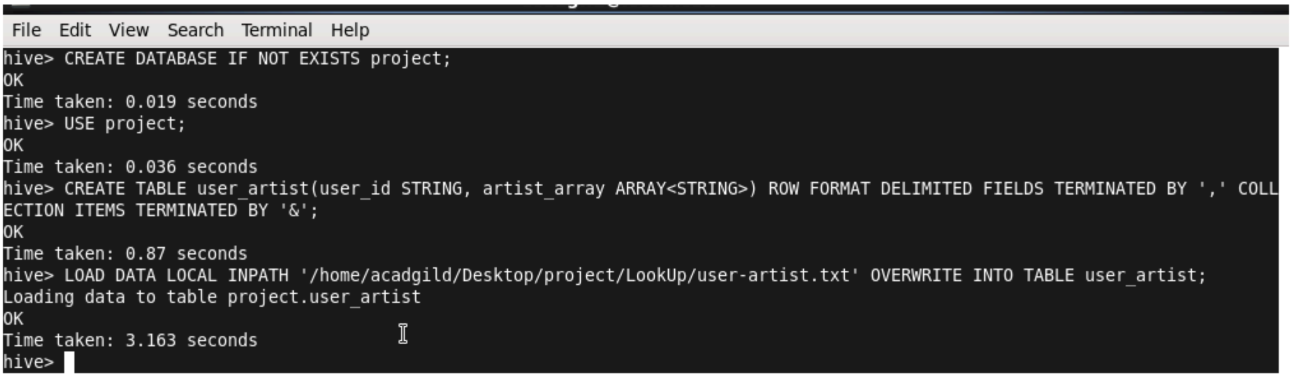
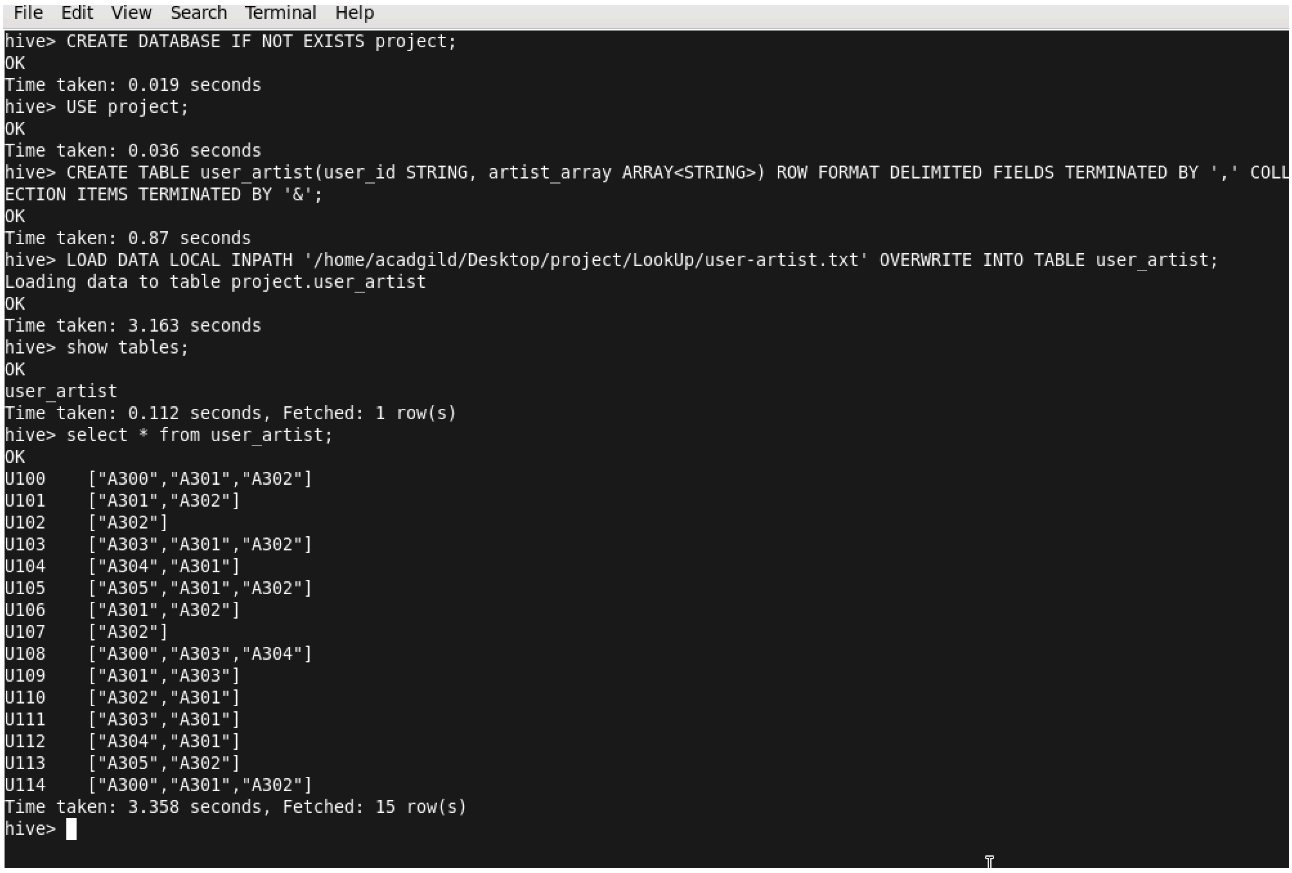
Music Data Analysis using Hadoop

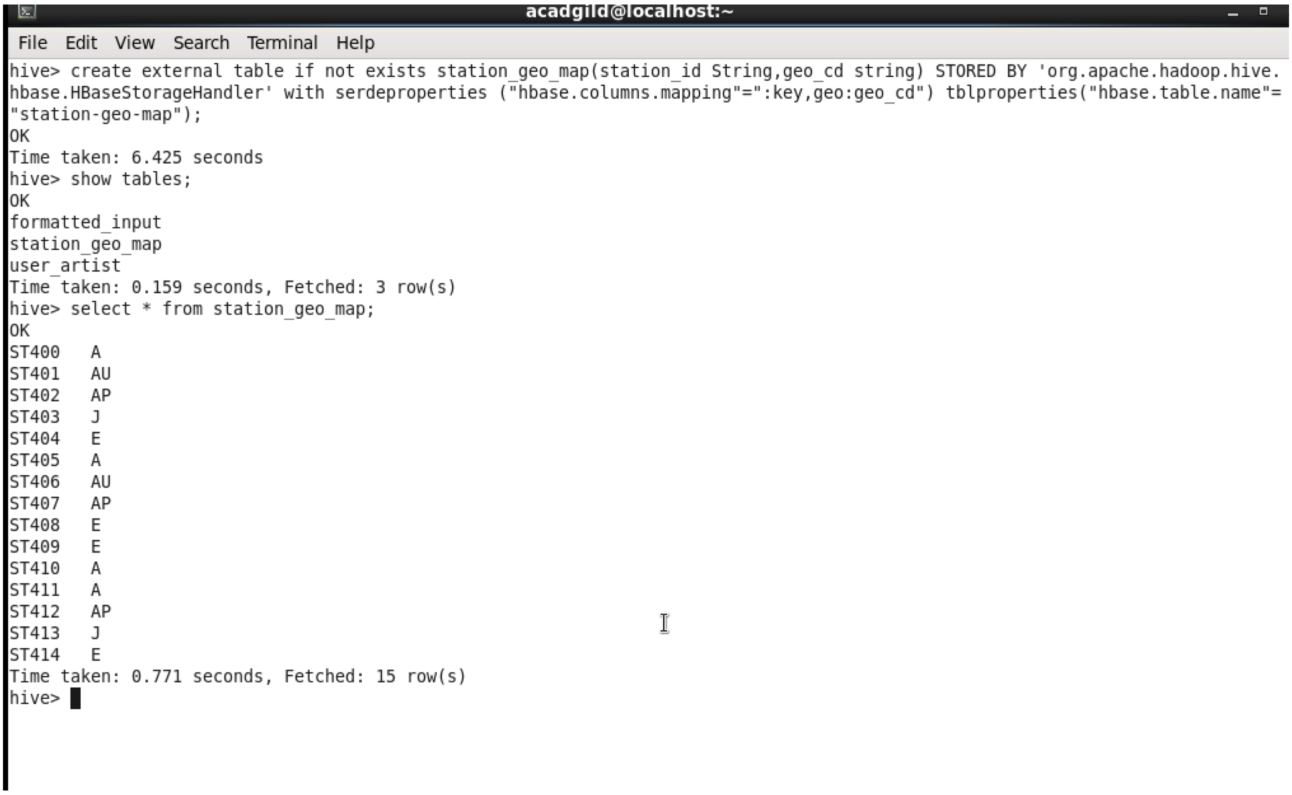
Creating Lookup table using hive



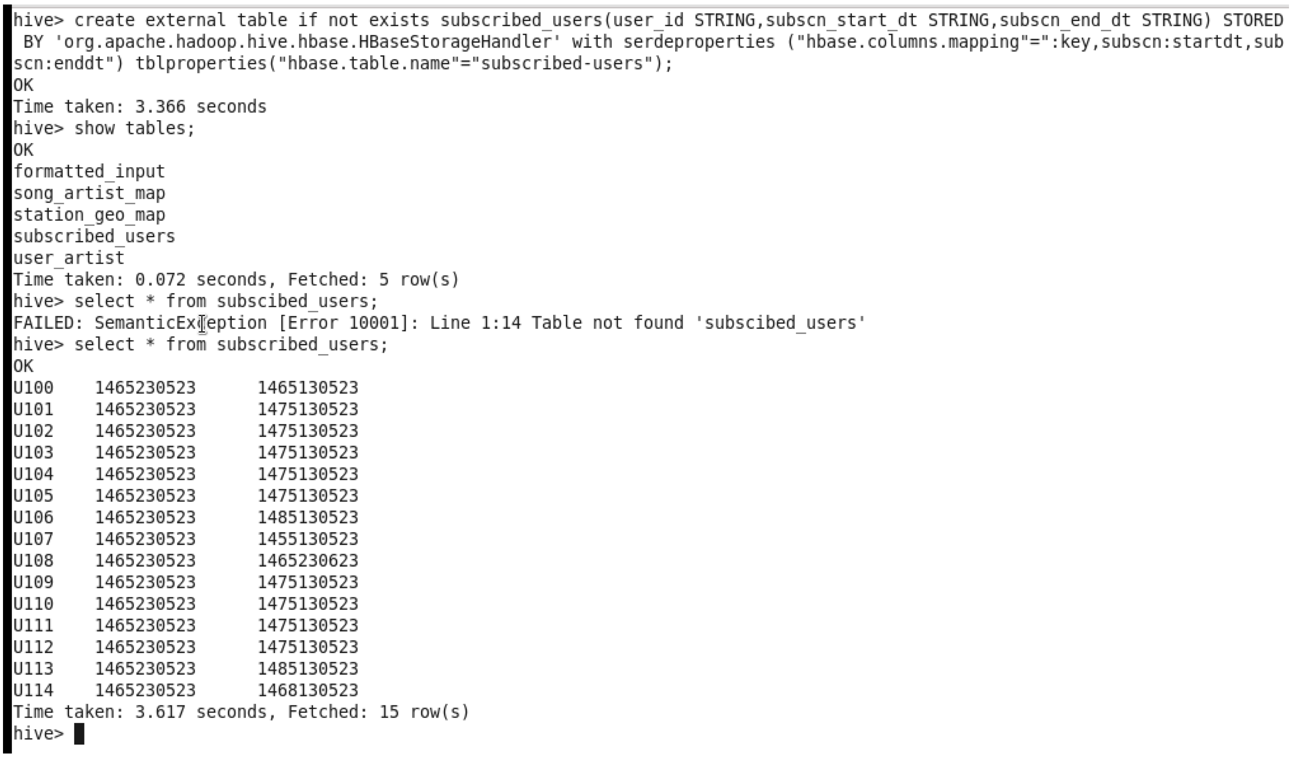
```
hive> CREATE DATABASE IF NOT EXISTS project;
OK
Time taken: 0.019 seconds
hive> USE project;
OK
Time taken: 0.036 seconds
hive> CREATE TABLE user_artist(user_id STRING, artist_array ARRAY<STRING>) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' COLL
ECTION ITEMS TERMINATED BY '&';
OK
Time taken: 0.87 seconds
hive> LOAD DATA LOCAL INPATH '/home/acadgild/Desktop/project/LookUp/user-artist.txt' OVERWRITE INTO TABLE user_artist;
Loading data to table project.user_artist
OK
Time taken: 3.163 seconds
hive> 
```



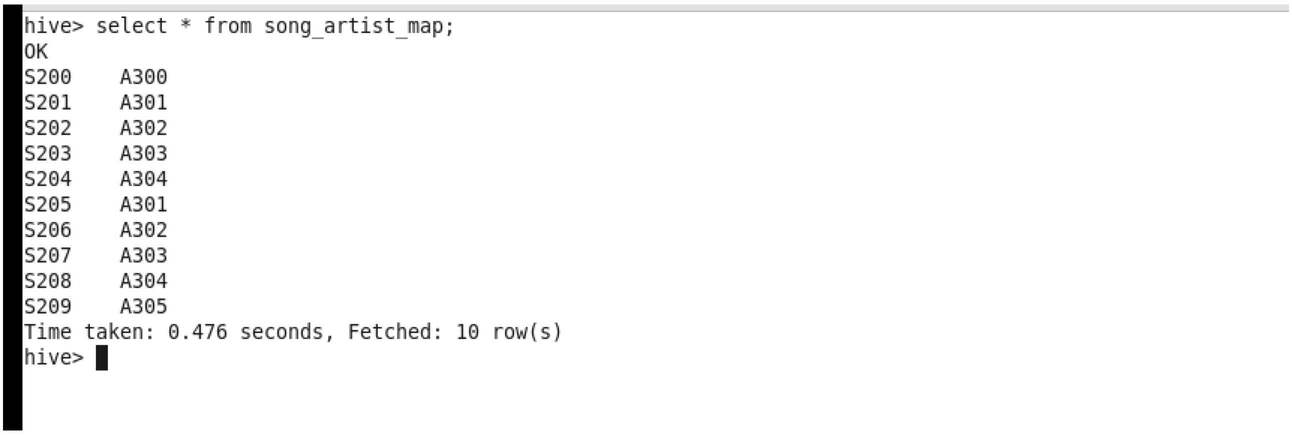
```
hive> CREATE DATABASE IF NOT EXISTS project;
OK
Time taken: 0.019 seconds
hive> USE project;
OK
Time taken: 0.036 seconds
hive> CREATE TABLE user_artist(user_id STRING, artist_array ARRAY<STRING>) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' COLL
ECTION ITEMS TERMINATED BY '&';
OK
Time taken: 0.87 seconds
hive> LOAD DATA LOCAL INPATH '/home/acadgild/Desktop/project/LookUp/user-artist.txt' OVERWRITE INTO TABLE user_artist;
Loading data to table project.user_artist
OK
Time taken: 3.163 seconds
hive> show tables;
OK
user_artist
Time taken: 0.112 seconds, Fetched: 1 row(s)
hive> select * from user_artist;
OK
U100    ["A300","A301","A302"]
U101    ["A301","A302"]
U102    ["A302"]
U103    ["A303","A301","A302"]
U104    ["A304","A301"]
U105    ["A305","A301","A302"]
U106    ["A301","A302"]
U107    ["A302"]
U108    ["A300","A303","A304"]
U109    ["A301","A303"]
U110    ["A302","A301"]
U111    ["A303","A301"]
U112    ["A304","A301"]
U113    ["A305","A302"]
U114    ["A300","A301","A302"]
Time taken: 3.358 seconds, Fetched: 15 row(s)
hive> 
```



```
hive> create external table if not exists station_geo_map(station_id String,geo_cd string) STORED BY 'org.apache.hadoop.hive.
hbase.HBaseStorageHandler' with serdeproperties ("hbase.columns.mapping"=":key,geo:geo_cd") tblproperties("hbase.table.name"=
"station-geo-map");
OK
Time taken: 6.425 seconds
hive> show tables;
OK
formatted_input
station_geo_map
