

MUSIC DATA ANALYSIS

Final Project



SEPTEMBER 8, 2018

Table of Contents

Introduction & Data-set.....	1
Look-Up Tables Files	3
Perform analysis on the Music Data	5
Step 1: Launch all necessary daemons	5
Step 2: Start Job Scheduling	7
Step 3: Populate Look-Up tables	7
Step 4: Perform Data Formatting	12
Step 5: Perform Data Enrichment and Cleaning	13
Step 6: Perform Data Analysis.....	18
Post Analysis	Error! Bookmark not defined.

Introduction & Data-set

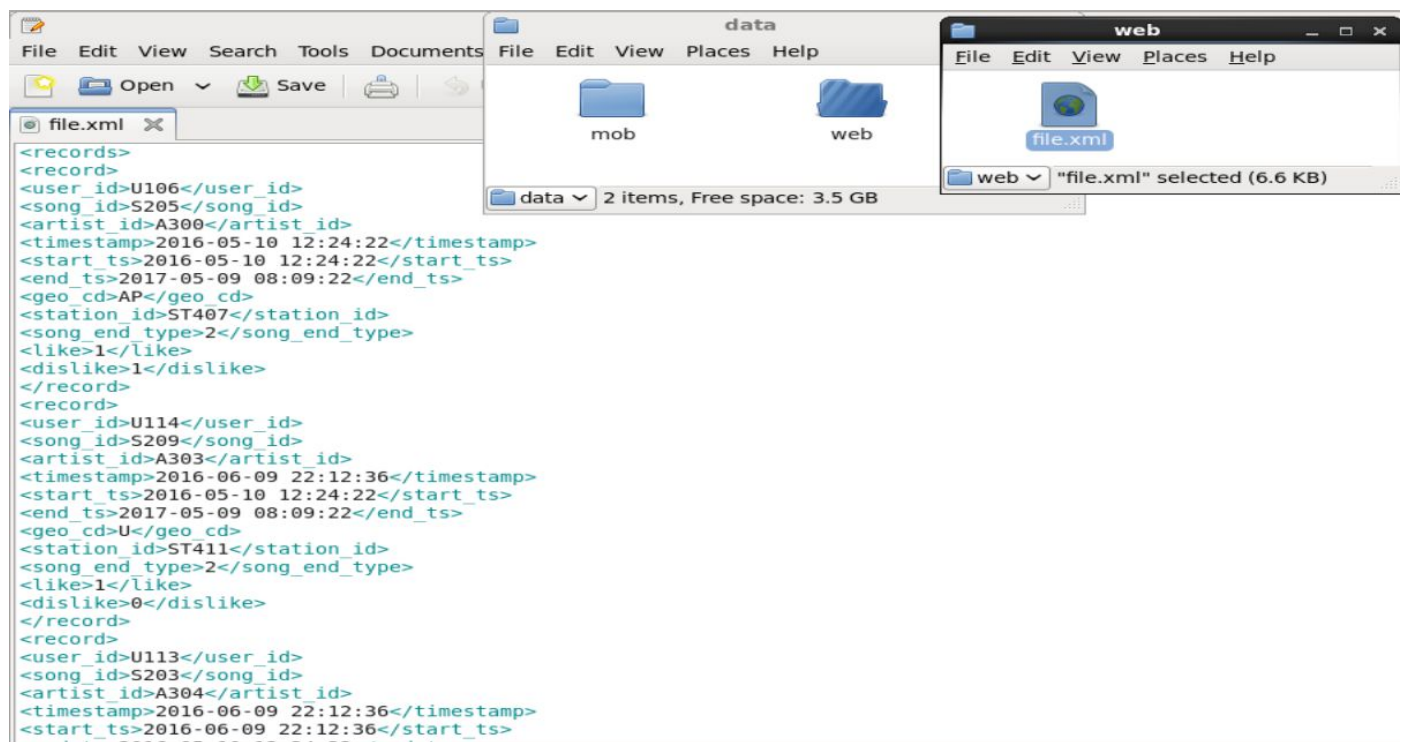
A leading music-catering company is planning to analyze large amount of data received from varieties of sources, namely mobile app and website to track the behavior of users, classify users, calculate royalties associated with the song and make appropriate business strategies. The file server receives data files periodically every 3 hours.

Data files contain below fields.

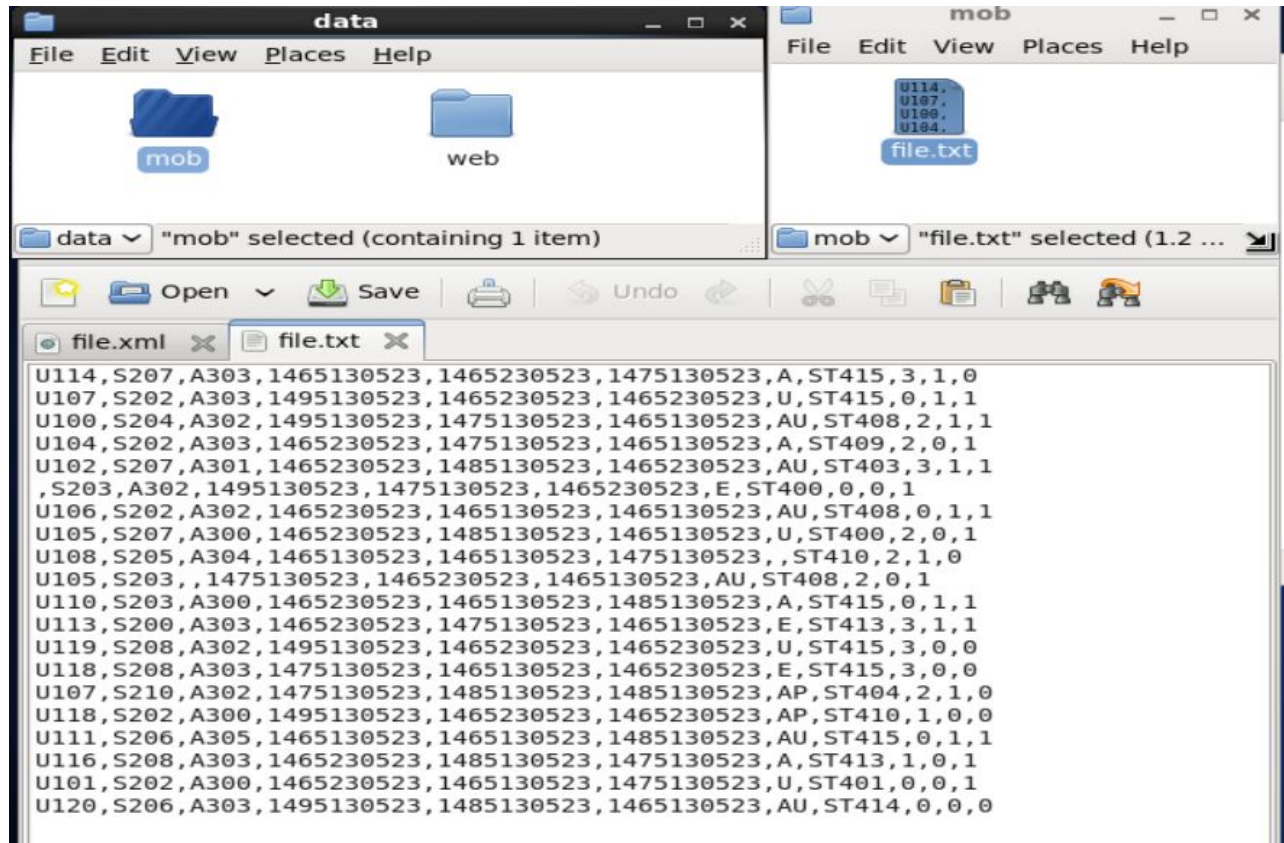
Column Name/Field Name	Column Description/Field Description
User_id	Unique identifier of every user
Song_id	Unique identifier of every song
Artist_id	Unique identifier of the lead artist of the song
Timestamp	Timestamp when the record was generated

