

# **FINAL PROJECT**

**MUSIC DATA Analysis**

**Using HADOOP ECO SYSTEM**

**By**

**MILIND GAUTAM**

## **Table of Contents**

<b>Serial no.</b>	<b>Topic</b>
1.0	Project Description
2.0	Data Files
2.1	Fields present in the data files
2.2	LookUp Tables
3.0	Data Ingestion and Initial Validation
3.1	Rules for data ingestion and data filtering
4.0	Data Enrichment
4.1	Rules for data enrichment
4.2	Post Enrichment
5.0	Data Analysis
5.1	Challenges and Optimization
6.0	Post Analysis
7.0	Design Architecture
8.0	Implementation
9.0	Highlights of the project
10.0	Conclusion

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

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

## 2.2 LookUp Tables :

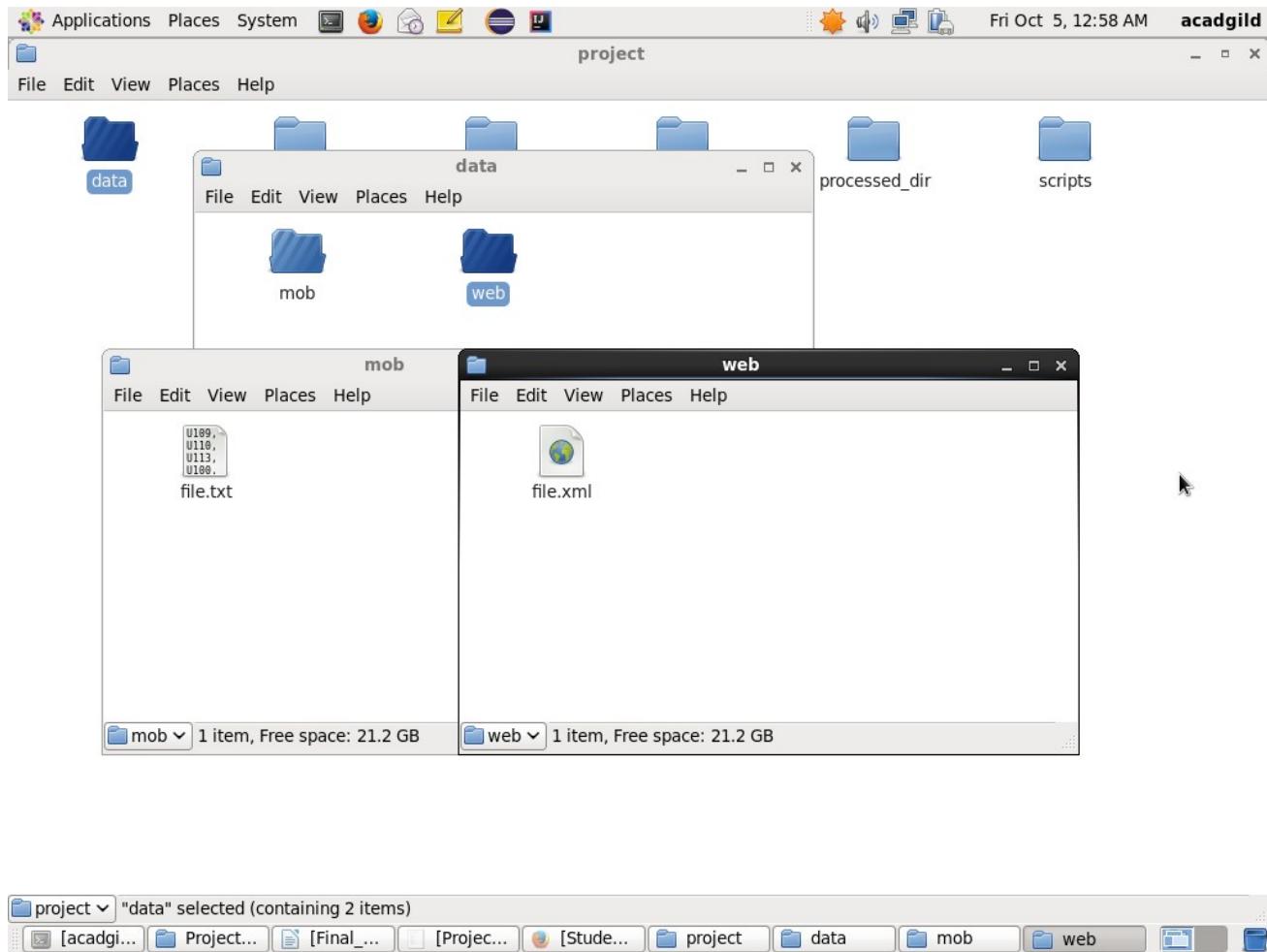
There are some existing look up tables present in NoSQL databases. They play an important role in data enrichment and analysis.

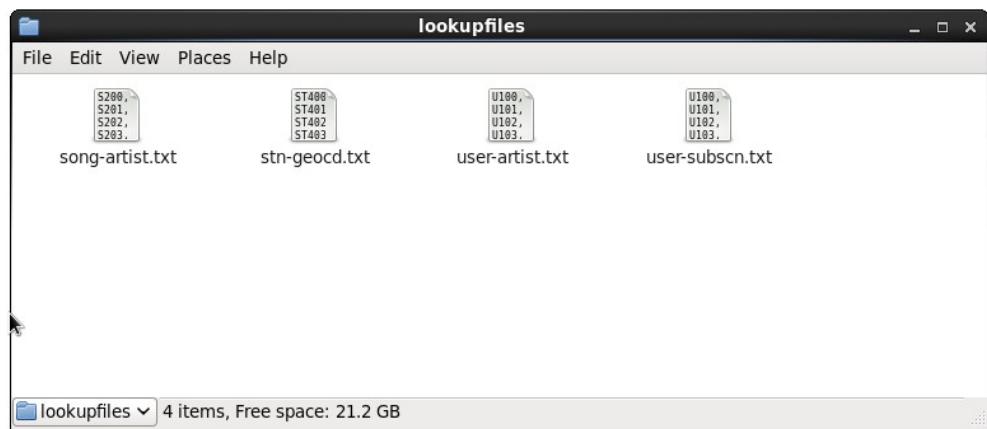
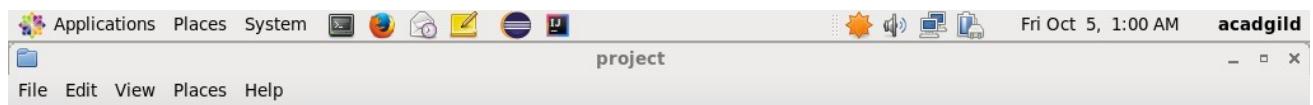
Table Name	Description
Station_Geo_Map	Contains mapping of a geo_cd with station_id
Subscribed_Users	Contains user_id, subscription_start_date and subscription_end_date. Contains details only for subscribed users
Song_Artist_Map	Contains mapping of song_id with artist_id alongwith royalty associated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





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

## **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 present in 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 :**

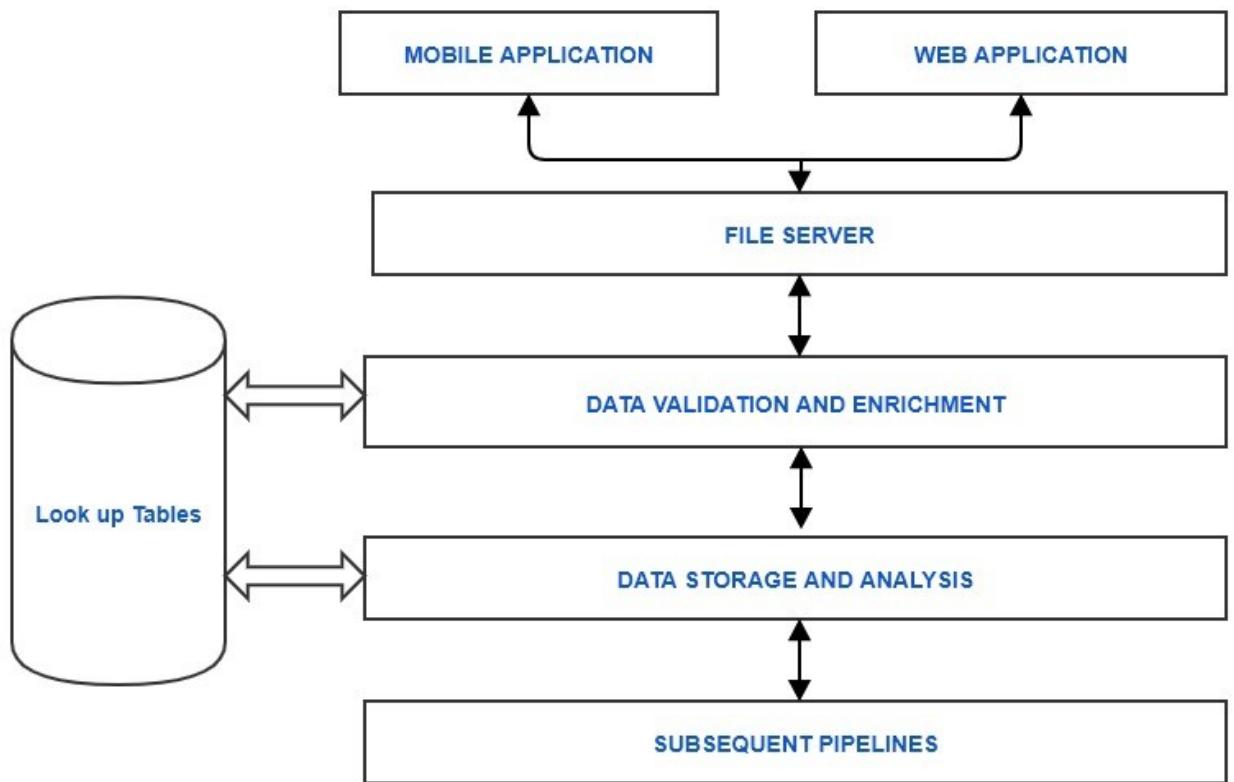
- LookUp tables are in NoSQL databases. Integrate them with the actual data flow.
- Try to make joins as less expensive as possible.
- Data Cleaning, Validation, Enrichment, Analysis and Post Analysis have to be automated.  
Try using schedulers.
- Appropriate logs have to maintain to track the behavior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

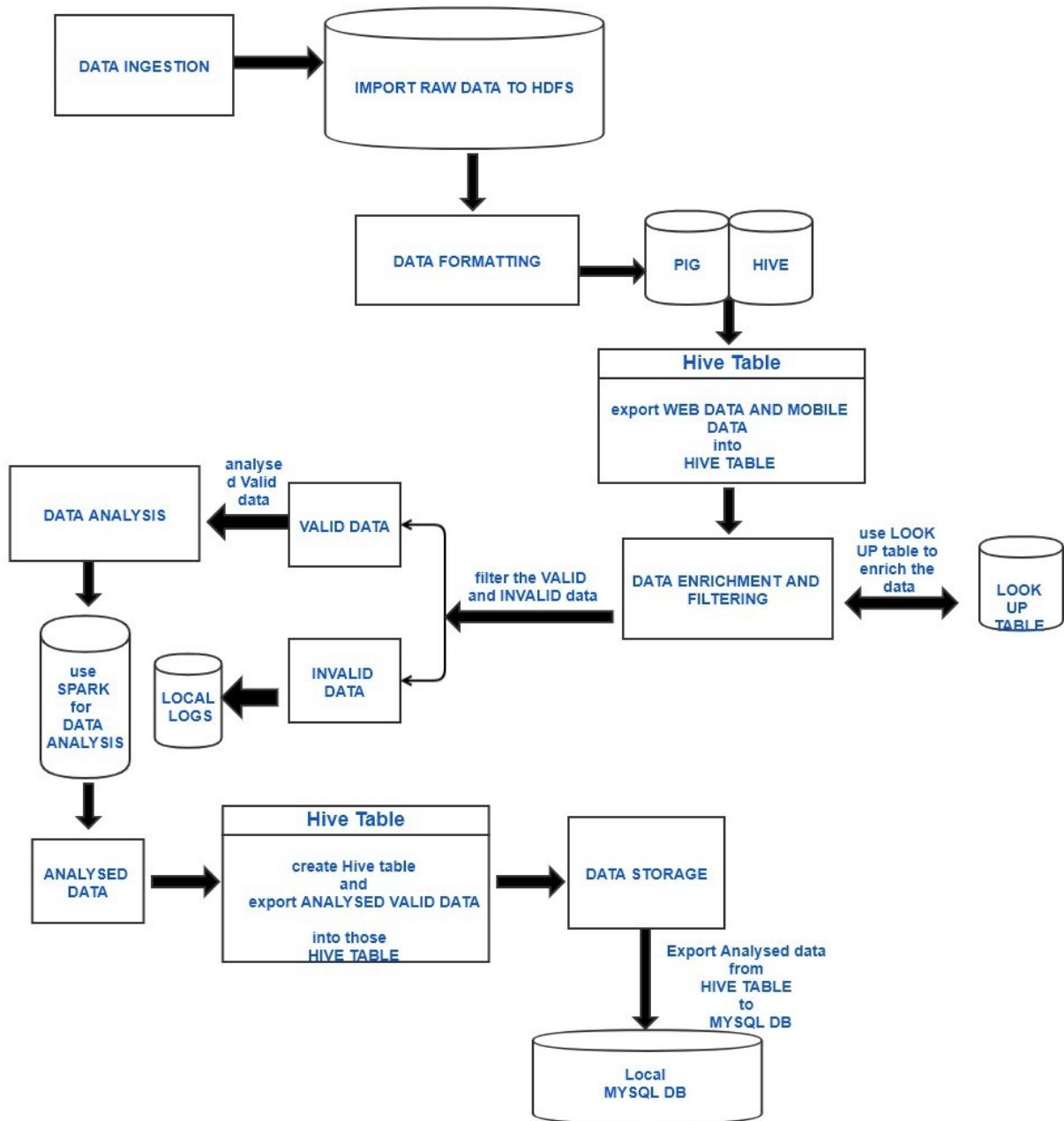
## 7.0 Design Architecture :



## 7.1 Low Level Design :



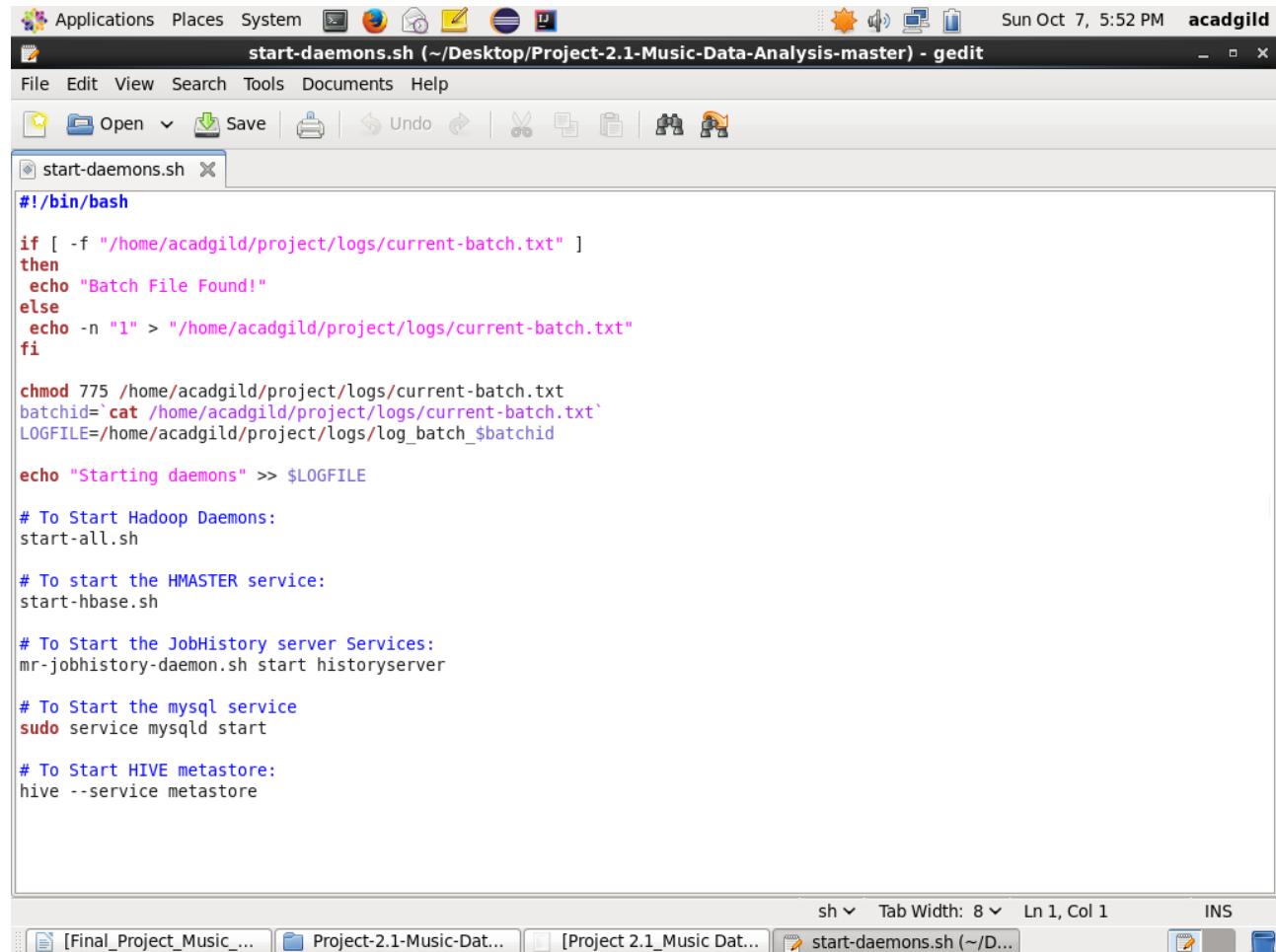
## 7.2 High Level Design :



## 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:



The screenshot shows a Gedit text editor window titled "start-daemons.sh (~/Desktop/Project-2.1-Music-Data-Analysis-master) - gedit". The window contains a shell script named "start-daemons.sh". The script starts with "#!/bin/bash" and includes logic to check if a log file exists, set permissions, and define variables for batch ID and log file. It then logs the start of daemons and lists several commands to start various Hadoop services like HDFS, YARN, and MapReduce, along with MySQL and Hive metastore.

```
#!/bin/bash

if [ -f "/home/acadgild/project/logs/current-batch.txt" ]
then
echo "Batch File Found!"
else
echo -n "1" > "/home/acadgild/project/logs/current-batch.txt"
fi

chmod 775 /home/acadgild/project/logs/current-batch.txt
batchid=`cat /home/acadgild/project/logs/current-batch.txt`
LOGFILE=/home/acadgild/project/logs/log_batch_$batchid

echo "Starting daemons" >> $LOGFILE

# To Start Hadoop Daemons:
start-all.sh

# To start the HMASTER service:
start-hbase.sh

# To Start the JobHistory server Services:
mr-jobhistory-daemon.sh start historyserver

# To Start the mysql service
sudo service mysqld start

# To Start HIVE metastore:
hive --service metastore
```

