FINAL PROJECT

MUSIC DATA Analysis

Using HADOOP ECO SYSTEM

By

MILIND GAUTAM

Table of Contents

1.0 Project Description :

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

2.0 Data Files :

Data set consists of user information , song details like :

- **song_id**
- **Artist_id**
- **number of likes and dislikes received for each song.**

2.1 Project Description :

Data files contain below fields :

Column Name/Field Name	Column Description/Field Description
User_id	**Unique identifier of every user**
Song_id	**Unique identifier of every song**
Artist_id	**Unique identifier of the lead artist of the song**
Timestamp	**Timestamp when the record was generated**
Start_ts	**Start timestamp when the song started to play**
End_ts	**End timestamp when the song was stopped**
Geo_cd	**Can be 'A' for USA region, 'AP' for asia pacific region,'J' for Japan region, 'E' for europe and 'AU' for australia region**
Station_id	**Unique identifier of the station from where the song was played**

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

2.2 LookUp Tables :

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

Table Name	Description
Station_Geo_Map	Contains mapping of a geo_cd with station_id
Subscribed_Users	Contains user_id, subscription_start_date and subscription_end_date. Contains details only for subscribed users
Song_Artist_Map	Contains mapping of song_id with artist_id alongwith royalty associated with each play of the song
User_Artist_Map	Contains an array of artist_id(s) followed by a user_id

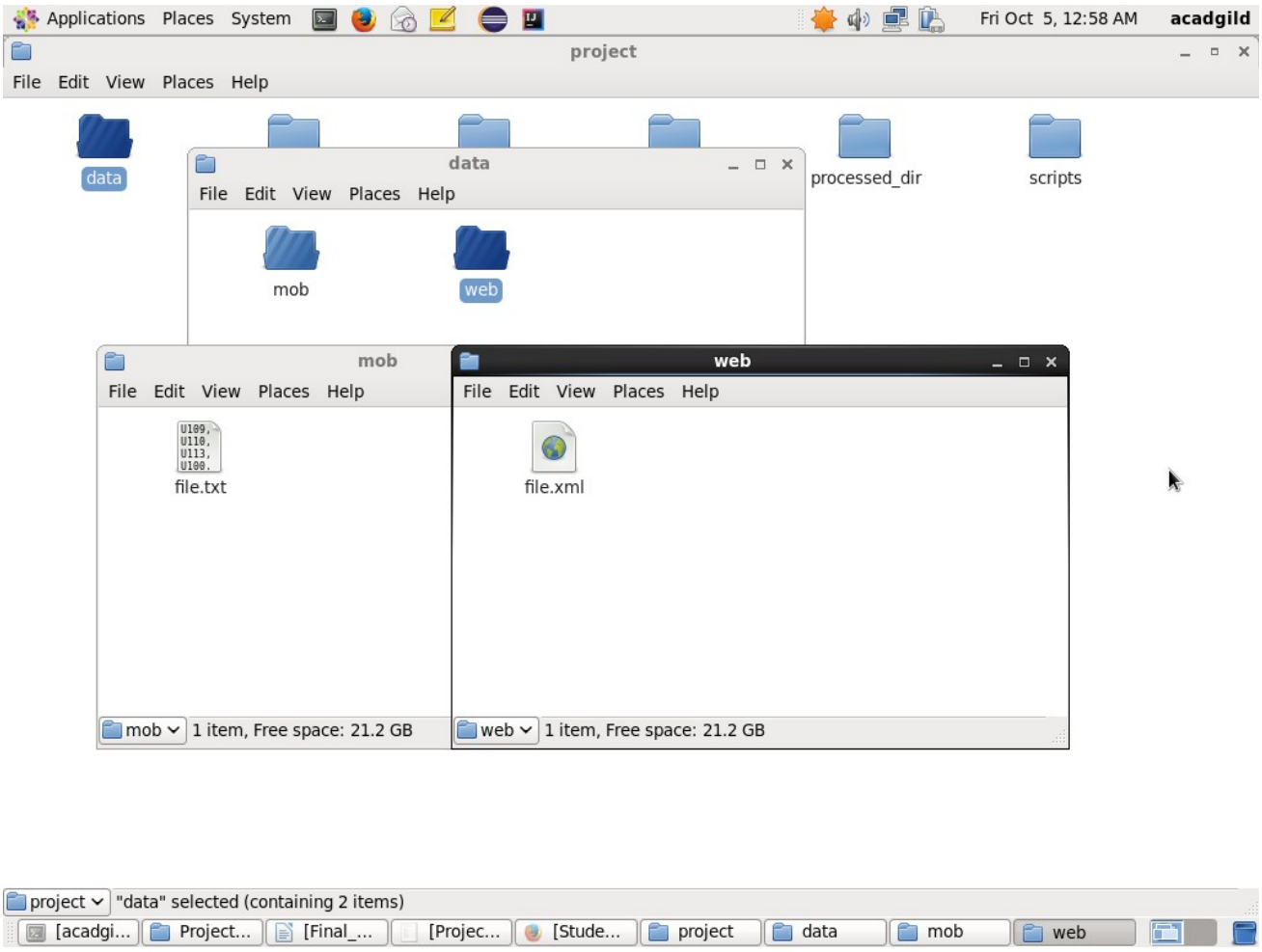
2.3. Data set :

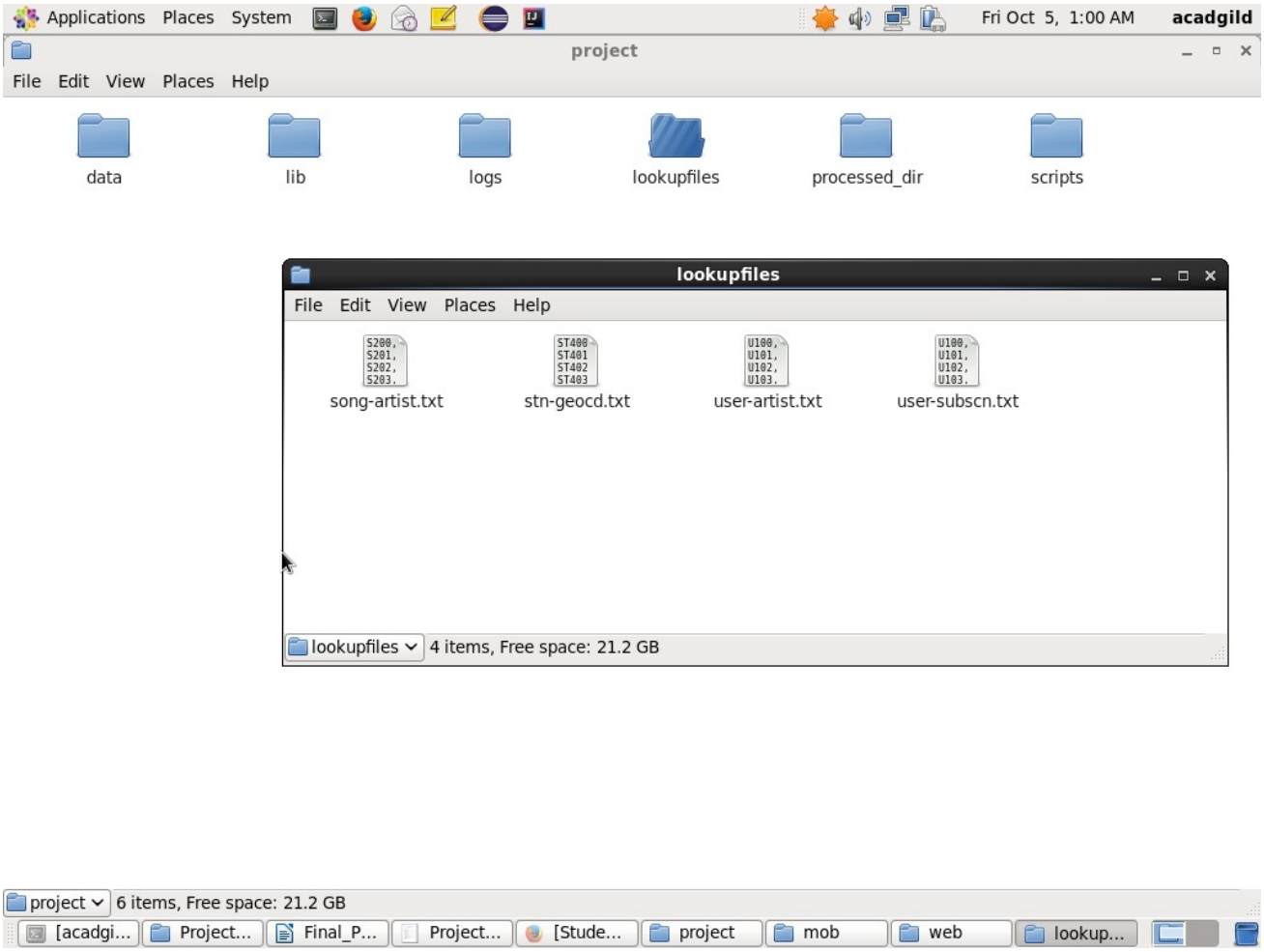
1. Data coming from web applications reside in /data/web and has xml format.

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

3. Data present in lookup directory should be used in HBase.

Screenshot :




Applications Places System
Fri Oct 5, 1:00 AM acadgild
project
File Edit View Places Help
data
lib
logs
lookupfiles
processed_dir
scripts
lookupfiles
File Edit View Places Help
S200,
S201,
S202,
S203.
song-artist.txt
ST400
ST401
ST402
ST403
stn-geocd.txt
U100,
U101,
U102,
U103.
user-artist.txt
U100,
U101,
U102,
U103.
user-subscn.txt
lookupfiles 4 items, Free space: 21.2 GB
project 6 items, Free space: 21.2 GB
[acadgi...
Project...
Final_P...
Project...
[Stude...
project
mob
web
lookup...

3.0 Data Ingestion and Initial Validation :

3.1 Rules for data ingestion and data filtering :

- Data coming from web applications reside in /data/web and has xml format.
- Data coming from mobile applications reside in /data/mob and has csv format.
- Data files come every 3 hours.
- All the timestamp fields in data coming from web application is of the format YYYY-MM-DD HH:MM:SS.
- All the timestamp fields in data coming from mobile application is a long integer interpreted as UNIX timestamps.
- Finally, all timestamps must have the format of a long integer to be interpreted as UNIX timestamps.
- If both like and dislike are 1, consider that record to be invalid.
- If any of the fields from User_id, Song_id, Timestamp, Start_ts, End_ts, Geo_cd is NULL or absent, consider that record to be **invalid**.
- If Song_end_type is NULL or absent, treat it to be 3
- Create a temporary identfier for all the data files received in the last 3 hours (may be an integer batch_id which is auto incremented or a string obtained after combining current date and current hour, to keep track of valid and invalid records per batch).

4.0 Data Enrichment :

4.1 Rules for data enrichment :

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

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

4.2 Post Enrichment :

- Move all valid records in /hadoop/processing_dir in HDFS and invalid records in Local File System at /usr/invalid directory.
- Maintain a copy of valid records in /usr/validated in Local File System. Run a cleaner everyday to clean validated files which are more than 7 days old.

5.0 Data Analysis (using SPARK for optimization) :

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.

- Determine top 10 station_id(s) where maximum number of songs were played, which were liked by unique users.
- 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 presentin Subscribed_users lookup table or has subscription_end_date earlier than the timestamp of the song played by him.
- Determine top 10 connected artists. Connected artists are those whose songs are most listened by the unique users who follow them.
- 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.
- Determine top 10 unsubscribed users who listened to the songs for the longest duration.

5.1 Challenges and Optimizations :

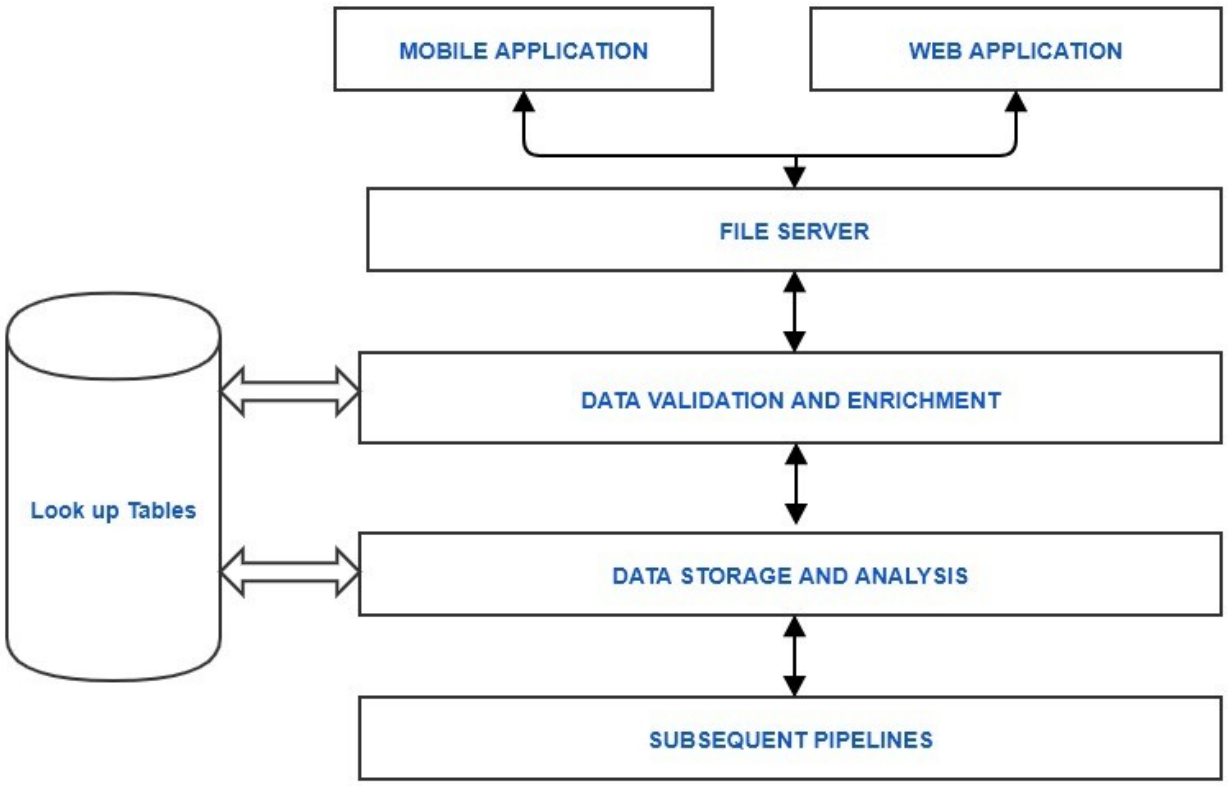
- LookUp tables are in NoSQL databases. Integrate them with the actual data flow.
- Try to make joins as less expensive as possible.
- Data Cleaning, Validation, Enrichment, Analysis and Post Analysis have to be automated. Try using schedulers.
- Appropriate logs have to maintain to track the behavior and overcome failures in the pipeline.

6.0 Post Analysis :

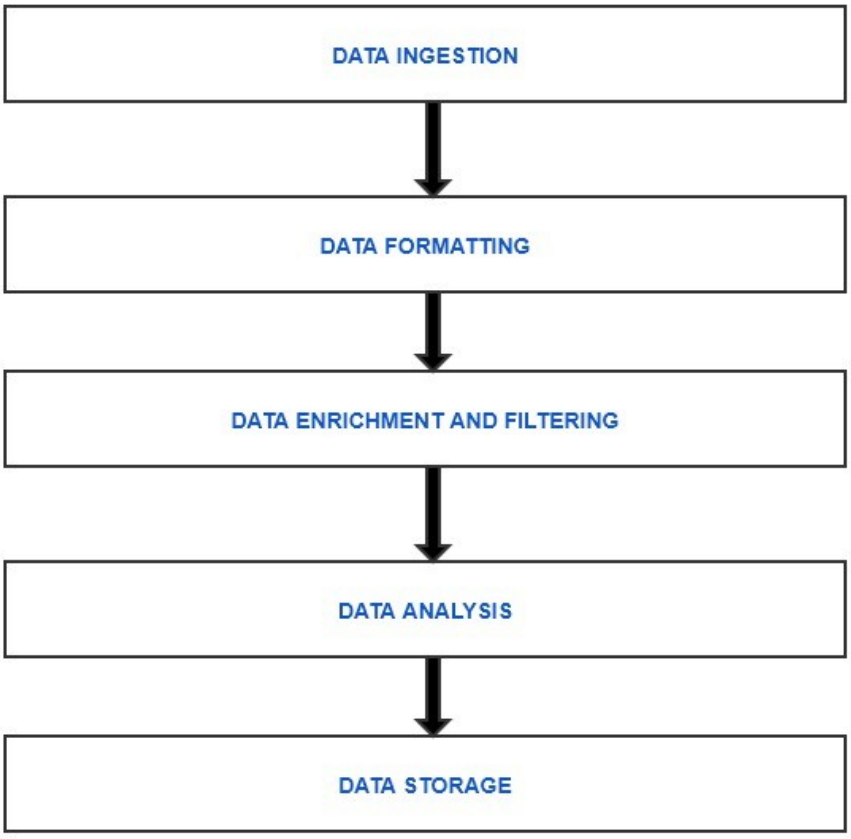
Once the analysis is complete, multiple actions can be taken place later on.

- **Moving result of analysis to the RDMS for data storage and quick retrieval.**

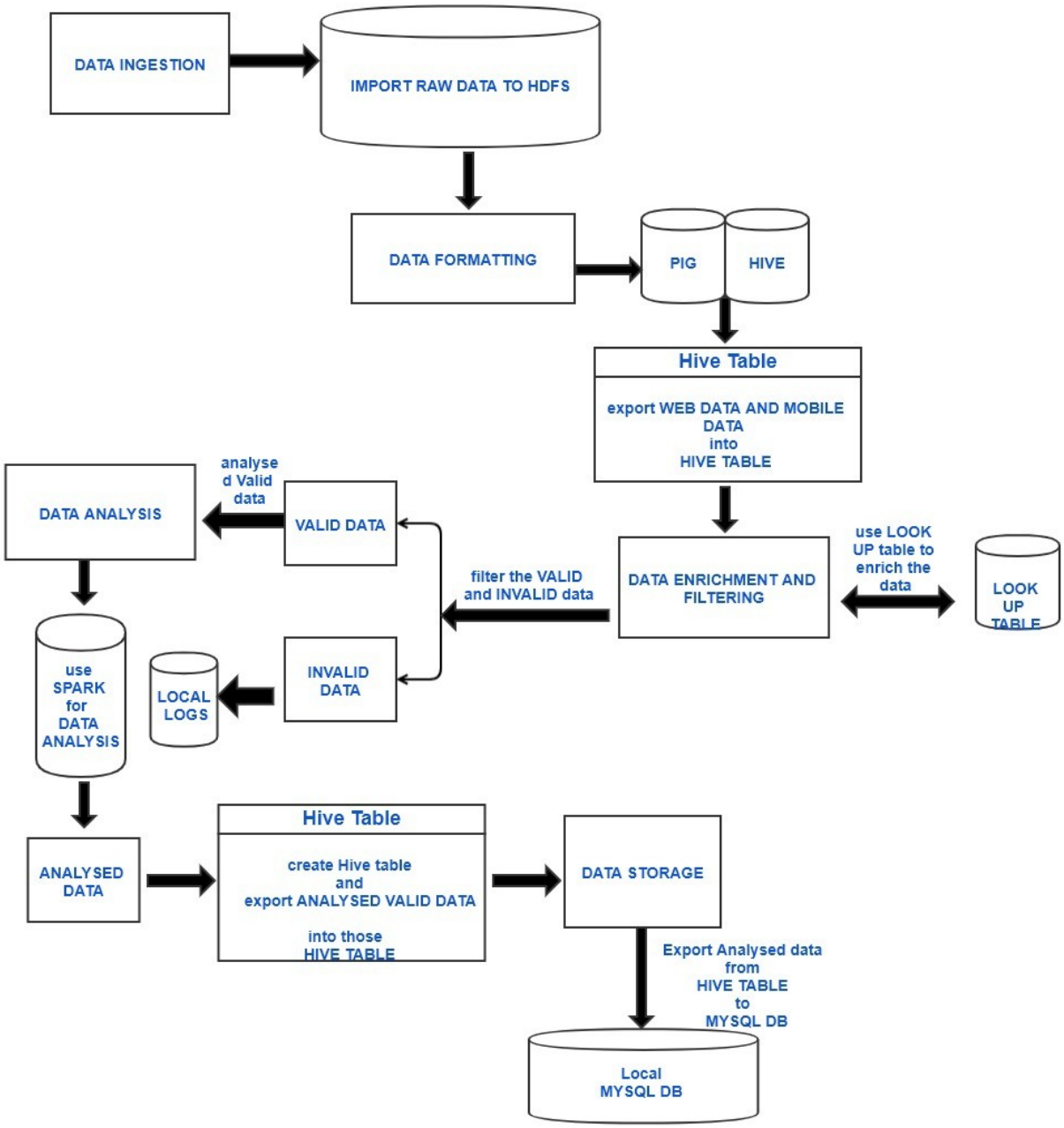
7.0 Design Architecture :



7.1 Low Level Design :



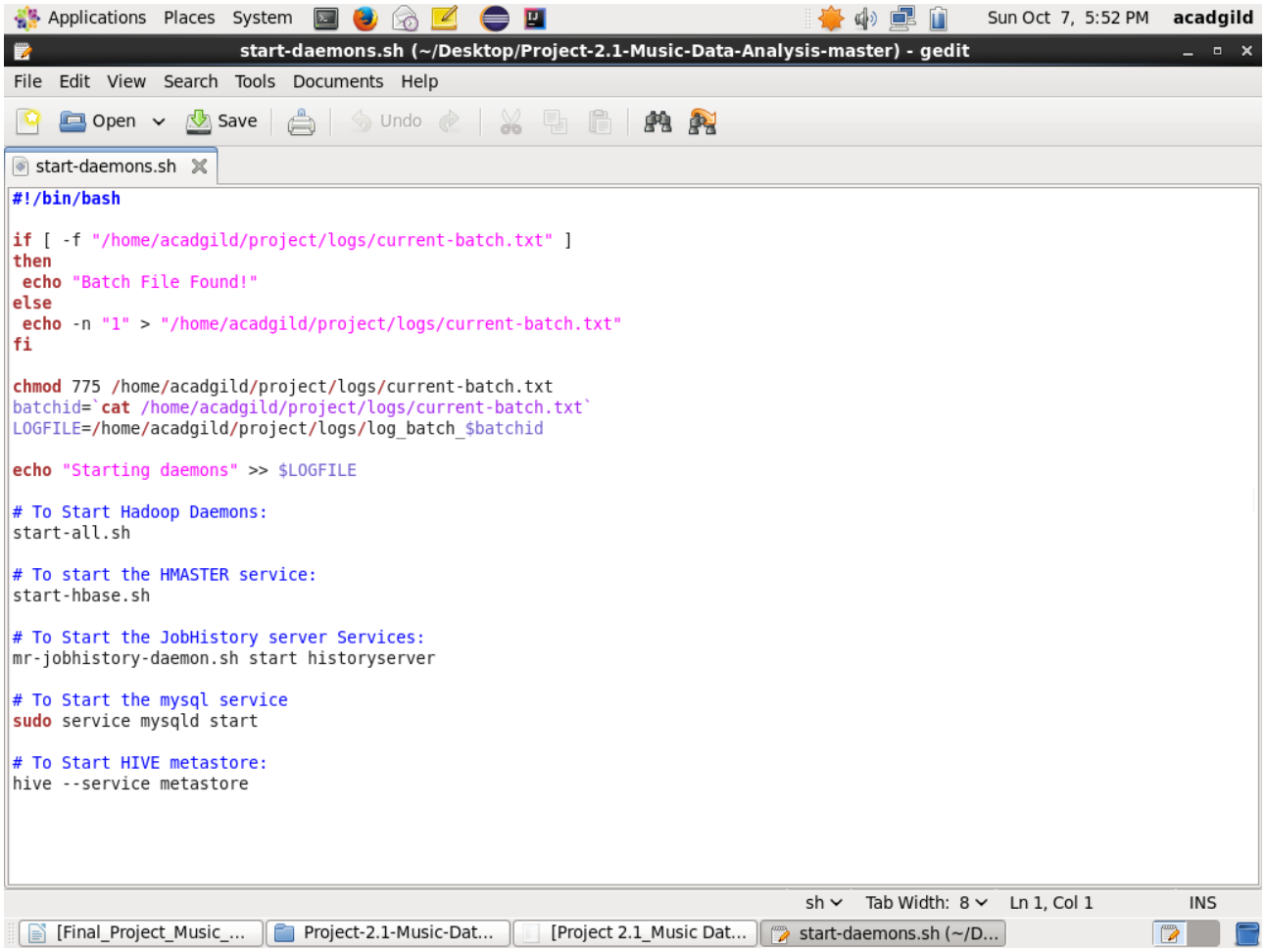
7.2 High Level Design :



8.0 Hadoop Eco system and Spark Implementation in Music Data Analysis :

1. We have created a batch file “start-daemon.sh” which starts the daemons such as hive, hbase, Mysql and rest of the all hadoop daemons.

Batch file script:



2. Starting all daemons :

sh start-daemon.sh

As per the batch file script all the hadoop daemons and the Hive, MySql and Hive daemons are started shown in the below terminal execution and screenshot :

Terminal Execution :

[acadgild@localhost scripts]$ sh start-daemons.sh

start-daemons.sh: line 7: /home/acadgild/project/logs/current-batch.txt: No such file or directory

chmod: cannot access `/home/acadgild/project/logs/current-batch.txt': No such file or directory

cat: /home/acadgild/project/logs/current-batch.txt: No such file or directory

start-daemons.sh: line 14: /home/acadgild/project/logs/log_batch_: No such file or directory

This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh

18/10/07 19:50:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Starting namenodes on [localhost]

localhost: starting namenode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-namenode-localhost.localdomain.out

localhost: starting datanode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-datanode-localhost.localdomain.out

Starting secondary namenodes [0.0.0.0]

0.0.0.0: starting secondarynamenode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-secondarynamenode-localhost.localdomain.out

18/10/07 19:52:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

starting yarn daemons

starting resourcemanager, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-resourcemanager-localhost.localdomain.out

localhost: starting nodemanager, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-nodemanager-localhost.localdomain.out

localhost: zookeeper running as process 4316. Stop it first.

running master, logging to /home/acadgild/install/hbase/hbase-1.4.4/logs/hbase-acadgild-master-localhost.localdomain.out

: running regionserver, logging to /home/acadgild/install/hbase/hbase-1.4.4/logs/hbase-acadgild-regionserver-localhost.localdomain.out

historyserver running as process 4562. Stop it first.

[sudo] password for acadgild:

Starting mysqld: [OK]

2018-10-07 19:56:00: Starting Hive Metastore Server

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.3-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]

Sun Oct 07 19:56:31 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.

Sun Oct 07 19:56:33 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.

Sun Oct 07 19:56:34 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.

