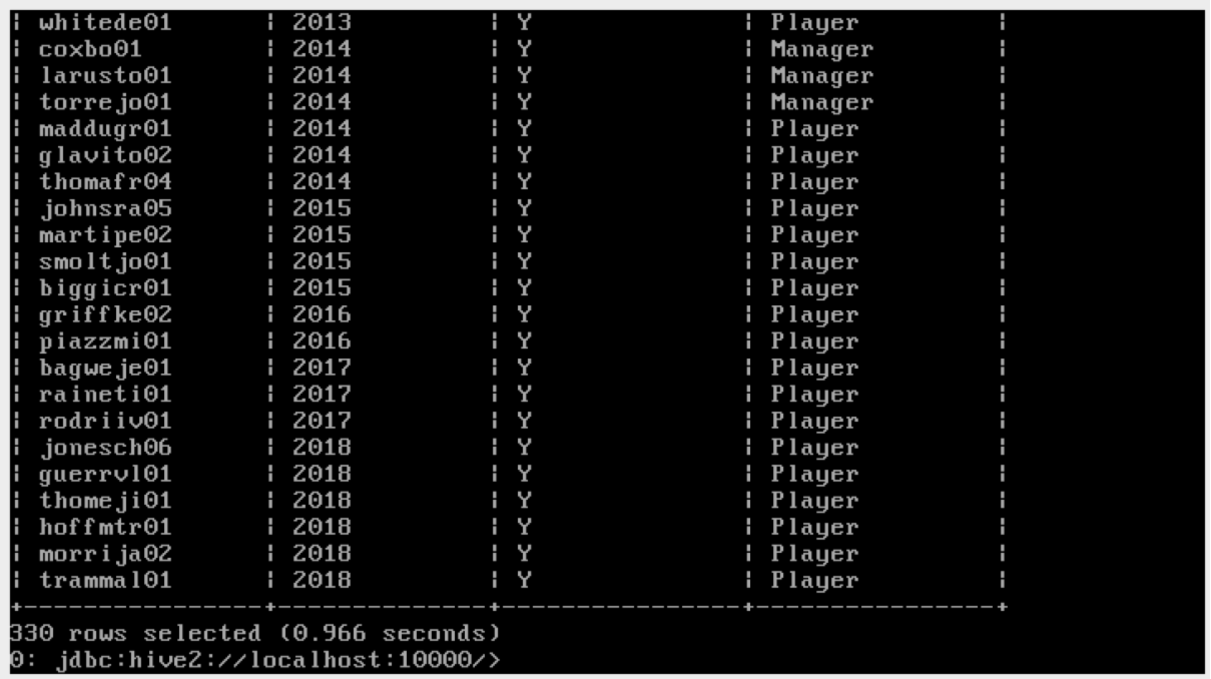
```
temp/HallOfFame.csv (state=42000,code=10000)
0: jdbc:hive2://localhost:10000/> load data inpath '/tmp/HallOfFame.csv' overwri
te into table fame;
```



```
¦ kentje01        ¦ 2018         ¦ N              ¦ Player             ¦
¦ sheffga01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ wagnebi02       ¦ 2018         ¦ N              ¦ Player             ¦
¦ rolensc01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ sosasa01        ¦ 2018         ¦ N              ¦ Player             ¦
¦ jonesan01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ moyerja01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ santajo01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ damonjo01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ matsuhi01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ carpech01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ woodke02        ¦ 2018         ¦ N              ¦ Player             ¦
¦ hernali01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ leeca01         ¦ 2018         ¦ N              ¦ Player             ¦
¦ hudsoor01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ huffau01        ¦ 2018         ¦ N              ¦ Player             ¦
¦ isrinja01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ lidgebr01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ millwke01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ zambrca01       ¦ 2018         ¦ N              ¦ Player             ¦
¦ morrija02       ¦ 2018         ¦ Y              ¦ Player             ¦
¦ trammal01       ¦ 2018         ¦ Y              ¦ Player             ¦
+-----------------+--------------+----------------+--------------------+
4,191 rows selected (4.273 seconds)
0: jdbc:hive2://localhost:10000/> _
```

