Project

Music Data Analysis

Data Ingestion and Initial Validation

- **Generate_mob_data**

Python script is used to generate random data

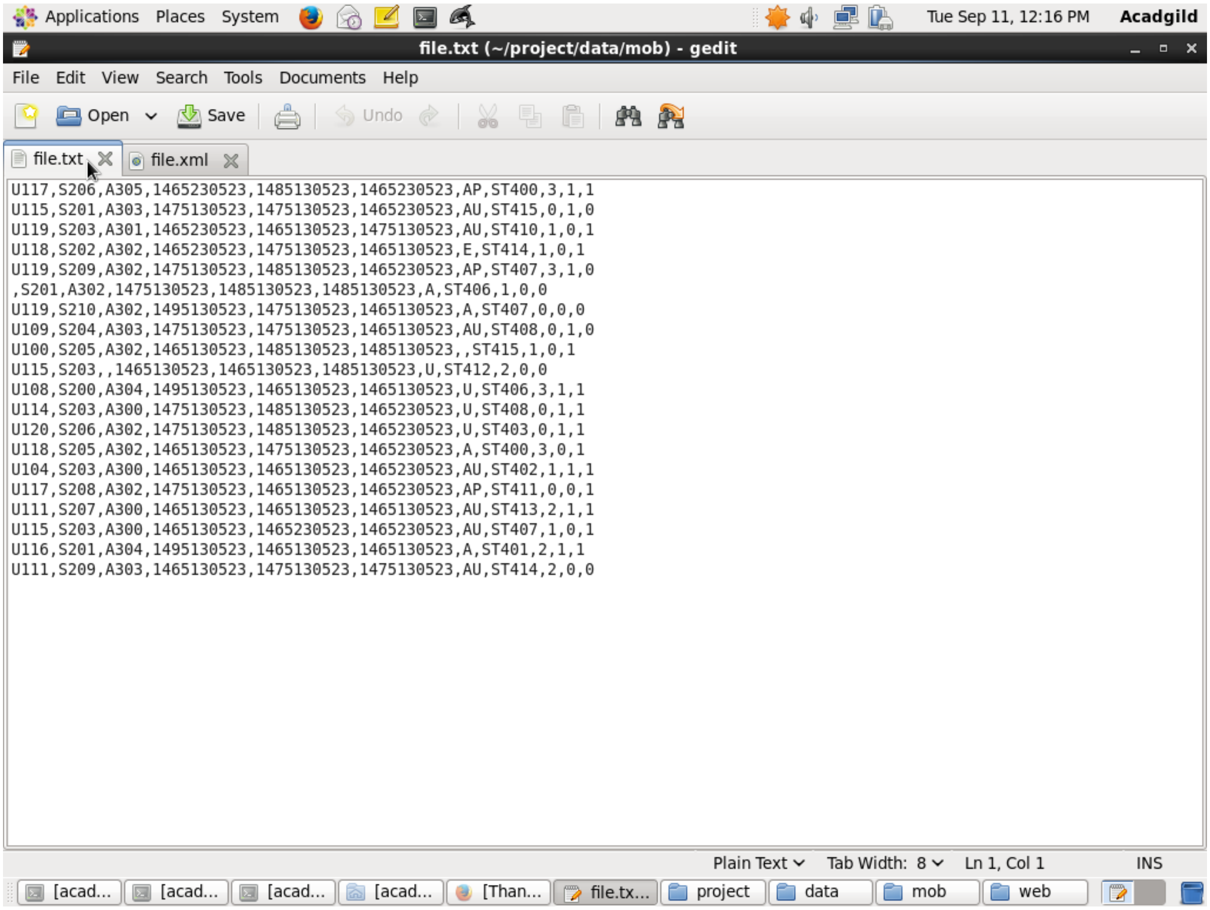
Command:

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

Code: (with comments)

generate_mob_data.py

Screen:



```
U117,S206,A305,1465230523,1485130523,1465230523,AP,ST400,3,1,1
U115,S201,A303,1475130523,1475130523,1465230523,AU,ST415,0,1,0
U119,S203,A301,1465230523,1465130523,1475130523,AU,ST410,1,0,1
U118,S202,A302,1465230523,1475130523,1465130523,E,ST414,1,0,1
U119,S209,A302,1475130523,1485130523,1465230523,AP,ST407,3,1,0
,S201,A302,1475130523,1485130523,1485130523,A,ST406,1,0,0
U119,S210,A302,1495130523,1475130523,1465130523,A,ST407,0,0,0
U109,S204,A303,1475130523,1475130523,1465130523,AU,ST408,0,1,0
U100,S205,A302,1465130523,1485130523,1485130523,,ST415,1,0,1
U115,S203,,1465130523,1465130523,1485130523,U,ST412,2,0,0
U108,S200,A304,1495130523,1465130523,1465130523,U,ST406,3,1,1
U114,S203,A300,1475130523,1485130523,1465230523,U,ST408,0,1,1
U120,S206,A302,1475130523,1485130523,1465230523,U,ST403,0,1,1
U118,S205,A302,1465130523,1475130523,1465230523,A,ST400,3,0,1
U104,S203,A300,1465130523,1465130523,1465230523,AU,ST402,1,1,1
U117,S208,A302,1475130523,1465130523,1465230523,AP,ST411,0,0,1
U111,S207,A300,1465130523,1465130523,1465130523,AU,ST413,2,1,1
U115,S203,A300,1465130523,1465230523,1465230523,AU,ST407,1,0,1
U116,S201,A304,1495130523,1465130523,1465130523,A,ST401,2,1,1
U111,S209,A303,1465130523,1475130523,1475130523,AU,ST414,2,0,0
```

