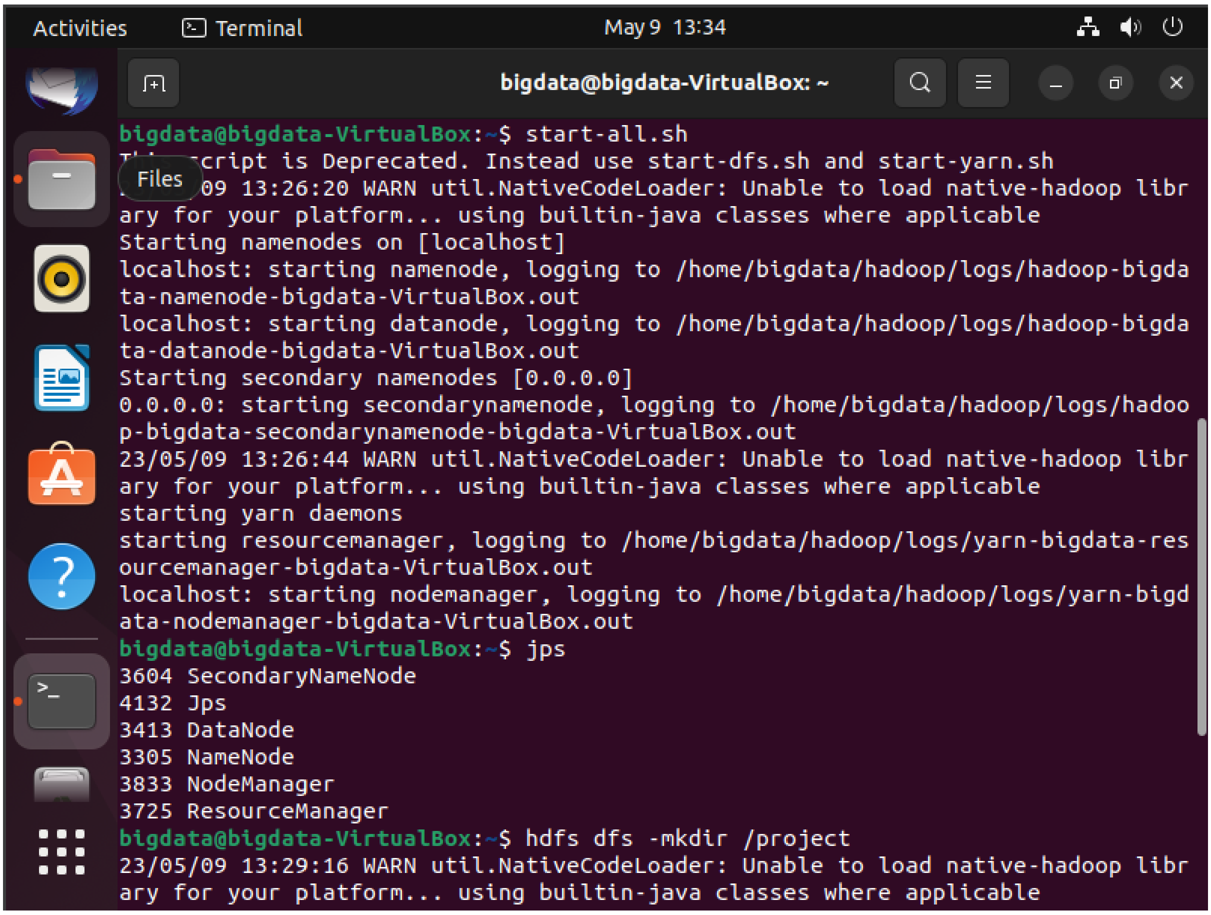
Bigdata Project

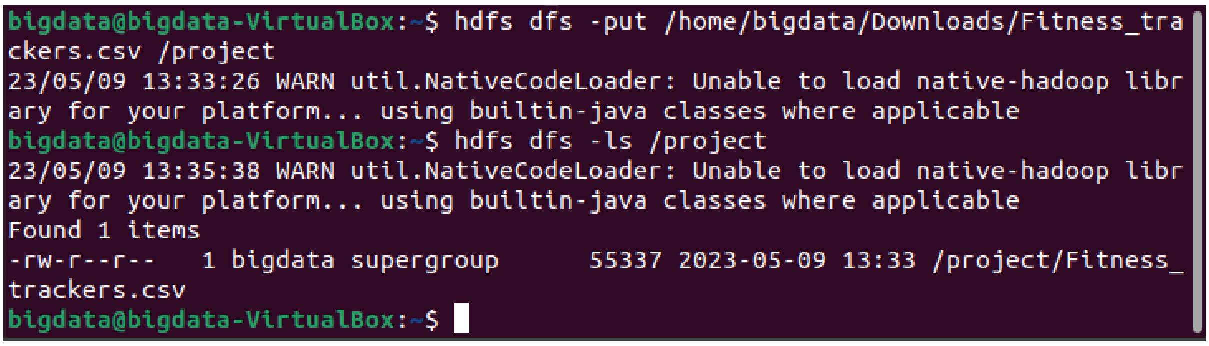
1. Add this dataset on HDFS

=> **Make a new directory on HDFS**
hdfs dfs -mkdir /project



```
bigdata@bigdata-VirtualBox:~$ start-all.sh
T[illegible]cript is Deprecated. Instead use start-dfs.sh and start-yarn.sh
[illegible]09 13:26:20 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/bigdata/hadoop/logs/hadoop-bigda
ta-namenode-bigdata-VirtualBox.out
localhost: starting datanode, logging to /home/bigdata/hadoop/logs/hadoop-bigda
ta-datanode-bigdata-VirtualBox.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/bigdata/hadoop/logs/hadoo
p-bigdata-secondarynamenode-bigdata-VirtualBox.out
23/05/09 13:26:44 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/bigdata/hadoop/logs/yarn-bigdata-res
ourcemanager-bigdata-VirtualBox.out
localhost: starting nodemanager, logging to /home/bigdata/hadoop/logs/yarn-bigd
ata-nodemanager-bigdata-VirtualBox.out
bigdata@bigdata-VirtualBox:~$ jps
3604 SecondaryNameNode
4132 Jps
3413 DataNode
3305 NameNode
3833 NodeManager
3725 ResourceManager
bigdata@bigdata-VirtualBox:~$ hdfs dfs -mkdir /project
23/05/09 13:29:16 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
```

=> **Send this data from local file system to HDFS**
Hdfs dfs -put /home/bigdata/Downloads/Fitness_trackers.csv /project

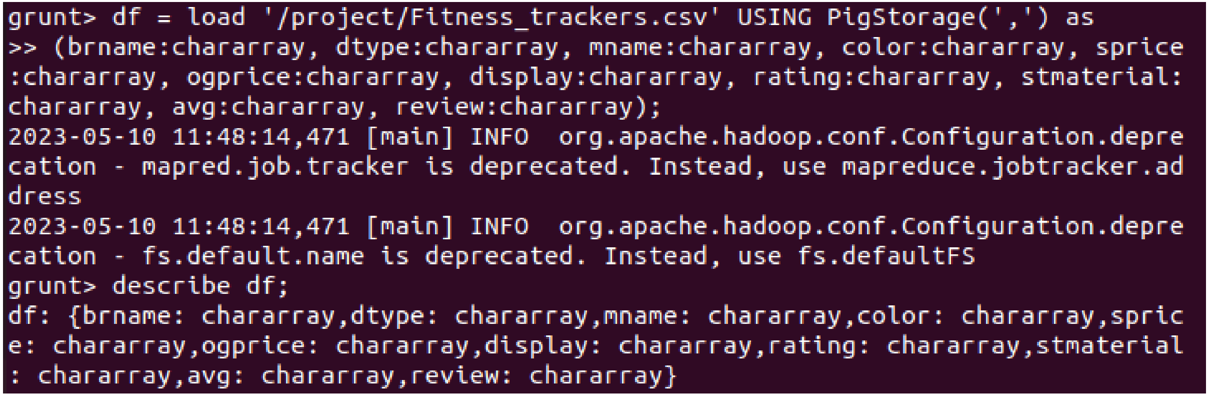


```
bigdata@bigdata-VirtualBox:~$ hdfs dfs -put /home/bigdata/Downloads/Fitness_tra
ckers.csv /project
