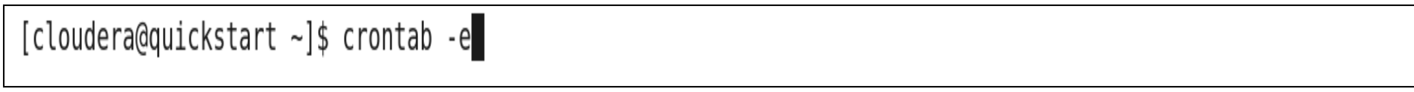
SPARK FINAL PROJECT

SETTING THE PERMISSION FOR BASH SCRIPT BEFORE EXECUTION

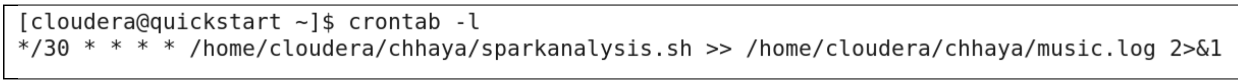
```
[cloudera@quickstart ~] chmod 755 /home/cloudera/chhaya/sparkanalysis.sh
```

EDIT THE CRON JOB TO CUSTOMIZE THE SCRIPT



```
[cloudera@quickstart ~]$ crontab -e
```

BROWSE THE CRONJOB POST EDITING



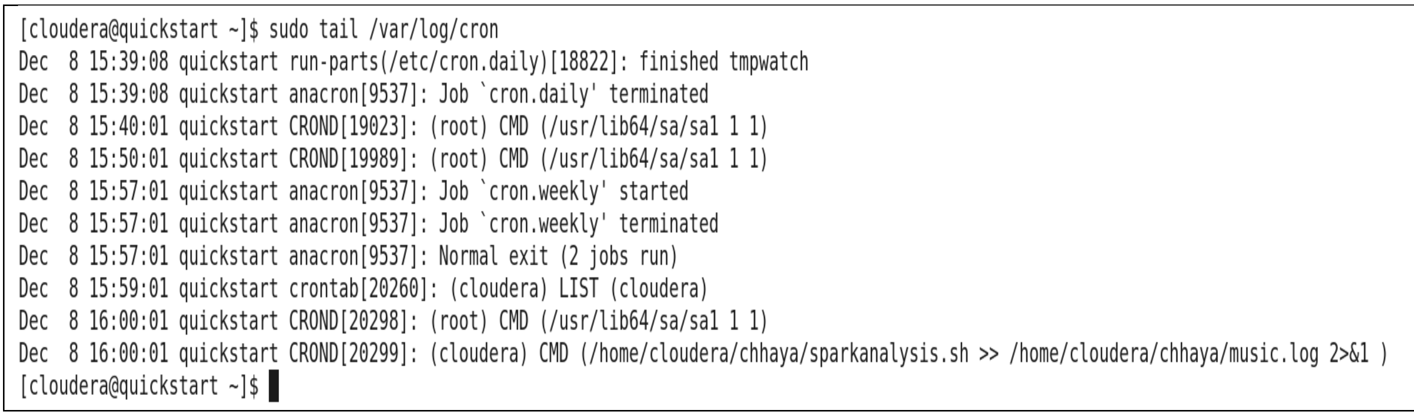
```
[cloudera@quickstart ~]$ crontab -l
*/30 * * * * /home/cloudera/chhaya/sparkanalysis.sh >> /home/cloudera/chhaya/music.log 2>&1
```

---SETTING THE TIME LIMIT FOR JOB TO RUN IN EVERY 30 MINS

---SETTING THE BASH SCRIPT WHICH CONTAINS ALL THE COMMANDS FOR HDFS, HBASE ,HIVE AND SPARK.

---KEEPING THE STDERR AND STDOUT to user defined log file.