user_artist
Time taken: 0.159 seconds, Fetched: 3 row(s)
hive> select * from station_geo_map;
OK
ST400   A
ST401   AU
ST402   AP
ST403   J
ST404   E
ST405   A
ST406   AU
ST407   AP
ST408   E
ST409   E
ST410   A
ST411   A
ST412   AP
ST413   J
ST414   E
Time taken: 0.771 seconds, Fetched: 15 row(s)
hive> 
```

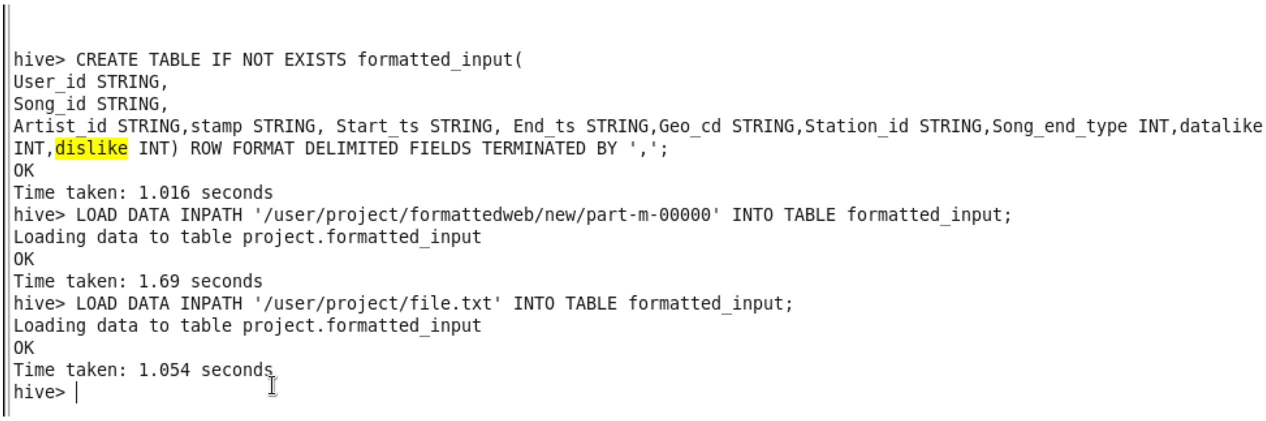


```
hive> create external table if not exists subscribed_users(user_id STRING,subscn_start_dt STRING,subscn_end_dt STRING) STORED
 BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' with serdeproperties ("hbase.columns.mapping"=":key,subscn:startdt,sub
scn:enddt") tblproperties("hbase.table.name"="subscribed-users");
OK
Time taken: 3.366 seconds
hive> show tables;
OK
formatted_input
song_artist_map
station_geo_map
subscribed_users
user_artist
Time taken: 0.072 seconds, Fetched: 5 row(s)
hive> select * from subscibed_users;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'subscibed_users'
hive> select * from subscribed_users;
OK
U100    1465230523      1465130523
U101    1465230523      1475130523
U102    1465230523      1475130523
U103    1465230523      1475130523
U104    1465230523      1475130523
U105    1465230523      1475130523
U106    1465230523      1485130523
U107    1465230523      1455130523
U108    1465230523      1465230623
U109    1465230523      1475130523
U110    1465230523      1475130523
U111    1465230523      1475130523
U112    1465230523      1475130523
U113    1465230523      1485130523
U114    1465230523      1468130523
Time taken: 3.617 seconds, Fetched: 15 row(s)
hive> 
```

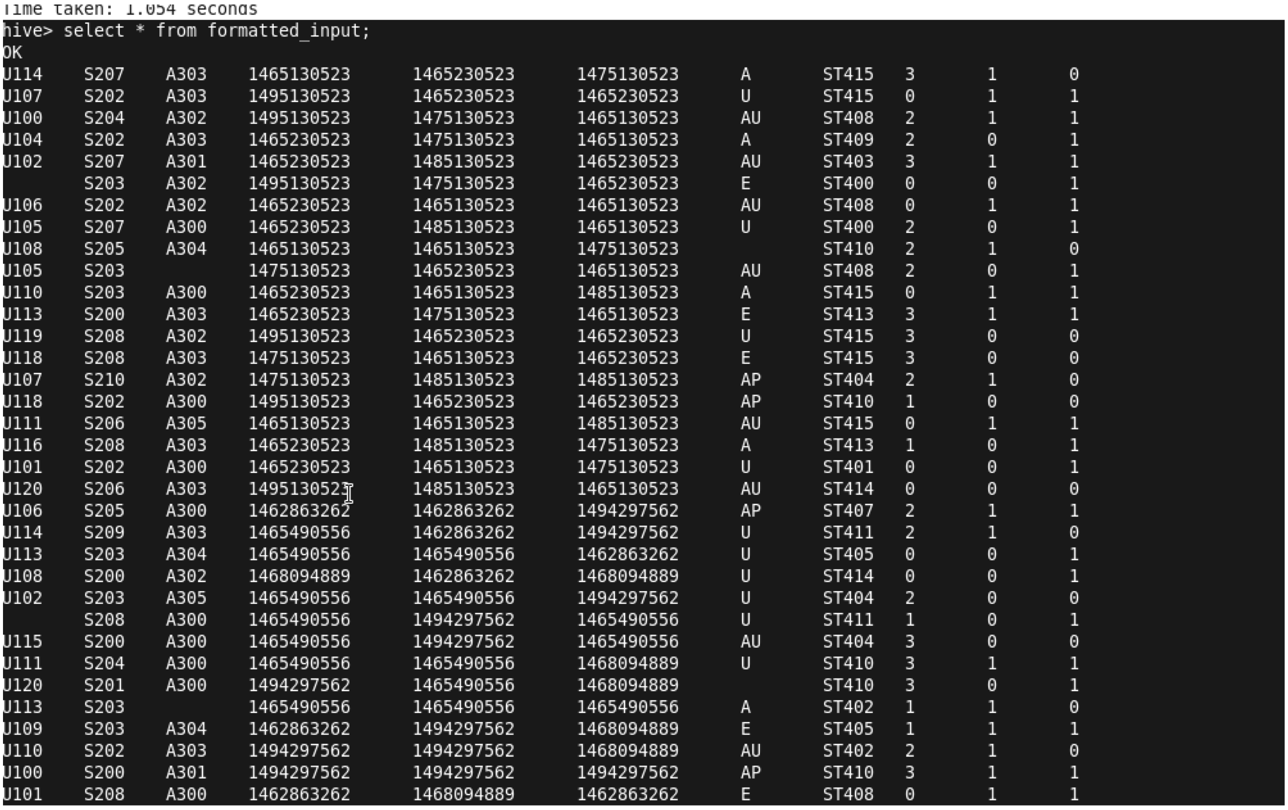


```
hive> select * from song_artist_map;
OK
S200    A300
S201    A301
S202    A302
S203    A303
S204    A304
S205    A301
S206    A302
S207    A303
S208    A304
S209    A305
Time taken: 0.476 seconds, Fetched: 10 row(s)
hive> 
```

Data Ingestion and Initial Validation

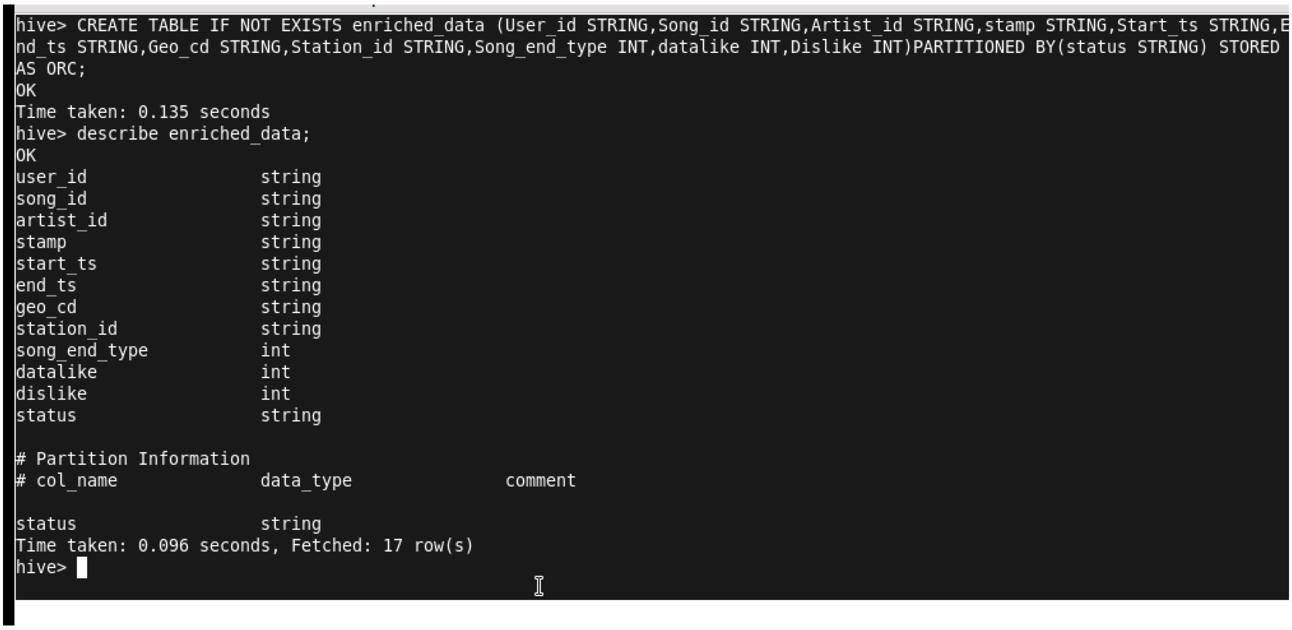


```
hive> CREATE TABLE IF NOT EXISTS formatted_input(
User_id STRING,
Song_id STRING,
Artist_id STRING,stamp STRING, Start_ts STRING, End_ts STRING,Geo_cd STRING,Station_id STRING,Song_end_type INT,datalike
INT,dislike INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 1.016 seconds