- **Generate_web_data**

Python script is used to generate random data

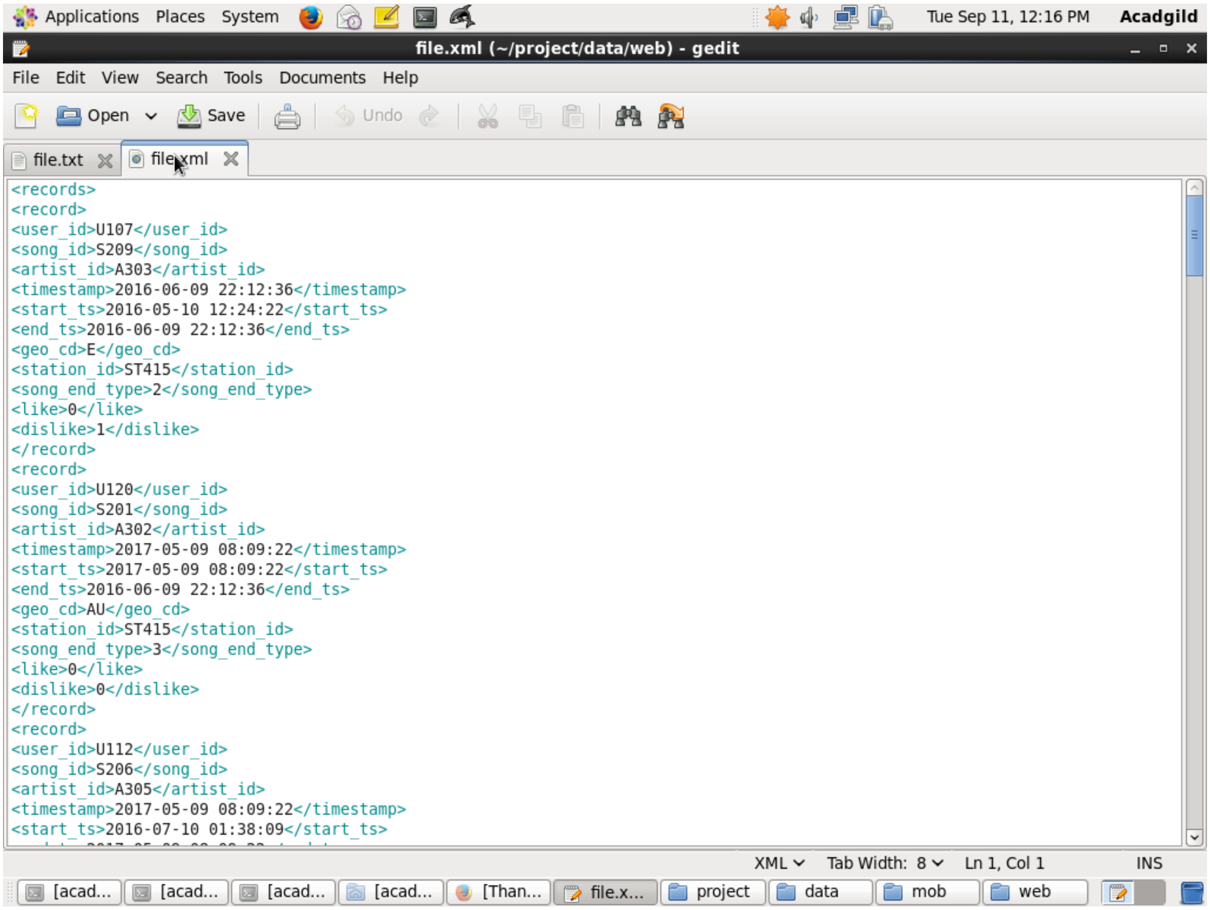
Command:

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

Code:

generate_web_data.py

Screen:



```
<records>
<record>
<user_id>U107</user_id>
<song_id>S209</song_id>
<artist_id>A303</artist_id>
<timestamp>2016-06-09 22:12:36</timestamp>
<start_ts>2016-05-10 12:24:22</start_ts>
<end_ts>2016-06-09 22:12:36</end_ts>
<geo_cd>E</geo_cd>
<station_id>ST415</station_id>
<song_end_type>2</song_end_type>
<like>0</like>
<dislike>1</dislike>
</record>
<record>
<user_id>U120</user_id>
<song_id>S201</song_id>
<artist_id>A302</artist_id>
<timestamp>2017-05-09 08:09:22</timestamp>
<start_ts>2017-05-09 08:09:22</start_ts>
<end_ts>2016-06-09 22:12:36</end_ts>
<geo_cd>AU</geo_cd>
<station_id>ST415</station_id>
<song_end_type>3</song_end_type>
<like>0</like>
<dislike>0</dislike>
</record>
<record>
<user_id>U112</user_id>
<song_id>S206</song_id>
<artist_id>A305</artist_id>
<timestamp>2017-05-09 08:09:22</timestamp>
<start_ts>2016-07-10 01:38:09</start_ts>
```

- **Starting Hadoop**

Command:

```
sh /home/acadgild/project/scripts/start-daemons.sh
```

Code:

start-daemons.sh

- **Populate-Lookup**

Command:

```
sh /home/acadgild/project/scripts/populate_lookup.sh
```

Code:

populate-lookup.sh user-artist.hql

Screen:

- **Data Formating:**

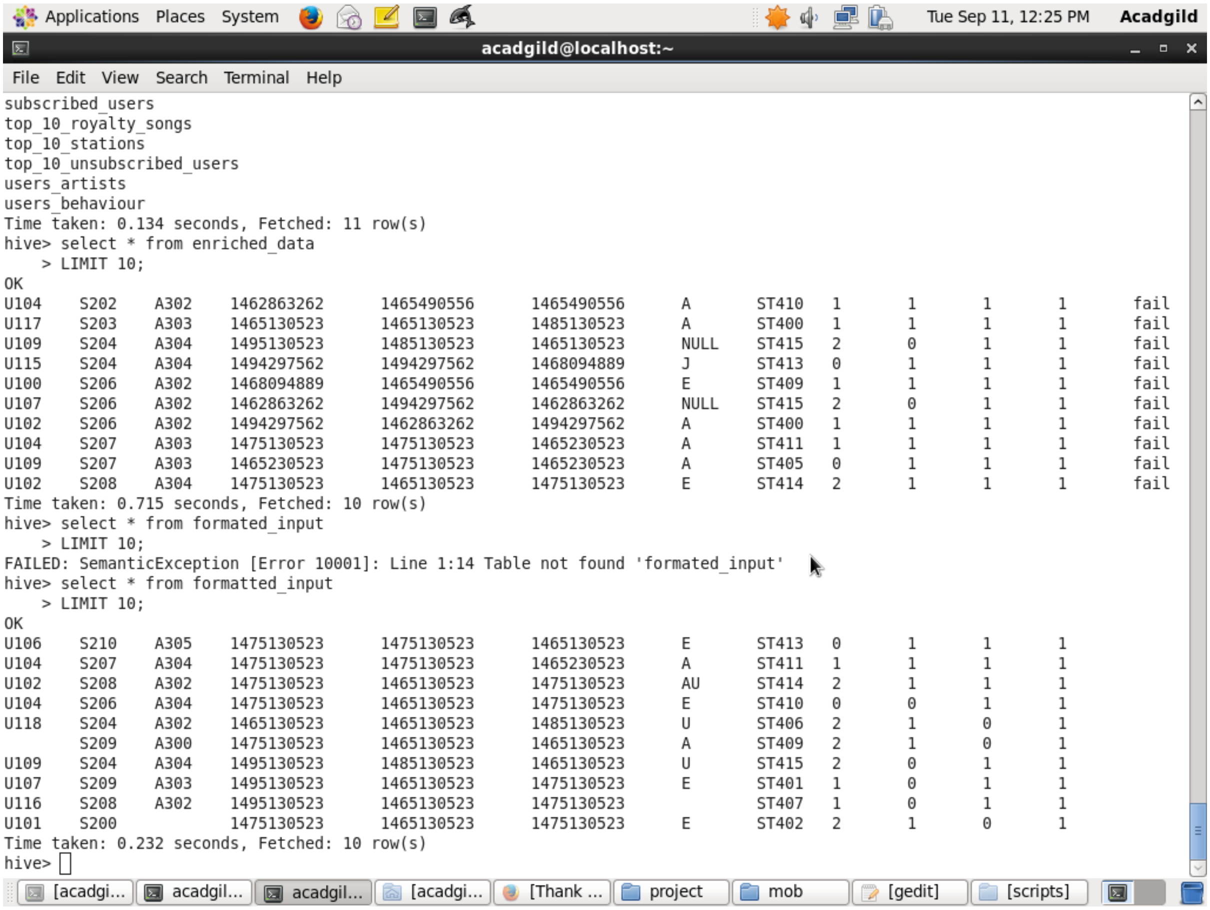
Command:

```
python /home/acadgild/project/scripts/data_formating.sh
```

