MUSIC DATA ANALYSIS

Using Spark, Hive, Hbase, Scoop & Hadoop

ABSTRACT

This project is on any song applications which we have seen in the market like gaana.com, wink, apple music and so on. We would fetch the data from this application to perform analysis like to know how many songs are heard by user and liked by the user, how many songs were completely listened by the user and how many were terminated in between and so on.

R Gracy Clara

Big data Hadoop and Spark Engineering

BIG DATA HADOOP AND SPARK DEVLOPMENT

MUSIC DATA ANALYSIS

PROJECT

Contents

Music Data Analysis 3

Project Description: 3

1.Data Files: 3

2.Look up Tables 4

3. Data Ingestion & Filtering 4

4. Data Enrichment: 5

5. Data Analysis 5

Project flow: 6

Explanation of the code: 7

Project execution 30

Output: 41

Exporting data from Hive to RDBMS(MySQL) using Sqoop 46

Steps to export the data from Hive to RDMBS(MySQL) 48

Solution to resolve the errors faced during project execution 58

MUSIC DATA ANALYSIS

Project Description:

This project is on any song applications which we have seen in the market like gaana.com, wink, apple music and so on. We would fetch the data from this application to perform analysis like to know how many songs are heard by user and liked by the user, how many songs were completely listened by the user and how many were terminated in between and so on.

To perform analysis on the data, let's see the required tables: -

1.Data Files:

These are fields that we would fetch from the application

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	2 means song was paused How the song was terminated? 0 means completed successfully 1 means song was skipped 3 means other type of failure like device issue, network error etc.
Like	0 means song was not liked, but song was played 1 means song was liked
Dislike	0 means song was not disliked 1 means song was disliked

2.Look up Tables

These tables are required to fill null or absent values in the data field, which is fetched from the above mentioned data files. For example, say from the above data extracted for some fields either Geo_cd or Station_id is Null or absent, then in such case we would look for the data in lookup tables to fetch the respective information in Station_Geo_Map lookup table which contains information on both geo_cd for respective station_id and vice versa, so when either of the information is not available, we can fetch the required data from the respective look up table. These look up tables are stored in HBase.

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

3. Data Ingestion & Filtering

- We are extracting from two sources namely:
  - From Web
  - From Mobile application
- Timestamps of the data from web application is YYYY-MM-DD HH:MM:SS.
- Timestamps of the data from mobile application is long integer, which is interpreted as UNIX timestamps.
- Both the timestamps are long integer interpreted as UNIX timestamps.
- 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

Rules for invalid records:

Columns	**Value of the data which is treated as invalid**
Like & Dislike	1
User_id, Song_id, Timestamp, Start_ts, End_ts, Geo_cd	Null or Absent
Song_end_type	Null or Absent

4. Data Enrichment:

- We would consider the below mentioned rules for enriching the data.
  - i.e. if both like and dislike is null or absent, we would assume the value as 0 & if both geo_cd and artist_id is null or absent, then we would consult the look tables to update the missing value, if the entry is not available in look up table as well, then we would consider it as invalid entry.
- After enrichment, we will have all valid and invalid records.
- We will move all valid records to processing_dir directory.

Data Enrichment rules:

Columns	Value of the data	Action to be taken
Like & Dislike	Null or Absent	Consider the value as 0
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. If corresponding lookup entry is not found, consider that record to be invalid

5. Data Analysis

It is not only the data which is important, rather it is the insight it can be used to generate important. Once we have made the data ready for analysis, we have to perform below analysis on a daily basis.

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

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.

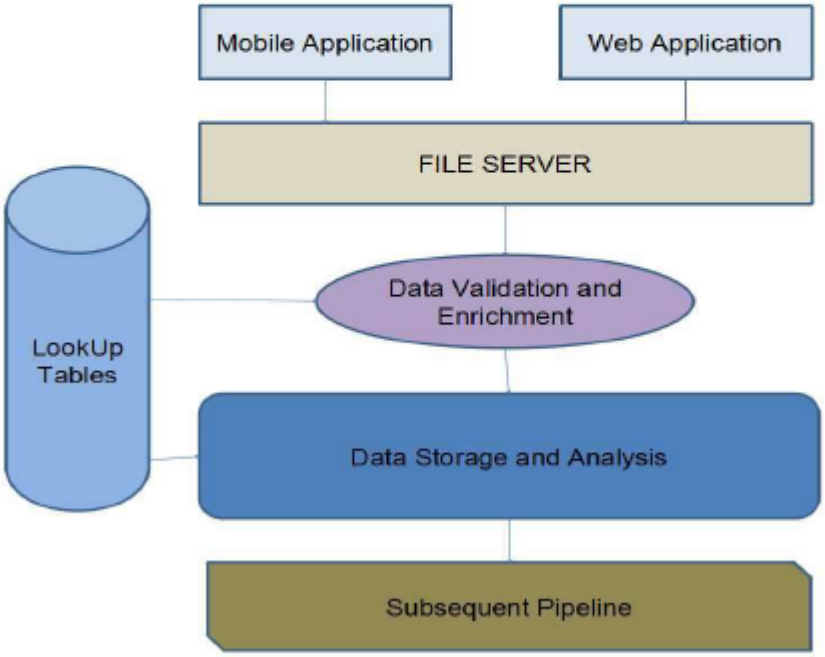
3. Determine top 10 connected artists. Connected artists are those whose songs are most listened by the unique users who follow them.

4. 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.

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

Project flow:

- Data is flowing from two sources I.e., from mobile application and web application.
- Data is flowing to file server.
- We apply data validation and enrichment (where we use lookup tables to extract missing data, for data enrichment and analysis.)
- Then we store data and perform analysis.
- We move the data to subsequent pipelines.



Explanation of the code:

1. Master script

```
# Script to Create data
echo "Preparing to execute python scripts to generate data..."
#if we already have the below directories we delete them.
rm -r /home/acadgild/examples/music/data/web
rm -r /home/acadgild/examples/music/data/mob
# we create the below directories for mobile and web application data
mkdir -p /home/acadgild/examples/music/data/web
mkdir -p /home/acadgild/examples/music/data/mob
#we execute the python scripts to create mobile and web application data
python /home/acadgild/examples/music/generate_web_data.py
python /home/acadgild/examples/music/generate_mob_data.py
echo "Data Generated Successfully !"
# Call Stop start daemon scripts to start hadoop daemons
echo "Starting the daemons...."
sh start-daemons.sh
# run jps commands to check the daemons
jps
echo "All hadoop daemons started !"
echo "Upload the look up tables now in Hbase..."
sh populate-lookup.sh
echo "Done with data population in look up tables !"
echo "Lets do some data formatting now...."
sh dataformatting.sh
echo "data formatting complete !"
echo "Creating hive tables on top of hbase tables for data enrichment and filtering..."
sh data_enrichment_filtering_schema.sh
echo "Hive table with Hbase Mapping Complete !"
echo "Let us do data enrichment as per the requirement..."
sh data_enrichment.sh
echo "Data Enrichment Complete"
echo "Lets run some use cases now..."
sh data_analysis.sh
echo "USE CASES COMPLETE !!"
```

2. generate_mob_data.py

```
from random import randint
from random import choice
#open the file called "file.txt" from the mentioned directory
file = open("/home/acadgild/examples/music/data/mob/file.txt", "w")
count = 20 #initialize the variable count as 20
#while loop till count goes less then 0, till then it will get the geo_cd list and other lists
while (count > 0):
        geo_cd_list=["A", "E", "AU", "AP", "U"]
        song_end_type_list=["0","1","2","3"]
        timestamp_list=["1465230523", "1465130523", "1475130523", "1495130523"]
        start_ts_list=["1465230523", "1465130523", "1475130523", "1485130523"]
        end_ts_list=["1465230523", "1465130523", "1475130523", "1485130523"]
#if modulo of 15 is zero then make userid as null if not make the user id with random
numbers #between 100 to 120 and song id between 200 to 210 random numbers
        if (count%15 == 0):
                user_id = ""
        else: user_id = "U" + str(randint(100,120))
        song_id = "S" + str(randint(200,210))
```

```
#if modulo of 11 is zero then make artist_id is null, if not make the artist_id with random
numbers #between 300 to 305 and choose any of values from timestamp list,start_ts,
end_ts
        if (count%11 == 0):
                artist_id = ""
        else:   artist_id = "A" + str(randint(300,305))
        timestamp = choice(timestamp_list)
        start_ts = choice(start_ts_list)
        end_ts = choice(end_ts_list)
#if modulo of 12 is zero then make geo_id is null, if not make the geo_id with list from
geo_cd_list   # station_id is made from random numbers between 400 to 415 and
Song_end_type is choosen from the list song_end_type_list
        if (count%12 == 0):
                geo_cd = ""
        else:
                geo_cd = choice(geo_cd_list)
        station_id = "ST" + str(randint(400,415))
        song_end_type = choice(song_end_type_list)
        like = str(randint(0,1)) #like variable which will save any random integer, either 0
or 1
        dislike = str(randint(0,1)) #dislike variable which will save any random integer,
either 0| 1
#write the data to the file and decrease the count by 1 and close the file
        file.write("%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s\n" % (user_id, song_id, artist_id,
timestamp, start_ts, end_ts, geo_cd, station_id, song_end_type, like, dislike))
        count = count-1
        file.close()
```

3. **Generate_web_data.py**
   Similar to the mobile data create web data. Here data is in XML format.

```
from random import randint
from random import choice
```

```
file = open("/home/acadgild/examples/music/data/web/file.xml", "w")
count = 20
file.write("<records>\n")
while (count > 0):
        geo_cd_list=["A", "E", "AU", "AP", "U"]
        song_end_type_list=["0","1","2","3"]
        timestamp_list=["2016-05-10 12:24:22", "2016-06-09 22:12:36", "2016-07-
10 01:38:09", "2017-05-09 08:09:22"]
        start_ts_list=["2016-05-10 12:24:22", "2016-06-09 22:12:36", "2016-07-10
01:38:09", "2017-05-09 08:09:22"]
        end_ts_list=["2016-05-10 12:24:22", "2016-06-09 22:12:36", "2016-07-10
01:38:09", "2017-05-09 08:09:22"]
        if (count%15 == 0):
                user_id = ""
        else:
                user_id = "U" + str(randint(100,120))

        song_id = "S" + str(randint(200,210))

        if (count%11 == 0):
                artist_id = ""
        else:
                artist_id = "A" + str(randint(300,305))
        timestamp = choice(timestamp_list)
        start_ts = choice(start_ts_list)
        end_ts = choice(end_ts_list)
        if (count%12 == 0):
                geo_cd = ""
       else:
               geo_cd = choice(geo_cd_list)
       station_id = "ST" + str(randint(400,415))
       song_end_type = choice(song_end_type_list)
       like = str(randint(0,1))
       dislike = str(randint(0,1))
       file.write("<record>\n")
       file.write("<user_id>%s</user_id>\n" % (user_id))
       file.write("<song_id>%s</song_id>\n" % (song_id))
       file.write("<artist_id>%s</artist_id>\n" % (artist_id))
       file.write("<timestamp>%s</timestamp>\n" % (timestamp))
       file.write("<start_ts>%s</start_ts>\n" % (start_ts))
       file.write("<end_ts>%s</end_ts>\n" % (end_ts))
       file.write("<geo_cd>%s</geo_cd>\n" % (geo_cd))
```

```
            file.write("<station_id>%s</station_id>\n" % (station_id))
            file.write("<song_end_type>%s</song_end_type>\n" % (song_end_type))
            file.write("<like>%s</like>\n" % (like))
            file.write("<dislike>%s</dislike>\n" % (dislike))
            file.write("</record>\n")
            count = count-1
    file.write("</records>")
    file.close()
```

4. **Start daemons.sh**

```
#!/bin/bash

#delete the logs from the mentioned directory

rm -r /home/acadgild/examples/music/logs

mkdir -p /home/acadgild/examples/music/logs

#search for the batch file, if it is found print batch file is found if not  change the
permissions of the #directory, cat the log file and start the respective daemons.

if [ -f "/home/acadgild/examples/music/logs/current-batch.txt" ]

then

 echo "Batch File Found!"

else

 echo -n "1" > "/home/acadgild/examples/music/logs/current-batch.txt"

fi

chmod 775 /home/acadgild/examples/music/logs/current-batch.txt

echo "After chmod"

batchid=`cat /home/acadgild/examples/music/logs/current-batch.txt`

echo "After batchid-->> "$batchid

LOGFILE=/home/acadgild/examples/music/logs/log_batch_$batchid

echo "Starting daemons" >> $LOGFILE

start-all.sh

start-hbase.sh

mr-jobhistory-daemon.sh start historyserver
```

cat /home/acadgild/examples/music/logs/current-batch.txt

5. Populate.lookup.sh

```
#!/bin/bash
#save the contents of the current-batch.txt to the variable "batched"
batchid=`cat /home/acadgild/examples/music/logs/current-batch.txt`
LOGFILE=/home/acadgild/examples/music/logs/log_batch_$batchid
echo "Creating LookUp Tables" >> $LOGFILE
echo "disable 'station-geo-map'" | hbase shell
echo "drop 'station-geo-map'" | hbase shell
echo "disable 'subscribed-users'" | hbase shell
echo "drop 'subscribed-users'" | hbase shell
echo "disable 'song-artist-map'" | hbase shell
echo "drop 'song-artist-map'" | hbase shell
echo "create 'station-geo-map', 'geo'" | hbase shell
echo "create 'subscribed-users', 'subscn'" | hbase shell
echo "create 'song-artist-map', 'artist'" | hbase shell
echo "Populating LookUp Tables" >> $LOGFILE
#read the contents of the file "stn-geocd.txt"
file="/home/acadgild/examples/music/lookupfiles/stn-geocd.txt"
while IFS= read -r line
do
 stnid=`echo $line | cut -d',' -f1`
 geocd=`echo $line | cut -d',' -f2`
 echo "put 'station-geo-map', '$stnid', 'geo:geo_cd', '$geocd'" | hbase shell
done <"$file"
#read the contents of the file "song-artist.txt"
file="/home/acadgild/examples/music/lookupfiles/song-artist.txt"
```

```
while IFS= read -r line
do
 songid=`echo $line | cut -d',' -f1`
 artistid=`echo $line | cut -d',' -f2`
 echo "put 'song-artist-map', '$songid', 'artist:artistid', '$artistid'" | hbase shell
done <"$file"
#read the contents of the file "user-subscn.txt"
file="/home/acadgild/examples/music/lookupfiles/user-subscn.txt"
while IFS= read -r line
do
 userid=`echo $line | cut -d',' -f1`
 startdt=`echo $line | cut -d',' -f2`
 enddt=`echo $line | cut -d',' -f3`
 echo "put 'subscribed-users', '$userid', 'subscn:startdt', '$startdt'" | hbase shell
 echo "put 'subscribed-users', '$userid', 'subscn:enddt', '$enddt'" | hbase shell
done <"$file"
#open the user-artist.hql file
hive -f /home/acadgild/examples/music/user-artist.hql
```