hive> LOAD DATA INPATH '/user/project/formattedweb/new/part-m-00000' INTO TABLE formatted_input;
Loading data to table project.formatted_input
OK
Time taken: 1.69 seconds
hive> LOAD DATA INPATH '/user/project/file.txt' INTO TABLE formatted_input;
Loading data to table project.formatted_input
OK
Time taken: 1.054 seconds
hive> 
```

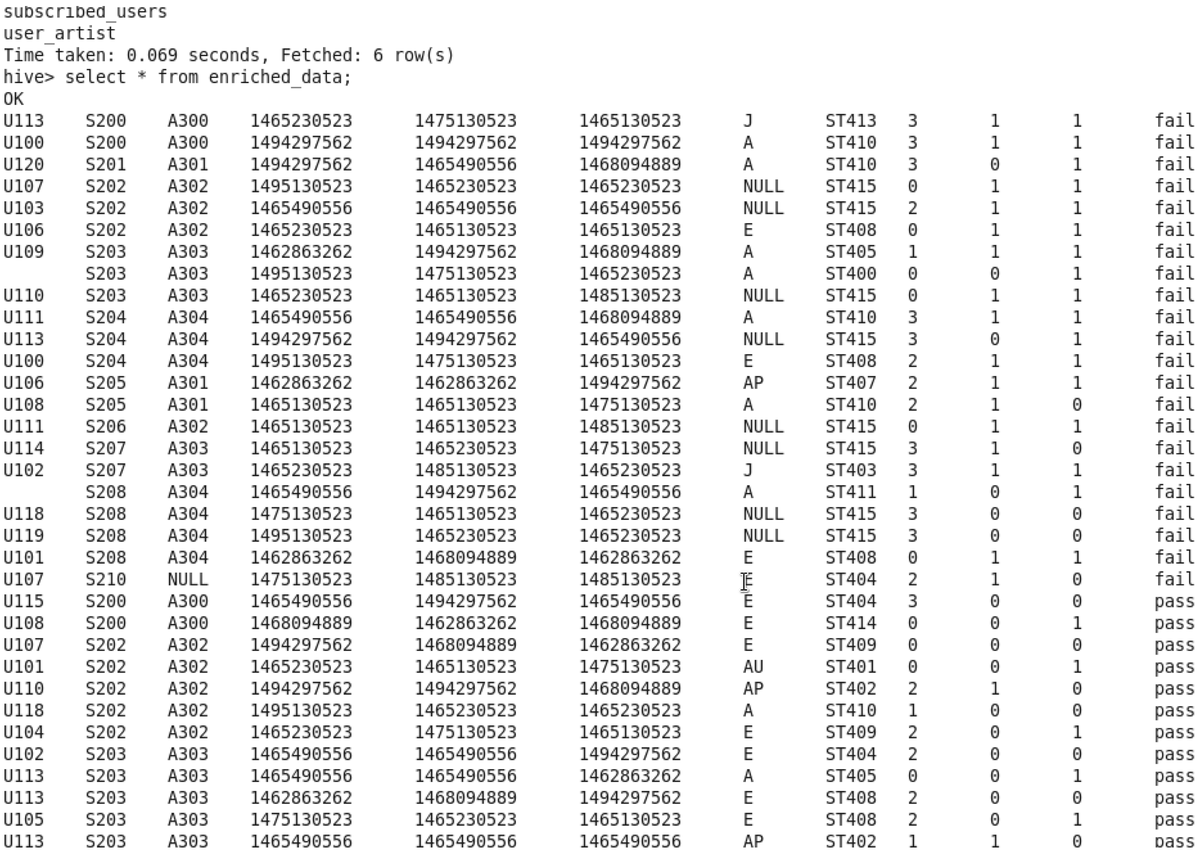


```
Time taken: 1.054 seconds
hive> select * from formatted_input;
OK
U114    S207    A303    1465130523      1465230523      1475130523      A       ST415   3       1       0
U107    S202    A303    1495130523      1465230523      1465230523      U       ST415   0       1       1
U100    S204    A302    1495130523      1475130523      1465130523      AU      ST408   2       1       1
U104    S202    A303    1465230523      1475130523      1465130523      A       ST409   2       0       1
U102    S207    A301    1465230523      1485130523      1465230523      AU      ST403   3       1       1
        S203    A302    1495130523      1475130523      1465230523      E       ST400   0       0       1
U106    S202    A302    1465230523      1465130523      1465130523      AU      ST408   0       1       1
U105    S207    A300    1465230523      1485130523      1465130523      U       ST400   2       0       1
U108    S205    A304    1465130523      1465130523      1475130523              ST410   2       1       0
U105    S203            1475130523      1465230523      1465130523      AU      ST408   2       0       1
U110    S203    A300    1465230523      1465130523      1485130523      A       ST415   0       1       1
U113    S200    A303    1465230523      1475130523      1465130523      E       ST413   3       1       1
U119    S208    A302    1495130523      1465230523      1465230523      U       ST415   3       0       0
U118    S208    A303    1475130523      1465130523      1465230523      E       ST415   3       0       0
U107    S210    A302    1475130523      1485130523      1485130523      AP      ST404   2       1       0
U118    S202    A300    1495130523      1465230523      1465230523      AP      ST410   1       0       0
U111    S206    A305    1465130523      1465130523      1485130523      AU      ST415   0       1       1
