

Final Project

Tables

```
Logging initialized using configuration in jar:file:/home/acadgild/install/hive-  
apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true  
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.  
hive> show databases;  
OK  
default  
project  
Time taken: 8.703 seconds, Fetched: 2 row(s)  
hive> use project;  
OK  
Time taken: 0.064 seconds  
hive> show tables;  
OK  
song_artist_map  
station_geo_map  
subscribed_users  
top_10_stations  
users_artists  
Time taken: 0.086 seconds, Fetched: 5 row(s)  
hive>
```

Problem Statement:

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

```
hive>
> SELECT
> station_id,
> COUNT(DISTINCT song_id) AS total_distinct_songs_played,
> COUNT(DISTINCT user_id) AS distinct_user_count
> FROM enriched_data
> WHERE status='pass'
> AND batchid=1
> AND liked=1
> GROUP BY station_id
> ORDER BY total_distinct_songs_played 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_20190129083204_5c937343-e2ef-4092-afb9-69e6b6caafb4
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_0007, Tracking URL = http://localhost:8088/proxy/application_1548684832124_0007/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1548684832124_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-01-29 08:32:15,523 Stage-1 map = 0%, reduce = 0%
2019-01-29 08:32:25,605 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.06 sec
2019-01-29 08:32:35,541 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.29 sec
MapReduce Total cumulative CPU time: 5 seconds 290 msec
Ended Job = job_1548684832124_0007
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_0008, Tracking URL = http://localhost:8088/proxy/application_1548684832124_0008/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1548684832124_0008
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-01-29 08:32:53,184 Stage-2 map = 0%, reduce = 0%
2019-01-29 08:33:01,047 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.42 sec
2019-01-29 08:33:12,088 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.56 sec
MapReduce Total cumulative CPU time: 3 seconds 560 msec
Ended Job = job_1548684832124_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.29 sec HDFS Read: 13013 HDFS Write: 246 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.56 sec HDFS Read: 6501 HDFS Write: 219 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 850 msec
OK
ST412 2 2
ST414 1 1
ST404 1 1
ST408 1 1
ST407 1 1
ST406 1 1
Time taken: 68.901 seconds, Fetched: 6 row(s)
hive>
```

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.

```
hive> INSERT OVERWRITE TABLE users_behaviour
> PARTITION(batchid=1)
> SELECT
> CASE WHEN (su.user_id IS NULL OR CAST(ed.timestamp1 AS DECIMAL(20,0)) > CAST(su.subscn_end_dt AS DECIMAL(20,0))) THEN 'UNSUBSCRIBED'
> WHEN (su.user_id IS NOT NULL AND CAST(ed.timestamp1 AS DECIMAL(20,0)) <= CAST(su.subscn_end_dt AS DECIMAL(20,0))) THEN 'SUBSCRIBED'
> END AS user_type,
> 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
> GROUP BY CASE WHEN (su.user_id IS NULL OR CAST(ed.timestamp1 AS DECIMAL(20,0)) > CAST(su.subscn_end_dt AS DECIMAL(20,0))) THEN 'UNSUBSCRIBED'
> WHEN (su.user_id IS NOT NULL AND CAST(ed.timestamp1 AS DECIMAL(20,0)) <= CAST(su.subscn_end_dt AS DECIMAL(20,0))) THEN 'SUBSCRIBED' END;
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_20190129083705_322a8ee6-83ce-40f3-9096-5465cce2058b
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_0009, Tracking URL = http://localhost:8088/proxy/application_1548684832124_0009/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1548684832124_0009
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2019-01-29 08:37:19,856 Stage-1 map = 0%, reduce = 0%
2019-01-29 08:37:19,856 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 1.89 sec
2019-01-29 08:37:39,598 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.32 sec
2019-01-29 08:37:49,586 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.68 sec
MapReduce Total cumulative CPU time: 8 seconds 680 msec
Ended Job = job_1548684832124_0009
Launching Job 2 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_0010, Tracking URL = http://localhost:8088/proxy/application_1548684832124_0010/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1548684832124_0010
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-01-29 08:38:07,355 Stage-2 map = 0%, reduce = 0%
2019-01-29 08:38:15,338 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.35 sec
2019-01-29 08:38:27,363 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 4.55 sec
MapReduce Total cumulative CPU time: 4 seconds 550 msec
Ended Job = job_1548684832124_0010
Loading data to table project.users_behaviour partition (batchid=1)
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 1 Cumulative CPU: 8.68 sec HDFS Read: 23407 HDFS Write: 166 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 4.55 sec HDFS Read: 6745 HDFS Write: 132 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 230 msec
OK
Time taken: 83.414 seconds
hive> show tables;
OK
enriched_data
formatted_input
song_artist_map
station_geo_map
subscribed_users
top_10_stations
users_artists
users_behaviour
Time taken: 0.085 seconds, Fetched: 8 row(s)
hive> select * from users_behaviour;
OK
SUBSCRIBED 121798567 1
UNSUBSCRIBED 89270260 1
Time taken: 0.246 seconds, Fetched: 2 row(s)
hive>
```

