

~~ACADGILD FINAL PROJECT~~

MyTunes Music Analysis

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

Fields present in the data files

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

LookUp Tables

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 alongwith royalty associated with each play of the song
User_Artist_Map	Contains an array of artist_id(s) followed by a user_id

➤ **Formatting by python script :-**

```
[acadgild@localhost scripts]$ vi generate_web_data.py
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost scripts]$ python generate_web_data.py
[acadgild@localhost scripts]$ ls -ltr
total 64
      . . . . .
```

```
□ <records>
  □ <record>
    <user_id>U106</user_id>
    <song_id>S205</song_id>
    <artist_id>A300</artist_id>
    <timestamp>2016-05-10 12:24:22</timestamp>
    <start_ts>2016-05-10 12:24:22</start_ts>
    <end_ts>2017-05-09 08:09:22</end_ts>
    <geo_cd>AP</geo_cd>
    <station_id>ST407</station_id>
    <song_end_type>2</song_end_type>
    <like>1</like>
    <dislike>1</dislike>
  -</record>
  □ <record>
    <user_id>U114</user_id>
    <song_id>S209</song_id>
    <artist_id>A303</artist_id>
    <timestamp>2016-06-09 22:12:36</timestamp>
    <start_ts>2016-05-10 12:24:22</start_ts>
    <end_ts>2017-05-09 08:09:22</end_ts>
    <geo_cd>U</geo_cd>
    <station_id>ST411</station_id>
    <song_end_type>2</song_end_type>
    <like>1</like>
    <dislike>0</dislike>
```

- Started all Daemon service like hive and hbase services

```

acadgild@localhost:~/project/scripts
File Edit View Search Terminal Help
[acadgild@localhost scripts]$ python generate_mob_data.py
[acadgild@localhost scripts]$ chmod 755 start-daemons.sh
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost scripts]$ ls -ltr
total 64
-rw-rw-r--. 1 acadgild acadgild 592 Sep  8 22:09 create_schema.sql
-rw-rw-r--. 1 acadgild acadgild 961 Sep  8 22:09 data_analysis.sh
-rw-rw-r--. 1 acadgild acadgild 873 Sep  8 22:09 create_hive_hbase_lookup.hql
-rw-rw-r--. 1 acadgild acadgild 1499 Sep  8 22:09 data_enrichment.sh
-rw-rw-r--. 1 acadgild acadgild 299 Sep  8 22:09 data_enrichment_filtering_schema.sh
-rw-rw-r--. 1 acadgild acadgild 1731 Sep  8 22:09 data_export.sh
-rw-rw-r--. 1 acadgild acadgild 402 Sep  8 22:09 dataformatting.sh
-rw-rw-r--. 1 acadgild acadgild 120 Sep  8 22:09 file_generation_in_python.txt
-rw-rw-r--. 1 acadgild acadgild 604 Sep  8 22:09 formatted_hive_load.hql
-rw-rw-r--. 1 acadgild acadgild 853 Sep  8 22:09 generate_mob_data.txt
-rw-rw-r--. 1 acadgild acadgild 1333 Sep  8 22:09 populate_lookup.sh
-rw-rw-r--. 1 acadgild acadgild 337 Sep  8 22:09 user-artist.hql
-rwxr-xr-x. 1 acadgild acadgild 412 Sep  8 22:09 start-daemons.sh
-rw-rw-r--. 1 acadgild acadgild 485 Sep  8 22:09 wrapper.sh
-rwxrwxrwx. 1 acadgild acadgild 1803 Sep  8 23:01 generate_web_data.py
-rwxrwxrwx. 1 acadgild acadgild 1165 Sep  8 23:12 generate_mob_data.py
[acadgild@localhost scripts]$ ./start-daemons.sh
./start-daemons.sh: line 7: /home/acadgild/project/logs/current-batch.txt: No such file or directory
chmod: cannot access '/home/acadgild/project/logs/current-batch.txt': No such file or directory
./start-daemons.sh: line 11: /home/acadgild/project/logs/current-batch.txt: No such file or directory
./start-daemons.sh: line 14: /home/acadgild/project/logs/log_batch_cat: No such file or directory
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
18/09/08 23:13:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using t
Starting namenodes on [localhost]
localhost: namenode running as process 4606. Stop it first.
localhost: datanode running as process 4699. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 4884. Stop it first.
18/09/08 23:13:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using t

```

➤ All services up and running

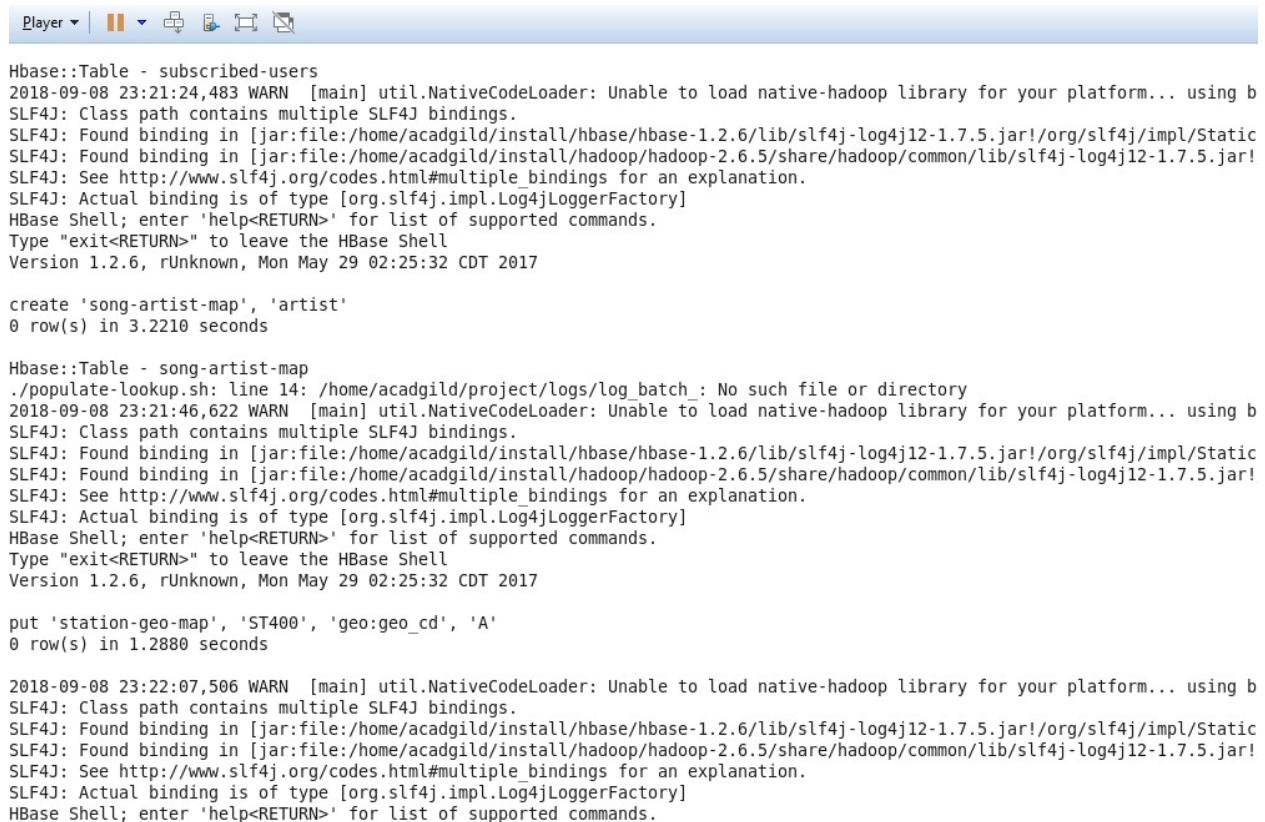
0.0.0.0: secondarynamenode running as process 4884. Stop it first.
18/09/08 23:13:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using bu
starting yarn daemons
resourcemanager running as process 5199. Stop it first.
localhost: nodemanager running as process 5299. Stop it first.
localhost: starting zookeeper, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-zookeeper
starting master, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-master-localhost.locald
starting regionserver, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-1-regionserver-lo
[acadgild@localhost scripts]\$ jps
8720 HRegionServer
5299 NodeManager
4884 SecondaryNameNode
8519 HQuorumPeer
4699 DataNode
8813 Jps
4606 NameNode
5199 ResourceManager
8607 HMaster
[acadgild@localhost scripts]\$ █