U116    S208    A303    1465230523      1485130523      1475130523      A       ST413   1       0       1
U101    S202    A300    1465230523      1465130523      1475130523      U       ST401   0       0       1
U120    S206    A303    1495130523      1485130523      1465130523      AU      ST414   0       0       0
U106    S205    A300    1462863262      1462863262      1494297562      AP      ST407   2       1       1
U114    S209    A303    1465490556      1462863262      1494297562      U       ST411   2       1       0
U113    S203    A304    1465490556      1465490556      1462863262      U       ST405   0       0       1
U108    S200    A302    1468094889      1462863262      1468094889      U       ST414   0       0       1
U102    S203    A305    1465490556      1465490556      1494297562      U       ST404   2       0       0
        S208    A300    1465490556      1494297562      1465490556      U       ST411   1       0       1
U115    S200    A300    1465490556      1494297562      1465490556      AU      ST404   3       0       0
U111    S204    A300    1465490556      1465490556      1468094889      U       ST410   3       1       1
U120    S201    A300    1494297562      1465490556      1468094889              ST410   3       0       1
U113    S203            1465490556      1465490556      1465490556      A       ST402   1       1       0
U109    S203    A304    1462863262      1494297562      1468094889      E       ST405   1       1       1
U110    S202    A303    1494297562      1494297562      1468094889      AU      ST402   2       1       0
U100    S200    A301    1494297562      1494297562      1494297562      AP      ST410   3       1       1
U101    S208    A300    1462863262      1468094889      1462863262      E       ST408   0       1       1
```

Data Enrichment

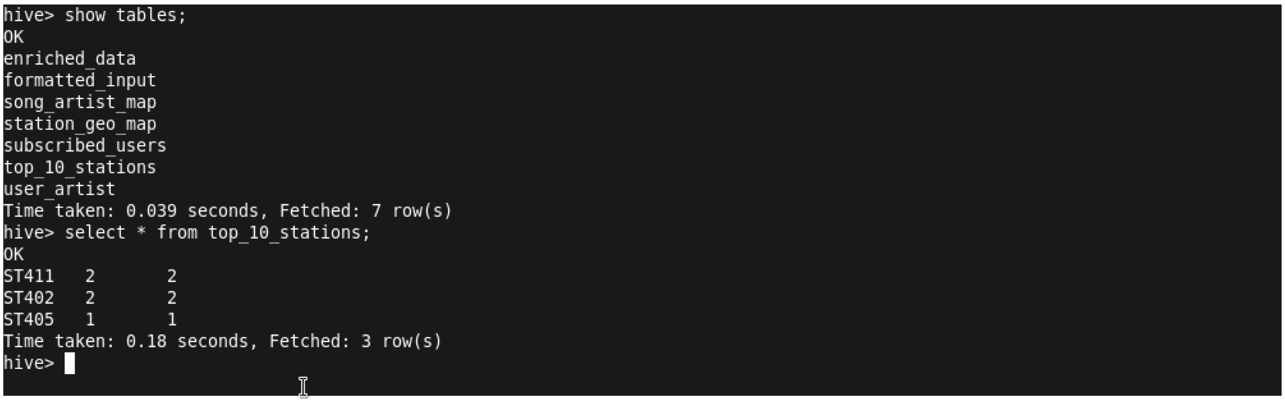


```
hive> CREATE TABLE IF NOT EXISTS enriched_data (User_id STRING,Song_id STRING,Artist_id STRING,stamp STRING,Start_ts STRING,E
nd_ts STRING,Geo_cd STRING,Station_id STRING,Song_end_type INT,datalike INT,Dislike INT)PARTITIONED BY(status STRING) STORED
AS ORC;
OK
Time taken: 0.135 seconds
hive> describe enriched_data;
OK