23/05/09 13:33:26 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
bigdata@bigdata-VirtualBox:~$ hdfs dfs -ls /project
23/05/09 13:35:38 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--   1 bigdata supergroup      55337 2023-05-09 13:33 /project/Fitness_
trackers.csv
bigdata@bigdata-VirtualBox:~$ 
```

2. Load the data in Pig and filter for Device Type SmartWatch.

=> **Start Pig, Load the data into pig**

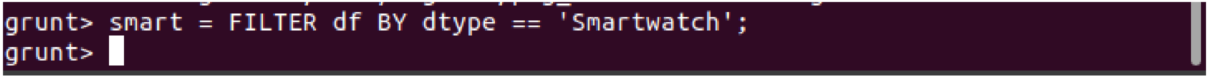
pig
df = load ‘/project/Fitness_trackers.csv’ USING PigStorage(‘,’) as (brname:chararray, dtype:chararray, mname:chararray , color:chararray, sprice:chararray, ogprice:chararray, display:chararray, rating:chararray, stmaterial:chararray, avg:chararray, review:chararray);



```
grunt> df = load '/project/Fitness_trackers.csv' USING PigStorage(',') as
>> (brname:chararray, dtype:chararray, mname:chararray, color:chararray, sprice
:chararray, ogprice:chararray, display:chararray, rating:chararray, stmaterial:
chararray, avg:chararray, review:chararray);
2023-05-10 11:48:14,471 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
2023-05-10 11:48:14,471 [main] INFO  org.apache.hadoop.conf.Configuration.depre
cation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> describe df;
df: {brname: chararray,dtype: chararray,mname: chararray,color: chararray,spric
e: chararray,ogprice: chararray,display: chararray,rating: chararray,stmaterial
: chararray,avg: chararray,review: chararray}
```

=> **Filter command to get the device type smartwatch**

Smart = FILTER df BY dtype == ‘Smartwatch’;

