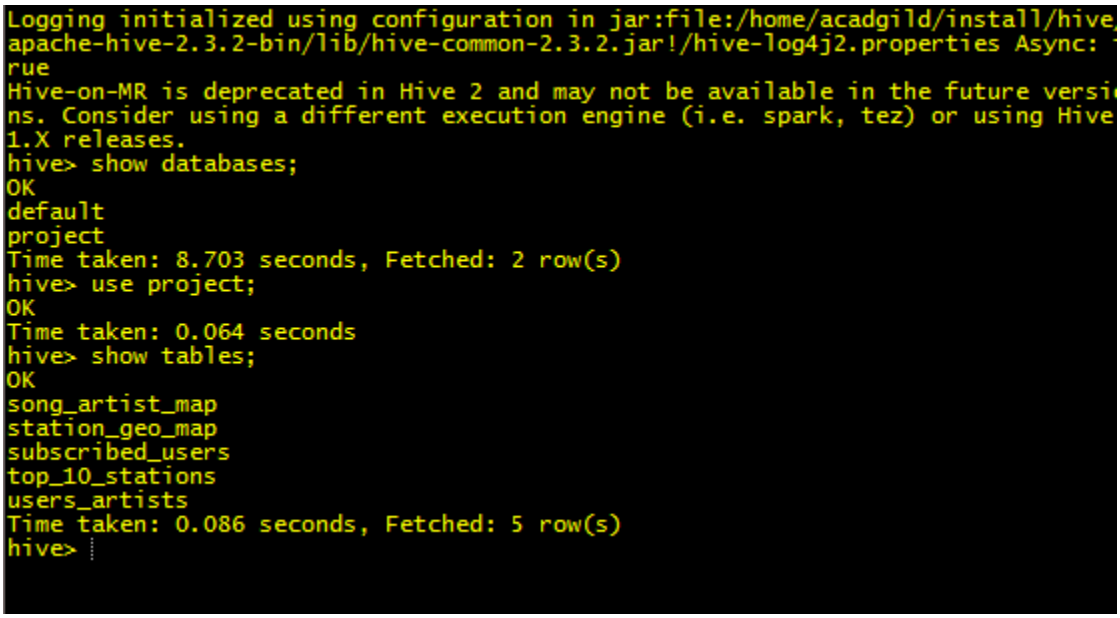
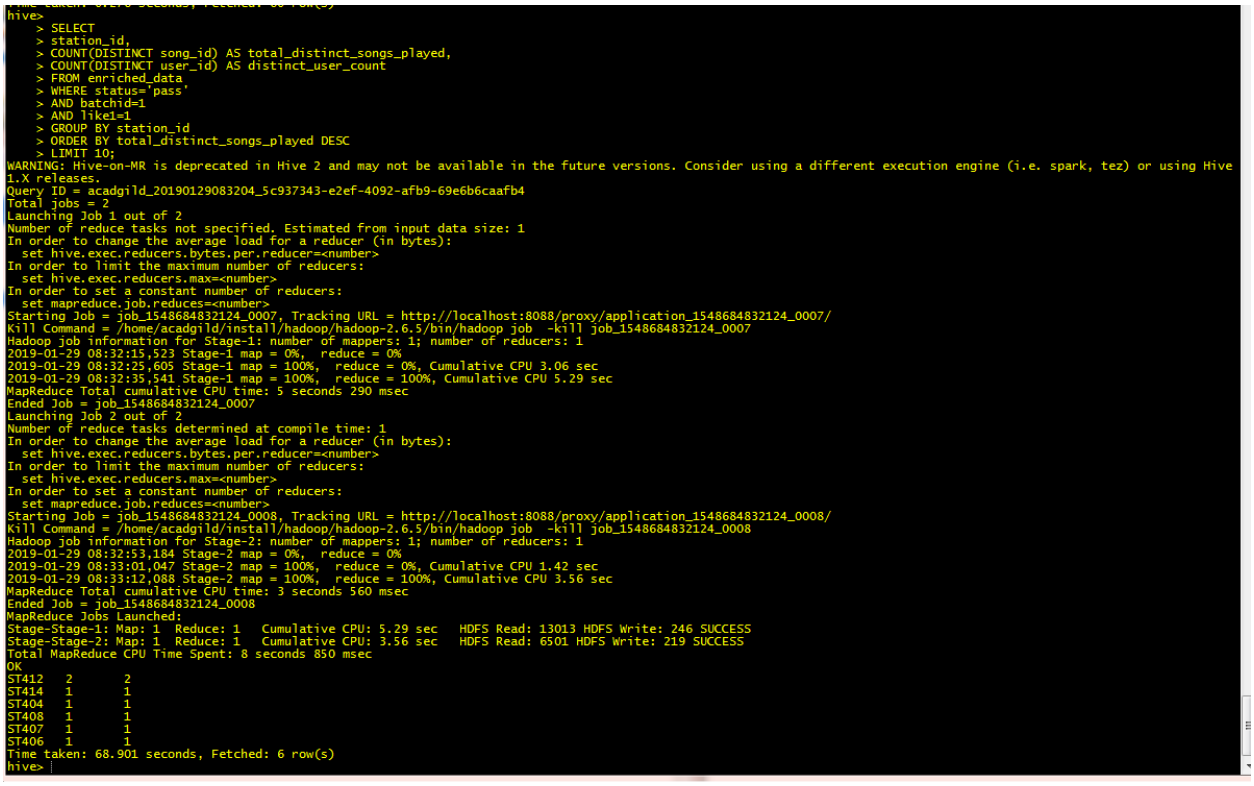
Final Project

Tables

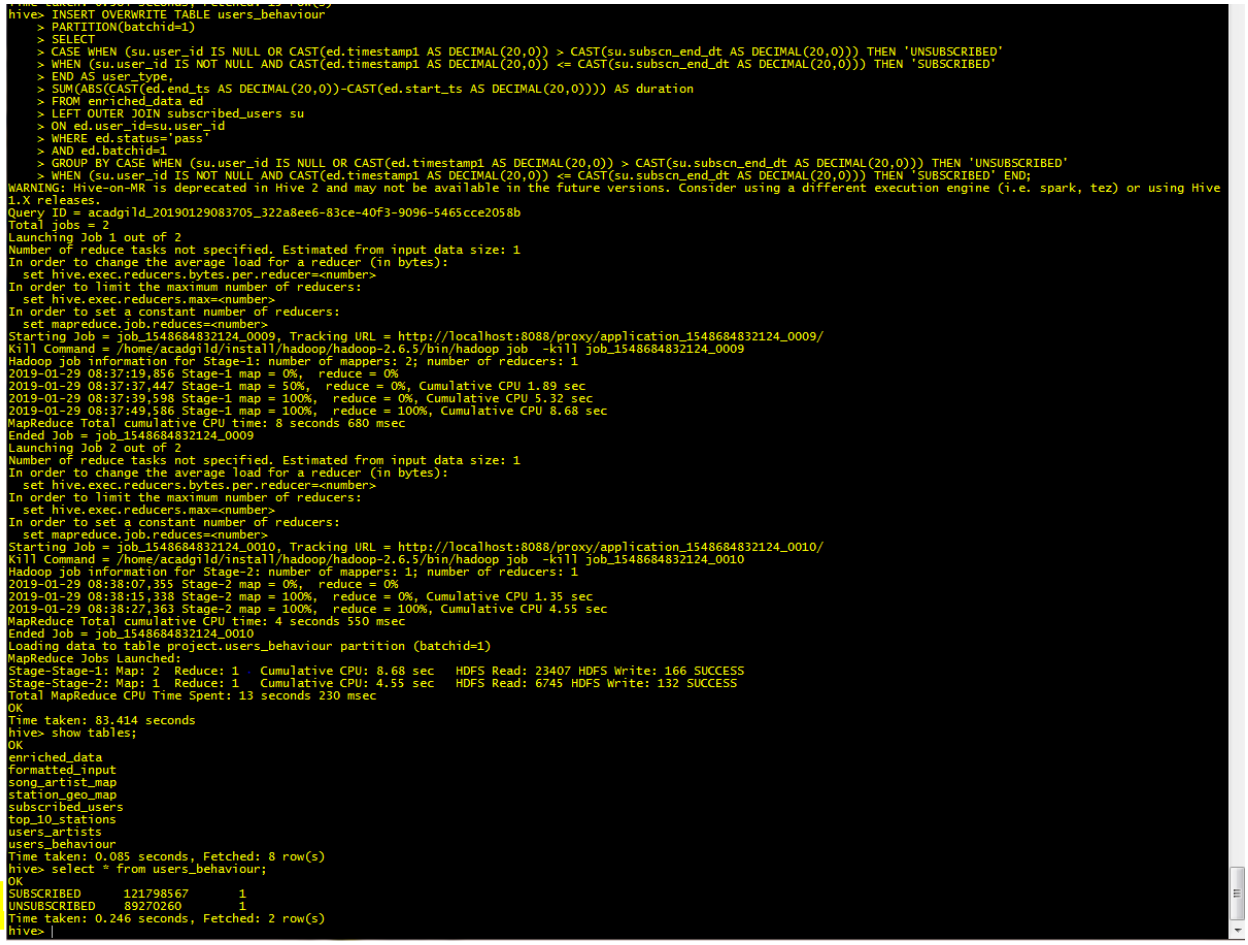


Problem Statement:

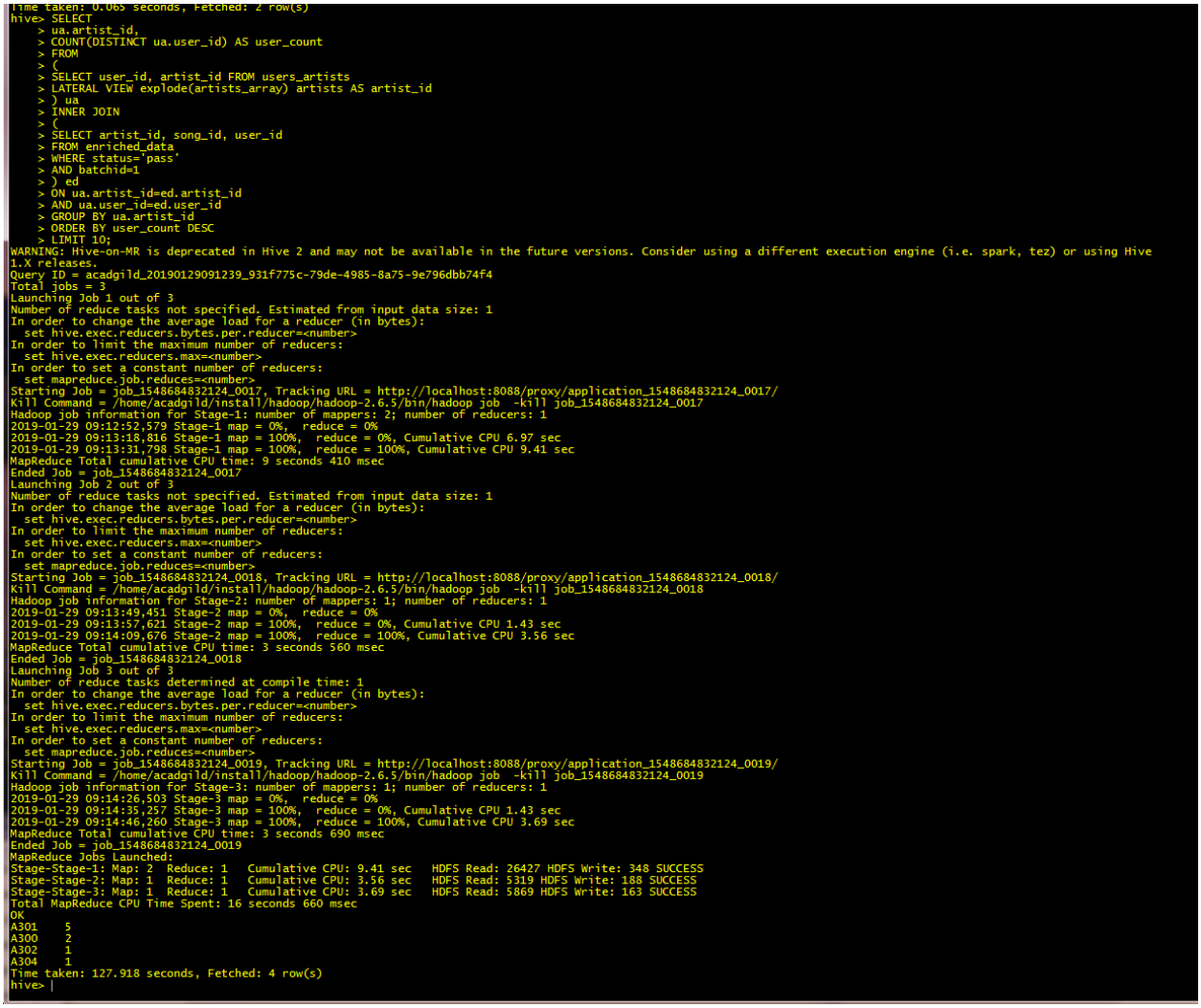
Data Analysis: Determine top 10 station_id(s) where maximum number of songs where placed, which were liked by unique users.



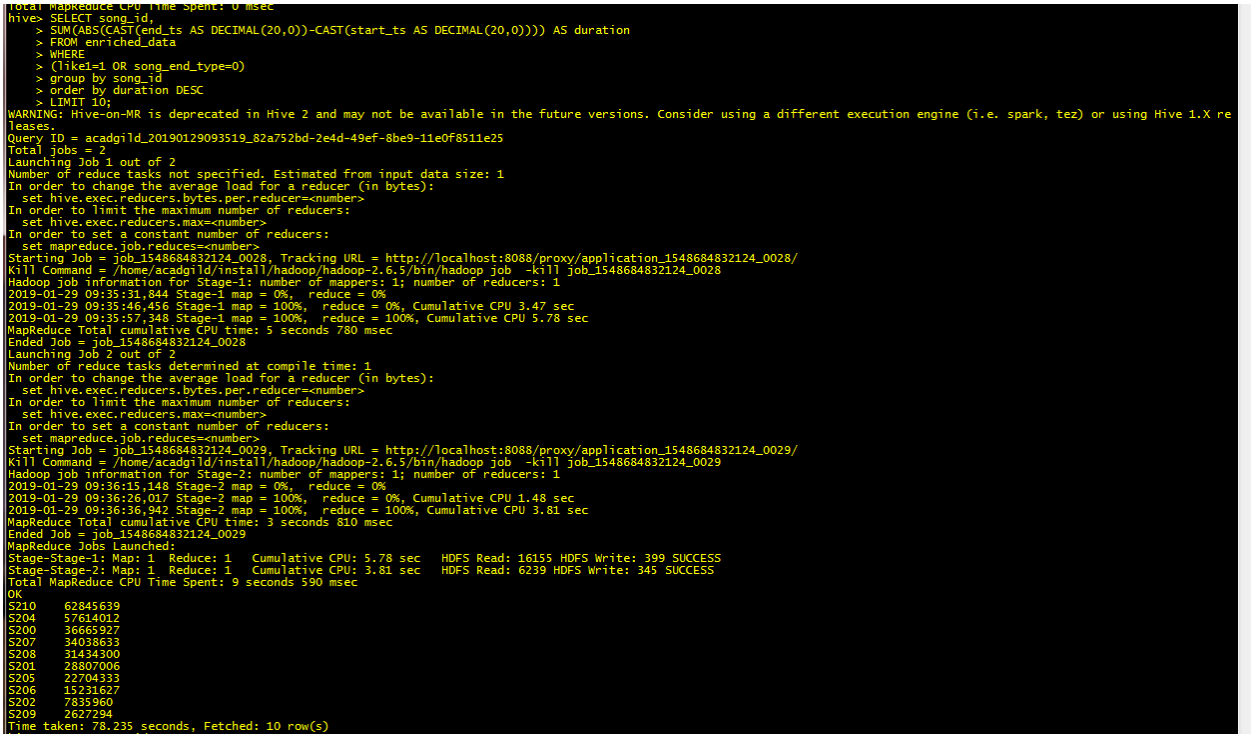
Determine total duration of songs played by each type of user, where type of user can be ‘subscribed’ or ‘unsubscribed’. An unsubscribed user is the one whose record is either not present in Subscribed_users lookup table or has subscription_end_date earlier than the timestamp of the song played by him.