MONITOR THE JOB WHEN IT EXECUTES



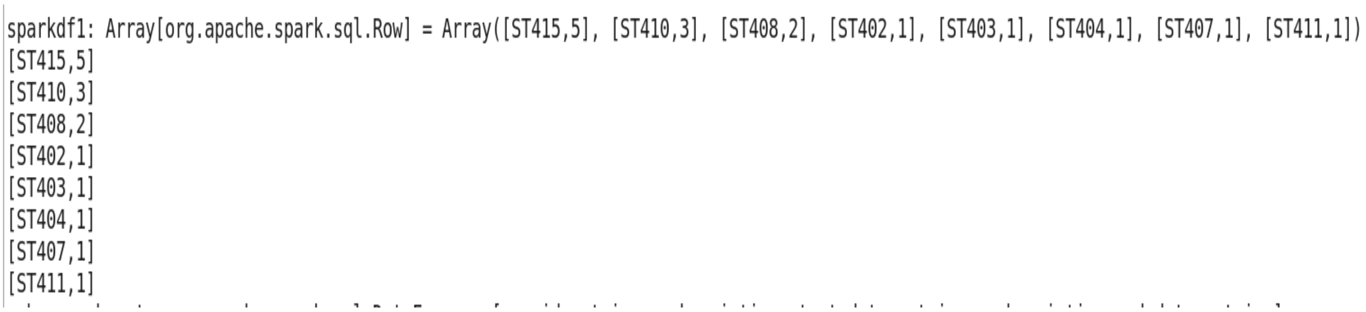
```
[cloudera@quickstart ~]$ sudo tail /var/log/cron
Dec  8 15:39:08 quickstart run-parts(/etc/cron.daily)[18822]: finished tmpwatch
Dec  8 15:39:08 quickstart anacron[9537]: Job `cron.daily' terminated
Dec  8 15:40:01 quickstart CROND[19023]: (root) CMD (/usr/lib64/sa/sa1 1 1)
Dec  8 15:50:01 quickstart CROND[19989]: (root) CMD (/usr/lib64/sa/sa1 1 1)
Dec  8 15:57:01 quickstart anacron[9537]: Job `cron.weekly' started
Dec  8 15:57:01 quickstart anacron[9537]: Job `cron.weekly' terminated
Dec  8 15:57:01 quickstart anacron[9537]: Normal exit (2 jobs run)
Dec  8 15:59:01 quickstart crontab[20260]: (cloudera) LIST (cloudera)
Dec  8 16:00:01 quickstart CROND[20298]: (root) CMD (/usr/lib64/sa/sa1 1 1)
Dec  8 16:00:01 quickstart CROND[20299]: (cloudera) CMD (/home/cloudera/chhaya/sparkanalysis.sh >> /home/cloudera/chhaya/music.log 2>&1 )
[cloudera@quickstart ~]$ 
```

OUTPUT FILE:/home/cloudera/chhaya/music.log

—OUTPUT FILE music.log is attached to GITHUB LINK for reference

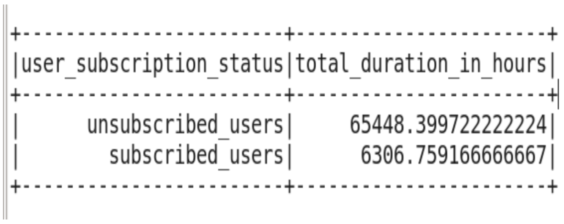
--This file will contain the output log as well as error logs.

SPARK QUERY1 OUTPUT:sparkdf1



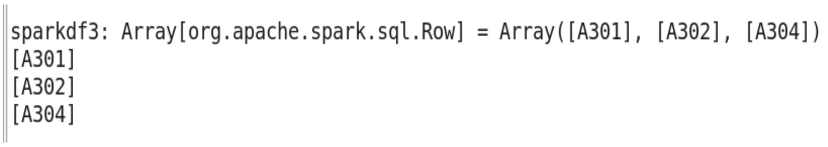
```
sparkdf1: Array[org.apache.spark.sql.Row] = Array([ST415,5], [ST410,3], [ST408,2], [ST402,1], [ST403,1], [ST404,1], [ST407,1], [ST411,1])
[ST415,5]
[ST410,3]
[ST408,2]
[ST402,1]
[ST403,1]
[ST404,1]
[ST407,1]
[ST411,1]
```