6. **User-artist.hql**

```
CREATE DATABASE IF NOT EXISTS project;
USE project;
#create table "user_artists"
CREATE TABLE users_artists
( user_id STRING,
artists_array ARRAY<STRING>)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
COLLECTION ITEMS TERMINATED BY '&';
```

```
#insert the contents of the file to above created table.
LOAD DATA LOCAL INPATH '/home/acadgild/project/lookupfiles/user-artist.txt'
OVERWRITE INTO TABLE users_artists;
```

7. Data formatting.sh

```
#!/bin/bash
batchid=`cat /home/acadgild/project/logs/current-batch.txt`
LOGFILE=/home/acadgild/project/logs/log_batch_$batchid
echo "Running script for data formatting..." >> $LOGFILE
#spark command to execute the spark program
spark-submit --packages com.databricks:spark-xml_2.10:0.4.1 \
--class DataFormatting \
--master local[2] \
/home/acadgild/project/scripts/MusicDataAnalysis/target/scala-
2.11/musicdataanalysis_2.11-1.0.jar $batchid
```

8. Data enrichment and filtering

```
#!/bin/bash
batchid=`cat /home/acadgild/project/logs/current-batch.txt`
LOGFILE=/home/acadgild/project/logs/log_batch_$batchid
echo "Creating hive tables on top of hbase tables for data enrichment and filtering..." >>
$LOGFILE
#calling the hive_hbase_lookup file to perform lookup operation
hive -f /home/acadgild/project/scripts/create_hive_hbase_lookup.hql
```

9. create_hive_hbase_lookup.hql

```
USE project;
#create an external hive tables by integrating it with Hbase for
#station_geo_map,subscribed_users,song_artist_map
create external table if not exists station_geo_map
(
station_id String,
```

```
geo_cd string
)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
with serdeproperties
("hbase.columns.mapping"=":key,geo:geo_cd")
tblproperties("hbase.table.name"="station-geo-map");
create external table if not exists subscribed_users
(
user_id STRING,
subscn_start_dt STRING,
subscn_end_dt STRING
)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
with serdeproperties
("hbase.columns.mapping"=":key,subscn:startdt,subscn:enddt")
tblproperties("hbase.table.name"="subscribed-users");
create external table if not exists song_artist_map
(
song_id STRING,
artist_id STRING
)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
with serdeproperties
("hbase.columns.mapping"=":key,artist:artistid")
tblproperties("hbase.table.name"="song-artist-map");
```

10. data enrichment.sh

```
#!/bin/bash

batchid=`cat /home/acadgild/project/logs/current-batch.txt`

LOGFILE=/home/acadgild/project/logs/log_batch_$batchid

VALIDDIR=/home/acadgild/project/processed_dir/valid/batch_$batchid

INVALIDDIR=/home/acadgild/project/processed_dir/invalid/batch_$batchid

echo "Running script for data enrichment and filtering..." >> $LOGFILE

#executing the dataenrichment class using the spark-submit command

spark-submit --class DataEnrichment \

--master local[2] \

--jars /usr/local/hive/lib/hive-hbase-handler-2.1.1.jar,/usr/local/hive/lib/hbase-client-1.1.1.jar,/usr/local/hive/lib/hbase-common-1.1.1.jar,/usr/local/hive/lib/hbase-hadoop-compat-1.1.1.jar,/usr/local/hive/lib/hbase-server-1.1.1.jar,/usr/local/hive/lib/hbase-protocol-1.1.1.jar,/usr/local/hive/lib/zookeeper-3.4.6.jar,/usr/local/hive/lib/guava-14.0.1.jar,/usr/local/hive/lib/htrace-core-3.1.0-incubating.jar \

/home/acadgild/project/scripts/MusicDataAnalysis/target/scala-2.11/musicdataanalysis_2.11-1.0.jar $batchid

#if batchid is valid then make the directory validdir, if not make invalid direcotry

if [ ! -d "$VALIDDIR" ]

then

mkdir -p "$VALIDDIR"

fi

if [ ! -d "$INVALIDDIR" ]
then

mkdir -p "$INVALIDDIR"

fi

echo "Copying valid and invalid records in local file system..." >> $LOGFILE

#save the valid files in valid directory and invalid files in invalid directory
```

```
hadoop fs -get
/user/hive/warehouse/project.db/enriched_data/batchid=$batchid/status=pass/*
$VALIDDIR

hadoop fs -get
/user/hive/warehouse/project.db/enriched_data/batchid=$batchid/status=fail/*
$INVALIDDIR

echo "Deleting older valid and invalid records from local file system..." >> $LOGFILE

find /home/acadgild/project/processed_dir/ -mtime +7 -exec rm {} \;
```

11. Data analysis.sh

```
#!/bin/bash

batchid=`cat /home/acadgild/project/logs/current-batch.txt`

LOGFILE=/home/acadgild/project/logs/log_batch_$batchid

VALIDDIR=/home/acadgild/project/processed_dir/valid/batch_$batchid

INVALIDDIR=/home/acadgild/project/processed_dir/invalid/batch_$batchid

echo "Running script for data enrichment and filtering..." >> $LOGFILE

#executing the dataAnalysis class using the spark-submit command

spark-submit --class DataAnalysis \

--master local[2] \

--jars /usr/local/hive/lib/hive-hbase-handler-2.1.1.jar,/usr/local/hive/lib/hbase-
client-1.1.1.jar,/usr/local/hive/lib/hbase-common-1.1.1.jar,/usr/local/hive/lib/hbase-
hadoop-compat-1.1.1.jar,/usr/local/hive/lib/hbase-server-
1.1.1.jar,/usr/local/hive/lib/hbase-protocol-1.1.1.jar,/usr/local/hive/lib/zookeeper-
3.4.6.jar,/usr/local/hive/lib/guava-14.0.1.jar,/usr/local/hive/lib/htrace-core-3.1.0-
incubating.jar \

/home/acadgild/project/scripts/MusicDataAnalysis/target/scala-
2.11/musicdataanalysis_2.11-1.0.jar $batchid
```

12. Data Enrichment.scala

```
//import statements

import org.apache.spark.{SparkConf, SparkContext}

import org.apache.spark.sql

//main function
```

```
object DataEnrichment {
 def main(args: Array[String]): Unit = {
//creating the spark context and hive context
   val conf = new SparkConf().setAppName("Data Formatting")
   val sc = new SparkContext(conf)
   val sqlContext = new org.apache.spark.sql.hive.HiveContext(sc)
   val batchId = args(0)
// creating a hive table with all required fields
 val create_hive_table = """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
                  """
//inserting the data to the table
```

```
val load_data = s"""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=$batchId
                """
  try {
     sqlContext.sql("SET hive.auto.convert.join=false")
     sqlContext.sql("SET hive.exec.dynamic.partition.mode=nonstrict")
     sqlContext.sql("USE project")

     sqlContext.sql(create_hive_table)
     sqlContext.sql(load_data)
    } catch{
    case e: Exception=>e.printStackTrace()
   } }}
```

13. DataFormatting.scala

```
import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.sql
//main function creating spark and hive context
object DataFormatting {
 def main(args: Array[String]): Unit = {
  val conf = new SparkConf().setAppName("Data Formatting")
  val sc = new SparkContext(conf)
  val sqlContext = new org.apache.spark.sql.hive.HiveContext(sc)
```

```
  val batchId = args(0)
//creating hive table
  val create_hive_table = """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 ','
                """
//inserting data to the table "mobile application data"
  val load_mob_data = s"""LOAD DATA LOCAL INPATH
'/home/acadgild/examples/music/data/mob/file.txt'
                INTO TABLE project.formatted_input PARTITION (batchid='$batchId')"""
//inserting web application data to the table
  val load_web_data = s"""INSERT INTO project.formatted_input
                PARTITION(batchid='$batchId')
```

```
                    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
                    """
    //formatting the xml data from the xml file and creating a temporary view on it
      try {
          val xmlData = sqlContext.read.format("com.databricks.spark.xml").option("rowTag",
"record").load("/home/acadgild/examples/music/data/web/file.xml")
          xmlData.createOrReplaceTempView("web_data")
          sqlContext.sql(create_hive_table)
          sqlContext.sql(load_mob_data)
          sqlContext.sql(load_web_data)
    }     catch{      case e: Exception=>e.printStackTrace()     } }}
```

14. DataAnalysis.scala

```
import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.sql
//main function which create hive and spark context
object DataAnalysis {
  def main(args: Array[String]): Unit = {

    val conf = new SparkConf().setAppName("Data Analysis")
    val sc = new SparkContext(conf)
    val sqlContext = new org.apache.spark.sql.hive.HiveContext(sc)
    val batchId = args(0)
```

```
//creating hive tables to analyze the data extracted so far
//finding top 10 stations
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"""
//loading the data to the created table
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"""
//analyzing users behaviour
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"""

//analyze data based on connected artists

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'
```

```
)
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"""
// create top 10 royalty songs
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"""
```

```
//create top 10 unsubscribed users
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"""
  try {
```

```
  sqlContext.sql("SET hive.auto.convert.join=false")
  sqlContext.sql("USE project")
  sqlContext.sql(create_top_10_stations)
  sqlContext.sql(load_top_10_stations)
  sqlContext.sql(create_users_behaviour)
  sqlContext.sql(load_users_behaviour)
  sqlContext.sql(create_connected_artists)
  sqlContext.sql(load_connected_artists)
  sqlContext.sql(create_top_10_royalty_songs)
  sqlContext.sql(load_top_10_royalty_songs)
  sqlContext.sql(create_top_10_unsubscribed_users)
  sqlContext.sql(load_top_10_unsubscribed_users)
 } catch{
 case e: Exception=>e.printStackTrace()
}}}
```

15. Create Schema.hql

```
CREATE DATABASE IF NOT EXISTS project;
USE project;
CREATE TABLE IF NOT EXISTS top_10_stations
(
station_id VARCHAR(50),
total_distinct_songs_played INT,
distinct_user_count INT
);
CREATE TABLE IF NOT EXISTS users_behaviour
(
user_type VARCHAR(50),
duration BIGINT
```

```
);
CREATE TABLE IF NOT EXISTS connected_artists
(
artist_id VARCHAR(50),
user_count INT
);
CREATE TABLE IF NOT EXISTS top_10_royalty_songs
(
song_id VARCHAR(50),
duration BIGINT
);
CREATE TABLE IF NOT EXISTS top_10_unsubscribed_users
(
user_id VARCHAR(50),
duration BIGINT
);
commit;
```

16. Data export.sh

```
#!/bin/bash
batchid=`cat /home/acadgild/example/music/logs/current-batch.txt`
LOGFILE=/home/acadgild/example/music/logs/log_batch_$batchid
echo "Creating mysql tables if not present..." >> $LOGFILE
mysql -u root < /home/acadgild/example/music/create_schema.sql
echo "Running sqoop job for data export..." >> $LOGFILE
#sqoop commands to export data stored in hive tables to RDBMS i.e. to MySQL.
sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table
top_10_stations --export-dir
hdfs://localhost:8020/user/hive/warehouse/project.db/top_10_stations/batchid=$batchi
d --input-fields-terminated-by ',' -m 1
```

sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table users_behaviour --export-dir hdfs://localhost:8020/user/hive/warehouse/project.db/users_behaviour/batchid=$batchid --input-fields-terminated-by ',' -m 1

sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table connected_artists --export-dir hdfs://localhost:8020/user/hive/warehouse/project.db/connected_artists/batchid=$batchid --input-fields-terminated-by ',' -m 1

sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table top_10_royalty_songs --export-dir hdfs://localhost:8020/user/hive/warehouse/project.db/top_10_royalty_songs/batchid=$batchid --input-fields-terminated-by ',' -m 1

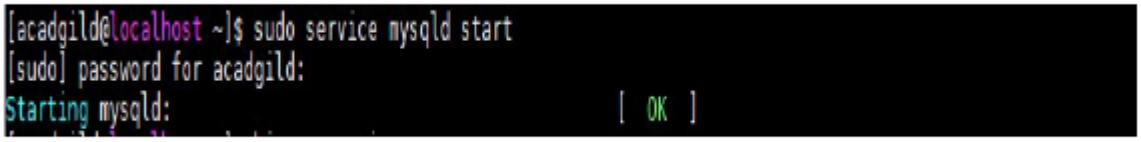
sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table top_10_unsubscribed_users --export-dir hdfs://localhost:8020/user/hive/warehouse/project.db/top_10_unsubscribed_users/batchid=$batchid --input-fields-terminated-by ',' -m 1

Project execution

Preliminary steps for execution:

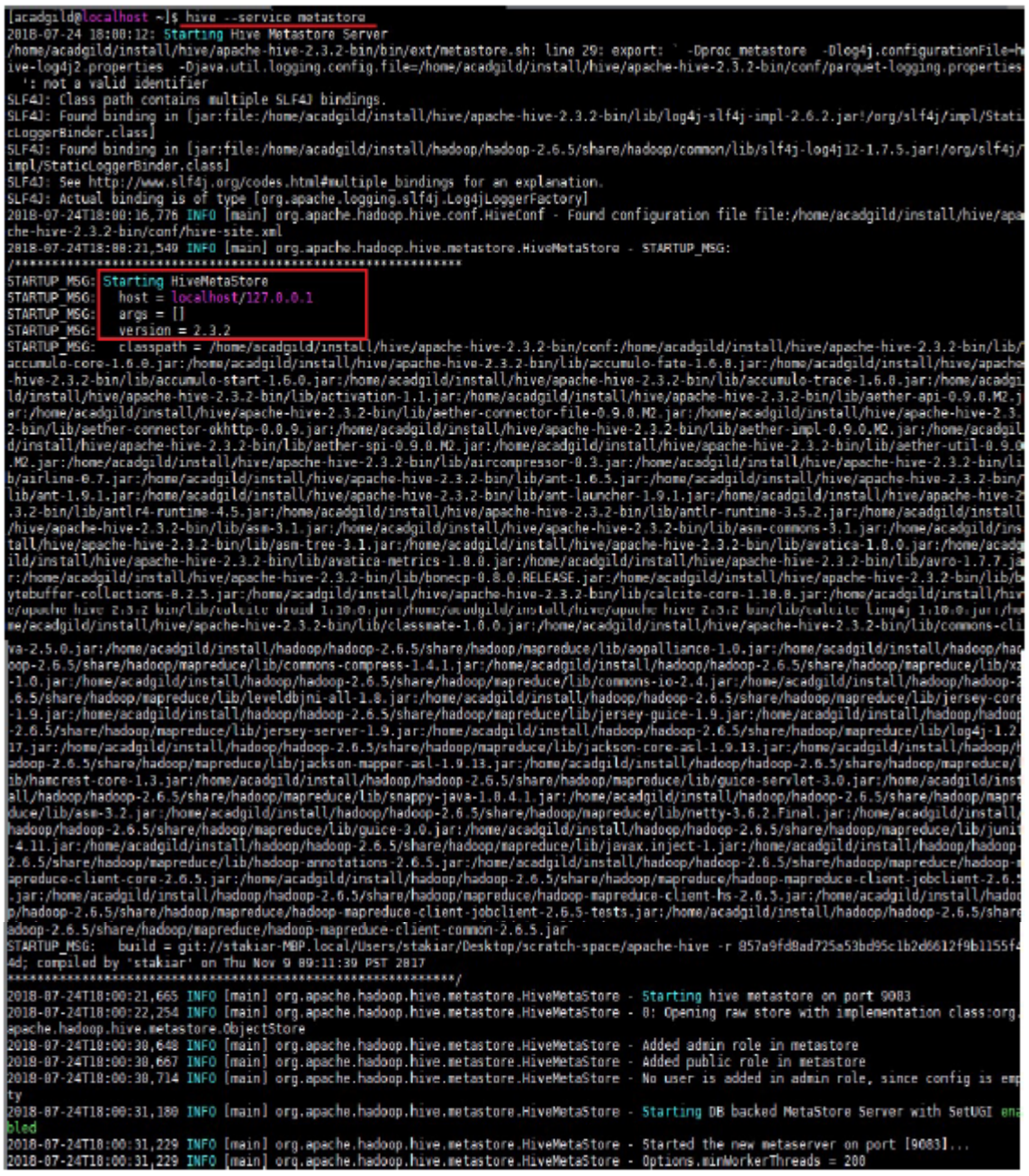
a) Check the numbers of process currently running:

b) Start mysqld services



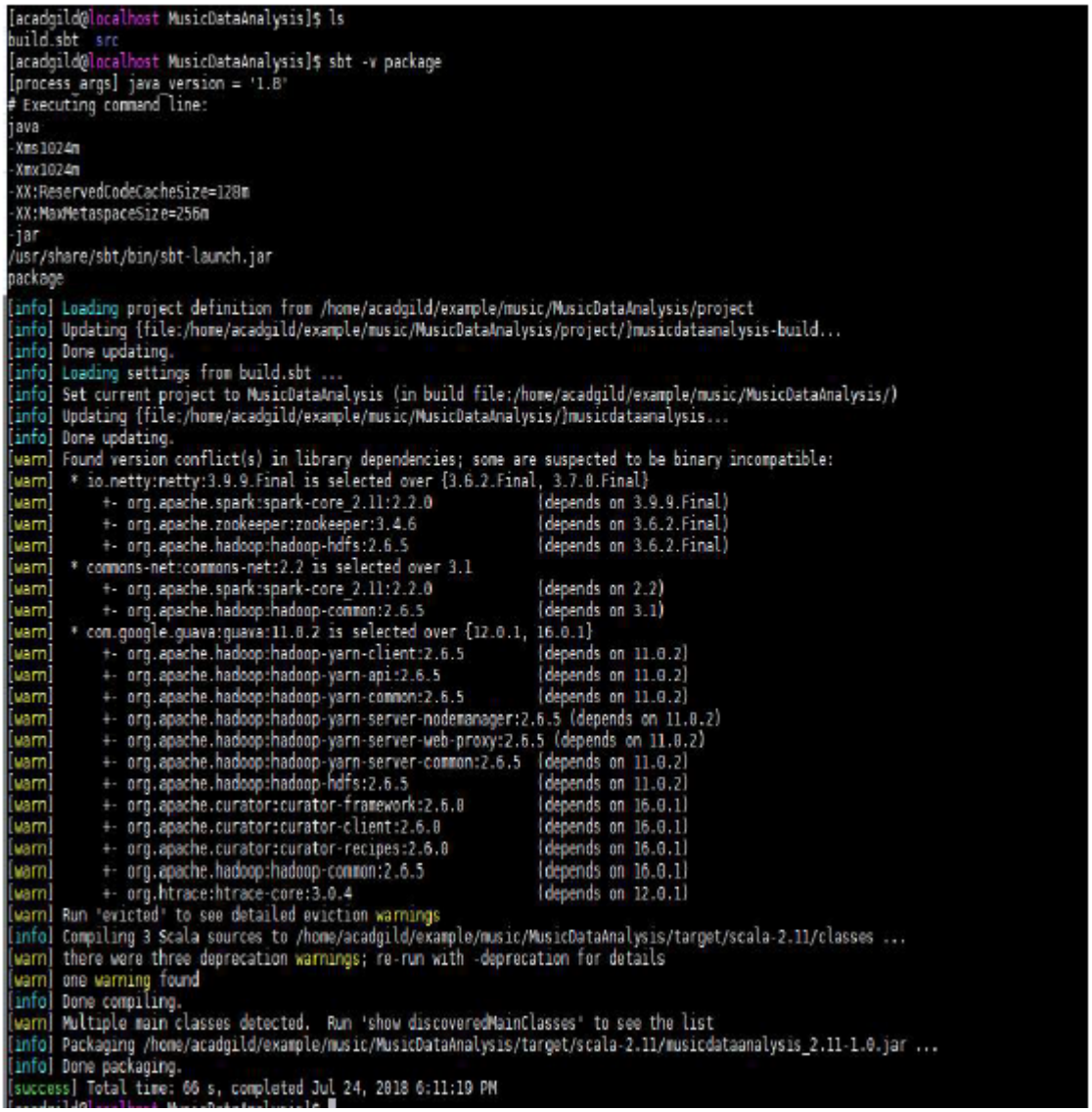
c) Starting hive store:

Let this be running in another tab



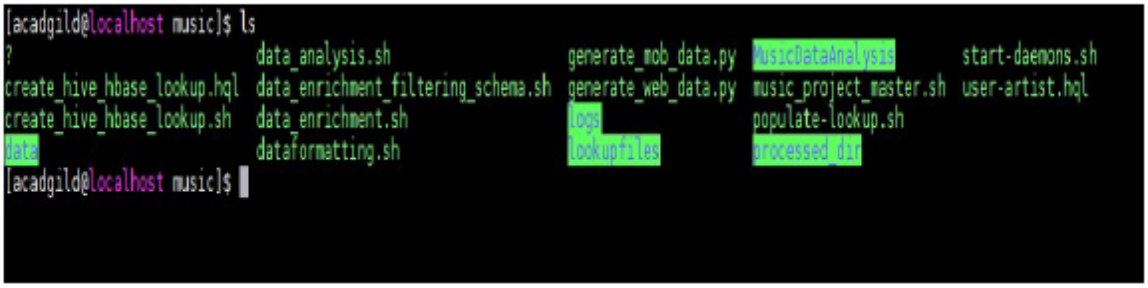
d) Open another tab, go the directory where we have all Scala class files:

Now we run the command "sbt –v package" which will download all the dependencies required


[acadgild@localhost MusicDataAnalysis]$ ls
build.sbt src
[acadgild@localhost MusicDataAnalysis]$ sbt -v package
[process_args] java version = '1.8'
Executing command line:
java
-Xms1024m
-Xmx1024m
-XX:ReservedCodeCacheSize=128m
-XX:MaxMetaspaceSize=256m
-jar
/usr/share/sbt/bin/sbt-launch.jar
package
[info] Loading project definition from /home/acadgild/example/music/MusicDataAnalysis/project
[info] Updating {file:/home/acadgild/example/music/MusicDataAnalysis/project/}musicdataanalysis-build...
[info] Done updating.
[info] Loading settings from build.sbt ...
[info] Set current project to MusicDataAnalysis (in build file:/home/acadgild/example/music/MusicDataAnalysis/)
[info] Updating {file:/home/acadgild/example/music/MusicDataAnalysis/}musicdataanalysis...
[info] Done updating.
[warn] Found version conflict(s) in library dependencies; some are suspected to be binary incompatible:
[warn] * io.netty:netty:3.9.9.Final is selected over {3.6.2.Final, 3.7.0.Final}
[warn]   +- org.apache.spark:spark-core_2.11:2.2.0   (depends on 3.9.9.Final)
[warn]   +- org.apache.zookeeper:zookeeper:3.4.6   (depends on 3.6.2.Final)
[warn]   +- org.apache.hadoop:hadoop-hdfs:2.6.5   (depends on 3.6.2.Final)
[warn] * commons-net:commons-net:2.2 is selected over 3.1
[warn]   +- org.apache.spark:spark-core_2.11:2.2.0   (depends on 2.2)
[warn]   +- org.apache.hadoop:hadoop-common:2.6.5   (depends on 3.1)
[warn] * com.google.guava:guava:11.0.2 is selected over {12.0.1, 16.0.1}
[warn]   +- org.apache.hadoop:hadoop-yarn-client:2.6.5   (depends on 11.0.2)
[warn]   +- org.apache.hadoop:hadoop-yarn-api:2.6.5   (depends on 11.0.2)
[warn]   +- org.apache.hadoop:hadoop-yarn-common:2.6.5   (depends on 11.0.2)
[warn]   +- org.apache.hadoop:hadoop-yarn-server-nodemanager:2.6.5 (depends on 11.0.2)
[warn]   +- org.apache.hadoop:hadoop-yarn-server-web-proxy:2.6.5 (depends on 11.0.2)
[warn]   +- org.apache.hadoop:hadoop-yarn-server-common:2.6.5 (depends on 11.0.2)
[warn]   +- org.apache.hadoop:hadoop-hdfs:2.6.5   (depends on 11.0.2)
[warn]   +- org.apache.curator:curator-framework:2.6.0   (depends on 16.0.1)
[warn]   +- org.apache.curator:curator-client:2.6.0   (depends on 16.0.1)
[warn]   +- org.apache.curator:curator-recipes:2.6.0   (depends on 16.0.1)
[warn]   +- org.apache.hadoop:hadoop-common:2.6.5   (depends on 16.0.1)
[warn]   +- org.htrace:htrace-core:3.0.4   (depends on 12.0.1)
[warn] Run 'evicted' to see detailed eviction warnings
[info] Compiling 3 Scala sources to /home/acadgild/example/music/MusicDataAnalysis/target/scala-2.11/classes ...
[warn] there were three deprecation warnings; re-run with -deprecation for details
[warn] one warning found
[info] Done compiling.
[warn] Multiple main classes detected. Run 'show discoveredMainClasses' to see the list
[info] Packaging /home/acadgild/example/music/MusicDataAnalysis/target/scala-2.11/musicdataanalysis_2.11-1.0.jar ...
[info] Done packaging.
[success] Total time: 66 s, completed Jul 24, 2018 6:11:19 PM

e) Now we can see folders called "project" & "target", let's go inside target folder and check if the jar file created has all required permissions:
   I. We can see the jar files does not have required permission to be executed
   II. To provide required permission, we execute chmod command to provide permissions.
   III. Now we can see that all files are converted to green color, which indicates that files have all required permissions to be executed.

f) Now let's go the directory where we have saved all the program codes and scripts and check if all the files have all permissions to execute the project:



Project execution:

Now let's execute the project by executing the master script

```
[acadgild@localhost music]$ ls
create_hive_hbase_lookup.hql  data_analysis.sh                  dataformatting.sh     lookupfiles          processed_dir
create_hive_hbase_lookup.sh   data_enrichment_filtering_schema.sh  generate_mob_data.py  MusicDataAnalysis    start-daemons.sh
create_schema.sql             data_enrichment.sh                generate_web_data.py  music_project_master.sh  top_10_stations.java
data                          data_export.sh                    logs                  populate-lookup.sh   user-artist.hql
[acadgild@localhost music]$
[acadgild@localhost music]$ ./music_project_master.sh
Preparing to execute python scripts to generate data...
Data Generated Successfully !
Starting the daemons....
After chmod
After batchid-->> 1
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
18/07/25 17:58:13 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-localh
ost.localdomain.out
18/07/25 17:58:45 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.localdom
ain.out
localhost: starting zookeeper, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-zookeeper-localhost.localdomain.o
ut
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
starting historyserver, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/mapred-acadgild-historyserver-localhost.localdomain.o
ut
```

```
12706 RunJar
4387 Jps
4245 HRegionServer
2984 NameNode
3417 ResourceManager
4154 HMaster
3083 DataNode
4348 JobHistoryServer
3518 NodeManager
3262 SecondaryNameNode
4062 HQuorumPeer
All hadoop daemons started !
Upload the look up tables now in Hbase...
2018-07-25 18:00:30,065 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerBinde
r.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
```

```
disable 'station-geo-map'
0 row(s) in 3.7150 seconds

2018-07-25 10:45:44,787 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerBinde
r.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

drop 'station-geo-map'
0 row(s) in 2.6530 seconds

2018-07-25 10:46:04,719 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerBinde
r.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
```

```
drop 'subscribed-users'
0 row(s) in 2.6990 seconds

2018-07-25 10:46:49,517 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerBinde
r.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

disable 'song-artist-map'
0 row(s) in 3.8580 seconds

2018-07-25 10:47:11,438 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerBinde
r.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
```

```
disable 'song-artist-map'
0 row(s) in 3.8580 seconds

2018-07-25 10:47:11,438 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerBinde
r.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

drop 'song-artist-map'
0 row(s) in 2.4560 seconds

2018-07-25 10:47:31,338 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerBinde
r.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 2.6040 seconds
```

Populating data:

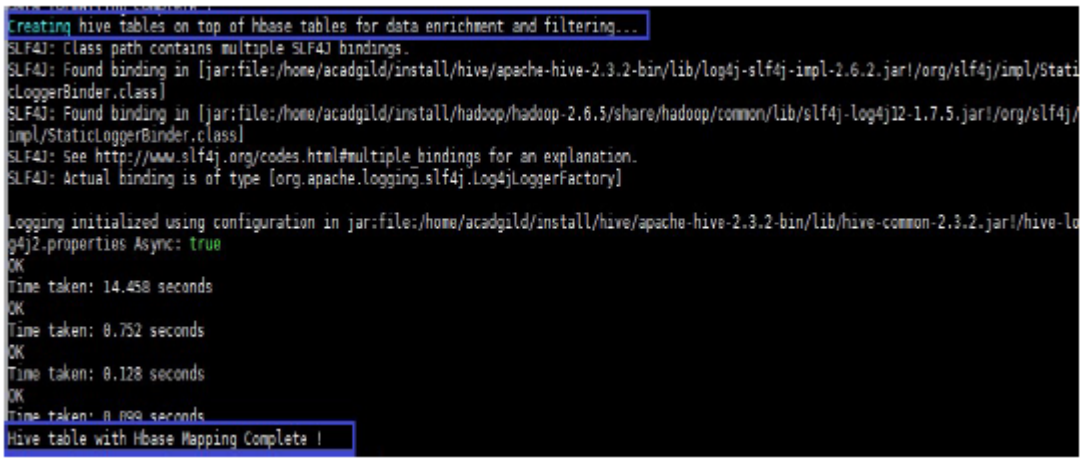
```
put 'station-geo-map', 'ST400', 'geo:geo_cd', 'A'
0 row(s) in 1.3700 seconds

2018-07-25 10:48:51,728 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerBinde
r.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 1.2450 seconds

2018-07-25 10:49:18,148 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-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/StaticLoggerBinde
r.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.
```

DataFormatting:

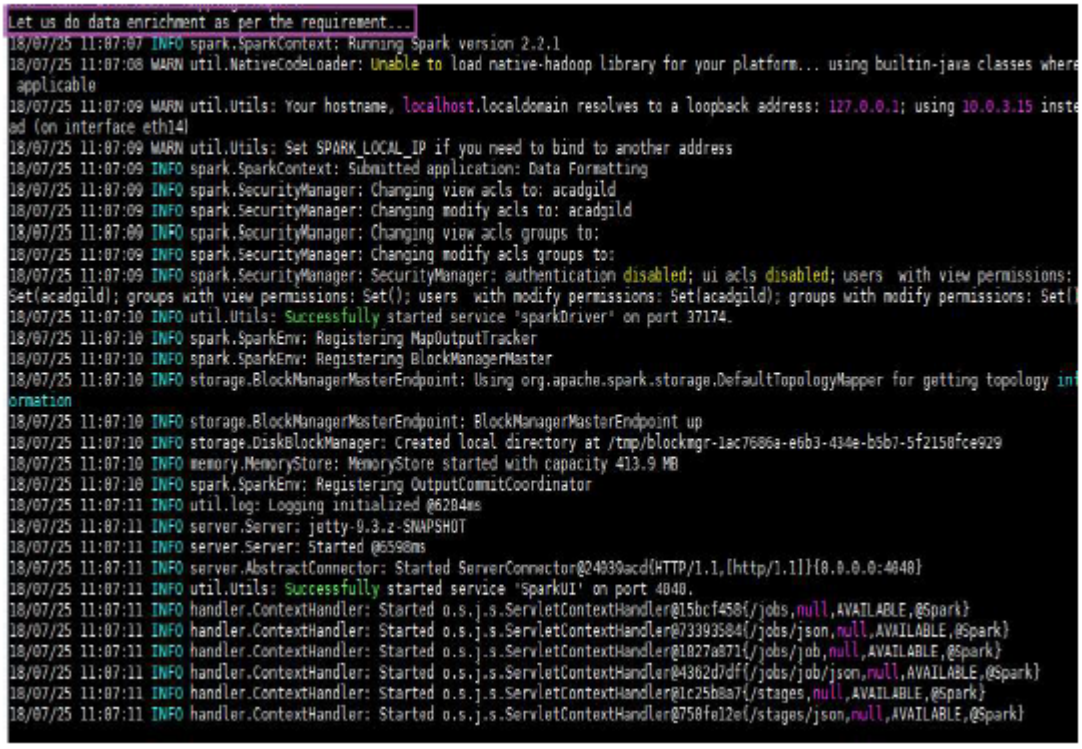
```
Done with data population in look up tables !
Lets do some data formatting now....
Ivy Default Cache set to: /home/acadgild/.ivy2/cache
The jars for the packages stored in: /home/acadgild/.ivy2/jars
:: loading settings :: url = jar:file:/home/acadgild/install/spark/spark-2.2.1-bin-hadoop2.7/jars/ivy-2.4.0.jar!/org/apache/ivy/core/set
tings/ivysettings.xml
com.databricks#spark-xml_2.10 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent;1.0
        confs: [default]
        found com.databricks#spark-xml_2.10;0.4.1 in central
:: resolution report :: resolve 520ms :: artifacts dl 20ms
        :: modules in use:
        com.databricks#spark-xml_2.10;0.4.1 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   1   |   0   |   0   |   0   ||   1   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent
        confs: [default]
        0 artifacts copied, 1 already retrieved (0kB/40ms)
18/07/25 11:05:50 INFO spark.SparkContext: Running Spark version 2.2.1
18/07/25 11:05:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
18/07/25 11:05:51 WARN util.Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 10.0.3.15 inste
ad (on interface eth1)
18/07/25 11:05:51 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
18/07/25 11:05:51 INFO spark.SparkContext: Submitted application: Data Formatting
18/07/25 11:05:52 INFO spark.SecurityManager: Changing view acls to: acadgild
18/07/25 11:05:52 INFO spark.SecurityManager: Changing modify acls to: acadgild
18/07/25 11:05:52 INFO spark.SecurityManager: Changing view acls groups to:
18/07/25 11:05:52 INFO spark.SecurityManager: Changing modify acls groups to:
18/07/25 11:05:52 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions:
Set(acadgild); groups with view permissions: Set(); users  with modify permissions: Set(acadgild); groups with modify permissions: Set()
18/07/25 11:05:53 INFO util.Utils: Successfully started service 'sparkDriver' on port 41365.
```

```
18/07/25 11:06:28 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:28 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:28 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:28 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:28 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:28 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:28 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:06:28 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:06:28 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:06:28 INFO metadata.Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/formatted_input/.hive-staging_h
ive_2018-07-25_11-06-23_658_2053715528423523164-1/-ext-10000/part-00000-efd2a0d1-e7dd-4342-9401-c7f838c4fa97-c000, dest: hdfs://localhos
t:8020/user/hive/warehouse/project.db/formatted_input/batchid=1/part-00000-efd2a0d1-e7dd-4342-9401-c7f838c4fa97-c000, Status:true
18/07/25 11:06:29 INFO execution.SparkSqlParser: Parsing command: `project`.`formatted_input`
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:06:29 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:06:29 INFO spark.SparkContext: Invoking stop() from shutdown hook
18/07/25 11:06:29 INFO server.AbstractConnector: Stopped Spark@14039fdc{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
18/07/25 11:06:29 INFO ui.SparkUI: Stopped Spark web UI at http://10.0.3.15:4040
18/07/25 11:06:29 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/07/25 11:06:29 INFO memory.MemoryStore: MemoryStore cleared
18/07/25 11:06:29 INFO storage.BlockManager: BlockManager stopped
18/07/25 11:06:29 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
18/07/25 11:06:29 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/07/25 11:06:29 INFO spark.SparkContext: Successfully stopped SparkContext
18/07/25 11:06:29 INFO util.ShutdownHookManager: Shutdown hook called
18/07/25 11:06:29 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-940200d0-9fb4-456c-a9e0-4ffbc27e37e7
data formatting complete !
```

Data Enrichment and filtering:



Data Enrichment



```
est: hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/batchid=1/status=pass/part-00177-967ffc16-7820-4d53-b8fa-28a6899
268bd.c000, Status:true
18/07/25 11:08:31 INFO metadata.Hive: Replacing src:hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-staging_hiv
e_2018-07-25_11-07-31_575_2375049059219045060-1/-ext-10000/batchid=1/status=pass/part-00199-967ffc16-7820-4d53-b8fa-28a6899268bd.c000, d
est: hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/batchid=1/status=pass/part-00199-967ffc16-7820-4d53-b8fa-28a6899
268bd.c000, Status:true
18/07/25 11:08:31 INFO metadata.Hive: New loading path = hdfs://localhost:8020/user/hive/warehouse/project.db/enriched_data/.hive-stagin
g_hive_2018-07-25_11-07-31_575_2375049059219045060-1/-ext-10000/batchid=1/status=pass with partSpec {batchid=1, status=pass}
18/07/25 11:08:31 INFO execution.SparkSqlParser: Parsing command: `project`.`enriched_data`
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:08:32 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:08:32 INFO spark.SparkContext: Invoking stop() from shutdown hook
18/07/25 11:08:32 INFO server.AbstractConnector: Stopped Spark@24039acd{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
18/07/25 11:08:32 INFO ui.SparkUI: Stopped Spark web UI at http://10.0.3.15:4040
18/07/25 11:08:32 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/07/25 11:08:32 INFO memory.MemoryStore: MemoryStore cleared
18/07/25 11:08:32 INFO storage.BlockManager: BlockManager stopped
18/07/25 11:08:32 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
18/07/25 11:08:32 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/07/25 11:08:32 INFO spark.SparkContext: Successfully stopped SparkContext
18/07/25 11:08:32 INFO util.ShutdownHookManager: Shutdown hook called
18/07/25 11:08:32 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-49c7357c-790e-4607-9ec6-b3db61d59d03
data_enrichment.sh: line 21: syntax error: unexpected end of file
Data Enrichment Complete
```

```
data_enrichment.sh: line 21: syntax error: unexpected end of file
Data Enrichment Complete
Lets run some use cases now...
18/07/25 11:08:35 INFO spark.SparkContext: Running Spark version 2.2.1
18/07/25 11:08:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
18/07/25 11:08:37 WARN util.Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 10.0.3.15 inste
ad (on interface eth14)
18/07/25 11:08:37 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
18/07/25 11:08:37 INFO spark.SparkContext: Submitted application: Data Analysis
18/07/25 11:08:37 INFO spark.SecurityManager: Changing view acls to: acadgild
18/07/25 11:08:37 INFO spark.SecurityManager: Changing modify acls to: acadgild
18/07/25 11:08:37 INFO spark.SecurityManager: Changing view acls groups to:
18/07/25 11:08:37 INFO spark.SecurityManager: Changing modify acls groups to:
18/07/25 11:08:37 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions:
Set(acadgild); groups with view permissions: Set(); users  with modify permissions: Set(acadgild); groups with modify permissions: Set()
18/07/25 11:08:38 INFO util.Utils: Successfully started service 'sparkDriver' on port 39638.
18/07/25 11:08:38 INFO spark.SparkEnv: Registering MapOutputTracker
18/07/25 11:08:38 INFO spark.SparkEnv: Registering BlockManagerMaster
18/07/25 11:08:38 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology inf
ormation
18/07/25 11:08:38 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
18/07/25 11:08:38 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-05d62f90-aa7a-4491-9873-eb34f22707a9
18/07/25 11:08:38 INFO memory.MemoryStore: MemoryStore started with capacity 413.9 MB
18/07/25 11:08:38 INFO spark.SparkEnv: Registering OutputCommitCoordinator
18/07/25 11:08:38 INFO util.log: Logging initialized @5381ms
18/07/25 11:08:39 INFO server.Server: jetty-9.3.z-SNAPSHOT
18/07/25 11:08:39 INFO server.Server: Started @5755ms
18/07/25 11:08:39 INFO server.AbstractConnector: Started ServerConnector@1cff4228{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
18/07/25 11:08:39 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@15bcf458{/jobs,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@73393584{/jobs/json,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@1827a871{/jobs/job,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@4362d7df{/jobs/job/json,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@1c25b8a7{/stages,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@750fe12e{/stages/json,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3e587920{/stages/stage,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@1e11bc55{/stages/stage/json,null,AVAILABLE,@Spark}
```

Data Analysis:

```
Lets run some use cases now...
18/07/25 11:08:35 INFO spark.SparkContext: Running Spark version 2.2.1
18/07/25 11:08:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
18/07/25 11:08:37 WARN util.Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 10.0.3.15 inste
ad (on interface eth14)
18/07/25 11:08:37 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
18/07/25 11:08:37 INFO spark.SparkContext: Submitted application: Data Analysis
18/07/25 11:08:37 INFO spark.SecurityManager: Changing view acls to: acadgild
18/07/25 11:08:37 INFO spark.SecurityManager: Changing modify acls to: acadgild
18/07/25 11:08:37 INFO spark.SecurityManager: Changing view acls groups to:
18/07/25 11:08:37 INFO spark.SecurityManager: Changing modify acls groups to:
18/07/25 11:08:37 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions:
Set(acadgild); groups with view permissions: Set(); users  with modify permissions: Set(acadgild); groups with modify permissions: Set()
18/07/25 11:08:38 INFO util.Utils: Successfully started service 'sparkDriver' on port 39638.
18/07/25 11:08:38 INFO spark.SparkEnv: Registering MapOutputTracker
18/07/25 11:08:38 INFO spark.SparkEnv: Registering BlockManagerMaster
18/07/25 11:08:38 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology inf
ormation
18/07/25 11:08:38 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
18/07/25 11:08:38 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-05d62f98-aa7a-4401-9873-eb34f22787a9
18/07/25 11:08:38 INFO memory.MemoryStore: MemoryStore started with capacity 413.9 MB
18/07/25 11:08:38 INFO spark.SparkEnv: Registering OutputCommitCoordinator
18/07/25 11:08:38 INFO util.log: Logging initialized @5381ms
18/07/25 11:08:39 INFO server.Server: jetty-9.3.z-SNAPSHOT
18/07/25 11:08:39 INFO server.Server: Started @5755ms
18/07/25 11:08:39 INFO server.AbstractConnector: Started ServerConnector@1cff4228{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
18/07/25 11:08:39 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@15bcf458{/jobs,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@73393584{/jobs/json,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@1827a871{/jobs/job,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@4362d7df{/jobs/job/json,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@1c25b8a7{/stages,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@750fe12e{/stages/json,null,AVAILABLE,@Spark}
18/07/25 11:08:39 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3e587920{/stages/stage,null,AVAILABLE,@Spark}
```

```
18/07/25 11:10:37 INFO datasources.FileFormatWriter: Job null committed.
18/07/25 11:10:37 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:10:37 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:10:37 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:10:37 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:10:37 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:10:37 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:10:37 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:10:37 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:10:37 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:10:37 INFO common.FileUtils: Creating directory if it doesn't exist: hdfs://localhost:8020/user/hive/warehouse/project.db/to
p_10_unsubscribed_users/batchid=1
18/07/25 11:10:37 INFO metadata.Hive: Renaming src: hdfs://localhost:8020/user/hive/warehouse/project.db/top_10_unsubscribed_users/.hive
-staging_hive_2018-07-25_11-10-29_243_2587906676680021037-1/-ext-10000/part-00000-ffaf1e6e-a732-4288-939c-0281e76ec106-c000, dest: hdfs:/
/localhost:8020/user/hive/warehouse/project.db/top_10_unsubscribed_users/batchid=1/part-00000-ffaf1e6e-a732-4288-939c-0281e76ec106-c000,
 Status:true
18/07/25 11:10:38 INFO execution.SparkSqlParser: Parsing command: `project`.`top_10_unsubscribed_users`
18/07/25 11:10:38 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:10:38 INFO parser.CatalystSqlParser: Parsing command: string
18/07/25 11:10:38 INFO parser.CatalystSqlParser: Parsing command: int
18/07/25 11:10:38 INFO spark.SparkContext: Invoking stop() from shutdown hook
18/07/25 11:10:38 INFO storage.BlockManagerInfo: Removed broadcast_25_piece0 on 10.0.3.15:44362 in memory (size: 20.6 KB, free: 413.6 MB
)
18/07/25 11:10:38 INFO storage.BlockManagerInfo: Removed broadcast_26_piece0 on 10.0.3.15:44362 in memory (size: 65.8 KB, free: 413.6 MB
)
18/07/25 11:10:38 INFO server.AbstractConnector: Stopped Spark@1cff4228{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
18/07/25 11:10:38 INFO ui.SparkUI: Stopped Spark web UI at http://10.0.3.15:4040
18/07/25 11:10:38 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/07/25 11:10:39 INFO memory.MemoryStore: MemoryStore cleared
18/07/25 11:10:39 INFO storage.BlockManager: BlockManager stopped
18/07/25 11:10:39 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
18/07/25 11:10:39 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/07/25 11:10:39 INFO spark.SparkContext: Successfully stopped SparkContext
18/07/25 11:10:39 INFO util.ShutdownHookManager: Shutdown hook called
18/07/25 11:10:39 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-d40b00a4-ab15-40bc-a261-f5be2c2e781a
USE CASES COMPLETE !!
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost music]$ ./music_project_master.sh
```