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



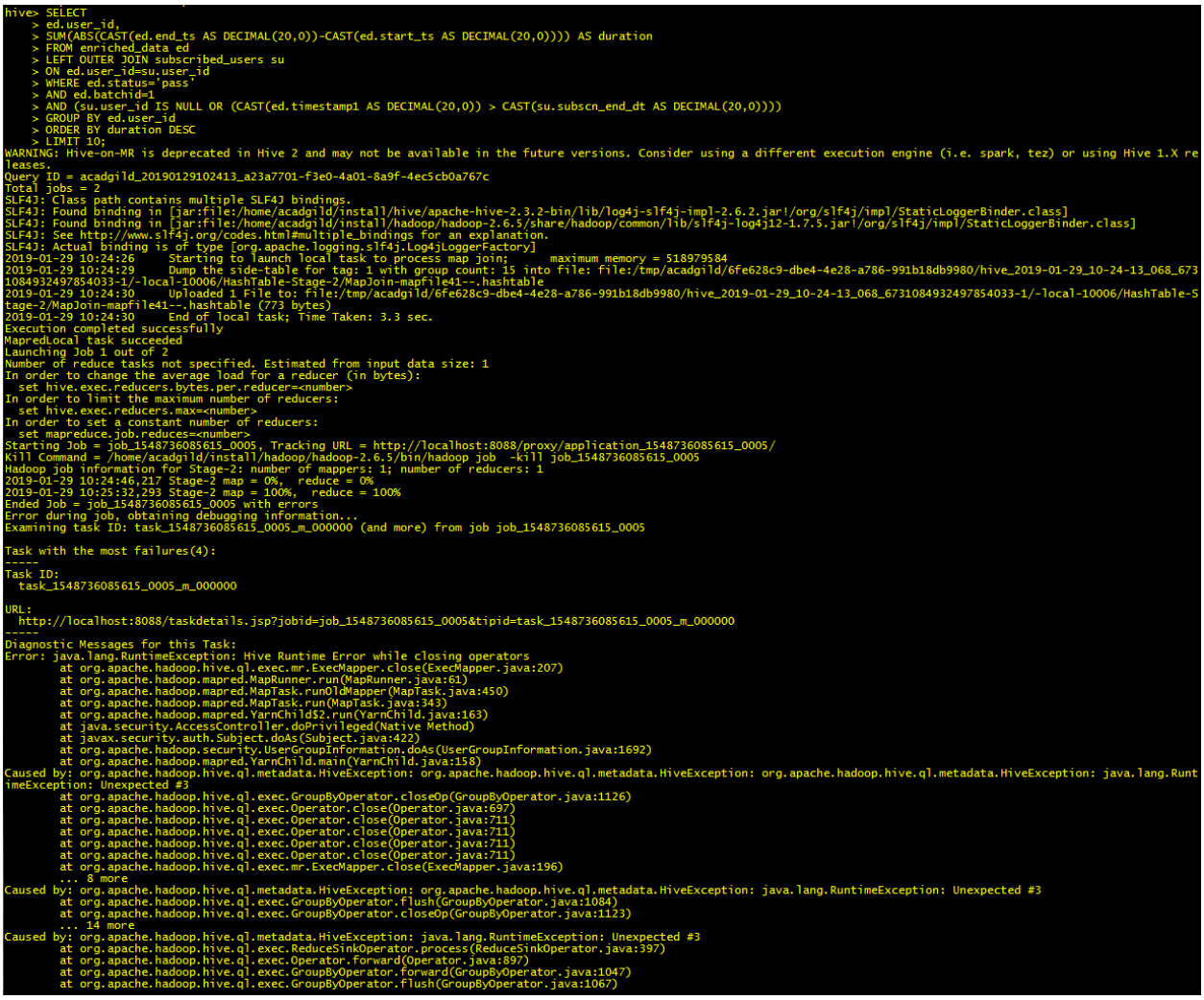
Determine top 10 songs that have generated the maximum revenue. Royalty applies to a song only if it was liked or was completed successfully or both.



```
Total MapReduce CPU Time Spent: 0 msec
hive> SELECT song_id,
    > SUM(ABS(CAST(end_ts AS DECIMAL(20,0))-CAST(start_ts AS DECIMAL(20,0)))) AS duration
    > FROM enriched_data
    > WHERE
    > (like1=1 OR song_end_type=0)
    > group by song_id
    > order by duration DESC