## **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](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]\$

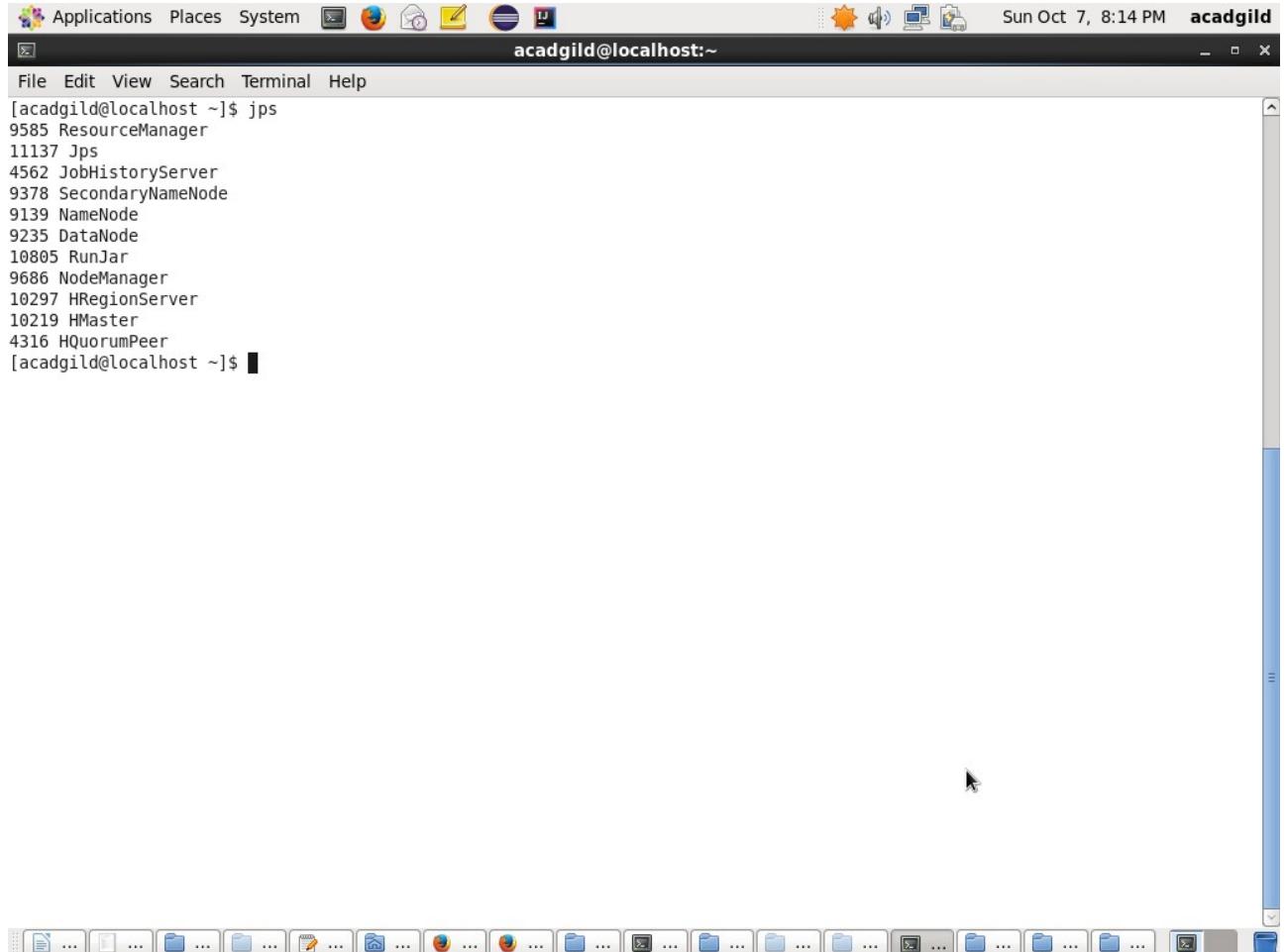
## Screenshot :

The screenshot shows a terminal window titled "acadgild@localhost:~/project/project/scripts". The window contains the following command and its 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 fi
le 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 libra
ry 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/ha
oop-2.6.5/logs/hadoop-acadgild-secondarynamenode-localhost.localdomain.out
18/10/07 19:52:40 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry 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-a
cadgild-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
```

The terminal window has a standard Linux desktop interface at the top with icons for Applications, Places, System, and various system status indicators. The bottom of the window shows a dock with various application icons.

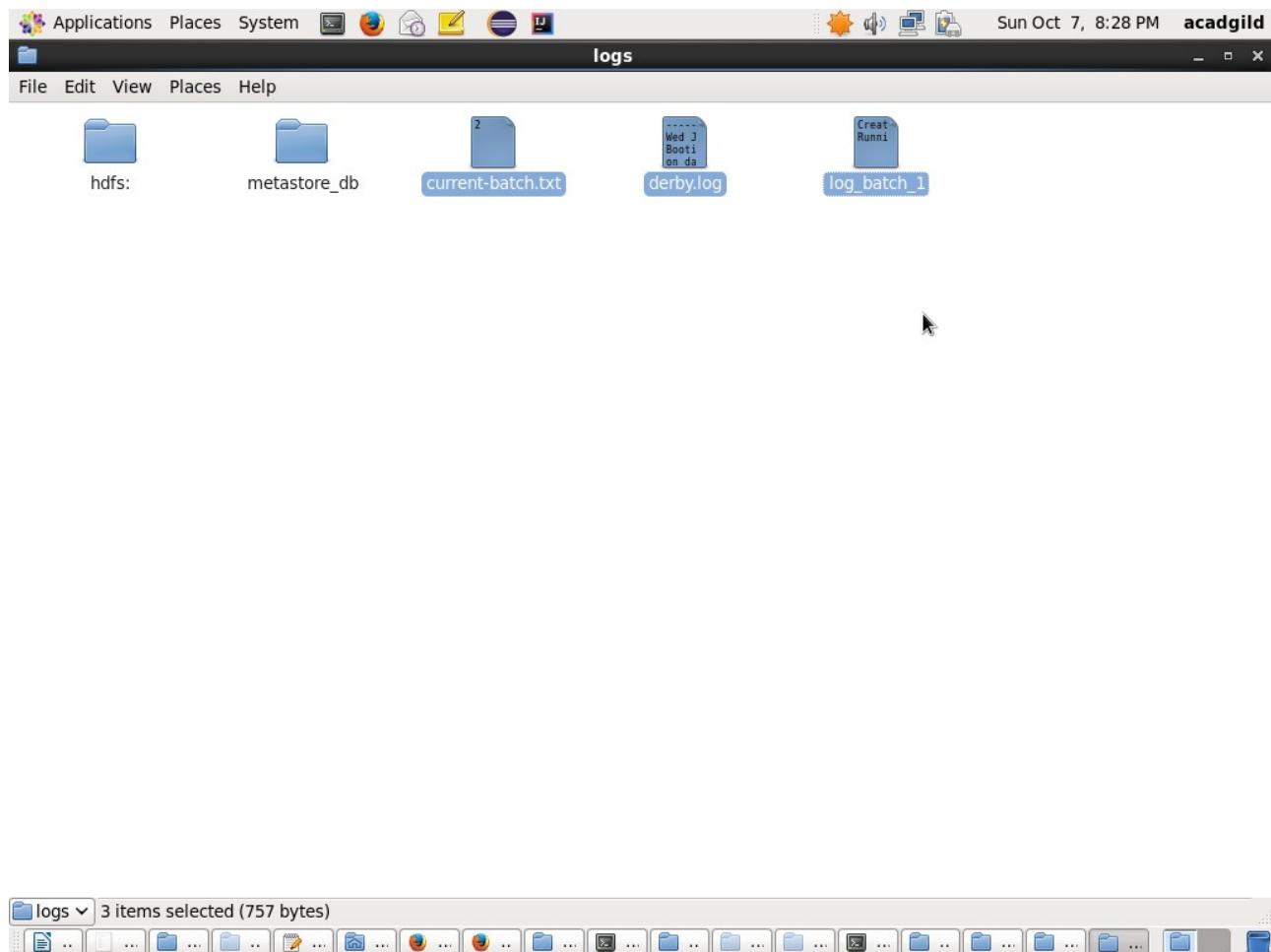
3. We can see the list active services using the jps command, see below screen shot and also Starting the hive metastore created a metastore\_db in the location where we desired,



The screenshot shows a terminal window titled "acadgild@localhost:~". The window contains the following text:

```
[acadgild@localhost ~]$ jps
9585 ResourceManager
11137 Jps
4562 JobHistoryServer
9378 SecondaryNameNode
9139 NameNode
9235 DataNode
10805 RunJar
9686 NodeManager
10297 HRegionServer
10219 HMaster
4316 QuorumPeer
[acadgild@localhost ~]$
```

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

**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 :**

The screenshot shows a Gedit text editor window titled "populate-lookup.sh (~/Desktop/project/scripts) - gedit". The window contains a bash script with syntax highlighting for commands like `cat`, `echo`, `hbase shell`, and `cut`. The script is used to create and populate HBase tables for station-geo-map, subscribed-users, and song-artist-map. It reads from files like "stn-geocd.txt" and "song-artist.txt" and writes logs to "log\_batch\_{batchid}.txt". The status bar at the bottom indicates "sh" as the file type, "Tab Width: 8", and "Ln 1, Col 1".

```
#!/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"
```

```
Applications Places System 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, rfe146eb48c24d56bcd2f669bb5ff8197e6c918b, 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, rfe146eb48c24d56bcd2f669bb5ff8197e6c918b, Sun Apr 22 20:42:02 PDT 2018

create 'subscribed-users', 'subscn'

```

```
Applications Places System File Edit View Search Terminal Help
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, rfe146eb48c24d56bcd2f669bb5ff8197e6c918b, Sun Apr 22 20:42:02 PDT 2018

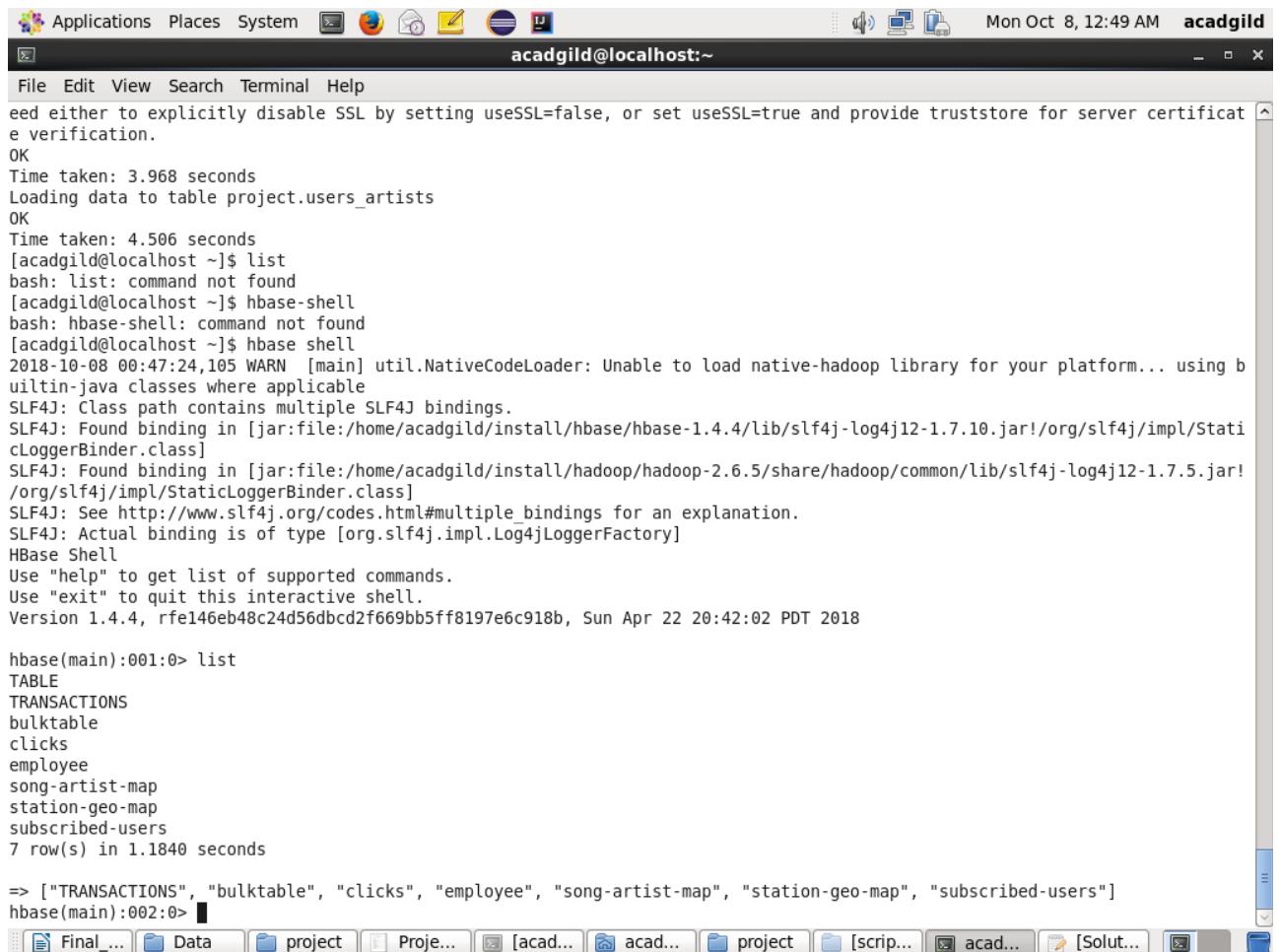
create 'song-artist-map', 'artist'
0 row(s) in 2.4600 seconds

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, rfe146eb48c24d56bcd2f669bb5ff8197e6c918b, Sun Apr 22 20:42:02 PDT 2018

put 'station-geo-map', 'ST400', 'geo:geo_cd', 'A'
0 row(s) in 3.1000 seconds
```

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

Lookup Tables in the hbase shell :



The screenshot shows a terminal window titled "acadgild@localhost:~". The window displays the output of an "hbase shell" session. The user runs the "list" command, which shows a list of tables: "TRANSACTIONS", "bulktable", "clicks", "employee", "song-artist-map", "station-geo-map", and "subscribed-users". The output also includes several warning messages from SLF4J about multiple bindings and native-hadoop library loading.

```
File Edit View Search Terminal Help
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
OK
Time taken: 3.968 seconds
Loading data to table project.users_artists
OK
Time taken: 4.506 seconds
[acadgild@localhost ~]$ list
bash: list: command not found
[acadgild@localhost ~]$ hbase-shell
bash: hbase-shell: command not found
[acadgild@localhost ~]$ hbase shell
2018-10-08 00:47:24,105 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.4.4/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
HBase Shell
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

hbase(main):001:0> list
TABLE
TRANSACTIONS
bulktable
clicks
employee
song-artist-map
station-geo-map
subscribed-users
7 row(s) in 1.1840 seconds