Output:

Let's check the output in respective tables:

HBase:

```
hbase(main):002:0> list
TABLE
SparkHBasesTable
song-artist-map
station-geo-map
subscribed-users
4 row(s) in 0.0510 seconds
```

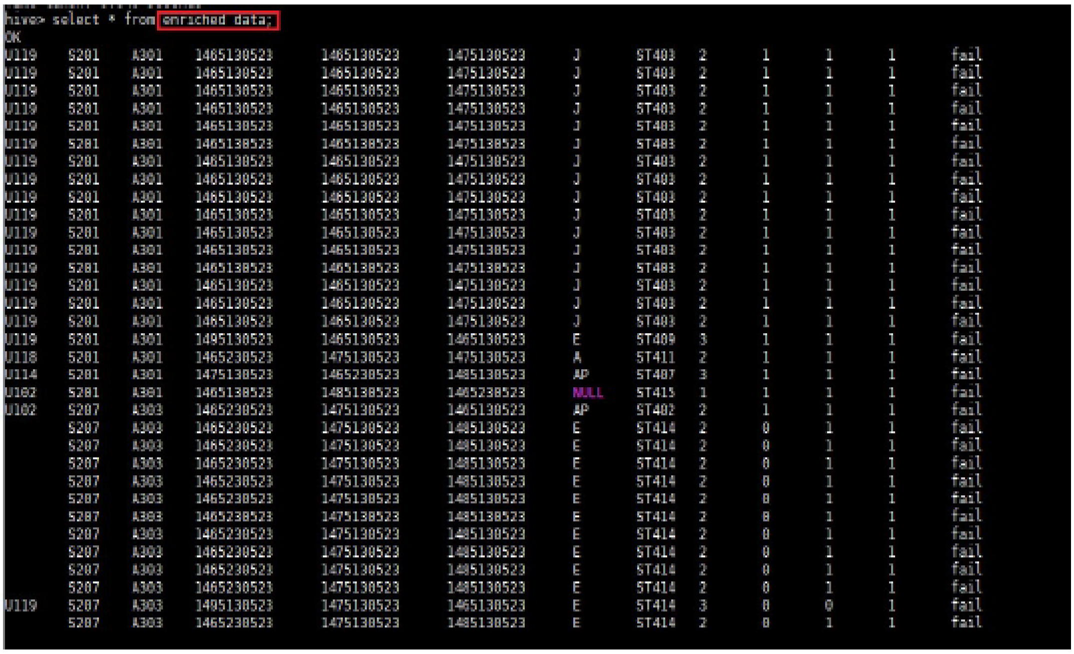
List of tables in Hbase

```
hbase(main):004:0> scan 'song-artist-map'
ROW                           COLUMN+CELL
 S200                         column=artist:artistid, timestamp=1532496197238, value=A300
 S201                         column=artist:artistid, timestamp=1532496216085, value=A301
 S202                         column=artist:artistid, timestamp=1532496233853, value=A302
 S203                         column=artist:artistid, timestamp=1532496251903, value=A303
 S204                         column=artist:artistid, timestamp=1532496260013, value=A304
 S205                         column=artist:artistid, timestamp=1532496287833, value=A301
 S206                         column=artist:artistid, timestamp=1532496306312, value=A302
 S207                         column=artist:artistid, timestamp=1532496324378, value=A303
 S208                         column=artist:artistid, timestamp=1532496342922, value=A304
 S209                         column=artist:artistid, timestamp=1532496362066, value=A305
10 row(s) in 0.1400 seconds

hbase(main):005:0> scan 'station-geo-map'
ROW                           COLUMN+CELL
 ST400                        column=geo:geo_cd, timestamp=1532495918086, value=A
 ST401                        column=geo:geo_cd, timestamp=1532495937056, value=AU
 ST402                        column=geo:geo_cd, timestamp=1532495955434, value=AP
 ST403                        column=geo:geo_cd, timestamp=1532495974102, value=J
 ST404                        column=geo:geo_cd, timestamp=1532495992558, value=E
 ST405                        column=geo:geo_cd, timestamp=1532496011402, value=A
 ST406                        column=geo:geo_cd, timestamp=1532496029862, value=AU
 ST407                        column=geo:geo_cd, timestamp=1532496048371, value=AP
 ST408                        column=geo:geo_cd, timestamp=1532496066679, value=E
 ST409                        column=geo:geo_cd, timestamp=1532496084662, value=E
 ST410                        column=geo:geo_cd, timestamp=1532496103989, value=A
 ST411                        column=geo:geo_cd, timestamp=1532496122377, value=A
 ST412                        column=geo:geo_cd, timestamp=1532496141693, value=AP
 ST413                        column=geo:geo_cd, timestamp=1532496159947, value=J
 ST414                        column=geo:geo_cd, timestamp=1532496178416, value=E
15 row(s) in 0.1280 seconds
```

```
hbase(main):006:0> scan 'subscribed-users'
ROW                           COLUMN+CELL
 U100                         column=subscn:enddt, timestamp=1532496397275, value=1465130523
 U100                         column=subscn:startdt, timestamp=1532496379484, value=1465230523
 U101                         column=subscn:enddt, timestamp=1532496433383, value=1475130523
 U101                         column=subscn:startdt, timestamp=1532496415326, value=1465230523
 U102                         column=subscn:enddt, timestamp=1532496469647, value=1475130523
 U102                         column=subscn:startdt, timestamp=1532496451588, value=1465230523
 U103                         column=subscn:enddt, timestamp=1532496505715, value=1475130523
 U103                         column=subscn:startdt, timestamp=1532496487692, value=1465230523
 U104                         column=subscn:enddt, timestamp=1532496541815, value=1475130523
 U104                         column=subscn:startdt, timestamp=1532496523763, value=1465230523
 U105                         column=subscn:enddt, timestamp=1532496577259, value=1475130523
 U105                         column=subscn:startdt, timestamp=1532496559561, value=1465230523
 U106                         column=subscn:enddt, timestamp=1532496613239, value=1485130523
 U106                         column=subscn:startdt, timestamp=1532496595162, value=1465230523
 U107                         column=subscn:enddt, timestamp=1532496648613, value=1455130523
 U107                         column=subscn:startdt, timestamp=1532496630917, value=1465230523
 U108                         column=subscn:enddt, timestamp=1532496686037, value=1465230623
 U108                         column=subscn:startdt, timestamp=1532496668099, value=1465230523
 U109                         column=subscn:enddt, timestamp=1532496726386, value=1475130523
 U109                         column=subscn:startdt, timestamp=1532496707431, value=1465230523
 U110                         column=subscn:enddt, timestamp=1532496764298, value=1475130523
 U110                         column=subscn:startdt, timestamp=1532496745701, value=1465230523
 U111                         column=subscn:enddt, timestamp=1532496800232, value=1475130523
 U111                         column=subscn:startdt, timestamp=1532496782367, value=1465230523
 U112                         column=subscn:enddt, timestamp=1532496836644, value=1475130523
 U112                         column=subscn:startdt, timestamp=1532496818771, value=1465230523
 U113                         column=subscn:enddt, timestamp=1532496872817, value=1485130523
 U113                         column=subscn:startdt, timestamp=1532496854530, value=1465230523
 U114                         column=subscn:enddt, timestamp=1532496908617, value=1468130523
 U114                         column=subscn:startdt, timestamp=1532496890876, value=1465230523
15 row(s) in 0.2940 seconds

hbase(main):007:0>
```

Hive:

```
hive> use Project;
OK
Time taken: 10.912 seconds
hive> show tables;
OK
connected_artists
enriched_data
formatted_input
song_artist_map
station_geo_map
subscribed_users
top_10_royalty_songs
top_10_stations
top_10_unsubscribed_users
users_artists
users_behaviour
Time taken: 0.395 seconds, Fetched: 11 row(s)
hive>
```

```
hive> select * from enriched data;
OK
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1465130523    1465130523    1475130523    J     ST403  2  1  1  1  fail
U119    S201    A301    1495130523    1465130523    1465130523    E     ST409  3  1  1  1  fail
U118    S201    A301    1465230523    1475130523    1475130523    A     ST411  2  1  1  1  fail
U114    S201    A301    1475130523    1465230523    1485130523    AP    ST407  3  1  1  1  fail
U102    S201    A301    1465130523    1485130523    1465230523    NULL  ST415  1  1  1  1  fail
U102    S207    A303    1465230523    1475130523    1465130523    AP    ST402  2  1  1  1  fail
        S207    A303    1465230523    1475130523    1485130523    E     ST414  2  0  1  1  fail
        S207    A303    1465230523    1475130523    1485130523    E     ST414  2  0  1  1  fail
        S207    A303    1465230523    1475130523    1485130523    E     ST414  2  0  1  1  fail
        S207    A303    1465230523    1475130523    1485130523    E     ST414  2  0  1  1  fail
        S207    A303    1465230523    1475130523    1485130523    E     ST414  2  0  1  1  fail
        S207    A303    1465230523    1475130523    1485130523    E     ST414  2  0  1  1  fail
        S207    A303    1465230523    1475130523    1485130523    E     ST414  2  0  1  1  fail
        S207    A303    1465230523    1475130523    1485130523    E     ST414  2  0  1  1  fail
        S207    A303    1465230523    1475130523    1485130523    E     ST414  2  0  1  1  fail
        S207    A303    1465230523    1475130523    1485130523    E     ST414  2  0  1  1  fail
U119    S207    A303    1495130523    1475130523    1465130523    E     ST414  3  0  0  1  fail
        S207    A303    1465230523    1475130523    1485130523    E     ST414  2  0  1  1  fail
```

```
hive> select * from formatted_input
    > ;
OK
U113    S207    A300    1475130523      1485130523      1475130523      AP      ST401   2       1       0       1
U118    S205    A300    1495130523      1465230523      1485130523      AP      ST408   0       1       1       1
U104    S209    A301    1465130523      1475130523      1475130523      AU      ST402   3       1       1       1
U110    S208    A305    1495130523      1485130523      1465230523      AP      ST409   2       1       1       1
U108    S203    A304    1475130523      1465130523      1475130523      AP      ST406   2       1       0       1
        S210    A303    1465130523      1475130523      1485130523      AU      ST411   0       0       0       1
U118    S210    A300    1465130523      1485130523      1465230523      U       ST401   2       1       1       1
U112    S209    A304    1465230523      1485130523      1475130523      E       ST403   2       1       1       1
U104    S200    A300    1475130523      1465230523      1475130523              ST406   1       1       0       1
U111    S210            1475130523      1485130523      1465230523      U       ST413   0       0       1       1
U111    S205    A302    1475130523      1485130523      1465230523      AP      ST412   0       1       0       1
U100    S205    A304    1465130523      1465130523      1465130523      U       ST415   2       1       0       1
U120    S210    A303    1465230523      1465230523      1465130523      AU      ST406   1       1       1       1
U110    S208    A302    1465230523      1465130523      1475130523      U       ST409   0       1       1       1
U103    S210    A301    1475130523      1465230523      1475130523      AP      ST409   2       1       1       1
U111    S205    A301    1465130523      1465230523      1465130523      AU      ST405   1       0       0       1
U117    S200    A300    1465230523      1485130523      1485130523      AU      ST407   1       1       1       1
U118    S208    A300    1495130523      1465130523      1475130523      E       ST413   1       0       1       1
U110    S201    A302    1465130523      1465130523      1465130523      E       ST406   2       0       0       1
U114    S201    A305    1475130523      1465230523      1485130523      E       ST407   3       1       1       1
U117    S200    A301    1495130523      1465130523      1485130523      A       ST409   2       1       0       1
U118    S206    A302    1465130523      1475130523      1485130523      A       ST411   3       0       1       1
U102    S207    A300    1465130523      1465230523      1485130523      AP      ST403   1       0       0       1
U110    S202    A300    1495130523      1465130523      1485130523      A       ST409   0       1       0       1
U107    S202    A303    1465130523      1475130523      1465230523      U       ST414   0       0       0       1
        S200    A304    1465230523      1465130523      1465130523      AU      ST412   2       0       0       1
U102    S207    A304    1465230523      1475130523      1465130523      A       ST402   2       1       1       1
U100    S206    A301    1465230523      1485130523      1465230523      AP      ST411   1       0       0       1
U120    S209    A303    1465130523      1465130523      1465130523              ST405   3       0       0       1
U120    S207            1495130523      1465230523      1465130523      U       ST410   3       1       1       1
U111    S204    A302    1495130523      1465230523      1485130523      AU      ST402   1       0       0       1
U105    S209    A303    1495130523      1475130523      1465230523      AU      ST408   2       1       1       1
U103    S206    A304    1465230523      1465130523      1475130523      A       ST402   1       1       1       1
U114    S205    A300    1465130523      1485130523      1485130523      AU      ST414   3       0       0       1
```

```
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: 1.095 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: 0.953 seconds, Fetched: 15 row(s)
hive>
```

```
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: 0.771 seconds, Fetched: 15 row(s)
hive>
```