Sun Oct 07 19:56:34 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.

Sun Oct 07 19:56:43 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.

Sun Oct 07 19:56:44 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.

Sun Oct 07 19:56:44 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.

Sun Oct 07 19:56:44 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.

Sun Oct 07 19:56:49 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.

Sun Oct 07 19:56:49 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.

Sun Oct 07 19:56:49 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.

Sun Oct 07 19:56:50 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.

Sun Oct 07 19:56:50 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.

Sun Oct 07 19:56:51 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.

Sun Oct 07 19:56:51 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.

Sun Oct 07 19:56:51 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.

Sun Oct 07 19:56:51 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.

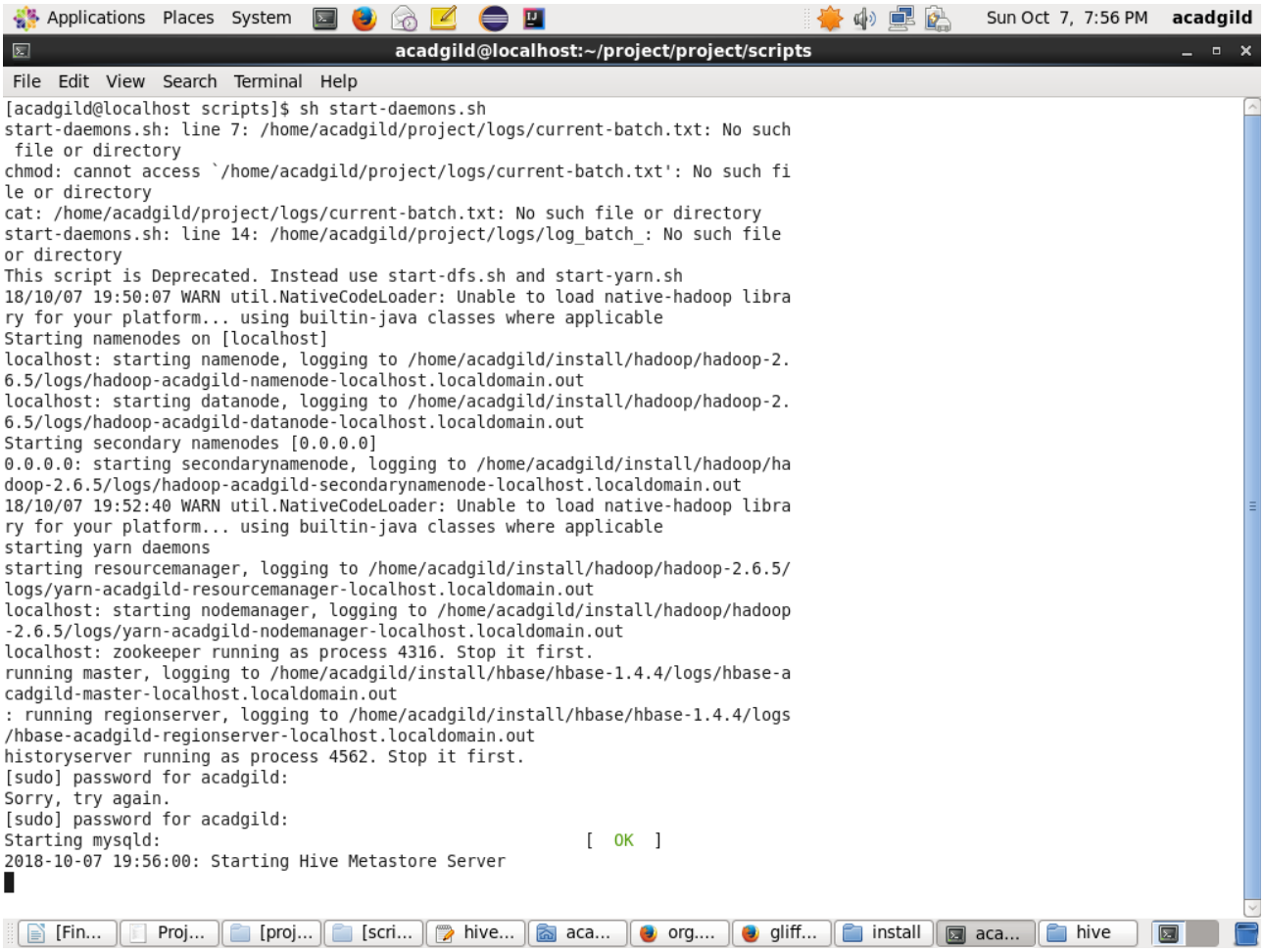
Sun Oct 07 19:56:52 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance with existing applications not 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.

Sun Oct 07 19:57:51 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.

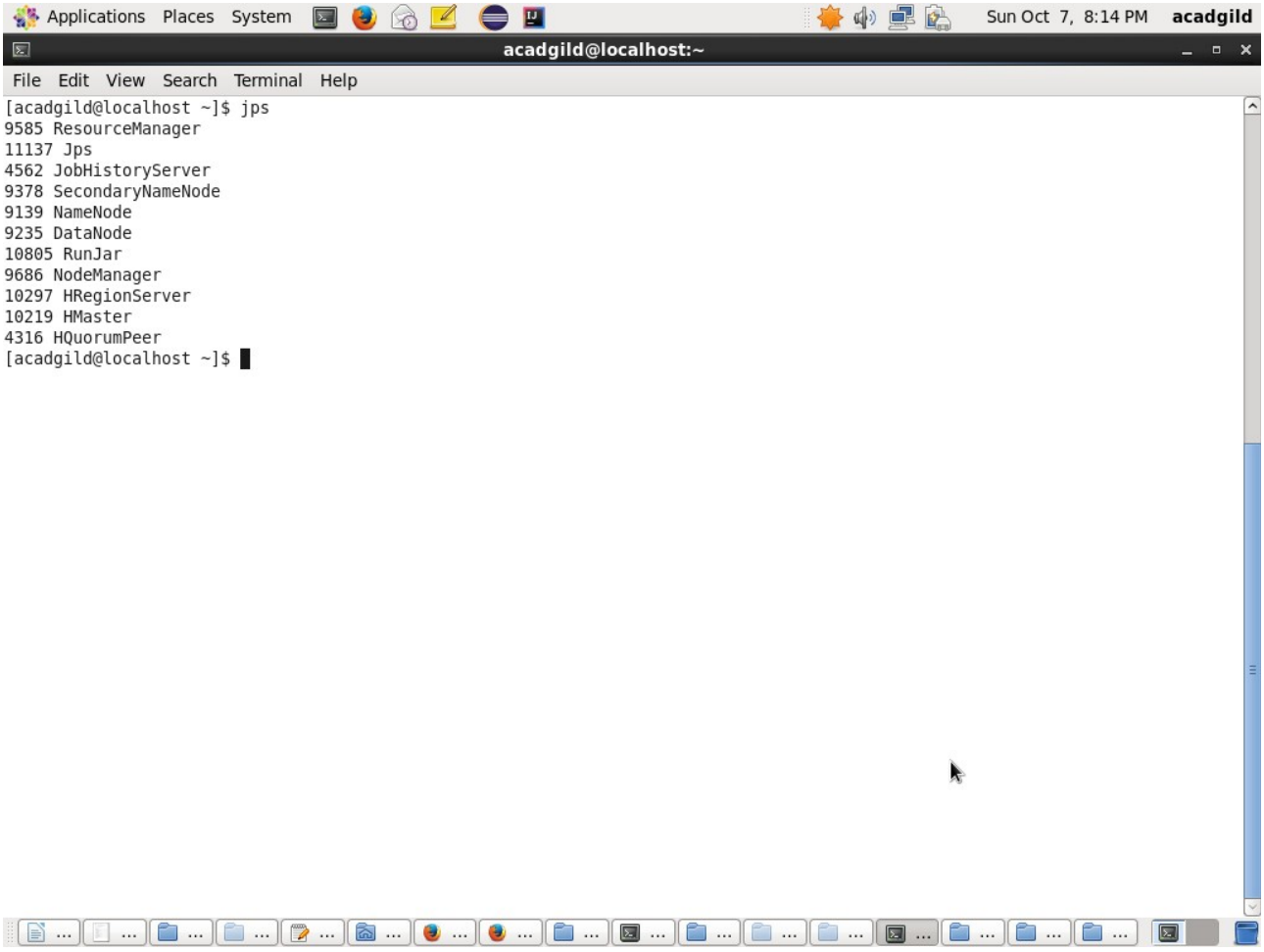
Sun Oct 07 19:57:52 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance with existing applications not 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.

[acadgild@localhost scripts]$

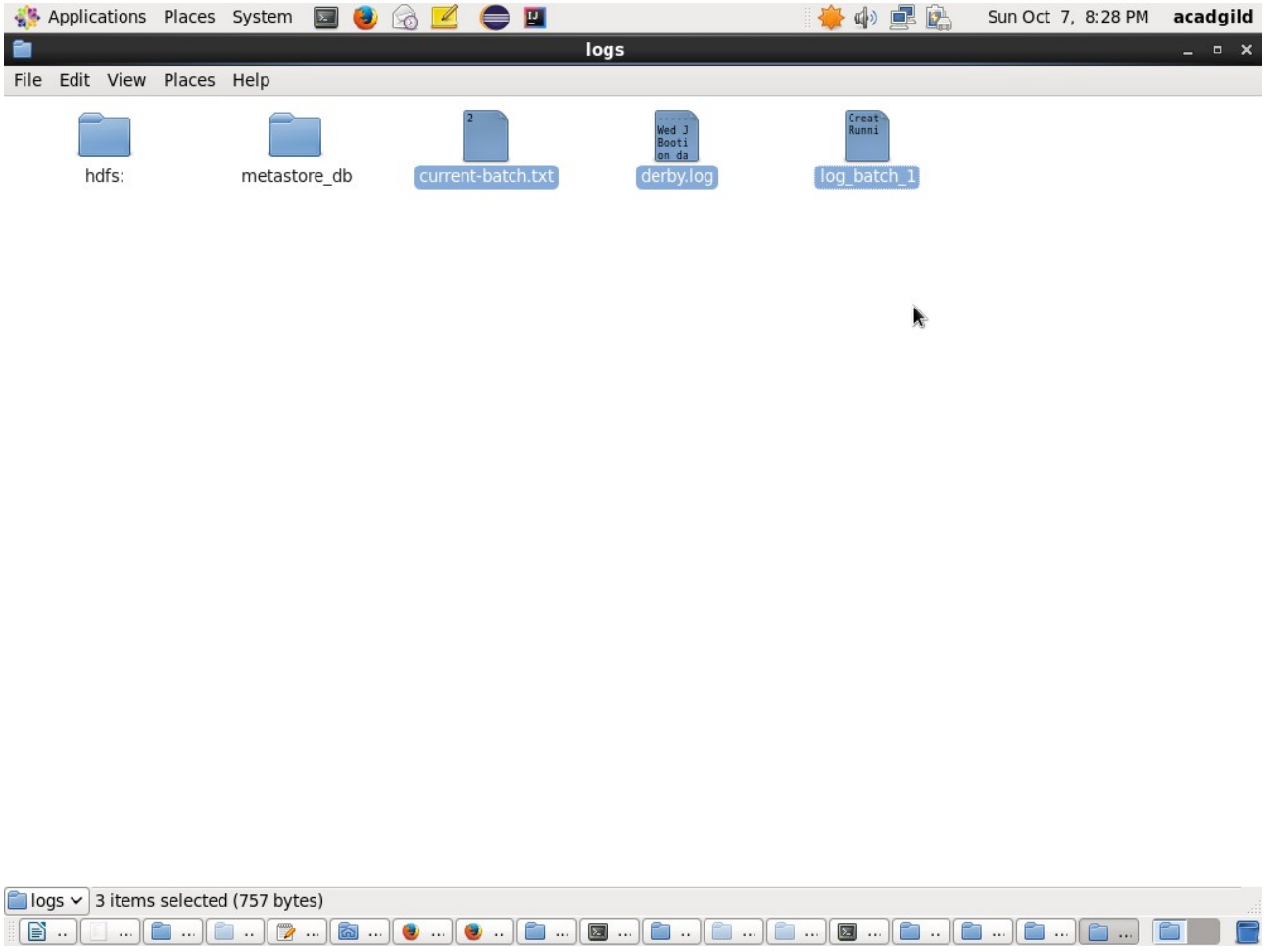
Screenshot :


Applications Places System
Sun Oct 7, 7:56 PM acadgild
acadgild@localhost:~/project/project/scripts
File Edit View Search Terminal Help
[acadgild@localhost scripts]$ sh start-daemons.sh
start-daemons.sh: line 7: /home/acadgild/project/logs/current-batch.txt: No such file or directory
chmod: cannot access `/home/acadgild/project/logs/current-batch.txt': No such file or directory
cat: /home/acadgild/project/logs/current-batch.txt: No such file or directory
start-daemons.sh: line 14: /home/acadgild/project/logs/log_batch_: No such file or directory
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
18/10/07 19:50:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-namenode-localhost.localdomain.out
localhost: starting datanode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-secondarynamenode-localhost.localdomain.out
18/10/07 19:52:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-resourcemanager-localhost.localdomain.out
localhost: starting nodemanager, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-nodemanager-localhost.localdomain.out
localhost: zookeeper running as process 4316. Stop it first.
running master, logging to /home/acadgild/install/hbase/hbase-1.4.4/logs/hbase-acadgild-master-localhost.localdomain.out
: running regionserver, logging to /home/acadgild/install/hbase/hbase-1.4.4/logs/hbase-acadgild-regionserver-localhost.localdomain.out
historyserver running as process 4562. Stop it first.
[sudo] password for acadgild:
Sorry, try again.
[sudo] password for acadgild:
Starting mysqld: [OK]
2018-10-07 19:56:00: Starting Hive Metastore Server
[Fin... Proj... [proj... [scri... hive... aca... org.... gliff... install aca... hive

3. We can see the list active services using the jps command, see below screen shot and also Starting the hive metastore created a metastore_db in the location where we desired,



4. The start-daemon.sh script will check whether the current-batch.txt file is available in the logs folder or not. If not it will create the file and dump value ‘1’ in that file and create LOGFILE with the current batchid.



5. Data Ingestion, Formatting, Enrichment and Filtering :

5.1 Stage 1: Data Ingestion

By using the “populate-lookup.sh” script we will create lookup tables in Hbase. These tables have to be used in following processes :

- **Data formatting**
- **Data enrichment**
- **Analysis stage**

Lookup Tables :

Sno.	Table Name	Description	Related File
1	**station-geo-map**	Contains mapping of a geo_cd with station_id	**stn-geocd.txt**
2	**Subscribed-users**	Contains user_id, subscription_start_date and subscription_end_date. Contains details only for subscribed users	**user-subscn.txt**
3	**song-artist-map**	Contains mapping of song_id with artist_id Along with royalty associated with each play of the song	**song-artist.txt**
4	**User-artist-map**	Contains an array of artist_id(s) followed by a user_id.	**user-artist.txt**

populate-lookup.sh script :

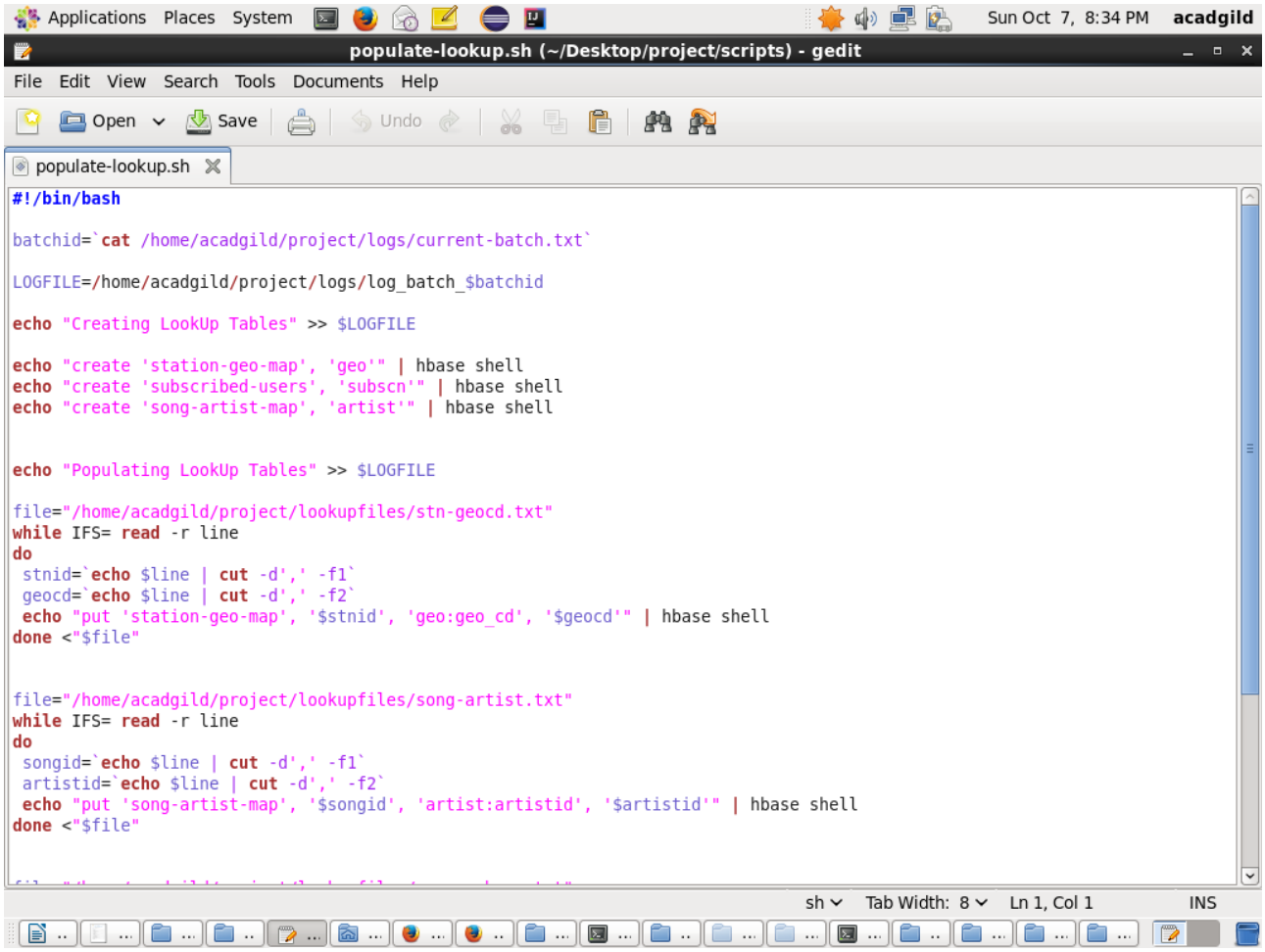
The “**populate-lookup.sh**” shell script creates the above 4 lookup tables in the Hbase and populate the data into the lookup tables from the dataset files.

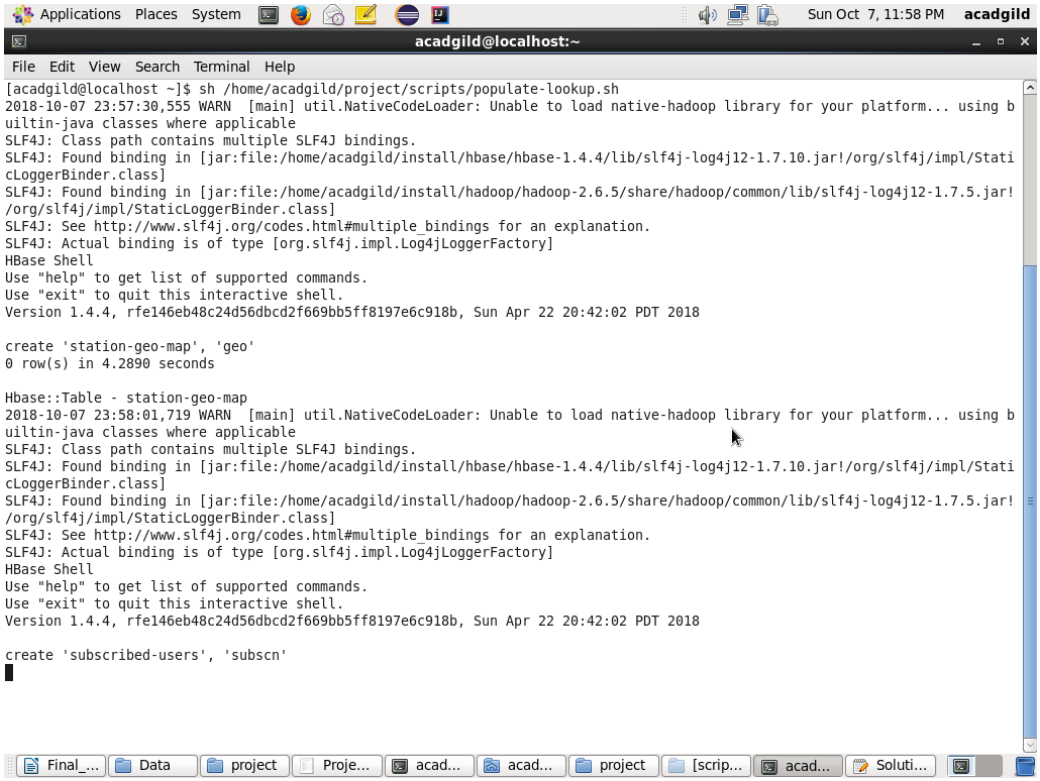
In the below screen shots, we can see the create-lookup.sh scripts and the following screen shots shows the tables creation and population of the data in the Hbase. Also, the values loaded into the Hbase Tables are also shown, please see the below screen shots.

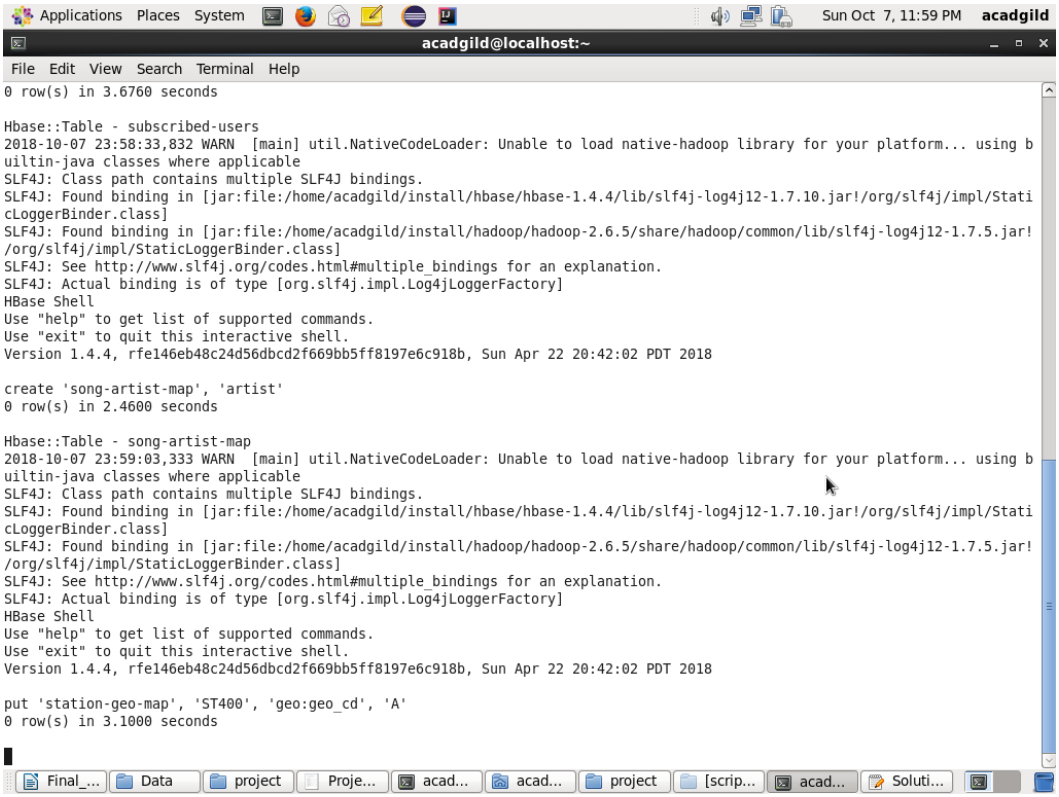
```
#!/bin/bash

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

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

echo "Creating LookUp Tables" >> $LOGFILE

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"


file="/home/acadgild/project/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"


file="/home/acadgild/project/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"

hive -f /home/acadgild/project/scripts/user-artist.hql
```

Screenshot :

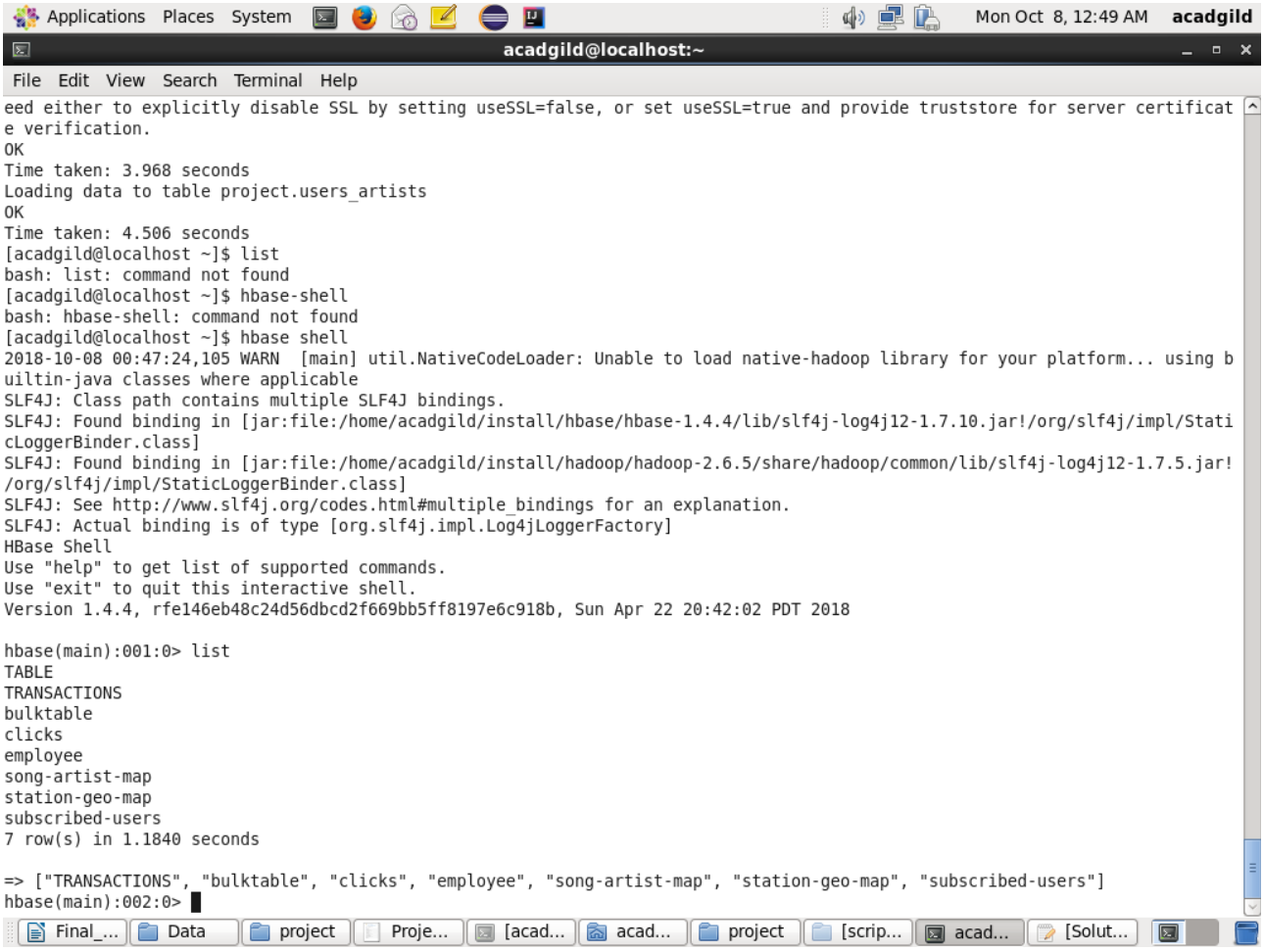

Applications Places System
Sun Oct 7, 8:34 PM acadgild
populate-lookup.sh (~/Desktop/project/scripts) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
populate-lookup.sh
#!/bin/bash
batchid=`cat /home/acadgild/project/logs/current-batch.txt`
LOGFILE=/home/acadgild/project/logs/log_batch_$batchid
echo "Creating LookUp Tables" >> $LOGFILE
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"
file="/home/acadgild/project/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"
sh Tab Width: 8 Ln 1, Col 1 INS


Applications Places System
Sun Oct 7, 11:58 PM acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ sh /home/acadgild/project/scripts/populate-lookup.sh
2018-10-07 23:57:30,555 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uiltin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.4.4/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/Stati
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.slf4j.impl.Log4jLoggerFactory]
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.4, rfe146eb48c24d56dbcd2f669bb5ff8197e6c918b, Sun Apr 22 20:42:02 PDT 2018
create 'station-geo-map', 'geo'
0 row(s) in 4.2890 seconds
Hbase::Table - station-geo-map
2018-10-07 23:58:01,719 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uiltin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.4.4/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/Stati
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.slf4j.impl.Log4jLoggerFactory]
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.4, rfe146eb48c24d56dbcd2f669bb5ff8197e6c918b, Sun Apr 22 20:42:02 PDT 2018
create 'subscribed-users', 'subscn'
Final_... Data project Proje... acad... acad... project [scrip... acad... Soluti...