Start_ts	Start timestamp when the song started to play
End_ts	End timestamp when the song was stopped
Geo_cd	Can be 'A' for USA region, 'AP' for asia pacific region, 'J' for Japan region, 'E' for europe and 'AU' for australia region
Station_id	Unique identifier of the station from where the song was played
Song_end_type	How the song was terminated. 0 means completed successfully 1 means song was skipped 2 means song was paused 3 means other type of failure like device issue, network error etc.
Like	0 means song was not liked 1 means song was liked
Dislike	0 means song was not disliked 1 means song was disliked

Data coming from web applications reside in /data/web and has xml format.



Data coming from mobile applications reside in /data/mob and has csv format.

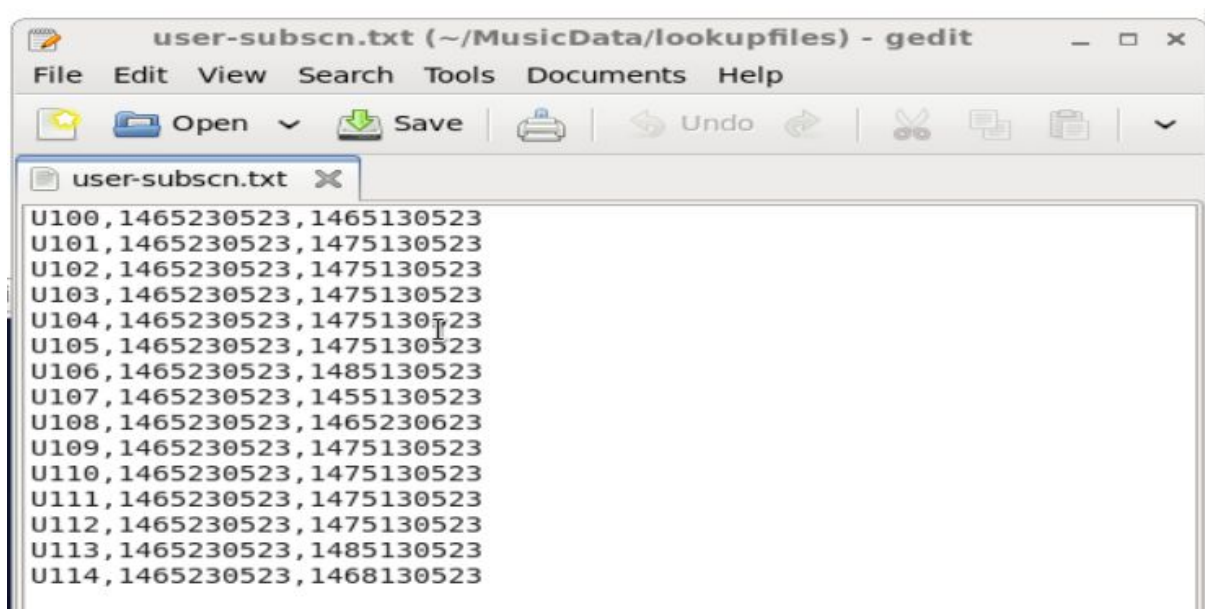
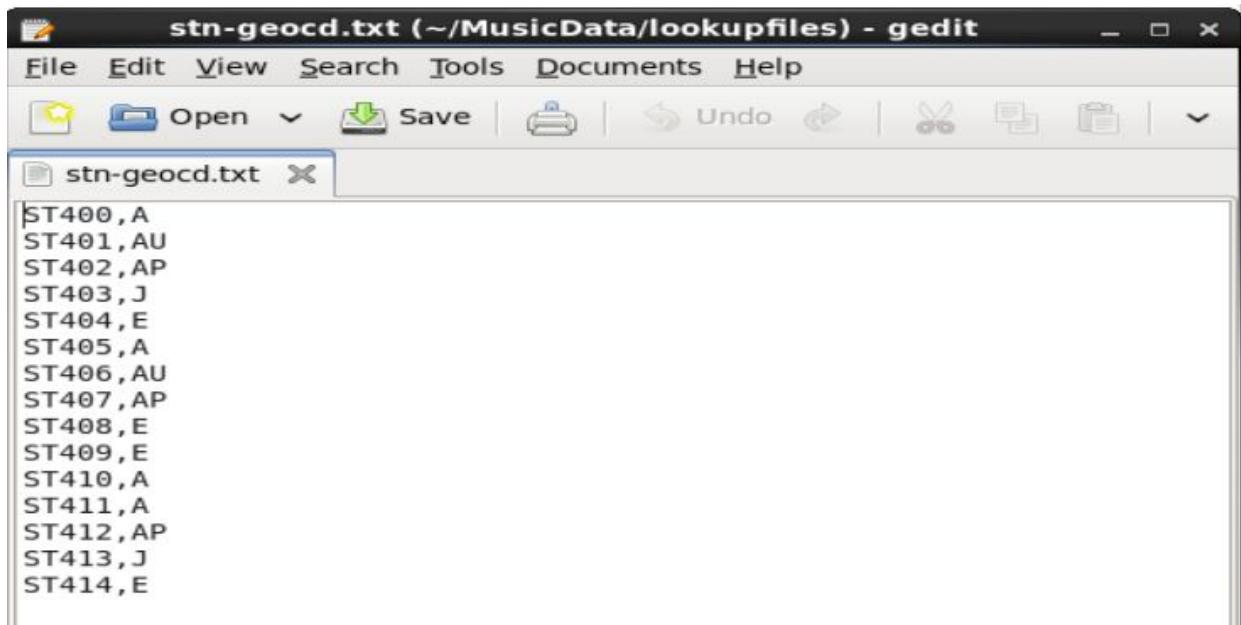
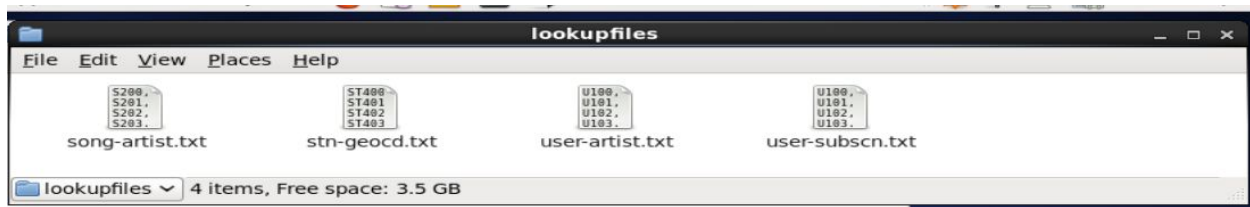