Data Analysis output:

```
hive> select * from top_10_royalty_songs
    > ;
OK
S208    653854132       1
S200    516700000       1
S207    359000000       1
S205    77195326        1
S203    67395326        1
S209    59900000        1
S202    50000000        1
S201    29800000        1
S204    0       1
S206    0       1
Time taken: 0.57 seconds, Fetched: 10 row(s)
hive> select * from top_10_stations;
OK
ST412   5       6       1
ST400   4       3       1
ST409   4       4       1
ST402   3       4       1
ST408   3       3       1
ST404   2       2       1
ST403   2       2       1
ST401   2       3       1
ST414   1       1       1
ST410   1       1       1
Time taken: 0.557 seconds, Fetched: 10 row(s)
hive> select * from top_10_unsubscribed_users;
OK
U101    643754132       1
U108    330000000       1
U107    328300000       1
U116    160100000       1
U114    67800468        1
U115    67295326        1
U118    59900000        1
U120    57395326        1
U117    30100000        1
```

```
hive> select * from users_behaviour;
OK
UNSUBSCRIBED    1824445252      1
SUBSCRIBED      -2043863706     1
Time taken: 0.434 seconds, Fetched: 2 row(s)
hive>
```

Exporting data from Hive to RDBMS(MySQL) using Sqoop

Now let's start exporting the data stored in hive tables to RDBMS I.e., MySQL using Sqoop.

Note: Make sure that your MySQL is up and running.

Steps:

- Let's login to MySQL database(RDBMS)
- Let's create required database and tables in MySQL.

```
[acadgild@localhost ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 17
Server version: 8.0.3-rc-log MySQL Community Server (GPL)

Copyright (c) 2000, 2017, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>
mysql> CREATE DATABASE IF NOT EXISTS project;
Query OK, 1 row affected (0.06 sec)

mysql> USE project;
Database changed
mysql> CREATE TABLE IF NOT EXISTS top_10_stations
    -> (
    -> station_id VARCHAR(50),
    -> total_distinct_songs_played INT,
    -> distinct_user_count INT
    -> );
Query OK, 0 rows affected (0.12 sec)

mysql> CREATE TABLE IF NOT EXISTS users_behaviour
    -> (
    -> user_type VARCHAR(50),
    -> duration BIGINT
    -> );
Query OK, 0 rows affected (0.10 sec)
```

```
mysql> CREATE TABLE IF NOT EXISTS users_behaviour
    -> (
    -> user_type VARCHAR(50),
    -> duration BIGINT
    -> );
Query OK, 0 rows affected (0.10 sec)

mysql> CREATE TABLE IF NOT EXISTS connected_artists
    -> (
    -> artist_id VARCHAR(50),
    -> user_count INT
    -> );
Query OK, 0 rows affected (0.04 sec)

mysql> CREATE TABLE IF NOT EXISTS top_10_royalty_songs
    -> (
    -> song_id VARCHAR(50),
    -> duration BIGINT
    -> );
Query OK, 0 rows affected (0.03 sec)

mysql> CREATE TABLE IF NOT EXISTS top_10_unsubscribed_users
    -> (
    -> user_id VARCHAR(50),
    -> duration BIGINT
    -> );
Query OK, 0 rows affected (0.03 sec)

mysql> commit;
Query OK, 0 rows affected (0.00 sec)

mysql>
```

- List of tables created:

```
mysql> show tables;
+-------------------------+
| Tables_in_project       |
+-------------------------+
| connected_artists       |
| top_10_royalty_songs    |
| top_10_stations         |
| top_10_unsubscribed_users |
| users_behaviour         |
+-------------------------+
5 rows in set (0.01 sec)

mysql>
```

Now, we have created all required databases and tables in MySQL, so now we can export the data stored in Hive to MySQL easily.

Steps to export the data from Hive to RDMBS(MySQL)

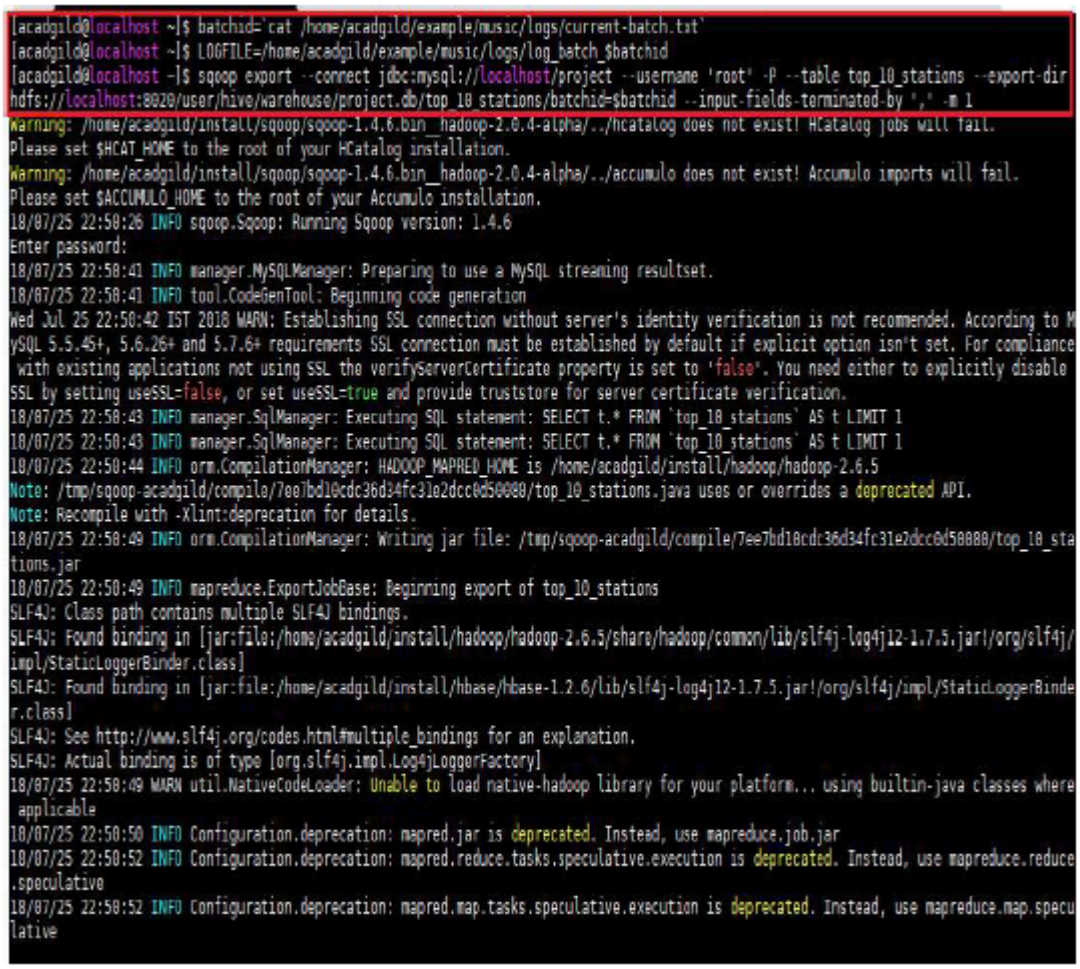
- Let's import the log file and batchid in the shell as follows:

batchid=`cat /home/acadgild/example/music/logs/current-batch.txt`

LOGFILE=/home/acadgild/example/music/logs/log_batch_$batchid

- Use the below Sqoop command to export the data stored in hive tables to MySql
  - **For top_10_stations table**

sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table top_10_stations --export-dir hdfs://localhost:8020/user/hive/warehouse/project.db/top_10_stations/batchid=$batchid --input-fields-terminated-by ',' -m 1



```
2018-07-26 12:54:24,211 INFO  [main] mapreduce.Job: Job job_1532585683668_0001 running in uber mode : false
2018-07-26 12:54:24,222 INFO  [main] mapreduce.Job:  map 0% reduce 0%
2018-07-26 12:54:38,274 INFO  [main] mapreduce.Job:  map 100% reduce 0%
2018-07-26 12:54:39,321 INFO  [main] mapreduce.Job: Job job_1532585683668_0001 completed successfully
2018-07-26 12:54:39,783 INFO  [main] mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=128262
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=306
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=10782
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=10782
                Total vcore-seconds taken by all map tasks=10782
                Total megabyte-seconds taken by all map tasks=11048768
        Map-Reduce Framework
                Map input records=9
                Map output records=9
                Input split bytes=213
                Spilled Records=0
                Failed shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=140
                CPU time spent (ms)=2300
                Physical memory (bytes) snapshot=104714240
                Virtual memory (bytes) snapshot=2061307904
                Total committed heap usage (bytes)=32571392
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
2018-07-26 12:54:39,721 INFO  [main] mapreduce.ExportJobBase: Transferred 306 bytes in 49.7357 seconds (6.1525 bytes/sec)
2018-07-26 12:54:39,734 INFO  [main] mapreduce.ExportJobBase: Exported 9 records.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost music]$
```

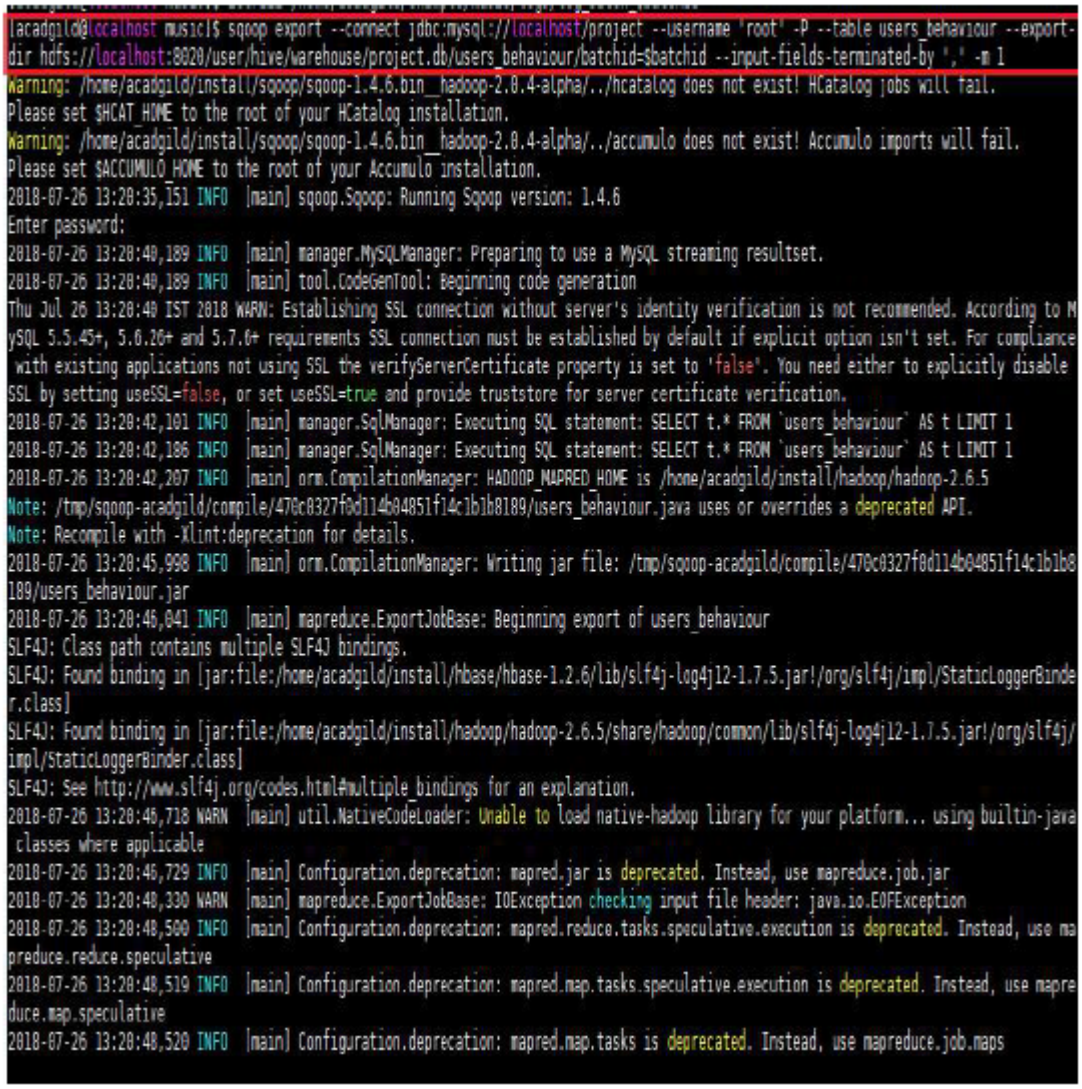
We can see that 9 records are exported from hive to RDBMS

- **Data is moved to RDBMS**

```
mysql> select * from top_10_stations;
+------------+-------------------------------+---------------------+
| station_id | total_distinct_songs_played | distinct_user_count |
+------------+-------------------------------+---------------------+
| ST411      |                             2 |                   1 |
| ST402      |                             1 |                   1 |
| ST414      |                             1 |                   1 |
| ST409      |                             1 |                   1 |
| ST410      |                             1 |                   1 |
| ST408      |                             1 |                   1 |
| ST413      |                             1 |                   1 |
| ST412      |                             1 |                   1 |
| ST406      |                             1 |                   1 |
+------------+-------------------------------+---------------------+
9 rows in set (0.07 sec)

mysql>
```

- **For Users_behaviour table:**

sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table users_behaviour --export-dir hdfs://localhost:8020/user/hive/warehouse/project.db/users_behaviour/batchid=$batch id --input-fields-terminated-by ',' -m 1


