Final Project Music Data Analysis

1st Stage: Creating the data using python scripts

Data come from two sources, web in xml format and mob in csv format.

Used python scripts to generate data

generate_web_data.py -- Generates some random data coming from web application

python /home/acadgild/project/scripts/generate_web_data.py

generate_mob_data.py -- Generates some random data coming from mobile application

python /home/acadgild/project/scripts/generate_mob_data.py

Data files screenshot

```
[acadgild@localhost project]$ cd data
[acadgild@localhost data]$ ll
total 8
drwxrwxr-x. 2 acadgild acadgild 4096 Jan 22 21:43 mob
drwxrwxr-x. 2 acadgild acadgild 4096 Jan 22 21:43 web
[acadgild@localhost data]$ pwd
/home/acadgild/project/data
[acadgild@localhost data]$ cd mob
[acadgild@localhost mob]$ ll
total 4
-rw-rw-r--. 1 acadgild acadgild 1236 Jan 22 21:43 file.txt
[acadgild@localhost mob]$ cat file.txt
U117,S206,A305,1465130523,1475130523,1465230523,U,ST401,1,1,0
U111,S210,A302,1495130523,1475130523,1465230523,AU,ST406,0,0,1
U101,S205,A301,1495130523,1475130523,1485130523,AU,ST413,2,0,1
U104,S204,A304,1495130523,1485130523,1475130523,U,ST400,0,1,0
U101,S206,A305,1495130523,1465130523,1465130523,U,ST415,0,1,0
,S202,A300,1475130523,1475130523,1485130523,E,ST400,0,1,1
U120,S200,A302,1465230523,1465130523,1465130523,A,ST400,0,0,0
U112,S201,A301,1465230523,1465230523,1475130523,A,ST403,3,0,0
U114,S202,A300,1465130523,1475130523,1465230523,,ST405,3,0,0
U120,S209,,1465130523,1465130523,1485130523,A,ST401,0,0,0
U118,S205,A305,1475130523,1485130523,1485130523,U,ST409,2,0,1
U104,S200,A300,1465130523,1485130523,1485130523,A,ST406,2,0,0
U106,S205,A305,1465230523,1465230523,1475130523,A,ST413,3,1,0
U107,S208,A304,1465130523,1465230523,1465130523,E,ST411,1,1,0
U117,S210,A303,1465130523,1465230523,1465130523,AP,ST414,2,0,0
U111,S203,A303,1495130523,1465130523,1475130523,A,ST404,0,1,0
U108,S201,A304,1465130523,1485130523,1465130523,AU,ST400,1,1,1
U115,S200,A304,1465230523,1475130523,1475130523,E,ST414,0,1,0
U109,S201,A300,1475130523,1465230523,1465130523,U,ST415,3,0,1
U118,S205,A302,1475130523,1485130523,1465230523,AP,ST401,3,1,1
[acadgild@localhost mob]$ cd ..
[acadgild@localhost data]$ cd web
[acadgild@localhost web]$ ll
total 8
-rw-rw-r--. 1 acadgild acadgild 6724 Jan 22 21:43 file.xml
```

2nd stage: Starting required services for the project Using start-daemons.sh

→ Starting below services

start-all.sh(starts hadoop)

start-hbase.sh(starts hbase)

mr-jobhistory-daemon.sh start historyserver (starts history server)

sql (starts sql for exporting data from hive to mysql)

Created a batch file to get the records for every 3 hours and log file for tracking. Batchid is present in every script for iteration.

Daemons and process screenshot

```
[acadgild@localhost scripts]$ jps
9666 SecondaryNameNode
9923 NodeManager
8803 HRegionServer
9317 NameNode
15977 RunJar
9419 DataNode
8588 HQuorumPeer
9149 JobHistoryServer
9821 ResourceManager
17823 HMaster
11167 Jps
```

3rd Stage: Populate lookup data using populate-lookup.sh

→ Populating data

We have 4 look up files in which 3 files will be loaded in hbase and the other one file will be loaded int hive. Table creation and schema and loading are there in populate-lookup.sh and user-artist.hql

From the look files(stn-geocd.txt, song-artist.txt, user-subscn.txt), load the data in hbase in below respective tables

song-artist-map

station-geo-map

subscribed-users

Screen shot showing how to create one of the hbase table and loading the lookupfile.

```
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

file="/home/acadgild/project/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"
```

Hbase screen shot showing the data from the lookup files.

```
song-artist-map
station-geo-map
subscribed-users
10 row(s) in 0.0600 seconds

=> ["SparkHBasesTable", "TRANSACTIONS", "bulktable", "click", "clicks", "employee", "htest", "song-artist-map", "station-geo-map", "subscribed-users"]
hbase(main):013:0> scan 'song-artist-map'
ROW                                    COLUMN+CELL
 S200                                  column=artist:artistid, timestamp=1548174593721, value=A300
 S201                                  column=artist:artistid, timestamp=1548174618586, value=A301
 S202                                  column=artist:artistid, timestamp=1548174642583, value=A302
 S203                                  column=artist:artistid, timestamp=1548174667220, value=A303
 S204                                  column=artist:artistid, timestamp=1548174693228, value=A304
 S205                                  column=artist:artistid, timestamp=1548174721320, value=A301
 S206                                  column=artist:artistid, timestamp=1548174748125, value=A302
 S207                                  column=artist:artistid, timestamp=1548174775322, value=A303
 S208                                  column=artist:artistid, timestamp=1548174801861, value=A304
 S209                                  column=artist:artistid, timestamp=1548174825222, value=A305
10 row(s) in 0.0620 seconds

hbase(main):014:0> scan 'station-geo-map'
ROW                                    COLUMN+CELL
 ST400                                 column=geo:geo_cd, timestamp=1548174205605, value=A
 ST401                                 column=geo:geo_cd, timestamp=1548174230421, value=AU
 ST402                                 column=geo:geo_cd, timestamp=1548174253372, value=AP
 ST403                                 column=geo:geo_cd, timestamp=1548174277451, value=J
 ST404                                 column=geo:geo_cd, timestamp=1548174300827, value=E
 ST405                                 column=geo:geo_cd, timestamp=1548174325442, value=A
 ST406                                 column=geo:geo_cd, timestamp=1548174351873, value=AU
 ST407                                 column=geo:geo_cd, timestamp=1548174378439, value=AP
 ST408                                 column=geo:geo_cd, timestamp=1548174405278, value=E
 ST409                                 column=geo:geo_cd, timestamp=1548174431851, value=E
 ST410                                 column=geo:geo_cd, timestamp=1548174457615, value=A
 ST411                                 column=geo:geo_cd, timestamp=1548174485131, value=A
 ST412                                 column=geo:geo_cd, timestamp=1548174508795, value=AP
 ST413                                 column=geo:geo_cd, timestamp=1548174536392, value=J
 ST414                                 column=geo:geo_cd, timestamp=1548174564968, value=E
15 row(s) in 0.0520 seconds
```

```
hbase(main):015:0> scan 'subscribed-users'
ROW                                    COLUMN+CELL
 U100                                  column=subscn:enddt, timestamp=1548174873114, value=1465130523
 U100                                  column=subscn:startdt, timestamp=1548174849503, value=1465230523
 U101                                  column=subscn:enddt, timestamp=1548174923622, value=1475130523
 U101                                  column=subscn:startdt, timestamp=1548174896861, value=1465230523
 U102                                  column=subscn:enddt, timestamp=1548174974055, value=1475130523
 U102                                  column=subscn:startdt, timestamp=1548174947615, value=1465230523
 U103                                  column=subscn:enddt, timestamp=1548175025844, value=1475130523
 U103                                  column=subscn:startdt, timestamp=1548174999533, value=1465230523
 U104                                  column=subscn:enddt, timestamp=1548175078972, value=1475130523
 U104                                  column=subscn:startdt, timestamp=1548175051880, value=1465230523
 U105                                  column=subscn:enddt, timestamp=1548175128405, value=1475130523
 U105                                  column=subscn:startdt, timestamp=1548175102595, value=1465230523
 U106                                  column=subscn:enddt, timestamp=1548175180994, value=1485130523
 U106                                  column=subscn:startdt, timestamp=1548175152541, value=1465230523
 U107                                  column=subscn:enddt, timestamp=1548175237391, value=1455130523
 U107                                  column=subscn:startdt, timestamp=1548175209495, value=1465230523
 U108                                  column=subscn:enddt, timestamp=1548175284272, value=1465230623
 U108                                  column=subscn:startdt, timestamp=1548175260098, value=1465230523
 U109                                  column=subscn:enddt, timestamp=1548175336225, value=1475130523
 U109                                  column=subscn:startdt, timestamp=1548175309120, value=1465230523
 U110                                  column=subscn:enddt, timestamp=1548175385387, value=1475130523
 U110                                  column=subscn:startdt, timestamp=1548175361684, value=1465230523
 U111                                  column=subscn:enddt, timestamp=1548145834492, value=1475130523
 U111                                  column=subscn:startdt, timestamp=1548175409699, value=1465230523
 U112                                  column=subscn:enddt, timestamp=1548145891631, value=1475130523
 U112                                  column=subscn:startdt, timestamp=1548145862922, value=1465230523
 U113                                  column=subscn:enddt, timestamp=1548145947583, value=1485130523
 U113                                  column=subscn:startdt, timestamp=1548145919657, value=1465230523
 U114                                  column=subscn:enddt, timestamp=1548146003549, value=1468130523
 U114                                  column=subscn:startdt, timestamp=1548145975690, value=1465230523
15 row(s) in 0.1810 seconds
```