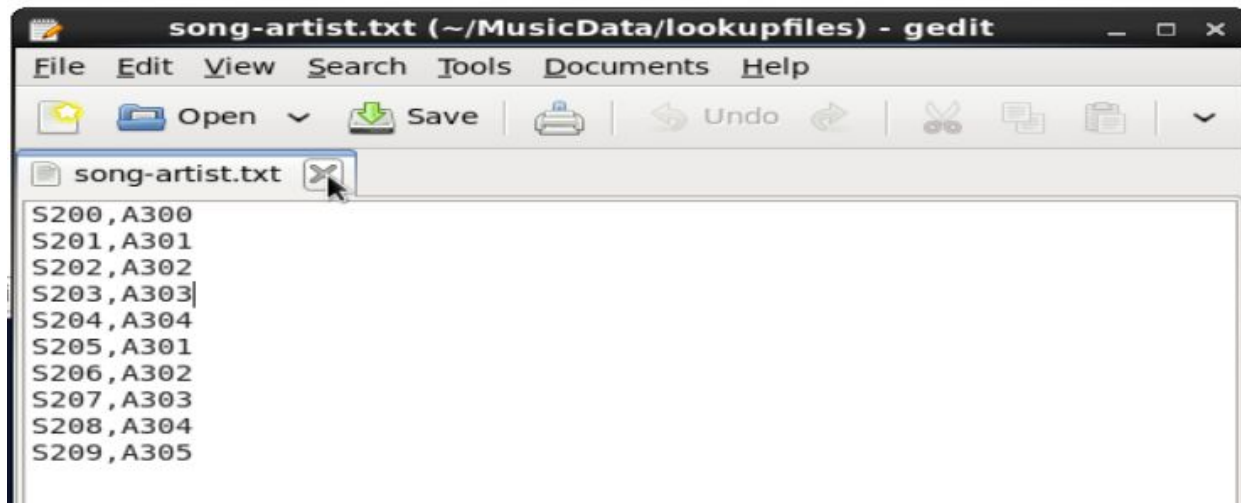
Look-Up Tables Files

There are some existing look up tables present in NoSQL databases. They play an important role in data enrichment and analysis.

Table Name	Description
Station_Geo_Map	Contains mapping of a geo_cd with station_id
Subscribed_Users	Contains user_id, subscription_start_date and subscription_end_date. Contains details only for subscribed users
Song_Artist_Map	Contains mapping of song_id with artist_id along with royalty associated with each play of the song

User_Artist_Map	Contains an array of artist_id(s) followed by a user_id
-----------------	---





```
S200,A300
S201,A301
S202,A302
S203,A303
S204,A304
S205,A301
S206,A302
S207,A303
S208,A304
S209,A305
```



```
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
```

Steps to perform data analysis on the Music Data

Step 1: Launch all necessary daemons

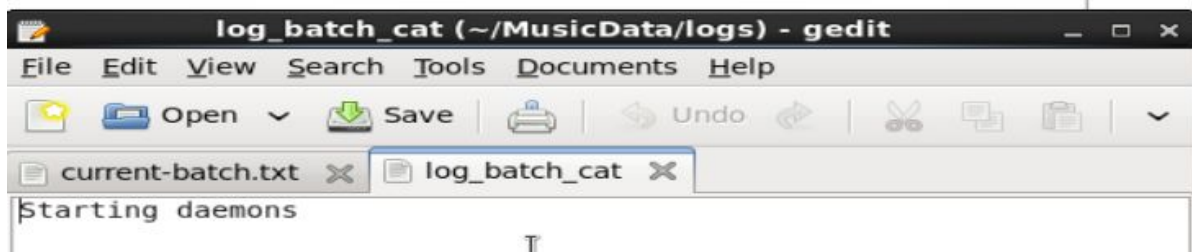
- Start Mysql service
- Execute start-daemons.sh script
- Start hive metaserver


```
acadgild@localhost:~  
File Edit View Search Terminal Help  
[acadgild@localhost ~]$ sudo service mysql start  
[sudo] password for acadgild:  
Starting mysql: [ OK ]  
[acadgild@localhost ~]$
```

```
acadgild@localhost:~/MusicData/files  
File Edit View Search Terminal Help  
#!/bin/bash  
  
if [ -f "/home/acadgild/MusicData/logs/current-batch.txt" ]  
then  
    echo "Batch File Found!"  
else  
    echo -n "1" > "/home/acadgild/MusicData/logs/current-batch.txt"  
fi  
  
chmod 775 /home/acadgild/MusicData/logs/current-batch.txt  
batchid=$(cat "/home/acadgild/MusicData/logs/current-batch.txt")  
LOGFILE=/home/acadgild/MusicData/logs/log_batch_$batchid  
  
echo "Starting daemons" >> $LOGFILE  
  
start-all.sh  
start-hbase.sh
```

```
[acadgild@localhost files]$ ./start-daemons.sh  
/home/acadgild/MusicData/logs/current-batch.txt: line 1: 1: command not found  
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh  
18/09/05 23:56:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
Starting namenodes on [localhost]  
localhost: starting namenode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-namenode-localhost.localdomain.out  
localhost: starting datanode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-datanode-localhost.localdomain.out  
Starting secondary namenodes [0.0.0.0]  
0.0.0.0: starting secondarynamenode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-secondarynamenode-localhost.localdomain.out  
18/09/05 23:57:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
starting yarn daemons  
starting resourcemanager, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-resourcemanager-localhost.localdomain.out  
localhost: starting nodemanager, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-nodemanager-localhost.localdomain.out  
localhost: starting zookeeper, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-zookeeper-localhost.localdomain.out  
starting master, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-master-localhost.localdomain.out  
starting regionserver, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-1-regionserver-localhost.localdomain.out  
You have new mail in /var/spool/mail/acadgild  
[acadgild@localhost files]$
```





Step 2: Start Job Scheduling

All the individual scripts ran here can be wrapped into a wrapper script and scheduled using crontab. Below is the command that can be used for scheduling.

```
* */3 * * * /home/acadgild/project/scripts/wrapper.sh
```

In the -e mode, Crontab schedules execution of commands by a regular user.

The statement above runs the wrapper.sh shell script every 3 hours.

Step 3: Populate Look-Up tables

Populate-lookup.sh script loads data from lookup tables to Hbase. Creates Hbase tables song-artist-map, station-geo-map and subscribed-users

```
acadgild@localhost:~/MusicData/files
File Edit View Search Terminal Help
[acadgild@localhost files]$ ./populate-lookup.sh
2018-09-06 00:47:52,195 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uilt-in java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017

create 'station-geo-map', 'geo'
0 row(s) in 5.4670 seconds

Hbase::Table - station-geo-map
2018-09-06 00:48:34,988 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uilt-in java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017
```



```

acadgild@localhost: ~/MusicData/files
File Edit View Search Terminal Help
create 'subscribed-users', 'subscn'
0 row(s) in 3.7490 seconds

Hbase::Table - subscribed-users
2018-09-06 00:49:21,976 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
ultin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017

create 'song-artist-map', 'artist'
0 row(s) in 3.3020 seconds

Hbase::Table - song-artist-map
2018-09-06 00:49:49,731 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
ultin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017

put 'station-geo-map', 'ST400', 'geo:geo_cd', 'A'
0 row(s) in 0.8100 seconds

2018-09-06 00:50:32,504 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
ultin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017

put 'station-geo-map', 'ST401', 'geo:geo_cd', 'AU'
0 row(s) in 2.9350 seconds

2018-09-06 00:51:26,145 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
ultin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017