➤ Loading all lookup tables in Hbase

```
[acadgild@localhost scripts]$ chmod 755 populate-lookup.sh
[acadgild@localhost scripts]$ dos2unix populate-lookup.sh
dos2unix: converting file populate-lookup.sh to UNIX format ...
[acadgild@localhost scripts]$ ./populate-lookup.sh
cat /home/acadgild/project/logs/current-batch.txt: No such file or directory
./populate-lookup.sh: line 7: /home/acadgild/project/logs/log_batch_: No such file or directory
2018-09-08 23:20:41.804 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
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/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.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 3.4410 seconds

Hbase::Table - station-geo-map
2018-09-08 23:21:03.591 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
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/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.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
```



```
Player | ▶ ⏴ ⏵ ⏷ ⏸ ⏹ ⏺ ⏻

Hbase::Table - subscribed-users
2018-09-08 23:21:24,483 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
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/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.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 3.2210 seconds

Hbase::Table - song-artist-map
./populate-lookup.sh: line 14: /home/acadgild/project/logs/log_batch_: No such file or directory
2018-09-08 23:21:46,622 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
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/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.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 1.2880 seconds

2018-09-08 23:22:07,506 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
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/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.slf4j.impl.Log4jLoggerFactory]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
```

```

Player | 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 1.0490 seconds

2018-09-08 23:22:25,774 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerE
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/s
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', 'ST402', 'geo:geo_cd', 'AP'
0 row(s) in 1.1810 seconds

2018-09-08 23:22:45,791 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerE
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/s
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 1.2790 seconds

2018-09-08 23:23:06,468 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerE
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/s
[acadgild@localhost:~] acadgild project acadgild@localhost:~ acadgild@localhost:~/
```

```

put 'station-geo-map', 'ST404', 'geo:geo_cd', 'E'
0 row(s) in 1.2420 seconds

2018-09-08 23:23:27,040 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
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/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/Sta
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', 'ST405', 'geo:geo_cd', 'A'
0 row(s) in 1.3840 seconds

2018-09-08 23:23:47,224 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
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/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/Sta
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', 'ST406', 'geo:geo_cd', 'AU'
0 row(s) in 1.3970 seconds

2018-09-08 23:24:06,899 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
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/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/Sta
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
```

Activate ▾
Go to Setting

```
2018-09-08 23:24:06,899 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform  
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/  
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j1  
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', 'ST407', 'geo:geo_cd', 'AP'
0 row(s) in 1.1920 seconds
```

```
2018-09-08 23:24:26,968 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform  
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/  
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j1  
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', 'ST408', 'geo:geo_cd', 'E'
0 row(s) in 1.1120 seconds
```

```
2018-09-08 23:24:47,236 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform  
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/  
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j1  
SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
```

A screenshot of a terminal window titled "Acadgild_64bit_2.6 - VMware Workstation 12 Player (Non-commercial use only)". The window shows the following text:

```
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: 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
```

The terminal has tabs at the top: "[acadgild@localhost:~]", "acadgild", "project", "[acadgild@localhost:~]", and "[acadgild@localhost:~]".

```
put 'station-geo-map', 'ST408', 'geo:geo_cd', 'E'  
0 row(s) in 1.1120 seconds
```

```
2018-09-08 23:24:47,236 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built
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.StaticLogg
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: 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', 'ST409', 'geo:geo_cd', 'E'  
0 row(s) in 1.0140 seconds
```

```
2018-09-08 23:25:07,019 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built
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/StaticLogg
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: 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', 'ST410', 'geo:geo_cd', 'A'  
0 row(s) in 1.0580 seconds
```

[acadgild@localhost:~] acadgild project [acadgild@localhost:~] [acadgild@localhost:~/...]

```
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:startdt', '1465230523'
0 row(s) in 1.0170 seconds

2018-09-08 23:38:35,103 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your pl
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: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-l
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:enddt', '1468130523'
0 row(s) in 1.0870 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
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-l
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/hi
OK
Time taken: 17.03 seconds
OK
Time taken: 0.075 seconds
OK
Time taken: 2.555 seconds
Loading data to table project.users_artists
OK
Time taken: 3.812 seconds
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost scripts]$ acadgild
```



➤ **Creating hive tables on top of hbase tables for data enrichment and filtering...**

```
-rw-rw-r-- 1 acadgild acadgild 299 Sep  8 23:41 data_enrichment_filtering_schema.sh
[acadgild@localhost scripts]$ chmod 755 data_enrichment_filtering_schema.sh
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost scripts]$ ./data_enrichment_filtering_schema.sh
cat: /home/acadgild/project/logs/current-batch.txt: No such file or directory
./data_enrichment_filtering_schema.sh: line 6: /home/acadgild/project/logs/log_batch_: No such file or directory
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/StaticLoggerFactory.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/StaticLoggerFactory.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!
OK
Time taken: 13.698 seconds
OK
Time taken: 6.777 seconds
OK
Time taken: 0.695 seconds
OK
Time taken: 0.631 seconds
[acadgild@localhost scripts]$
```

acadgild@localhost:~/project/sci

[acadgild@localhost:~] [acadgild] [project] [acadgild@localhost:~] [acadgild@localhost:~/...]