→*Create table user-artists in hive and loading user-artist.txt file into hive*

```
[acadgild@localhost scripts]$ cat /home/acadgild/project/scripts/user-artist.hql
CREATE DATABASE IF NOT EXISTS project;

USE project;

CREATE TABLE users_artists
(
user_id STRING,
artists_array ARRAY<STRING>
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
COLLECTION ITEMS TERMINATED BY '&';

LOAD DATA LOCAL INPATH '/home/acadgild/project/lookupfiles/user-artist.txt'
OVERWRITE INTO TABLE users_artists;
```

```
hive> show tables;
OK
song_artist_map
station_geo_map
subscribed_users
users_artists
Time taken: 0.178 seconds, Fetched: 4 row(s)
hive> select * from users_artists;
OK
U100    ["A300","A301","A302"]
U101    ["A301","A302"]
U102    ["A302"]
U103    ["A303","A301","A302"]
U104    ["A304","A301"]
U105    ["A305","A301","A302"]
U106    ["A301","A302"]
U107    ["A302"]
U108    ["A300","A303","A304"]
U109    ["A301","A303"]
U110    ["A302","A301"]
U111    ["A303","A301"]
U112    ["A304","A301"]
U113    ["A305","A302"]
U114    ["A300","A301","A302"]
Time taken: 1.034 seconds, Fetched: 15 row(s)
hive>
```

4th stage: Formatting the data using dataformatting.sh

This script contains dataformatting.pig and hive scripts.

Dataformatting.pig converts the .xml file that is generated from web source into a csv format.

Other file from mobile will be in csv format so no need to change that.

Pig script screenshot

```
[acadgild@localhost scripts]$ cat /home/acadgild/project/scripts/dataformatting.pig
REGISTER /home/acadgild/project/lib/piggybank.jar;

DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();

A =  LOAD '/user/acadgild/project/batch${batchid}/web/' using org.apache.pig.piggybank.storage.XMLLoader('record') as (x:chararray);

B = FOREACH A GENERATE TRIM(XPath(x, 'record/user_id')) AS user_id,
    TRIM(XPath(x, 'record/song_id')) AS song_id,
    TRIM(XPath(x, 'record/artist_id')) AS artist_id,
    ToUnixTime(ToDate(TRIM(XPath(x, 'record/timestamp')),'yyyy-MM-dd HH:mm:ss')) AS timestamp,
    ToUnixTime(ToDate(TRIM(XPath(x, 'record/start_ts')),'yyyy-MM-dd HH:mm:ss')) AS start_ts,
    ToUnixTime(ToDate(TRIM(XPath(x, 'record/end_ts')),'yyyy-MM-dd HH:mm:ss')) AS end_ts,
    TRIM(XPath(x, 'record/geo_cd')) AS geo_cd,
    TRIM(XPath(x, 'record/station_id')) AS station_id,
    TRIM(XPath(x, 'record/song_end_type')) AS song_end_type,
    TRIM(XPath(x, 'record/like')) AS like,
    TRIM(XPath(x, 'record/dislike')) AS dislike;

STORE B INTO '/user/acadgild/project/batch${batchid}/formattedweb/' USING PigStorage(',');
```

Screen shot show the formattedweb(file generated from web) and mob(file generated from mobile) directory which contain two files in respective folder

```
[acadgild@localhost ~]$ hadoop dfs -ls /user/acadgild/project/batch1
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/01/27 17:18:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
drwxr-xr-x   - acadgild supergroup          0 2019-01-27 17:17 /user/acadgild/project/batch1/formattedweb
drwxr-xr-x   - acadgild supergroup          0 2019-01-27 17:16 /user/acadgild/project/batch1/mob
drwxr-xr-x   - acadgild supergroup          0 2019-01-27 17:15 /user/acadgild/project/batch1/web
[acadgild@localhost ~]$ hadoop dfs -ls /user/acadgild/project/batch1/web
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/01/27 17:18:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup       6724 2019-01-27 17:15 /user/acadgild/project/batch1/web/file.xml
[acadgild@localhost ~]$ hadoop dfs -ls /user/acadgild/project/batch1/mob
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/01/27 17:19:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup       1236 2019-01-27 17:16 /user/acadgild/project/batch1/mob/file.txt
-rw-r--r--   1 acadgild supergroup       1844 2019-01-27 17:16 /user/acadgild/project/batch1/mob/pig_1548187765226.log
[acadgild@localhost ~]$ hadoop dfs -ls /user/acadgild/project/batch1/formattedweb
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/01/27 17:19:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2019-01-27 17:17 /user/acadgild/project/batch1/formattedweb/_SUCCESS
-rw-r--r--   1 acadgild supergroup       1244 2019-01-27 17:17 /user/acadgild/project/batch1/formattedweb/part-m-00000
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

The formatted_hive_load.hql script will help in data loading from the two files in to one table called formatted_input.