```

```

acadgild@localhost:~/MusicData/files
File Edit View Search Terminal Help
put 'station-geo-map', 'ST402', 'geo:geo_cd', 'AP'
0 row(s) in 2.6940 seconds

2018-09-06 00:52:04,441 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uiltin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017

put 'station-geo-map', 'ST403', 'geo:geo_cd', 'J'
0 row(s) in 0.7720 seconds

2018-09-06 00:52:35,929 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uiltin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017

put 'station-geo-map', 'ST404', 'geo:geo_cd', 'E'
0 row(s) in 1.0560 seconds

2018-09-06 00:53:20,474 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uiltin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static

```

```

put Devices Help

2018-09-08 19:22:35,947 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uiltin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017

put 'subscribed-users', 'U114', 'subscn:enndt', '1468130523'
0 row(s) in 1.3440 seconds

SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j
/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]

Logging initialized using configuration in jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.j
ar!/hive-log4j2.properties Async: true
OK
Time taken: 23.042 seconds
OK
Time taken: 0.172 seconds
OK
Time taken: 3.909 seconds
Loading data to table project.users_artists
OK
Time taken: 6.448 seconds
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost files]$

```

```
hbase(main):001:0> list
TABLE
SparkHBasesTable
song-artist-map
station-geo-map
subscribed-users
4 row(s) in 1.7190 seconds

=> ["SparkHBasesTable", "song-artist-map", "station-geo-map", "subscribed-users"]
```

```
hbase(main):002:0> scan 'song-artist-map'
ROW COLUMN+CELL
S200 column=artist:artistid, timestamp=1536413899381, value=A300
S201 column=artist:artistid, timestamp=1536413919251, value=A301
S202 column=artist:artistid, timestamp=1536413938323, value=A302
S203 column=artist:artistid, timestamp=1536413959317, value=A303
S204 column=artist:artistid, timestamp=1536413978244, value=A304
S205 column=artist:artistid, timestamp=1536413999045, value=A301
S206 column=artist:artistid, timestamp=1536414016798, value=A302
S207 column=artist:artistid, timestamp=1536414042581, value=A303
S208 column=artist:artistid, timestamp=1536414063712, value=A304
S209 column=artist:artistid, timestamp=1536414084344, value=A305
10 row(s) in 1.0120 seconds
```

```
hbase(main):003:0> scan 'station-geo-map'
ROW COLUMN+CELL
ST400 column=geo:geo_cd, timestamp=1536413584597, value=A
ST401 column=geo:geo_cd, timestamp=1536413606754, value=AU
ST402 column=geo:geo_cd, timestamp=1536413637009, value=AP
ST403 column=geo:geo_cd, timestamp=1536413661266, value=J
ST404 column=geo:geo_cd, timestamp=1536413679330, value=E
ST405 column=geo:geo_cd, timestamp=1536413698719, value=A
ST406 column=geo:geo_cd, timestamp=1536413716961, value=AU
ST407 column=geo:geo_cd, timestamp=1536413740805, value=AP
ST408 column=geo:geo_cd, timestamp=1536413759082, value=E
ST409 column=geo:geo_cd, timestamp=1536413784220, value=E
ST410 column=geo:geo_cd, timestamp=1536413802422, value=A
ST411 column=geo:geo_cd, timestamp=1536413825557, value=A
ST412 column=geo:geo_cd, timestamp=1536413843685, value=AP
ST413 column=geo:geo_cd, timestamp=1536413862000, value=J
ST414 column=geo:geo_cd, timestamp=1536413881031, value=E
15 row(s) in 0.8260 seconds
```

```
hbase(main):004:0> scan 'subscribed-users'
```

```

hbase(main):004:0> scan 'subscribed-users'
ROW                                COLUMN+CELL
U100                               column=subscn:enddt, timestamp=1536414123043, value=1465130523
U100                               column=subscn:startdt, timestamp=1536414103949, value=1465230523
U101                               column=subscn:enddt, timestamp=1536414169641, value=1475130523
U101                               column=subscn:startdt, timestamp=1536414149892, value=1465230523
U102                               column=subscn:enddt, timestamp=1536414217966, value=1475130523
U102                               column=subscn:startdt, timestamp=1536414196984, value=1465230523
U103                               column=subscn:enddt, timestamp=1536414258802, value=1475130523
U103                               column=subscn:startdt, timestamp=1536414238555, value=1465230523
U104                               column=subscn:enddt, timestamp=1536414300973, value=1475130523
U104                               column=subscn:startdt, timestamp=1536414279985, value=1465230523
U105                               column=subscn:enddt, timestamp=1536414342825, value=1475130523
U105                               column=subscn:startdt, timestamp=1536414321428, value=1465230523
U106                               column=subscn:enddt, timestamp=1536414386528, value=1485130523
U106                               column=subscn:startdt, timestamp=1536414364984, value=1465230523
U107                               column=subscn:enddt, timestamp=1536414430379, value=1455130523
U107                               column=subscn:startdt, timestamp=1536414407867, value=1465230523
U108                               column=subscn:enddt, timestamp=1536414477078, value=1465230623
U108                               column=subscn:startdt, timestamp=1536414452074, value=1465230523
U109                               column=subscn:enddt, timestamp=1536414521540, value=1475130523
U109                               column=subscn:startdt, timestamp=1536414498971, value=1465230523
U110                               column=subscn:enddt, timestamp=1536414565682, value=1475130523
U110                               column=subscn:startdt, timestamp=1536414542327, value=1465230523
U111                               column=subscn:enddt, timestamp=1536414611343, value=1475130523
U111                               column=subscn:startdt, timestamp=1536414589198, value=1465230523
U112                               column=subscn:enddt, timestamp=1536414656462, value=1475130523
U112                               column=subscn:startdt, timestamp=1536414636121, value=1465230523
U113                               column=subscn:enddt, timestamp=1536414714021, value=1485130523
U113                               column=subscn:startdt, timestamp=1536414693713, value=1465230523
U114                               column=subscn:enddt, timestamp=1536414760922, value=1468130523
U114                               column=subscn:startdt, timestamp=1536414740738, value=1465230523
15 row(s) in 1.5030 seconds

hbase(main):005:0> █

```

Hive script user-artist.hql populates hive table user-artists by fetching data from lookup table user-artist.

```

hive> show databases;
OK
default
project
Time taken: 15.541 seconds, Fetched: 2 row(s)
hive> use project;
OK
Time taken: 0.148 seconds
hive> show tables;
OK
users_artists
Time taken: 0.33 seconds, Fetched: 1 row(s)
hive> select * from users_artists;
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: 8.477 seconds, Fetched: 15 row(s)
hive> █

```

acacnild@localhost acacnild@localhost [acacnild] [MusicData] [logs]