There are 4,191 entries.

```
SELECT * from fame where inducted = “Y” and category = “Player” or category “Manager”;
```



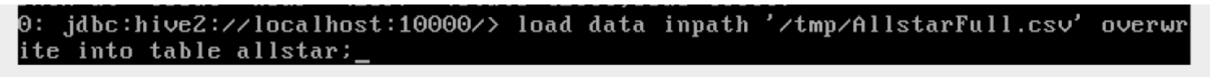
```
¦ whitede01       ¦ 2013         ¦ Y               ¦ Player          ¦
¦ coxbo01         ¦ 2014         ¦ Y               ¦ Manager         ¦
¦ larusto01       ¦ 2014         ¦ Y               ¦ Manager         ¦
¦ torrejo01       ¦ 2014         ¦ Y               ¦ Manager         ¦
¦ maddugr01       ¦ 2014         ¦ Y               ¦ Player          ¦
¦ glavito02       ¦ 2014         ¦ Y               ¦ Player          ¦
¦ thomafr04       ¦ 2014         ¦ Y               ¦ Player          ¦
¦ johnsra05       ¦ 2015         ¦ Y               ¦ Player          ¦
¦ martipe02       ¦ 2015         ¦ Y               ¦ Player          ¦
¦ smoltjo01       ¦ 2015         ¦ Y               ¦ Player          ¦
¦ biggicr01       ¦ 2015         ¦ Y               ¦ Player          ¦
¦ griffke02       ¦ 2016         ¦ Y               ¦ Player          ¦
¦ piazzmi01       ¦ 2016         ¦ Y               ¦ Player          ¦
¦ bagweje01       ¦ 2017         ¦ Y               ¦ Player          ¦
¦ raineti01       ¦ 2017         ¦ Y               ¦ Player          ¦
¦ rodriiv01       ¦ 2017         ¦ Y               ¦ Player          ¦
¦ jonesch06       ¦ 2018         ¦ Y               ¦ Player          ¦
¦ guerrvl01       ¦ 2018         ¦ Y               ¦ Player          ¦
¦ thomeji01       ¦ 2018         ¦ Y               ¦ Player          ¦
¦ hoffmtr01       ¦ 2018         ¦ Y               ¦ Player          ¦
¦ morrija02       ¦ 2018         ¦ Y               ¦ Player          ¦
¦ trammal01       ¦ 2018         ¦ Y               ¦ Player          ¦
+-----------------+--------------+-----------------+-----------------+
330 rows selected (0.966 seconds)
0: jdbc:hive2://localhost:10000/>
```

330 players and managers were inducted to the Hall of Fame.