```
[acadgild@localhost scripts]$ cat /home/acadgild/project/scripts/formatted_hive_load.hql
USE project;

CREATE TABLE IF NOT EXISTS formatted_input
(
User_id STRING,
Song_id STRING,
Artist_id STRING,
Time_stamp STRING,
Start_ts STRING,
End_ts STRING,
Geo_cd STRING,
Station_id STRING,
Song_end_type INT,
Likes INT,
Dislikes INT
)
PARTITIONED BY
(batchid INT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';

LOAD DATA INPATH '/user/acadgild/project/batch${hiveconf:batchid}/formattedweb/'
INTO TABLE formatted_input PARTITION (batchid=${hiveconf:batchid});

LOAD DATA INPATH '/user/acadgild/project/batch${hiveconf:batchid}/mob/'
INTO TABLE formatted_input PARTITION (batchid=${hiveconf:batchid});
```

```
hive> select * from formatted_input;
OK
U117    S206    A305    1465130523      1475130523      1465230523      U       ST401   1       1       0       1
U111    S210    A302    1495130523      1475130523      1465230523      AU      ST406   0       0       1       1
U101    S205    A301    1495130523      1475130523      1485130523      AU      ST413   2       0       1       1
U104    S204    A304    1495130523      1485130523      1475130523      U       ST400   0       1       0       1
U101    S206    A305    1495130523      1465130523      1465130523      U       ST415   0       1       0       1
        S202    A300    1475130523      1475130523      1485130523      E       ST400   0       1       1       1
U120    S200    A302    1465230523      1465130523      1465130523      A       ST400   0       0       0       1
U112    S201    A301    1465230523      1465230523      1475130523      A       ST403   3       0       0       1
U114    S202    A300    1465130523      1475130523      1465230523              ST405   3       0       0       1
U120    S209            1465130523      1465130523      1485130523      A       ST401   0       0       0       1
U118    S205    A305    1475130523      1485130523      1485130523      U       ST409   2       0       1       1
U104    S200    A300    1465130523      1485130523      1485130523      A       ST406   2       0       0       1
U106    S205    A305    1465230523      1465230523      1475130523      A       ST413   3       1       0       1
U107    S208    A304    1465130523      1465230523      1465130523      E       ST411   1       1       0       1
U117    S210    A303    1465130523      1465230523      1465130523      AP      ST414   2       0       0       1
U111    S203    A303    1495130523      1465130523      1475130523      A       ST404   0       1       0       1
U108    S201    A304    1465130523      1485130523      1465130523      AU      ST400   1       1       1       1
U115    S200    A304    1465230523      1475130523      1475130523      E       ST414   0       1       0       1
U109    S201    A300    1475130523      1465230523      1465130523      U       ST415   3       0       1       1
U118    S205    A302    1475130523      1485130523      1465230523      AP      ST401   3       1       1       1
U109    S204    A300    1465490556      1462863262      1494297562      AP      ST401   0       1       0       1
U112    S206    A303    1494297562      1465490556      1468094889      AU      ST403   0       1       1       1
U101    S205    A304    1462863262      1494297562      1468094889      AU      ST414   0       0       0       1
U107    S201    A302    1462863262      1462863262      1462863262      AU      ST413   0       0       1       1
U110    S206    A302    1494297562      1465490556      1468094889      AU      ST401   0       0       0       1
        S208    A301    1465490556      1465490556      1462863262      AP      ST405   0       0       0       1
U111    S208    A302    1494297562      1494297562      1462863262      A       ST408   2       0       0       1
U111    S205    A300    1494297562      1465490556      1468094889      AP      ST403   3       0       0       1
U111    S205    A300    1468094889      1468094889      1468094889              ST409   2       0       0       1
U103    S204            1465490556      1465490556      1462863262      AU      ST405   2       0       1       1
U119    S209    A301    1494297562      1462863262      1494297562      AP      ST403   3       0       0       1
U103    S208    A305    1494297562      1468094889      1465490556      A       ST415   0       0       1       1
U116    S206    A302    1494297562      1494297562      1465490556      E       ST414   3       0       0       1
U118    S203    A303    1494297562      1465490556      1468094889      AP      ST404   0       1       0       1
U117    S203    A302    1465490556      1468094889      1462863262      AU      ST410   0       0       1       1
U117    S201    A303    1462863262      1494297562      1462863262      AP      ST415   3       0       1       1
U113    S204    A301    1462863262      1462863262      1462863262      E       ST412   1       0       1       1
U102    S206    A305    1462863262      1494297562      1468094889      A       ST410   1       0       0       1
U116    S205    A303    1494297562      1462863262      1468094889      U       ST400   1       0       1       1
```

Output screenshots of dataformatting.sh

```
[acadgild@localhost scripts]$ sh dataformatting.sh
19/01/23 00:24:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/23 00:24:02 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/acadgild/project/batch1/web
19/01/23 00:24:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
rm: `/user/acadgild/project/batch1/formattedweb/': No such file or directory
19/01/23 00:24:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/23 00:24:15 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/acadgild/project/batch1/mob
19/01/23 00:24:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/23 00:24:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/23 00:24:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/23 00:24:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/23 00:24:46 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
19/01/23 00:24:46 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
19/01/23 00:24:46 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2019-01-23 00:24:46,998 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2019-01-23 00:24:46,999 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/project/scripts/pig_1548183286995.log
SLF4J: Class path contains multiple SLF4J bindings.
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: 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: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2019-01-23 00:24:48,726 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes w
ere applicable
2019-01-23 00:24:49,533 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2019-01-23 00:24:50,115 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2019-01-23 00:24:50,115 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2019-01-23 00:24:50,116 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:8020
2019-01-23 00:24:51,957 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-dataformatting.pig-53eba90c-6558-45a0-8131-f9e29ba44eae
2019-01-23 00:24:51,957 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
```

```
2019-01-23 00:25:44,490 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to
 job history server
2019-01-23 00:25:44,646 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2019-01-23 00:25:44,652 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2019-01-23 00:25:44,668 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to
 job history server
2019-01-23 00:25:44,920 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2019-01-23 00:25:44,923 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.5   0.16.0  acadgild        2019-01-23 00:24:56     2019-01-23 00:25:44     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        AliasF
eature  Outputs
job_1548141993219_0003  1       0       14      14      14      14      0       0       0       0       A,B     MAP_ONLY        /user/acadgild/project/batch1/formatte
dweb,

Input(s):
Successfully read 20 records (7114 bytes) from: "/user/acadgild/project/batch1/web"

Output(s):
Successfully stored 20 records (1244 bytes) in: "/user/acadgild/project/batch1/formattedweb"

Counters:
Total records written : 20
Total bytes written : 1244
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1548141993219_0003


2019-01-23 00:25:44,929 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2019-01-23 00:25:44,938 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to
 job history server
2019-01-23 00:25:45,081 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.0.1:8032
2019-01-23 00:25:45,101 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to
 job history server
```

```
2019-01-23 00:25:45,235 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting t
 job history server
2019-01-23 00:25:45,378 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2019-01-23 00:25:45,476 [main] INFO  org.apache.pig.Main - Pig script completed in 59 seconds and 552 milliseconds (59552 ms)
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
OK
Time taken: 20.812 seconds
OK
Time taken: 1.068 seconds
Loading data to table project.formatted_input partition (batchid=1)
OK
Time taken: 6.413 seconds
Loading data to table project.formatted_input partition (batchid=1)
OK
Time taken: 3.054 seconds
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost scripts]$
```

5th stage: Replication the hbase data into hive

Using create_hive_hbase_lookup.hql we are creating three external tables like below that are created similar in hbase.

song_artist_map

station_geo_map

subscribed_users

```
hive> show tables;
OK
formatted_input
song_artist_map
station_geo_map
subscribed_users
users_artists
Time taken: 0.092 seconds, Fetched: 5 row(s)
hive> select * from song_artist_map;
OK
S200    A300
S201    A301
S202    A302
S203    A303
S204    A304
S205    A301
S206    A302
S207    A303
S208    A304
S209    A305
Time taken: 0.889 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: 0.828 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: 0.727 seconds, Fetched: 15 row(s)
hive>
```

Screesn shot of script loading one of the hive table from hbase