Code:

dataformatting.sh dataformatting.pig formatted_hive_load.hql

Screen:


Applications Places System
Tue Sep 11, 12:25 PM
Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgi... acadgil... acadgil... [acadgi... [Thank ... project mob [gedit] [scripts]

```
subscribed_users
top_10_royalty_songs
top_10_stations
top_10_unsubscribed_users
users_artists
users_behaviour
Time taken: 0.134 seconds, Fetched: 11 row(s)
hive> select * from enriched_data
    > LIMIT 10;
OK
U104    S202    A302    1462863262      1465490556      1465490556      A       ST410   1       1       1       1       fail
U117    S203    A303    1465130523      1465130523      1485130523      A       ST400   1       1       1       1       fail
U109    S204    A304    1495130523      1485130523      1465130523      NULL    ST415   2       0       1       1       fail
U115    S204    A304    1494297562      1494297562      1468094889      J       ST413   0       1       1       1       fail
U100    S206    A302    1468094889      1465490556      1465490556      E       ST409   1       1       1       1       fail
U107    S206    A302    1462863262      1494297562      1462863262      NULL    ST415   2       0       1       1       fail
U102    S206    A302    1494297562      1462863262      1494297562      A       ST400   1       1       1       1       fail
U104    S207    A303    1475130523      1475130523      1465230523      A       ST411   1       1       1       1       fail
U109    S207    A303    1465230523      1475130523      1465230523      A       ST405   0       1       1       1       fail
U102    S208    A304    1475130523      1465130523      1475130523      E       ST414   2       1       1       1       fail
Time taken: 0.715 seconds, Fetched: 10 row(s)
hive> select * from formated_input
    > LIMIT 10;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'formated_input'
hive> select * from formatted_input
    > LIMIT 10;
OK
U106    S210    A305    1475130523      1475130523      1465130523      E       ST413   0       1       1       1
U104    S207    A304    1475130523      1475130523      1465230523      A       ST411   1       1       1       1
U102    S208    A302    1475130523      1465130523      1475130523      AU      ST414   2       1       1       1
U104    S206    A304    1475130523      1465130523      1475130523      E       ST410   0       0       1       1
U118    S204    A302    1465130523      1465130523      1485130523      U       ST406   2       1       0       1
        S209    A300    1475130523      1465130523      1465130523      A       ST409   2       1       0       1
U109    S204    A304    1495130523      1485130523      1465130523      U       ST415   2       0       1       1
U107    S209    A303    1495130523      1465130523      1475130523      E       ST401   1       0       1       1
U116    S208    A302    1495130523      1465130523      1475130523              ST407   1       0       1       1
U101    S200            1475130523      1465130523      1475130523      E       ST402   2       1       0       1
Time taken: 0.232 seconds, Fetched: 10 row(s)
hive> 
```

- **Hive lookup:**

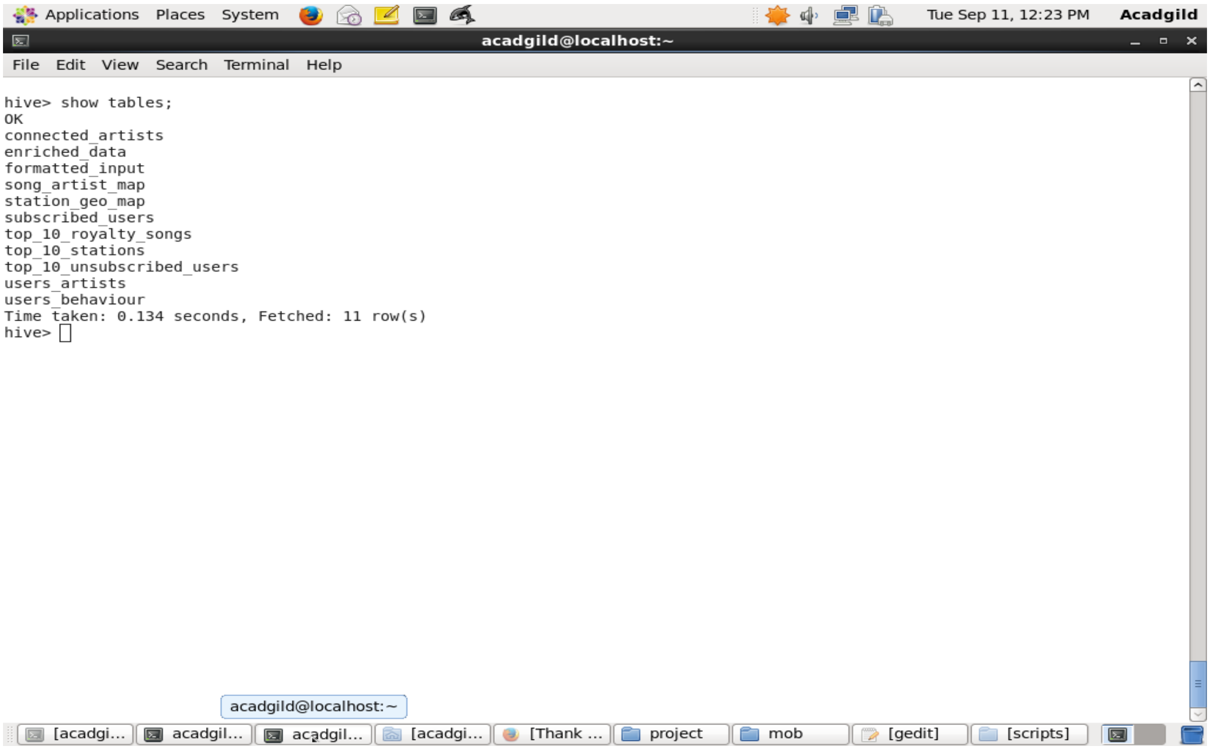
Command:

Hive -f /home/acadgild/project/scripts/create_hive_hbase_lookup.hql

Code:

create_hive_hbase_lookup.hql

Screen:



Applications Places System
Tue Sep 11, 12:23 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
acadgild@localhost:~
[acadgi… acadgil… acadgil… [acadgi… [Thank … project mob [gedit] [scripts]

```
hive> show tables;
OK
connected_artists
enriched_data
formatted_input
song_artist_map
station_geo_map
subscribed_users
top_10_royalty_songs
top_10_stations
top_10_unsubscribed_users
users_artists
users_behaviour
Time taken: 0.134 seconds, Fetched: 11 row(s)
hive>
```

Data Enrichment

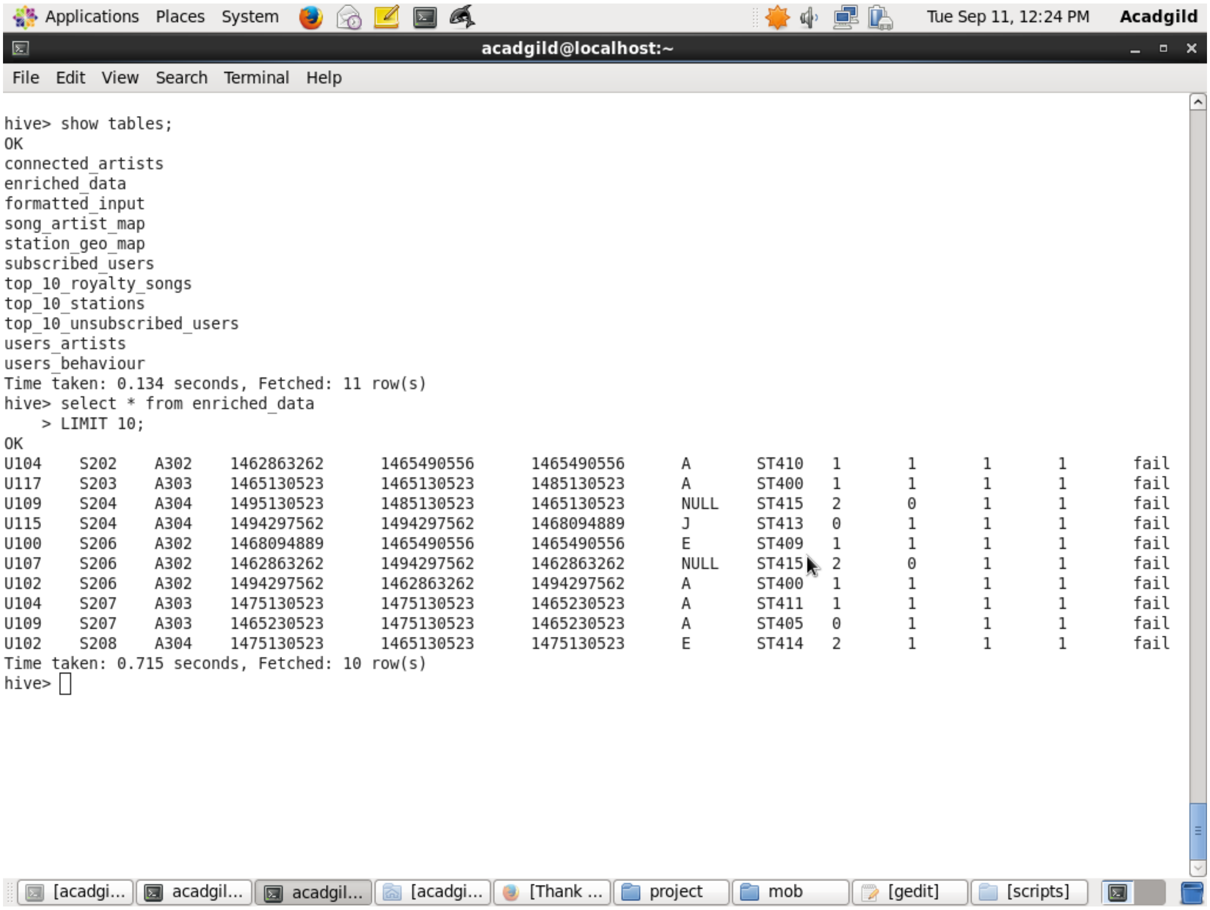
Command:

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

Code:

data_enrichment.sh data_enrichment.hql

Screen:



```
hive> show tables;
OK
connected_artists
enriched_data
formatted_input
song_artist_map
station_geo_map
subscribed_users
top_10_royalty_songs
top_10_stations
top_10_unsubscribed_users
users_artists
users_behaviour
Time taken: 0.134 seconds, Fetched: 11 row(s)
hive> select * from enriched_data
    > LIMIT 10;
OK
U104    S202    A302    1462863262      1465490556      1465490556      A       ST410   1       1       1       1       fail
U117    S203    A303    1465130523      1465130523      1485130523      A       ST400   1       1       1       1       fail
U109    S204    A304    1495130523      1485130523      1465130523      NULL    ST415   2       0       1       1       fail
U115    S204    A304    1494297562      1494297562      1468094889      J       ST413   0       1       1       1       fail
U100    S206    A302    1468094889      1465490556      1465490556      E       ST409   1       1       1       1       fail
U107    S206    A302    1462863262      1494297562      1462863262      NULL    ST415   2       0       1       1       fail
U102    S206    A302    1494297562      1462863262      1494297562      A       ST400   1       1       1       1       fail
U104    S207    A303    1475130523      1475130523      1465230523      A       ST411   1       1       1       1       fail
U109    S207    A303    1465230523      1475130523      1465230523      A       ST405   0       1       1       1       fail
U102    S208    A304    1475130523      1465130523      1475130523      E       ST414   2       1       1       1       fail
Time taken: 0.715 seconds, Fetched: 10 row(s)
hive> 
```

Data Analysis

1. Determine top 10 station_id(s) where maximum number of songs were played, which were liked by unique users.
2. Determine total duration of songs played by each type of user, where type of user can be **'subscribed'** or **'unsubscribed'**. An unsubscribed user is the one whose record is either not present in **Subscribed_users** lookup table or has *subscription_end_date* earlier than the *timestamp* of the song played by him.
3. Determine top 10 connected artists. Connected artists are those whose songs are most listened by the unique users who follow them.
4. Determine top 10 songs who have generated the maximum revenue. Royalty applies to a song only if it was *liked* or was *completed successfully* or both.

Command:

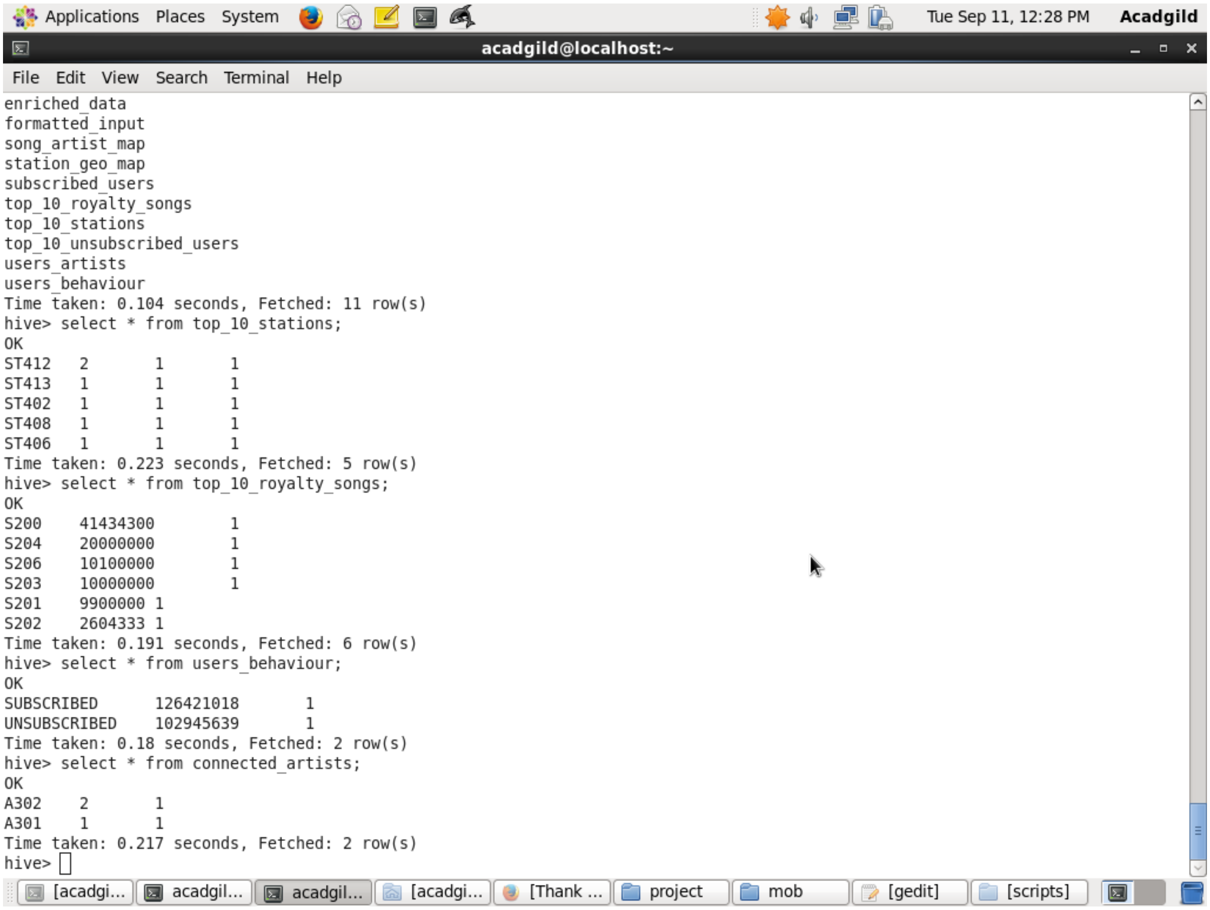
sh /home/acadgild/project/scripts/data_analysis.sh