user_id                 string
song_id                 string
artist_id               string
stamp                   string
start_ts                string
end_ts                  string
geo_cd                  string
station_id              string
song_end_type           int
datalike                int
dislike                 int
status                  string

# Partition Information
# col_name              data_type               comment

status                  string
Time taken: 0.096 seconds, Fetched: 17 row(s)
hive>
```

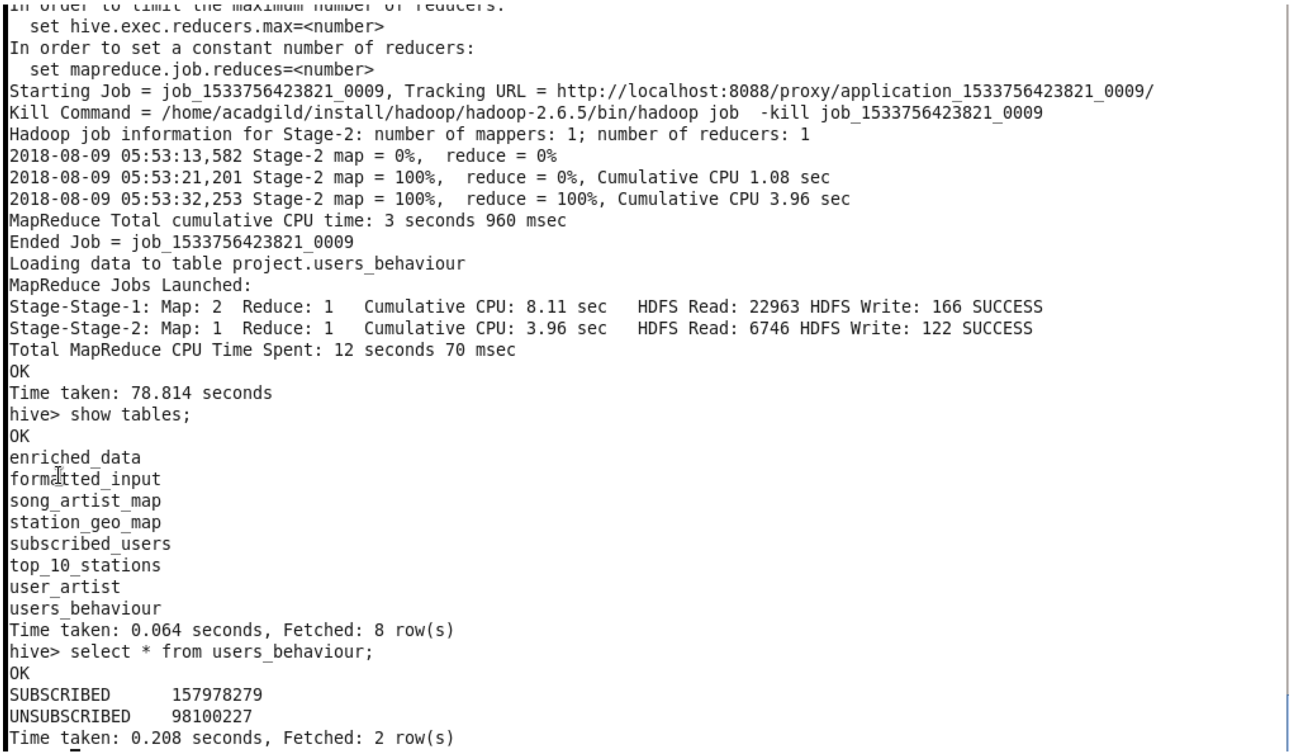


```
subscribed_users
user_artist
Time taken: 0.069 seconds, Fetched: 6 row(s)
hive> select * from enriched_data;
OK
U113    S200    A300    1465230523      1475130523      1465130523      J       ST413   3       1       1       fail
U100    S200    A300    1494297562      1494297562      1494297562      A       ST410   3       1       1       fail
U120    S201    A301    1494297562      1465490556      1468094889      A       ST410   3       0       1       fail
U107    S202    A302    1495130523      1465230523      1465230523      NULL    ST415   0       1       1       fail
U103    S202    A302    1465490556      1465490556      1465490556      NULL    ST415   2       1       1       fail
U106    S202    A302    1465230523      1465130523      1465130523      E       ST408   0       1       1       fail
U109    S203    A303    1462863262      1494297562      1468094889      A       ST405   1       1       1       fail
        S203    A303    1495130523      1475130523      1465230523      A       ST400   0       0       1       fail
U110    S203    A303    1465230523      1465130523      1485130523      NULL    ST415   0       1       1       fail
U111    S204    A304    1465490556      1465490556      1468094889      A       ST410   3       1       1       fail
U113    S204    A304    1494297562      1494297562      1465490556      NULL    ST415   3       0       1       fail
U100    S204    A304    1495130523      1475130523      1465130523      E       ST408   2       1       1       fail
U106    S205    A301    1462863262      1462863262      1494297562      AP      ST407   2       1       1       fail
