HQL TASKS ON HIVE

The purpose of this document is to highlight the following

- Importing data from RDS to HDFS
- Creation of HIVE Tables
- Loading of Data from RDS and Text files
- Performing HIVE Queries on HIVE tables
- Insights gained while performing various hive tasks on HIVE Tables.

1. Importing RDS Table data into HDFS

1.1 Importing app_events table data into HDFS

1.1.1 Command

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table app_events --target-dir /user/hadoop/mlctest/app_events --username student -P -m 1

The system will prompt for password and please provide **STUDENT123**

1.1.2 Scoop import

```
(base) sanjana_mantri@Sanjanas-MacBook-Air Capstone % chmod 400 csd_pair.pem
(base) sanjana_mantri@Sanjanas-MacBook-Air Capstone % ssh -i ~/csd_pair.pem hadoop@ec2-3-235-23-1.compute-1.amazonaws.com
Warning: Identity file /Users/sanjana_mantri/csd_pair.pem not accessible: No such file or directory.
The authenticity of host 'ec2-3-235-23-1.compute-1.amazonaws.com (3.235.23.1)' can't be established.
ECDSA key fingerprint is SHA256:7grVk4O8PmUrl2VmWzcLtJnm1GKPJ7NDJVUkVyhJncA.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-3-235-23-1.compute-1.amazonaws.com,3.235.23.1' (ECDSA) to the list of known hosts.
hadoop@ec2-3-235-23-1.compute-1.amazonaws.com: Permission denied (publickey,gssapi-keyex,gssapi-with-mic).
(base) sanjana_mantri@Sanjanas-MacBook-Air Capstone % ssh -i csd_pair.pem hadoop@ec2-3-235-23-1.compute-1.amazonaws.com
Last login: Thu Oct 14 16:28:15 2021

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
13 package(s) needed for security, out of 44 available
Run "sudo yum update" to apply all updates.
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M::::::M         M::::::M R::::::::::::::R
EE::::EEEEEEEEE:::E M:::::::M       M:::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M::::::::M     M::::::::M RR::::R      R::::R
  E::::E             M:::::M:::M   M:::M:::::M   R:::R      R::::R
  E::::::EEEEEEEEEE  M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E:::::::::::::::E  M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E::::::EEEEEEEEEE  M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE::::::EEEEEEEE:::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR
```

1.1.3 HDFS storage

The following command will list the HDFS contents

hadoop fs -ls mlctest/app_events

```
21/10/14 16:31:10 INFO mapreduce.Job: The url to track the job: http://ip-172-31-75-0.ec2.internal:20888/proxy/application_1634228715435_0001/
21/10/14 16:31:10 INFO mapreduce.Job: Running job: job_1634228715435_0001
21/10/14 16:31:16 INFO mapreduce.Job: Job job_1634228715435_0001 running in uber mode : false
21/10/14 16:31:16 INFO mapreduce.Job:  map 0% reduce 0%
21/10/14 16:32:11 INFO mapreduce.Job:  map 100% reduce 0%
21/10/14 16:32:11 INFO mapreduce.Job: Job job_1634228715435_0001 completed successfully
21/10/14 16:32:11 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=228485
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=1037267620
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=5045472
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=52557
                Total vcore-milliseconds taken by all map tasks=52557
                Total megabyte-milliseconds taken by all map tasks=161455104
        Map-Reduce Framework
                Map input records=32473067
                Map output records=32473067
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=186
                CPU time spent (ms)=34910
                Physical memory (bytes) snapshot=1049788416
                Virtual memory (bytes) snapshot=4638875648
                Total committed heap usage (bytes)=653262848
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=1037267620
21/10/14 16:32:11 INFO mapreduce.ImportJobBase: Transferred 989.2155 MB in 66.6493 seconds (14.8421 MB/sec)
21/10/14 16:32:11 INFO mapreduce.ImportJobBase: Retrieved 32473067 records.
[hadoop@ip-172-31-75-0 ~]$
```