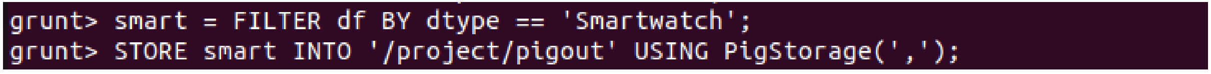


```
grunt> smart = FILTER df BY dtype == 'Smartwatch';
grunt> 
```

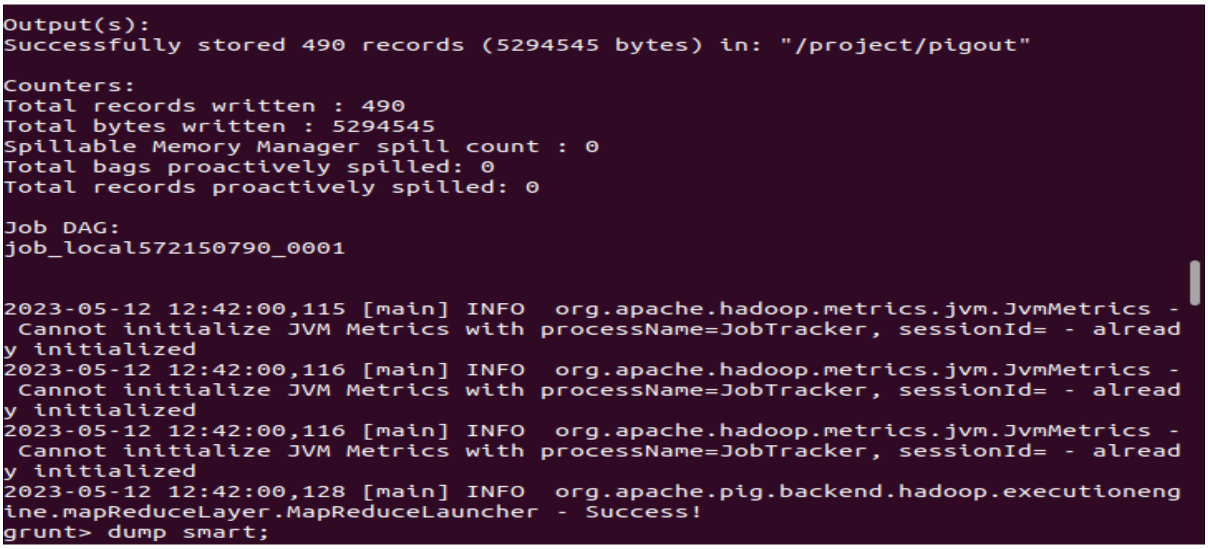
3. Store the result in HDFS

=> **Store the result that you get after filtering**

STORE smart INTO ‘/project/pigout’ USING PigStorage(‘,’);



```
grunt> smart = FILTER df BY dtype == 'Smartwatch';
grunt> STORE smart INTO '/project/pigout' USING PigStorage(',');
```

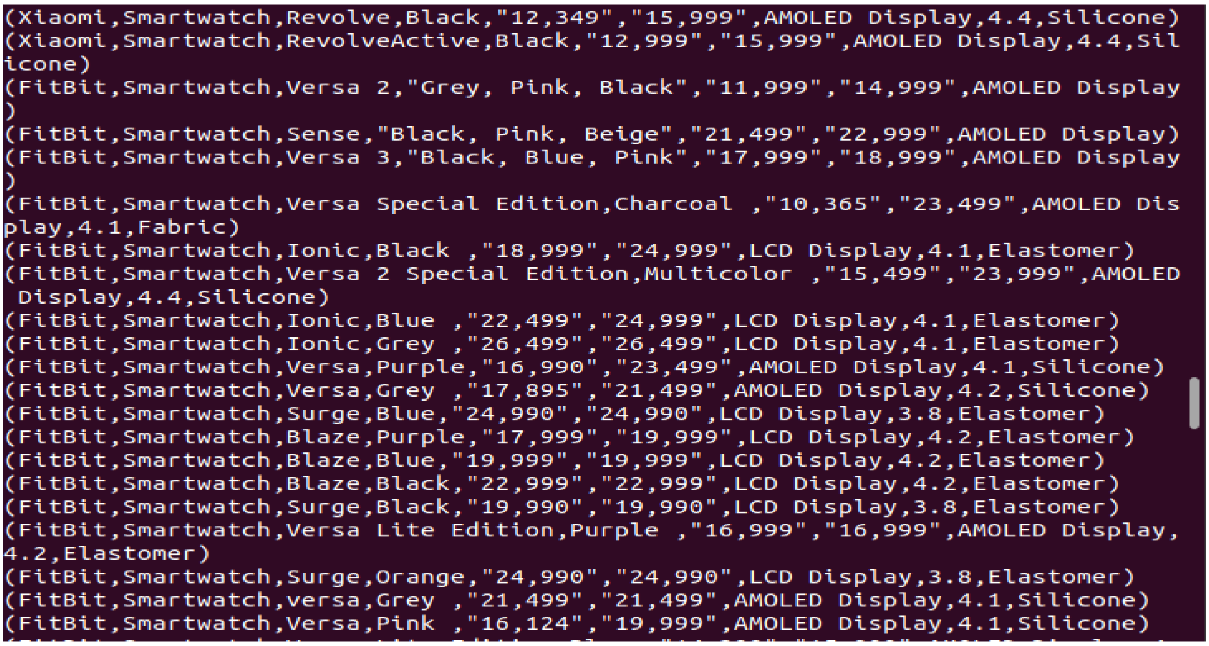


```
Output(s):
Successfully stored 490 records (5294545 bytes) in: "/project/pigout"

Counters:
Total records written : 490
Total bytes written : 5294545
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local572150790_0001


2023-05-12 12:42:00,115 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
 Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alread
y initialized
2023-05-12 12:42:00,116 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
 Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alread
y initialized
2023-05-12 12:42:00,116 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
 Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alread
y initialized
2023-05-12 12:42:00,128 [main] INFO  org.apache.pig.backend.hadoop.executioneng
ine.mapReduceLayer.MapReduceLauncher - Success!
grunt> dump smart;
```