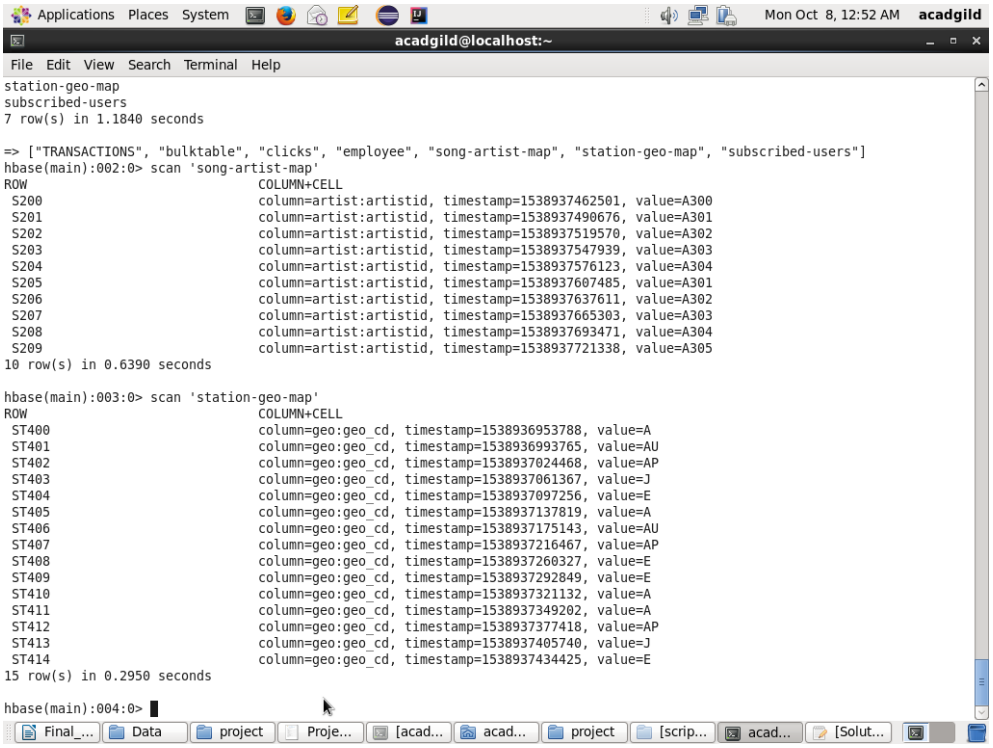

Applications Places System
Sun Oct 7, 11:59 PM acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
0 row(s) in 3.6760 seconds
Hbase::Table - subscribed-users
2018-10-07 23:58:33,832 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uiltin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.4.4/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/Stati
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.slf4j.impl.Log4jLoggerFactory]
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.4, rfe146eb48c24d56dbcd2f669bb5ff8197e6c918b, Sun Apr 22 20:42:02 PDT 2018
create 'song-artist-map', 'artist'
0 row(s) in 2.4600 seconds
Hbase::Table - song-artist-map
2018-10-07 23:59:03,333 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uiltin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.4.4/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/Stati
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.slf4j.impl.Log4jLoggerFactory]
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.4, rfe146eb48c24d56dbcd2f669bb5ff8197e6c918b, Sun Apr 22 20:42:02 PDT 2018
put 'station-geo-map', 'ST400', 'geo:geo_cd', 'A'
0 row(s) in 3.1000 seconds
Final_... Data project Proje... acad... acad... project [scrip... acad... Soluti...

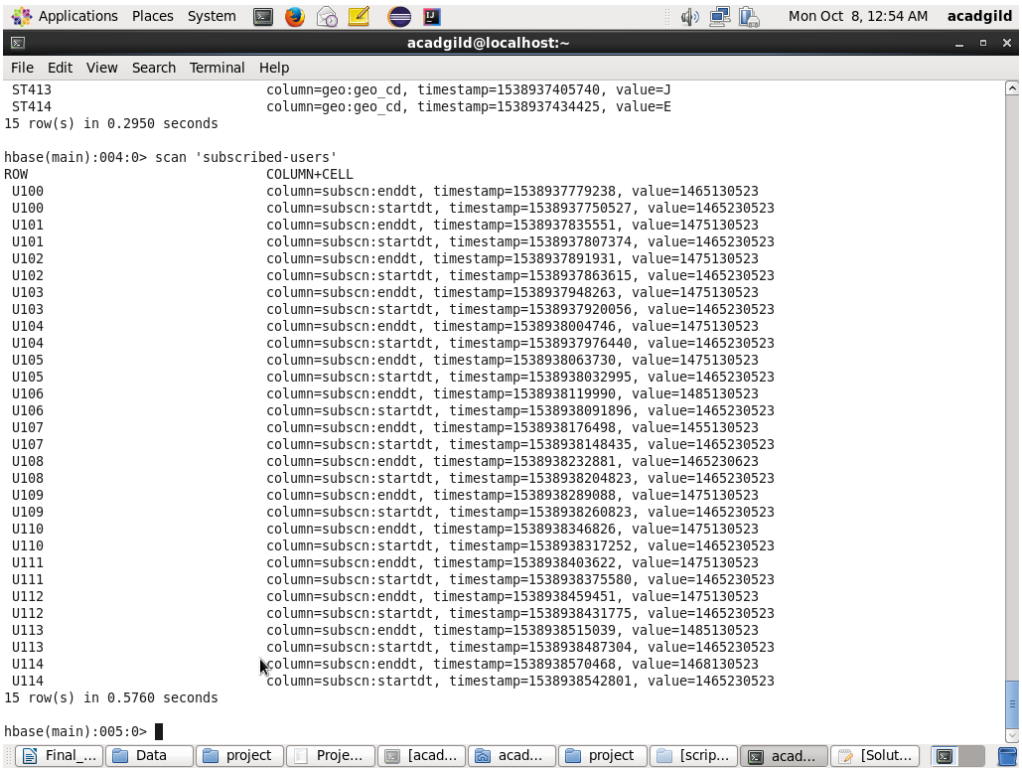
We can see the lookup tables created using the “populate-lookup.sh” in the below screen shot,

Lookup Tables in the hbase shell :



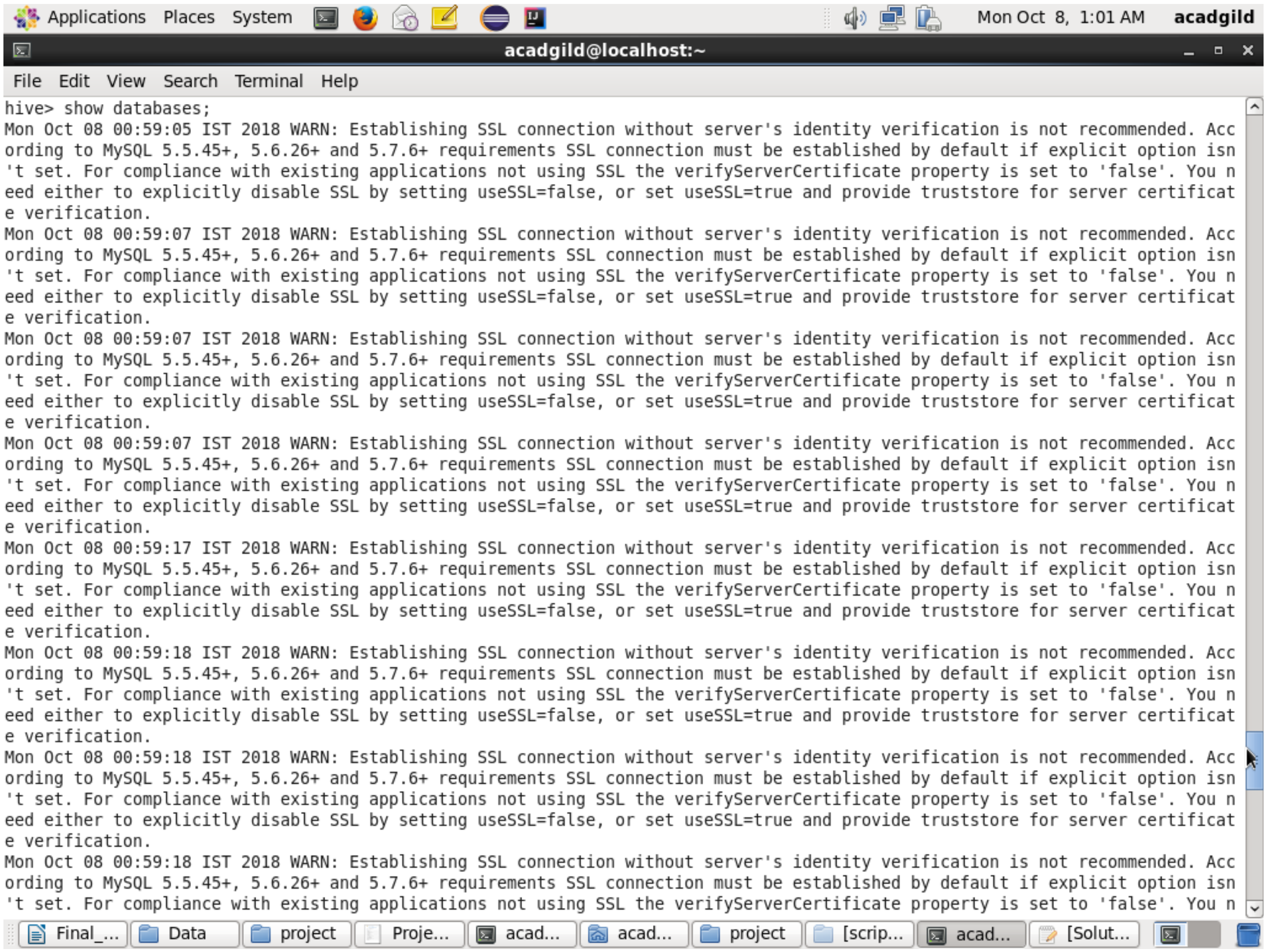
The values loaded in the Lookup tables are shown below :

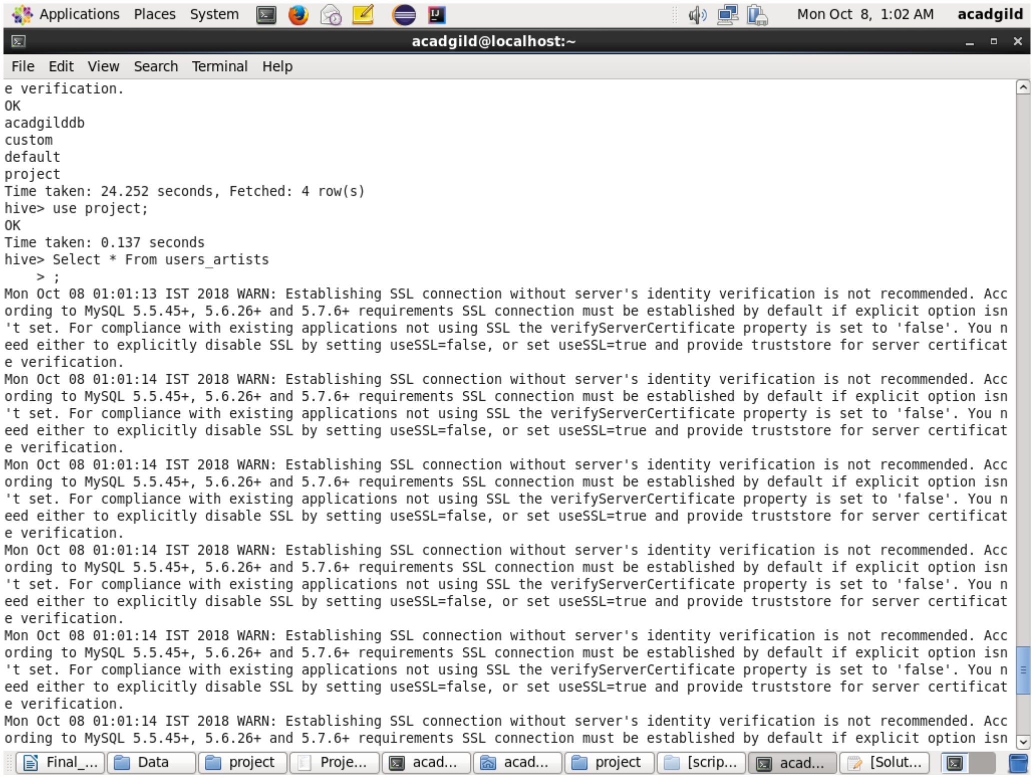




We have successfully created the lookup tables in the Hbase.

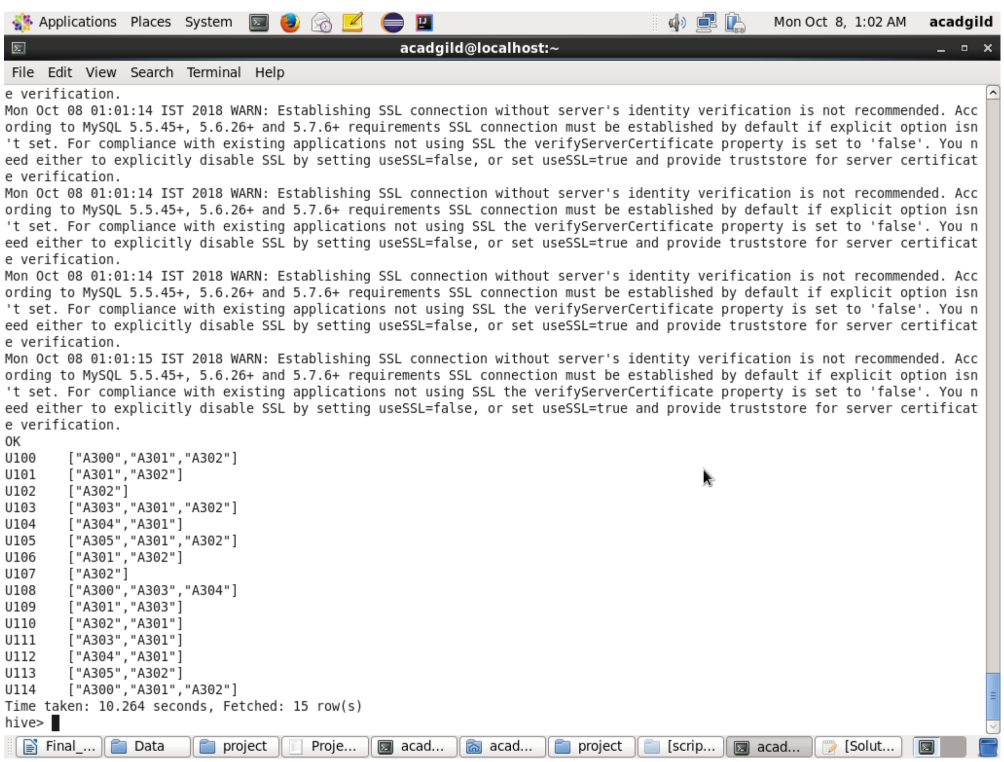
The populate-lookup.sh also creates a lookup table “users_artists” in the HIVE, loading the data from the user-artist.txt, the below screen shot shows that the table has been created in the HIVE.





```
e verification.
OK
acadgilddb
custom
default
project
Time taken: 24.252 seconds, Fetched: 4 row(s)
hive> use project;
OK
Time taken: 0.137 seconds
hive> Select * From users_artists
    > ;
Mon Oct 08 01:01:13 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:01:14 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:01:14 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:01:14 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:01:14 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:01:14 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn
```

hive> Select * From users_artists;



```
e verification.
Mon Oct 08 01:01:14 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:01:14 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:01:14 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:01:15 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
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: 10.264 seconds, Fetched: 15 row(s)
hive>
```

Now we need to link theses lookup tables in hive using the Hbase Storage Handler.

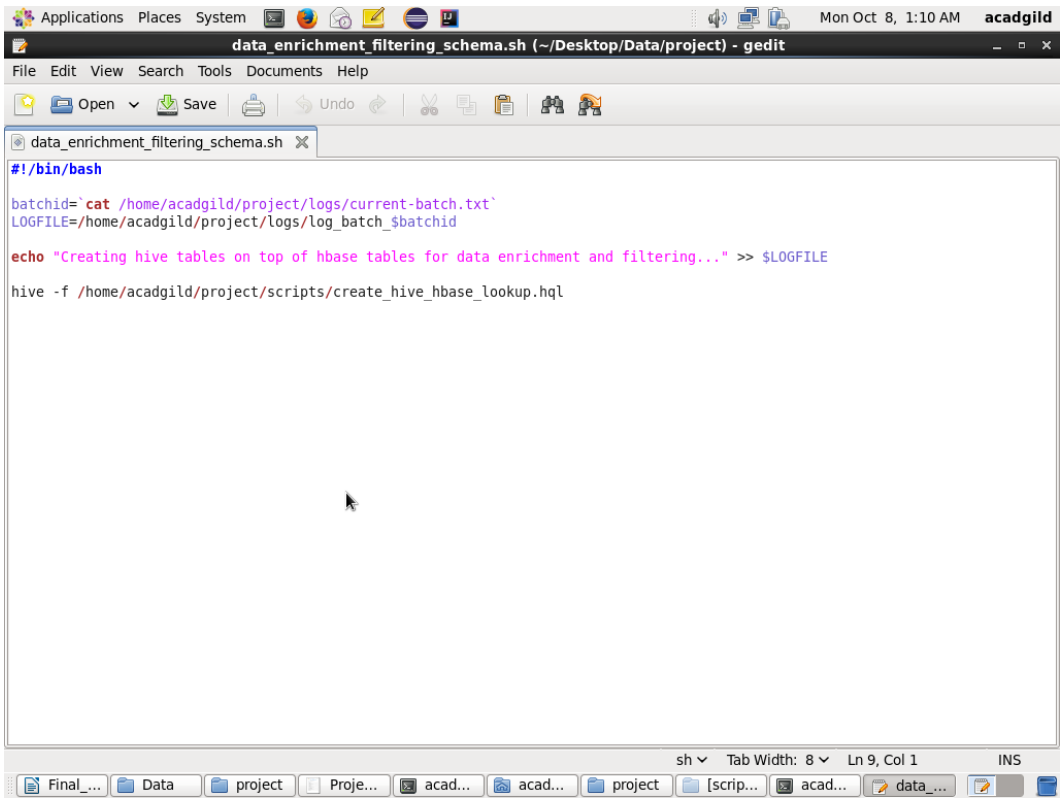
With the help of "**data_enrichment_filtering_schema.sh**" file we will create hive tables on the top of Hbase tables using "**create_hive_hbase_lookup.hql**".

<u>Creating Hive Tables on the top of Hbase</u>:

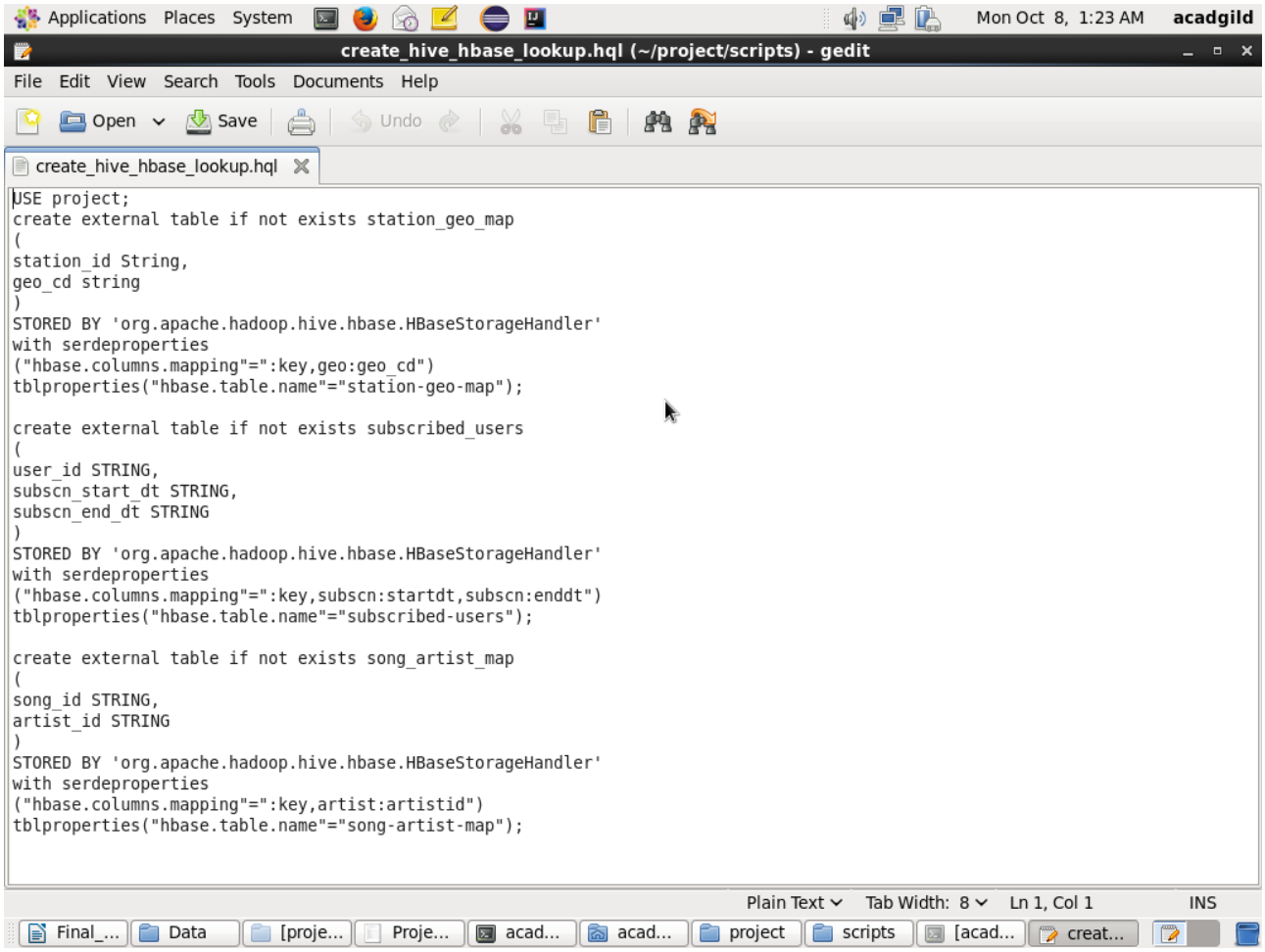
In this section with the help of Hbase storage handler & SerDe properties we are creating the hive external tables by matching the columns of Hbase tables to hive tables.