Code:

data_analysis.sh   data_analysis.hql   data_export.sh   create_schema.sql

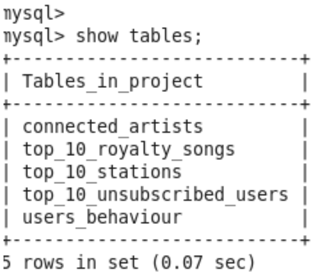
Screen:

Hive:


Applications Places System
Tue Sep 11, 12:28 PM
Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgi... acadgil... acadgil... [acadgi... [Thank ... project mob [gedit] [scripts]

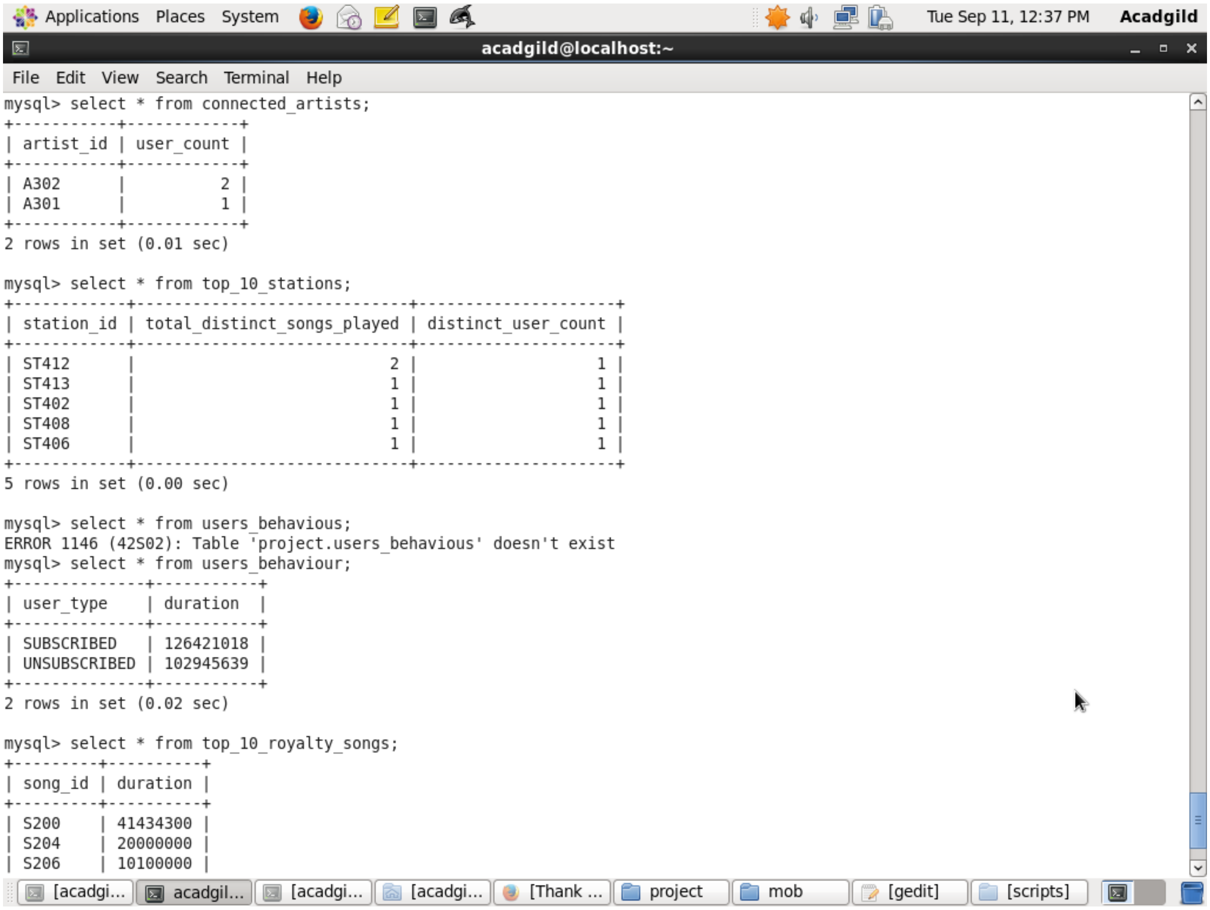
```
enriched_data
formatted_input
song_artist_map
station_geo_map
subscribed_users
top_10_royalty_songs
top_10_stations
top_10_unsubscribed_users
users_artists
users_behaviour
Time taken: 0.104 seconds, Fetched: 11 row(s)
hive> select * from top_10_stations;
OK
ST412   2       1       1
ST413   1       1       1
ST402   1       1       1
ST408   1       1       1
ST406   1       1       1
Time taken: 0.223 seconds, Fetched: 5 row(s)
hive> select * from top_10_royalty_songs;
OK
S200    41434300        1
S204    20000000        1
S206    10100000        1
S203    10000000        1
S201    9900000 1
S202    2604333 1
Time taken: 0.191 seconds, Fetched: 6 row(s)
hive> select * from users_behaviour;
OK
SUBSCRIBED      126421018       1
UNSUBSCRIBED    102945639       1
Time taken: 0.18 seconds, Fetched: 2 row(s)
hive> select * from connected_artists;
OK
A302    2       1
A301    1       1
Time taken: 0.217 seconds, Fetched: 2 row(s)
hive> 
```