Step 4: Perform Data Enrichment

Creates lookup tables in hive and imports data from hbase lookup tables.

```
acadgild@localhost:~/MusicData/files
File Edit View Search Terminal Help
[acadgild@localhost ~]$ cd MusicData/files/
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost files]$
[acadgild@localhost files]$
[acadgild@localhost files]$
[acadgild@localhost files]$
[acadgild@localhost files]$ ./data_enrichment filtering_schema.sh
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]

Logging initialized using configuration in jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
OK
Time taken: 121.082 seconds
OK
Time taken: 56.045 seconds
OK
Time taken: 2.906 seconds
OK
Time taken: 3.046 seconds
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost files]$
```

```
1.X releases.
hive> show databases;
OK
default
project
Time taken: 31.608 seconds, Fetched: 2 row(s)
hive> use project;
OK
Time taken: 0.214 seconds
hive> show tables;
OK
song_artist_map
station_geo_map
subscribed_users
users_artists
Time taken: 0.464 seconds, Fetched: 4 row(s)
hive>
```

```

Time taken: 17.431 seconds, Fetched: 10 row(s)
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: 17.431 seconds, Fetched: 10 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: 2.512 seconds, Fetched: 15 row(s)
hive> █

Time taken: 2.145 seconds, Fetched: 15 row(s)
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: 2.145 seconds, Fetched: 15 row(s)
hive> █

```

Step 5: Perform Data Formatting

Creates a table formatted_input in hive and populates it with mobile and web data.


```
18/09/10 19:23:50 INFO deprecation: mapred.input.dir.recursive is deprecated. Instead, use mapreduce.input.fileinputformat.input.dir.rec
18/09/10 19:23:59 INFO metastore: Trying to connect to metastore with URI thrift://localhost:9083
18/09/10 19:24:19 INFO metastore: Connected to metastore.
18/09/10 19:24:37 INFO SessionState: Created local directory: /tmp/4206e71f-c9f5-4391-9e6b-9bfadc766485_resources
18/09/10 19:24:37 INFO SessionState: Created HDFS directory: /tmp/hive/acadgild/4206e71f-c9f5-4391-9e6b-9bfadc766485
18/09/10 19:24:38 INFO SessionState: Created local directory: /tmp/acadgild/4206e71f-c9f5-4391-9e6b-9bfadc766485
18/09/10 19:24:38 INFO SessionState: Created HDFS directory: /tmp/hive/acadgild/4206e71f-c9f5-4391-9e6b-9bfadc766485/_tmp_space.db
18/09/10 19:24:38 INFO HiveClientImpl: Warehouse location for Hive client (version 1.2.1) is /user/hive/warehouse
18/09/10 19:24:39 INFO SessionState: Created local directory: /tmp/e65bb229-4239-4481-8380-0d7e961b856f_resources
18/09/10 19:24:39 INFO SessionState: Created HDFS directory: /tmp/hive/acadgild/e65bb229-4239-4481-8380-0d7e961b856f
18/09/10 19:24:39 INFO SessionState: Created local directory: /tmp/acadgild/e65bb229-4239-4481-8380-0d7e961b856f
18/09/10 19:24:39 INFO SessionState: Created HDFS directory: /tmp/hive/acadgild/e65bb229-4239-4481-8380-0d7e961b856f/_tmp_space.db
18/09/10 19:24:39 INFO HiveClientImpl: Warehouse location for Hive client (version 1.2.1) is /user/hive/warehouse
18/09/10 19:25:01 INFO SparkSqlParser: Parsing command: web_data
18/09/10 19:25:11 INFO BlockManagerInfo: Removed broadcast_1_piece0 on 10.0.2.15:37167 in memory (size: 2.6 KB, free: 111.2 MB)
18/09/10 19:25:11 INFO BlockManagerInfo: Removed broadcast_0_piece0 on 10.0.2.15:37167 in memory (size: 14.6 KB, free: 111.2 MB)
18/09/10 19:25:13 INFO SparkSqlParser: Parsing command: CREATE TABLE IF NOT EXISTS project.formatted_input (
  User_id STRING, Song_id STRING, Artist_id STRING, Timestamp STRING,
  Start_ts STRING, End_ts STRING, Geo_cd STRING, Station_id STRING,
  Song_end_type INT, Like INT, Dislike INT ) PARTITIONED BY (batchid INT)
  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
18/09/10 19:25:56 INFO SparkSqlParser: Parsing command: LOAD DATA LOCAL INPATH '/home/acadgild/MusicData/data/mob/file.txt'
  INTO TABLE project.formatted_input PARTITION (batchid='1')
18/09/10 19:26:11 INFO FileUtils: Creating directory if it doesn't exist: hdfs://localhost:8020/user/hive/warehouse/project.db/formatted
18/09/10 19:26:12 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 19:26:15 INFO Hive: Renaming src: file:/home/acadgild/MusicData/data/mob/file.txt, dest: hdfs://localhost:8020/user/hive/wareh
18/09/10 19:26:18 INFO SparkSqlParser: Parsing command: INSERT INTO project.formatted_input PARTITION (batchid='1')
  SELECT user_id, song_id, artist_id, unix_timestamp(timestamp,'yyyy-MM-dd HH:mm:ss')
  AS timestamp, unix_timestamp(start_ts,'yyyy-MM-dd HH:mm:ss') AS start_ts,
  unix_timestamp(end_ts,'yyyy-MM-dd HH:mm:ss') AS end_ts, geo_cd, station_id,
  song_end_type, like, dislike FROM web_data
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: int
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: string
```

```
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: int
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: int
18/09/10 19:26:21 INFO CatalystSqlParser: Parsing command: int
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: int
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: string
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: int
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: int
18/09/10 19:26:24 INFO CatalystSqlParser: Parsing command: int
18/09/10 19:26:25 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 128.4 KB, free 111.1 MB)
18/09/10 19:26:25 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 14.6 KB, free 111.0 MB)
18/09/10 19:26:26 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 10.0.2.15:37167 (size: 14.6 KB, free: 111.2 MB)
18/09/10 19:26:26 INFO SparkContext: Created broadcast 2 from newAPIHadoopFile at XmlFile.scala:46
18/09/10 19:26:26 INFO FileUtils: Creating directory if it doesn't exist: hdfs://localhost:8020/user/hive/warehouse/project.db/formatted
18/09/10 19:26:35 INFO CodeGenerator: Code generated in 5190.90528 ms
18/09/10 19:26:35 INFO deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
18/09/10 19:26:35 INFO deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
18/09/10 19:26:35 INFO deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
18/09/10 19:26:35 INFO deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
18/09/10 19:26:35 INFO deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
18/09/10 19:26:35 INFO FileInputFormat: Total input paths to process : 1
18/09/10 19:26:36 INFO SparkContext: Starting job: sql at DataFormatting.scala:36
18/09/10 19:26:36 INFO DAGScheduler: Got job 1 (sql at DataFormatting.scala:36) with 1 output partitions
18/09/10 19:26:36 INFO DAGScheduler: Final stage: ResultStage 1 (sql at DataFormatting.scala:36)
18/09/10 19:26:36 INFO DAGScheduler: Parents of final stage: List()
18/09/10 19:26:36 INFO DAGScheduler: Missing parents: List()
18/09/10 19:26:36 INFO DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[11] at sql at DataFormatting.scala:36), which has no mis
18/09/10 19:26:36 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size 62.3 KB, free 111.0 MB)
18/09/10 19:26:36 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 23.4 KB, free 111.0 MB)
18/09/10 19:26:36 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on 10.0.2.15:37167 (size: 23.4 KB, free: 111.2 MB)
18/09/10 19:26:36 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1012
18/09/10 19:26:36 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (MapPartitionsRDD[11] at sql at DataFormatting.scala:
```

```

18/09/10 19:26:36 INFO DAGScheduler: Job 1 (sql at DataFormatting.scala:36) is submitted to the driver with 1 output partitions
18/09/10 19:26:36 INFO DAGScheduler: Final stage: ResultStage 1 (sql at DataFormatting.scala:36)
18/09/10 19:26:36 INFO DAGScheduler: Parents of final stage: List()
18/09/10 19:26:36 INFO DAGScheduler: Missing parents: List()
18/09/10 19:26:36 INFO DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[11] at sql at DataFormatting.scala:36), which has no missing tasks: List()
18/09/10 19:26:36 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size 62.3 KB, free 111.0 MB)
18/09/10 19:26:36 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 23.4 KB, free 111.0 MB)
18/09/10 19:26:36 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on 10.0.2.15:37167 (size: 23.4 KB, free: 111.2 MB)
18/09/10 19:26:36 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1012
18/09/10 19:26:36 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (MapPartitionsRDD[11] at sql at DataFormatting.scala:36)
18/09/10 19:26:36 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
18/09/10 19:26:36 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, partition 0, PROCESS_LOCAL, 5482 bytes)
18/09/10 19:26:36 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
18/09/10 19:26:37 INFO deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
18/09/10 19:26:37 INFO deprecation: mapred.output.value.class is deprecated. Instead, use mapreduce.job.output.value.class
18/09/10 19:26:37 INFO deprecation: mapred.working.dir is deprecated. Instead, use mapreduce.job.working.dir
18/09/10 19:26:37 INFO NewHadoopRDD: Input split: file:/home/acadgild/MusicData/data/web/file.xml:0+6716
18/09/10 19:26:39 INFO FileOutputCommitter: Saved output of task 'attempt_201809101926_0001_m_000000_0' to hdfs://localhost:8020/user/hive/warehouse/project.db/formatted_input/.hive-staging_hive_201809101926_0001_m_000000_0
18/09/10 19:26:39 INFO SparkHadoopMapRedUtil: attempt_201809101926_0001_m_000000_0: Committed
18/09/10 19:26:39 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1382 bytes result sent to driver
18/09/10 19:26:39 INFO DAGScheduler: ResultStage 1 (sql at DataFormatting.scala:36) finished in 2.649 s
18/09/10 19:26:39 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 2650 ms on localhost (1/1)
18/09/10 19:26:39 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/09/10 19:26:39 INFO DAGScheduler: Job 1 finished: sql at DataFormatting.scala:36, took 3.148274 s
18/09/10 19:26:42 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 19:26:42 INFO Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/formatted_input/.hive-staging_hive_201809101926_0001_m_000000_0 to hdfs://localhost:8020/user/hive/warehouse/project.db/formatted_input/formatted_input
18/09/10 19:26:43 INFO SparkContext: Invoking stop() from shutdown hook
18/09/10 19:26:43 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/09/10 19:26:44 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/09/10 19:26:44 INFO MemoryStore: MemoryStore cleared
18/09/10 19:26:44 INFO BlockManager: BlockManager stopped
18/09/10 19:26:44 INFO BlockManagerMaster: BlockManagerMaster stopped
18/09/10 19:26:44 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/09/10 19:26:44 INFO SparkContext: Successfully stopped SparkContext
18/09/10 19:26:44 INFO ShutdownHookManager: Shutdown hook called
18/09/10 19:26:44 INFO ShutdownHookManager: Deleting directory /tmp/spark-588024f5-08de-4066-912c-b2e16b5b6dfd