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

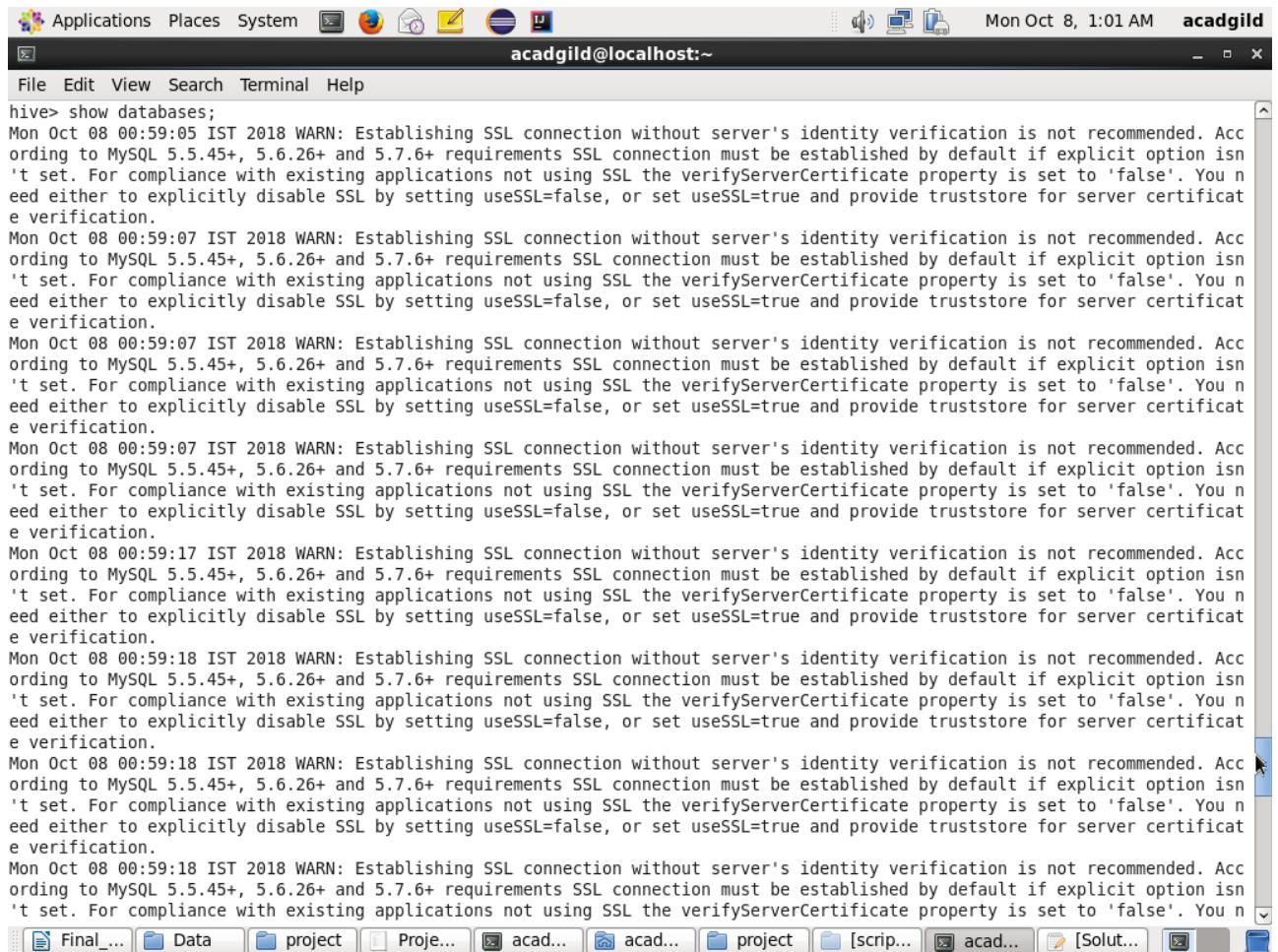
The values loaded in the Lookup tables are shown below :

```
Applications Places System File Edit View Search Terminal Help acadgild@localhost:~  
station-geo-map  
subscribed-users  
7 row(s) in 1.1840 seconds  
>> ["TRANSACTIONS", "bulktable", "clicks", "employee", "song-artist-map", "station-geo-map", "subscribed-users"]  
hbase(main):002:0> scan 'song-artist-map'  
ROW COLUMN+CELL  
S200 column=artist:artistid, timestamp=1538937462501, value=A300  
S201 column=artist:artistid, timestamp=1538937490676, value=A301  
S202 column=artist:artistid, timestamp=1538937519570, value=A302  
S203 column=artist:artistid, timestamp=1538937547939, value=A303  
S204 column=artist:artistid, timestamp=1538937576123, value=A304  
S205 column=artist:artistid, timestamp=1538937607485, value=A301  
S206 column=artist:artistid, timestamp=1538937637611, value=A302  
S207 column=artist:artistid, timestamp=1538937665303, value=A303  
S208 column=artist:artistid, timestamp=1538937693471, value=A304  
S209 column=artist:artistid, timestamp=1538937721338, value=A305  
10 row(s) in 0.6390 seconds  
hbase(main):003:0> scan 'station-geo-map'  
ROW COLUMN+CELL  
ST400 column=geo:geo_cd, timestamp=1538936953788, value=A  
ST401 column=geo:geo_cd, timestamp=1538936993765, value=AU  
ST402 column=geo:geo_cd, timestamp=1538937024468, value=AP  
ST403 column=geo:geo_cd, timestamp=1538937061367, value=J  
ST404 column=geo:geo_cd, timestamp=1538937097256, value=E  
ST405 column=geo:geo_cd, timestamp=1538937137819, value=A  
ST406 column=geo:geo_cd, timestamp=1538937175143, value=AU  
ST407 column=geo:geo_cd, timestamp=1538937216467, value=AP  
ST408 column=geo:geo_cd, timestamp=15389372669327, value=E  
ST409 column=geo:geo_cd, timestamp=1538937292849, value=E  
ST410 column=geo:geo_cd, timestamp=1538937321132, value=A  
ST411 column=geo:geo_cd, timestamp=1538937349282, value=A  
ST412 column=geo:geo_cd, timestamp=1538937377418, value=AP  
ST413 column=geo:geo_cd, timestamp=1538937405740, value=J  
ST414 column=geo:geo_cd, timestamp=1538937434425, value=E  
15 row(s) in 0.2950 seconds  
hbase(main):004:0>
```

```
Applications Places System File Edit View Search Terminal Help acadgild@localhost:~  
ST413 column=geo:geo_cd, timestamp=1538937405740, value=J  
ST414 column=geo:geo_cd, timestamp=1538937434425, value=E  
15 row(s) in 0.2950 seconds  
hbase(main):004:0> scan 'subscribed-users'  
ROW COLUMN+CELL  
U100 column=subscn:enddt, timestamp=1538937779238, value=1465130523  
U100 column=subscn:startdt, timestamp=1538937750527, value=1465230523  
U101 column=subscn:enddt, timestamp=1538937835551, value=1475130523  
U101 column=subscn:startdt, timestamp=1538937807374, value=1465230523  
U102 column=subscn:enddt, timestamp=1538937891931, value=1475130523  
U102 column=subscn:startdt, timestamp=1538937863615, value=1465230523  
U103 column=subscn:enddt, timestamp=1538937948263, value=1475130523  
U103 column=subscn:startdt, timestamp=1538937920056, value=1465230523  
U104 column=subscn:enddt, timestamp=1538938004746, value=1475130523  
U104 column=subscn:startdt, timestamp=1538937976440, value=1465230523  
U105 column=subscn:enddt, timestamp=1538938063730, value=1475130523  
U105 column=subscn:startdt, timestamp=1538938032995, value=1465230523  
U106 column=subscn:enddt, timestamp=1538938119990, value=1485130523  
U106 column=subscn:startdt, timestamp=1538938091896, value=1465230523  
U107 column=subscn:enddt, timestamp=1538938176498, value=1455130523  
U107 column=subscn:startdt, timestamp=1538938148435, value=1465230523  
U108 column=subscn:enddt, timestamp=1538938232881, value=1465230623  
U108 column=subscn:startdt, timestamp=1538938204823, value=1465230523  
U109 column=subscn:enddt, timestamp=1538938289088, value=1475130523  
U109 column=subscn:startdt, timestamp=1538938260823, value=1465230523  
U110 column=subscn:enddt, timestamp=15389383346826, value=1475130523  
U110 column=subscn:startdt, timestamp=1538938317252, value=1465230523  
U111 column=subscn:enddt, timestamp=1538938403622, value=1475130523  
U111 column=subscn:startdt, timestamp=1538938375580, value=1465230523  
U112 column=subscn:enddt, timestamp=1538938459451, value=1475130523  
U112 column=subscn:startdt, timestamp=1538938431775, value=1465230523  
U113 column=subscn:enddt, timestamp=1538938515039, value=1485130523  
U113 column=subscn:startdt, timestamp=1538938487304, value=1465230523  
U114 column=subscn:enddt, timestamp=1538938570468, value=1468130523  
U114 column=subscn:startdt, timestamp=1538938542801, value=1465230523  
15 row(s) in 0.5760 seconds  
hbase(main):005:0>
```

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.



A screenshot of a Linux terminal window titled "acadgild@localhost:~". The window contains the following text:

```
hive> show databases;
Mon Oct 08 00:59:05 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.
Mon Oct 08 00:59:07 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.
Mon Oct 08 00:59:07 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.
Mon Oct 08 00:59:07 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.
Mon Oct 08 00:59:17 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.
Mon Oct 08 00:59:18 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.
Mon Oct 08 00:59:18 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.
Mon Oct 08 00:59:18 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.
```

The terminal window has a standard Linux desktop interface with icons for Applications, Places, System, and various system status indicators. The title bar shows the session name "acadgild" and the date/time "Mon Oct 8, 1:01 AM". The bottom of the window shows a row of open file tabs.

```
Applications Places System Mon Oct 8, 1:02 AM acadgild
File Edit View Search Terminal Help
e verification.
OK
acadgildbb
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. 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.
Mon Oct 08 01:01:14 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.
Mon Oct 08 01:01:14 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.
Mon Oct 08 01:01:14 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.
Mon Oct 08 01:01:14 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.
Mon Oct 08 01:01:14 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.
```

```
hive> Select * From users artists;
```

```
Applications Places System acadgild@localhost:~ Mon Oct 8, 1:02 AM acadgild
File Edit View Search Terminal Help
e verification.
Mon Oct 08 01:01:14 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.
Mon Oct 08 01:01:14 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.
Mon Oct 08 01:01:14 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.
Mon Oct 08 01:01:15 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.
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 these lookup tables in hive using the Hbase Storage Handler.

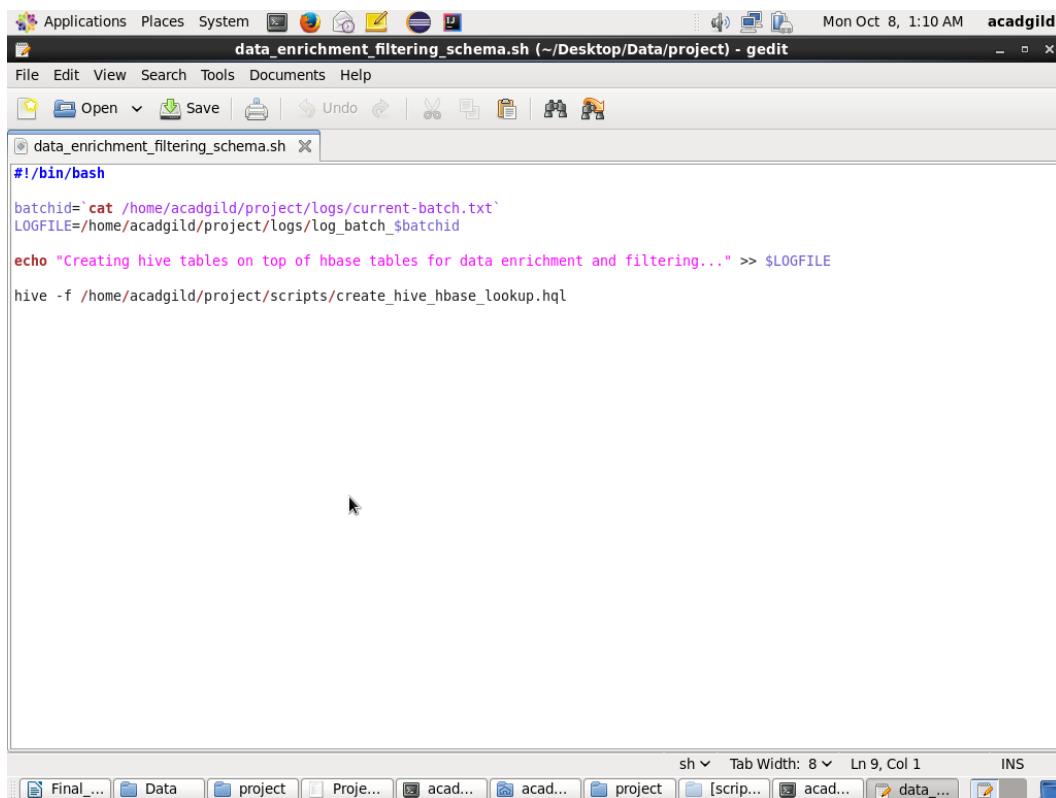
With the help of “**data\_enrichment\_filtering\_schema.sh**” file we will create hive tables on the top of Hbase tables using “**create\_hive\_hbase\_lookup.hql**”.

### **Creating Hive Tables on the top of Hbase:**

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.

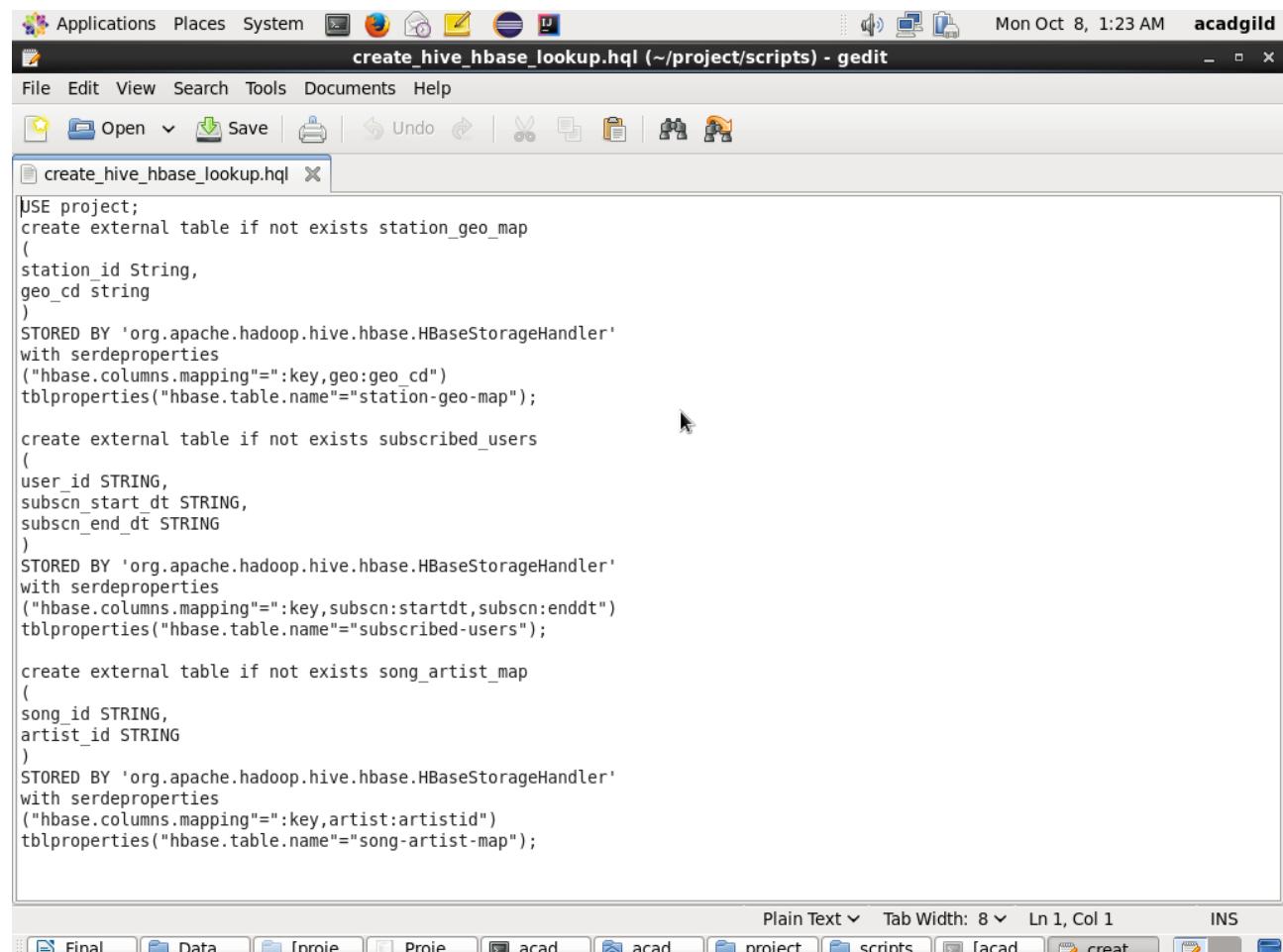
Run the script: **./data\_enrichment\_filtering\_schema.sh**,

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.



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

## create\_hive\_hbase\_lookup.hql



The screenshot shows a Gedit text editor window with the file 'create\_hive\_hbase\_lookup.hql' open. The window title bar reads 'create\_hive\_hbase\_lookup.hql (~/project/scripts) - gedit'. The status bar at the bottom right shows 'Mon Oct 8, 1:23 AM acadgild'. The main text area contains the following Hive/HBase schema definition:

```
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");

create external table if not exists subscribed_users
(
user_id STRING,
subscn_start_dt STRING,
subscn_end_dt STRING
)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
with serdeproperties
("hbase.columns.mapping"=:key,subscn:startdt,subscn:enddt")
tblproperties("hbase.table.name"="subscribed-users");

create external table if not exists song_artist_map
(
song_id STRING,
artist_id STRING
)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
with serdeproperties
("hbase.columns.mapping"=:key,artist:artistid")
tblproperties("hbase.table.name"="song-artist-map");
```

The status bar at the bottom also displays 'Plain Text' and 'Tab Width: 8'.

The below screenshot we can see tables getting created in hive by running the “**“data\_enrichement\_filtering\_schema.sh** file”

```
File Edit View Search Terminal Help
Time taken: 0.981 seconds
[acadgild@localhost ~]$ sh /home/acadgild/project/scripts/data_enrichment_filtering_schema.sh
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.3-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/docs/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-log4j.properties Async: true
Mon Oct 08 01:30:26 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 01:30:28 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.
```

Applications Places System

Mon Oct 8, 1:31 AM acadgild

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. 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.
```

```
Mon Oct 08 01:30:28 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.
```

```
Mon Oct 08 01:30:38 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.
```

```
Mon Oct 08 01:30:38 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.
```

```
Mon Oct 08 01:30:38 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.
```

```
Mon Oct 08 01:30:38 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.
```

```
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 ~]$
```

Hive>Show Tables;

```
[x] Applications Places System [ ] Mon Oct 8, 1:38 AM acadgild
File Edit View Search Terminal Help
e verification.
OK
acadgilddb
custom
default
project
Time taken: 24.204 seconds, Fetched: 4 row(s)
hive> use project;
OK
Time taken: 0.138 seconds
hive> show tables;
Mon Oct 08 01:37:47 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 01:37:47 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 01:37:47 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 01:37:47 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 01:37:47 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 01:37:48 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
```

```
Applications Places System acadgild@localhost:~  
File Edit View Search Terminal Help  
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 01:37:47 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 01:37:47 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 01:37:47 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 01:37:48 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 01:37:48 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 01:37:48 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  
song_artist_map  
station_geo_map  
subscribed_users  
users_artists  
Time taken: 1.853 seconds, Fetched: 4 row(s)  
hive>   
Final... Data project [Proje... [acad... acad... [proje... scripts acad... Soluti...  
[  ]
```

```
hive>Select * From song_artist_map
```

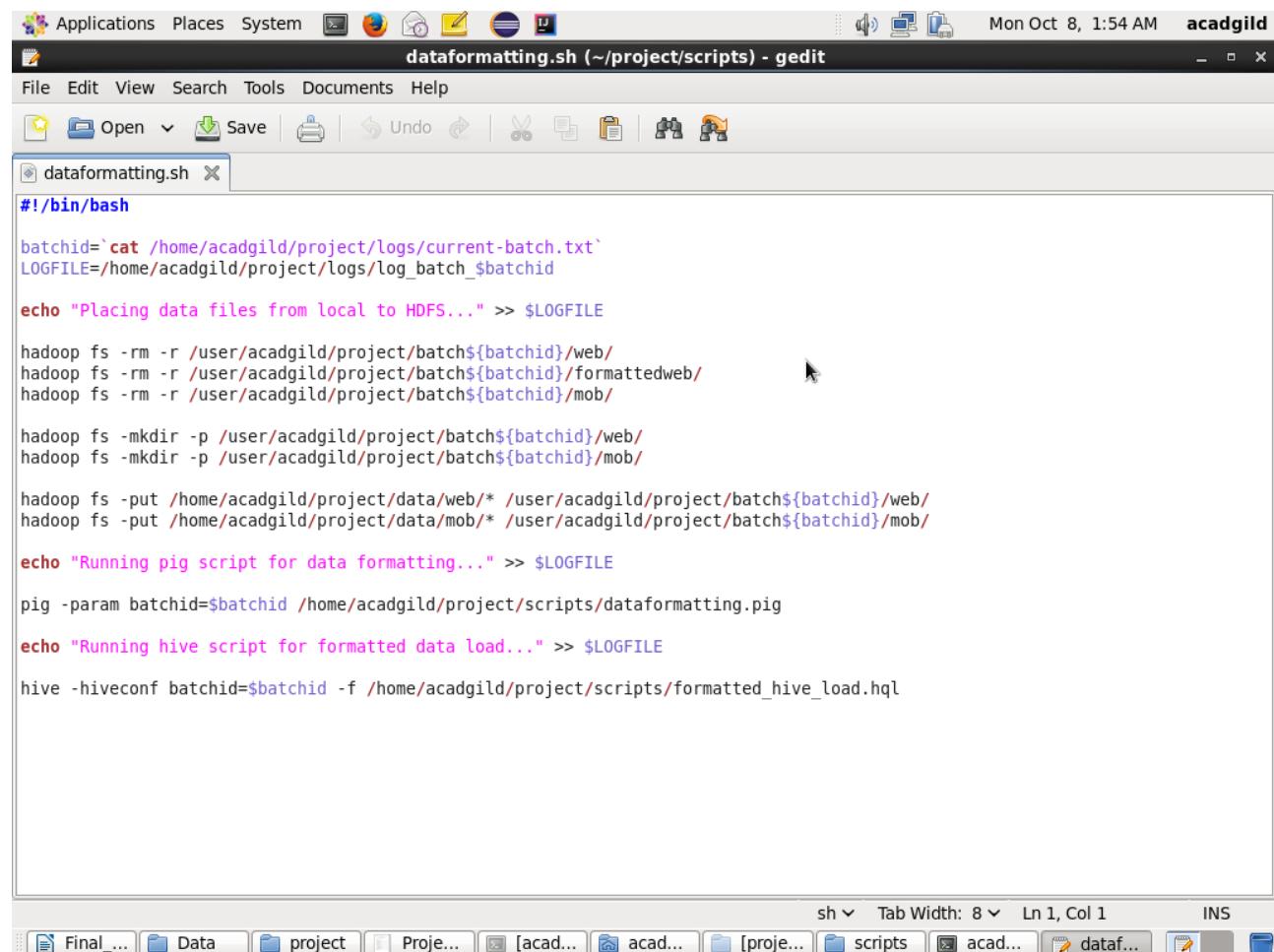
```
hive>Select * From station_geo_map
```

```
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**



The screenshot shows a Gedit text editor window titled "dataformatting.sh (~/project/scripts) - gedit". The window contains a bash script with various commands for data processing and file operations. The script includes commands for reading a batch ID from a file, creating log files, using Hadoop to move files to HDFS, and running Pig and Hive scripts for data formatting and loading.