U108    S205    A301    1465130523      1465130523      1475130523      A       ST410   2       1       0       fail
U111    S206    A302    1465130523      1465130523      1485130523      NULL    ST415   0       1       1       fail
U114    S207    A303    1465130523      1465230523      1475130523      NULL    ST415   3       1       0       fail
U102    S207    A303    1465230523      1485130523      1465230523      J       ST403   3       1       1       fail
        S208    A304    1465490556      1494297562      1465490556      A       ST411   1       0       1       fail
U118    S208    A304    1475130523      1465130523      1465230523      NULL    ST415   3       0       0       fail
U119    S208    A304    1495130523      1465230523      1465230523      NULL    ST415   3       0       0       fail
U101    S208    A304    1462863262      1468094889      1462863262      E       ST408   0       1       1       fail
U107    S210    NULL    1475130523      1485130523      1485130523      E       ST404   2       1       0       fail
U115    S200    A300    1465490556      1494297562      1465490556      E       ST404   3       0       0       pass
U108    S200    A300    1468094889      1462863262      1468094889      E       ST414   0       0       1       pass
U107    S202    A302    1494297562      1468094889      1462863262      E       ST409   0       0       0       pass
U101    S202    A302    1465230523      1465130523      1475130523      AU      ST401   0       0       1       pass
U110    S202    A302    1494297562      1494297562      1468094889      AP      ST402   2       1       0       pass
U118    S202    A302    1495130523      1465230523      1465230523      A       ST410   1       0       0       pass
U104    S202    A302    1465230523      1475130523      1465130523      E       ST409   2       0       1       pass
U102    S203    A303    1465490556      1465490556      1494297562      E       ST404   2       0       0       pass
U113    S203    A303    1465490556      1465490556      1462863262      A       ST405   0       0       1       pass
U113    S203    A303    1462863262      1468094889      1494297562      E       ST408   2       0       0       pass