```

Process finished with exit code 0

```

hive>
> show tables;
OK
formatted_input
song_artist_map
station_geo_map
subscribed_users
users_artists
Time taken: 24.002 seconds, Fetched: 5 row(s)
hive>

```



```
> select * from formatted_input;
```

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

Creates a hive table enriched_data by joining formatted_input, station_geo_map and song_artist_map tables.

```
18/09/10 20:34:16 INFO HiveClientImpl: Warehouse location for Hive client (version 1.2.1) is /user/hive/warehouse
18/09/10 20:35:38 INFO SparkSqlParser: Parsing command: SET hive.exec.dynamic.partition.mode=nonstrict
18/09/10 20:35:39 INFO SparkSqlParser: Parsing command: USE project
18/09/10 20:35:39 INFO SparkSqlParser: Parsing command: CREATE TABLE IF NOT EXISTS enriched_data(
  User_id STRING, Song_id STRING, Artist_id STRING, Timestamp STRING, Start_ts STRING,
  End_ts STRING, Geo_cd STRING, Station_id STRING, Song_end_type INT, Like INT,
  Dislike INT ) PARTITIONED BY (batchid INT,status STRING) STORED AS ORC
18/09/10 20:35:56 INFO SparkSqlParser: Parsing command: INSERT OVERWRITE TABLE enriched_data PARTITION (batchid, status) SELECT i.user_id,
  i.song_id, sa.artist_id, i.timestamp, i.start_ts, i.end_ts, sg.geo_cd, i.station_id,
  IF (i.song_end_type IS NULL, 3, i.song_end_type) AS song_end_type, IF (i.like IS NULL, 0, i.like) AS like,
  IF (i.dislike IS NULL, 0, i.dislike) AS dislike, i.batchid, IF(i.like=1 AND i.dislike=1)
  OR i.user_id IS NULL OR i.song_id IS NULL OR i.timestamp IS NULL OR i.start_ts IS NULL
  OR i.end_ts IS NULL OR i.geo_cd IS NULL OR i.user_id='' OR i.song_id='' OR i.timestamp=''
  OR i.start_ts='' OR i.end_ts='' OR i.geo_cd='' OR sg.geo_cd IS NULL OR sg.geo_cd=''
  OR sa.artist_id IS NULL OR sa.artist_id='', 'fail', 'pass') AS status
FROM formatted_input i LEFT OUTER JOIN station_geo_map sg ON i.station_id = sg.station_id
LEFT OUTER JOIN song_artist_map sa ON i.song_id = sa.song_id WHERE i.batchid=1
18/09/10 20:36:05 INFO CatalystSqlParser: Parsing command: int
18/09/10 20:36:05 INFO CatalystSqlParser: Parsing command: string
18/09/10 20:36:05 INFO CatalystSqlParser: Parsing command: string
18/09/10 20:36:05 INFO CatalystSqlParser: Parsing command: string
18/09/10 20:36:05 INFO CatalystSqlParser: Parsing command: string
18/09/10 20:36:05 INFO CatalystSqlParser: Parsing command: string
18/09/10 20:36:05 INFO CatalystSqlParser: Parsing command: string
18/09/10 20:36:05 INFO CatalystSqlParser: Parsing command: string
18/09/10 20:36:05 INFO CatalystSqlParser: Parsing command: int
18/09/10 20:36:05 INFO CatalystSqlParser: Parsing command: int
18/09/10 20:36:05 INFO CatalystSqlParser: Parsing command: int
18/09/10 20:36:14 INFO deprecation: mapred.job.tracker.persist.jobstatus.hours is deprecated. Instead, use mapreduce.jobtracker.persist
18/09/10 20:36:14 INFO deprecation: mapred.heartbeats.in.second is deprecated. Instead, use mapreduce.jobtracker.heartbeats.in.second
18/09/10 20:36:14 INFO deprecation: topology.node.switch.mapping.impl is deprecated. Instead, use net.topology.node.switch.mapping.impl
18/09/10 20:36:14 INFO deprecation: mapred.skip.map.max.skip.records is deprecated. Instead, use mapreduce.map.skip.maxrecords
18/09/10 20:36:14 INFO deprecation: mapred.job.tracker.jobhistory.lru.cache.size is deprecated. Instead, use mapreduce.jobtracker.jobhis
18/09/10 20:36:14 INFO deprecation: mapred.skip.attempts.to.start.skipping is deprecated. Instead, use mapreduce.task.skip.start.attempt
18/09/10 20:36:14 INFO deprecation: mapred.tasktracker.map.tasks.maximum is deprecated. Instead, use mapreduce.tasktracker.map.tasks.ma
18/09/10 20:36:14 INFO deprecation: mapred.map.child.log.level is deprecated. Instead, use mapreduce.map.log.level
18/09/10 20:36:14 INFO deprecation: mapred.local.dir.minspacestart is deprecated. Instead, use mapreduce.tasktracker.local.dir.minspaces
18/09/10 20:36:14 INFO deprecation: tasktracker.http.threads is deprecated. Instead, use mapreduce.tasktracker.http.threads
18/09/10 20:36:14 INFO deprecation: tasktracker.http.threads is deprecated. Instead, use mapreduce.tasktracker.http.threads
```

```

18/09/10 20:41:37 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_2018-09-
18/09/10 20:41:38 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 20:41:38 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_2018-09-
18/09/10 20:42:23 INFO Hive: New loading path = hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_20
18/09/10 20:43:04 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 20:43:05 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_2018-09-
18/09/10 20:43:05 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 20:43:05 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_2018-09-
18/09/10 20:43:07 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 20:43:09 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_2018-09-
18/09/10 20:43:09 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 20:43:09 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_2018-09-
18/09/10 20:43:09 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 20:43:09 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_2018-09-
18/09/10 20:43:09 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 20:43:10 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 20:43:11 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_2018-09-
18/09/10 20:43:11 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 20:43:11 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_2018-09-
18/09/10 20:43:11 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 20:43:12 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_2018-09-
18/09/10 20:43:12 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/10 20:43:12 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_2018-09-
18/09/10 20:43:35 INFO Hive: New loading path = hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hive_20
18/09/10 20:43:36 INFO SparkContext: Invoking stop() from shutdown hook
18/09/10 20:43:38 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/09/10 20:43:44 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/09/10 20:43:48 INFO MemoryStore: MemoryStore cleared
18/09/10 20:43:48 INFO BlockManager: BlockManager stopped
18/09/10 20:43:48 INFO BlockManagerMaster: BlockManagerMaster stopped
18/09/10 20:43:48 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/09/10 20:43:49 INFO SparkContext: Successfully stopped SparkContext
18/09/10 20:43:49 INFO ShutdownHookManager: Shutdown hook called
18/09/10 20:43:49 INFO ShutdownHookManager: Deleting directory /tmp/spark-19b3f0b9-a7dc-4aad-b472-dcfcecd9d9f1

Process finished with exit code 0