```
#!/bin/bash

batchid=`cat /home/acadgild/project/logs/current-batch.txt`
LOGFILE=/home/acadgild/project/logs/log_batch_${batchid}

echo "Placing data files from local to HDFS..." >> $LOGFILE

hadoop fs -rm -r /user/acadgild/project/batch${batchid}/web/
hadoop fs -rm -r /user/acadgild/project/batch${batchid}/formattedweb/
hadoop fs -rm -r /user/acadgild/project/batch${batchid}/mob/

hadoop fs -mkdir -p /user/acadgild/project/batch${batchid}/web/
hadoop fs -mkdir -p /user/acadgild/project/batch${batchid}/mob/

hadoop fs -put /home/acadgild/project/data/web/* /user/acadgild/project/batch${batchid}/web/
hadoop fs -put /home/acadgild/project/data/mob/* /user/acadgild/project/batch${batchid}/mob/

echo "Running pig script for data formatting..." >> $LOGFILE
pig -param batchid=$batchid /home/acadgild/project/scripts/dataformatting.pig
echo "Running hive script for formatted data load..." >> $LOGFILE
hive -hiveconf batchid=$batchid -f /home/acadgild/project/scripts/formatted_hive_load.hql
```

The screenshot shows a terminal window titled "acadgild@localhost:~". The window displays the output of a "sh /home/acadgild/project/scripts/dataformatting.sh" command. The output includes several "WARN" messages from the "util.NativeCodeLoader" class, indicating issues with loading native-hadoop libraries. It also shows "INFO" messages from the "pig.ExecTypeProvider" class, specifying the execution type as LOCAL or MAPREDUCE. The log concludes with the Apache Pig version information: "Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49". The terminal window has a standard Linux desktop interface with icons for Applications, Places, System, and various system status indicators at the top. The bottom of the window shows a dock with icons for various files and applications.

```
[acadgild@localhost ~]$ sh /home/acadgild/project/scripts/dataformatting.sh
18/10/08 01:56:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
rm: '/user/acadgild/project/batch2/web/': No such file or directory
18/10/08 01:56:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
rm: '/user/acadgild/project/batch2/formattedweb/': No such file or directory
18/10/08 01:56:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
rm: '/user/acadgild/project/batch2/mob/': No such file or directory
18/10/08 01:56:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/10/08 01:56:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/10/08 01:56:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/10/08 01:56:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/10/08 01:57:00 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/10/08 01:57:00 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/10/08 01:57:00 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-10-08 01:57:00,668 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-10-08 01:57:00,668 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1538944020661.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-1.4.4/lib/slf4j-log4j12-1.7.10.jar!/*org/slf4j/impl/Stati
cLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-10-08 01:57:02,657 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
2018-10-08 01:57:03,614 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-10-08 01:57:04,168 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead,
use mapreduce.jobtracker.address
2018-10-08 01:57:04,168 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead,
use fs.defaultFS
```

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 batch ID's

## Dataformatting.pig

The screenshot shows a Gedit text editor window titled "dataformatting.pig (~/project/scripts) - gedit". The window contains a Pig Latin script with the following code:

```
REGISTER /home/acadgild/project/lib/piggybank.jar;
DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();
A = LOAD '/user/acadgild/project/batch${batchid}/web/' using org.apache.pig.piggybank.storageXMLLoader('record') as (x:chararray);
B = FOREACH A GENERATE TRIM(XPath(x, 'record/user_id')) AS user_id,
    TRIM(XPath(x, 'record/song_id')) AS song_id,
    TRIM(XPath(x, 'record/artist_id')) AS artist_id,
    ToUnixTime(ToDate(TRIM(XPath(x, 'record/timestamp')),'yyyy-MM-dd HH:mm:ss')) AS timestp,
    ToUnixTime(ToDate(TRIM(XPath(x, 'record/start_ts')),'yyyy-MM-dd HH:mm:ss')) AS start_ts,
    ToUnixTime(ToDate(TRIM(XPath(x, 'record/end_ts')),'yyyy-MM-dd HH:mm:ss')) AS end_ts,
    TRIM(XPath(x, 'record/geo_cd')) AS geo_cd,
    TRIM(XPath(x, 'record/station_id')) AS station_id,
    TRIM(XPath(x, 'record/song_end_type')) AS song_end_type,
    TRIM(XPath(x, 'record/like')) AS like,
    TRIM(XPath(x, 'record/dislike')) AS dislike;
STORE B INTO '/user/acadgild/project/batch${batchid}/formattedweb/' USING PigStorage(',');
```

The status bar at the bottom of the editor shows the file path "Loading file '/home/acadgild/project/scripts/dataformatting.pig'" and other status information like "Plain Text", "Tab Width: 8", "Ln 1, Col 1", and "INS".

## formatted\_hive\_load.hql

The screenshot shows a Gedit text editor window titled "formatted\_hive\_load.hql (~/project/scripts) - gedit". The window contains the following HiveQL code:

```
USE project;
CREATE TABLE IF NOT EXISTS formatted_input
(
user_id STRING,
song_id STRING,
artist_id STRING,
`timestamp` STRING,
start_ts STRING,
end_ts STRING,
geo_cd STRING,
station_id STRING,
song_end_type INT,
`like`| INT,
dislike INT
)
PARTITIONED BY
(batchid INT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';

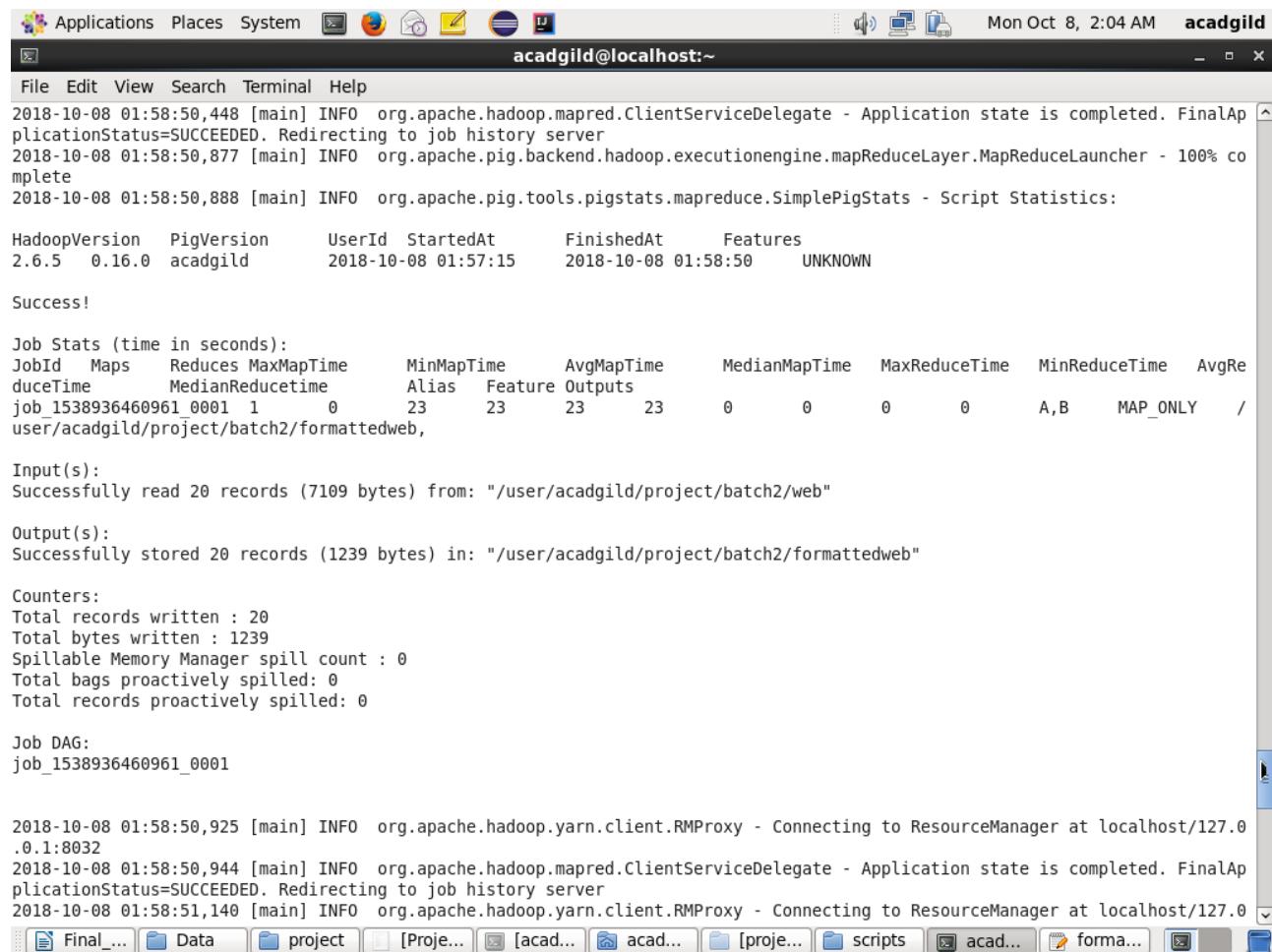
LOAD DATA INPATH '/user/acadgild/project/batch${hiveconf:batchid}/formattedweb/'
INTO TABLE formatted_input PARTITION (batchid=${hiveconf:batchid});

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

The status bar at the bottom of the editor shows "Plain Text" and "Tab Width: 8". The cursor is positioned near the end of the first partition definition.

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:

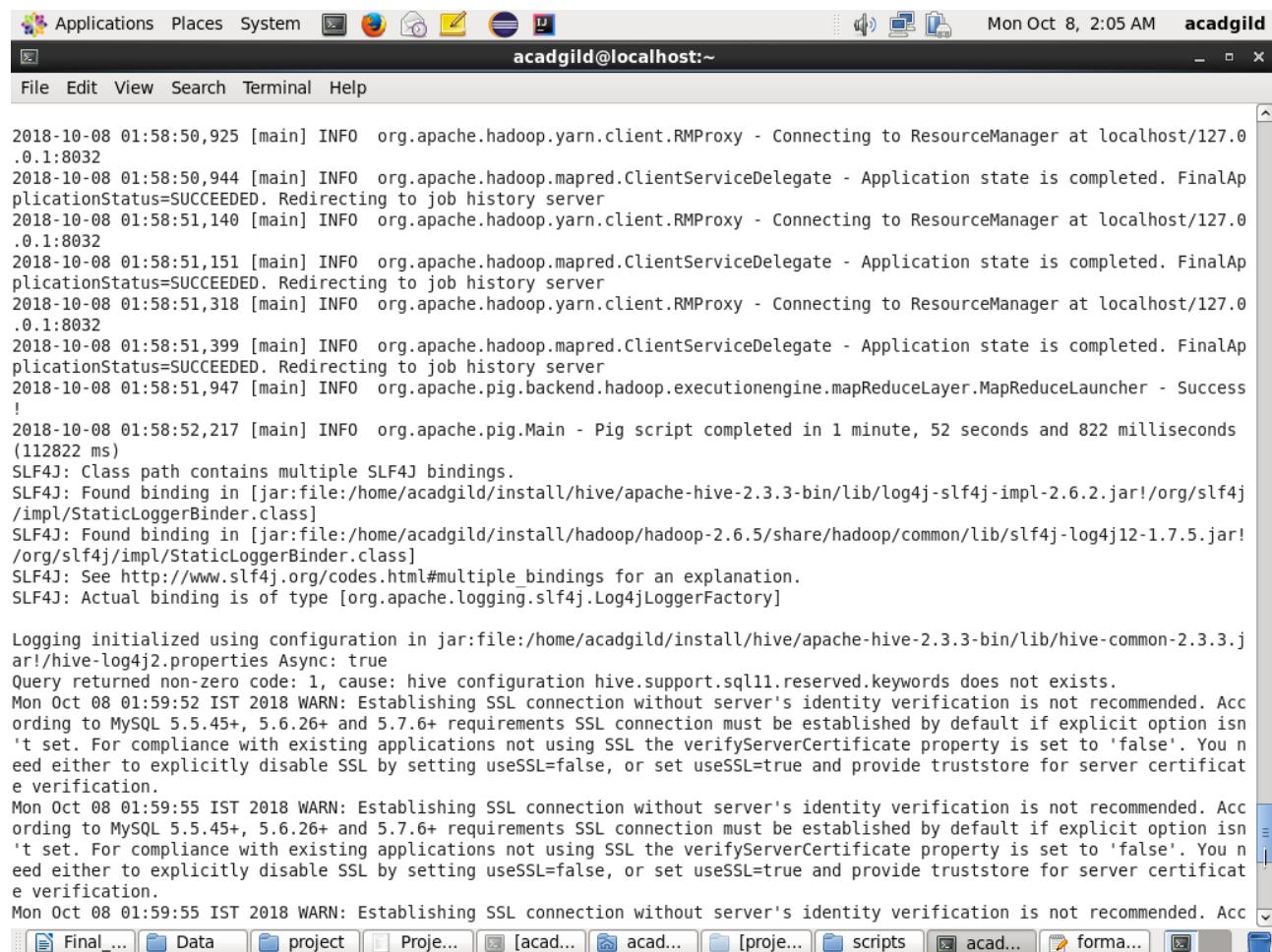


The screenshot shows a terminal window titled "acadgild@localhost:~". The terminal displays the output of a Pig script. The output includes logs from Hadoop and Pig, statistics for the job, and details about input and output. The script was run on a local host and completed successfully.

```
File Edit View Search Terminal Help
2018-10-08 01:58:50,448 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2018-10-08 01:58:50,877 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% co
mplete
2018-10-08 01:58:50,888 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.5 0.16.0 acadgild 2018-10-08 01:57:15 2018-10-08 01:58:50 UNKNOWN
Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgRe
duceTime MedianReducetime Alias Feature Outputs
job_1538936460961_0001 1 0 23 23 23 23 0 0 0 A,B MAP_ONLY /
user/acadgild/project/batch2/formattedweb,
Input(s):
Successfully read 20 records (7109 bytes) from: "/user/acadgild/project/batch2/web"
Output(s):
Successfully stored 20 records (1239 bytes) in: "/user/acadgild/project/batch2/formattedweb"
Counters:
Total records written : 20
Total bytes written : 1239
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_1538936460961_0001

2018-10-08 01:58:50,925 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0
.0.1:8032
2018-10-08 01:58:50,944 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2018-10-08 01:58:51,140 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0
```

## Hive script successfully load the data into hive terminal :



The screenshot shows a terminal window titled "acadgild@localhost:~". The window displays the output of a Hive script. The log includes several INFO messages from org.apache.hadoop.mapred.ClientServiceDelegate indicating successful application completion and redirection to job history servers. It also shows a success message from org.apache.pig.backend.hadoop.executionengine.mapReduceLauncher. The log concludes with multiple WARN messages from MySQL regarding SSL connection security, which are repeated multiple times.

```
2018-10-08 01:58:50,925 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.1:8032
2018-10-08 01:58:50,944 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-10-08 01:58:51,140 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.1:8032
2018-10-08 01:58:51,151 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-10-08 01:58:51,318 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at localhost/127.0.1:8032
2018-10-08 01:58:51,399 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2018-10-08 01:58:51,947 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLauncher - Success !
2018-10-08 01:58:52,217 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 52 seconds and 822 milliseconds (112822 ms)
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.jar!/hive-log4j2.properties Async: true
Query returned non-zero code: 1, cause: hive configuration hive.support.sql11.reserved.keywords does not exists.
Mon Oct 08 01:59:52 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.
Mon Oct 08 01:59:55 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.
Mon Oct 08 01:59:55 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
```

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:

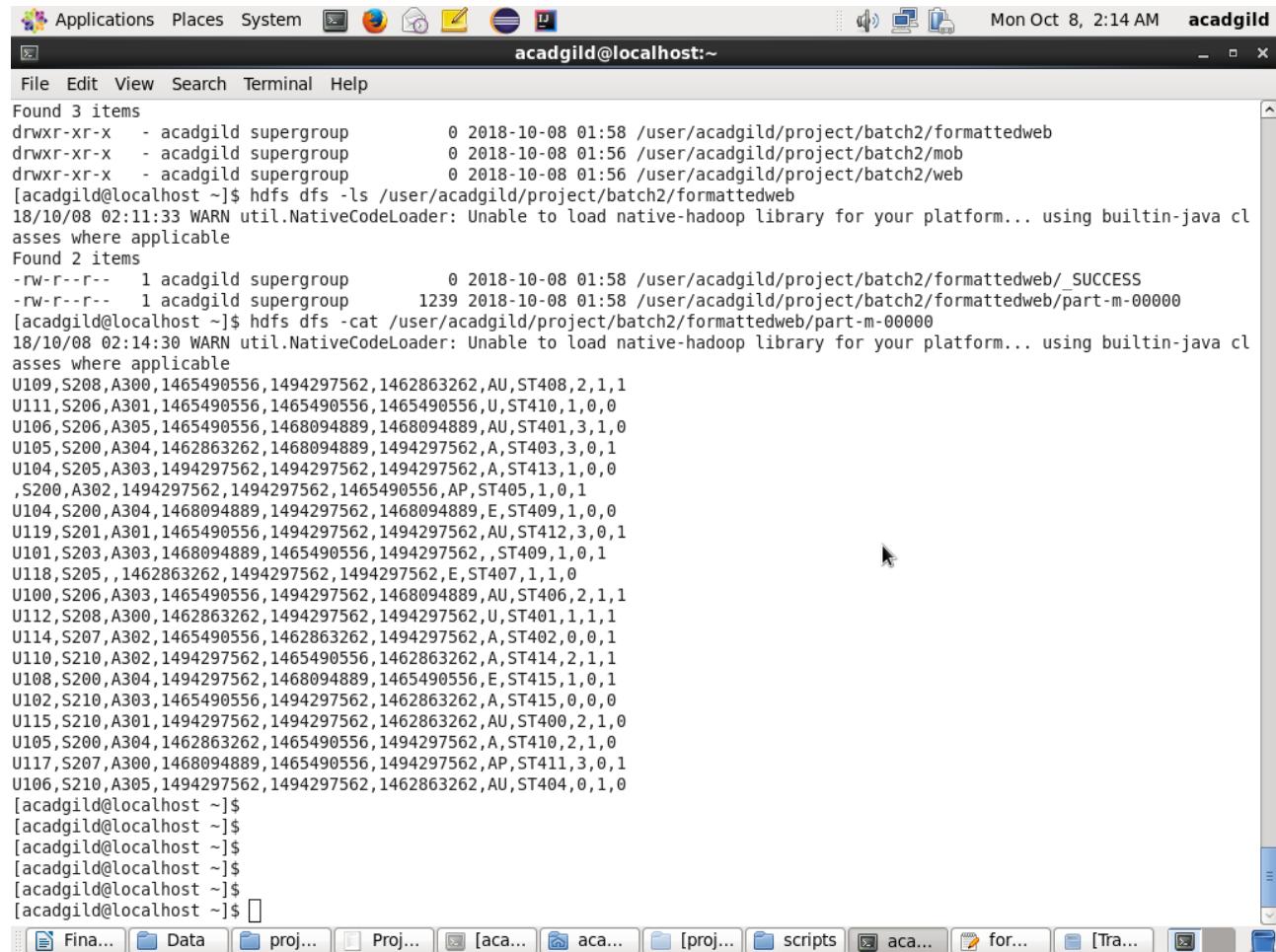


```
Applications Places System Mon Oct 8, 2:11 AM acadgild
acadgild@localhost:~ File Edit View Search Terminal Help
Mon Oct 08 02:00:05 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL 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.
Mon Oct 08 02:00:05 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.
Mon Oct 08 02:00:05 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.
Mon Oct 08 02:00:05 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.
Mon Oct 08 02:00:05 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 ~]$ hdfs dfs -ls /user/acadgild/project
18/10/08 02:10:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x  - acadgild supergroup          0 2018-10-08 01:58 /user/acadgild/project/batch2
[acadgild@localhost ~]$ hdfs dfs -ls /user/acadgild/project/batch2
18/10/08 02:10:26 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 2018-10-08 01:58 /user/acadgild/project/batch2/formattedweb
drwxr-xr-x  - acadgild supergroup          0 2018-10-08 01:56 /user/acadgild/project/batch2/mob
drwxr-xr-x  - acadgild supergroup          0 2018-10-08 01:56 /user/acadgild/project/batch2/web
[acadgild@localhost ~]$ hdfs dfs -ls /user/acadgild/project/batch2/formattedweb
18/10/08 02:11:33 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 2018-10-08 01:58 /user/acadgild/project/batch2/formattedweb/_SUCCESS
-rw-r--r--  1 acadgild supergroup        1239 2018-10-08 01:58 /user/acadgild/project/batch2/formattedweb/part-m-00000
[acadgild@localhost ~]$
```

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
```



A screenshot of a Linux terminal window titled "acadgild@localhost:~". The window shows the output of the command "hdfs dfs -cat /user/acadgild/project/batch1/formattedweb/part-m-00000". The output consists of several lines of text representing file contents, primarily numerical values separated by commas and colons. The terminal interface includes a menu bar with File, Edit, View, Search, Terminal, Help, and a toolbar with icons for Applications, Places, System, and various system status indicators.

```
Found 3 items
drwxr-xr-x - acadgild supergroup          0 2018-10-08 01:58 /user/acadgild/project/batch2/formattedweb
drwxr-xr-x - acadgild supergroup          0 2018-10-08 01:56 /user/acadgild/project/batch2/mob
drwxr-xr-x - acadgild supergroup          0 2018-10-08 01:56 /user/acadgild/project/batch2/web
[acadgild@localhost ~]$ hdfs dfs -ls /user/acadgild/project/batch2/formattedweb
18/10/08 02:11:33 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 2018-10-08 01:58 /user/acadgild/project/batch2/formattedweb/_SUCCESS
-rw-r--r-- 1 acadgild supergroup    1239 2018-10-08 01:58 /user/acadgild/project/batch2/formattedweb/part-m-00000
[acadgild@localhost ~]$ hdfs dfs -cat /user/acadgild/project/batch2/formattedweb/part-m-00000
18/10/08 02:14:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
U109,S208,A300,1465490556,1494297562,1462863262,AU,ST408,2,1,1
U111,S206,A301,1465490556,1465490556,1465490556,U,ST410,1,0,0
U106,S206,A305,1465490556,1468094889,1468094889,AU,ST401,3,1,0
U105,S200,A304,1462863262,1468094889,1494297562,A,ST403,3,0,1
U104,S205,A303,1494297562,1494297562,1494297562,A,ST413,1,0,0
,S200,A302,1494297562,1494297562,1465490556,AP,ST405,1,0,1
U104,S200,A304,1468094889,1494297562,1468094889,E,ST409,1,0,0
U119,S201,A301,1465490556,1494297562,1494297562,AU,ST412,3,0,1
U101,S203,A303,1468094889,1465490556,1494297562,,ST409,1,0,1
U118,S205,,1462863262,1494297562,1494297562,E,ST407,1,1,0
U100,S206,A303,1465490556,1494297562,1468094889,AU,ST406,2,1,1
U112,S208,A300,1462863262,1494297562,1494297562,U,ST401,1,1,1
U114,S207,A302,1465490556,1462863262,1494297562,A,ST402,0,0,1
U110,S210,A302,1494297562,1465490556,1462863262,A,ST414,2,1,1
U108,S200,A304,1494297562,1468094889,1465490556,E,ST415,1,0,1
U102,S210,A303,1465490556,1494297562,1462863262,A,ST415,0,0,0
U115,S210,A301,1494297562,1494297562,1462863262,AU,ST400,2,1,0
U105,S200,A304,1462863262,1465490556,1494297562,A,ST410,2,1,0
U117,S207,A300,1468094889,1465490556,1494297562,AP,ST411,3,0,1
U106,S210,A305,1494297562,1494297562,1462863262,AU,ST404,0,1,0
[acadgild@localhost ~]$
[acadgild@localhost ~]$
[acadgild@localhost ~]$
[acadgild@localhost ~]$
[acadgild@localhost ~]$
[acadgild@localhost ~]$
```

The new Tables has been created and show below :

```
Applications Places System Tue Oct 9, 1:10 AM acadgild
acadgild@localhost:~ File Edit View Search Terminal Help
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. 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.
Tue Oct 09 01:10:37 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.
Tue Oct 09 01:10:37 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.
Tue Oct 09 01:10:37 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.
Tue Oct 09 01:10:37 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.
Tue Oct 09 01:10:38 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.
[T...][a...][c...][sc...][F...][...][pr...][ac...][ac...][p...][s...][ac...][Pr...]
```

```
Applications Places System Tue Oct 9, 1:11 AM acadgild
acadgild@localhost:~ File Edit View Search Terminal Help
e verification.
Tue Oct 09 01:10:37 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.
Tue Oct 09 01:10:37 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.
Tue Oct 09 01:10:37 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.
Tue Oct 09 01:10:37 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.
Tue Oct 09 01:10:38 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.
Tue Oct 09 01:10:38 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.
OK
formatted input
song_artist_map
station_geo_map
subscribed_users
users_artists
Time taken: 2.515 seconds, Fetched: 5 row(s)
hive> [S...][a...][c...][sc...][F...][...][pr...][ac...][ac...][p...][s...][ac...][Pr...]
```

## DataFormatting.sh output in hive terminal,

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

```
Applications Places System Terminal Help
acadgild@localhost:~ 
File Edit View Search Terminal Help
Time taken: 2.515 seconds, Fetched: 5 row(s)
hive> select * from formatted_input;
OK
U105    S203    A304    1495130523    1485130523    1485130523    A    ST404    1    1    1    2
U104    S208    A303    1465130523    1485130523    1465130523    E    ST408    3    1    0    2
U105    S202    A305    1475130523    1465130523    1485130523    E    ST410    0    0    1    2
U117    S204    A302    1495130523    1465230523    1465230523    U    ST404    3    0    1    2
U115    S201    A303    1465230523    1475130523    1465230523    AU   ST414    3    0    1    2
U205    S203    A303    1475130523    1475130523    1485130523    A    ST406    1    1    0    2
U106    S200    A303    1465130523    1485130523    1465130523    AP   ST413    1    1    0    2
U119    S204    A305    1475130523    1475130523    1475130523    AU   ST400    3    1    1    2
U102    S209    A300    1465130523    1475130523    1475130523    ST401    2    1    0    2
U117    S207    A304    1495130523    1485130523    1465130523    E    ST411    1    0    1    2
U108    S208    A300    1465130523    1465230523    1465130523    E    ST406    2    0    0    2
U120    S203    A305    1475130523    1465230523    1465230523    AU   ST411    1    0    0    2
U115    S202    A305    1465130523    1465130523    1475130523    A    ST401    1    0    1    2
U115    S200    A304    1465130523    1465130523    1485130523    AP   ST407    3    1    0    2
U111    S207    A305    1495130523    1465230523    1465130523    AP   ST405    2    1    1    2
U107    S202    A302    1465230523    1465130523    1475130523    E    ST402    2    1    1    2
U119    S204    A300    1475130523    1465130523    1485130523    E    ST403    2    1    1    2
U102    S209    A303    1465230523    1465130523    1465130523    U    ST414    3    0    1    2
U101    S209    A304    1475130523    1475130523    1485130523    AU   ST415    3    0    0    2
U106    S202    A301    1475130523    1485130523    1465130523    E    ST403    3    0    1    2
U100    S200    A300    1494297562    1468094889    1494297562    AP   ST403    0    0    1    2
U101    S206    A302    1468094889    1468094889    1494297562    U    ST405    3    1    1    2
U116    S208    A300    1468094889    1468094889    1462863262    AU   ST400    3    0    1    2
U109    S200    A305    1468094889    1462863262    1494297562    AU   ST405    2    1    1    2
U106    S208    A304    1462863262    1468094889    1465490556    AU   ST405    1    1    1    2
U203    A301    1462863262    1494297562    1465490556    AU   ST405    1    0    0    2
U106    S202    A300    1465490556    1494297562    1462863262    AP   ST410    0    0    1    2
U120    S202    A301    1494297562    1468094889    1494297562    U    ST406    3    1    0    2
U117    S208    A305    1494297562    1462863262    1462863262    ST408    0    0    1    2
U109    S200    1465490556    1494297562    1462863262    AP   ST408    3    0    0    2
U103    S210    A300    1462863262    1465490556    1462863262    E    ST415    2    1    1    2
U108    S200    A303    1465490556    1468094889    1465490556    AP   ST400    2    1    1    2
U101    S209    A302    1494297562    1465490556    1468094889    E    ST404    3    0    0    2
U116    S210    A300    1494297562    1468094889    1494297562    AP   ST415    3    0    0    2
U105    S203    A304    1494297562    1465490556    1468094889    AU   ST401    3    0    1    2
U103    S200    A303    1494297562    1465490556    1462863262    AP   ST402    2    1    1    2
```

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**

The screenshot shows a Gedit text editor window titled "data\_enrichment.sh (~/project/scripts) - gedit". The window contains a bash script with syntax highlighting. The script performs several tasks: it reads the current batch ID from a log file, sets up log and data directories, runs a Hive query to copy data, creates valid and invalid directories if they don't exist, copies valid and invalid records to the local file system, and finally deletes old records from the local file system. The script uses echo statements to log its actions to a log file.

```
#!/bin/bash

batchid=`cat /home/acadgild/project/logs/current-batch.txt`
LOGFILE=/home/acadgild/project/logs/log_batch_$batchid
VALIDDIR=/home/acadgild/project/processed_dir/valid/batch_$batchid
INVALIDDIR=/home/acadgild/project/processed_dir/invalid/batch_$batchid

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

hive -hiveconf batchid=$batchid -f /home/acadgild/project/scripts/data_enrichment.hql

if [ ! -d "$VALIDDIR" ]
then
mkdir -p "$VALIDDIR"
fi

if [ ! -d "$INVALIDDIR" ]
then
mkdir -p "$INVALIDDIR"
fi

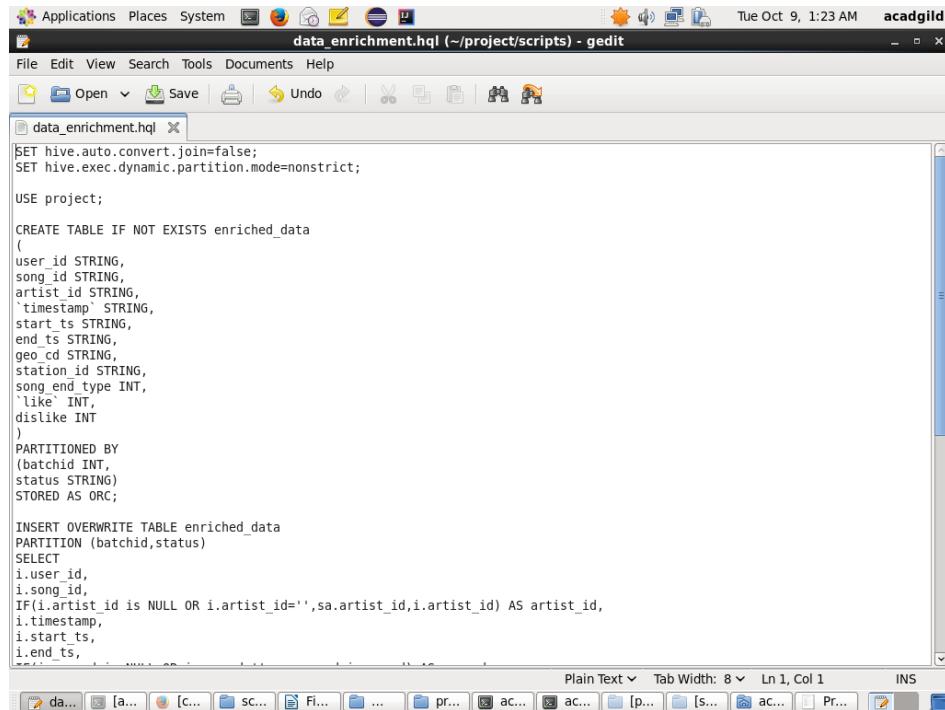
echo "Copying valid and invalid records in local file system..." >> $LOGFILE

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

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

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

## **data\_enrichment.hql :**

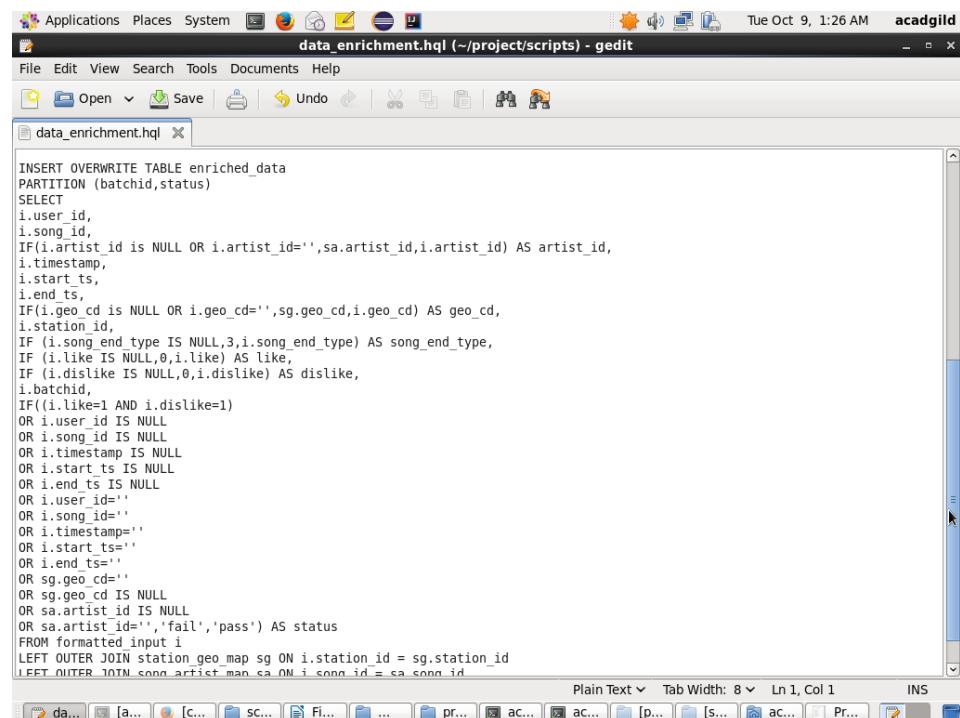


```
SET hive.auto.convert.join=false;
SET hive.exec.dynamic.partition.mode=nonstrict;

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,
```



```
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
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
```