➤ **Details of DB and Tables under Hive and it's Data**

```
File Edit View Search Terminal Help
[acadgild@localhost ~]$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hado
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-on-MR is deprecated in Hive 2 and may not be available in the future versions. Con
eases.
hive> show databases;
OK
acadgild
default
project
Time taken: 25.186 seconds, Fetched: 3 row(s)
hive> use project;
OK
Time taken: 0.088 seconds
hive> show tables;
OK
song_artist_map
station_geo_map
subscribed_users
users_artists
Time taken: 0.125 seconds, Fetched: 4 row(s)
hive> select * from song_artist_map limit 5;
OK
S200      A300
S201      A301
S202      A302
S203      A303
S204      A304
Time taken: 8.964 seconds, Fetched: 5 row(s)
```

```
hive> select * from station_geo_map limit 5;
OK
ST400    A
ST401    AU
ST402    AP
ST403    J
ST404    E
Time taken: 1.083 seconds, Fetched: 5 row(s)
hive> select * from subscribed_users limit 5;
OK
U100    1465230523    1465130523
U101    1465230523    1475130523
U102    1465230523    1475130523
U103    1465230523    1475130523
U104    1465230523    1475130523
Time taken: 0.749 seconds, Fetched: 5 row(s)
hive> select * from users_artists limit 5;
OK
U100    ["A300","A301","A302"]
U101    ["A301","A302"]
U102    ["A302"]
U103    ["A303","A301","A302"]
U104    ["A304","A301"]
Time taken: 0.605 seconds, Fetched: 5 row(s)
hive> █
```

➤ Loading csv and xml file to format table

Data Ingestion and Initial Validation

Rules for data ingestion and data filtering

1. Data coming from web applications reside in /data/web and has xml format.
2. Data coming from mobile applications reside in /data/mob and has csv format.
3. Data files come every 3 hours.
4. All the timestamp fields in data coming from web application is of the format YYYY-MM-DD HH:MM:SS.
5. All the timestamp fields in data coming from mobile application is a long integer interpreted as UNIX timestamps.
6. Finally, all timestamps must have the format of a long integer to be interpreted as UNIX timestamps.

```

build.sbt x fmthiveload.scala x
1 import org.apache.spark.sql.SparkSession
2
3
4 object fmthiveload {
5   def main(args: Array[String]): Unit = {
6     val sparkSession = SparkSession.builder()
7       .master("local")
8       .appName("spark")
9       .config("spark.sql.warehouse.dir", "user/hive/warehouse")
10      .config("hive.metastore.uris", "thrift://localhost:9083")
11      .enableHiveSupport()
12      .getOrCreate()
13
14     val listOfDbs = sparkSession.sqlContext.sql("show databases")
15     val xmlload= sparkSession.sqlContext.read.format("com.databricks.spark.xml")
16       .option("rowtag","record")
17       .load( path = "file:///home/acadgild/project/data/web/file.xml")
18     xmlload.createOrReplaceTempView("webfile")
19     xmlload.show()
20
21
22
23     listOfDbs.show(numRows = 4, truncate = false)
24
25
26     //sparkSession.sqlContext.sql("use project")
27     //sparkSession.sqlContext.sql("show tables").show()--insert into project.formatted_input
28     sparkSession.sqlContext.sql(sqlText = "LOAD DATA LOCAL INPATH '/home/acadgild/project/data/mob/file.txt' into table project.formatted_input")
29     sparkSession.sqlContext.sql(sqlText = " insert into project.formatted_input select user_id,song_id,artist_id,unix_timestamp(timestamp,'yyyy-mm-dd hh:mm:ss') as 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," +
30       "geo_cd,station_id,song_end_type,like,dislike from webpage ")
31   }
32 }
33

```

artist_id dislike	end_ts	geo_cd	like	song_end_type	song_id	start_ts	station_id	timestamp	user_id
A300	0 2017-05-09 08:09:22	U	0	3	S200 2016-06-09 22:12:36	ST400 2016-07-10 01:38:09	U110		
A305	0 2016-05-10 12:24:22	AP	1	2	S200 2016-06-09 22:12:36	ST415 2016-05-10 12:24:22	U118		
A300	1 2016-05-10 12:24:22	E	1	1	S202 2016-07-10 01:38:09	ST406 2016-07-10 01:38:09	U113		
A300	1 2016-07-10 01:38:09	E	1	1	S205 2017-05-09 08:09:22	ST411 2016-07-10 01:38:09	U101		
A304	1 2016-07-10 01:38:09	AP	1	1	S207 2016-06-09 22:12:36	ST414 2016-07-10 01:38:09	U103		
A301	1 2016-06-09 22:12:36	AU	1	1	S204 2016-06-09 22:12:36	ST411 2016-05-10 12:24:22	null		
A300	0 2016-07-10 01:38:09	E	1	1	S200 2016-06-09 22:12:36	ST415 2016-05-10 12:24:22	U115		
A303	0 2017-05-09 08:09:22	U	0	3	S210 2016-05-10 12:24:22	ST414 2017-05-09 08:09:22	U106		
A304	0 2016-06-09 22:12:36	null	0	2	S202 2017-05-09 08:09:22	ST404 2016-06-09 22:12:36	U117		
null	0 2016-06-09 22:12:36	E	0	1	S203 2017-05-09 08:09:22	ST405 2016-05-10 12:24:22	U108		
A304	1 2016-05-10 12:24:22	AP	0	2	S201 2017-05-09 08:09:22	ST405 2017-05-09 08:09:22	U115		
A301	0 2016-05-10 12:24:22	A	1	1	S202 2016-07-10 01:38:09	ST404 2017-05-09 08:09:22	U112		
A301	1 2016-06-09 22:12:36	AP	1	0	S201 2016-06-09 22:12:36	ST404 2016-06-09 22:12:36	U119		
A300	0 2016-05-10 12:24:22	AP	1	3	S210 2016-07-10 01:38:09	ST400 2016-05-10 12:24:22	U106		
A302	1 2016-07-10 01:38:09	U	0	3	S208 2016-06-09 22:12:36	ST407 2017-05-09 08:09:22	U116		
A303	1 2017-05-09 08:09:22	AP	1	2	S208 2016-05-10 12:24:22	ST413 2016-06-09 22:12:36	U106		
A304	0 2016-07-10 01:38:09	A	0	3	S200 2016-07-10 01:38:09	ST411 2016-06-09 22:12:36	U114		
A303	0 2017-05-09 08:09:22	U	1	0	S200 2016-07-10 01:38:09	ST400 2016-05-10 12:24:22	U114		
A302	1 2016-06-09 22:12:36	AU	0	2	S200 2016-05-10 12:24:22	ST411 2016-06-09 22:12:36	U120		
A303	1 2017-05-09 08:09:22	U	0	0	S204 2016-05-10 12:24:22	ST412 2016-06-09 22:12:36	U116		

➤ After enrichment loading into Enrich_load_data table

7. If both *like* and *dislike* are 1, consider that record to be **invalid**.
8. If any of the fields from User_id, Song_id, Timestamp, Start_ts, End_ts, Geo_cd is **NULL** or *absent*, consider that record to be invalid.
9. If Song_end_type is **NULL** or *absent*, treat it to be 3.

Create a temporary identifier for all the data files received in the last 3 hours (may be an integer batch_id which is auto incremented or a string obtained after combining current date and current hour, to keep track of valid and invalid records per batch).

Data Enrichment

Rules for data enrichment

1. If any of *likeordislike* is **NULL** or *absent*, consider it as 0.
 2. If fields like *Geo_cd* and *Artist_id* are **NULL** or *absent*, consult the lookup tables for fields *Station_id* and *Song_id* respectively to get the values of *Geo_cd* and *Artist_id*.
 3. If corresponding lookup entry is not found, consider that record to be invalid.