```

```

> show tables
> ;

```

OK

enriched_data

formatted_input

song_artist_map

station_geo_map

subscribed_users

users_artists

Time taken: 19.941 seconds, Fetched: 6 row(s)

hive> █


```

> select * from enriched_data;
OK
U120 S201 A301 1494297562 1465490556 1468094889 A ST410 3 0 1 1 fail
U102 S207 A303 1465230523 1485130523 1465230523 J ST403 3 1 1 1 fail
U114 S207 A303 1465130523 1465230523 1475130523 NULL ST415 3 1 0 1 fail
U106 S202 A302 1465230523 1465130523 1465130523 E ST408 0 1 1 1 fail
U107 S202 A302 1495130523 1465230523 1465230523 NULL ST415 0 1 1 1 fail
U103 S202 A302 1465490556 1465490556 1465490556 NULL ST415 2 1 1 1 fail
U111 S204 A304 1465490556 1465490556 1468094889 A ST410 3 1 1 1 fail
U100 S204 A304 1495130523 1475130523 1465130523 E ST408 2 1 1 1 fail
U113 S204 A304 1494297562 1494297562 1465490556 NULL ST415 3 0 1 1 fail
U111 S206 A302 1465130523 1465130523 1485130523 NULL ST415 0 1 1 1 fail
NULL S208 A304 1465490556 1494297562 1465490556 A ST411 1 0 1 1 fail
U101 S208 A304 1462863262 1468094889 1462863262 E ST408 0 1 1 1 fail
U119 S208 A304 1495130523 1465230523 1465230523 NULL ST415 3 0 0 1 fail
U118 S208 A304 1475130523 1465130523 1465230523 NULL ST415 3 0 0 1 fail
U107 S210 NULL 1475130523 1485130523 1485130523 E ST404 2 1 0 1 fail
U108 S205 A301 1465130523 1465130523 1475130523 A ST410 2 1 0 1 fail
U106 S205 A301 1462863262 1462863262 1494297562 AP ST407 2 1 1 1 fail
U100 S200 A300 1494297562 1494297562 1494297562 A ST410 3 1 1 1 fail
U113 S200 A300 1465230523 1475130523 1465130523 J ST413 3 1 1 1 fail
S203 A303 1495130523 1475130523 1465230523 A ST400 0 0 1 1 fail
U109 S203 A303 1462863262 1494297562 1468094889 A ST405 1 1 1 1 fail
U110 S203 A303 1465230523 1465130523 1485130523 NULL ST415 0 1 1 1 fail
U105 S207 A303 1465230523 1485130523 1465130523 A ST400 2 0 1 1 pass
U110 S202 A302 1494297562 1494297562 1468094889 AP ST402 2 1 0 1 pass
U104 S202 A302 1465230523 1475130523 1465130523 E ST409 2 0 1 1 pass
U107 S202 A302 1494297562 1468094889 1462863262 E ST409 0 0 0 1 pass
U118 S202 A302 1495130523 1465230523 1465230523 A ST410 1 0 0 1 pass
U101 S202 A302 1465230523 1465130523 1475130523 AU ST401 0 0 1 1 pass
U103 S204 A304 1468094889 1494297562 1465490556 A ST411 2 1 0 1 pass
U114 S209 A305 1465490556 1462863262 1494297562 A ST411 2 1 0 1 pass
U120 S206 A302 1495130523 1485130523 1465130523 E ST414 0 0 0 1 pass
U106 S206 A302 1494297562 1465490556 1462863262 A ST405 3 1 0 1 pass
U116 S208 A304 1465230523 1485130523 1475130523 J ST413 1 0 1 1 pass

```

Step 6: Perform Data Analysis

Run queries in enriched_data table to get results.

```

val create_top_10_stations = """CREATE TABLE IF NOT EXISTS top_10_stations(
  station_id STRING, total_distinct_songs_played INT, distinct_user_count INT )
  PARTITIONED BY (batchid INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE"""

val load_top_10_stations = s"""INSERT OVERWRITE TABLE top_10_stations PARTITION(batchid='$batchId')
  SELECT station_id, COUNT(DISTINCT song_id) AS total_distinct_songs_played, COUNT
  (DISTINCT user_id) AS distinct_user_count FROM enriched_data WHERE status='pass'
  AND batchid='$batchId' AND like=1 GROUP BY station_id ORDER BY total_distinct_songs_played
  DESC LIMIT 10"""

```

```

sparkSession.sqlContext.sql( sqlText = "SELECT station_id FROM top_10_stations").show()