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

```
USE project;
```

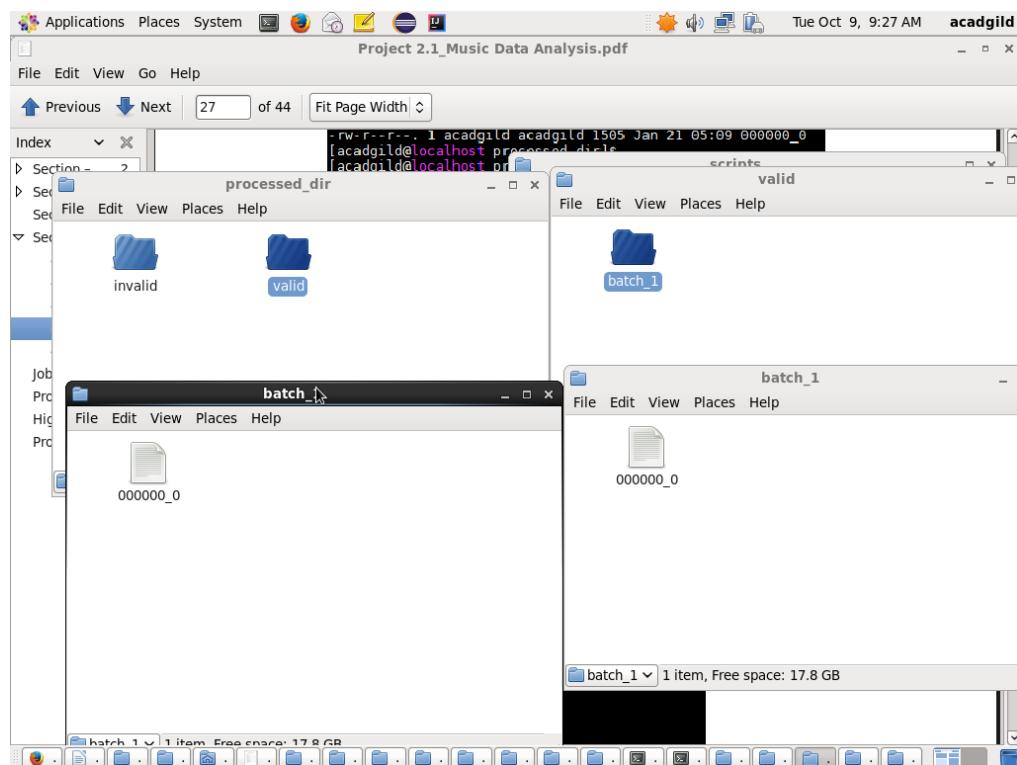
```
CREATE TABLE IF NOT EXISTS enriched_data  
(  
    user_id STRING,  
    song_id STRING,  
    artist_id STRING,  
    `timestamp` STRING,  
    start_ts STRING,  
    end_ts STRING,  
    geo_cd STRING,  
    station_id STRING,  
    song_end_type INT,  
    `like` INT,  
    dislike INT  
)  
PARTITIONED BY  
(batchid INT,  
status STRING)  
STORED AS ORC;
```

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

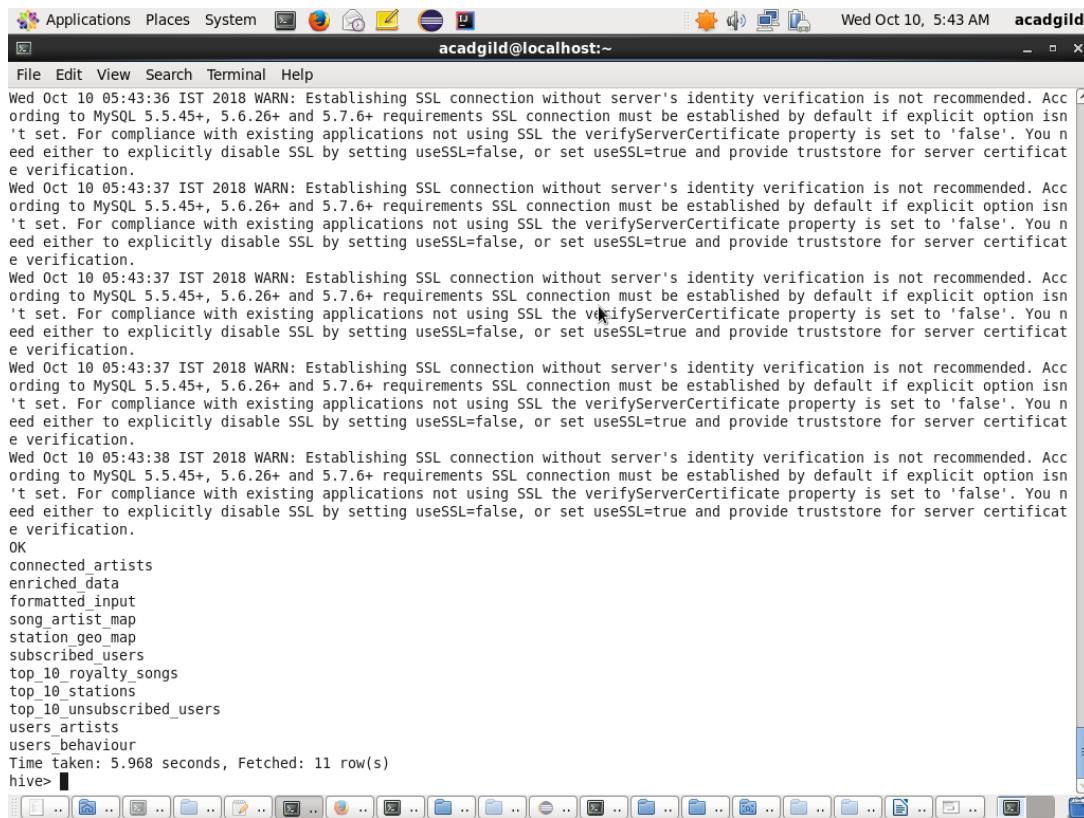
```
i.song_id,  
IF(i.artist_id is NULL OR i.artist_id=",sa.artist_id,i.artist_id) AS artist_id,  
i.`timestamp`,  
i.start_ts,  
i.end_ts,  
IF(i.geo_cd is NULL OR i.geo_cd=",sg.geo_cd,i.geo_cd) AS geo_cd,  
i.station_id,  
IF (i.song_end_type IS NULL,3,i.song_end_type) AS song_end_type,  
IF (i.`like` 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.**



```
File Edit View Search Terminal Help
Wed Oct 10 05:43:36 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.
Wed Oct 10 05:43:37 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.
Wed Oct 10 05:43:37 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.
Wed Oct 10 05:43:37 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.
Wed Oct 10 05:43:38 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.
OK
connected_artists
enriched_data
formatted_input
song_artist_map
station_geo_map
subscribed_users
top_10_royalty_songs
top_10_stations
top_10_unsubscribed_users
users_artists
usersBehaviour
Time taken: 5.968 seconds, Fetched: 11 row(s)
hive> [REDACTED]
```

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;
```

```
Applications Places System Firefox Calc Evolution Picasa Sunflower Terminal acdgild@localhost:~ File Edit View Search Terminal Help
hive> select * from enriched_data;
OK
U103 S200 A303 1494297562 1465490556 1462863262 AP ST402 2 1 1 1 2 fail
U109 S200 A305 1468094889 1462863262 1494297562 AU ST405 2 1 1 1 2 fail
U108 S200 A303 1465490556 1468094889 1465490556 AP ST400 2 1 1 1 2 fail
U107 S202 A302 1465230523 1465130523 1475130523 E ST402 2 1 1 1 2 fail
S203 A301 1462863262 1494297562 1465490556 AU ST405 1 0 0 0 2 fail
U105 S203 A304 1495130523 1485130523 1485130523 A ST404 1 1 1 1 2 fail
U119 S204 A305 1475130523 1475130523 1475130523 AU ST400 3 1 1 1 2 fail
U119 S204 A300 1475130523 1465130523 1485130523 E ST403 2 1 1 1 2 fail
U117 S205 A301 1465490556 1468094889 1465490556 U ST415 1 0 0 1 2 fail
S205 A303 1475130523 1475130523 1485130523 A ST406 1 1 0 2 2 fail
U101 S206 A302 1468094889 1468094889 1494297562 U ST405 3 1 1 1 2 fail
U111 S207 A305 1495130523 1465230523 1465130523 AP ST405 2 1 1 1 2 fail
U106 S208 A304 1462863262 1468094889 1465490556 AU ST405 1 1 1 1 2 fail
U101 S209 A304 1475130523 1475130523 1485130523 AU ST415 3 0 0 2 2 fail
U103 S210 A300 1462863262 1465490556 1462863262 E ST415 2 1 1 1 2 fail
U116 S210 A300 1494297562 1468094889 1494297562 AP ST415 3 0 0 2 2 fail
U109 S200 A300 1465490556 1494297562 1462863262 AP ST408 3 0 0 0 2 pass
U100 S200 A300 1494297562 1468094889 1494297562 AP ST403 0 0 0 1 2 pass
U115 S200 A304 1465130523 1465130523 1485130523 AP ST407 3 1 0 2 2 pass
U106 S200 A303 1465130523 1485130523 1465130523 AP ST413 1 1 0 2 2 pass
U101 S200 A304 1462863262 1494297562 1494297562 AP ST413 2 0 0 0 2 pass
U115 S201 A303 1465230523 1475130523 1465230523 AU ST414 3 0 1 2 2 pass
U106 S202 A301 1475130523 1485130523 1465130523 E ST403 3 0 1 2 2 pass
U115 S202 A305 1465130523 1465130523 1475130523 A ST401 1 0 1 2 2 pass
U106 S202 A300 1465490556 1494297562 1462863262 AP ST410 0 0 1 2 2 pass
U105 S202 A305 1475130523 1465130523 1485130523 E ST410 0 0 1 2 2 pass
U120 S202 A301 1494297562 1468094889 1494297562 U ST406 3 1 0 2 2 pass
U120 S203 A305 1475130523 1475130523 1465230523 AU ST411 1 0 0 2 2 pass
U105 S203 A304 1494297562 1465490556 1468094889 AU ST401 3 0 1 2 2 pass
U117 S204 A302 1495130523 1465230523 1465230523 U ST404 3 0 1 2 2 pass
U108 S204 A301 1468094889 1465490556 1494297562 AU ST410 0 1 0 2 2 pass
U117 S207 A303 1495130523 1485130523 1465130523 E ST411 1 0 1 2 2 pass
U108 S208 A300 1465130523 1465230523 1465130523 E ST406 2 0 0 2 2 pass
U117 S208 A305 1494297562 1462863262 1462863262 E ST408 0 0 1 2 2 pass
U104 S208 A303 1465130523 1485130523 1465130523 E ST408 3 1 0 2 2 pass
U116 S208 A300 1468094889 1468094889 1462863262 AU ST400 3 0 1 2 2 pass
U101 S209 A302 1494297562 1465490556 1468094889 E ST404 3 0 0 2 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

The screenshot shows a Gedit window with the title "data\_analysis.sh (~/project/scripts) - gedit". The menu bar includes "File", "Edit", "View", "Search", "Tools", "Documents", and "Help". The toolbar contains icons for "Open", "Save", "Undo", "Redo", "Cut", "Copy", "Paste", "Find", and "Replace". There are two tabs open: "data\_analysis.sh" and "data\_export.sh". The "data\_analysis.sh" tab contains the following script:

```
batchid=`cat /home/acadgild/project/logs/current-batch.txt`  
LOGFILE=/home/acadgild/project/logs/log_batch_$batchid  
  
echo "Running script for data analysis in Spark..." >> $LOGFILE  
  
chmod 775 /home/acadgild/project/lib/sparkanalysis.jar  
  
/home/acadgild/install/spark/spark-2.1.0-bin-hadoop2.6/bin/spark-submit \  
--class /home/acadgild/project/scripts/Spark_analysis \  
--master local[2] \  
--driver-class-path /home/acadgild/install/hive/apache-hive-2.3.3-bin/lib/hive-hbase-handler-2.3.3.jar:/home/acadgild/install/hbase/hbase-1.4.4/lib/* \  
/home/acadgild/project/lib/sparkanalysis.jar $batchid  
  
/home/acadgild/install/spark/spark-2.1.0-bin-hadoop2.6/bin/spark-submit \  
--class /home/acadgild/project/scripts/Spark_analysis_2 \  
--master local[2] \  
--driver-class-path /home/acadgild/install/hive/apache-hive-2.3.3-bin/lib/hive-hbase-handler-2.3.3.jar:/home/acadgild/install/hbase/hbase-1.4.4/lib/* \  
/home/acadgild/project/lib/sparkanalysis.jar $batchid  
  
sh /home/acadgild/project/scripts/data_export.sh  
  
echo "Incrementing batchid..." >> $LOGFILE  
  
batchid=`expr $batchid + 1`  
echo -n $batchid > /home/acadgild/project/logs/current-batch.txt
```

## Spark\_analysis.scala

```
import org.apache.spark.sql.SQLContext
import org.apache.hadoop.hive.
import org.apache.spark.sql.hive.HiveContext

object Spark_analysis {

def main(args: Array[String]): Unit = {
    val sparkSession = SparkSession.builder()
        .master("local[2]")
        .appName("Data Analysis Main_1")
        .config("spark.sql.warehouse.dir","/user/hive/warehouse")
        .config("hive.metastore.uris","thrift://127.0.0.1:9083")
        .enableHiveSupport()
        .getOrCreate()

    val batchId = args(0)

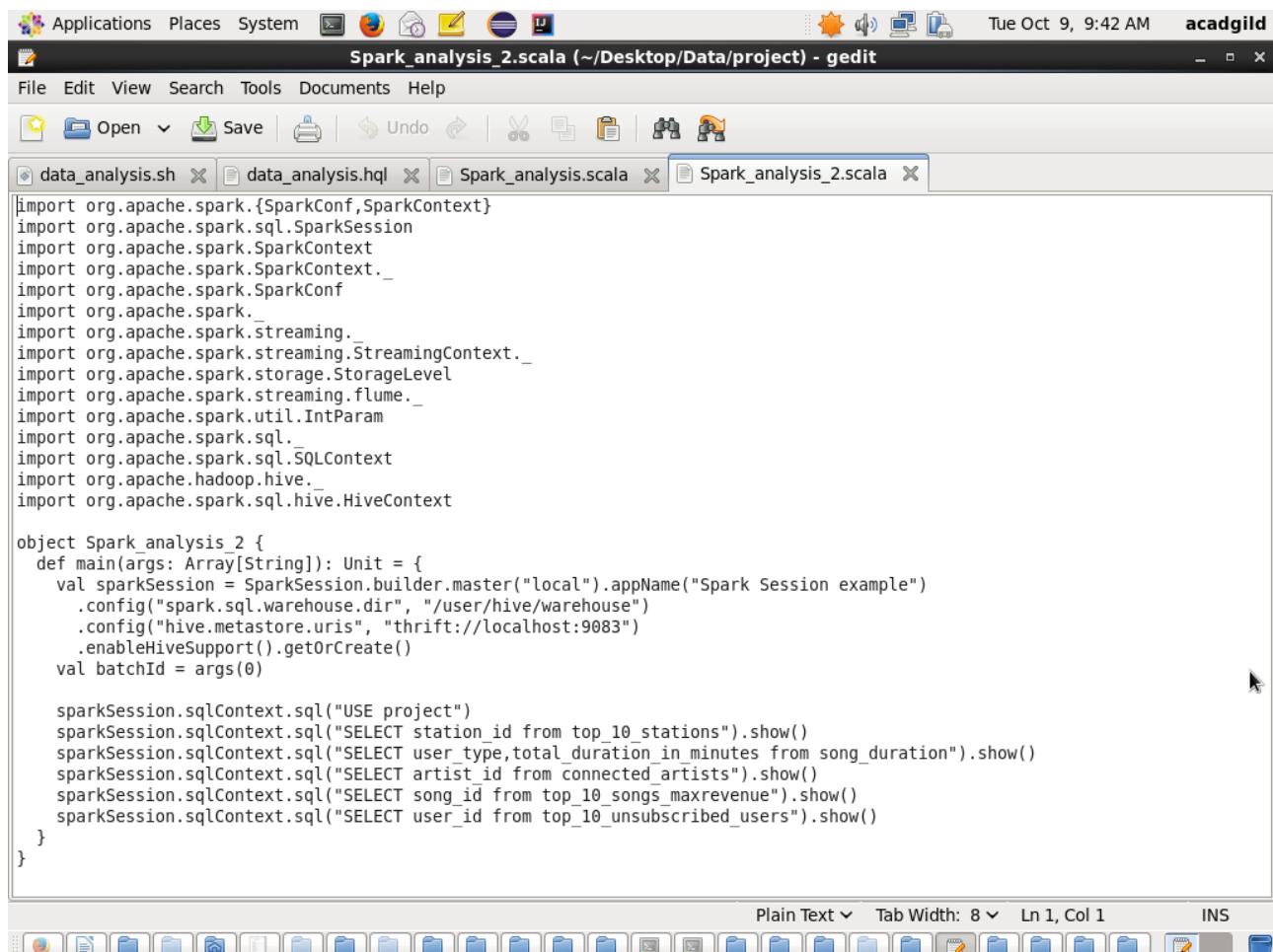
    //<<<<<<----- PROBLEM 1 - Creation of table and Insertion of data ----->>>>>>>
    //Determine top 10 station_id(s) where maximum number of songs were played, which were liked by unique users.

    val set_properties = sparkSession.sqlContext.sql("set hive.auto.convert.join=false")

    val use_project_database = sparkSession.sqlContext.sql("USE project")

    val create hive_table_top_10_stations = sparkSession.sqlContext.sql("CREATE TABLE IF NOT EXISTS
project.top_10_stations"+
        "("+
        " station_id STRING,"+
        " total_distinct_songs_played INT,"+
        " distinct_user_count INT"+
        ")"+
        " PARTITIONED BY (batchid INT)"+
        " ROW FORMAT DELIMITED"+
```

## Spark\_analysis\_2.scala



The screenshot shows a Gedit text editor window titled "Spark\_analysis\_2.scala (~/Desktop/Data/project) - gedit". The window contains Scala code for a Spark session example. The code imports various Spark modules and defines a main method that creates a SparkSession, configures it for a local master and a specific Hive metastore, and then runs several SQL queries on the session.