```
[acadgild@localhost scripts]$ cat create_hive_hbase_lookup.hql
USE project;
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");
```

6th stage: Enriching the data using dataenrichment.sh

In this script we are creating enriched data.

I have removed all the null values and if like and dislikes are 1 consider it as invalid record.

We are loading the data from formatted_input by joining staion_geo_map and song_artist_map into enriched_data.

The valid and invalid are kept in separate folder and deleting the records that are 7 days old.

Screen shot of enriched_data table

```
hive> select * from enriched_data;
OK
U112    S200    A300    1462863262      1494297562      1494297562      AP      ST412   3       1       1       1       fail
U117    S201    A301    1462863262      1494297562      1462863262      NULL    ST415   3       0       1       1       fail
U109    S201    A301    1475130523      1465230523      1465130523      NULL    ST415   3       0       1       1       fail
U108    S201    A301    1465130523      1485130523      1465130523      A       ST400   1       1       1       1       fail
        S202    A302    1475130523      1475130523      1485130523      A       ST400   0       1       1       1       fail
U114    S202    A302    1465130523      1475130523      1465230523      A       ST405   3       0       0       1       fail
U111    S205    A301    1468094889      1468094889      1468094889      E       ST409   2       0       0       1       fail
U118    S205    A301    1475130523      1485130523      1465230523      AU      ST401   3       1       1       1       fail
U112    S206    A302    1494297562      1465490556      1468094889      J       ST403   0       1       1       1       fail
U101    S206    A302    1495130523      1465130523      1465130523      NULL    ST415   0       1       0       1       fail
U103    S208    A304    1494297562      1468094889      1465490556      NULL    ST415   0       0       1       1       fail
        S208    A304    1465490556      1465490556      1462863262      A       ST405   0       0       0       1       fail
U117    S210    NULL    1465130523      1465230523      1465130523      E       ST414   2       0       0       1       fail
U111    S210    NULL    1495130523      1475130523      1465230523      AU      ST406   0       0       1       1       fail
U120    S200    A300    1465230523      1465130523      1465130523      A       ST400   0       0       0       1       pass
U115    S200    A300    1465230523      1475130523      1475130523      E       ST414   0       1       0       1       pass
U104    S200    A300    1465130523      1485130523      1485130523      AU      ST406   2       0       0       1       pass
U112    S201    A301    1465230523      1465230523      1475130523      J       ST403   3       0       0       1       pass
U107    S201    A301    1462863262      1462863262      1462863262      J       ST413   0       0       1       1       pass
U117    S203    A303    1465490556      1468094889      1462863262      A       ST410   0       0       1       1       pass
U118    S203    A303    1494297562      1465490556      1468094889      E       ST404   0       1       0       1       pass
U111    S203    A303    1495130523      1465130523      1475130523      E       ST404   0       1       0       1       pass
U103    S204    A304    1465490556      1465490556      1462863262      A       ST405   2       0       1       1       pass
U113    S204    A304    1462863262      1462863262      1462863262      AP      ST412   1       0       1       1       pass
U104    S204    A304    1495130523      1485130523      1475130523      A       ST400   0       1       0       1       pass
U109    S204    A304    1465490556      1462863262      1494297562      AU      ST401   0       1       0       1       pass
U111    S205    A301    1494297562      1465490556      1468094889      J       ST403   3       0       0       1       pass
U118    S205    A301    1475130523      1485130523      1485130523      E       ST409   2       0       1       1       pass
U116    S205    A301    1494297562      1462863262      1468094889      A       ST400   1       0       1       1       pass
U101    S205    A301    1462863262      1494297562      1468094889      E       ST414   0       0       0       1       pass
U101    S205    A301    1495130523      1475130523      1485130523      J       ST413   2       0       1       1       pass
U106    S205    A301    1465230523      1465230523      1475130523      J       ST413   3       1       0       1       pass
U116    S206    A302    1494297562      1494297562      1465490556      E       ST414   3       0       0       1       pass
U102    S206    A302    1462863262      1494297562      1468094889      A       ST410   1       0       0       1       pass
U117    S206    A302    1465130523      1475130523      1465230523      AU      ST401   1       1       0       1       pass
U110    S206    A302    1494297562      1465490556      1468094889      AU      ST401   0       0       0       1       pass
U107    S208    A304    1465130523      1465230523      1465130523      A       ST411   1       1       0       1       pass
U111    S208    A304    1494297562      1494297562      1462863262      E       ST408   2       0       0       1       pass
U119    S209    A305    1494297562      1462863262      1494297562      J       ST403   3       0       0       1       pass
U120    S209    A305    1465130523      1465130523      1485130523      AU      ST401   0       0       0       1       pass
Time taken: 0.323 seconds, Fetched: 40 row(s)
```

Output screenshots of data enrichment scripts

```
[acadgild@localhost scripts]$ sh data_enrichment.sh
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
OK
Time taken: 8.785 seconds
OK
Time taken: 0.912 seconds
No Stats for project@formatted_input, Columns: start_ts, song_id, time_stamp, user_id, end_ts, station_id, geo_cd, dislikes, song_end_type, likes
No Stats for project@station_geo_map, Columns: station_id, geo_cd
No Stats for project@song_artist_map, Columns: song_id, artist_id
WARNING: 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. spark, tez) or using Hi
ve 1.X releases.
Query ID = acadgild_20190127182353_66194a18-ddc5-423d-b375-b1d1f7b2833a
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548588431358_0002, Tracking URL = http://localhost:8088/proxy/application_1548588431358_0002/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548588431358_0002
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 1
2019-01-27 18:24:26,488 Stage-1 map = 0%,  reduce = 0%
2019-01-27 18:24:51,519 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 3.72 sec
2019-01-27 18:24:53,761 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.79 sec
2019-01-27 18:25:02,798 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.87 sec
MapReduce Total cumulative CPU time: 8 seconds 870 msec
Ended Job = job_1548588431358_0002
```

```
MapReduce Total cumulative CPU time: 8 seconds 870 msec
Ended Job = job_1548588431358_0002
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548588431358_0003, Tracking URL = http://localhost:8088/proxy/application_1548588431358_0003/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548588431358_0003
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 1
2019-01-27 18:25:24,128 Stage-2 map = 0%,  reduce = 0%
2019-01-27 18:25:40,730 Stage-2 map = 50%,  reduce = 0%, Cumulative CPU 1.54 sec
2019-01-27 18:25:42,910 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 4.49 sec
2019-01-27 18:25:54,138 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 8.26 sec
MapReduce Total cumulative CPU time: 8 seconds 260 msec
Ended Job = job_1548588431358_0003
Loading data to table project.enriched_data partition (batchid=null, status=null)


Loaded : 2/2 partitions.
         Time taken to load dynamic partitions: 1.077 seconds
         Time taken for adding to write entity : 0.002 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 1   Cumulative CPU: 8.87 sec   HDFS Read: 49558 HDFS Write: 3085 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 1   Cumulative CPU: 8.26 sec   HDFS Read: 24168 HDFS Write: 3185 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 130 msec
OK
Time taken: 123.664 seconds
19/01/27 18:25:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/27 18:26:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
You have new mail in /var/spool/mail/acadgild
```

Null values which are present still will be moved to invalid directory.

Query for inserting into table enriched_data