Run the script: **./data_enrichment_filtering_schema.sh**,

The script will run the "**create_hive_hbase_lookup.hql**" which will create the HIVE external tables with the help of **Hbase storage handler & SerDe properties**. The hive external tables will match the columns of **Hbase** tables to **HIVE** tables.



create_hive_hbase_lookup.hql



The below screenshot we can see tables getting created in hive by running the “**data_enrichement_filtering_schema.sh** file”

Applications Places System Mon Oct 8, 1:31 AM acadgild

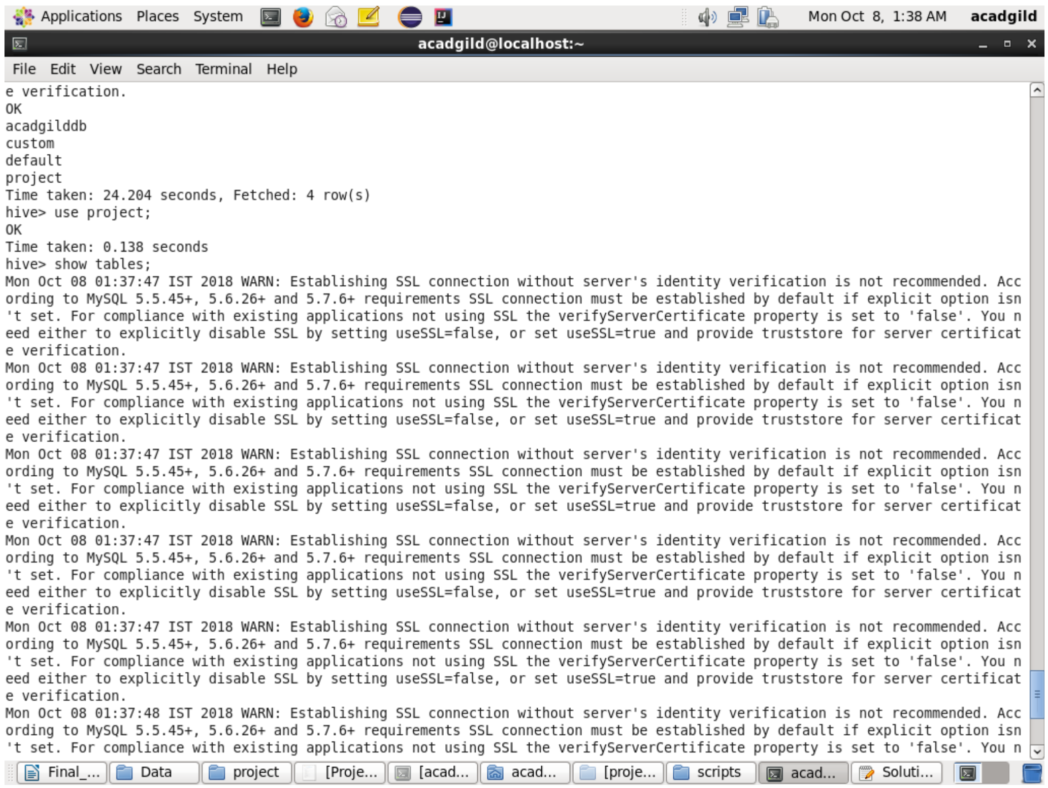
acadgild@localhost:~

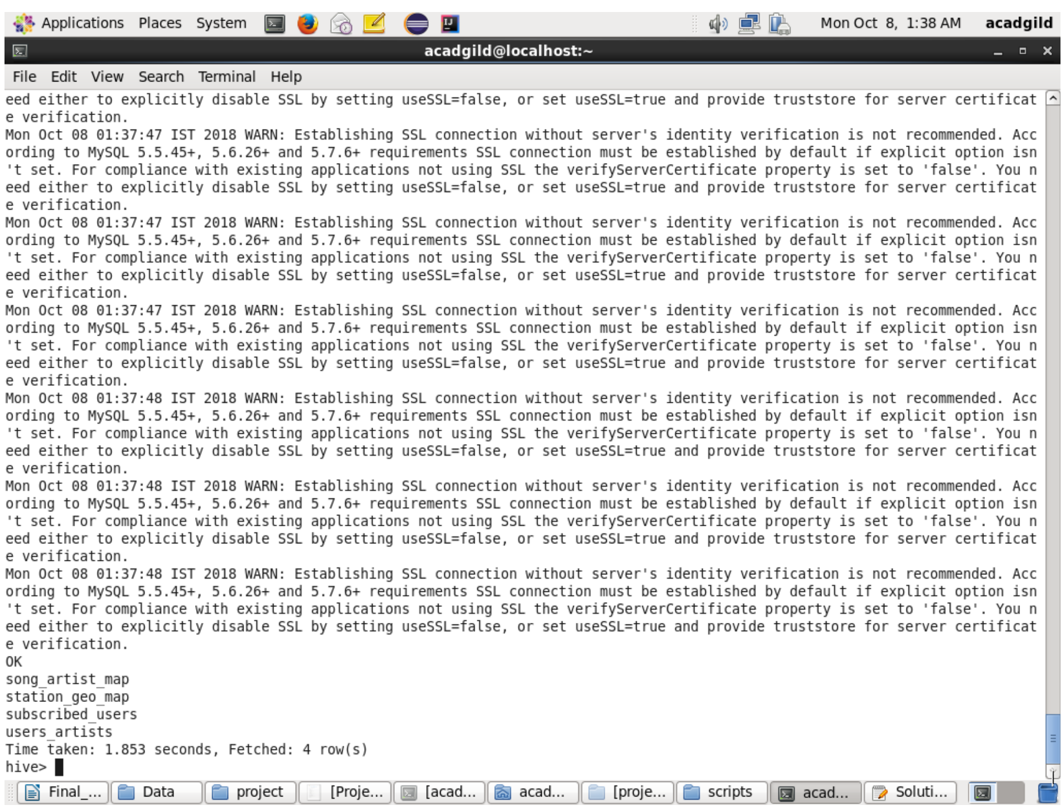
File Edit View Search Terminal Help

```
Mon Oct 08 01:30:28 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:30:28 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:30:38 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:30:38 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:30:38 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Mon Oct 08 01:30:38 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
OK
Time taken: 27.119 seconds
OK
Time taken: 1.433 seconds
OK
Time taken: 0.137 seconds
OK
Time taken: 0.128 seconds
[acadgild@localhost ~]$
```

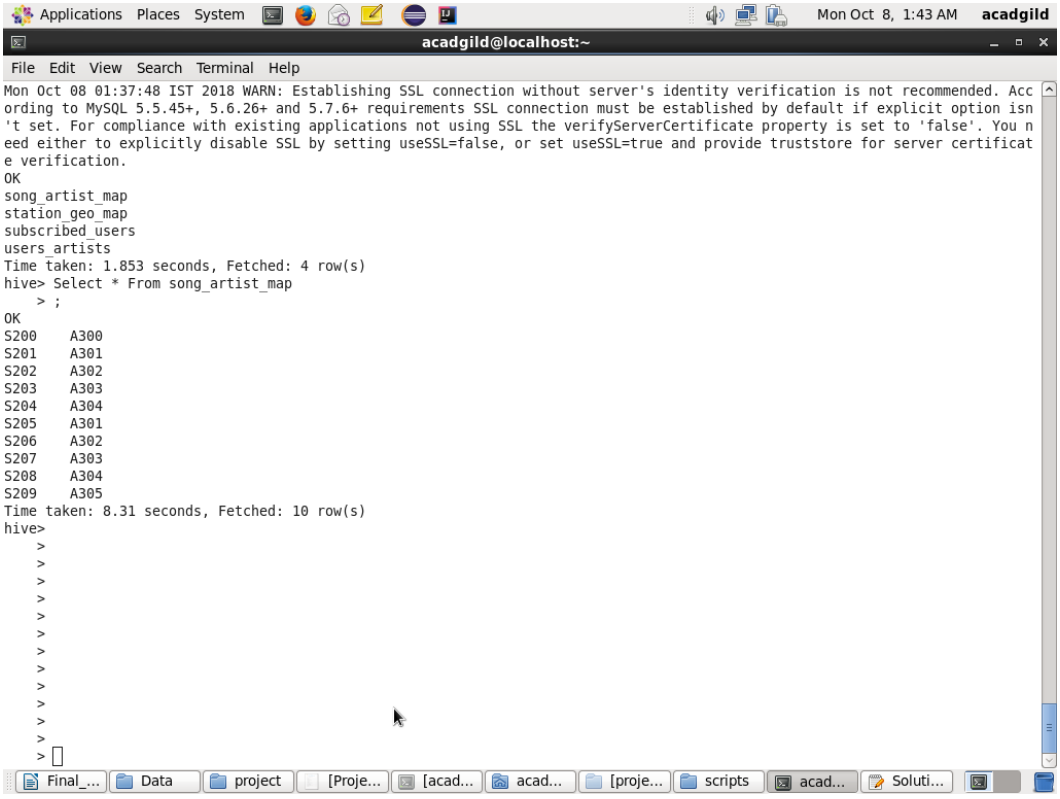
Final_... Data project Proje... acad... acad... [proje... [scrip... acad... Soluti...

Hive>Show Tables;

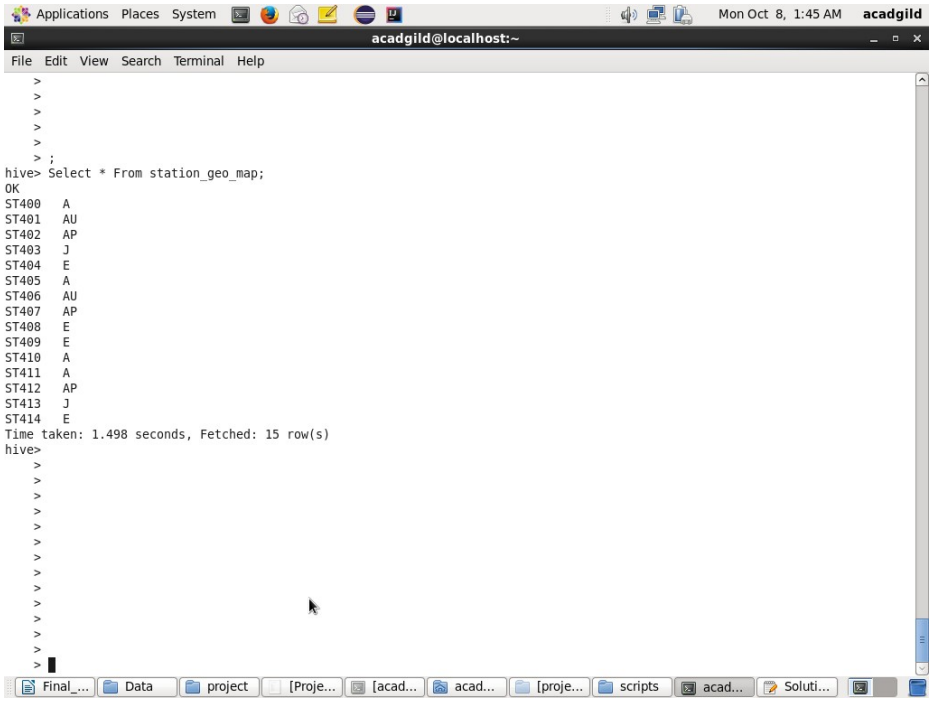




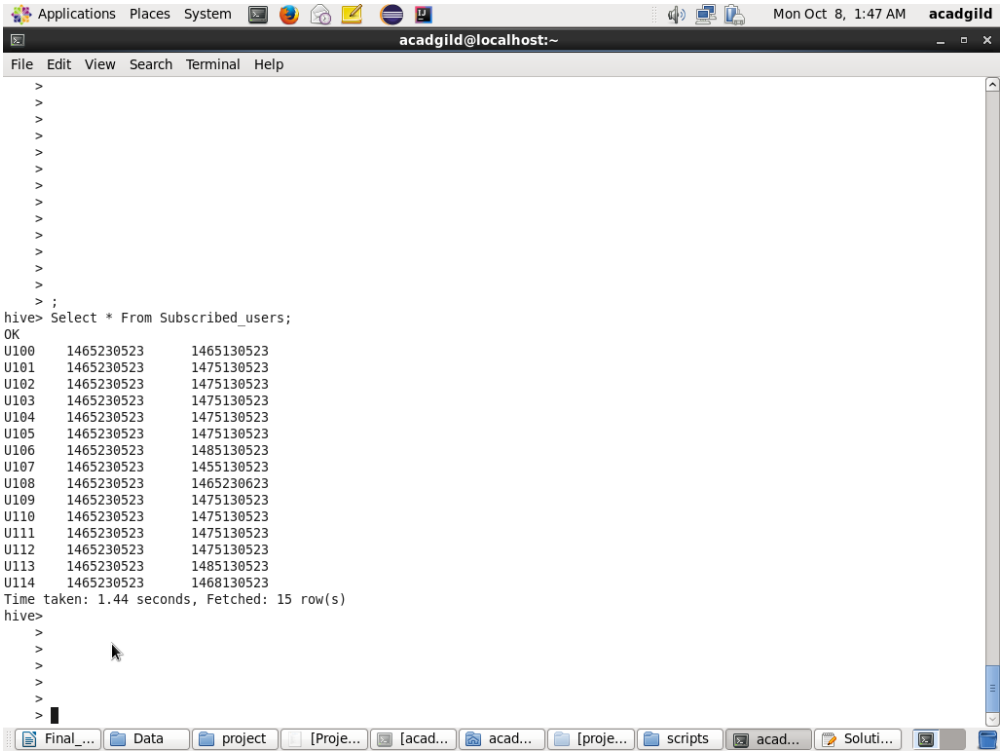
hive>Select * From song_artist_map



hive>Select * From station_geo_map



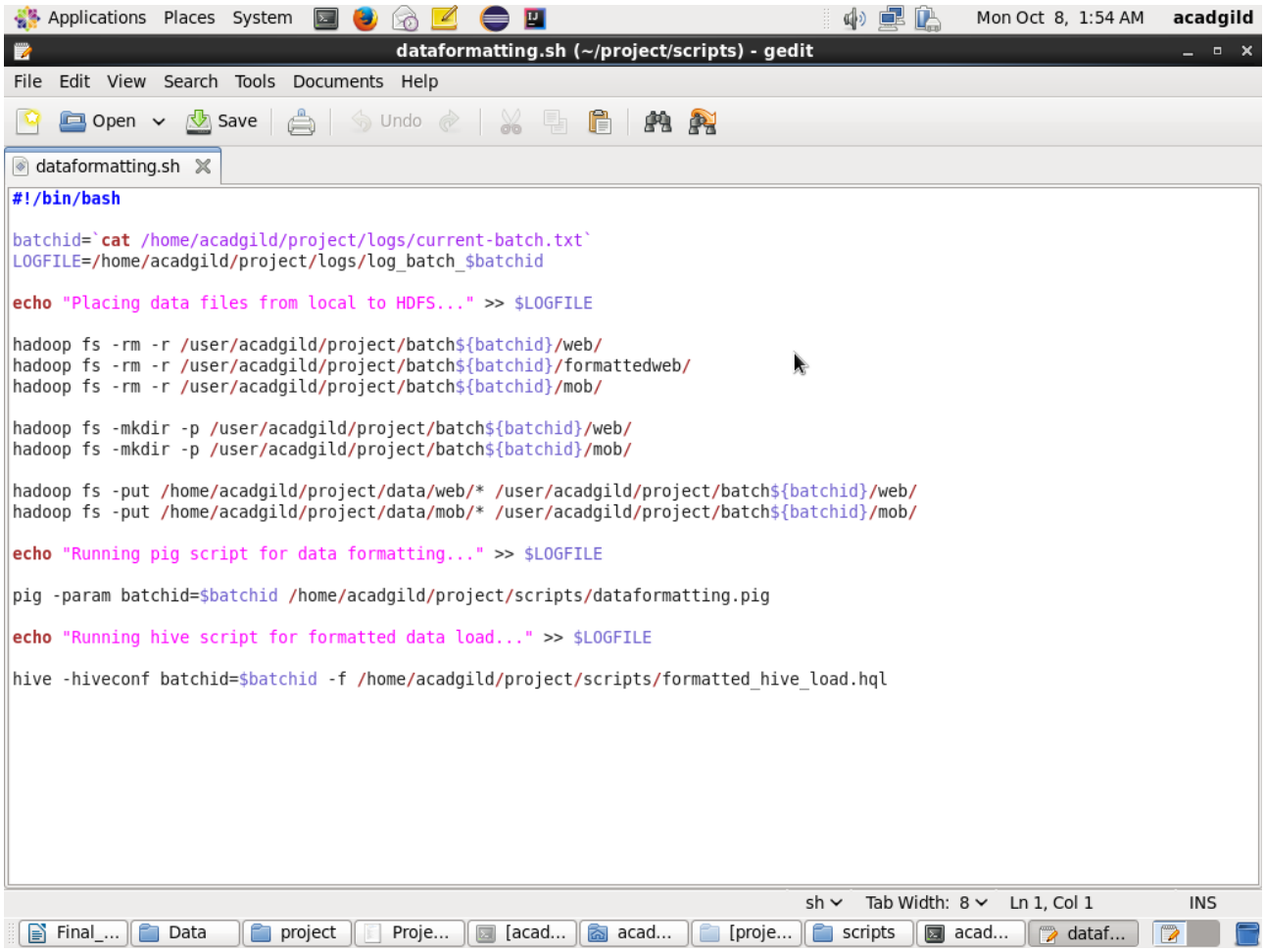
hive> Select * From Subscribed_users;

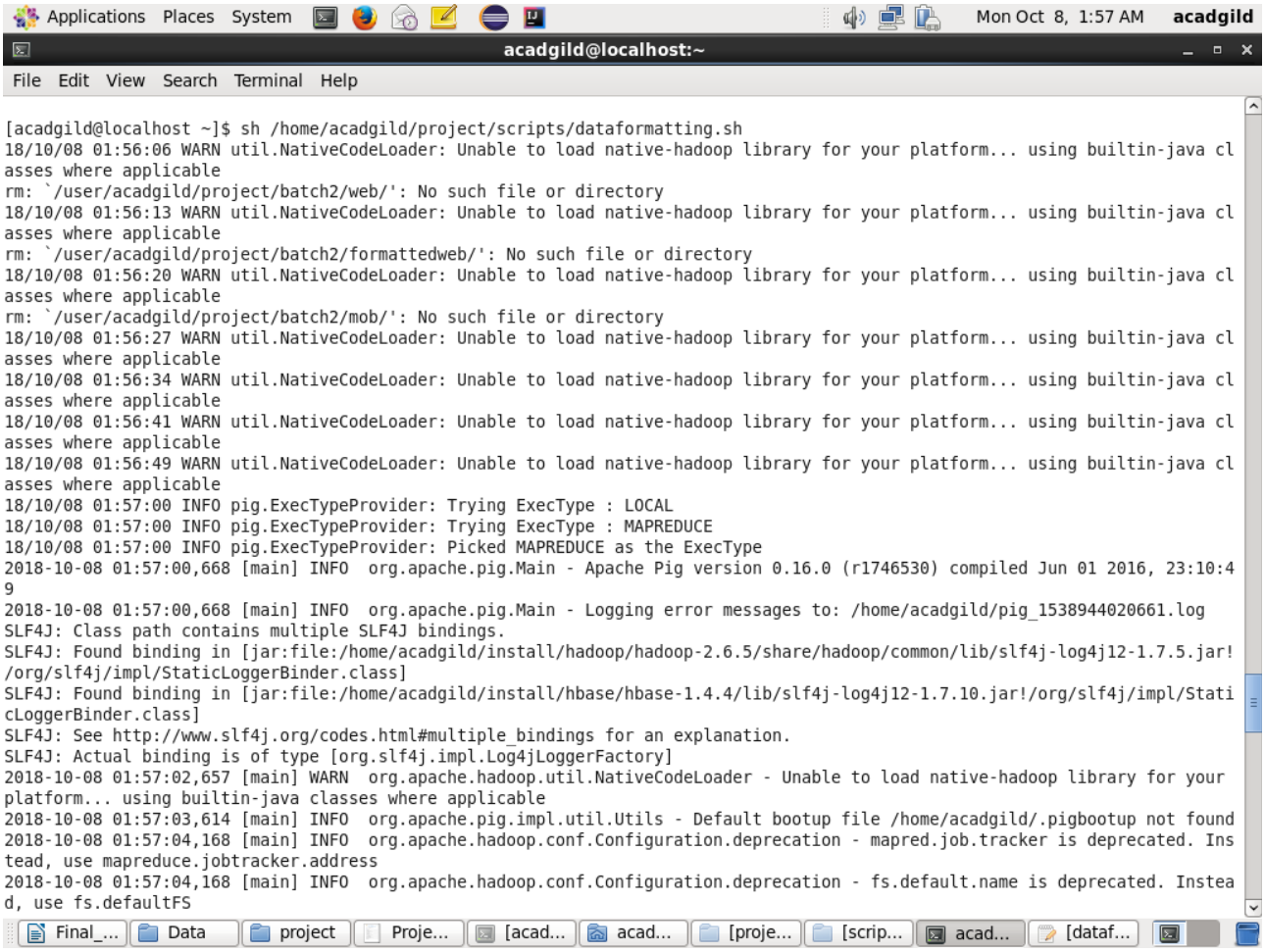


5.2 Stage 2: Data Formatting

In this stage we are merging the data coming from both web applications and mobile applications and create a common table for analyzing purpose and create partitioned data based on batchid, since we are running this scripts for every 3 hours.

Run the script: **./dataformatting.sh**





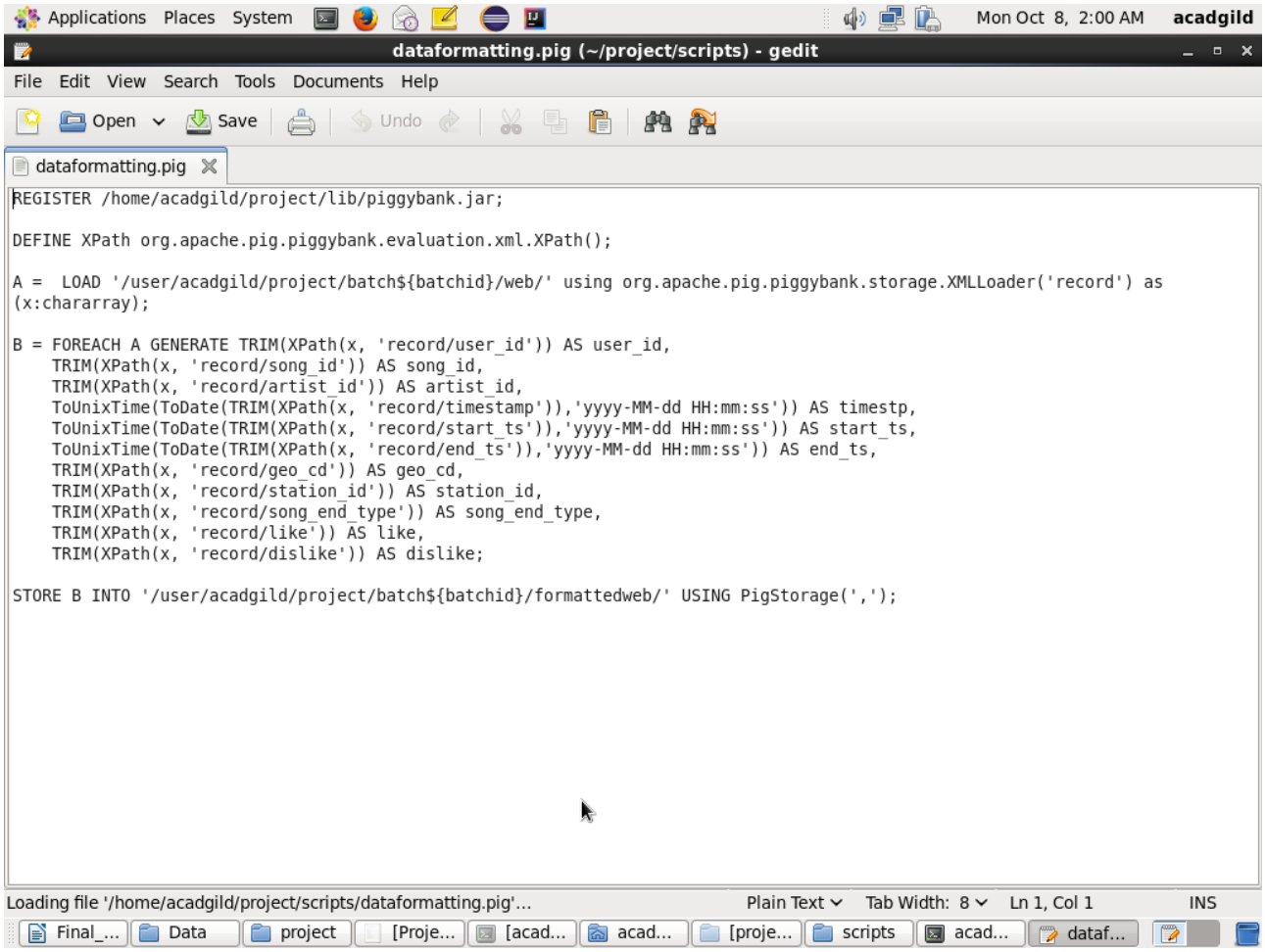
We are running two scripts to format the data. They are:

Dataformatting.pig

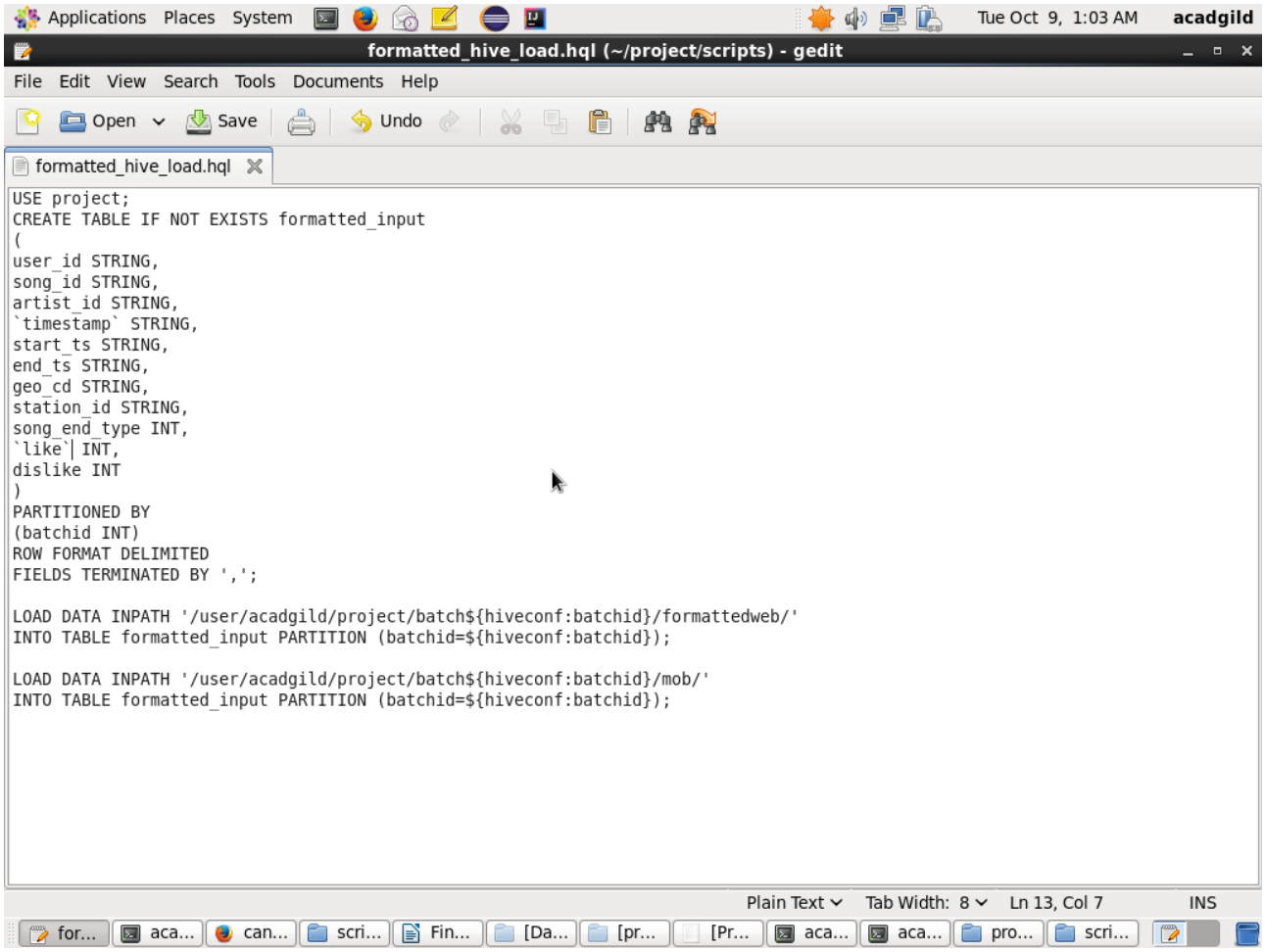
Formatted_hive_load.hql

Pig script to parse the data from coming from **web_data.xml** to **csv** format and partition both web and mob data based on based on batch ID’s