[acadgild@localhost music]$ sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table users_behaviour --export-dir hdfs://localhost:8020/user/hive/warehouse/project.db/users_behaviour/batchid=$batchid --input-fields-terminated-by ',' -m 1
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
2018-07-26 13:28:35,151 INFO [main] sqoop.Sqoop: Running Sqoop version: 1.4.6
Enter password:
2018-07-26 13:28:40,189 INFO [main] manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2018-07-26 13:28:40,189 INFO [main] tool.CodeGenTool: Beginning code generation
Thu Jul 26 13:28:40 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
2018-07-26 13:28:42,101 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `users_behaviour` AS t LIMIT 1
2018-07-26 13:28:42,186 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `users_behaviour` AS t LIMIT 1
2018-07-26 13:28:42,207 INFO [main] orm.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/install/hadoop/hadoop-2.6.5
Note: /tmp/sqoop-acadgild/compile/470c8327f8d114b04851f14c1b1b8189/users_behaviour.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2018-07-26 13:28:45,998 INFO [main] orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/470c8327f8d114b04851f14c1b1b8189/users_behaviour.jar
2018-07-26 13:28:46,041 INFO [main] mapreduce.ExportJobBase: Beginning export of users_behaviour
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.
2018-07-26 13:28:46,718 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2018-07-26 13:28:46,729 INFO [main] Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2018-07-26 13:28:48,330 WARN [main] mapreduce.ExportJobBase: IOException checking input file header: java.io.EOFException
2018-07-26 13:28:48,500 INFO [main] Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
2018-07-26 13:28:48,519 INFO [main] Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
2018-07-26 13:28:48,520 INFO [main] Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps

```
2018-07-26 13:21:44,710 INFO  [main] mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=256444
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=29978
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=602
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=2
                Other local map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=41248
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=41248
                Total vcore-seconds taken by all map tasks=41248
                Total megabyte-seconds taken by all map tasks=42237952
        Map-Reduce Framework
                Map input records=2
                Map output records=2
                Input split bytes=29928
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=401
                CPU time spent (ms)=5600
                Physical memory (bytes) snapshot=239767552
                Virtual memory (bytes) snapshot=4122615808
                Total committed heap usage (bytes)=121765888
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
2018-07-26 13:21:44,732 INFO  [main] mapreduce.ExportJobBase: Transferred 29.2754 KB in 56.1676 seconds (533.7243 bytes/sec)
```

- **Data moved to RDBMS:**

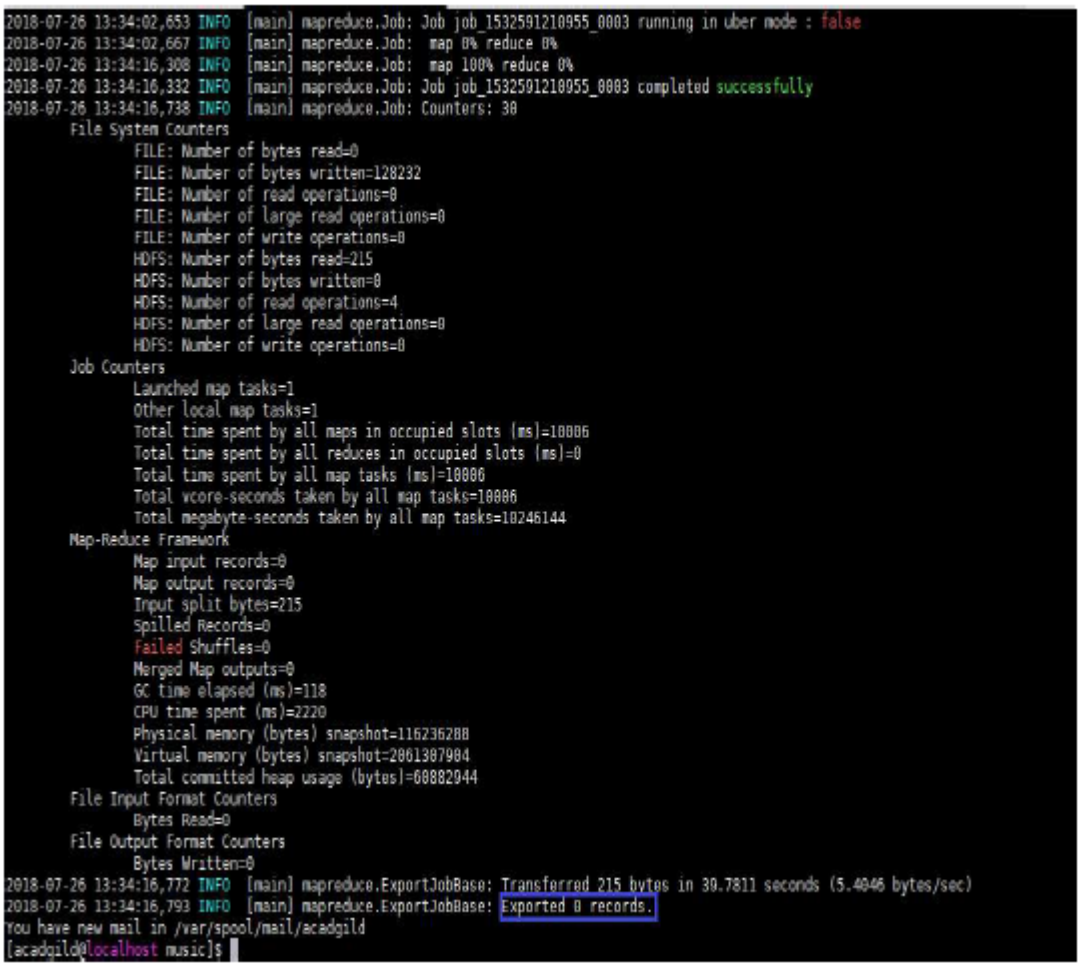
```
mysql> select * from users_behaviour
    -> ;
Empty set (0.00 sec)

mysql> select * from users_behaviour;
+-------------+----------+
| user_type   | duration |
+-------------+----------+
| UNSUBSCRIBED | 180945866 |
| SUBSCRIBED   | 120041306 |
+-------------+----------+
2 rows in set (0.00 sec)

mysql>
```

- **For connected_artists table from hive: there are no records for this table in hive as there are no connected_artists recorded.**

sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table connected_artists --export-dir hdfs://localhost:8020/user/hive/warehouse/project.db/connected_artists/batchid=$batchid --input-fields-terminated-by ',' -m 1

```
hive> select * from connected_artists;
OK
Time taken: 6.842 seconds
hive>
```


[acadgild@localhost music]$ sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table connected_artists --expo
t-dir hdfs://localhost:8020/user/hive/warehouse/project.db/connected_artists/batchid=$batchid --input-fields-terminated-by ',' -m 1
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
2018-07-26 13:33:23,375 INFO [main] sqoop.Sqoop: Running Sqoop version: 1.4.6
Enter password:
2018-07-26 13:33:28,405 INFO [main] manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2018-07-26 13:33:28,406 INFO [main] tool.CodeGenTool: Beginning code generation
Thu Jul 26 13:33:28 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to
ySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For complianc
with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable
SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
2018-07-26 13:33:30,277 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `connected_artists` AS t LIMIT 1
2018-07-26 13:33:30,364 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `connected_artists` AS t LIMIT 1
2018-07-26 13:33:30,380 INFO [main] orm.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/install/hadoop/hadoop-2.6.5
Note: /tmp/sqoop-acadgild/compile/72081881eb9bee05efd1734d8c168cb7/connected_artists.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2018-07-26 13:33:34,386 INFO [main] orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/72081881eb9bee05efd1734d8c16
cb7/connected_artists.jar
2018-07-26 13:33:34,410 INFO [main] mapreduce.ExportJobBase: Beginning export of connected_artists
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/StaticLoggerBind
r.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.
2018-07-26 13:33:34,933 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
classes where applicable
2018-07-26 13:33:34,946 INFO [main] Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2018-07-26 13:33:36,737 WARN [main] mapreduce.ExportJobBase: IOException checking input file header: java.io.EOFException
2018-07-26 13:33:36,910 INFO [main] Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use m
preduce.reduce.speculative
2018-07-26 13:33:36,926 INFO [main] Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapr
duce.map.speculative
2018-07-26 13:33:36,930 INFO [main] Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps



We can see that 0 records are exported from hive to RDBMS

- **Data in RDBMS**

- **For top_10_royalty_songs table:**

sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table top_10_royalty_songs --export-dir hdfs://localhost:8020/user/hive/warehouse/project.db/top_10_royalty_songs/batchid=$batchid --input-fields-terminated-by ',' -m 1

```
[acadgild@localhost music]$ sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table top_10_royalty_songs --ex
port-dir hdfs://localhost:8020/user/hive/warehouse/project.db/top_10_royalty_songs/batchid=$batchid --input-fields-terminated-by ',' -m
1
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
2018-07-26 13:40:09,063 INFO  [main] sqoop.Sqoop: Running Sqoop version: 1.4.6
Enter password:
2018-07-26 13:40:13,090 INFO  [main] manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2018-07-26 13:40:13,093 INFO  [main] tool.CodeGenTool: Beginning code generation
Thu Jul 26 13:40:13 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to M
ySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance
 with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable
SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
2018-07-26 13:40:14,957 INFO  [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `top_10_royalty_songs` AS t LIMIT 1
2018-07-26 13:40:15,044 INFO  [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `top_10_royalty_songs` AS t LIMIT 1
2018-07-26 13:40:15,064 INFO  [main] orm.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/install/hadoop/hadoop-2.6.5
Note: /tmp/sqoop-acadgild/compile/81ad140ba64046e68829c29f63528b44/top_10_royalty_songs.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2018-07-26 13:40:18,910 INFO  [main] orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/81ad140ba64046e68829c29f63528
b44/top_10_royalty_songs.jar
2018-07-26 13:40:18,952 INFO  [main] mapreduce.ExportJobBase: Beginning export of top_10_royalty_songs
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/StaticLoggerBinde
r.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.
2018-07-26 13:40:19,598 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
 classes where applicable
2018-07-26 13:40:19,529 INFO  [main] Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2018-07-26 13:40:21,509 INFO  [main] Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use ma
preduce.reduce.speculative
2018-07-26 13:40:21,525 INFO  [main] Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapre
duce.map.speculative
2018-07-26 13:40:21,527 INFO  [main] Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2018-07-26 13:40:21,796 INFO  [main] client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
```

```
2018-07-26 13:40:43,132 INFO  [main] mapreduce.Job: Job job_1532591210955_0004 running in uber mode : false
2018-07-26 13:40:43,137 INFO  [main] mapreduce.Job:  map 0% reduce 0%
2018-07-26 13:40:56,192 INFO  [main] mapreduce.Job:  map 100% reduce 0%
2018-07-26 13:40:57,224 INFO  [main] mapreduce.Job: Job job_1532591210955_0004 completed successfully
2018-07-26 13:40:57,674 INFO  [main] mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=128240
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=317
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=9414
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=9414
                Total vcore-seconds taken by all map tasks=9414
                Total megabyte-seconds taken by all map tasks=9639936
        Map-Reduce Framework
                Map input records=7
                Map output records=7
                Input split bytes=218
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=120
                CPU time spent (ms)=2150
                Physical memory (bytes) snapshot=117022720
                Virtual memory (bytes) snapshot=2061307904
                Total committed heap usage (bytes)=60882944
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
2018-07-26 13:40:57,700 INFO  [main] mapreduce.ExportJobBase: Transferred 317 bytes in 36.1206 seconds (8.7761 bytes/sec)
2018-07-26 13:40:57,713 INFO  [main] mapreduce.ExportJobBase: Exported 7 records.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost music]$
```

We can see that 7 records are exported from hive to RDBMS

- **Data exported to RDBMS**

```
mysql> select * from top_10_royalty_songs;
Empty set (0.00 sec)

mysql> select * from top_10_royalty_songs;
+---------+----------+
| song_id | duration |
+---------+----------+
| S201    | 38807006 |
| S207    | 28807006 |
| S208    | 15131627 |
| S209    | 10000000 |
| S203    | 10000000 |
| S202    |  5231627 |
| S206    |  2004333 |
+---------+----------+
7 rows in set (0.00 sec)
```

- **For top_10_unsibscribed_users table**

sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table top_10_unsubscribed_users --export-dir hdfs://localhost:8020/user/hive/warehouse/project.db/top_10_unsubscribed_users/batchid=$batchid --input-fields-terminated-by ',' -m 1

```
[acadgild@localhost music]$ sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table top_10_unsubscribed_users
 --export-dir hdfs://localhost:8020/user/hive/warehouse/project.db/top_10_unsubscribed_users/batchid=$batchid --input-fields-terminated-
by ',' -m 1
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
2018-07-26 13:45:25,327 INFO  [main] sqoop.Sqoop: Running Sqoop version: 1.4.6
Enter password:
2018-07-26 13:45:31,086 INFO  [main] manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2018-07-26 13:45:31,087 INFO  [main] tool.CodeGenTool: Beginning code generation
Thu Jul 26 13:45:31 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to M
ySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance
 with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable
SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
2018-07-26 13:45:33,035 INFO  [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `top_10_unsubscribed_users` AS t LIMIT
 1
2018-07-26 13:45:33,115 INFO  [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `top_10_unsubscribed_users` AS t LIMIT
 1
2018-07-26 13:45:33,130 INFO  [main] orm.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/install/hadoop/hadoop-2.6.5
Note: /tmp/sqoop-acadgild/compile/751c5dd20c8926bf147fca00453b202a/top_10_unsubscribed_users.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2018-07-26 13:45:36,898 INFO  [main] orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/751c5dd20c8926bf147fca00453b2
02a/top_10_unsubscribed_users.jar
2018-07-26 13:45:36,916 INFO  [main] mapreduce.ExportJobBase: Beginning export of top_10_unsubscribed_users
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/StaticLoggerBinde
r.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.
2018-07-26 13:45:37,509 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
 classes where applicable
2018-07-26 13:45:37,528 INFO  [main] Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2018-07-26 13:45:39,606 INFO  [main] Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use ma
preduce.reduce.speculative
2018-07-26 13:45:39,623 INFO  [main] Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapre
```

```
2018-07-26 13:46:08,108 INFO  [main] mapreduce.Job: Job job_1532501210955_0005 running in uber mode : false
2018-07-26 13:46:08,113 INFO  [main] mapreduce.Job:  map 0% reduce 0%
2018-07-26 13:46:11,950 INFO  [main] mapreduce.Job:  map 100% reduce 0%
2018-07-26 13:46:13,003 INFO  [main] mapreduce.Job: Job job_1532501210955_0005 completed successfully
2018-07-26 13:46:13,315 INFO  [main] mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=128260
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=329
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=8663
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=8663
                Total vcore-seconds taken by all map tasks=8663
                Total megabyte-seconds taken by all map tasks=8870912
        Map-Reduce Framework
                Map input records=8
                Map output records=8
                Input split bytes=223
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=99
                CPU time spent (ms)=2010
                Physical memory (bytes) snapshot=115450048
                Virtual memory (bytes) snapshot=2061307904
                Total committed heap usage (bytes)=60002944
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
2018-07-26 13:46:13,351 INFO  [main] mapreduce.ExportJobBase: Transferred 329 bytes in 33.6729 seconds (9.7705 bytes/sec)
2018-07-26 13:46:13,359 INFO  [main] mapreduce.ExportJobBase: Exported 8 records.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost music]$
```