```
INSERT OVERWRITE TABLE enriched_data
PARTITION (batchid, status)
SELECT
i.user_id,
i.song_id,
sa.artist_id,
i.time_stamp,
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.likes IS NULL, 0, i.likes) AS likes,
IF (i.dislikes IS NULL, 0, i.dislikes) AS dislikes,
i.batchid,
IF((i.likes=1 AND i.dislikes=1)
OR i.user_id IS NULL
OR i.song_id IS NULL
OR i.time_stamp 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.time_stamp=''
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=${hiveconf:batchid};
You have new mail in /var/spool/mail/acadgild
```

7th Stage data analysis using data_analysis.sh

In this script we are performing 5 uses cases of the project

and exporting the data into Mysql using sqoop export. We are increasing the batch id to 2 and so on for next iteration.

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

Hive Query to determine top 10 station ids

```
INSERT OVERWRITE TABLE top_10_stations
PARTITION(batchid=${hiveconf:batchid})
SELECT
station_id,
COUNT(DISTINCT song_id) AS total_distinct_songs_played,
COUNT(DISTINCT user_id) AS distinct_user_count
FROM enriched_data
WHERE status='pass'
AND batchid=${hiveconf:batchid}
AND likes=1
GROUP BY station_id
ORDER BY total_distinct_songs_played DESC
LIMIT 10;
```

Hive output top_10_stations

```
hive> select * from top_10_stations;
OK
ST401   2       2       1
ST414   1       1       1
ST413   1       1       1
ST411   1       1       1
ST404   1       2       1
ST400   1       1       1
Time taken: 0.187 seconds, Fetched: 6 row(s)
```

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.

Hive Query for determining total duration of song by sub/unsub users

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

Hive Result

```
hive> select * from users_behaviour;
OK
SUBSCRIBED      106266940       1
UNSUBSCRIBED    169951859       1
Time taken: 0.19 seconds, Fetched: 2 row(s)
```

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

Hive query for finding top 10 connected artists

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

Hive result for top connected artists

```
hive> select * from connected_artists;
OK
A301    4       1
A302    2       1
A303    1       1
A304    1       1
Time taken: 0.213 seconds, Fetched: 4 row(s)
```

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.

Hive query for top 10 songs that generated max revenue

```
INSERT OVERWRITE TABLE top_10_royalty_songs
PARTITION(batchid=${hiveconf:batchid})
SELECT song_id,
SUM(ABS(CAST(end_ts AS DECIMAL(20,0))-CAST(start_ts AS DECIMAL(20,0)))) AS duration
FROM enriched_data
WHERE status='pass'
AND batchid=${hiveconf:batchid}
AND (likes=1 OR song_end_type=0)
GROUP BY song_id
ORDER BY duration DESC
LIMIT 10;
```

Hive Result for top royalty

```
hive> select * from top_10_royalty_songs;
OK
S204    41434300        1
S205    36102673        1
S209    20000000        1
S203    17835960        1
S206    12504333        1
S208    100000  1
S200    0       1
S201    0       1
Time taken: 0.505 seconds, Fetched: 8 row(s)
hive>
```

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

Hive Query for 10 unsubscribed users who listened to the songs for the longest duration

```
INSERT OVERWRITE TABLE top_10_unsubscribed_users
PARTITION(batchid=${hiveconf:batchid})
SELECT
ed.user_id,
SUM(ABS(CAST(ed.end_ts AS DECIMAL(20,0))-CAST(ed.start_ts AS DECIMAL(20,0)))) AS duration
FROM enriched_data ed
LEFT OUTER JOIN subscribed_users su
ON ed.user_id=su.user_id
WHERE ed.status='pass'
AND ed.batchid=${hiveconf:batchid}
AND (su.user_id IS NULL OR (CAST(ed.time_stamp AS DECIMAL(20,0)) > CAST(su.subscn_end_dt AS DECIMAL(20,0))))
GROUP BY ed.user_id
ORDER BY duration DESC
LIMIT 10;
```

Hive result for top 10 unsubscribed users

```
hive> select * from top_10_unsubscribed_users;
OK
U111    44038633        1
U116    34038633        1
U119    31434300        1
U120    20000000        1
U117    15131627        1
U101    10000000        1
U104    10000000        1
U118    2604333 1
U110    2604333 1
U107    100000  1
Time taken: 0.391 seconds, Fetched: 10 row(s)
hive>
```


OUTPUT SCREENSHOTS for one ANALYSIS

Screenshots for one of the analysis of top_10_unsubscribed_users in hive

```
hive> INSERT OVERWRITE TABLE top_10_unsubscribed_users
    > PARTITION(batchid=2)
    > SELECT
    > ed.user_id,
    > SUM(ABS(end_ts - start_ts)) 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=2
    > AND (su.user_id IS NULL OR (CAST(ed.time_stamp AS DECIMAL(20,0)) > CAST(su.subscn_end_dt AS DECIMAL(20,0))))
    > GROUP BY ed.user_id
    > ORDER BY duration DESC
    > LIMIT 10;
WARNING: 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. spark, tez) or using Hi
ve 1.X releases.
Query ID = acadgild_20190130122414_c8d40ca6-6af3-4670-a96c-e4d6189d7a5a
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548821963264_0018, Tracking URL = http://localhost:8088/proxy/application_1548821963264_0018/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548821963264_0018
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2019-01-30 12:24:38,938 Stage-1 map = 0%,  reduce = 0%
2019-01-30 12:25:14,531 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 3.94 sec
2019-01-30 12:25:16,981 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 10.16 sec
2019-01-30 12:25:33,549 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 15.46 sec
2019-01-30 12:25:36,097 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 17.3 sec
MapReduce Total cumulative CPU time: 17 seconds 300 msec
Ended Job = job_1548821963264_0018
Launching Job 2 out of 3
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
```

```
MapReduce Total cumulative CPU time: 17 seconds 300 msec
Ended Job = job_1548821963264_0018
Launching Job 2 out of 3
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548821963264_0019, Tracking URL = http://localhost:8088/proxy/application_1548821963264_0019/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548821963264_0019
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-01-30 12:26:01,072 Stage-2 map = 0%,  reduce = 0%
2019-01-30 12:26:16,749 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.6 sec
2019-01-30 12:26:34,458 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 6.66 sec
MapReduce Total cumulative CPU time: 6 seconds 660 msec
Ended Job = job_1548821963264_0019
Launching Job 3 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548821963264_0020, Tracking URL = http://localhost:8088/proxy/application_1548821963264_0020/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548821963264_0020
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2019-01-30 12:27:00,179 Stage-3 map = 0%,  reduce = 0%
2019-01-30 12:27:15,980 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.46 sec
2019-01-30 12:27:34,748 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 7.61 sec
MapReduce Total cumulative CPU time: 8 seconds 90 msec
Ended Job = job_1548821963264_0020
Loading data to table project.top_10_unsubscribed_users partition (batchid=2)
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 17.3 sec   HDFS Read: 23591 HDFS Write: 546 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 6.66 sec   HDFS Read: 5249 HDFS Write: 546 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 8.09 sec   HDFS Read: 7047 HDFS Write: 240 SUCCESS
Total MapReduce CPU Time Spent: 32 seconds 50 msec
OK
Time taken: 203.276 seconds
hive>
```

Out Put screen shots of Data analysis script :