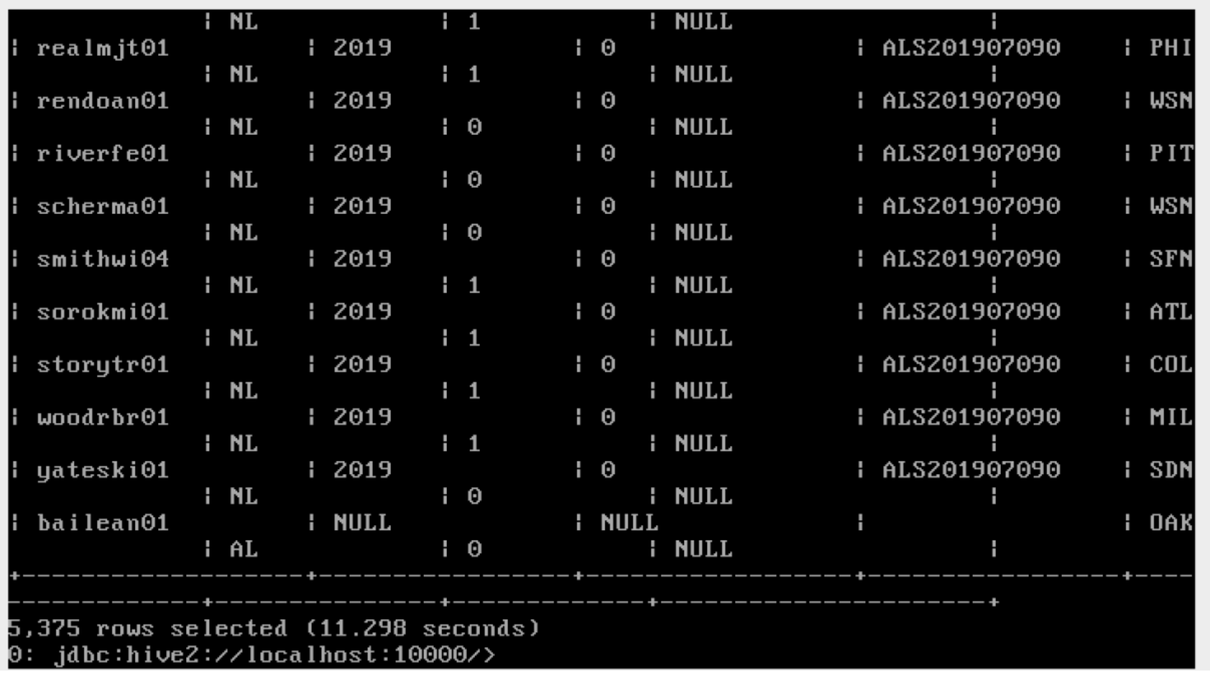
g. Provide a list of all players for any team and from any year. For example, print the list of players who played for Chicago Cubs in 2000.

To answer this question, use the AllstarFull.csv table. For uploading data into Hive proceed as with previous tables.

Successful transfer of data from AllstarFull.csv into fame Hive table.