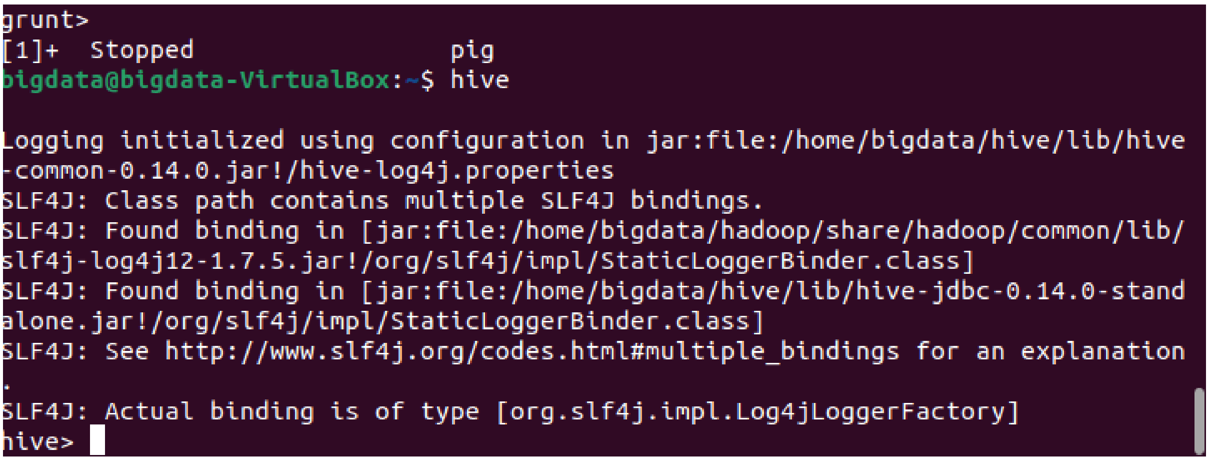


```
(Xiaomi,Smartwatch,Revolve,Black,"12,349","15,999",AMOLED Display,4.4,Silicone)
(Xiaomi,Smartwatch,RevolveActive,Black,"12,999","15,999",AMOLED Display,4.4,Sil
icone)
(FitBit,Smartwatch,Versa 2,"Grey, Pink, Black","11,999","14,999",AMOLED Display
)
(FitBit,Smartwatch,Sense,"Black, Pink, Beige","21,499","22,999",AMOLED Display)
(FitBit,Smartwatch,Versa 3,"Black, Blue, Pink","17,999","18,999",AMOLED Display
)
(FitBit,Smartwatch,Versa Special Edition,Charcoal ,"10,365","23,499",AMOLED Dis
play,4.1,Fabric)
(FitBit,Smartwatch,Ionic,Black ,"18,999","24,999",LCD Display,4.1,Elastomer)
(FitBit,Smartwatch,Versa 2 Special Edition,Multicolor ,"15,499","23,999",AMOLED
 Display,4.4,Silicone)
(FitBit,Smartwatch,Ionic,Blue ,"22,499","24,999",LCD Display,4.1,Elastomer)
(FitBit,Smartwatch,Ionic,Grey ,"26,499","26,499",LCD Display,4.1,Elastomer)
(FitBit,Smartwatch,Versa,Purple,"16,990","23,499",AMOLED Display,4.1,Silicone)
(FitBit,Smartwatch,Versa,Grey ,"17,895","21,499",AMOLED Display,4.2,Silicone)
(FitBit,Smartwatch,Surge,Blue,"24,990","24,990",LCD Display,3.8,Elastomer)
(FitBit,Smartwatch,Blaze,Purple,"17,999","19,999",LCD Display,4.2,Elastomer)
(FitBit,Smartwatch,Blaze,Blue,"19,999","19,999",LCD Display,4.2,Elastomer)
(FitBit,Smartwatch,Blaze,Black,"22,999","22,999",LCD Display,4.2,Elastomer)
(FitBit,Smartwatch,Surge,Black,"19,990","19,990",LCD Display,3.8,Elastomer)
(FitBit,Smartwatch,Versa Lite Edition,Purple ,"16,999","16,999",AMOLED Display,
4.2,Elastomer)
(FitBit,Smartwatch,Surge,Orange,"24,990","24,990",LCD Display,3.8,Elastomer)
(FitBit,Smartwatch,versa,Grey ,"21,499","21,499",AMOLED Display,4.1,Silicone)
(FitBit,Smartwatch,Versa,Pink ,"16,124","19,999",AMOLED Display,4.1,Silicone)
```