```
Last login: Sun Jan 27 16:56:48 2019 from 10.0.2.2
[acadgild@localhost ~]$ sh /home/acadgild/project/scripts/data_analysis.sh
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
OK
Time taken: 18.959 seconds
OK
Time taken: 1.678 seconds
WARNING: 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. spark, tez) or using Hi
ve 1.X releases.
Query ID = acadgild_20190127200435_ea7c3f21-64f6-4c4e-8105-1e13b9614c56
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548588431358_0004, Tracking URL = http://localhost:8088/proxy/application_1548588431358_0004/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548588431358_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-01-27 20:05:17,841 Stage-1 map = 0%,  reduce = 0%
2019-01-27 20:05:37,165 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.84 sec
2019-01-27 20:05:56,635 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.55 sec
MapReduce Total cumulative CPU time: 9 seconds 550 msec
Ended Job = job_1548588431358_0004
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
```

```
Ended Job = job_1548588431358_0005
Loading data to table project.top_10_stations partition (batchid=1)
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.55 sec   HDFS Read: 12959 HDFS Write: 246 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 9.88 sec   HDFS Read: 7433 HDFS Write: 149 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 430 msec
OK
Time taken: 153.895 seconds
OK
Time taken: 0.274 seconds
WARNING: 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. spark, tez) or using Hi
ve 1.X releases.
Query ID = acadgild_20190127200709_0a11f8ca-d259-4708-93d0-d521cb19c64b
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548588431358_0006, Tracking URL = http://localhost:8088/proxy/application_1548588431358_0006/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548588431358_0006
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2019-01-27 20:08:00,627 Stage-1 map = 0%,  reduce = 0%
2019-01-27 20:08:35,092 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 4.03 sec
2019-01-27 20:08:37,569 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 10.12 sec
2019-01-27 20:08:54,562 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 15.15 sec
2019-01-27 20:08:56,950 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 17.15 sec
MapReduce Total cumulative CPU time: 17 seconds 150 msec
Ended Job = job_1548588431358_0006
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548588431358_0007, Tracking URL = http://localhost:8088/proxy/application_1548588431358_0007/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548588431358_0007
```

```
MapReduce Total cumulative CPU time: 17 seconds 150 msec
Ended Job = job_1548588431358_0006
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548588431358_0007, Tracking URL = http://localhost:8088/proxy/application_1548588431358_0007/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548588431358_0007
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-01-27 20:09:24,675 Stage-2 map = 0%,  reduce = 0%
2019-01-27 20:09:40,585 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.71 sec
2019-01-27 20:10:01,324 Stage-2 map = 100%,  reduce = 67%, Cumulative CPU 7.27 sec
2019-01-27 20:10:03,715 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 8.97 sec
MapReduce Total cumulative CPU time: 8 seconds 970 msec
Ended Job = job_1548588431358_0007
Loading data to table project.users_behaviour partition (batchid=1)
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 17.15 sec   HDFS Read: 36111 HDFS Write: 166 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 8.97 sec   HDFS Read: 6751 HDFS Write: 133 SUCCESS
Total MapReduce CPU Time Spent: 26 seconds 120 msec
OK
Time taken: 176.957 seconds
OK
Time taken: 0.202 seconds
WARNING: 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. spark, tez) or using Hi
ve 1.X releases.
Query ID = acadgild_20190127201006_6f959b59-e5f1-42ac-8f74-39d76777ca96
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548588431358_0008, Tracking URL = http://localhost:8088/proxy/application_1548588431358_0008/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548588431358_0008
```

```
Starting Job = job_1548588431358_0008, Tracking URL = http://localhost:8088/proxy/application_1548588431358_0008/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548588431358_0008
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2019-01-27 20:10:31,081 Stage-1 map = 0%,  reduce = 0%
2019-01-27 20:11:09,613 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 5.49 sec
2019-01-27 20:11:10,993 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 11.62 sec
2019-01-27 20:11:33,763 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 16.68 sec
MapReduce Total cumulative CPU time: 16 seconds 680 msec
Ended Job = job_1548588431358_0008
Launching Job 2 out of 3
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548588431358_0009, Tracking URL = http://localhost:8088/proxy/application_1548588431358_0009/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548588431358_0009
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-01-27 20:12:01,313 Stage-2 map = 0%,  reduce = 0%
2019-01-27 20:12:19,425 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.87 sec
2019-01-27 20:12:29,628 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.71 sec
MapReduce Total cumulative CPU time: 4 seconds 710 msec
Ended Job = job_1548588431358_0009
Launching Job 3 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1548588431358_0010, Tracking URL = http://localhost:8088/proxy/application_1548588431358_0010/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548588431358_0010
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2019-01-27 20:12:47,464 Stage-3 map = 0%,  reduce = 0%
2019-01-27 20:12:55,121 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 1.27 sec
2019-01-27 20:13:05,999 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 4.21 sec
MapReduce Total cumulative CPU time: 4 seconds 210 msec
Ended Job = job_1548588431358_0010
```

8th stage:Tables in mysql in project database

Tables in mysql

```
mysql> use project;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
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 (0.00 sec)

mysql>
```

Exporting the data in my sql using sqoop

1.top 10 stations

*sqoop export *

*--connect jdbc:mysql://localhost/project *

*--username 'root' *

*--password Root@123 *

*--table top_10_stations *

*--export-dir=/user/hive/warehouse/project.db/top_10_stations/batchid=1 *

*--input-fields-terminated-by ',' *

-m 1

```
mysql> select * from top_10_stations;
+------------+----------------------------+---------------------+
| station_id | total_distinct_songs_played | distinct_user_count |
+------------+----------------------------+---------------------+
| ST401      |                          2 |                   2 |
| ST414      |                          1 |                   1 |
| ST413      |                          1 |                   1 |
| ST411      |                          1 |                   1 |
| ST404      |                          1 |                   2 |
| ST400      |                          1 |                   1 |
+------------+----------------------------+---------------------+
6 rows in set (0.00 sec)
```

2. Subscribed and Unsubscribed users

*sqoop export *

*--connect jdbc:mysql://localhost/project *

*--username 'root' *

*--table users_behaviour *

*--export-dir =/user/hive/warehouse/project.db/users_behaviour/batchid=$batchid *

*--input-fields-terminated-by ',' *

-m 1

Batch id is one because it's for first iteration

```
mysql> select * from users_behaviour;
+-------------+----------+
| user_type   | duration |
+-------------+----------+
| SUBSCRIBED  | 106266940 |
| UNSUBSCRIBED | 169951859 |
+-------------+----------+
2 rows in set (0.00 sec)
```

3.Connected artists

*sqoop export *

*--connect jdbc:mysql://localhost/project *

*--username 'root' *

*--table connected_artists *

--export-dir

*=/user/hive/warehouse/project.db/connected_artists/batchid=$batchid *

*--input-fields-terminated-by ',' *

-m 1

```
mysql> select * from connected_artists;
+----------+-----------+
| artist_id | user_count |
+----------+-----------+
| A301     |         4 |
| A302     |         2 |
| A303     |         1 |
| A304     |         1 |
+----------+-----------+
4 rows in set (0.00 sec)
```

-

4.Royalty song

*sqoop export *

*--connect jdbc:mysql://localhost/project *

*--username 'root' *

*--password Root@123 *

*--table top_10_royalty_songs *

*--export-dir=/user/hive/warehouse/project.db/top_10_royalty_songs/batchid=1 *

*--input-fields-terminated-by ',' *

-m 1