NULL or absent field	Look up field	Look up table (Table from which record can be updated)
Geo_cd	Station_id	Station_Geo_Map
Artist_id	Song_id	Song_Artist_Map

➤ Data loaded under formatted table

Logs of above script

```
fmthiveload x
18/09/13 11:44:44 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/09/13 11:44:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/09/13 11:44:44 INFO SQLHadoopMapReduceCommitProtocol: Using output committer class org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
18/09/13 11:44:45 INFO FileOutputCommitter: Saved output of task 'attempt_20180913114444_0004_m_000165_0' to hdfs://localhost:8020/user/hive/warehouse
18/09/13 11:44:45 INFO SparkHadoopMapRedUtil: attempt_20180913114444_0004_m_000165_0: Committed
18/09/13 11:44:45 INFO Executor: Finished task 165.0 in stage 4.0 (TID 401). 4662 bytes result sent to driver
18/09/13 11:44:45 INFO TaskSetManager: Starting task 177.0 in stage 4.0 (TID 402, localhost, executor driver, partition 177, ANY, 8082 bytes)
18/09/13 11:44:45 INFO TaskSetManager: Finished task 165.0 in stage 4.0 (TID 401) in 639 ms on localhost (executor driver) (198/200)
18/09/13 11:44:45 INFO Executor: Running task 177.0 in stage 4.0 (TID 402)
18/09/13 11:44:45 INFO ShuffleBlockFetcherIterator: Getting 3 non-empty blocks out of 200 blocks
18/09/13 11:44:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/09/13 11:44:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/09/13 11:44:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/09/13 11:44:45 INFO SQLHadoopMapReduceCommitProtocol: Using output committer class org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
18/09/13 11:44:45 INFO FileOutputCommitter: Saved output of task 'attempt_20180913114445_0004_m_000177_0' to hdfs://localhost:8020/user/hive/warehouse
18/09/13 11:44:45 INFO SparkHadoopMapRedUtil: attempt_20180913114445_0004_m_000177_0: Committed
18/09/13 11:44:45 INFO Executor: Finished task 177.0 in stage 4.0 (TID 402). 4602 bytes result sent to driver
18/09/13 11:44:45 INFO TaskSetManager: Starting task 199.0 in stage 4.0 (TID 403, localhost, executor driver, partition 199, ANY, 8082 bytes)
18/09/13 11:44:45 INFO TaskSetManager: Finished task 177.0 in stage 4.0 (TID 402) in 420 ms on localhost (executor driver) (199/200)
18/09/13 11:44:45 INFO Executor: Running task 199.0 in stage 4.0 (TID 403)
18/09/13 11:44:45 INFO ShuffleBlockFetcherIterator: Getting 6 non-empty blocks out of 200 blocks
18/09/13 11:44:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/09/13 11:44:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/09/13 11:44:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/09/13 11:44:45 INFO SQLHadoopMapReduceCommitProtocol: Using output committer class org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
18/09/13 11:44:46 INFO FileOutputCommitter: Saved output of task 'attempt_20180913114445_0004_m_000199_0' to hdfs://localhost:8020/user/hive/warehouse
18/09/13 11:44:46 INFO SparkHadoopMapRedUtil: attempt_20180913114445_0004_m_000199_0: Committed
18/09/13 11:44:46 INFO Executor: Finished task 199.0 in stage 4.0 (TID 403). 4602 bytes result sent to driver
18/09/13 11:44:46 INFO TaskSetManager: Finished task 199.0 in stage 4.0 (TID 403) in 712 ms on localhost (executor driver) (200/200)
18/09/13 11:44:46 INFO DAGScheduler: ResultStage 4 (sql at fmthiveload.scala:37) finished in 54.852 s
```

```
fmthive[local] 11:44:51 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/13 11:44:51 INFO Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/enrich_load_data/.hive-staging_hive_2018-09-13_11-41-47_415
18/09/13 11:44:51 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/13 11:44:51 INFO Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/enrich_load_data/.hive-staging_hive_2018-09-13_11-41-47_415
18/09/13 11:44:51 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/13 11:44:51 INFO Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/enrich_load_data/.hive-staging_hive_2018-09-13_11-41-47_415
18/09/13 11:44:51 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/13 11:44:52 INFO Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/enrich_load_data/.hive-staging_hive_2018-09-13_11-41-47_415
18/09/13 11:44:52 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/13 11:44:52 INFO Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/enrich_load_data/.hive-staging_hive_2018-09-13_11-41-47_415
18/09/13 11:44:52 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/13 11:44:52 INFO Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/enrich_load_data/.hive-staging_hive_2018-09-13_11-41-47_415
18/09/13 11:44:52 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/13 11:44:52 INFO Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/enrich_load_data/.hive-staging_hive_2018-09-13_11-41-47_415
18/09/13 11:44:52 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/13 11:44:52 INFO Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/enrich_load_data/.hive-staging_hive_2018-09-13_11-41-47_415
18/09/13 11:44:52 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/13 11:44:52 INFO SparkContext: Invoking stop() from shutdown hook
18/09/13 11:44:55 INFO SparkUI: Stopped Spark web UI at http://192.168.64.129:4040
18/09/13 11:44:56 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/09/13 11:44:56 INFO MemoryStore: MemoryStore cleared
18/09/13 11:44:56 INFO BlockManager: BlockManager stopped
18/09/13 11:44:56 INFO BlockManagerMaster: BlockManagerMaster stopped
18/09/13 11:44:57 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/09/13 11:44:57 INFO SparkContext: Successfully stopped SparkContext
18/09/13 11:44:57 INFO ShutdownHookManager: Shutdown hook called
18/09/13 11:44:57 INFO ShutdownHookManager: Deleting directory /tmp/spark-40b200d9-78bb-4d9c-9972-6630c207440f
```