```

18/09/10 21:14:58 INFO CatalystSqlParser: Parsing command: SELECT user_type,duration FROM users_behavior

```

+-----+
|station_id|
+-----+
|      ST402|
|      ST411|
|      ST405|
+-----+

```

18/09/10 21:14:58 INFO CatalystSqlParser: Parsing command: int
18/09/10 21:14:58 INFO CatalystSqlParser: Parsing command: string
18/09/10 21:14:58 INFO CatalystSqlParser: Parsing command: int

```

val create_users_behaviour = """CREATE TABLE IF NOT EXISTS users_behaviour( user_type STRING, duration INT )
PARTITIONED BY (batchid INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE"""

val load_users_behaviour = s"""INSERT OVERWRITE TABLE users_behaviour PARTITION(batchid='$batchId')
SELECT CASE WHEN (su.user_id IS NULL OR CAST(ed.timestamp AS DECIMAL(20,0)) > CAST(su.subscn_end_dt AS DECIMAL(20,0)))
THEN 'UNSUBSCRIBED' WHEN (su.user_id IS NOT NULL AND CAST(ed.timestamp AS DECIMAL(20,0)) <= CAST(su.subscn_end_dt
AS DECIMAL(20,0))) THEN 'SUBSCRIBED' END AS user_type, SUM(ABS(CAST(ed.end_ts AS DECIMAL(20,0))-CAST(ed.start_ts AS DECIMAL(20,0))
AS duration FROM enriched_data ed LEFT OUTER JOIN subscribed_users su ON ed.user_id=su.user_id WHERE ed.status='pass' AND
ed.batchid='$batchId' GROUP BY CASE WHEN (su.user_id IS NULL OR CAST(ed.timestamp AS DECIMAL(20,0)) > CAST(su.subscn_end_dt
AS DECIMAL(20,0))) THEN 'UNSUBSCRIBED' WHEN (su.user_id IS NOT NULL AND CAST(ed.timestamp AS DECIMAL(20,0)) <= CAST
(su.subscn_end_dt AS DECIMAL(20,0))) THEN 'SUBSCRIBED' END"""

```

```
sparkSession.sqlContext.sql( sqlText = "SELECT user_type,duration FROM users_behaviour").show()
```

```
18/09/10 21:15:15 INFO SparkSqlParser: Parsing command: SELECT artist_id FROM connected_artists
```

```

+-----+-----+
| user_type| duration|
+-----+-----+
|UNSUBSCRIBED| 98100227|
| SUBSCRIBED|157978279|
+-----+-----+

```

```

18/09/10 21:15:16 INFO CatalystSqlParser: Parsing command: int
18/09/10 21:15:16 INFO CatalystSqlParser: Parsing command: string
18/09/10 21:15:16 INFO CatalystSqlParser: Parsing command: int

```

```

val create_connected_artists = """CREATE TABLE IF NOT EXISTS connected_artists( artist_id STRING, user_count INT )
PARTITIONED BY (batchid INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE"""

val load_connected_artists = s"""INSERT OVERWRITE TABLE connected_artists PARTITION(batchid='$batchId')
SELECT ua.artist_id, COUNT(DISTINCT ua.user_id) AS user_count FROM ( SELECT user_id, artist_id FROM users_artists
LATERAL VIEW explode(artists_array) artists AS artist_id ) ua INNER JOIN ( SELECT artist_id, song_id, user_id FROM enriched_data
WHERE status='pass' AND batchid='$batchId' ) ed ON ua.artist_id=ed.artist_id AND ua.user_id=ed.user_id GROUP BY ua.artist_id
ORDER BY user_count DESC LIMIT 10"""

```

```
sparkSession.sqlContext.sql( sqlText = "SELECT artist_id FROM connected_artists").show()
```

```

18/09/10 21:15:20 INFO DAGScheduler: ResultStage 23 (show at DataAnalysis.scala:74) finished in 1.455 s
18/09/10 21:15:20 INFO DAGScheduler: Job 9 finished: show at DataAnalysis.scala:74, took 1.928186 s

```

```

+-----+
|artist_id|
+-----+
|      A302|
|      A300|
+-----+

```

```

18/09/10 21:15:20 INFO SparkSqlParser: Parsing command: SELECT song_id FROM top_10_royalty_songs
18/09/10 21:15:21 INFO CatalystSqlParser: Parsing command: int
18/09/10 21:15:21 INFO CatalystSqlParser: Parsing command: string

```

```

val create_top_10_royalty_songs = """CREATE TABLE IF NOT EXISTS top_10_royalty_songs( song_id STRING, duration INT )
PARTITIONED BY (batchid INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE"""

val load_top_10_royalty_songs = s"""INSERT OVERWRITE TABLE top_10_royalty_songs PARTITION(batchid='$batchId')
SELECT song_id, SUM(ABS(CAST(end_ts AS DECIMAL(20,0))-CAST(start_ts AS DECIMAL(20,0)))) AS duration
FROM enriched_data WHERE status='pass' AND batchid='$batchId' AND (like=1 OR song_end_type=0) GROUP BY song_id
ORDER BY duration DESC LIMIT 10"""

```

```
sparkSession.sqlContext.sql( sqlText = "SELECT song_id FROM top_10_royalty_songs").show()
```



```

18/09/10 21:15:24 INFO SparkSqlParser: Parsing command: SELECT user_id FROM top_10_unsubscribed_users
+-----+
|song_id|
+-----+
|  S202 |
|  S209 |
|  S204 |
|  S206 |
|  S200 |
|  S203 |
+-----+

18/09/10 21:15:25 INFO CatalystSqlParser: Parsing command: int
18/09/10 21:15:25 INFO CatalystSqlParser: Parsing command: string
18/09/10 21:15:25 INFO CatalystSqlParser: Parsing command: int

```

```

val create_top_10_unsubscribed_users = """CREATE TABLE IF NOT EXISTS top_10_unsubscribed_users( user_id STRING, duration INT )
PARTITIONED BY (batchid INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE"""

val load_top_10_unsubscribed_users = s"""INSERT OVERWRITE TABLE top_10_unsubscribed_users PARTITION(batchid='$batchId')
SELECT ed.user_id, SUM(ABS(CAST(ed.end_ts AS DECIMAL(20,0))-CAST(ed.start_ts AS DECIMAL(20,0)))) AS duration
FROM enriched_data ed LEFT OUTER JOIN subscribed_users su ON ed.user_id=su.user_id WHERE ed.status='pass'
AND ed.batchid='$batchId' AND (su.user_id IS NULL OR (CAST(ed.timestamp AS DECIMAL(20,0)) > CAST(su.subscn_end_dt AS DECIMAL(20,0)))
GROUP BY ed.user_id ORDER BY duration DESC LIMIT 10"""

```

```

sparkSession.sqlContext.sql( sqlText = "SELECT user_id FROM top_10_unsubscribed_users ").show()

```

```

18/09/10 21:15:26 INFO DAGScheduler: ResultStage 25 (show at DataAnalysis.scala:76) finished in 0.048 s
18/09/10 21:15:26 INFO DAGScheduler: Job 11 finished: show at DataAnalysis.scala:76, took 0.087741 s

```

```

+-----+
|user_id|
+-----+
|  U115 |
|  U110 |
|  U120 |
|  U116 |
|  U107 |
|  U108 |
|  U106 |
|  U118 |
+-----+

```

```

18/09/10 21:15:26 INFO SparkContext: Invoking stop() from shutdown hook

```