```
mysql> use project;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> select * from top_10_royalty_songs;
+---------+----------+
| song_id | duration |
+---------+----------+
| S204    | 41434300 |
| S205    | 36102673 |
| S209    | 20000000 |
| S203    | 17835960 |
| S206    | 12504333 |
| S208    |   100000 |
| S200    |        0 |
| S201    |        0 |
+---------+----------+
8 rows in set (0.00 sec)
```

5. Top 10 unsubscribed users playing the song

*sqoop export *

*--connect jdbc:mysql://localhost/project *

*--username 'root' *

*--password Root@123 *

*--table top_10_unsubscribed_users *

*--export-dir=/user/hive/warehouse/project.db/top_10_unsubscribed_users/batchid=$batchid *

*--input-fields-terminated-by ',' *

-m 1

```
mysql> select * from top_10_unsubscribed_users;
+---------+----------+
| user_id | duration |
+---------+----------+
| U111    | 44038633 |
| U116    | 34038633 |
| U119    | 31434300 |
| U120    | 20000000 |
| U117    | 15131627 |
| U101    | 10000000 |
| U104    | 10000000 |
| U118    |  2604333 |
| U110    |  2604333 |
| U107    |   100000 |
+---------+----------+
10 rows in set (0.00 sec)
```

Screen shot of data exporting via scoop for one table top_10_royalty_songs

```
[acadgild@localhost scripts]$ sqoop export \
> --connect jdbc:mysql://localhost/project \
> --username 'root' \
> --password Root@123 \
> --table top_10_royalty_songs \
> --export-dir=/user/hive/warehouse/project.db/top_10_royalty_songs/batchid=1 \
> --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.
19/01/30 12:00:35 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
19/01/30 12:00:35 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
19/01/30 12:00:35 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
19/01/30 12:00:35 INFO tool.CodeGenTool: Beginning code generation
Wed Jan 30 12:00:36 IST 2019 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 verifyServerCe
rtificate 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 certific
ate verification.
19/01/30 12:00:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `top_10_royalty_songs` AS t LIMIT 1
19/01/30 12:00:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `top_10_royalty_songs` AS t LIMIT 1
19/01/30 12:00:39 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/install/hadoop/hadoop-2.6.5
Note: /tmp/sqoop-acadgild/compile/c6bdd0c930f3be2cfaa30c9b72ac96df/top_10_royalty_songs.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
19/01/30 12:00:47 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/c6bdd0c930f3be2cfaa30c9b72ac96df/top_10_royalty_songs.jar
19/01/30 12:00:47 INFO 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/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
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: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
19/01/30 12:00:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/30 12:00:48 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
19/01/30 12:00:51 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
19/01/30 12:00:51 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
19/01/30 12:00:51 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
19/01/30 12:00:51 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
19/01/30 12:00:57 INFO input.FileInputFormat: Total input paths to process : 1
```



```
19/01/30 12:01:23 INFO mapreduce.Job:  map 0% reduce 0%
19/01/30 12:01:42 INFO mapreduce.Job:  map 100% reduce 0%
19/01/30 12:01:42 INFO mapreduce.Job: Job job_1548821963264_0011 completed successfully
19/01/30 12:01:42 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=127620
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=272
                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)=15259
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=15259
                Total vcore-milliseconds taken by all map tasks=15259
                Total megabyte-milliseconds taken by all map tasks=15625216
        Map-Reduce Framework
                Map input records=8
                Map output records=8
                Input split bytes=173
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=192
                CPU time spent (ms)=3820
                Physical memory (bytes) snapshot=119164928
                Virtual memory (bytes) snapshot=2061332480
                Total committed heap usage (bytes)=62980096
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
19/01/30 12:01:43 INFO mapreduce.ExportJobBase: Transferred 272 bytes in 51.3977 seconds (5.2921 bytes/sec)
19/01/30 12:01:43 INFO mapreduce.ExportJobBase: Exported 8 records.
You have new mail in /var/spool/mail/acadgild
```

9th Stage:Scheduling the crontab

All the scripts python ,data formatting data , start-daemons, enrichment and analysis scripts are placed in project.sh and scheduled in crontab for iterations

```
[acadgild@localhost scripts]$ crontab -l
* * * * * /home/acadgild/install/data/dfs/simple/update-acadgildvm.sh
* */3 * * * /home/acadgild/project/scripts/project.sh
[acadgild@localhost scripts]$
```

```
[acadgild@localhost scripts]$ cat project.sh
#!/bin/bash

python /home/acadgild/project/scripts/generate_web_data.py
python /home/acadgild/project/scripts/generate_mob_data.py
sh /home/acadgild/project/scripts/start-daemons.sh
sh /home/acadgild/project/scripts/dataformatting.sh

sh /home/acadgild/project/scripts/data_enrichment.sh
sh /home/acadgild/project/scripts/data_analysis.sh
```

Not included the populate-lookup.sh as the records in files in lookupfiles directory are static and already loaded once.

After 2nd iteration batch 2

Hdfs 2nd iteration files

```
[acadgild@localhost scripts]$ hdfs dfs -ls /user/acadgild/project
19/01/30 00:32:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - acadgild supergroup          0 2019-01-27 17:32 /user/acadgild/project/batch1
drwxr-xr-x   - acadgild supergroup          0 2019-01-30 00:02 /user/acadgild/project/batch2
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost scripts]$ hdfs dfs -ls /user/acadgild/project/batch2
19/01/30 00:32:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
drwxr-xr-x   - acadgild supergroup          0 2019-01-30 00:03 /user/acadgild/project/batch2/formattedweb
drwxr-xr-x   - acadgild supergroup          0 2019-01-30 00:03 /user/acadgild/project/batch2/mob
drwxr-xr-x   - acadgild supergroup          0 2019-01-30 00:01 /user/acadgild/project/batch2/web
```

Hive results

Station id changed after second iteration. The values will get repeated in station id field as I didn't remove the previous iteration file and didn't drop the table so that we can compare previous run

```
hive> select * from top_10_stations;
OK
ST401   2       2       1
ST414   1       1       1
ST413   1       1       1
ST411   1       1       1
ST404   1       2       1
ST400   1       1       1
ST414   2       2       2
ST404   2       3       2
ST411   2       2       2
ST401   2       2       2
ST400   2       2       2
ST413   1       1       2
ST409   1       1       2
Time taken: 0.824 seconds, Fetched: 13 row(s)
hive> select * from users_behaviour;
OK
SUBSCRIBED      106266940       1
UNSUBSCRIBED    169951859       1
SUBSCRIBED      189035540       2
UNSUBSCRIBED    293337791       2
Time taken: 0.612 seconds, Fetched: 4 row(s)
hive> select * from connected_artists;
OK
A301    4       1
A302    2       1
A303    1       1
A304    1       1
A301    6       2
A302    3       2
A303    2       2
A304    1       2
A300    1       2
Time taken: 0.585 seconds, Fetched: 9 row(s)
hive>
```

Here we can see song id S200 has some value (royalty after second iterartion)

The values will get repeated in song id field as I didn't remove the previous iteration file and didn't drop the table because to compare the values.