4. Load the output of Pig in Hive. Display the data sorted by Rating.

=> **Start the Hive shell**

hive



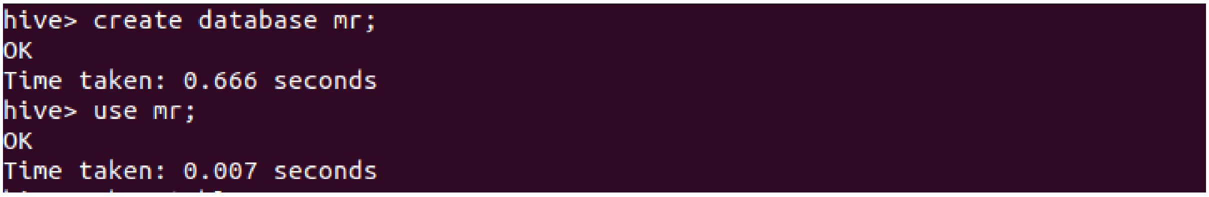
```
grunt>
[1]+  Stopped                 pig
bigdata@bigdata-VirtualBox:~$ hive

Logging initialized using configuration in jar:file:/home/bigdata/hive/lib/hive
-common-0.14.0.jar!/hive-log4j.properties
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/bigdata/hadoop/share/hadoop/common/lib/
slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/bigdata/hive/lib/hive-jdbc-0.14.0-stand
alone.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.slf4j.impl.Log4jLoggerFactory]
hive> 
```

=> **Create a database, Select the database**

create database mr;

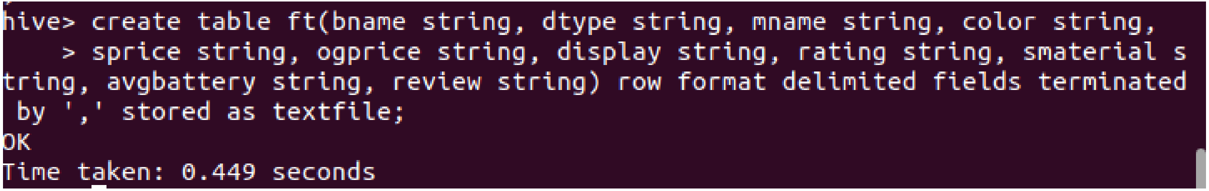
use mr;



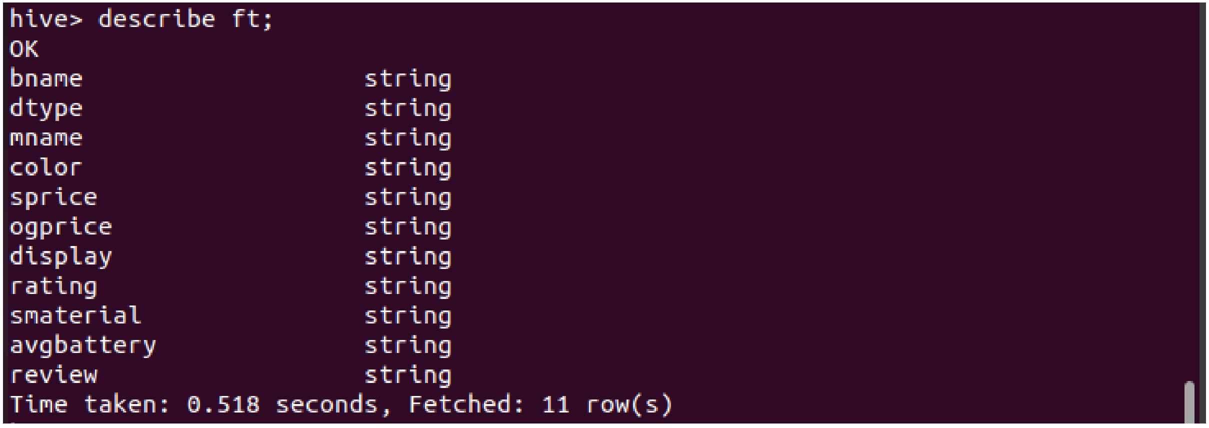
```
hive> create database mr;
OK
Time taken: 0.666 seconds
hive> use mr;
OK
Time taken: 0.007 seconds
```

=> **Create a table with the correct schema**

create table ft(bname string, dtype string, mname string, color string, sprice string, ogprice string, display string, rating string, smaterial string, avgbattery string, review string) row format delimited fields terminated by ‘,’ stored as textfile;