Dataformatting.pig

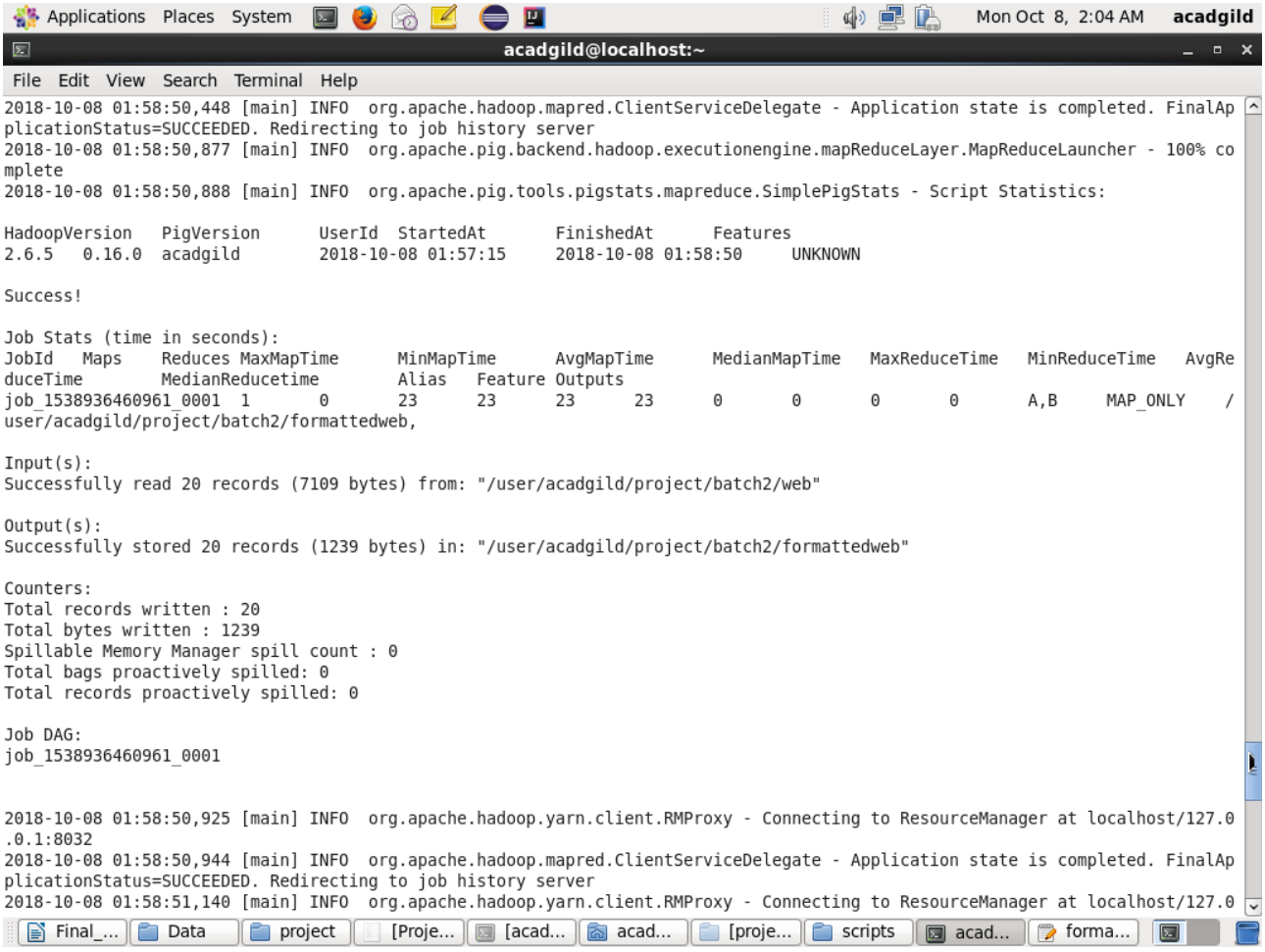


formatted_hive_load.hql

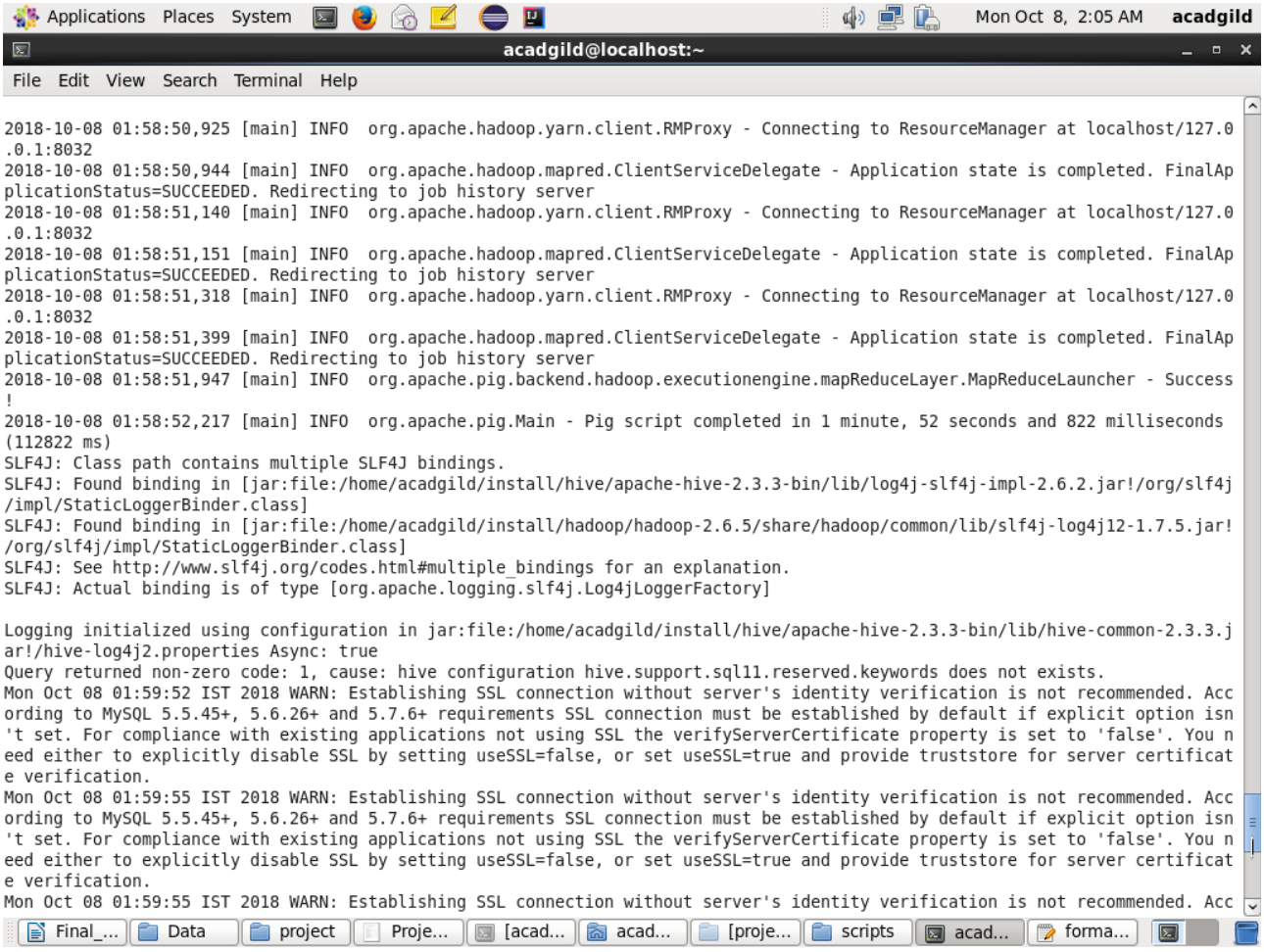


In the below screenshot we can see the data both the scripts in action, first pig script will parse the data and then hive script will load the data into hive terminal successfully.

Pig script successful completion:



Hive script successfully load the data into hive terminal :



In the above screenshot we can see the **dataformatting.pig** along with the **formatted_hive_load.hql**

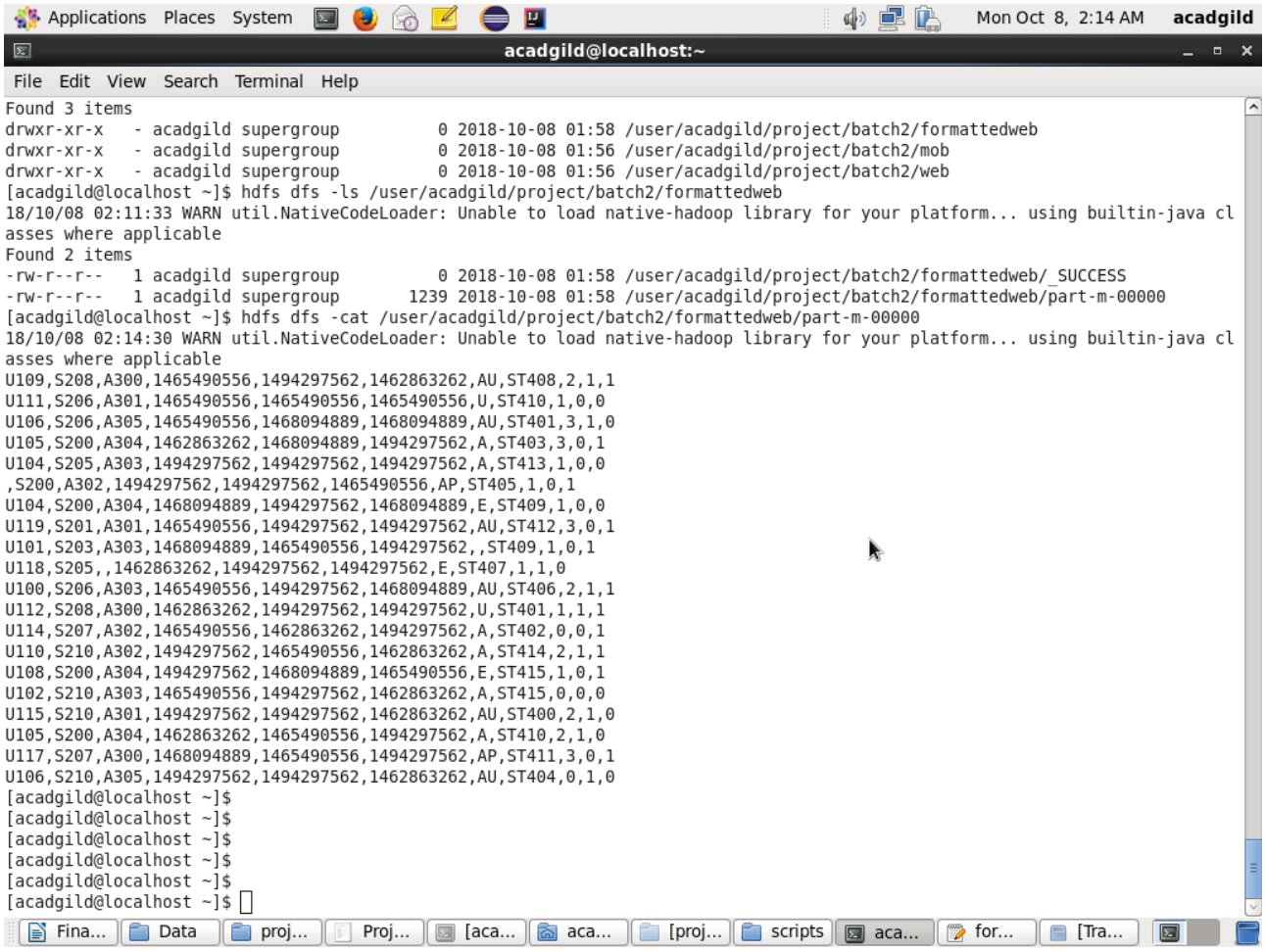
executed successfully.

The output of **dataformatting.sh** script in HDFS folders:

The output of the **formattedweb data** obtained from the **Dataformatting.pig** is shown in the below screenshot,

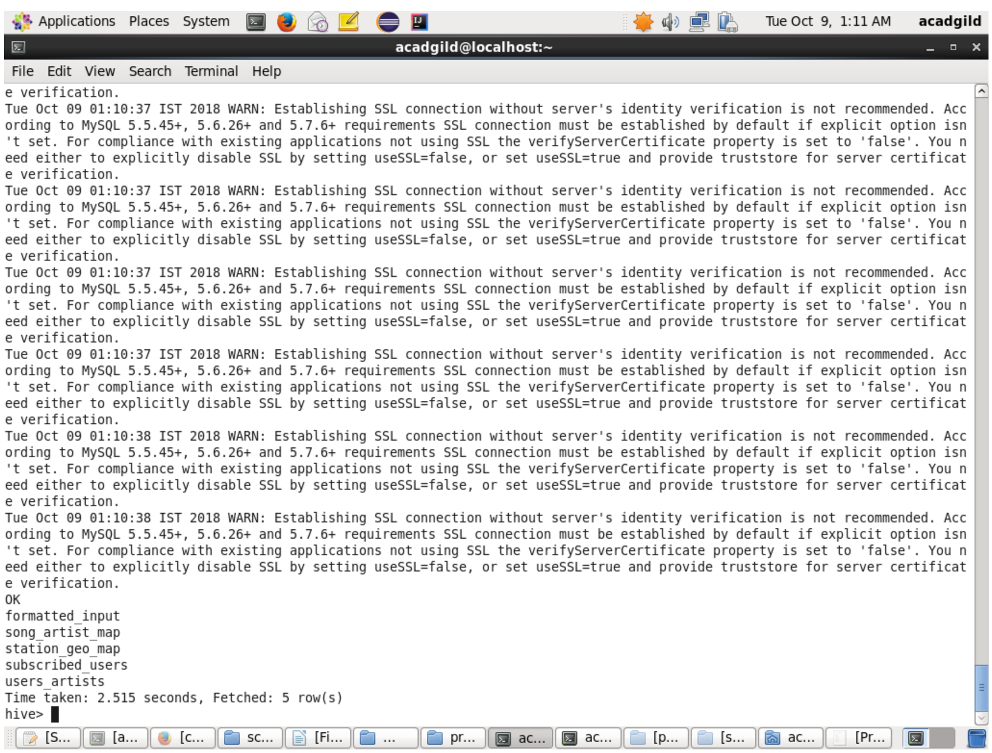
Command :

hdfs dfs -cat /user/acadgild/project/batch1/formattedweb/part-m-00000



The new Tables has been created and show below :

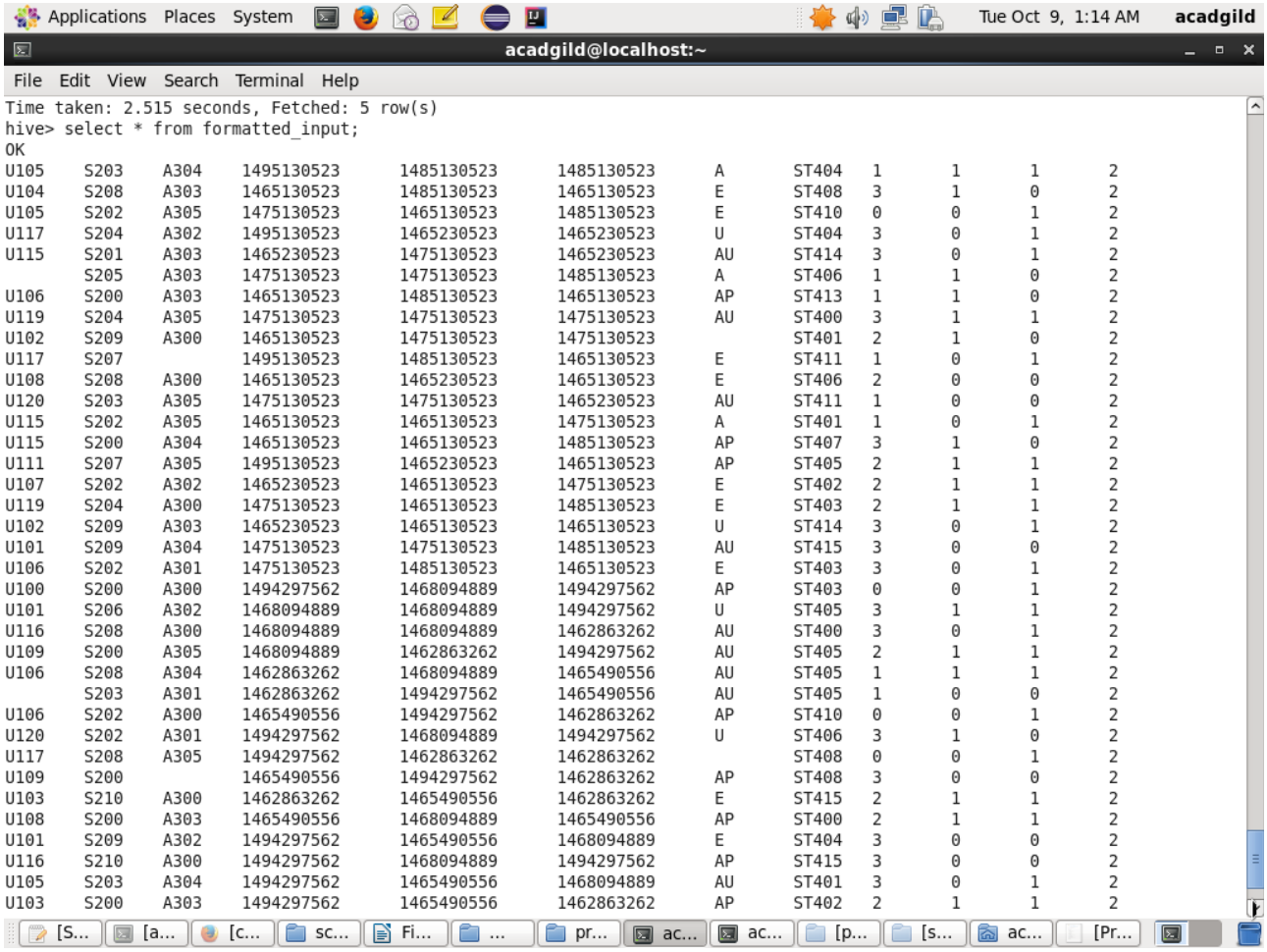
```
Time taken: 22.246 seconds
hive> show tables;
Tue Oct 09 01:10:37 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn
't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:37 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:37 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:37 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:37 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:37 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:38 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:38 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn
```



```
e verification.
Tue Oct 09 01:10:37 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:37 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:37 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:37 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:38 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 01:10:38 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
OK
formatted_input
song_artist_map
station_geo_map
subscribed_users
users_artists
Time taken: 2.515 seconds, Fetched: 5 row(s)
hive>
```

DataFormatting.sh output in hive terminal,

hive> select * from formatted_input;



In the above screenshot, we can see the formatted input data with some null values in **user_id**, **aritist_id and geo_cd** columns which we will fill the enrichment script based on rules of enrichment for **artist_id and geo_cd** only. We will get neglect **user_id** because they didn't mentioned anything about user_id for enrichment purpose.

Data formatting phase is executed successfully by loading both **mobile and web data** and partitioned based on **batchid.**

5.3 Stage 3: Data Enrichment & Filtering

In this stage, we will enrich the data coming from web and mobile applications using the lookup table stored in Hbase and divide the records based on the enrichment rules into ‘pass’ and ‘fail’ records.

Rules for data enrichment :

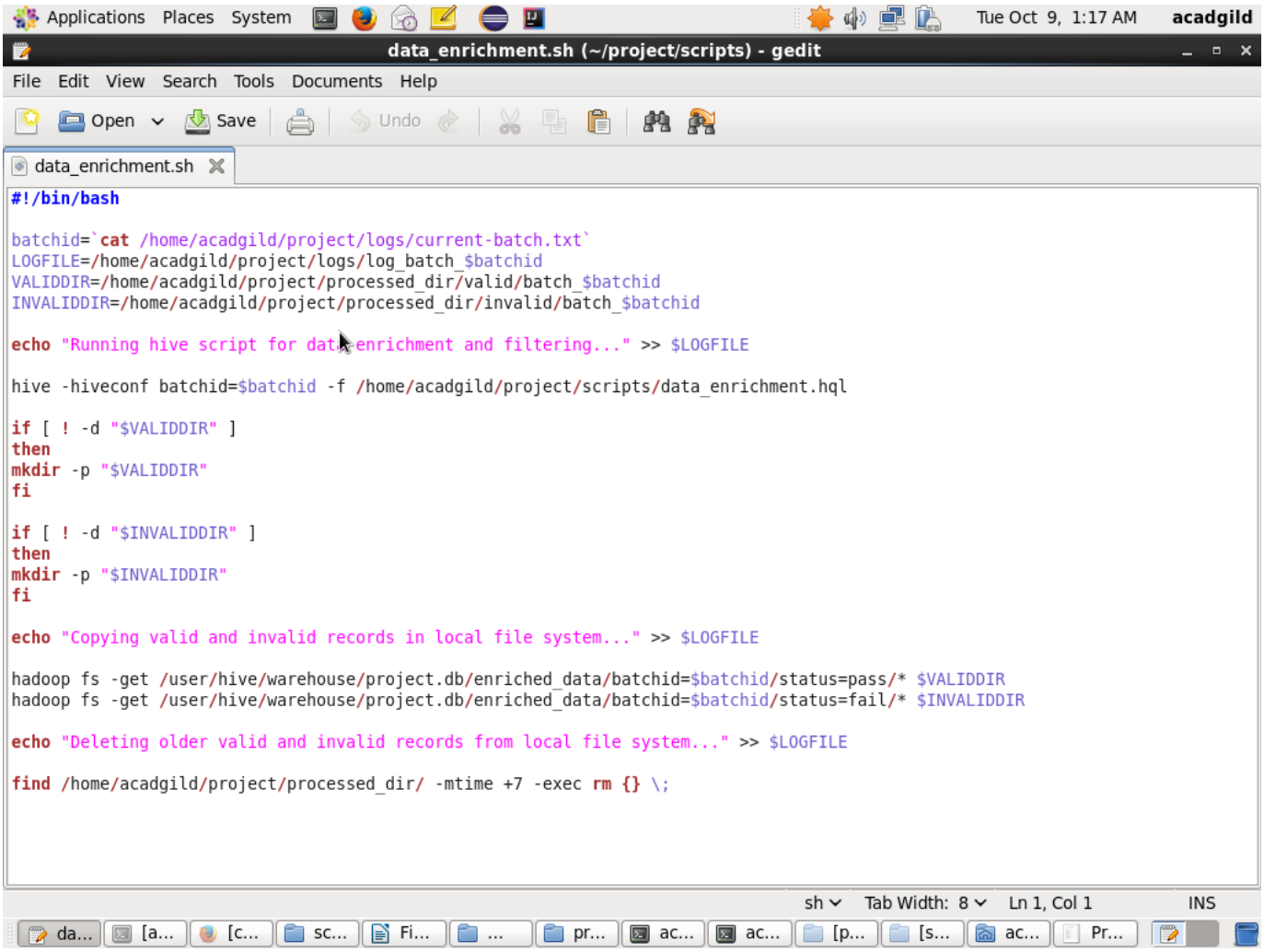
1. If any of **like** or **dislike** is **NULL** or **absent**, consider it as **0**.

2. If fields like **Geo_cd** and **Artist_id** are **NULL** or **absent**, consult the lookup tables for fields **Station_id** and **Song_id** respectively to get the values of **Geo_cd** and **Artist_id**.