```
hive> select * from top_10_royalty_songs;
OK
S204    41434300        1
S205    36102673        1
S209    20000000        1
S203    17835960        1
S206    12504333        1
S208    100000  1
S200    0       1
S201    0       1
S203    51874593        2
S204    41434300        2
S205    36102673        2
S201    28807006        2
S200    22604333        2
S208    20100000        2
S209    20000000        2
S206    12504333        2
S202    0       2
Time taken: 0.71 seconds, Fetched: 17 row(s)
hive> select * from top_10_unsubscribed_users;
OK
U111    44038633        1
U116    34038633        1
U119    31434300        1
U120    20000000        1
U117    15131627        1
U101    10000000        1
U104    10000000        1
U118    2604333 1
U110    2604333 1
U107    100000  1
U111    44038633        2
U119    41434300        2
U117    37735960        2
U118    36642966        2
U116    34038633        2
U115    34038633        2
U120    30000000        2
U101    10000000        2
U108    10000000        2
U104    10000000        2
Time taken: 0.323 seconds, Fetched: 20 row(s)
```

MySQL results

```
mysql> select * from top_10_stations;
+------------+---------------------------+--------------------+
| station_id | total_distinct_songs_played | distinct_user_count |
+------------+---------------------------+--------------------+
| ST401      |                         2 |                  2 |
| ST414      |                         1 |                  1 |
| ST413      |                         1 |                  1 |
| ST411      |                         1 |                  1 |
| ST404      |                         1 |                  2 |
| ST400      |                         1 |                  1 |
| ST401      |                         2 |                  2 |
| ST414      |                         1 |                  1 |
| ST413      |                         1 |                  1 |
| ST411      |                         1 |                  1 |
| ST404      |                         1 |                  2 |
| ST400      |                         1 |                  1 |
| ST414      |                         2 |                  2 |
| ST404      |                         2 |                  3 |
| ST411      |                         2 |                  2 |
| ST401      |                         2 |                  2 |
| ST400      |                         2 |                  2 |
| ST413      |                         1 |                  1 |
| ST409      |                         1 |                  1 |
+------------+---------------------------+--------------------+
19 rows in set (0.01 sec)

mysql> select * from users_behaviour;
+--------------+-----------+
| user_type    | duration  |
+--------------+-----------+
| SUBSCRIBED   | 106266940 |
| UNSUBSCRIBED | 169951859 |
| SUBSCRIBED   | 189035540 |
| UNSUBSCRIBED | 293337791 |
+--------------+-----------+
4 rows in set (0.00 sec)
```

```
mysql> select * from connected_artists;
+-----------+------------+
| artist_id | user_count |
+-----------+------------+
| A301      |          4 |
| A302      |          2 |
| A303      |          1 |
| A304      |          1 |
| A301      |          6 |
| A302      |          3 |
| A303      |          2 |
| A304      |          1 |
| A300      |          1 |
+-----------+------------+
9 rows in set (0.00 sec)

mysql>
```

Folders and scripts used for project

drwxrwxr-x. 4 acadgild acadgild 4096 Jan 22 21:43 data (Mobile and web files generated by python)

```
drwxrwxr-x. 2 acadgild acadgild 4096 Jan 29 23:30 mob
drwxrwxr-x. 2 acadgild acadgild 4096 Jan 22 21:43 web
[acadgild@localhost data]$ pwd
/home/acadgild/project/data
```

drwxrwxr-x. 2 acadgild acadgild 4096 Jan 22 22:57 lib (piggybank.jar file pig formatting the xml file)

drwxrwxr-x. 2 acadgild acadgild 4096 Jan 30 12:59 logs (log folder)

drwxrwxr-x. 2 acadgild acadgild 4096 Sep 25 2017 lookupfiles (lookup files)

```
[acadgild@localhost lookupfiles]$ ll
total 32
-rw-rw-r--. 1 acadgild acadgild 100 Mar 14  2017 song-artist.txt
-rw-rw-r--. 1 acadgild acadgild 108 Mar 14  2017 song-artist.txt~
-rw-rw-r--. 1 acadgild acadgild 125 Mar 14  2017 stn-geocd.txt
-rw-rw-r--. 1 acadgild acadgild 138 Mar 14  2017 stn-geocd.txt~
-rw-rw-r--. 1 acadgild acadgild 240 Mar 14  2017 user-artist.txt
-rw-rw-r--. 1 acadgild acadgild 253 Mar 14  2017 user-artist.txt~
-rw-rw-r--. 1 acadgild acadgild 405 Mar 14  2017 user-subscn.txt
-rw-rw-r--. 1 acadgild acadgild 418 Mar 14  2017 user-subscn.txt~
[acadgild@localhost lookupfiles]$ pwd
/home/acadgild/project/lookupfiles
```

drwxrwxr-x. 4 acadgild acadgild 4096 Jan 23 01:22 processed_dir (valid/invalid records)

```
[acadgild@localhost processed_dir]$ ll
total 8
drwxrwxr-x. 85 acadgild acadgild 4096 Jan 30 12:59 invalid
drwxrwxr-x. 85 acadgild acadgild 4096 Jan 30 12:59 valid
[acadgild@localhost processed_dir]$ pwd
/home/acadgild/project/processed_dir
```

drwxrwxr-x. 3 acadgild acadgild 4096 Jan 30 11:46 scripts (all the scripts for project)

-rwxrwxr-x. 1 acadgild acadgild 11139 Mar 14 2017 connected_artists.java

-rwxrwxr-x. 1 acadgild acadgild 872 Mar 14 2017 create_hive_hbase_lookup.hql

-rwxrwxr-x. 1 acadgild acadgild 592 Mar 14 2017 create_schema.sql

-rwxrwxr-x. 1 acadgild acadgild 2358 Jan 30 00:16 data_analysis.hql

-rwxrwxr-x. 1 acadgild acadgild 3560 Jan 30 00:12 data_analysis.hql_backup

-rwxrwxr-x. 1 acadgild acadgild 508 Jan 30 00:52 data_analysis.sh

-rwxrwxr-x. 1 acadgild acadgild 299 Mar 14 2017 data_enrichment_filtering_schema.sh

-rwxrwxr-x. 1 acadgild acadgild 1346 Jan 27 18:22 data_enrichment.hql

-rwxrwxr-x. 1 acadgild acadgild 982 Mar 14 2017 data_enrichment.sh

-rwxrwxr-x. 1 acadgild acadgild 996 Jan 30 09:32 data_export.sh

-rwxrwxr-x. 1 acadgild acadgild 1743 Jan 30 00:13 data_export.sh_backup

```
-rwxrwxr-x. 1 acadgild acadgild  1046 Mar 14  2017 dataformatting.pig
-rwxrwxr-x. 1 acadgild acadgild   892 Jan 30 09:54 dataformatting.sh
-rwxrwxr-x. 1 acadgild acadgild   607 Jan 29 23:23 formatted_hive_load.hql
-rwxrwxr-x. 1 acadgild acadgild  1193 Mar 14  2017 generate_mob_data.py
-rwxrwxr-x. 1 acadgild acadgild  1831 Mar 14  2017 generate_web_data.py
-rwxrwxr-x. 1 acadgild acadgild  1334 Mar 14  2017 populate-lookup.sh
-rwxrwxr-x. 1 acadgild acadgild   339 Jan 30 00:20 project.sh
-rwxrwxr-x. 1 acadgild acadgild   456 Mar 14  2017 start-daemons.sh
-rwxrwxr-x. 1 acadgild acadgild 10964 Mar 14  2017 top_10_royalty_songs.java
-rw-rw-r--. 1 acadgild acadgild 14612 Jan 29 20:48 top_10_stations.java
-rwxrwxr-x. 1 acadgild acadgild 13188 Mar 14  2017 top_10_unsubscribed_users.java
-rwxrwxr-x. 1 acadgild acadgild   337 Mar 14  2017 user-artist.hql
-rw-rw-r--. 1 acadgild acadgild 10989 Jan 29 22:16 users_behaviour.java
```