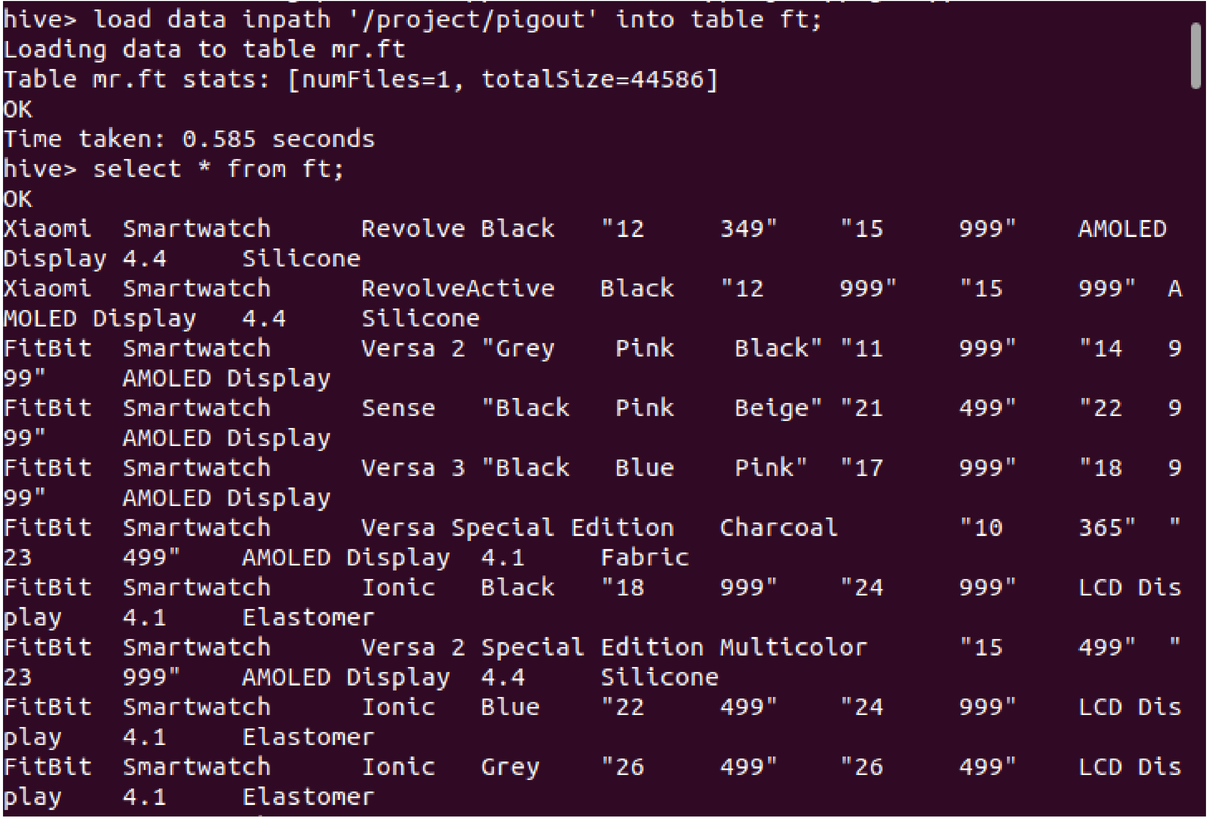
```
hive> create table ft(bname string, dtype string, mname string, color string,
    > sprice string, ogprice string, display string, rating string, smaterial s
tring, avgbattery string, review string) row format delimited fields terminated
 by ',' stored as textfile;
OK
Time taken: 0.449 seconds
```



```
hive> describe ft;
OK
bname                   string
dtype                   string
mname                   string
color                   string
sprice                  string
ogprice                 string
display                 string
rating                  string
smaterial               string
avgbattery              string
review                  string
Time taken: 0.518 seconds, Fetched: 11 row(s)
```