```
import org.apache.spark.{SparkConf,SparkContext}
import org.apache.spark.sql.SparkSession
import org.apache.spark.SparkContext
import org.apache.spark.SparkContext._
import org.apache.spark.SparkConf
import org.apache.spark._
import org.apache.spark.streaming._
import org.apache.spark.streaming.StreamingContext._
import org.apache.spark.storage.StorageLevel
import org.apache.spark.streaming.flume._
import org.apache.spark.util.IntParam
import org.apache.spark.sql._
import org.apache.spark.sql.SQLContext
import org.apache.hadoop.hive._
import org.apache.spark.sql.hive.HiveContext

object Spark_analysis_2 {
  def main(args: Array[String]): Unit = {
    val sparkSession = SparkSession.builder.master("local").appName("Spark Session example")
      .config("spark.sql.warehouse.dir", "/user/hive/warehouse")
      .config("hive.metastore.uris", "thrift://localhost:9083")
      .enableHiveSupport().getOrCreate()
    val batchId = args(0)

    sparkSession.sqlContext.sql("USE project")
    sparkSession.sqlContext.sql("SELECT station_id from top_10_stations").show()
    sparkSession.sqlContext.sql("SELECT user_type,total_duration_in_minutes from song_duration").show()
    sparkSession.sqlContext.sql("SELECT artist_id from connected_artists").show()
    sparkSession.sqlContext.sql("SELECT song_id from top_10_songs_maxrevenue").show()
    sparkSession.sqlContext.sql("SELECT user_id from top_10_unsubscribed_users").show()
  }
}
```

```

File Edit View Terminal Help
[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.jar!/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. 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.
Tue Oct 09 12:52:27 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.
Tue Oct 09 12:52:27 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.
Tue Oct 09 12:52:38 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.
Tue Oct 09 12:52:38 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.

```

```

File Edit View Terminal Help
Tue Oct 09 12:55 PM acadgild@localhost:~
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 need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate 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 execution 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 :

```
Applications Places System Wed Oct 10, 7:18 AM acadgild
acadgild@localhost:~ File Edit View Search Terminal Help
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.
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 Hive 1.X releases.
Query ID = acadgild_20181010030425_c6a61611-6651-4f99-be81-e800ac820cf
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_1539113996591_0024, Tracking URL = http://localhost:8088/proxy/application_1539113996591_0024/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1539113996591_0024
Hadoop job information for Stage-1: number of mappers: 0; number of reducers: 1
2018-10-10 03:05:08,784 Stage-1 map = 0%,  reduce = 0%
2018-10-10 03:05:23,497 Stage-1 map = 0%,  reduce = 100%
2018-10-10 03:05:24,695 Stage-1 map = 0%,  reduce = 0%
Ended Job = job_1539113996591_0024 with errors
Error during job, obtaining debugging information...
FAILED: Execution Error, return code 2 from org.apache.hadoop.hive.ql.exec.mr.MapReduceTask
MapReduce Jobs Launched:
Stage-Stage-1: Reduce: 1 HDFS Read: 0 HDFS Write: 0 FAIL
Total MapReduce CPU Time Spent: 0 msec
[acadgild@localhost ~]$ sh /home/acadgild/project/scripts/data_analysis.sh
zip warning: name not matched: META-INF/*.DSA
zip warning: name not matched: META-INF/*.RSA
zip warning: name not matched: META-INF/*.SF

zip error: Nothing to do! (/home/acadgild/project/lib/sparkanalysis.jar)
Warning: Local jar /home/acadgild/hbase/hbase-1.4.4/lib/* does not exist, skipping
```

## The tables have also been created in the Hive:

```
Applications Places System Wed Oct 10, 5:43 AM acadgild
acadgild@localhost:~ File Edit View Search Terminal Help
Wed Oct 10 05:43:36 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.
Wed Oct 10 05:43:37 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.
Wed Oct 10 05:43:37 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.
Wed Oct 10 05:43:37 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.
Wed Oct 10 05:43:38 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.
OK
connected_artists
enriched_data
formatted_input
song_artist_map
station_geo_map
subscribed_users
top_10_royalty_songs
top_10_stations
top_10_unsubscribed_users
users_artists
users_behaviour
Time taken: 5.968 seconds, Fetched: 11 row(s)
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.

```
Time taken: 0.097 seconds, Fetched: 11 row(s)
hive> Select * From connected_artists;
OK
connected_artists.artist_id      connected_artists.user_count      connected_artists.batchid
A383    2          1
A382    2          1
A380    1          1
Time taken: 0.225 seconds, Fetched: 3 row(s)
hive> Select * From top_10_royalty_songs;
OK
top_10_royalty_songs.song_id      top_10_royalty_songs.duration      top_10_royalty_songs.batchid
S288    22627294      1
S287    28000000      1
S286    19000000      1
S289    15254588      1
S280    99000000      1
S284    2604333      1
S282    100000      1
S285    0          1
Time taken: 0.237 seconds, Fetched: 8 row(s)
hive> Select * From top_10_stations;
OK
top_10_stations.station_id      top_10_stations.total_distinct_songs_played      top_10_stations.distinct_user_count      top_10_stations.batchid
ST487    2          3          1
ST414    1          1          1
ST411    1          1          1
ST402    1          2          1
ST406    1          1          1
ST405    1          1          1
Time taken: 0.336 seconds, Fetched: 6 row(s)
hive> Select * From top_10_unsubscribed_users;
OK
top_10_unsubscribed_users.user_id      top_10_unsubscribed_users.duration      top_10_unsubscribed_users.batchid
U117    20000000      1
U118    20000000      1
U110    20000000      1
U120    12627294      1
U115    12527294      1
```

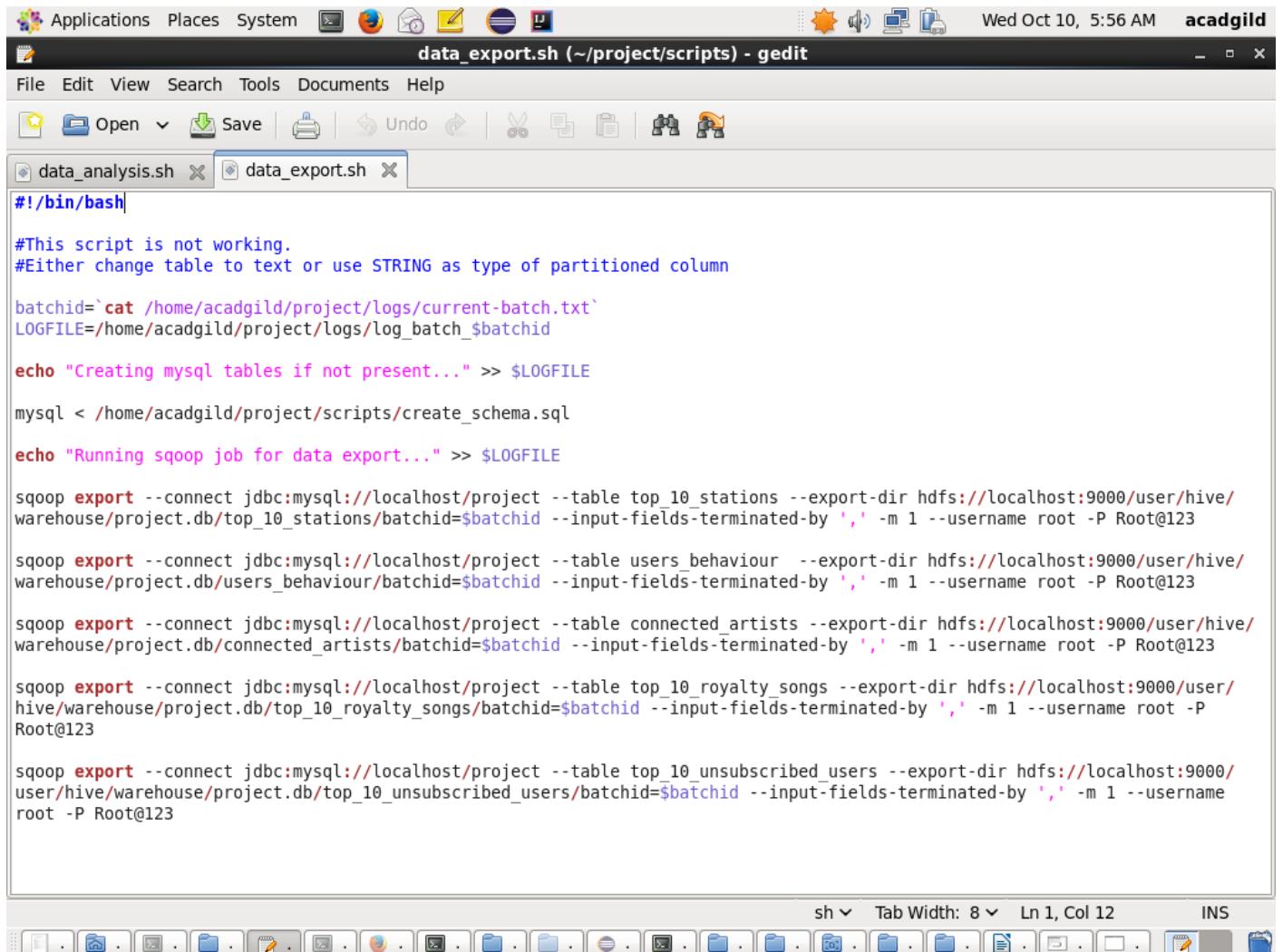
## Output from top\_10\_unsubscribed\_users, usersBehaviour.

```
Time taken: 0.237 seconds, Fetched: 8 row(s)
hive> Select * From top_10_stations;
OK
top_10_stations.station_id      top_10_stations.total_distinct_songs_played      top_10_stations.distinct_user_count      top_10_stations.batchid
ST407    2          3          1
ST414    1          1          1
ST411    1          1          1
ST482    1          2          1
ST406    1          1          1
ST405    1          1          1
Time taken: 0.336 seconds, Fetched: 6 row(s)
hive> Select * From top_10_unsubscribed_users;
OK
top_10_unsubscribed_users.user_id      top_10_unsubscribed_users.duration      top_10_unsubscribed_users.batchid
U117    20000000          1
U118    20000000          1
U110    20000000          1
U120    12527294          1
U115    12527294          1
U187    10000000          1
U188    5231627           1
U189    2604333           1
U186    2604333           1
U100    0                 1
Time taken: 0.275 seconds, Fetched: 10 row(s)
hive> Select * From usersBehaviour;
OK
users behaviour.user_type      users behaviour.duration      users behaviour.batchid
SUBSCRIBED    93861594          1
UNSUBSCRIBED  105594881         1
Time taken: 0.274 seconds, Fetched: 2 row(s)
hive>
```

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**.



```
#!/bin/bash

#This script is not working.
#Either change table to text or use STRING as type of partitioned column

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

echo "Creating mysql tables if not present..." >> $LOGFILE

mysql < /home/acadgild/project/scripts/create_schema.sql

echo "Running sqoop job for data export..." >> $LOGFILE

sqoop export --connect jdbc:mysql://localhost/project --table top_10_stations --export-dir hdfs://localhost:9000/user/hive/warehouse/project.db/top_10_stations/batchid=$batchid --input-fields-terminated-by ',' -m 1 --username root -P Root@123

sqoop export --connect jdbc:mysql://localhost/project --table users behaviour --export-dir hdfs://localhost:9000/user/hive/warehouse/project.db/users_behaviour/batchid=$batchid --input-fields-terminated-by ',' -m 1 --username root -P Root@123

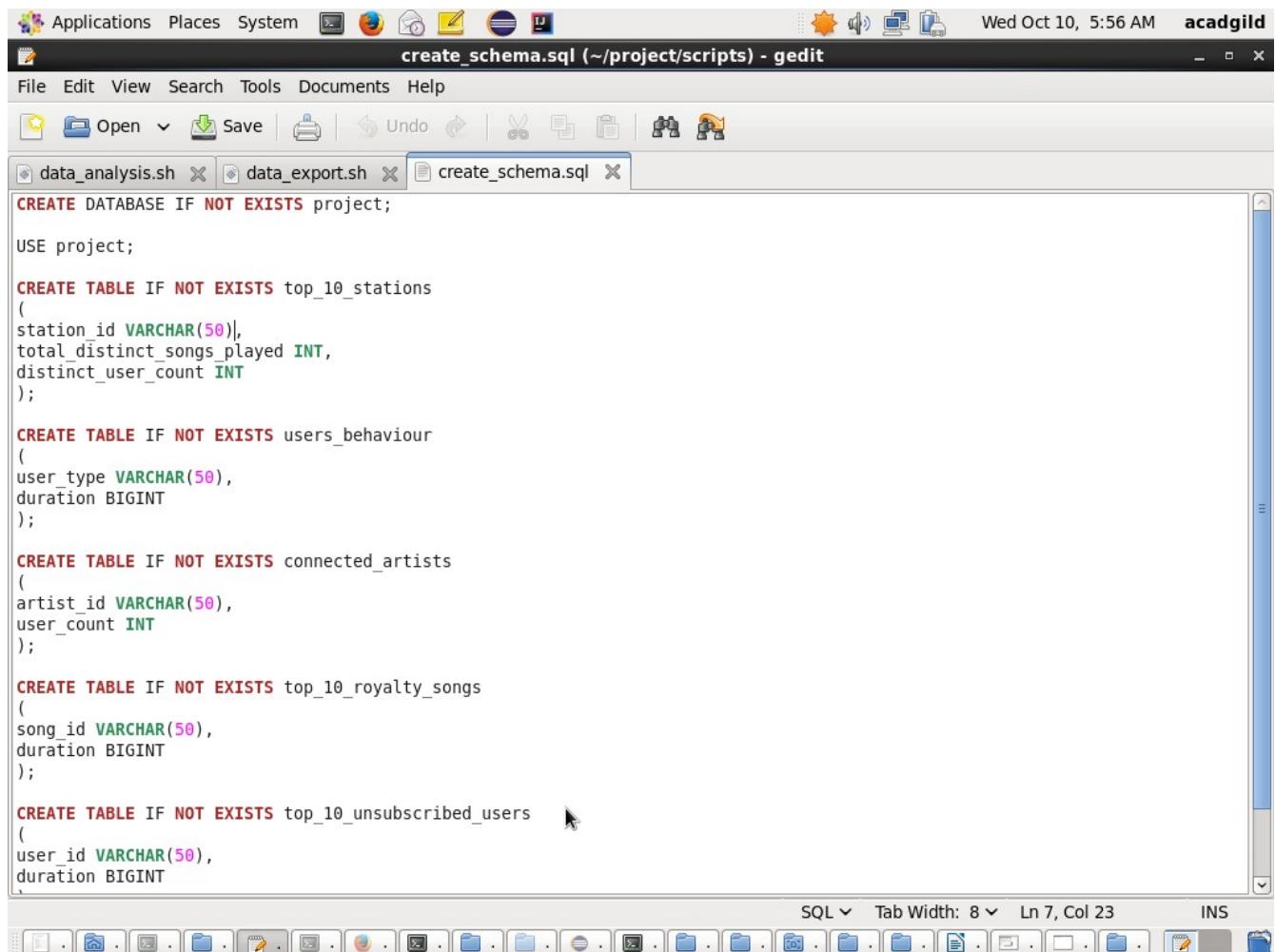
sqoop export --connect jdbc:mysql://localhost/project --table connected_artists --export-dir hdfs://localhost:9000/user/hive/warehouse/project.db/connected_artists/batchid=$batchid --input-fields-terminated-by ',' -m 1 --username root -P Root@123

sqoop export --connect jdbc:mysql://localhost/project --table top_10_royalty_songs --export-dir hdfs://localhost:9000/user/hive/warehouse/project.db/top_10_royalty_songs/batchid=$batchid --input-fields-terminated-by ',' -m 1 --username root -P Root@123

sqoop export --connect jdbc:mysql://localhost/project --table top_10_unsubscribed_users --export-dir hdfs://localhost:9000/user/hive/warehouse/project.db/top_10_unsubscribed_users/batchid=$batchid --input-fields-terminated-by ',' -m 1 --username root -P Root@123
```

### **create\_schema.sql :**

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



The screenshot shows a Gnome desktop environment with a window titled "create\_schema.sql (~/project/scripts) - gedit". The window contains a SQL script for creating a database and several tables. The script includes creating a database named "project", selecting it, and then creating five tables: "top\_10\_stations", "usersBehaviour", "connectedArtists", "top\_10\_royalty\_songs", and "top\_10\_unsubscribed\_users". Each table has specific column definitions like VARCHAR(50) or INT.

```
CREATE DATABASE IF NOT EXISTS project;
USE project;

CREATE TABLE IF NOT EXISTS top_10_stations
(
station_id VARCHAR(50),
total_distinct_songs_played INT,
distinct_user_count INT
);

CREATE TABLE IF NOT EXISTS usersBehaviour
(
user_type VARCHAR(50),
duration BIGINT
);

CREATE TABLE IF NOT EXISTS connectedArtists
(
artist_id VARCHAR(50),
user_count INT
);

CREATE TABLE IF NOT EXISTS top_10_royalty_songs
(
song_id VARCHAR(50),
duration BIGINT
);

CREATE TABLE IF NOT EXISTS top_10_unsubscribed_users
(
user_id VARCHAR(50),
duration BIGINT
```

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.

```
mysql>
mysql> use project;
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> Select * From top_10_stations;
+-----+-----+-----+
| station_id | total_distinct_songs_played | distinct_user_count |
+-----+-----+-----+
| ST407      | 2                  | 3                |
| ST414      | 1                  | 1                |
| ST411      | 1                  | 1                |
| ST402      | 1                  | 2                |
| ST406      | 1                  | 1                |
| ST405      | 1                  | 1                |
+-----+-----+-----+
6 rows in set (0.00 sec)

mysql> Select * From connected_artists;
+-----+-----+
| artist_id | user_count |
+-----+-----+
| A303      | 2          |
| A302      | 2          |
| A300      | 1          |
+-----+-----+
3 rows in set (0.00 sec)
```

**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)
```