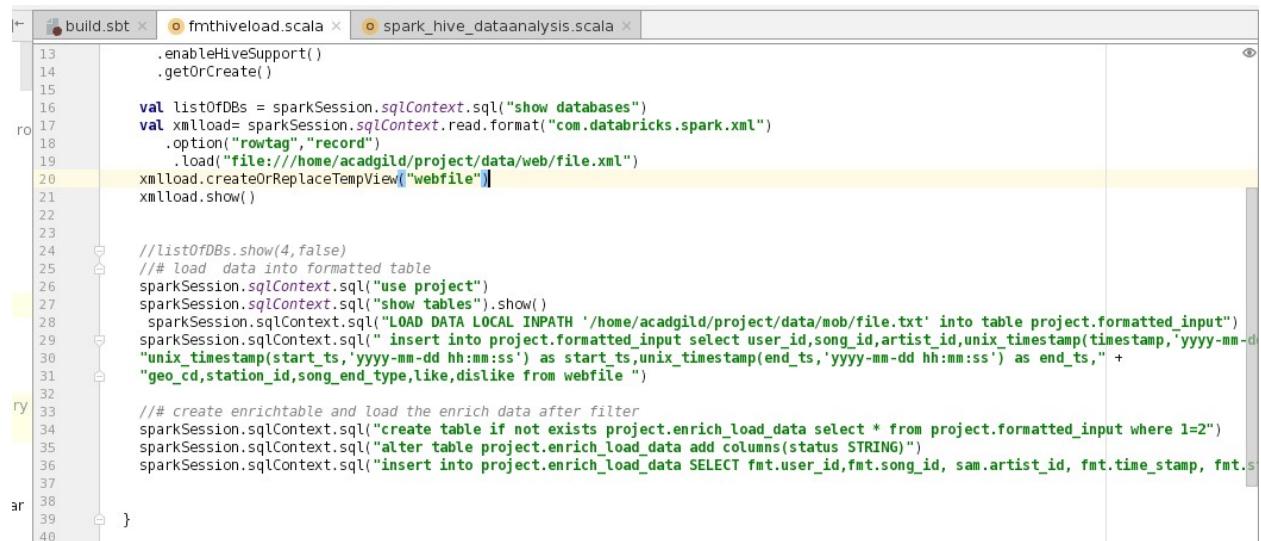
- Loaded data into formatted table in hive

```

hive> select * from formatted_input;
OK
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| user_id | song_id | artist_id | unix_timestamp | start_ts | end_ts | geo_cd | station_id | song_end_type | like | dislike |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| S210 | A305 | 1465230523 | 1485130523 | 1465230523 | AU | ST402 | 3 | 1 | 1 | |
| S210 | A305 | 1465130523 | 1485130523 | 1465230523 | E | ST414 | 0 | 0 | 1 |
| S202 | A304 | 1495130523 | 1475130523 | 1465230523 | A | ST406 | 2 | 0 | 1 |
| S203 | A300 | 1495130523 | 1485130523 | 1465130523 | U | ST411 | 1 | 0 | 1 |
| S201 | A303 | 1465130523 | 1475130523 | 1465230523 | AP | ST409 | 0 | 0 | 0 |
| S203 | A301 | 1465230523 | 1485130523 | 1475130523 | E | ST406 | 3 | 0 | 0 |
| S205 | A304 | 1475130523 | 1465130523 | 1475130523 | AP | ST412 | 1 | 0 | 1 |
| S210 | A304 | 1465230523 | 1475130523 | 1465130523 | A | ST401 | 1 | 0 | 1 |
| S209 | A302 | 1495130523 | 1485130523 | 1465230523 | ST404 | 3 | 0 | 0 | 0 |
| S208 | 1465130523 | 1465230523 | 1475130523 | AU | ST407 | 2 | 1 | 0 |
| S210 | A301 | 1495130523 | 1475130523 | 1465130523 | U | ST409 | 3 | 1 | 1 |
| S208 | A304 | 1495130523 | 1485130523 | 1465230523 | A | ST401 | 1 | 1 | 1 |
| S206 | A300 | 1465130523 | 1475130523 | 1465230523 | A | ST405 | 0 | 0 | 0 |
| S203 | A300 | 1475130523 | 1465230523 | 1465130523 | AU | ST412 | 1 | 1 | 0 |
| S203 | A305 | 1465230523 | 1465130523 | 1465130523 | E | ST401 | 1 | 0 | 0 |
| S203 | A302 | 1465130523 | 1465230523 | 1475130523 | A | ST403 | 2 | 1 | 0 |
| S207 | A300 | 1465130523 | 1485130523 | 1465230523 | U | ST403 | 1 | 0 | 1 |
| S207 | A305 | 1465130523 | 1465130523 | 1485130523 | A | ST409 | 3 | 0 | 1 |
| S205 | A302 | 1475130523 | 1475130523 | 1465230523 | AP | ST400 | 1 | 0 | 0 |
| S200 | A301 | 1465230523 | 1465230523 | 1465130523 | E | ST400 | 2 | 1 | 1 |
| S200 | A300 | 1468094889 | 1465490556 | 1494297562 | U | ST400 | 3 | 0 | 0 |
| S200 | A305 | 1462863262 | 1465490556 | 1462863262 | AP | ST415 | 2 | 1 | 0 |
| S202 | A300 | 1468094889 | 1468094889 | 1462863262 | E | ST406 | 1 | 1 | 1 |
| S205 | A300 | 1468094889 | 1494297562 | 1468094889 | E | ST411 | 1 | 1 | 1 |
| S207 | A304 | 1468094889 | 1465490556 | 1468094889 | AP | ST414 | 1 | 1 | 1 |
| NULL | S204 | A301 | 1462863262 | 1465490556 | 1465490556 | AU | ST411 | 1 | 1 | 1 |
| S200 | A300 | 1462863262 | 1465490556 | 1468094889 | E | ST415 | 1 | 1 | 0 |
| S210 | A303 | 1494297562 | 1462863262 | 1494297562 | U | ST414 | 3 | 0 | 0 |
| S202 | A304 | 1465490556 | 1494297562 | 1465490556 | NULL | ST404 | 2 | 0 | 0 |
| S203 | NULL | 1462863262 | 1494297562 | 1465490556 | E | ST405 | 1 | 0 | 0 |
| S201 | A304 | 1494297562 | 1494297562 | 1462863262 | AP | ST405 | 2 | 0 | 1 |
| S202 | A301 | 1494297562 | 1468094889 | 1462863262 | A | ST404 | 1 | 1 | 0 |
| S201 | A301 | 1465490556 | 1465490556 | 1465490556 | AP | ST404 | 0 | 1 | 1 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

Data loaded into enrich table



```

build.sbt x fmthiveload.scala x spark_hive_dataanalysis.scala x
13   .enableHiveSupport()
14   .getOrCreate()
15
16   val listOfDBs = sparkSession.sqlContext.sql("show databases")
17   val xmlload= sparkSession.sqlContext.read.format("com.databricks.spark.xml")
18   .option("rowtag","record")
19   .load("file:///home/acadgild/project/data/web/file.xml")
20   xmlload.createOrReplaceTempView("webfile")
21   xmlload.show()
22
23
24   //listOfDBs.show(4,false)
25   ## load data into formatted table
26   sparkSession.sqlContext.sql("use project")
27   sparkSession.sqlContext.sql("show tables").show()
28   sparkSession.sqlContext.sql("LOAD DATA LOCAL INPATH '/home/acadgild/project/data/mob/file.txt' into table project.formatted_input")
29   sparkSession.sqlContext.sql(" insert into project.formatted_input select user_id,song_id,artist_id,unix_timestamp(timestamp,'yyyy-mm-d
30   "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," +
31   "geo_cd,station_id,song_end_type,like,dislike from webfile ")
32
33   ## create enrichable and load the enrich data after filter
34   sparkSession.sqlContext.sql("create table if not exists project.enrich_load_data select * from project.formatted_input where 1=2")
35   sparkSession.sqlContext.sql("alter table project.enrich_load_data add columns(status STRING)")
36   sparkSession.sqlContext.sql("insert into project.enrich_load_data SELECT fmt.user_id,fmt.song_id, sam.artist_id, fmt.time_stamp, fmt.s
37
38
39
40
}

```