=> **Load the output of Pig that you got after filtering into that table**

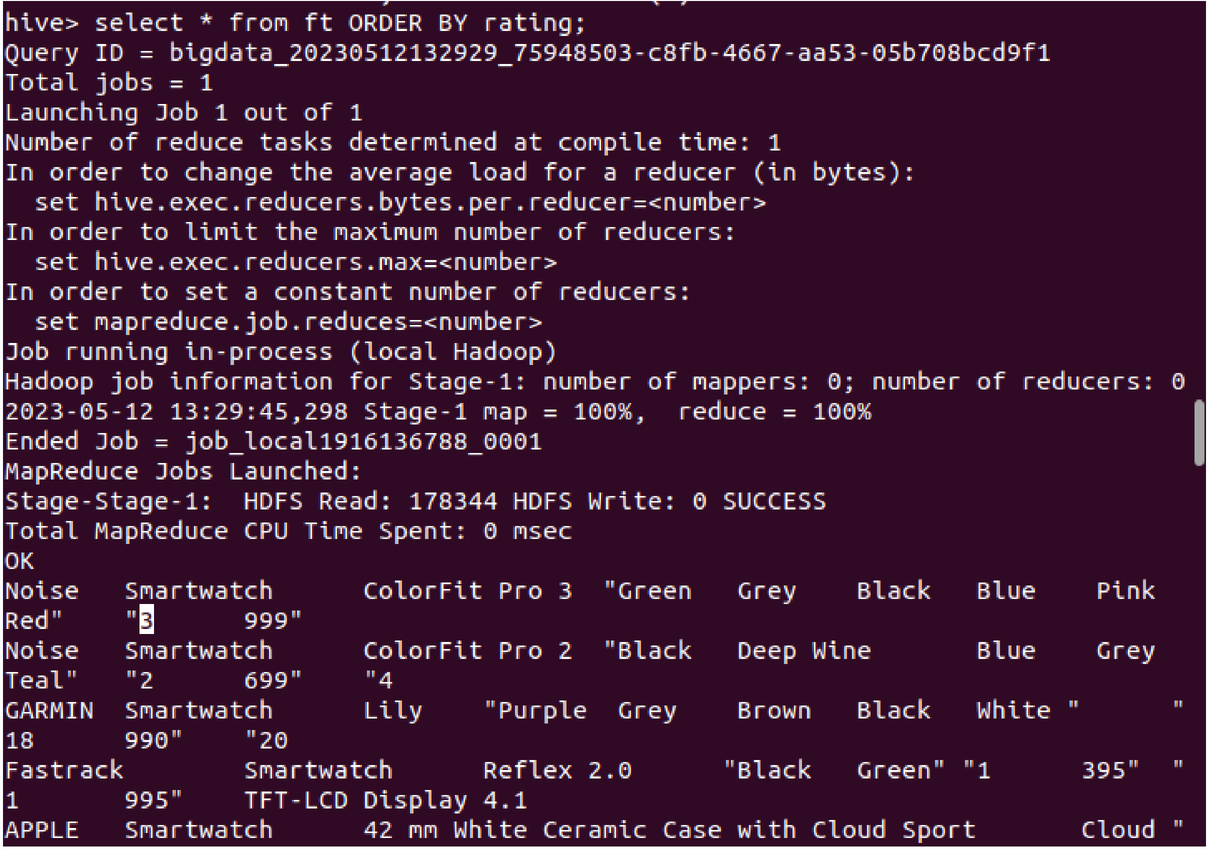
load data inpath ‘/project/pigout’ into table ft;



```
hive> load data inpath '/project/pigout' into table ft;
Loading data to table mr.ft
Table mr.ft stats: [numFiles=1, totalSize=44586]
OK
Time taken: 0.585 seconds
hive> select * from ft;
OK
Xiaomi  Smartwatch      Revolve Black   "12     349"    "15     999"    AMOLED
Display 4.4     Silicone
Xiaomi  Smartwatch      RevolveActive   Black   "12     999"    "15     999"  A
MOLED Display   4.4     Silicone
FitBit  Smartwatch      Versa 2 "Grey    Pink    Black" "11     999"    "14   9
99"     AMOLED Display
FitBit  Smartwatch      Sense   "Black   Pink    Beige" "21     499"    "22   9
99"     AMOLED Display
FitBit  Smartwatch      Versa 3 "Black   Blue    Pink"  "17     999"    "18   9
99"     AMOLED Display
FitBit  Smartwatch      Versa Special Edition   Charcoal        "10     365"  "
23      499"    AMOLED Display  4.1     Fabric
FitBit  Smartwatch      Ionic   Black   "18     999"    "24     999"    LCD Dis
play    4.1     Elastomer
FitBit  Smartwatch      Versa 2 Special Edition Multicolor      "15     499"  "
23      999"    AMOLED Display  4.4     Silicone
FitBit  Smartwatch      Ionic   Blue    "22     499"    "24     999"    LCD Dis
play    4.1     Elastomer
FitBit  Smartwatch      Ionic   Grey    "26     499"    "26     499"    LCD Dis
play    4.1     Elastomer
```

=> **Write a select query to sort the data by Rating (order by clause)**

select * from ft ORDER BY rating;

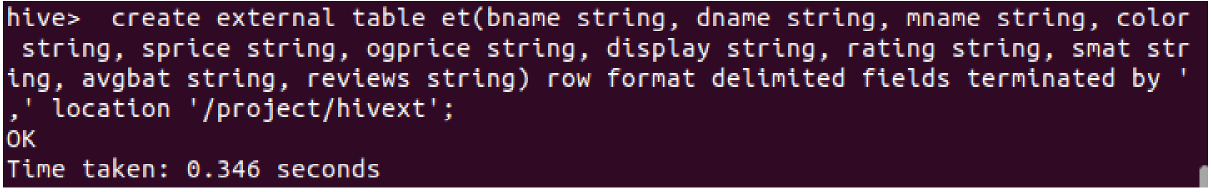


```
hive> select * from ft ORDER BY rating;
Query ID = bigdata_20230512132929_75948503-c8fb-4667-aa53-05b708bcd9f1
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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>
Job running in-process (local Hadoop)
Hadoop job information for Stage-1: number of mappers: 0; number of reducers: 0
2023-05-12 13:29:45,298 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1916136788_0001
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 178344 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Noise   Smartwatch      ColorFit Pro 3  "Green   Grey    Black   Blue    Pink
Red"    "3      999"
Noise   Smartwatch      ColorFit Pro 2  "Black   Deep Wine       Blue    Grey
Teal"   "2      699"    "4
GARMIN  Smartwatch      Lily    "Purple  Grey    Brown   Black   White "      "
18      990"    "20
Fastrack        Smartwatch      Reflex 2.0      "Black   Green" "1      395"  "
1       995"    TFT-LCD Display 4.1
APPLE   Smartwatch      42 mm White Ceramic Case with Cloud Sport       Cloud "
```