U105    S203    A303    1475130523      1465230523      1465130523      E       ST408   2       0       1       pass
U113    S203    A303    1465490556      1465490556      1465490556      AP      ST402   1       1       0       pass
```

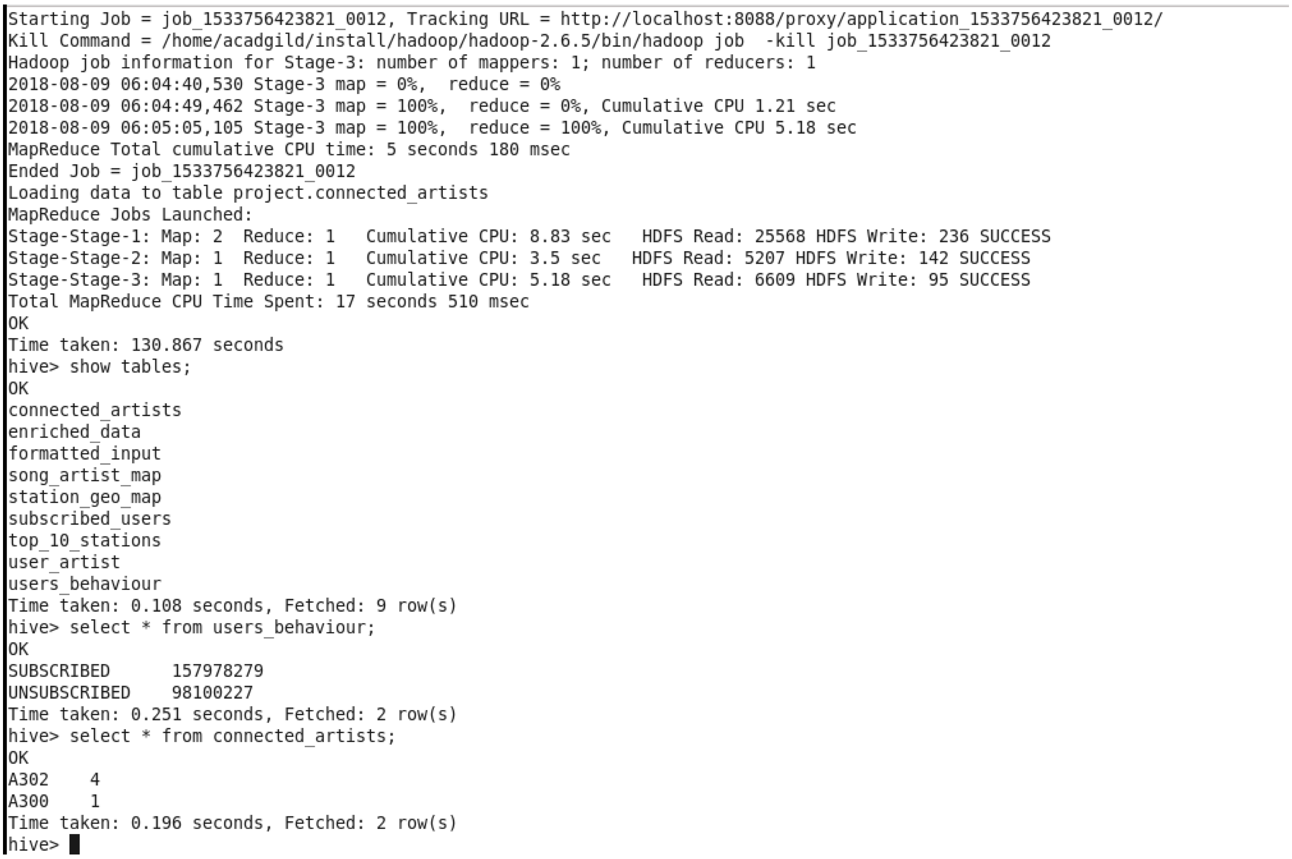
Data Analysis



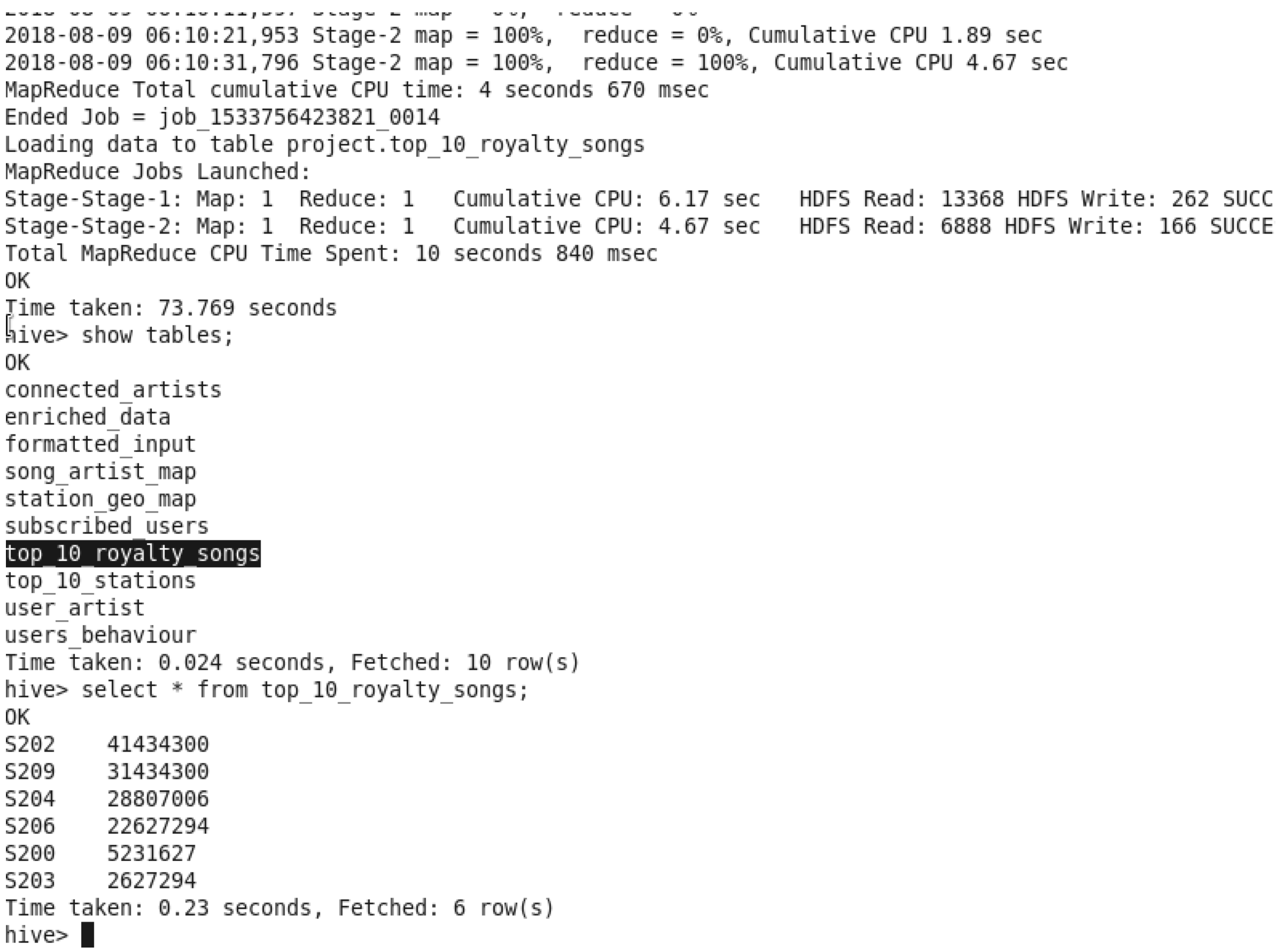
```
hive> show tables;
OK
enriched_data
formatted_input
song_artist_map
station_geo_map
subscribed_users
top_10_stations
user_artist
Time taken: 0.039 seconds, Fetched: 7 row(s)
hive> select * from top_10_stations;
OK
ST411   2       2
ST402   2       2
ST405   1       1
Time taken: 0.18 seconds, Fetched: 3 row(s)
hive>
```



```
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_1533756423821_0009, Tracking URL = http://localhost:8088/proxy/application_1533756423821_0009/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1533756423821_0009
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-08-09 05:53:13,582 Stage-2 map = 0%,  reduce = 0%
2018-08-09 05:53:21,201 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.08 sec
2018-08-09 05:53:32,253 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.96 sec
MapReduce Total cumulative CPU time: 3 seconds 960 msec
Ended Job = job_1533756423821_0009
Loading data to table project.users_behaviour
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 8.11 sec   HDFS Read: 22963 HDFS Write: 166 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.96 sec   HDFS Read: 6746 HDFS Write: 122 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 70 msec
OK
Time taken: 78.814 seconds
hive> show tables;
OK
enriched_data
formatted_input
song_artist_map
station_geo_map
subscribed_users
top_10_stations
user_artist
users_behaviour
Time taken: 0.064 seconds, Fetched: 8 row(s)
hive> select * from users_behaviour;
OK
SUBSCRIBED      157978279
UNSUBSCRIBED    98100227
Time taken: 0.208 seconds, Fetched: 2 row(s)
```



```
Starting Job = job_1533756423821_0012, Tracking URL = http://localhost:8088/proxy/application_1533756423821_0012/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1533756423821_0012
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2018-08-09 06:04:40,530 Stage-3 map = 0%,  reduce = 0%