```

hive> select * from enrich_load_data;
OK
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| U119 | S201 | A301 | 1465490556 | 1465490556 | 1465490556 | E | ST404 | 0 | 1 | 1 | False |
| U115 | S201 | A301 | 1494297562 | 1494297562 | 1462863262 | A | ST405 | 2 | 0 | 1 | True  |
| U105 | S201 | A301 | 1465130523 | 1475130523 | 1465230523 | E | ST409 | 0 | 0 | 0 | True  |
| U103 | S207 | A303 | 1468094889 | 1465490556 | 1468094889 | E | ST414 | 1 | 1 | 1 | False |
| U113 | S207 | A303 | 1465130523 | 1485130523 | 1465230523 | J | ST403 | 1 | 0 | 1 | True  |
| U115 | S207 | A303 | 1465130523 | 1465130523 | 1485130523 | E | ST409 | 3 | 0 | 1 | True  |
| U117 | S202 | A302 | 1465490556 | 1494297562 | 1465490556 | E | ST404 | 2 | 0 | 0 | False |
| U112 | S202 | A302 | 1494297562 | 1468094889 | 1462863262 | E | ST404 | 1 | 1 | 0 | True  |
| U114 | S202 | A302 | 1495130523 | 1475130523 | 1465230523 | AU | ST406 | 2 | 0 | 1 | True  |
| U113 | S202 | A302 | 1468094889 | 1468094889 | 1462863262 | AU | ST406 | 1 | 1 | 1 | False |
| NULL | S204 | A304 | 1462863262 | 1465490556 | 1465490556 | A | ST411 | 1 | 1 | 1 | False |
| U116 | S204 | A304 | 1465490556 | 1462863262 | 1494297562 | AP | ST412 | 0 | 0 | 1 | True  |
| U119 | S209 | A305 | 1495130523 | 1485130523 | 1465230523 | E | ST404 | 3 | 0 | 0 | False |
| U103 | S206 | A302 | 1465130523 | 1475130523 | 1465230523 | A | ST405 | 0 | 0 | 0 | True  |
| U104 | S208 | A304 | 1495130523 | 1485130523 | 1465230523 | AU | ST401 | 1 | 1 | 1 | False |
| U105 | S208 | A304 | 1465130523 | 1465230523 | 1475130523 | AP | ST407 | 2 | 1 | 0 | True  |
| U116 | S208 | A304 | 1494297562 | 1465490556 | 1468094889 | AP | ST407 | 3 | 0 | 1 | True  |
| U106 | S208 | A304 | 1465490556 | 1462863262 | 1494297562 | J | ST413 | 2 | 1 | 1 | False |
| U103 | S210 | NULL | 1465230523 | 1485130523 | 1465230523 | AP | ST402 | 3 | 1 | 1 | False |
| U106 | S210 | NULL | 1462863262 | 1468094889 | 1462863262 | A | ST400 | 3 | 1 | 0 | False |
| U103 | S210 | NULL | 1465130523 | 1485130523 | 1465230523 | E | ST414 | 0 | 0 | 1 | False |
| U106 | S210 | NULL | 1494297562 | 1462863262 | 1494297562 | E | ST414 | 3 | 0 | 0 | False |
| U101 | S210 | NULL | 1495130523 | 1475130523 | 1465130523 | E | ST409 | 3 | 1 | 1 | False |
| U100 | S210 | NULL | 1465230523 | 1475130523 | 1465130523 | AU | ST401 | 1 | 0 | 1 | False |
| U101 | S205 | A301 | 1475130523 | 1475130523 | 1465230523 | A | ST400 | 1 | 0 | 0 | True  |
| U101 | S205 | A301 | 1468094889 | 1494297562 | 1468094889 | A | ST411 | 1 | 1 | 1 | False |
| U108 | S205 | A301 | 1475130523 | 1465130523 | 1475130523 | AP | ST412 | 1 | 0 | 1 | True  |
| U104 | S200 | A300 | 1465230523 | 1465230523 | 1465130523 | A | ST400 | 2 | 1 | 1 | False |
| U110 | S200 | A300 | 1468094889 | 1465490556 | 1494297562 | A | ST400 | 3 | 0 | 0 | True  |
| U114 | S200 | A300 | 1462863262 | 1468094889 | 1494297562 | A | ST400 | 0 | 1 | 0 | True  |
| U114 | S200 | A300 | 1465490556 | 1468094889 | 1468094889 | A | ST411 | 3 | 0 | 0 | True  |
| U120 | S200 | A300 | 1465490556 | 1462863262 | 1465490556 | A | ST412 | 2 | 0 | 1 | True  |
| U118 | S200 | A300 | 1462863262 | 1465490556 | 1462863262 | NULL | ST415 | 2 | 1 | 0 | False |
| U115 | S200 | A300 | 1462863262 | 1465490556 | 1468094889 | NULL | ST415 | 1 | 1 | 0 | False |

```

➤ DataAnalysis Queries

```

project ~/IdeaProjects/project
  +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
  | 9   | .config("hive.metastore.uris","thrift://localhost:9083") |
  | 10  | .enableHiveSupport() |
  | 11  | .getOrCreate() |
  | 12  | |
  | 13  | sparkSession.sql(sqlText = "use project") |
  | 14  | sparkSession.sql(sqlText = "INSERT INTO TABLE project.top_10_stations SELECT station_id, count(DISTINCT song_id) AS total_distinct_songs FROM project.enrich_load_data") |
  | 15  | //sparkSession.sql("select * from project.enrich_load_data").show(40)//INSERT OVERWRITE TABLE top_10_stations |
  | 16  | sparkSession.sql(sqlText = "SELECT CASE WHEN (su.user_id IS NULL OR CAST(eid.timestamp AS DECIMAL(20,0)) > CAST(su.subscr_end_dt AS DECIMAL(20,0))) THEN 0 ELSE 1 END AS duration FROM project.enrich_load_data su JOIN project.top_10_stations eid ON su.station_id = eid.station_id") |
  | 17  | sparkSession.sql(sqlText = "SELECT ua.artist_id, COUNT(DISTINCT ua.user_id) AS user_count FROM ( SELECT user_id, artist_id FROM project.enrich_load_data ua ) ua GROUP BY ua.artist_id") |
  | 18  | sparkSession.sql(sqlText = "SELECT song_id, SUM(ABS(CAST(end_ts AS DECIMAL(20,0))-CAST(start_ts AS DECIMAL(20,0)))) AS duration FROM project.enrich_load_data WHERE start_ts < end_ts") |
  | 19  | sparkSession.sql(sqlText = "SELECT eld.user_id, SUM(ABS(CAST(eld.end_ts AS DECIMAL(20,0))-CAST(eld.start_ts AS DECIMAL(20,0)))) AS duration FROM project.enrich_load_data eld WHERE eld.end_ts > eld.start_ts") |
  | 20  | } |
  +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
  spark_hive_dataanalysis > main(args: Array[String])

```

tun: spark_hive_dataanalysis

```

18/09/13 13:16:48 INFO DAGScheduler: ResultStage 3 (sql at spark_hive_dataanalysis.scala:14) finished in 23.021 s
18/09/13 13:16:48 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
18/09/13 13:16:49 INFO DAGScheduler: Job 0 finished: sql at spark_hive_dataanalysis.scala:14, took 140.268757 s
18/09/13 13:16:49 INFO FileFormatWriter: Job null committed.
18/09/13 13:16:49 INFO FileFormatWriter: Finished processing stats for job null.
18/09/13 13:16:56 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/13 13:16:56 INFO Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/top_10_stations/.hive-staging_hive_2018-09-13_13-10_304_74739900348277160-1--18/09/13 13:18:45 INFO BlockManagerInfo: Removed broadcast_4_piece0 on 192.168.64.129:38301 in memory (size: 77.7 KB, free: 349.1 MB)
18/09/13 13:18:50 INFO SparkContext: Invoking stop() from shutdown hook
18/09/13 13:19:11 INFO SparkUI: Stopped Spark web UI at http://192.168.64.129:4040
18/09/13 13:19:27 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/09/13 13:19:32 INFO MemoryStore: MemoryStore cleared
18/09/13 13:19:32 INFO BlockManager: BlockManager stopped
18/09/13 13:19:32 INFO BlockManagerMaster: BlockManagerMaster stopped
18/09/13 13:19:33 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/09/13 13:19:33 INFO SparkContext: Successfully stopped SparkContext
18/09/13 13:19:33 INFO ShutdownHookManager: Shutdown hook called
18/09/13 13:19:34 INFO ShutdownHookManager: Deleting directory /tmp/spark-f62783b5-c8da-4564-acbf-71a3c5c9515f

```

Process finished with exit code 0

1. Determine top 10 station_id(s) where maximum number of songs were played, which were liked by unique users.