SPARK QUERY2 OUTPUT: sparkdf2



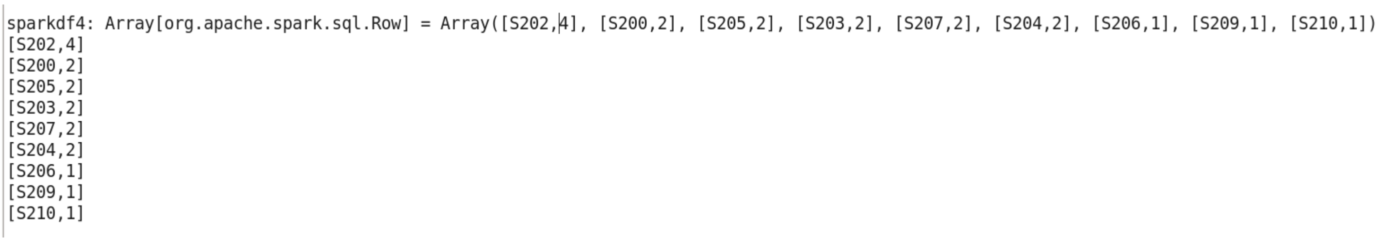
```
+------------------------+----------------------+
|user_subscription_status|total_duration_in_hours|
+------------------------+----------------------+
|      unsubscribed_users|     65448.399722222224|
|        subscribed_users|      6306.759166666667|
+------------------------+----------------------+
```

SPARK QUERY3 OUTPUT: sparkdf3



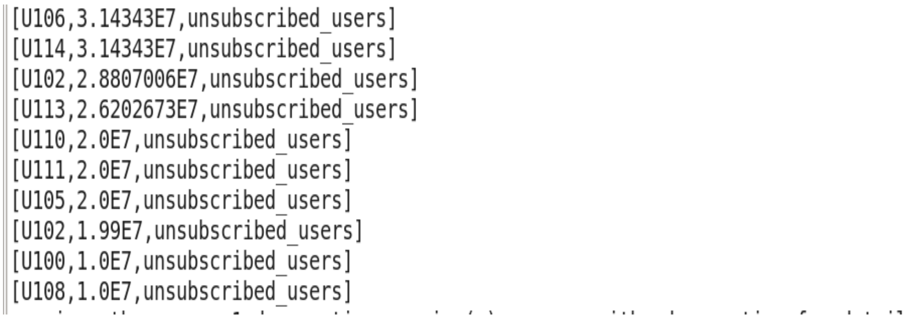
```
sparkdf3: Array[org.apache.spark.sql.Row] = Array([A301], [A302], [A304])
[A301]
[A302]
[A304]
```

SPARK QUERY4 OUTPUT: sparkdf4



```
sparkdf4: Array[org.apache.spark.sql.Row] = Array([S202,4], [S200,2], [S205,2], [S203,2], [S207,2], [S204,2], [S206,1], [S209,1], [S210,1])
[S202,4]
[S200,2]
[S205,2]
[S203,2]
[S207,2]
[S204,2]
[S206,1]
[S209,1]
[S210,1]
```

SPARK QUERY5 OUTPUT: sparkdf5



```
[U106,3.14343E7,unsubscribed_users]
[U114,3.14343E7,unsubscribed_users]
[U102,2.8807006E7,unsubscribed_users]
[U113,2.6202673E7,unsubscribed_users]
[U110,2.0E7,unsubscribed_users]
[U111,2.0E7,unsubscribed_users]
[U105,2.0E7,unsubscribed_users]
[U102,1.99E7,unsubscribed_users]
[U100,1.0E7,unsubscribed_users]
[U108,1.0E7,unsubscribed_users]
```

BASHSCRIPT DESCRIPTION : sparkanalysis.sh

This bashscript contains all commands for HBASE , HDFS,HIVE AND SPARK.

It calls below files which are attached to GITHUB DIRECTORY.

Hbasescript1.txt and hbasescripts2.txt contain HBASE COMMANDS.

Musiclatest.scala contains all the SPARK Commands.

Musiclatest.hql contains all the HIVE Commands.

All the HDFS commands are also written in same bash script.

HDFS commands to bulk load hbase tables are also written in same bash scripts.