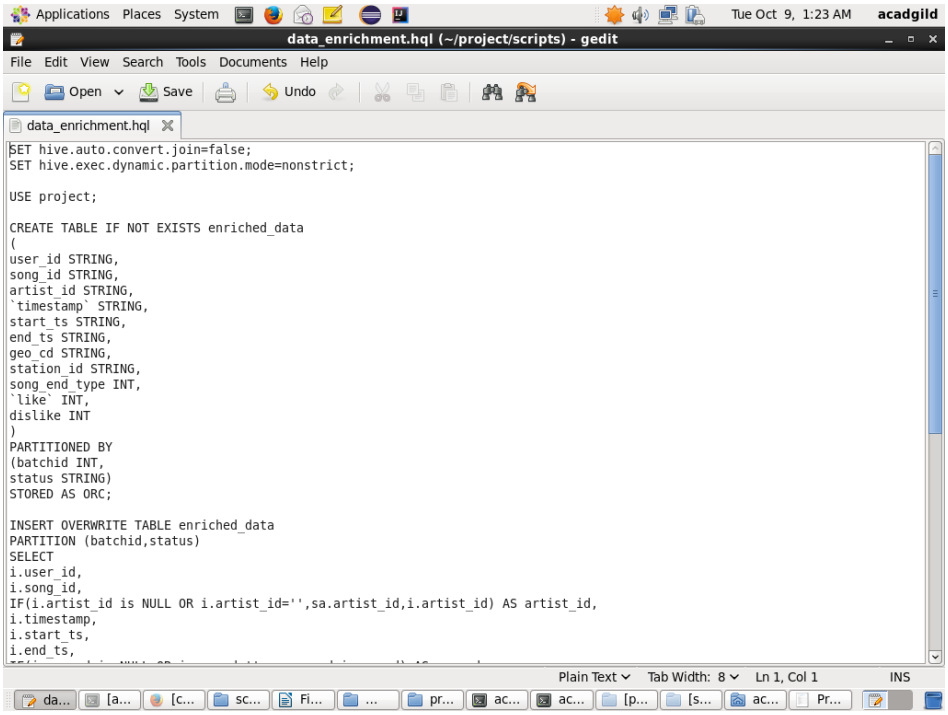
3. If corresponding lookup entry is not found, consider that record to be **invalid**

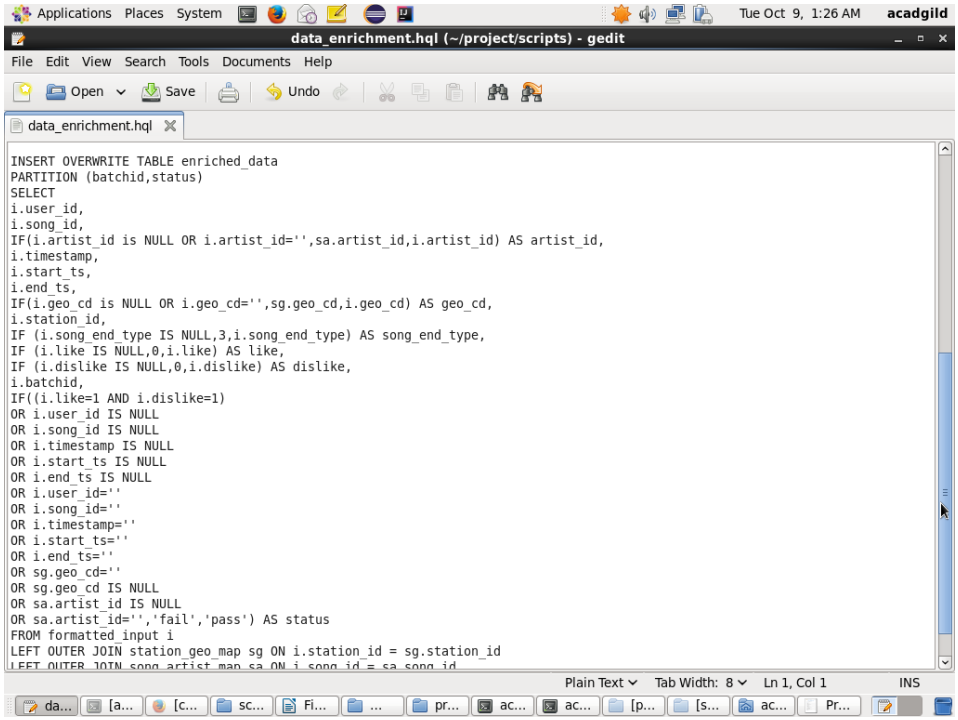
So based on the enrichment rules we will fill the null **geo_cd** and **artist_id** values with the help of corresponding lookup values in **song-artist-map** and **station-geo-map** tables in Hive-Hbase tables.

data_enrichment.sh



data_enrichment.hql :



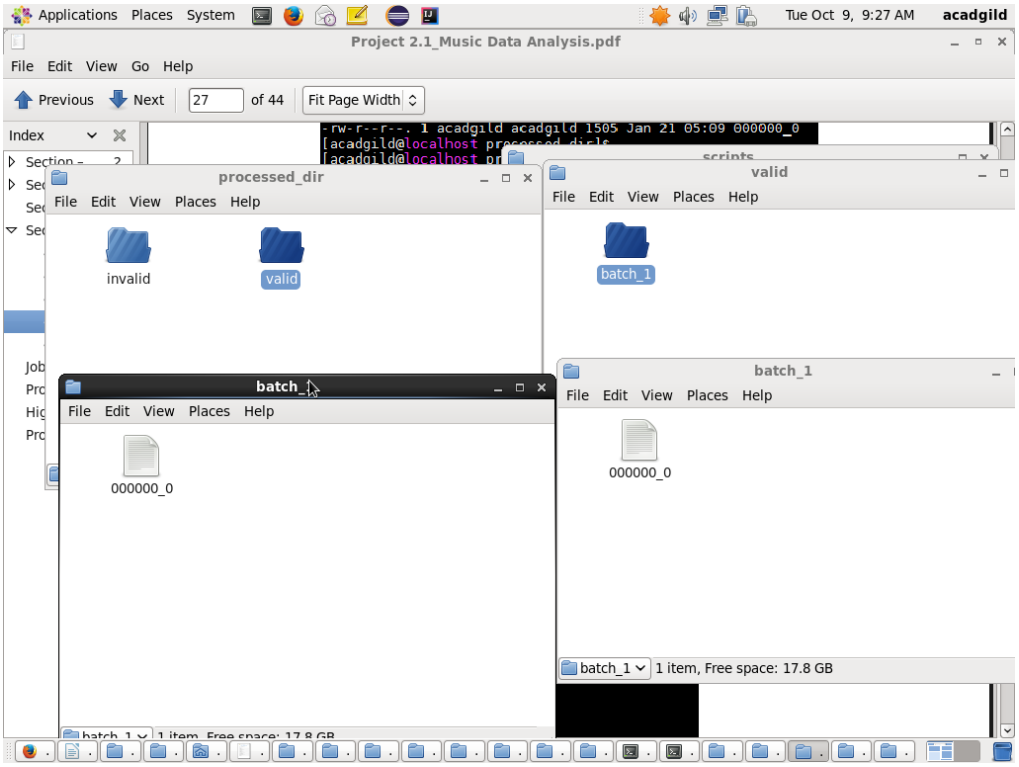


```
SET hive.auto.convert.join=false;
SET hive.exec.dynamic.partition.mode=nonstrict;

USE project;

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;

INSERT OVERWRITE TABLE enriched_data
PARTITION (batchid,status)
SELECT
i.user_id,
```

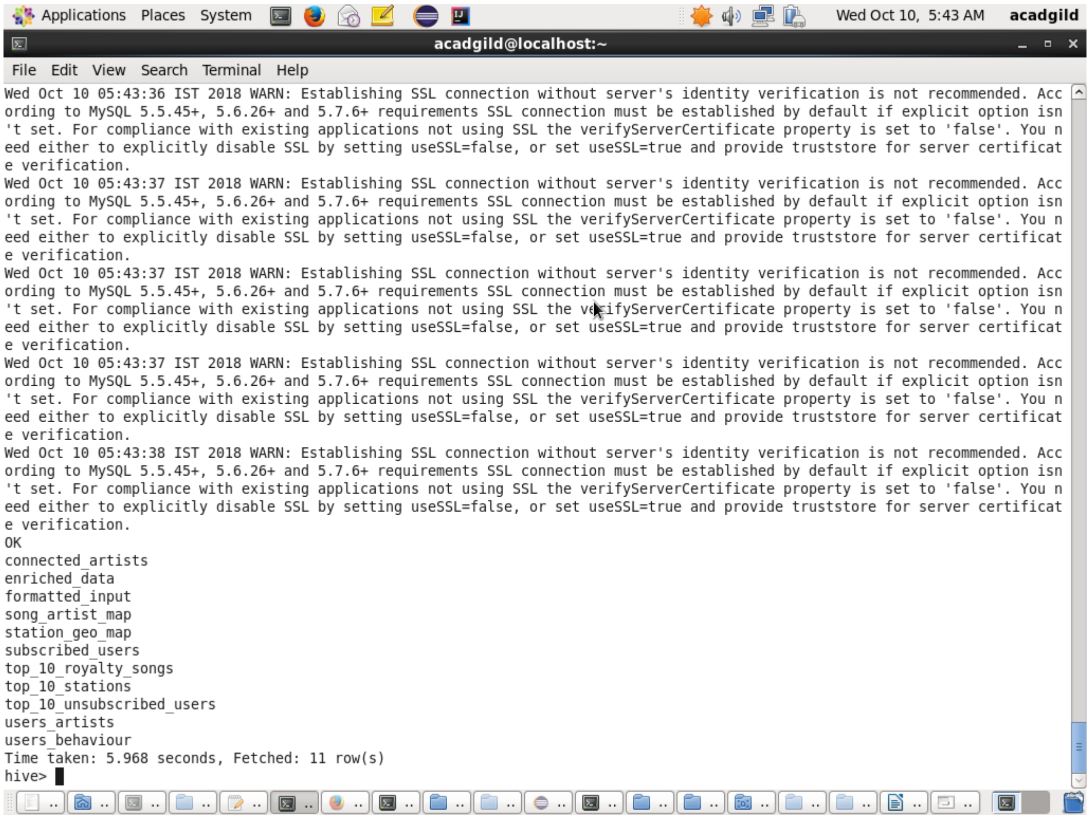
```
i.song_id,
IF(i.artist_id is NULL OR i.artist_id='',sa.artist_id,i.artist_id) AS artist_id,
i.`timestamp`,
i.start_ts,
i.end_ts,
IF(i.geo_cd is NULL OR i.geo_cd='',sg.geo_cd,i.geo_cd) AS geo_cd,
i.station_id,
IF (i.song_end_type IS NULL,3,i.song_end_type) AS song_end_type,
IF (i.`like` IS NULL,0,i.`like`) 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.user_id=''
OR i.song_id=''
OR i.`timestamp`=''
OR i.start_ts=''
OR i.end_ts=''
OR sg.geo_cd=''
OR sg.geo_cd IS NULL
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};

At the end script will automatically divide the records based on status pass & fail and dump the result into **processed_dir folder** with **valid** and **invalid** folders

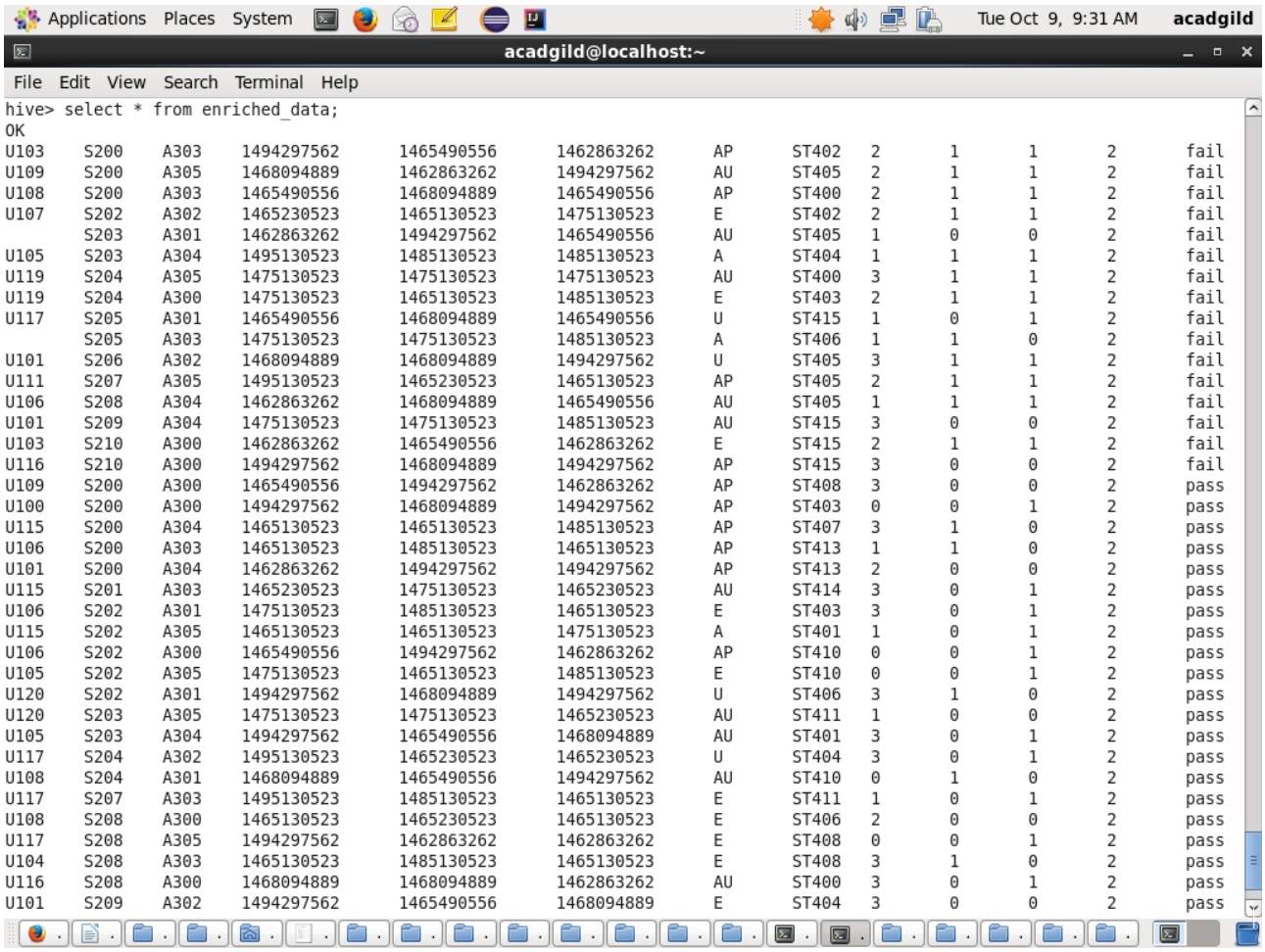


Now we can check whether the data properly loaded in the hive terminal or not.



In the below screenshot, we have data for **enriched_data table** where we filled the **null** values of **artist_id** and **geo_cd** of formatted input with the help of lookup tables

hive>select * From enriched_data;




```
hive> select * from enriched_data;
OK
U103    S200    A303    1494297562      1465490556      1462863262      AP      ST402   2       1       1       2       fail
U109    S200    A305    1468094889      1462863262      1494297562      AU      ST405   2       1       1       2       fail
U108    S200    A303    1465490556      1468094889      1465490556      AP      ST400   2       1       1       2       fail
U107    S202    A302    1465230523      1465130523      1475130523      E       ST402   2       1       1       2       fail
        S203    A301    1462863262      1494297562      1465490556      AU      ST405   1       0       0       2       fail
U105    S203    A304    1495130523      1485130523      1485130523      A       ST404   1       1       1       2       fail
U119    S204    A305    1475130523      1475130523      1475130523      AU      ST400   3       1       1       2       fail
U119    S204    A300    1475130523      1465130523      1485130523      E       ST403   2       1       1       2       fail
U117    S205    A301    1465490556      1468094889      1465490556      U       ST415   1       0       1       2       fail
        S205    A303    1475130523      1475130523      1485130523      A       ST406   1       1       0       2       fail
U101    S206    A302    1468094889      1468094889      1494297562      U       ST405   3       1       1       2       fail
U111    S207    A305    1495130523      1465230523      1465130523      AP      ST405   2       1       1       2       fail
U106    S208    A304    1462863262      1468094889      1465490556      AU      ST405   1       1       1       2       fail
U101    S209    A304    1475130523      1475130523      1485130523      AU      ST415   3       0       0       2       fail
U103    S210    A300    1462863262      1465490556      1462863262      E       ST415   2       1       1       2       fail
U116    S210    A300    1494297562      1468094889      1494297562      AP      ST415   3       0       0       2       fail
U109    S200    A300    1465490556      1494297562      1462863262      AP      ST408   3       0       0       2       pass
U100    S200    A300    1494297562      1468094889      1494297562      AP      ST403   0       0       1       2       pass
U115    S200    A304    1465130523      1465130523      1485130523      AP      ST407   3       1       0       2       pass
U106    S200    A303    1465130523      1485130523      1465130523      AP      ST413   1       1       0       2       pass
U101    S200    A304    1462863262      1494297562      1494297562      AP      ST413   2       0       0       2       pass
U115    S201    A303    1465230523      1475130523      1465230523      AU      ST414   3       0       1       2       pass
U106    S202    A301    1475130523      1485130523      1465130523      E       ST403   3       0       1       2       pass
U115    S202    A305    1465130523      1465130523      1475130523      A       ST401   1       0       1       2       pass
U106    S202    A300    1465490556      1494297562      1462863262      AP      ST410   0       0       1       2       pass
U105    S202    A305    1475130523      1465130523      1485130523      E       ST410   0       0       1       2       pass
U120    S202    A301    1494297562      1468094889      1494297562      U       ST406   3       1       0       2       pass
U120    S203    A305    1475130523      1475130523      1465230523      AU      ST411   1       0       0       2       pass
U105    S203    A304    1494297562      1465490556      1468094889      AU      ST401   3       0       1       2       pass
U117    S204    A302    1495130523      1465230523      1465230523      U       ST404   3       0       1       2       pass
U108    S204    A301    1468094889      1465490556      1494297562      AU      ST410   0       1       0       2       pass
U117    S207    A303    1495130523      1485130523      1465130523      E       ST411   1       0       1       2       pass
U108    S208    A300    1465130523      1465230523      1465130523      E       ST406   2       0       0       2       pass
U117    S208    A305    1494297562      1462863262      1462863262      E       ST408   0       0       1       2       pass
U104    S208    A303    1465130523      1485130523      1465130523      E       ST408   3       1       0       2       pass
U116    S208    A300    1468094889      1468094889      1462863262      AU      ST400   3       0       1       2       pass
U101    S209    A302    1494297562      1465490556      1468094889      E       ST404   3       0       0       2       pass
```

By applying the provided rules, we have successfully accomplished Data enrichment and Filtering stage.

5.4 Stage 4 : Data Analysis using Spark

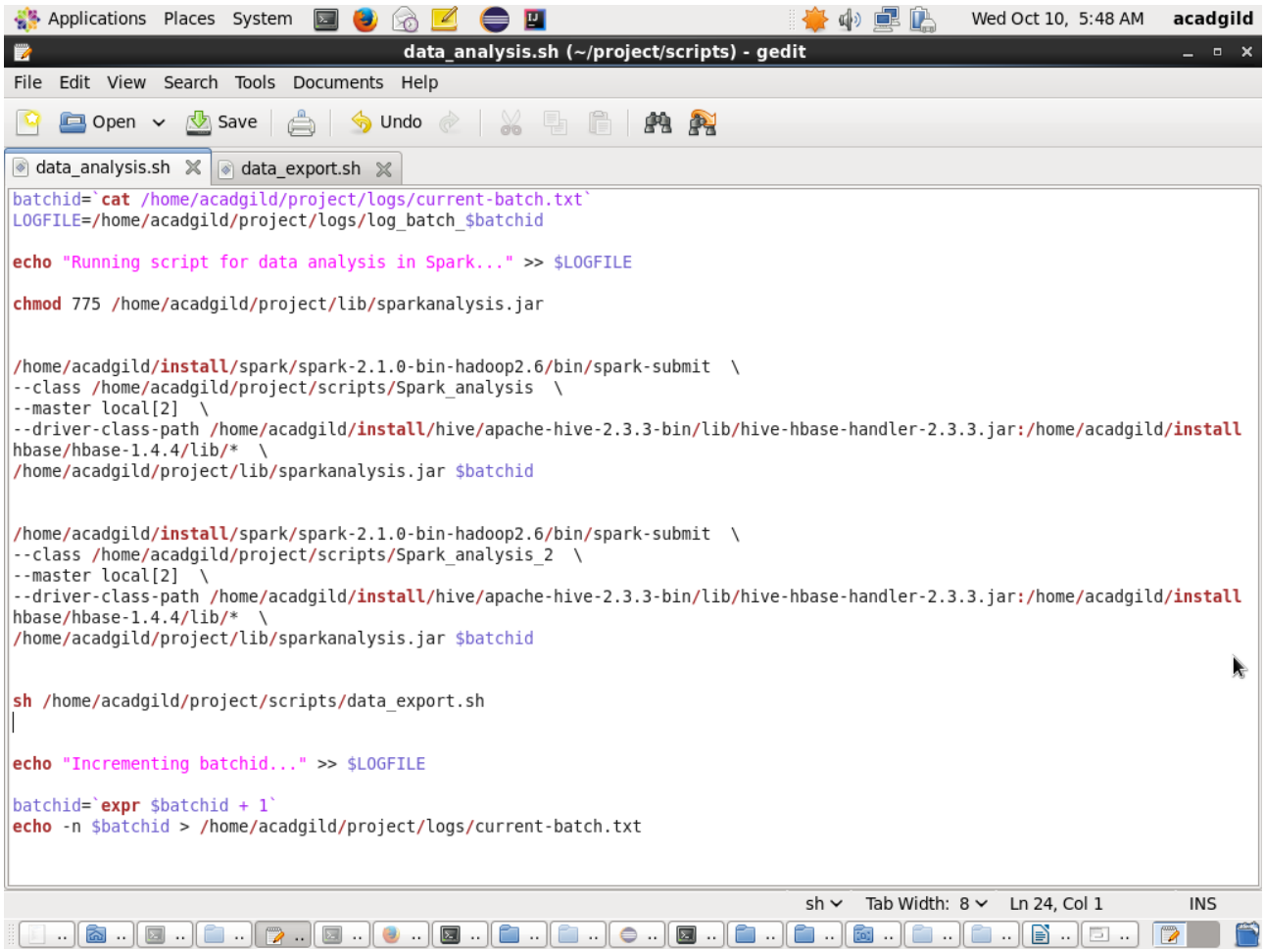
In this stage we will do analysis on enriched data using Spark SQL and run the program using Spark Submit command.

Before running the spark-submit command we have to zip –d command to remove the bad manifests in created spark project jar file to avoid the invalid Signature exception. We used two spark-submits for analysis.

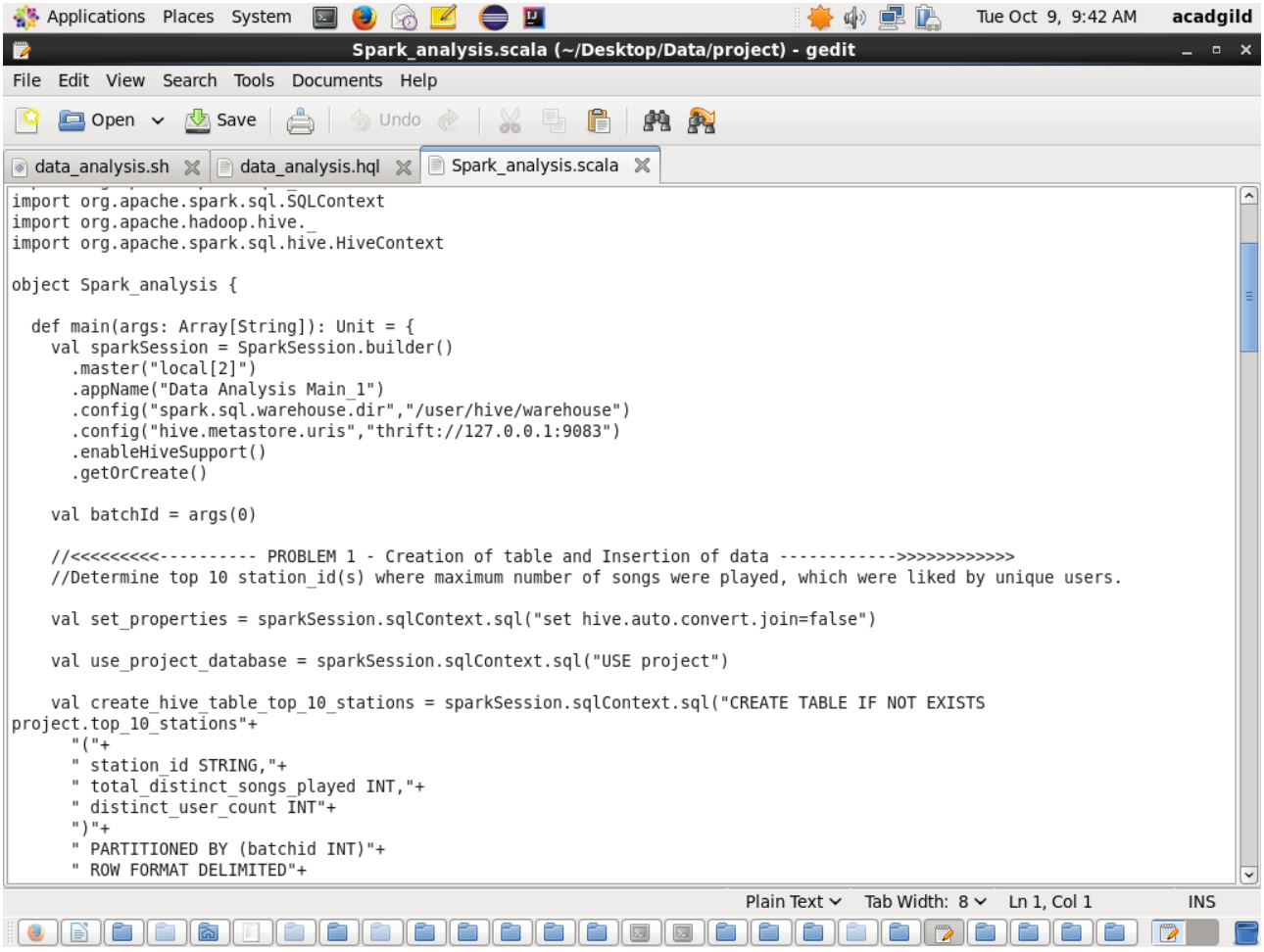
a. Spark_analysis for creating tables for each query/problem statement.

b. Spark_analysis_2 for displaying results for each query in terminal.

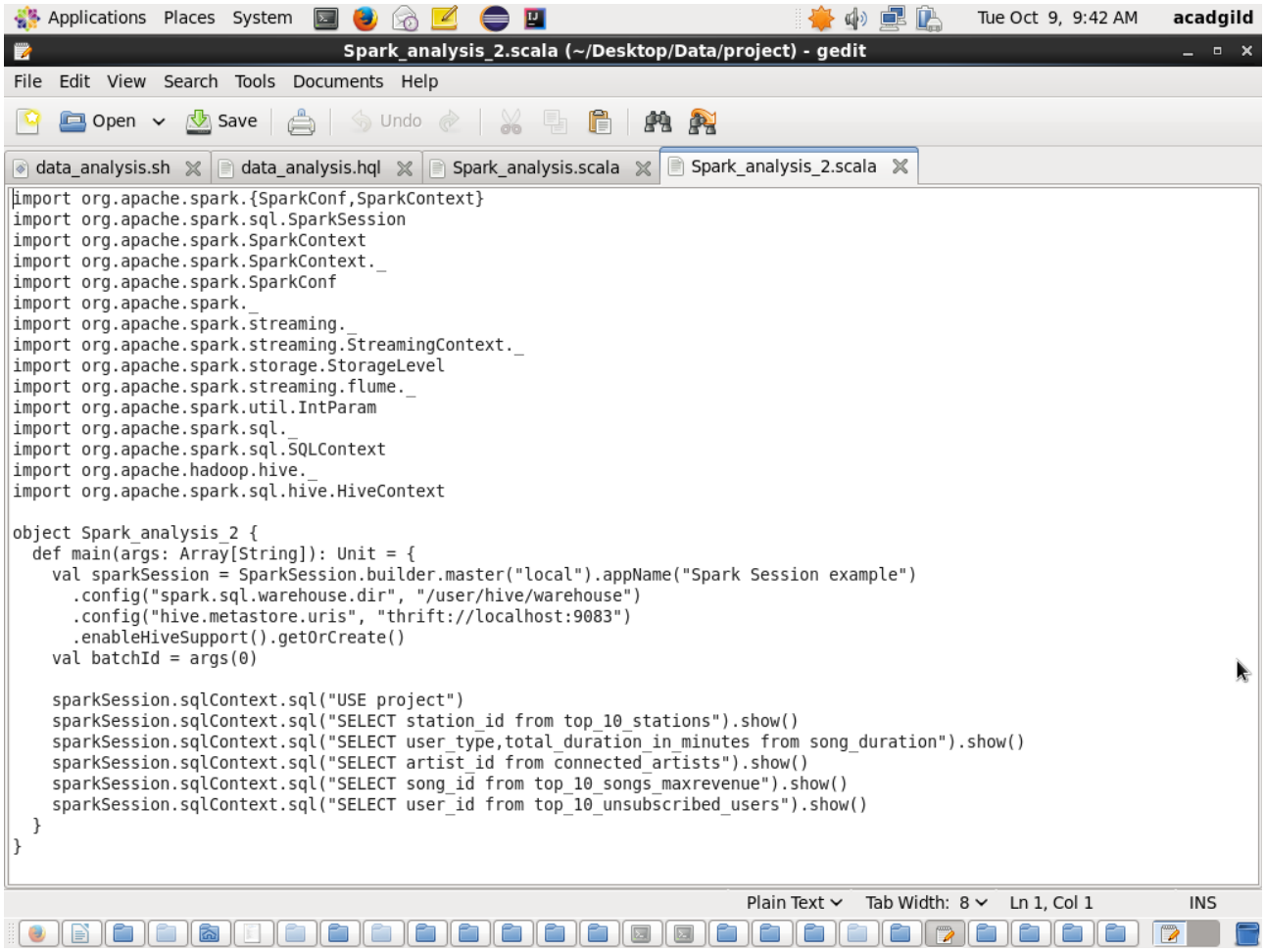
DataAnalysis.sh



Spark_analysis.scala



Spark_analysis_2.scala

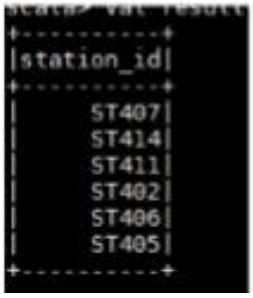