```
hive> select * from top_10_stations;
OK
ST400    1      1
ST404    1      1
ST403    1      1
ST407    1      1
ST412    1      1
Time taken: 21.057 seconds, Fetched: 5 row(s)
hive> █
```

2. Determine total duration of songs played by each type of user, where type of user can be '**subscribed**' or '**unsubscribed**'. An unsubscribed user is the one whose record is either not present in **Subscribed_users** lookup table or has *subscription_end_date* earlier than the *timestamp* of the song played by him.

```

18/09/13 13:36:29 INFO TaskSetManager: Finished task 74.0 in stag
18/09/13 13:36:29 INFO TaskSchedulerImpl: Removed TaskSet 19.0, w
18/09/13 13:36:29 INFO DAGScheduler: ResultStage 19 (show at spar
18/09/13 13:36:29 INFO DAGScheduler: Job 4 finished: show at spar
+-----+
| user_type| duration|
+-----+
|UNSUBSCRIBED|133331854|
| SUBSCRIBED|163216685|
+-----+
18/09/13 13:36:39 INFO CodeGenerator: Code generated in 172.28553
18/09/13 13:36:40 INFO CodeGenerator: Code generated in 193.59978
18/09/13 13:36:40 INFO CodeGenerator: Code generated in 163.11504
18/09/13 13:36:40 INFO CodeGenerator: Code generated in 314.50033

```

1. Determine top 10 connected artists. Connected artists are those whose songs are most listened by the unique users who follow them.

```

18/09/13 13:37:39 INFO DAGScheduler: ResultStage 24 (show at spark_hive_dataanalysis.scala:17) finish
18/09/13 13:37:39 INFO DAGScheduler: Job 6 finished: show at spark_hive_dataanalysis.scala:17, took 4
+-----+
|artist_id|user_count|
+-----+
|    A301|      2|
|    A302|      2|
|    A303|      2|
|    A300|      1|
+-----+
18/09/13 13:37:42 INFO CodeGenerator: Code generated in 247.644891 ms
18/09/13 13:37:43 INFO CodeGenerator: Code generated in 338.758282 ms
18/09/13 13:37:43 INFO CodeGenerator: Code generated in 347.047082 ms
18/09/13 13:37:43 INFO HashAggregateExec: spark.sql.codegen.aggregate.map.twolevel.enabled is set to

```

1. Determine top 10 songs who have generated the maximum revenue. Royalty applies to a song only if it was *liked* or was *completed successfully* or both.

```

18/09/13 13:38:10 INFO TaskSchedulerImpl: Removed TaskSet 27.0, whose
18/09/13 13:38:10 INFO DAGScheduler: ResultStage 27 (show at spark_hiv
18/09/13 13:38:10 INFO DAGScheduler: Job 7 finished: show at spark_hiv
+-----+-----+
|song_id|duration|
+-----+-----+
| S204|31434300|
| S200|26202673|
| S203|199000000|
| S201| 99000000|
| S206| 99000000|
| S208| 99000000|
| S202| 5231627|
+-----+-----+
18/09/13 13:38:17 INFO CodeGenerator: Code generated in 126.485646 ms
18/09/13 13:38:17 INFO HashAggregateExec: spark.sql.codegen.aggregate.
18/09/13 13:38:17 INFO CodeGenerator: Code generated in 232.070772 ms

```

1. Determine top 10 unsubscribed users who listened to the songs for the longest duration.

```

18/09/13 13:38:56 INFO TaskSchedulerImpl: Removed TaskSet 31.0, whose tasks have all
18/09/13 13:38:56 INFO DAGScheduler: ResultStage 31 (show at spark_hive_dataanalysis.
18/09/13 13:38:56 INFO DAGScheduler: Job 8 finished: show at spark_hive_dataanalysis.
+-----+-----+
|user_id|duration|
+-----+-----+
| U115|51434300|
| U116|34038633|
| U114|299000000|
| U108|100000000|
| U112| 5231627|
| U120| 2627294|
| U117| 100000|
+-----+-----+
18/09/13 13:38:58 INFO SparkContext: Invoking stop() from shutdown hook
18/09/13 13:38:59 INFO SparkUI: Stopped Spark web UI at http://192.168.64.129:4040
18/09/13 13:38:59 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint
18/09/13 13:39:00 INFO MemoryStore: MemoryStore cleared

```

The screenshot shows a terminal window with two tabs: "build.sbt" and "fmthiveload.scala". The "fmthiveload.scala" tab is active, displaying the following Scala code:

```
val databaseName = "acadgilddb"
val default = "default"
val project = "project"
```

Below the code, the terminal output is shown, starting with informational messages from the "CodeGenerator" and "ContextCleaner" classes.

```
18/09/12 01:33:05 INFO CodeGenerator: Code generated in 213.279187 ms
18/09/12 01:33:05 INFO CodeGenerator: Code generated in 141.50823 ms
+-----+
|databaseName|
+-----+
|acadgilddb |
|default      |
|project      |
+-----+
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 28
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 44
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 25
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 48
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 31
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 40
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 33
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 51
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 49
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 47
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 26
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 29
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 42
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 30
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 36
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 27
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 41
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 34
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 32
18/09/12 01:33:12 INFO BlockManagerInfo: Removed broadcast_3_piece0 on 192.168.
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 52
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 45
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 37
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 46
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 50
18/09/12 01:33:12 INFO ContextCleaner: Cleaned accumulator 38
```

```

fmthiveload <
18/09/12 01:34:10 INFO TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
18/09/12 01:34:10 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, localhost, executor driver, partition 0, PROCESS_LOCAL, 7934 bytes
18/09/12 01:34:10 INFO Executor: Running task 0.0 in stage 2.0 (TID 2)
18/09/12 01:34:10 INFO NewHadoopRDD: Input split: file:/home/acadgild/project/data/web/file.xml:0+6719
18/09/12 01:34:10 INFO SQLHadoopMapReduceCommitProtocol: Using output committer class org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
18/09/12 01:34:15 INFO FileOutputCommitter: Saved output of task 'attempt_20180912013410_0002_m_000000_0' to hdfs://localhost:8020/user/hive/war
18/09/12 01:34:15 INFO SparkHadoopMapRedUtil: attempt_20180912013410_0002_m_000000_0: Committed
18/09/12 01:34:15 INFO Executor: Finished task 0.0 in stage 2.0 (TID 2). 2219 bytes result sent to driver
18/09/12 01:34:15 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 4798 ms on localhost (executor driver) (1/1)
18/09/12 01:34:15 INFO DAGScheduler: ResultStage 2 (sql at fmthiveload.scala:28) finished in 5.433 s
18/09/12 01:34:15 INFO DAGScheduler: Job 2 finished: sql at fmthiveload.scala:28, took 5.464132 s
18/09/12 01:34:15 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
18/09/12 01:34:15 INFO FileFormatWriter: Job null committed.
18/09/12 01:34:15 INFO FileFormatWriter: Finished processing stats for job null.
18/09/12 01:34:16 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/12 01:34:16 INFO Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/formatted_input/.hive-staging_hive_2018-09-12_01
18/09/12 01:34:20 INFO CodeGenerator: Code generated in 57.828141 ms
++
||
++
++