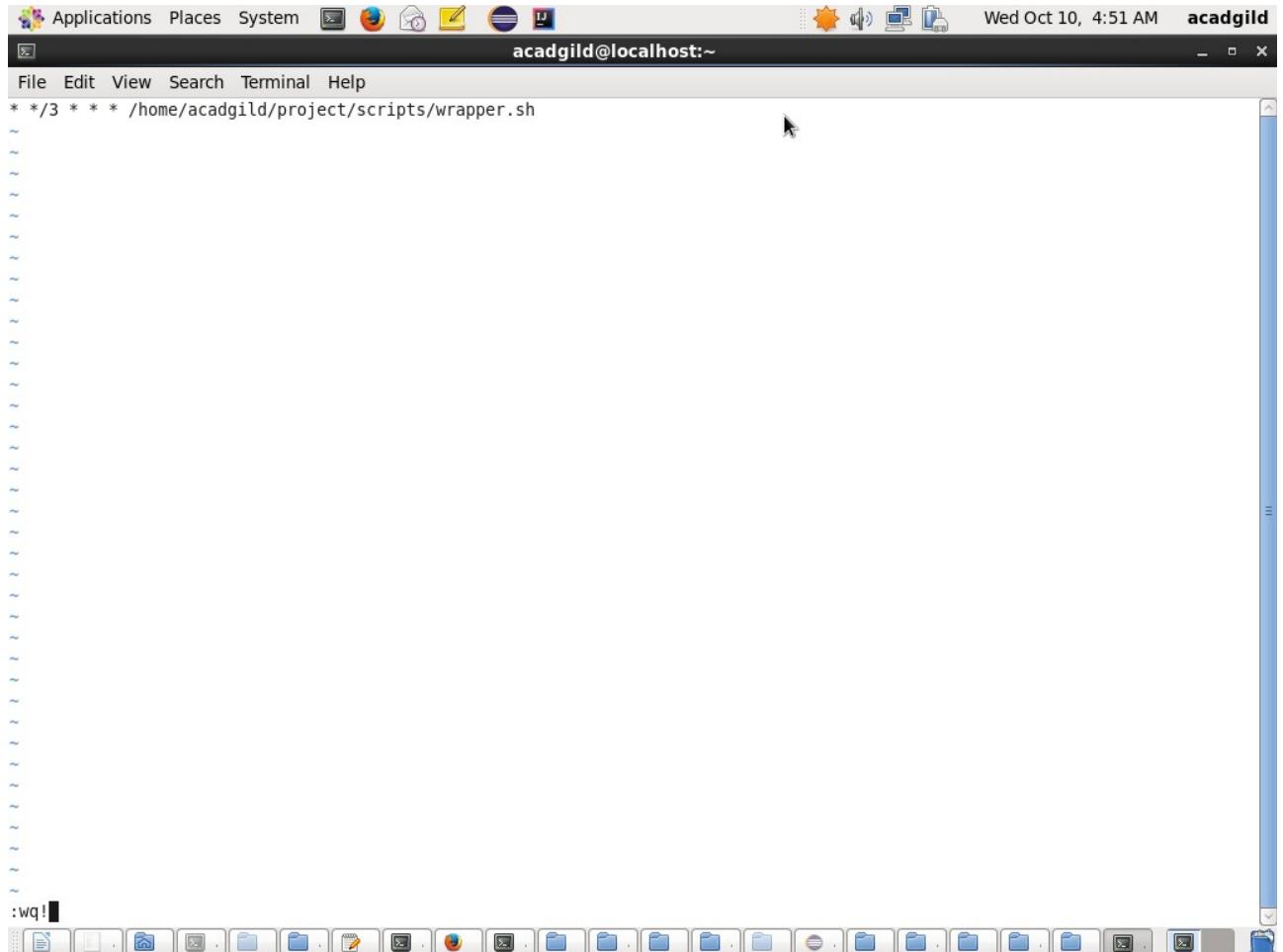
**Output from top\_10\_unsubscribed\_users and usersBehaviour :**

```
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 usersBehaviour;
+-----+-----+
| 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.



The screenshot shows a Linux desktop environment with a terminal window open. The terminal window title is "acadgild@localhost:~". The command entered is "\* \* /3 \* \* \* /home/acadgild/project/scripts/wrapper.sh". The terminal window displays numerous tilde (~) characters, indicating many lines of output have been omitted. The bottom of the terminal window shows the command ":wq!". Below the terminal window is a docked application bar with icons for various desktop applications like a file manager, browser, and system tools. The desktop background is a light blue color.

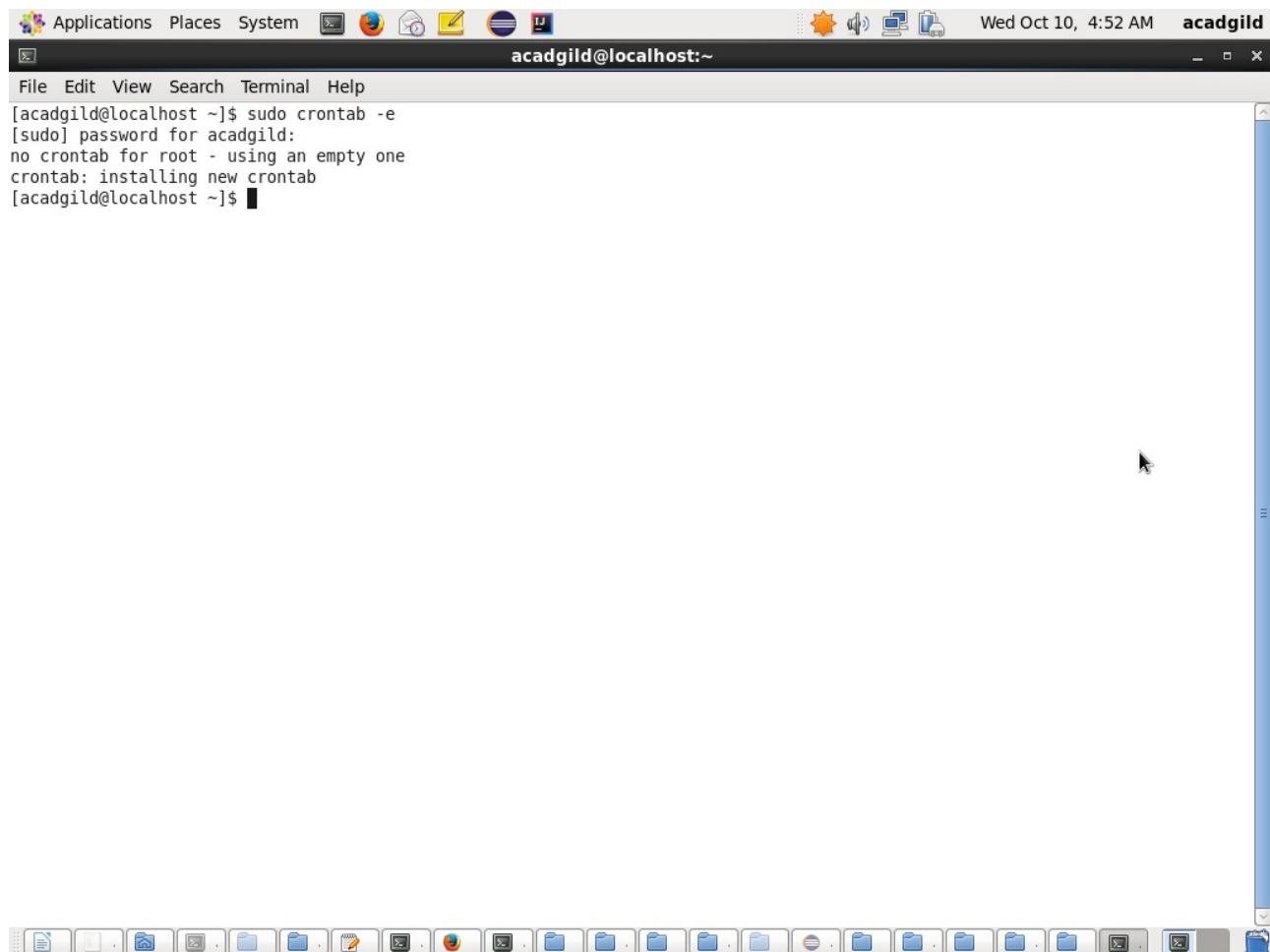
Applications Places System  Wed Oct 10, 7:39 AM acadgild

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...
```



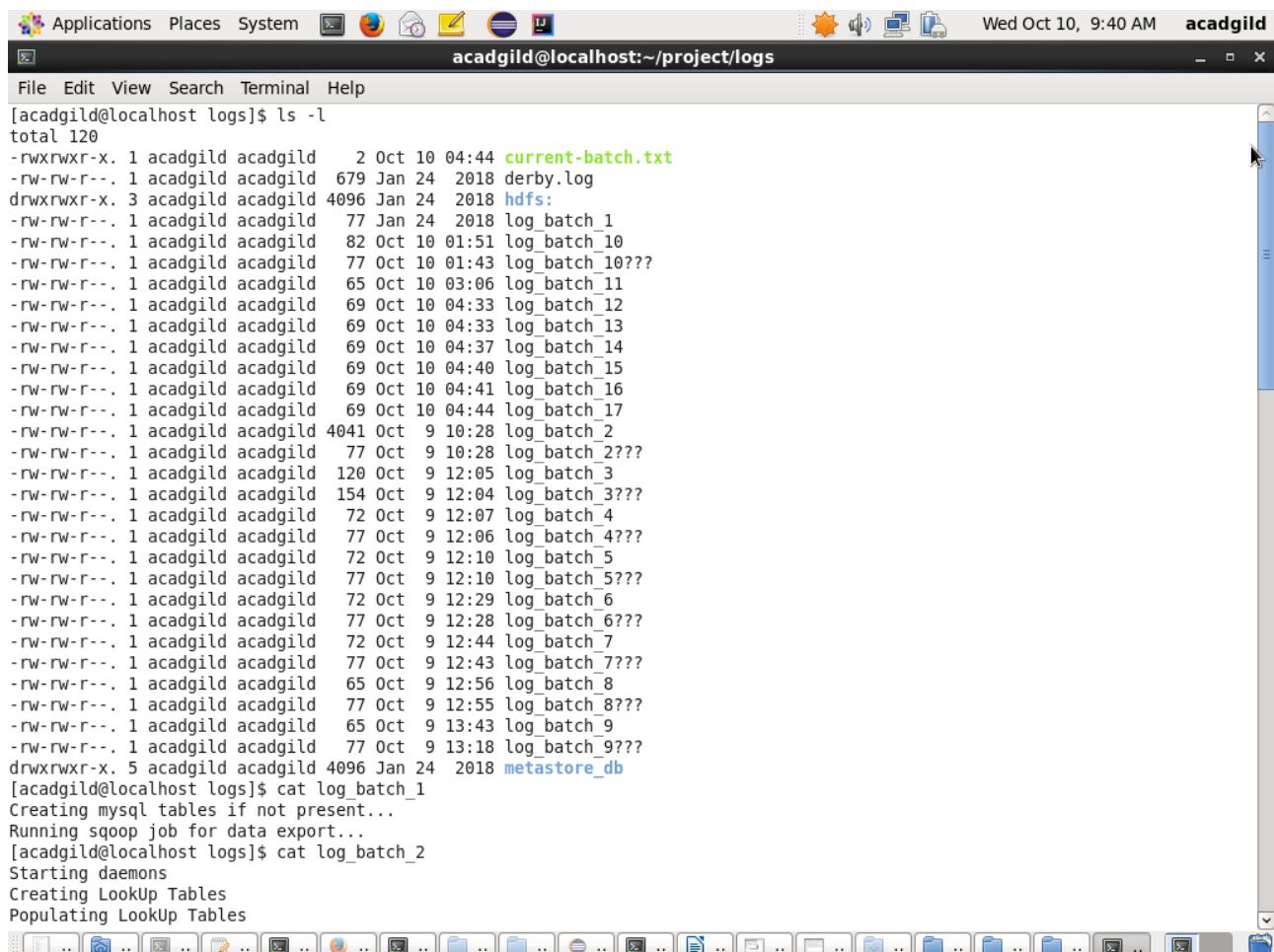
We can check logs to track the behavior of the operations we have done on the data and overcome failures in the pipeline and we can see the **batchid** incremented value in **current-batch.txt**



A screenshot of a Linux desktop environment. At the top, there is a panel with icons for Applications, Places, System, and various system status indicators. The date and time are shown as "Wed Oct 10, 4:52 AM". The terminal window title is "acadgild@localhost:~". The terminal content shows the command "sudo crontab -e" being run, followed by a password prompt, and the message "no crontab for root - using an empty one" and "crontab: installing new crontab". The desktop background is a light blue gradient.

```
[acadgild@localhost ~]$ sudo crontab -e
[sudo] password for acadgild:
no crontab for root - using an empty one
crontab: installing new crontab
[acadgild@localhost ~]$
```

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

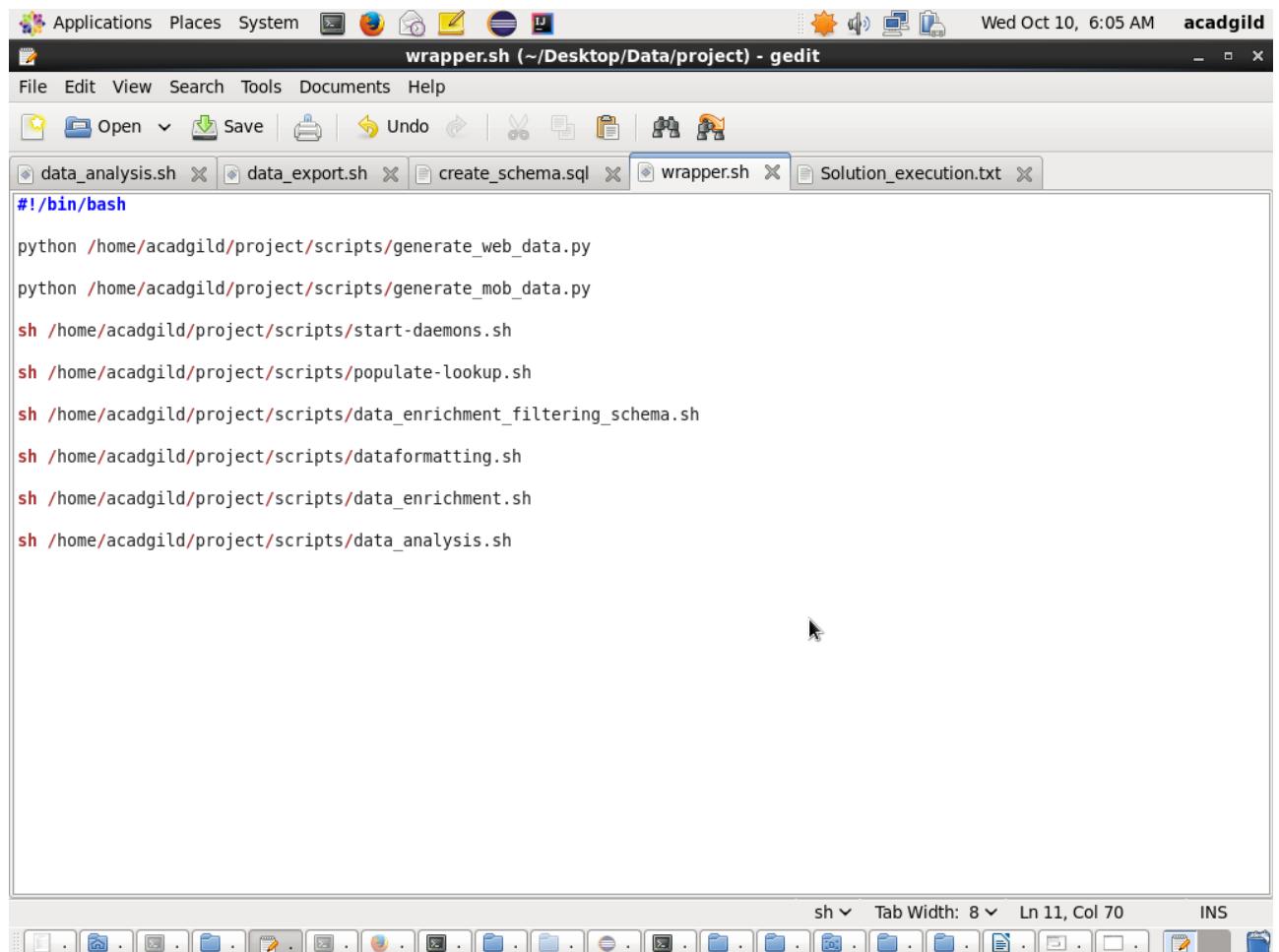


A screenshot of a Linux desktop environment showing a terminal window. The terminal title is "acadgild@localhost:~/project/logs". The window contains a command-line session output:

```
[acadgild@localhost logs]$ ls -l
total 120
-rwxrwxr-x. 1 acadgild acadgild 2 Oct 10 04:44 current-batch.txt
-rw-rw-r--. 1 acadgild acadgild 679 Jan 24 2018 derby.log
drwxrwxr-x. 3 acadgild acadgild 4096 Jan 24 2018 hdfs:
-rw-rw-r--. 1 acadgild acadgild 77 Jan 24 2018 log_batch_1
-rw-rw-r--. 1 acadgild acadgild 82 Oct 10 01:51 log_batch_10
-rw-rw-r--. 1 acadgild acadgild 77 Oct 10 01:43 log_batch_10???
-rw-rw-r--. 1 acadgild acadgild 65 Oct 10 03:06 log_batch_11
-rw-rw-r--. 1 acadgild acadgild 69 Oct 10 04:33 log_batch_12
-rw-rw-r--. 1 acadgild acadgild 69 Oct 10 04:33 log_batch_13
-rw-rw-r--. 1 acadgild acadgild 69 Oct 10 04:37 log_batch_14
-rw-rw-r--. 1 acadgild acadgild 69 Oct 10 04:40 log_batch_15
-rw-rw-r--. 1 acadgild acadgild 69 Oct 10 04:41 log_batch_16
-rw-rw-r--. 1 acadgild acadgild 69 Oct 10 04:44 log_batch_17
-rw-rw-r--. 1 acadgild acadgild 4041 Oct 9 10:28 log_batch_2
-rw-rw-r--. 1 acadgild acadgild 77 Oct 9 10:28 log_batch_2???
-rw-rw-r--. 1 acadgild acadgild 120 Oct 9 12:05 log_batch_3
-rw-rw-r--. 1 acadgild acadgild 154 Oct 9 12:04 log_batch_3???
-rw-rw-r--. 1 acadgild acadgild 72 Oct 9 12:07 log_batch_4
-rw-rw-r--. 1 acadgild acadgild 77 Oct 9 12:06 log_batch_4???
-rw-rw-r--. 1 acadgild acadgild 72 Oct 9 12:10 log_batch_5
-rw-rw-r--. 1 acadgild acadgild 77 Oct 9 12:10 log_batch_5???
-rw-rw-r--. 1 acadgild acadgild 72 Oct 9 12:29 log_batch_6
-rw-rw-r--. 1 acadgild acadgild 77 Oct 9 12:28 log_batch_6???
-rw-rw-r--. 1 acadgild acadgild 72 Oct 9 12:44 log_batch_7
-rw-rw-r--. 1 acadgild acadgild 77 Oct 9 12:43 log_batch_7???
-rw-rw-r--. 1 acadgild acadgild 65 Oct 9 12:56 log_batch_8
-rw-rw-r--. 1 acadgild acadgild 77 Oct 9 12:55 log_batch_8???
-rw-rw-r--. 1 acadgild acadgild 65 Oct 9 13:43 log_batch_9
-rw-rw-r--. 1 acadgild acadgild 77 Oct 9 13:18 log_batch_9???
drwxrwxr-x. 5 acadgild acadgild 4096 Jan 24 2018 metastore_db
[acadgild@localhost logs]$ cat log_batch_1
Creating mysql tables if not present...
Running sqoop job for data export...
[acadgild@localhost logs]$ cat log_batch_2
Starting daemons
Creating LookUp Tables
Populating LookUp Tables
```

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 screenshot shows a Gedit text editor window titled "wrapper.sh (~/Desktop/Data/project) - gedit". The window contains the following code:

```
#!/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/populate-lookup.sh
sh /home/acadgild/project/scripts/data_enrichment_filtering_schema.sh
sh /home/acadgild/project/scripts/dataformatting.sh
sh /home/acadgild/project/scripts/data_enrichment.sh
sh /home/acadgild/project/scripts/data_analysis.sh
```

The status bar at the bottom of the editor shows "sh" and "INS".

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 screenshot shows a terminal window titled "acadgild@localhost:~". The window contains the following text:

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

The terminal window has a standard Linux desktop interface with icons for Applications, Places, System, and various system status indicators at the top. The bottom of the window features a toolbar with icons for file operations like Open, Save, Print, and others. The title bar also displays the user's name and the date and time.

The **crontab job scheduler** will run the **wrapper.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.