2018-08-09 06:04:49,462 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 1.21 sec
2018-08-09 06:05:05,105 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 5.18 sec
MapReduce Total cumulative CPU time: 5 seconds 180 msec
Ended Job = job_1533756423821_0012
Loading data to table project.connected_artists
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 8.83 sec   HDFS Read: 25568 HDFS Write: 236 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.5 sec   HDFS Read: 5207 HDFS Write: 142 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 5.18 sec   HDFS Read: 6609 HDFS Write: 95 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 510 msec
OK
Time taken: 130.867 seconds
hive> show tables;
OK
connected_artists
enriched_data
formatted_input
song_artist_map
station_geo_map
subscribed_users
top_10_stations
user_artist
users_behaviour
Time taken: 0.108 seconds, Fetched: 9 row(s)
hive> select * from users_behaviour;
OK
SUBSCRIBED      157978279
UNSUBSCRIBED    98100227
Time taken: 0.251 seconds, Fetched: 2 row(s)
hive> select * from connected_artists;
OK
A302    4
A300    1
Time taken: 0.196 seconds, Fetched: 2 row(s)
hive>
```



```
2018-08-09 06:10:11,557 Stage-2 map = 0%,  reduce = 0%
2018-08-09 06:10:21,953 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.89 sec
2018-08-09 06:10:31,796 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.67 sec
MapReduce Total cumulative CPU time: 4 seconds 670 msec
Ended Job = job_1533756423821_0014
Loading data to table project.top_10_royalty_songs
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.17 sec   HDFS Read: 13368 HDFS Write: 262 SUCC
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.67 sec   HDFS Read: 6888 HDFS Write: 166 SUCCE
Total MapReduce CPU Time Spent: 10 seconds 840 msec
OK
Time taken: 73.769 seconds
hive> show tables;
OK
connected_artists
enriched_data
formatted_input
song_artist_map
station_geo_map
subscribed_users
top_10_royalty_songs
top_10_stations
user_artist
users_behaviour
Time taken: 0.024 seconds, Fetched: 10 row(s)
hive> select * from top_10_royalty_songs;
OK
S202    41434300
S209    31434300
S204    28807006
S206    22627294
S200    5231627
S203    2627294
Time taken: 0.23 seconds, Fetched: 6 row(s)
hive> 
```