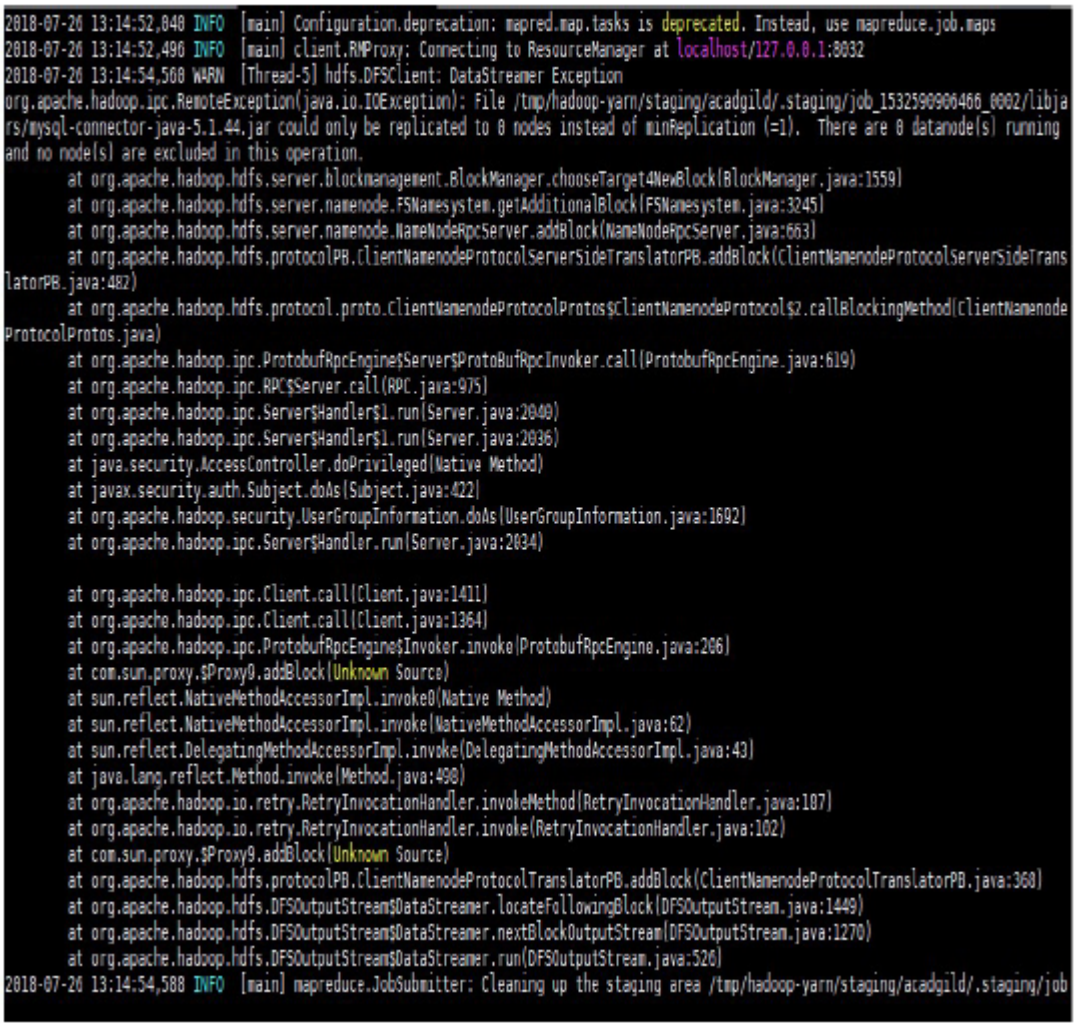
We can see that 8 records are exported from hive to RDBMS

- **Data is exported to RDBMS**

```
mysql> select * from top_10_unsubscribed_users;
Empty set (0.00 sec)

mysql> select * from top_10_unsubscribed_users;
+---------+----------+
| user_id | duration |
+---------+----------+
| U115    | 51434300 |
| U120    | 44038633 |
| U114    | 38807006 |
| U111    | 28887006 |
| U117    | 10008008 |
| U100    |  5231627 |
| U108    |  2627294 |
| U113    |        0 |
+---------+----------+
8 rows in set (0.00 sec)

mysql>
```

Solution to resolve the errors faced during project execution:

- **Had faced this error has shown below:**



```
2018-07-26 13:14:52,848 INFO  [main] Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2018-07-26 13:14:52,496 INFO  [main] client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-07-26 13:14:54,568 WARN  [Thread-5] hdfs.DFSClient: DataStreamer Exception
org.apache.hadoop.ipc.RemoteException(java.io.IOException): File /tmp/hadoop-yarn/staging/acadgild/.staging/job_1532590906466_0002/libja
rs/mysql-connector-java-5.1.44.jar could only be replicated to 0 nodes instead of minReplication (=1).  There are 0 datanode(s) running
and no node(s) are excluded in this operation.
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.chooseTarget4NewBlock(BlockManager.java:1559)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getAdditionalBlock(FSNamesystem.java:3245)
        at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.addBlock(NameNodeRpcServer.java:663)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.addBlock(ClientNamenodeProtocolServerSideTrans
latorPB.java:482)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingMethod(ClientNamenode
ProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:619)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:975)
        at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2040)
        at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2036)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1692)
        at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2034)

        at org.apache.hadoop.ipc.Client.call(Client.java:1411)
        at org.apache.hadoop.ipc.Client.call(Client.java:1364)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:206)
        at com.sun.proxy.$Proxy9.addBlock(Unknown Source)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:187)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:102)
        at com.sun.proxy.$Proxy9.addBlock(Unknown Source)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.addBlock(ClientNamenodeProtocolTranslatorPB.java:368)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.locateFollowingBlock(DFSOutputStream.java:1449)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.nextBlockOutputStream(DFSOutputStream.java:1270)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:526)
2018-07-26 13:14:54,588 INFO  [main] mapreduce.JobSubmitter: Cleaning up the staging area /tmp/hadoop-yarn/staging/acadgild/.staging/job
```

The above error was because of corrupted HBase, initially HBase looked fine, but when I executed the project, I got this error, I checked if HBase and Hive are working properly or not. Then when I tried "list" command in HBase, then I realized that HBase is corrupted.As shown in below screenshot.

```
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hbase shell
2018-07-26 13:12:06,612 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
 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/StaticLoggerBind
r.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

hbase(main):001:0>
hbase(main):001:0> list
TABLE

ERROR: Can't get master address from ZooKeeper; znode data == null

Here is some help for this command:
List all tables in hbase. Optional regular expression parameter could
be used to filter the output. Examples:

  hbase> list
  hbase> list 'abc.*'
  hbase> list 'ns:abc.*'
  hbase> list 'ns:.*'
```

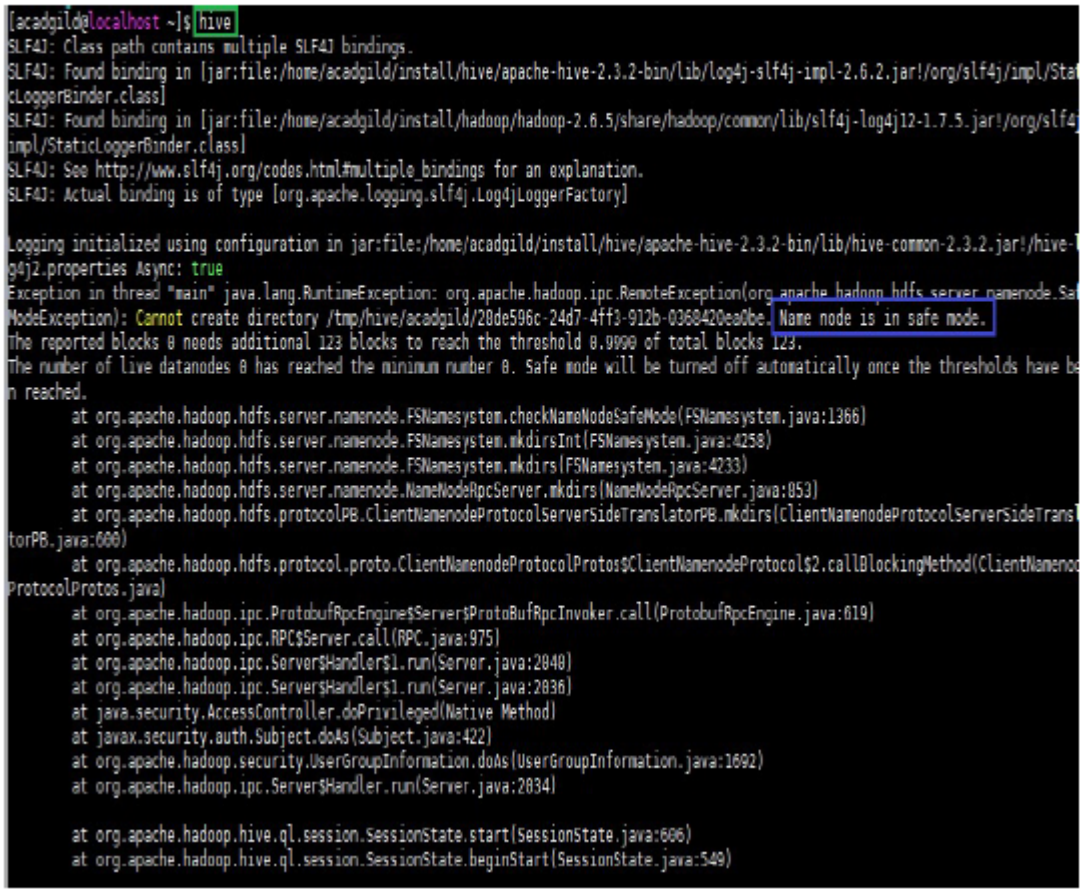
Error was because of Hbase, which was corrupted

Solution: I restarted HBase to resolve the above error

```
[acadgild@localhost ~]$ stop-hbase.sh
stopping hbasecat: /tmp/hbase-acadgild-master.pid: No such file or directory

localhost: stopping zookeeper.
[acadgild@localhost ~]$ start-hbase.sh
localhost: starting zookeeper, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-zookeeper-localhost.localdomain.
ut
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 ~]$
[acadgild@localhost ~]$ hbase shell
2018-07-26 13:18:32,558 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
 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/StaticLoggerBind
r.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

hbase(main):001:0> list
TABLE
song-artist-map
station-geo-map
subscribed-users
3 row(s) in 0.6470 seconds

=> ["song-artist-map", "station-geo-map", "subscribed-users"]
hbase(main):002:0>
```

Restart Hbase

- While starting hive, name node was in safe mode. Thereby, hive shell was not up and running.

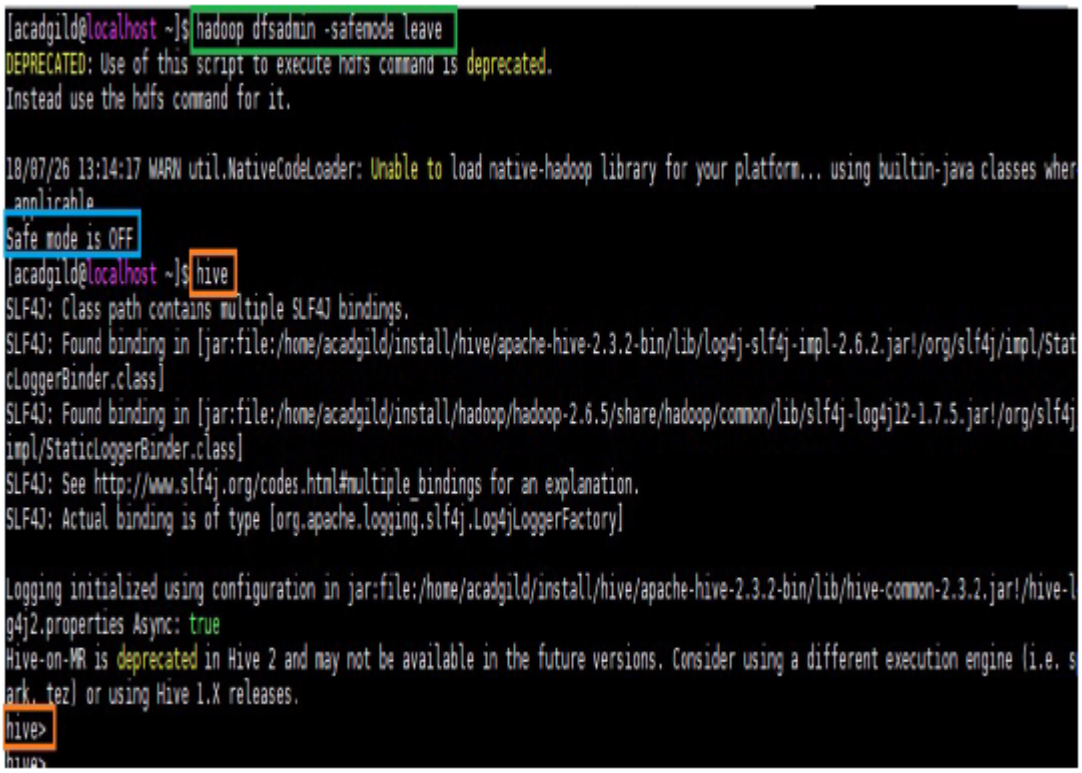


Solution:

I. Use this command to remove the name node from safe mode .

hadoop dfsadmin -safemode leave

II. Restart Hive shell


[acadgild@localhost ~]$ hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
18/07/26 13:14:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wher
applicable
Safe mode is OFF
[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/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/Stat
cLoggerBinder.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-l
g4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. s
ark, tez) or using Hive 1.X releases.
hive>