5. Store that results in HDFS.

=> **Create an external table**

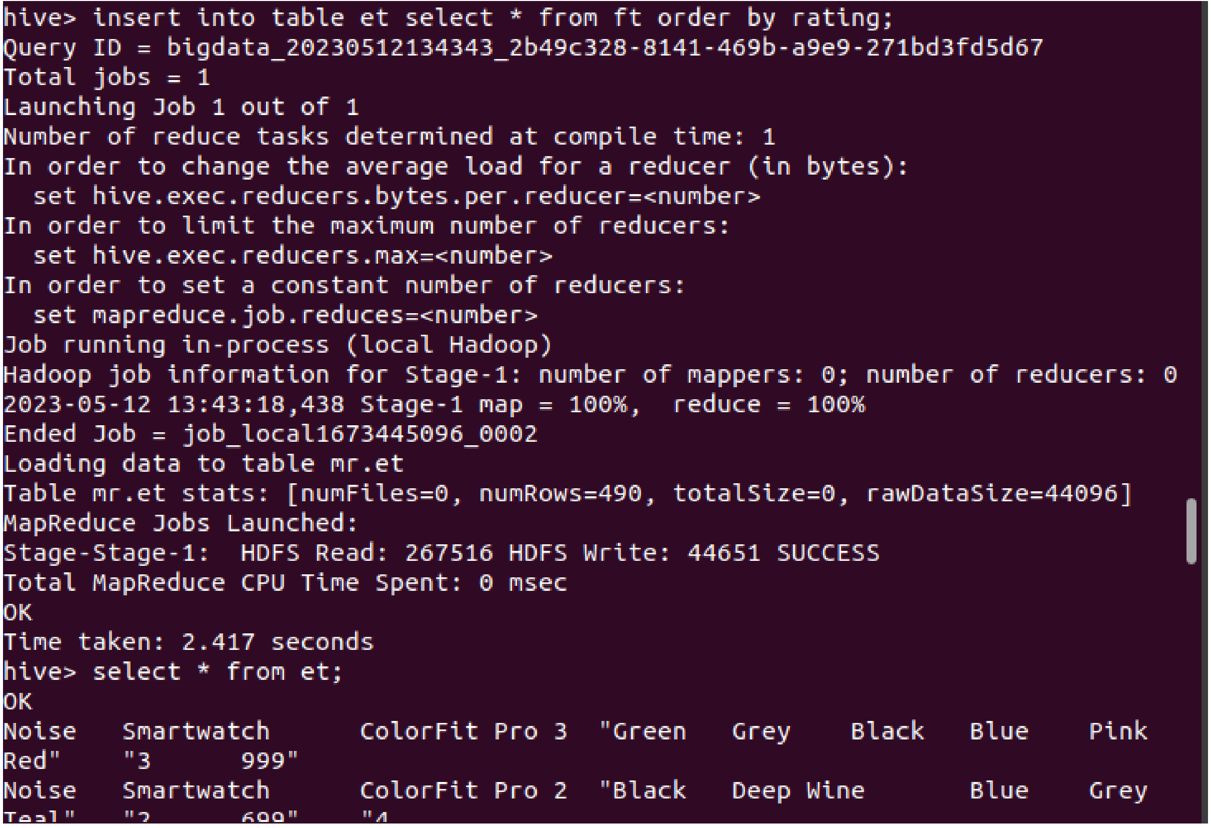
Create external table et(bname string, dtype string, mname string, color string, sprice string, ogprice string, display string, rating string, smat string, avgbat string, reviews string) row format delimited fields terminated by ‘,’ location ‘/project/hivext’;



```
hive>  create external table et(bname string, dname string, mname string, color
 string, sprice string, ogprice string, display string, rating string, smat str
ing, avgbat string, reviews string) row format delimited fields terminated by '
,' location '/project/hivext';
OK
Time taken: 0.346 seconds
```

=> **Insert the result of the select query to sort the data by rating into that external table**

Insert into table et select * from ft order by rating;



```
hive> insert into table et select * from ft order by rating;
Query ID = bigdata_20230512134343_2b49c328-8141-469b-a9e9-271bd3fd5d67
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>
Job running in-process (local Hadoop)
Hadoop job information for Stage-1: number of mappers: 0; number of reducers: 0
2023-05-12 13:43:18,438 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1673445096_0002
Loading data to table mr.et
Table mr.et stats: [numFiles=0, numRows=490, totalSize=0, rawDataSize=44096]
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 267516 HDFS Write: 44651 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 2.417 seconds
hive> select * from et;
OK
Noise   Smartwatch      ColorFit Pro 3  "Green   Grey    Black   Blue    Pink
Red"    "3      999"
Noise   Smartwatch      ColorFit Pro 2  "Black   Deep Wine       Blue    Grey
Teal"   "2      699"    "4
```