18/09/12 01:34:20 INFO SparkContext: Invoking stop() from shutdown hook
18/09/12 01:34:20 INFO SparkUI: Stopped Spark web UI at http://192.168.0.5:4040
18/09/12 01:34:21 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/09/12 01:34:21 INFO MemoryStore: MemoryStore cleared
18/09/12 01:34:21 INFO BlockManager: BlockManager stopped
18/09/12 01:34:21 INFO BlockManagerMaster: BlockManagerMaster stopped
18/09/12 01:34:21 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/09/12 01:34:22 INFO SparkContext: Successfully stopped SparkContext
18/09/12 01:34:22 INFO ShutdownHookManager: Shutdown hook called
18/09/12 01:34:22 INFO ShutdownHookManager: Deleting directory /tmp/spark-c0990380-3a56-4414-a9e6-0b33aelba2e6

Process finished with exit code 0

```

Act

Go to

```

[acadgild@localhost mob]$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/l
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hai
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-log4j2.properties
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using Tez or Spark.
hive> show databases;
OK
acadgild
default
project
Time taken: 43.489 seconds, Fetched: 3 row(s)
hive> use project
> ;
OK
Time taken: 0.187 seconds
hive> show tables;
OK
formatted_input
song_artist_map
station_geo_map
subscribed_users
users_artists
Time taken: 0.32 seconds, Fetched: 5 row(s)

```

```

hive> select * from formatted_input;
OK
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| U103 | S210 | A305 | 1465230523 | 1485130523 | 1465230523 | AU | ST402 | 3 | 1 |
| U103 | S210 | A305 | 1465130523 | 1485130523 | 1465230523 | E | ST414 | 0 | 0 |
| U114 | S202 | A304 | 1495130523 | 1475130523 | 1465230523 | A | ST406 | 2 | 0 |
| U114 | S203 | A300 | 1495130523 | 1485130523 | 1465130523 | U | ST411 | 1 | 0 |
| U105 | S201 | A303 | 1465130523 | 1475130523 | 1465230523 | AP | ST409 | 0 | 0 |
|       | S203 | A301 | 1465230523 | 1485130523 | 1475130523 | E | ST406 | 3 | 0 |
| U108 | S205 | A304 | 1475130523 | 1465130523 | 1475130523 | AP | ST412 | 1 | 0 |
| U100 | S210 | A304 | 1465230523 | 1475130523 | 1465130523 | A | ST401 | 1 | 0 |
| U119 | S209 | A302 | 1495130523 | 1485130523 | 1465230523 |       | ST404 | 3 | 0 |
| U105 | S208 |       | 1465130523 | 1465230523 | 1475130523 | AU | ST407 | 2 | 1 |
| U101 | S210 | A301 | 1495130523 | 1475130523 | 1465130523 | U | ST409 | 3 | 1 |
| U104 | S208 | A304 | 1495130523 | 1485130523 | 1465230523 | A | ST401 | 1 | 1 |
| U103 | S206 | A300 | 1465130523 | 1475130523 | 1465230523 | A | ST405 | 0 | 0 |
| U104 | S203 | A300 | 1475130523 | 1475130523 | 1465130523 | AU | ST412 | 1 | 0 |
| U117 | S203 | A305 | 1465230523 | 1465230523 | 1465130523 | E | ST401 | 1 | 0 |
| U111 | S203 | A302 | 1465130523 | 1465230523 | 1475130523 | A | ST403 | 2 | 1 |
| U113 | S207 | A300 | 1465130523 | 1485130523 | 1465230523 | U | ST403 | 1 | 0 |
| U115 | S207 | A305 | 1465130523 | 1465130523 | 1485130523 | A | ST409 | 3 | 0 |
| U101 | S205 | A302 | 1475130523 | 1475130523 | 1465230523 | AP | ST400 | 1 | 0 |
| U104 | S200 | A301 | 1465230523 | 1465230523 | 1465130523 | E | ST400 | 2 | 1 |
| U110 | S200 | A300 | 1468094889 | 1465490556 | 1494297562 | U | ST400 | 3 | 0 |
| U118 | S200 | A305 | 1462863262 | 1465490556 | 1462863262 | AP | ST415 | 2 | 1 |
| U113 | S202 | A300 | 1468094889 | 1468094889 | 1462863262 | E | ST406 | 1 | 1 |
| U101 | S205 | A300 | 1468094889 | 1494297562 | 1468094889 | E | ST411 | 1 | 1 |
| U103 | S207 | A304 | 1468094889 | 1465490556 | 1468094889 | AP | ST414 | 1 | 1 |
| NULL | S204 | A301 | 1462863262 | 1465490556 | 1465490556 | AU | ST411 | 1 | 1 |
| U115 | S200 | A300 | 1462863262 | 1465490556 | 1468094889 | E | ST415 | 1 | 0 |
| U106 | S210 | A303 | 1494297562 | 1462863262 | 1494297562 | U | ST414 | 3 | 0 |
| U117 | S202 | A304 | 1465490556 | 1494297562 | 1465490556 | NULL | ST404 | 2 | 0 |
| U108 | S203 | NULL | 1462863262 | 1494297562 | 1465490556 | E | ST405 | 1 | 0 |
| U115 | S201 | A304 | 1494297562 | 1494297562 | 1462863262 | AP | ST405 | 2 | 0 |
| U112 | S202 | A301 | 1494297562 | 1468094889 | 1462863262 | A | ST404 | 1 | 1 |
| U119 | S201 | A301 | 1465490556 | 1465490556 | 1465490556 | AP | ST404 | 0 | 1 |
| U106 | S210 | A300 | 1462863262 | 1468094889 | 1462863262 | AP | ST400 | 3 | 1 |
| U116 | S208 | A302 | 1494297562 | 1465490556 | 1468094889 | U | ST407 | 3 | 0 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

Create table with only structure

```

18/09/12 02:11:00 INFO FileFormatWriter: Finished processing stats for job null.
18/09/12 02:11:03 INFO SessionState: Could not get hdfsEncryptionShim, it is only applicable to hdfs filesystem.
18/09/12 02:11:03 INFO Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enrich_load_data/.hi
18/09/12 02:11:15 INFO CodeGenerator: Code generated in 335.578202 ms
18/09/12 02:11:16 INFO CodeGenerator: Code generated in 329.105987 ms
+-----+-----+-----+
| database|   tableName|isTemporary|
+-----+-----+-----+
| project|enrich_load_data|    false|
| project|formatted_input|    false|
| project|song_artist_map|    false|
| project|station_geo_map|    false|
| project|subscribed_users|    false|
| project|users_artists|    false|
+-----+-----+-----+
18/09/12 02:11:16 INFO SparkContext: Invoking stop() from shutdown hook
18/09/12 02:11:17 INFO SparkUI: Stopped Spark web UI at http://192.168.0.5:4040
18/09/12 02:11:17 TINFO ManOutputTrackerMasterEndpoint: ManOutputTrackerMasterEndpoint stopped!

```