    > LIMIT 10;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20190129093519_82a752bd-2e4d-49ef-8be9-11e0f8511e25
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_1548684832124_0028, Tracking URL = http://localhost:8088/proxy/application_1548684832124_0028/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548684832124_0028
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-01-29 09:35:31,844 Stage-1 map = 0%,  reduce = 0%
2019-01-29 09:35:46,456 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.47 sec
2019-01-29 09:35:57,348 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.78 sec
MapReduce Total cumulative CPU time: 5 seconds 780 msec
Ended Job = job_1548684832124_0028
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_1548684832124_0029, Tracking URL = http://localhost:8088/proxy/application_1548684832124_0029/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548684832124_0029
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-01-29 09:36:15,148 Stage-2 map = 0%,  reduce = 0%
2019-01-29 09:36:26,017 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.48 sec
2019-01-29 09:36:36,942 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.81 sec
MapReduce Total cumulative CPU time: 3 seconds 810 msec
Ended Job = job_1548684832124_0029
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.78 sec   HDFS Read: 16155 HDFS Write: 399 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.81 sec   HDFS Read: 6239 HDFS Write: 345 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 590 msec
OK
S210    62845639
S204    57614012
S200    36665927
S207    34038633
S208    31434300
S201    28807006
S205    22704333
S206    15231627
S202    7835960
S209    2627294
Time taken: 78.235 seconds, Fetched: 10 row(s)
```

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

Unresolved problem with query execution. I believe the problem is with the heap size, but was unable to get it resolved.



```
hive> SELECT
    > ed.user_id,
    > SUM(ABS(CAST(ed.end_ts AS DECIMAL(20,0))-CAST(ed.start_ts AS DECIMAL(20,0)))) AS duration
    > FROM enriched_data ed
    > LEFT OUTER JOIN subscribed_users su
    > ON ed.user_id=su.user_id
    > WHERE ed.status='pass'
    > AND ed.batchid=1
    > AND (su.user_id IS NULL OR (CAST(ed.timestamp1 AS DECIMAL(20,0)) > CAST(su.subscn_end_dt AS DECIMAL(20,0))))
    > GROUP BY ed.user_id
    > ORDER BY duration DESC
    > LIMIT 10;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X re
leases.