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

```
Time taken: 0.063 seconds, Fetched: 2 row(s)
hive> SELECT
> ua.artist_id,
> COUNT(DISTINCT ua.user_id) AS user_count
> FROM
> (
> SELECT user_id, artist_id FROM users_artists
> LATERAL VIEW explode(artists_array) artists AS artist_id
> ) ua
> INNER JOIN
> (
> SELECT artist_id, song_id, user_id
> FROM enriched_data
> WHERE status='pass'
> AND batchid=1
> ) ed
> ON ua.artist_id=ed.artist_id
> AND ua.user_id=ed.user_id
> GROUP BY ua.artist_id
> ORDER BY user_count 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_20190129091239_931f775c-79de-4985-8a75-9e796dbb74f4
Total jobs = 3
Launching Job 1 out of 3
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_0017, Tracking URL = http://localhost:8088/proxy/application_1548684832124_0017/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1548684832124_0017
Hadoop job information for Stage:1: number of mappers: 2; number of reducers: 1
2019-01-29 09:12:52,579 Stage-1 map = 0%, reduce = 0%
2019-01-29 09:13:31,798 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.97 sec
2019-01-29 09:13:31,798 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 9.41 sec
MapReduce Total cumulative CPU time: 9 seconds 410 msec
Ended Job = job_1548684832124_0017
Launching Job 2 out of 3
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_0018, Tracking URL = http://localhost:8088/proxy/application_1548684832124_0018/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1548684832124_0018
Hadoop job information for Stage:2: number of mappers: 1; number of reducers: 1
2019-01-29 09:13:49,451 Stage-2 map = 0%, reduce = 0%
2019-01-29 09:13:57,621 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.43 sec
2019-01-29 09:14:09,676 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.56 sec
MapReduce Total cumulative CPU time: 3 seconds 560 msec
Ended Job = job_1548684832124_0018
Launching Job 3 out of 3
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_0019, Tracking URL = http://localhost:8088/proxy/application_1548684832124_0019/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1548684832124_0019
Hadoop job information for Stage:3: number of mappers: 1; number of reducers: 1
2019-01-29 09:14:26,503 Stage-3 map = 0%, reduce = 0%
2019-01-29 09:14:35,257 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 1.43 sec
2019-01-29 09:14:46,260 Stage-3 map = 100%, reduce = 100%, Cumulative CPU 3.69 sec
MapReduce Total cumulative CPU time: 3 seconds 690 msec
Ended Job = job_1548684832124_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 1 Cumulative CPU: 9.41 sec HDFS Read: 26427 HDFS Write: 348 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.56 sec HDFS Read: 5319 HDFS Write: 188 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 3.69 sec HDFS Read: 5869 HDFS Write: 163 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 660 msec
OK
A301 5
A300 2
A302 1
A304 1
Time taken: 127.918 seconds, Fetched: 4 row(s)
hive> |
```

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
> (liked=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 who 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 unsubscribed_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 releases.
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-Stage-2/MapJoin-mapfile41--.hashtable (773 bytes)
2019-01-29 10:24:30 End of local task; Time Taken: 3.3 sec.
Execution completed successfully
Mapreduce local 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 10:25:22,292 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.q1.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.mapred.YarnChild.main(YarnChild.java:158)
Caused by: org.apache.hadoop.hive.q1.metadata.HiveException: org.apache.hadoop.hive.q1.metadata.HiveException: java.lang.RuntimeException: Unexpected #3
    at org.apache.hadoop.hive.q1.exec.GroupByOperator.closeOp(GroupByOperator.java:1126)
    at org.apache.hadoop.hive.q1.exec.Operator.close(Operator.java:697)
    at org.apache.hadoop.hive.q1.exec.Operator.close(Operator.java:711)
    at org.apache.hadoop.hive.q1.exec.Operator.close(Operator.java:711)
    at org.apache.hadoop.hive.q1.exec.Operator.close(Operator.java:711)
    at org.apache.hadoop.hive.q1.exec.Operator.close(Operator.java:711)
    at org.apache.hadoop.hive.q1.exec.mr.ExecMapper.close(ExecMapper.java:196)
    ... 8 more
Caused by: org.apache.hadoop.hive.q1.metadata.HiveException: org.apache.hadoop.hive.q1.metadata.HiveException: java.lang.RuntimeException: Unexpected #3
    at org.apache.hadoop.hive.q1.exec.GroupByOperator.flush(GroupByOperator.java:1084)
    at org.apache.hadoop.hive.q1.exec.GroupByOperator.closeOp(GroupByOperator.java:1123)
    ... 14 more
Caused by: org.apache.hadoop.hive.q1.metadata.HiveException: java.lang.RuntimeException: Unexpected #3
    at org.apache.hadoop.hive.q1.exec.ReduceSinkOperator.process(ReduceSinkOperator.java:397)
    at org.apache.hadoop.hive.q1.exec.Operator.forward(Operator.java:897)
    at org.apache.hadoop.hive.q1.exec.GroupByOperator.forward(GroupByOperator.java:1047)
    at org.apache.hadoop.hive.q1.exec.GroupByOperator.flush(GroupByOperator.java:1067)
```