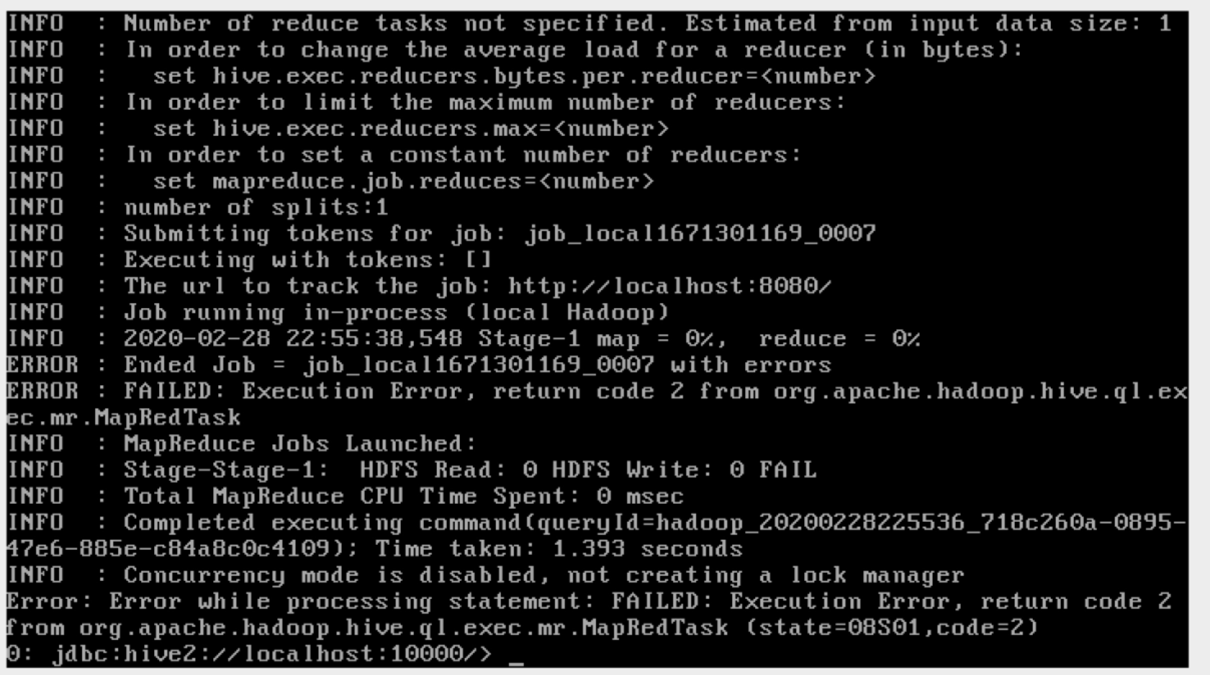


```
0: jdbc:hive2://localhost:10000/> load data inpath '/tmp/AllstarFull.csv' overwr
ite into table allstar;_
```



```
              ¦ NL            ¦ 1            ¦ NULL                 ¦
¦ realmjt01          ¦ 2019            ¦ 0                  ¦ ALS201907090     ¦ PHI
              ¦ NL            ¦ 1            ¦ NULL                 ¦
¦ rendoan01          ¦ 2019            ¦ 0                  ¦ ALS201907090     ¦ WSN
              ¦ NL            ¦ 0            ¦ NULL                 ¦
¦ riverfe01          ¦ 2019            ¦ 0                  ¦ ALS201907090     ¦ PIT
              ¦ NL            ¦ 0            ¦ NULL                 ¦
¦ scherma01          ¦ 2019            ¦ 0                  ¦ ALS201907090     ¦ WSN
              ¦ NL            ¦ 0            ¦ NULL                 ¦
¦ smithwi04          ¦ 2019            ¦ 0                  ¦ ALS201907090     ¦ SFN
              ¦ NL            ¦ 1            ¦ NULL                 ¦
¦ sorokmi01          ¦ 2019            ¦ 0                  ¦ ALS201907090     ¦ ATL
              ¦ NL            ¦ 1            ¦ NULL                 ¦
¦ storytr01          ¦ 2019            ¦ 0                  ¦ ALS201907090     ¦ COL
              ¦ NL            ¦ 1            ¦ NULL                 ¦
¦ woodrbr01          ¦ 2019            ¦ 0                  ¦ ALS201907090     ¦ MIL
              ¦ NL            ¦ 1            ¦ NULL                 ¦
¦ yateski01          ¦ 2019            ¦ 0                  ¦ ALS201907090     ¦ SDN
              ¦ NL            ¦ 0            ¦ NULL                 ¦
¦ bailean01          ¦ NULL            ¦ NULL               ¦                  ¦ OAK
              ¦ AL            ¦ 0            ¦ NULL                 ¦
+-------------------+-----------------+------------------+-----------------+----
-------------+--------------+------------+---------------------+
5,375 rows selected (11.298 seconds)
0: jdbc:hive2://localhost:10000/>
```

```
SELECT playerID, Count(*) FROM allstar GROUP BY yearID GROUP BY teamID;
```



```
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_local1671301169_0007
INFO  : Executing with tokens: []
INFO  : The url to track the job: http://localhost:8080/
INFO  : Job running in-process (local Hadoop)
INFO  : 2020-02-28 22:55:38,548 Stage-1 map = 0%,  reduce = 0%
ERROR : Ended Job = job_local1671301169_0007 with errors
ERROR : FAILED: Execution Error, return code 2 from org.apache.hadoop.hive.ql.ex
ec.mr.MapRedTask
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1:  HDFS Read: 0 HDFS Write: 0 FAIL
INFO  : Total MapReduce CPU Time Spent: 0 msec
INFO  : Completed executing command(queryId=hadoop_20200228225536_718c260a-0895-
47e6-885e-c84a8c0c4109); Time taken: 1.393 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
Error: Error while processing statement: FAILED: Execution Error, return code 2
from org.apache.hadoop.hive.ql.exec.mr.MapRedTask (state=08S01,code=2)
0: jdbc:hive2://localhost:10000/> _
```

For some reason GROUP BY commands raise an error.