Query ID = acadgild_20190129102413_a23a7701-f3e0-4a01-8a9f-4ec5cb0a767c
Total jobs = 2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
2019-01-29 10:24:26     Starting to launch local task to process map join;      maximum memory = 518979584
2019-01-29 10:24:29     Dump the side-table for tag: 1 with group count: 15 into file: file:/tmp/acadgild/6fe628c9-dbe4-4e28-a786-991b18db9980/hive_2019-01-29_10-24-13_068_673
1084932497854033-1/-local-10006/HashTable-Stage-2/MapJoin-mapfile41--.hashtable
2019-01-29 10:24:30     Uploaded 1 File to: file:/tmp/acadgild/6fe628c9-dbe4-4e28-a786-991b18db9980/hive_2019-01-29_10-24-13_068_6731084932497854033-1/-local-10006/HashTable-S
tage-2/MapJoin-mapfile41--.hashtable (773 bytes)
2019-01-29 10:24:30     End of local task; Time Taken: 3.3 sec.
Execution completed successfully
MapredLocal task succeeded
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_1548736085615_0005, Tracking URL = http://localhost:8088/proxy/application_1548736085615_0005/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1548736085615_0005
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-01-29 10:24:46,217 Stage-2 map = 0%,  reduce = 0%
2019-01-29 10:25:32,293 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_1548736085615_0005 with errors
Error during job, obtaining debugging information...
Examining task ID: task_1548736085615_0005_m_000000 (and more) from job job_1548736085615_0005

Task with the most failures(4):
-----
Task ID:
  task_1548736085615_0005_m_000000

URL:
  http://localhost:8088/taskdetails.jsp?jobid=job_1548736085615_0005&tipid=task_1548736085615_0005_m_000000
-----
Diagnostic Messages for this Task:
Error: java.lang.RuntimeException: Hive Runtime Error while closing operators
        at org.apache.hadoop.hive.ql.exec.mr.ExecMapper.close(ExecMapper.java:207)
        at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
        at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:450)
        at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
        at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:163)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1692)
        at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)
Caused by: org.apache.hadoop.hive.ql.metadata.HiveException: org.apache.hadoop.hive.ql.metadata.HiveException: org.apache.hadoop.hive.ql.metadata.HiveException: java.lang.Runt
imeException: Unexpected #3
        at org.apache.hadoop.hive.ql.exec.GroupByOperator.closeOp(GroupByOperator.java:1126)
        at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:697)
        at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:711)
        at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:711)
        at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:711)
        at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:711)
        at org.apache.hadoop.hive.ql.exec.mr.ExecMapper.close(ExecMapper.java:196)
        ... 8 more
Caused by: org.apache.hadoop.hive.ql.metadata.HiveException: org.apache.hadoop.hive.ql.metadata.HiveException: java.lang.RuntimeException: Unexpected #3
        at org.apache.hadoop.hive.ql.exec.GroupByOperator.flush(GroupByOperator.java:1084)
        at org.apache.hadoop.hive.ql.exec.GroupByOperator.closeOp(GroupByOperator.java:1123)
        ... 14 more
Caused by: org.apache.hadoop.hive.ql.metadata.HiveException: java.lang.RuntimeException: Unexpected #3
        at org.apache.hadoop.hive.ql.exec.ReduceSinkOperator.process(ReduceSinkOperator.java:397)
        at org.apache.hadoop.hive.ql.exec.Operator.forward(Operator.java:897)
        at org.apache.hadoop.hive.ql.exec.GroupByOperator.forward(GroupByOperator.java:1047)
        at org.apache.hadoop.hive.ql.exec.GroupByOperator.flush(GroupByOperator.java:1067)
```