```
[acadgild@localhost ~]$ sh /home/acadgild/project/scripts/data_analysis.sh
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.3-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j
/impl/StaticLoggerBinder.class]
SLF4J: Found 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-2.3.3-bin/lib/hive-common-2.3.3.j
ar!/hive-log4j2.properties Async: true
Tue Oct 09 12:52:25 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 12:52:27 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 12:52:27 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 12:52:27 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 12:52:38 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
Tue Oct 09 12:52:38 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording 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 n
```

```
ording 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 n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
OK
Time taken: 3.804 seconds
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181009125245_0b717b06-f42c-4484-b74f-0d6737422ce7
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_1539056459717_0003, Tracking URL = http://localhost:8088/proxy/application_1539056459717_0003/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1539056459717_0003
Hadoop job information for Stage-1: number of mappers: 0; number of reducers: 1
2018-10-09 12:53:29,681 Stage-1 map = 0%,  reduce = 0%
2018-10-09 12:53:50,935 Stage-1 map = 0%,  reduce = 100%, Cumulative CPU 3.91 sec
MapReduce Total cumulative CPU time: 3 seconds 910 msec
Ended Job = job_1539056459717_0003
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:
  set mapreduce.job.reduces=<number>
Starting Job = job_1539056459717_0004, Tracking URL = http://localhost:8088/proxy/application_1539056459717_0004/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1539056459717_0004
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-10-09 12:54:24,443 Stage-2 map = 0%,  reduce = 0%
2018-10-09 12:54:46,061 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.44 sec
2018-10-09 12:55:07,196 Stage-2 map = 100%,  reduce = 67%, Cumulative CPU 7.96 sec
```

Query 1:

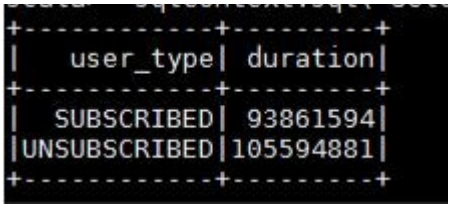
Determine top 10 station_id(s) where maximum number of songs were played, which were liked by unique users.



station_id
ST407
ST414
ST411
ST402
ST406
ST405

Query 2:

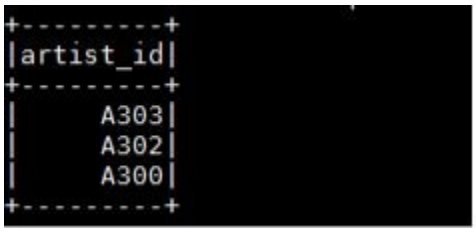
Determine total duration of songs played by each type of user, where type of user can be 'subscribed' or 'unsubscribed'. An unsubscribed user is the one whose record is either not present in Subscribed_users lookup table or has subscription_end_date earlier than the timestamp of the song played by him.



user_type	duration
SUBSCRIBED	93861594
UNSUBSCRIBED	105594881

Query 3:

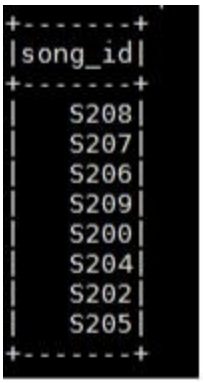
Determine top 10 connected artists. Connected artists are those whose songs are most listened by the unique users who follow them.



artist_id
A303
A302
A300

Query 4:

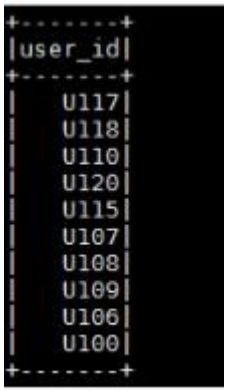
Determine top 10 songs who have generated the maximum revenue. Royalty applies to a song only if it was liked or was completed successfully or both



song_id
S208
S207
S206
S209
S200
S204
S202
S205

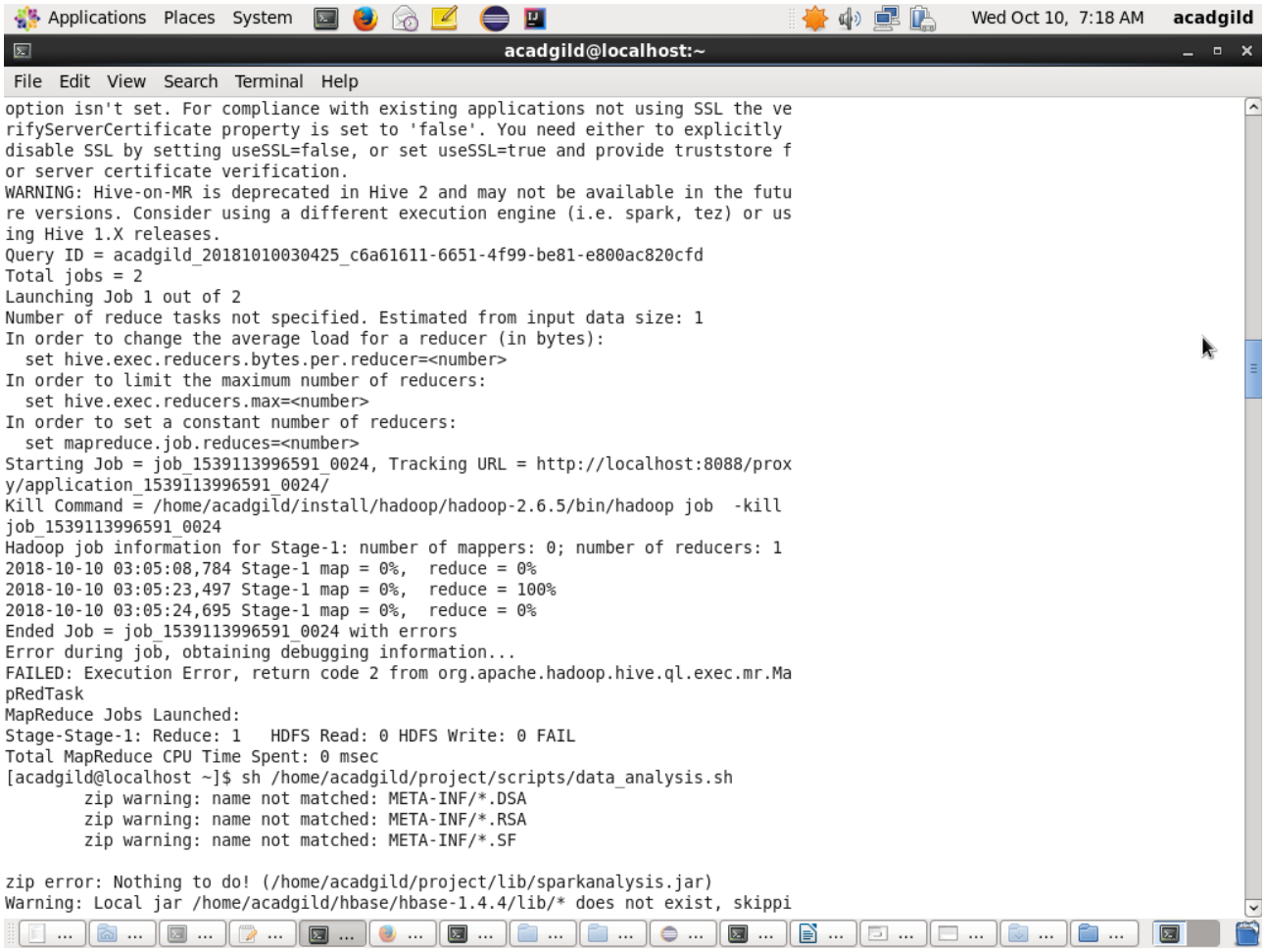
Query 5:

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

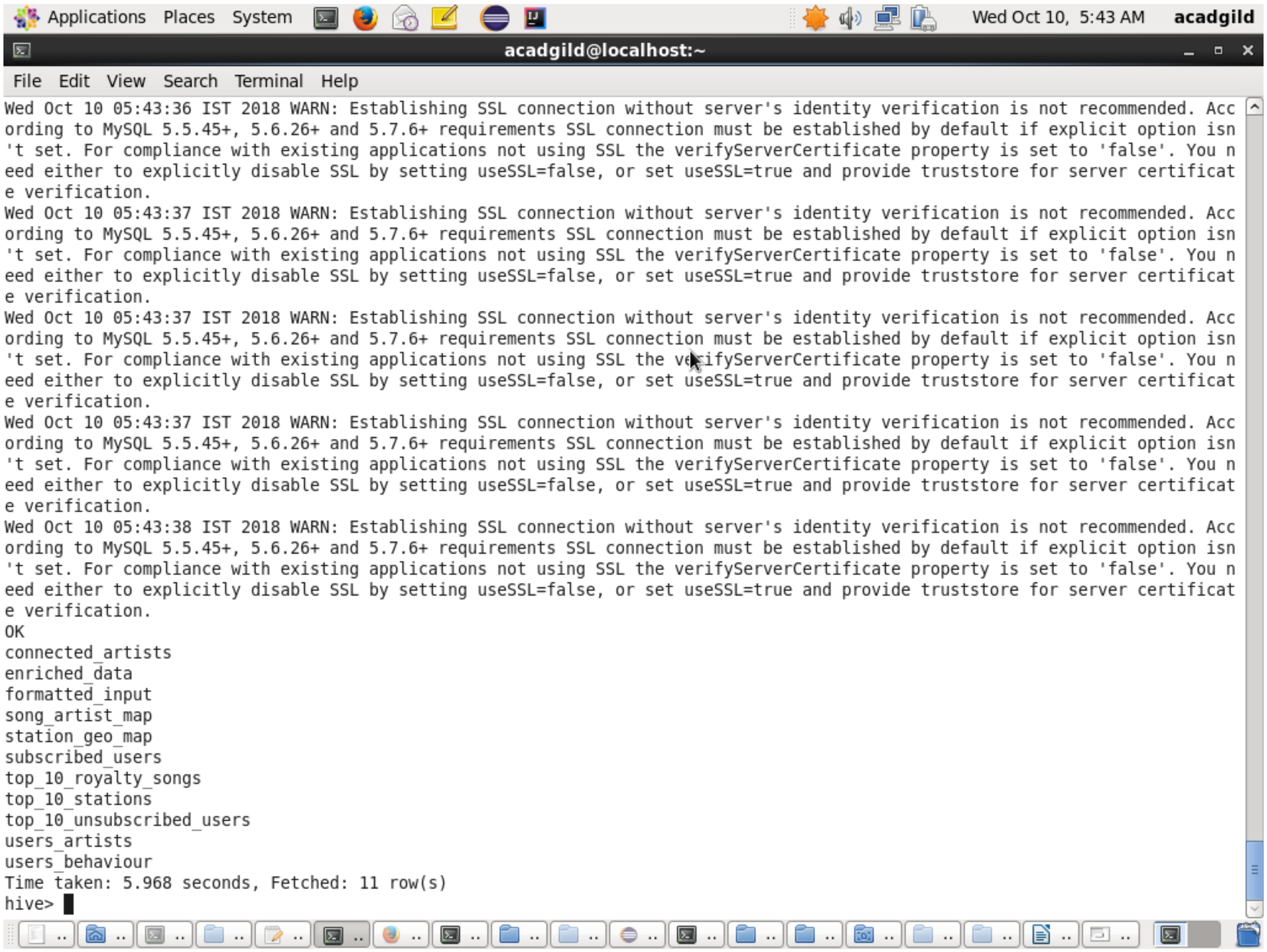


user_id
U117
U118
U110
U120
U115
U107
U108
U109
U106
U100

Table Creation in HIVE and Data analysis using HIVE :



The tables have also been created in the Hive:

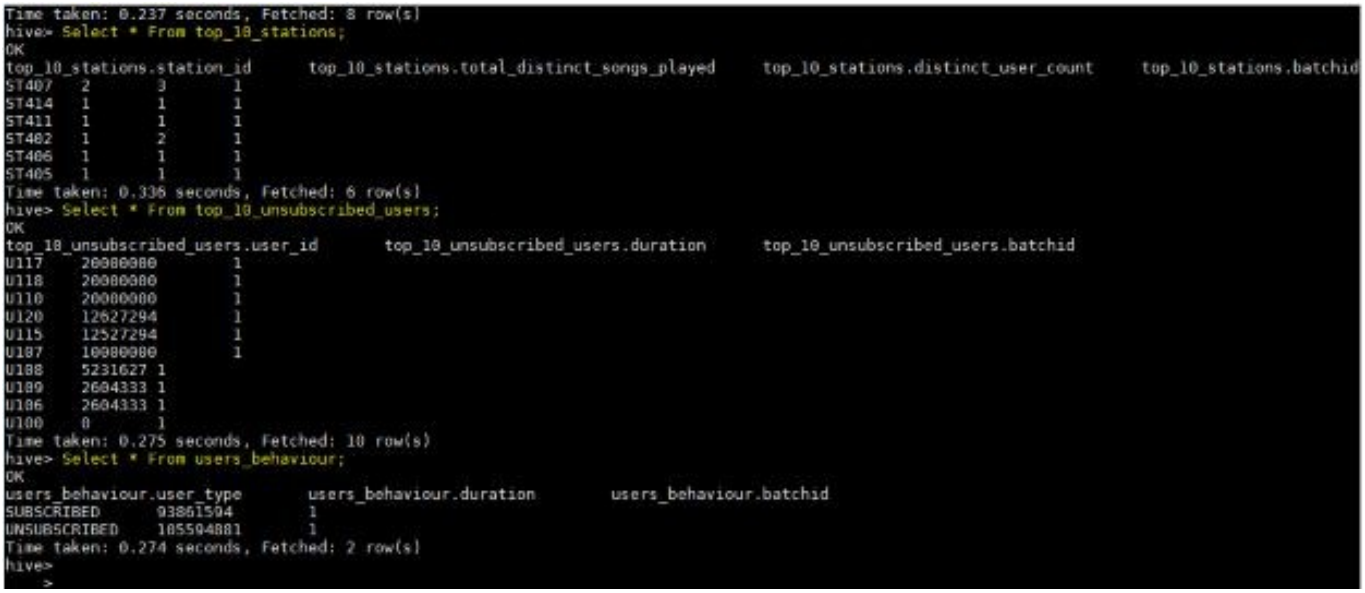


We have seen all the spark queries creating the tables for each query. So Data Analysis using Spark is executed successfully.

The data analysis result is shown in the Hive tables below in the screen shot :

Output from, connected_artists, top_10_royalty_songs, top_10_stations.

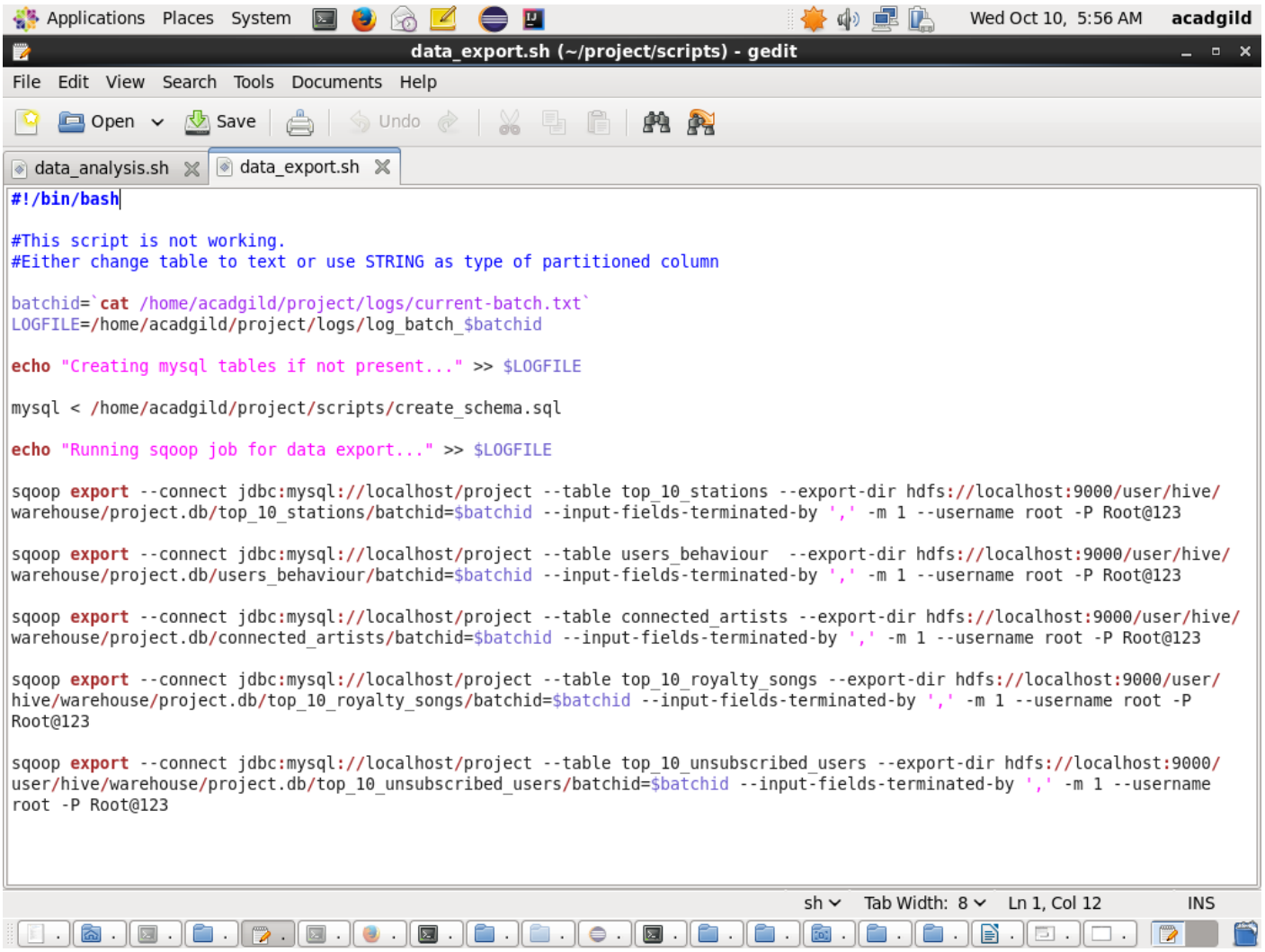
Output from top_10_unsubscribed_users, users_behaviour.



Now, we need to export all the data to the **MYSQL** using **sqoop**, run the script **data_export.sh** .

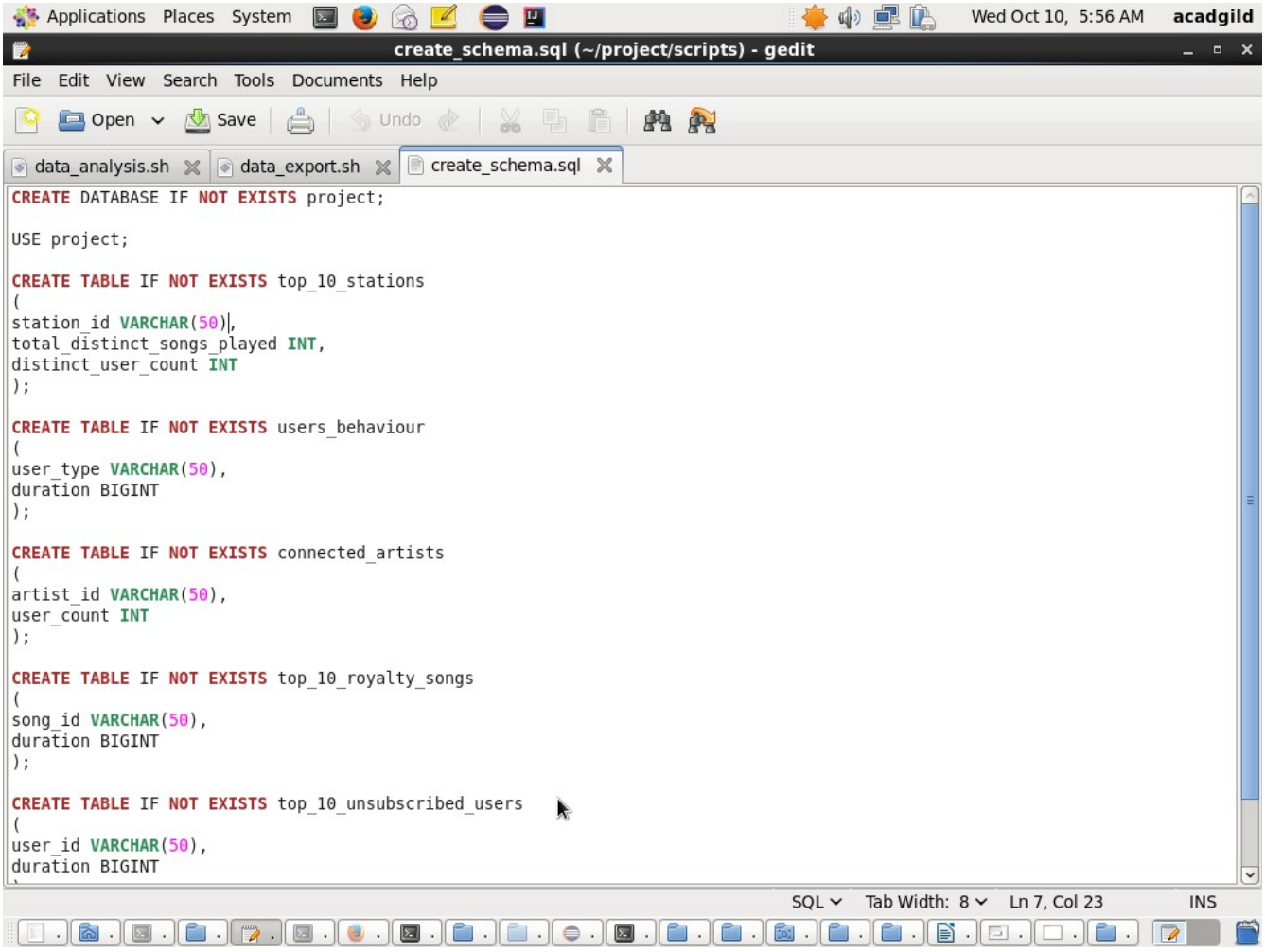
5.5 Stage 5 : Data Storage in MYSQL

Using the bash file shown below, **data_export.sh** we are going to **export** the data **from the hive tables** into **mysql** using **Sqoop export**.



create_schema.sql :

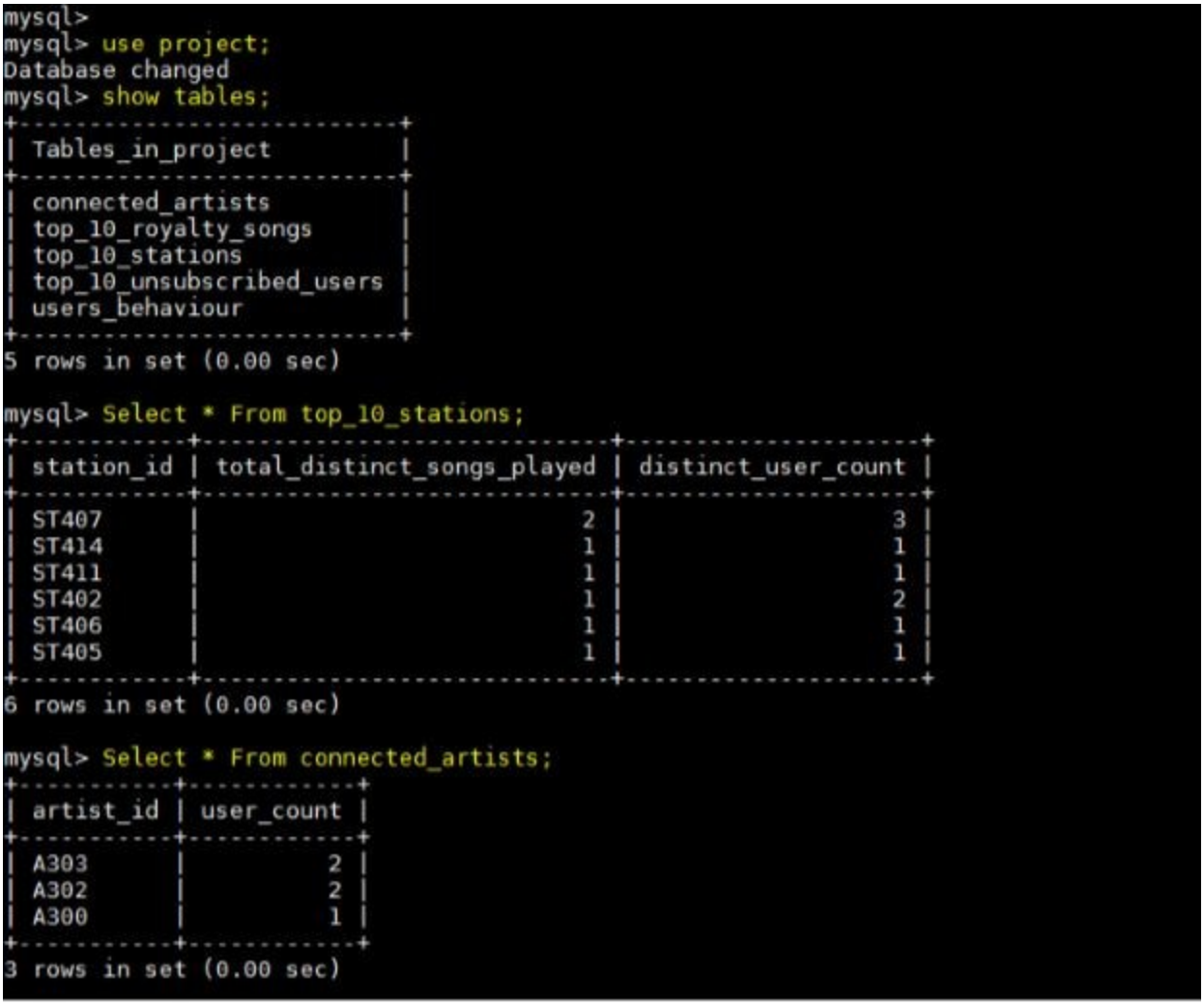
Make sure that you logged in to MySql. The below schema will create the database and tables in the MySQl.



Now we can see the data exported successfully into the MYSQL Database for all the 5 queries.

The sqoop export command exported the tables from the hive and it stored in the Mysql. The below screen shot show the successful Sqoop export from hive to mysql.

The data base project had been exported from the hive.



top_10_royalty_songs :

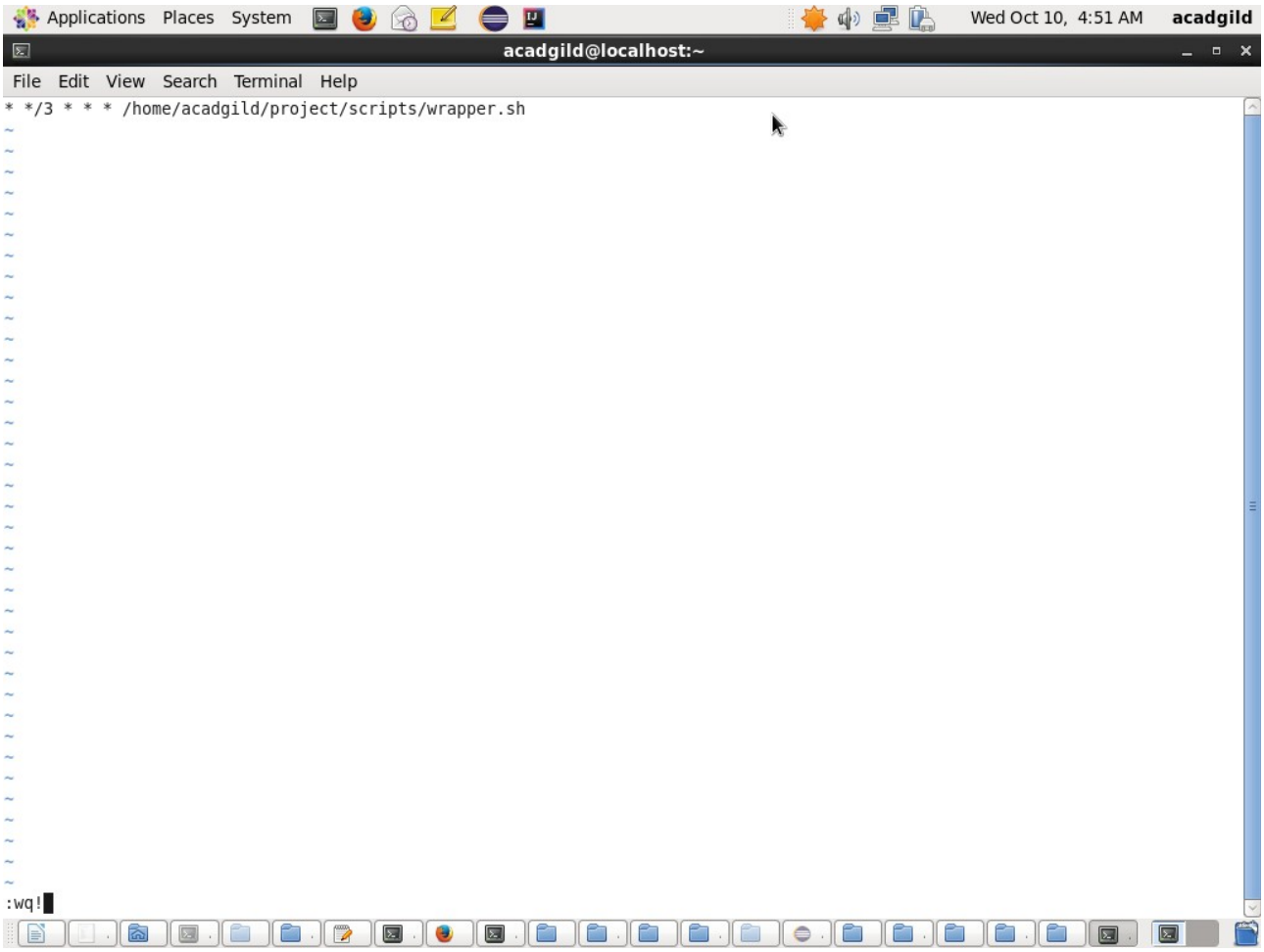
```
mysql> Select * From top_10_royalty_songs;
+---------+----------+
| song_id | duration |
+---------+----------+
| S208    | 22627294 |
| S207    | 20000000 |
| S206    | 19900000 |
| S209    | 15254588 |
| S200    |  9900000 |
| S204    |  2604333 |
| S202    |   100000 |
| S205    |        0 |
+---------+----------+
8 rows in set (0.00 sec)
```

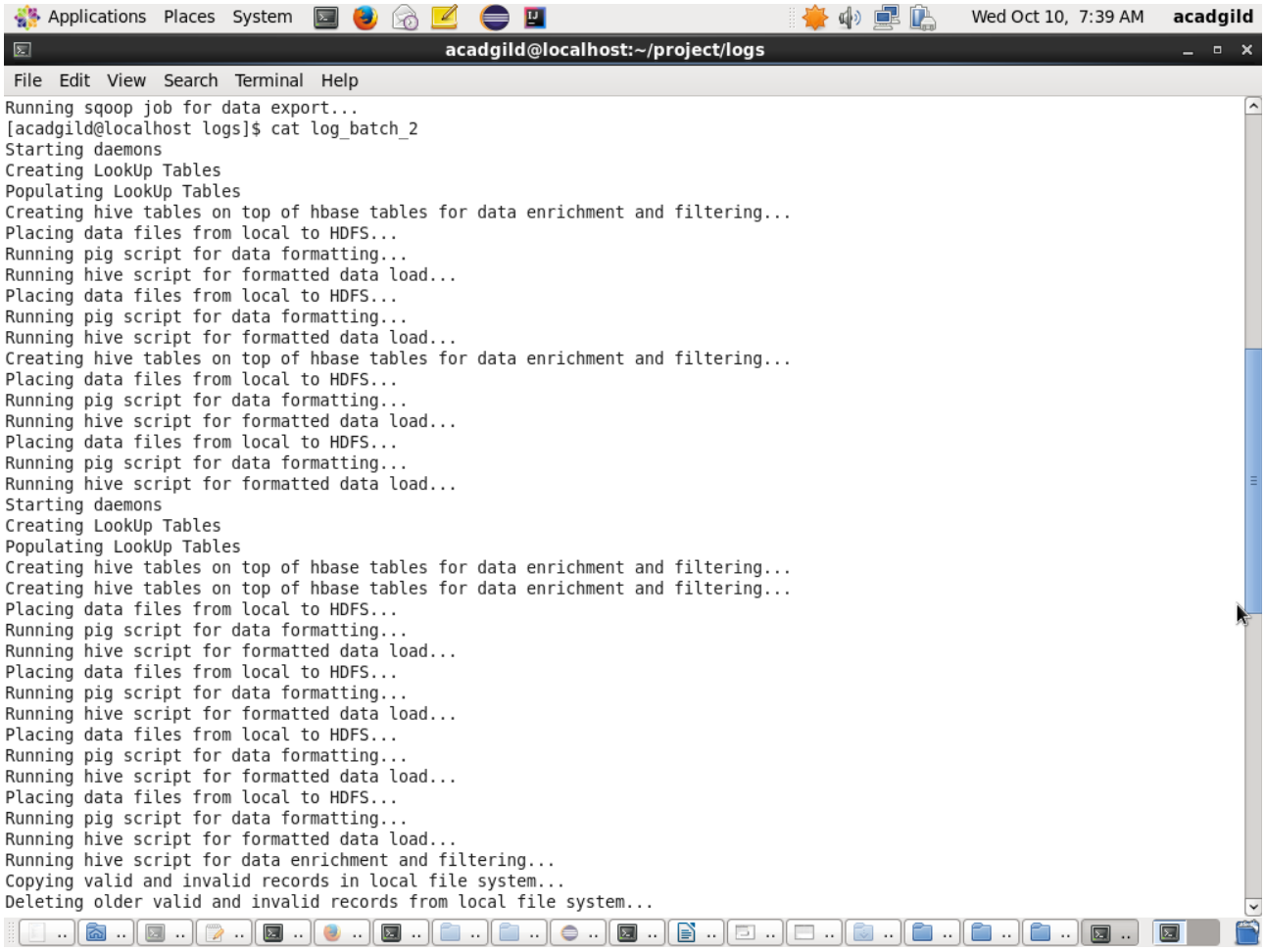
Output from top_10_unsubscribed_users and users_behaviour :