Mysql create schema:



```
mysql>
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 (0.07 sec)
```

Mysql Analysis:

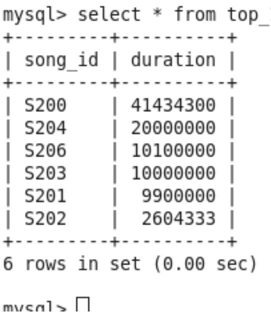

Applications Places System
Tue Sep 11, 12:37 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgi... acadgil... [acadgi... [acadgi... [Thank ... project mob [gedit] [scripts]

```
mysql> select * from connected_artists;
+-----------+------------+
| artist_id | user_count |
+-----------+------------+
| A302      |          2 |
| A301      |          1 |
+-----------+------------+
2 rows in set (0.01 sec)

mysql> select * from top_10_stations;
+------------+-----------------------------+---------------------+
| station_id | total_distinct_songs_played | distinct_user_count |
+------------+-----------------------------+---------------------+
| ST412      |                           2 |                   1 |
| ST413      |                           1 |                   1 |
| ST402      |                           1 |                   1 |
| ST408      |                           1 |                   1 |
| ST406      |                           1 |                   1 |
+------------+-----------------------------+---------------------+
5 rows in set (0.00 sec)

mysql> select * from users_behavious;
ERROR 1146 (42S02): Table 'project.users_behavious' doesn't exist
mysql> select * from users_behaviour;
+--------------+-----------+
| user_type    | duration  |
+--------------+-----------+
| SUBSCRIBED   | 126421018 |
| UNSUBSCRIBED | 102945639 |
+--------------+-----------+
2 rows in set (0.02 sec)

mysql> select * from top_10_royalty_songs;
+---------+----------+
| song_id | duration |
+---------+----------+
| S200    | 41434300 |
| S204    | 20000000 |
| S206    | 10100000 |
```



```
mysql> select * from top_10_royalty_songs;
+---------+----------+
| song_id | duration |
+---------+----------+
| S200    | 41434300 |
| S204    | 20000000 |
| S206    | 10100000 |
| S203    | 10000000 |
| S201    |  9900000 |
| S202    |  2604333 |
+---------+----------+
6 rows in set (0.00 sec)

mysql> 
```

Job scheduling:

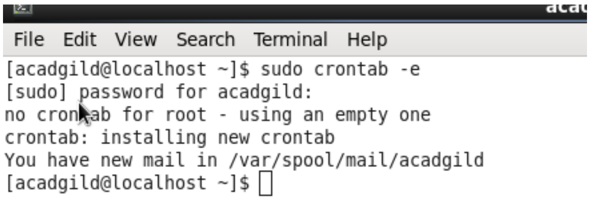
Command:

- Open cron tab -e
- Add the code for scheduling:
  * */3 * * * /home/acadgild/project/scripts/wrapper.sh

Code:

wrapper.sh

Screen:



```
[acadgild@localhost ~]$ sudo crontab -e
[sudo] password for acadgild:
no crontab for root - using an empty one
crontab: installing new crontab
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ 
```