```
mysql> Select * From top_10_unsubscribed_users;
+---------+----------+
| user_id | duration |
+---------+----------+
| U117    | 20000000 |
| U118    | 20000000 |
| U110    | 20000000 |
| U120    | 12627294 |
| U115    | 12527294 |
| U107    | 10000000 |
| U108    |  5231627 |
| U109    |  2604333 |
| U106    |  2604333 |
| U100    |        0 |
+---------+----------+
10 rows in set (0.01 sec)

mysql> Select * From users_behaviour;
+--------------+-----------+
| user_type    | duration  |
+--------------+-----------+
| SUBSCRIBED   |  93861594 |
| UNSUBSCRIBED | 105594881 |
+--------------+-----------+
2 rows in set (0.00 sec)
```

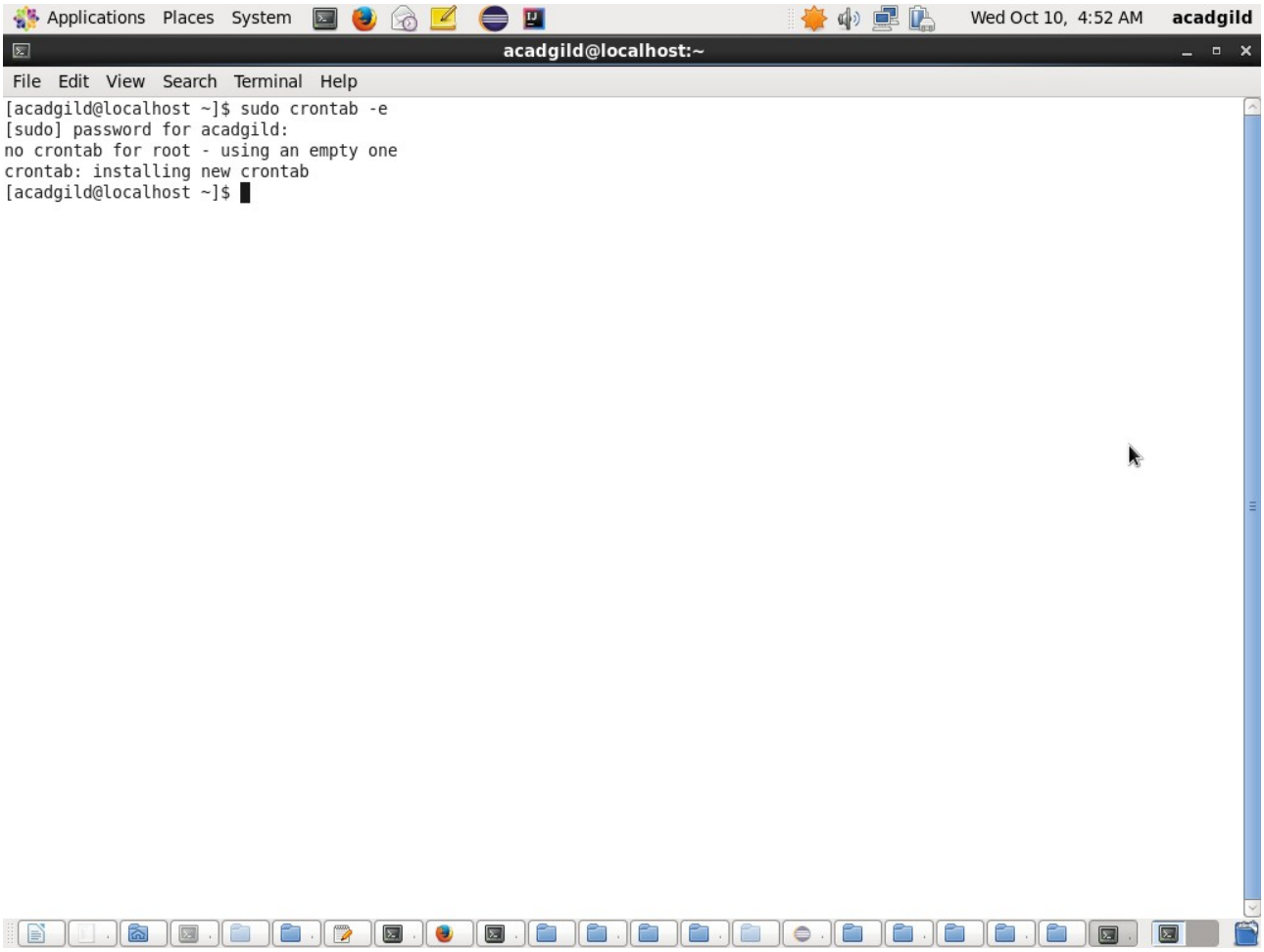
6. Job Scheduling :

After exporting data into MySQL batchid will be incremented to additional 1 means one batch of data operations is successfully completed and new batch of data will be loaded for the analysis after every 3 hours.

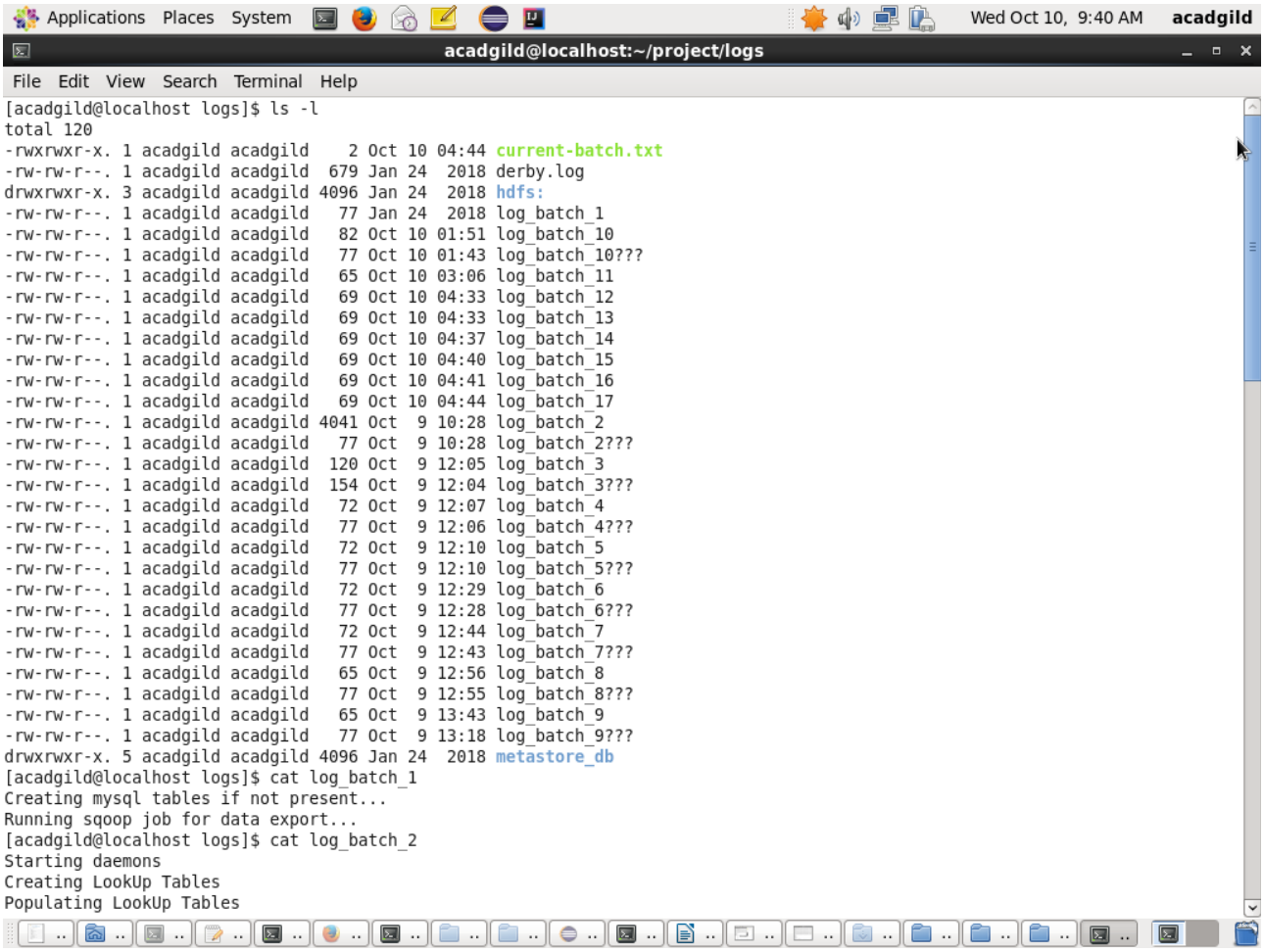



Applications Places System
Wed Oct 10, 7:39 AM acadgild
acadgild@localhost:~/project/logs
File Edit View Search Terminal Help
Running sqoop job for data export...
[acadgild@localhost logs]$ cat log_batch_2
Starting daemons
Creating LookUp Tables
Populating LookUp Tables
Creating hive tables on top of hbase tables for data enrichment and filtering...
Placing data files from local to HDFS...
Running pig script for data formatting...
Running hive script for formatted data load...
Placing data files from local to HDFS...
Running pig script for data formatting...
Running hive script for formatted data load...
Creating hive tables on top of hbase tables for data enrichment and filtering...
Placing data files from local to HDFS...
Running pig script for data formatting...
Running hive script for formatted data load...
Placing data files from local to HDFS...
Running pig script for data formatting...
Running hive script for formatted data load...
Starting daemons
Creating LookUp Tables
Populating LookUp Tables
Creating hive tables on top of hbase tables for data enrichment and filtering...
Creating hive tables on top of hbase tables for data enrichment and filtering...
Placing data files from local to HDFS...
Running pig script for data formatting...
Running hive script for formatted data load...
Placing data files from local to HDFS...
Running pig script for data formatting...
Running hive script for formatted data load...
Placing data files from local to HDFS...
Running pig script for data formatting...
Running hive script for formatted data load...
Placing data files from local to HDFS...
Running pig script for data formatting...
Running hive script for formatted data load...
Running hive script for data enrichment and filtering...
Copying valid and invalid records in local file system...
Deleting older valid and invalid records from local file system...

We can check logs to track the behavior of the operations we have done on the data and overcome failures in the pipeline and we can see the **batchid** incremented value in **current-batch.txt**

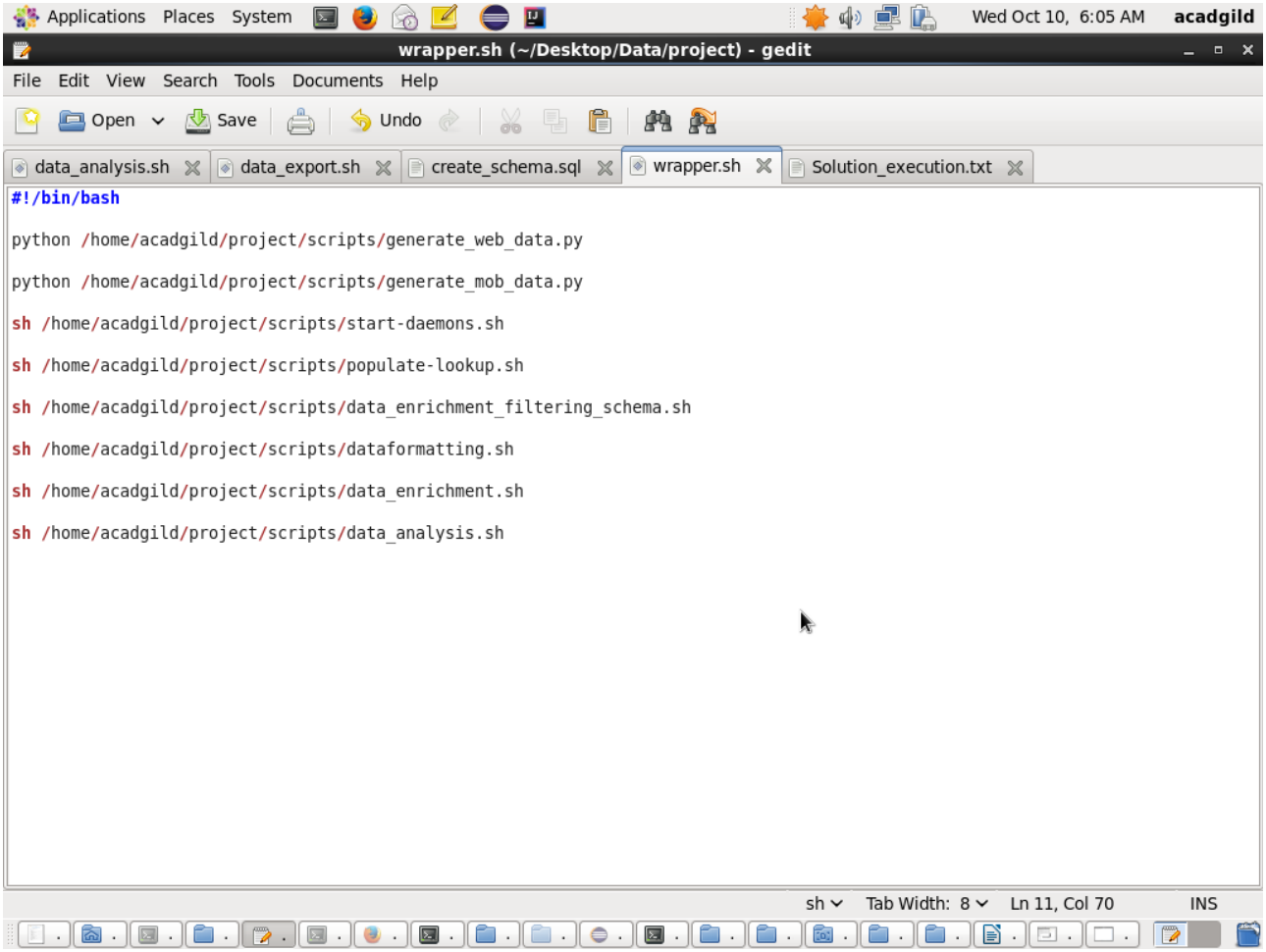


The log file captured all the data and steps we performed so far :



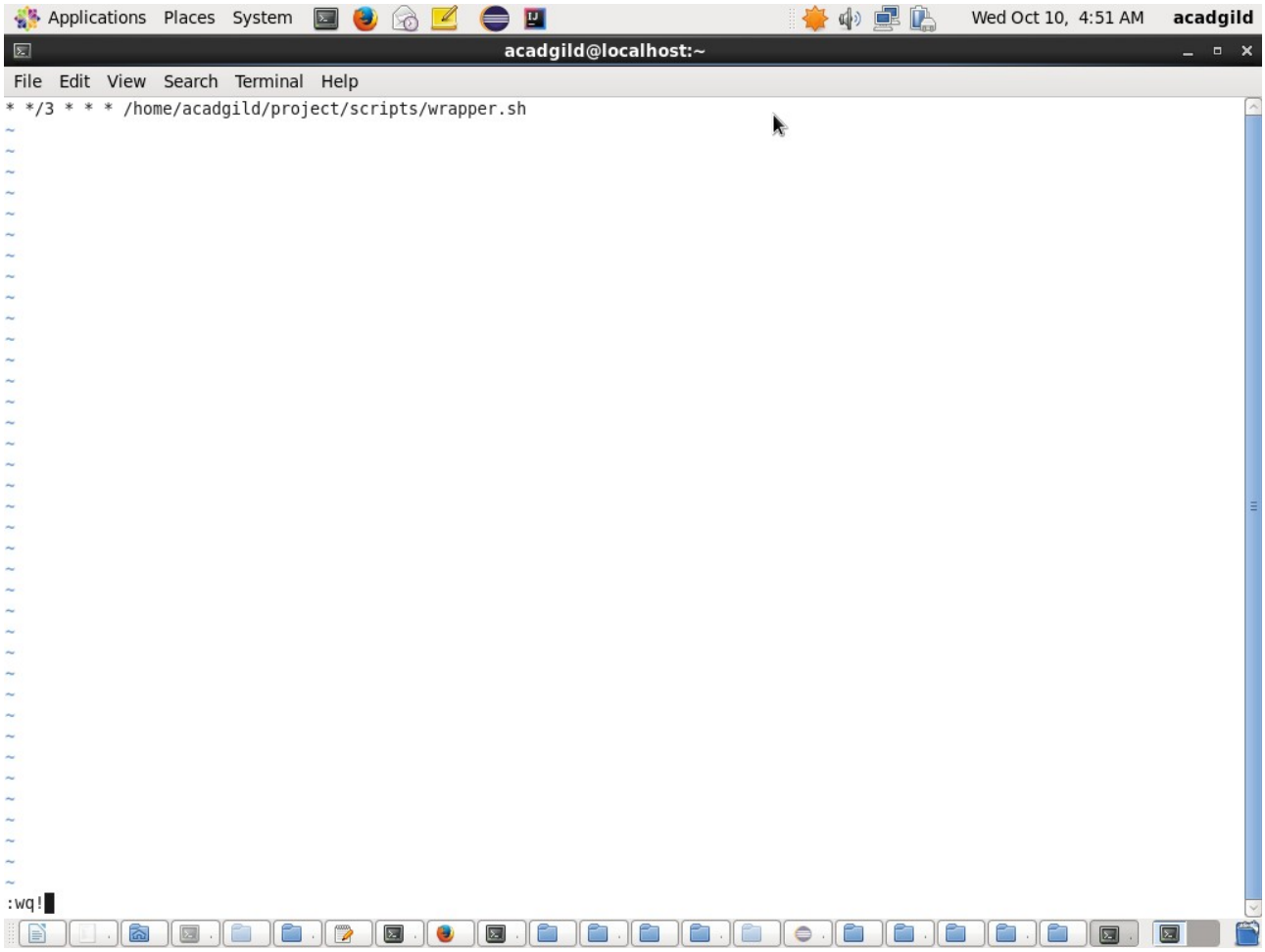
Wrapping all the scripts inside the single script file and scheduling this file to run at the periodic interval of every 3 hours.

wrapper.sh



The **wrapper.sh** will be running for every 3 hours as per the job scheduling done below, as per the above order the **wrapper.sh** will run the scripts.

Creating Crontab to schedule the wrapper.sh script to run for every 3 hour interval.



The **crontab job scheduler** will run the **wrappr.sh** every 3 hours and for every 3 hours we will get incremental batch ID’s.

Hence, as per the request this job scheduling has been done.

9. Highlights of the Project

- No join of query is used while analysis. Data is already enriched with new fields and using broadcast maps on Lookup tables so as to avoid any join.

- We used full automated bash scripts from start to end.

10. Project End Conclusion:

So we performed all the data operations as per the sequence mentioned in the **wrapper.sh** file and obtained results successfully for one of the leading music company.