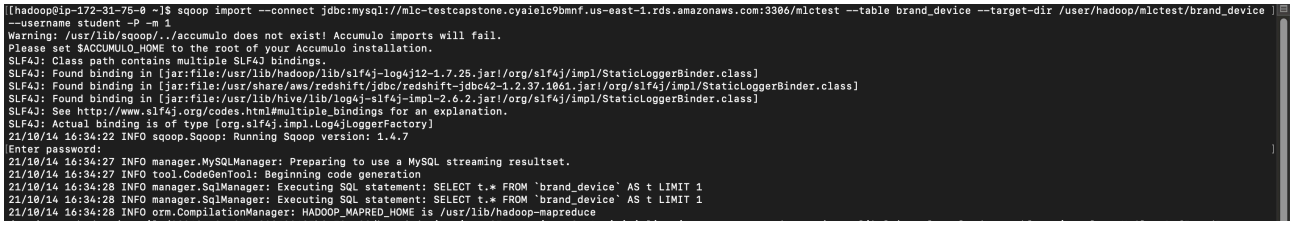
```
[hadoop@ip-172-31-75-0 ~]$ hadoop fs -ls mlctest/app_events
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2021-10-14 16:32 mlctest/app_events/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup 1037267620 2021-10-14 16:32 mlctest/app_events/part-m-00000
[hadoop@ip-172-31-75-0 ~]$
```

1.2 Importing brand_device table data into HDFS

1.2.1 Command

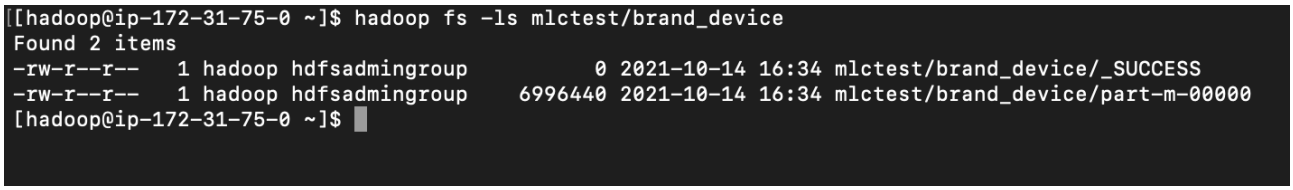
sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table brand_device --target-dir /user/hadoop/mlctest/brand_device --username student -P -m 1
The system will prompt for password and please provide **STUDENT123**

1.2.2 Scoop import



1.2.3 HDFS storage

The following command will list the HDFS contents
hadoop fs -ls mlctest/brand_device

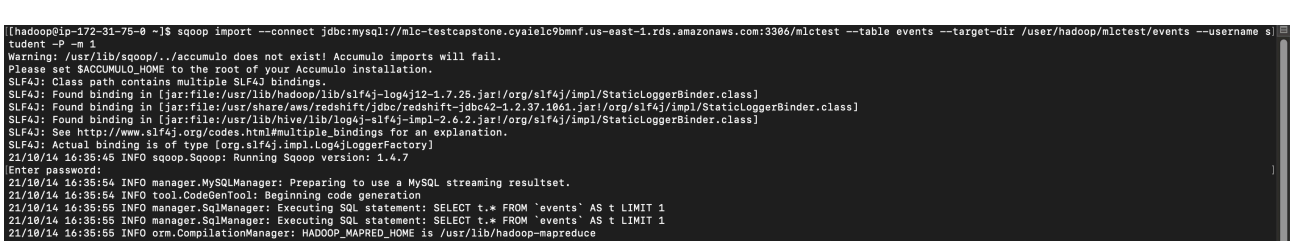


1.3 Importing events table data into HDFS

1.3.1 Command

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table events --target-dir /user/hadoop/mlctest/events --username student -P -m 1
The system will prompt for password and please provide **STUDENT123**

1.3.2 Scoop import



1.3.3 HDFS storage

The following command will list the HDFS contents
hadoop fs -ls mlctest/events

```
[[hadoop@ip-172-31-75-0 ~]$ hadoop fs -ls mlctest/events
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2021-10-14 16:36 mlctest/events/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup  194985245 2021-10-14 16:36 mlctest/events/part-m-00000
[hadoop@ip-172-31-75-0 ~]$
```

1.4 Importing train table data into HDFS

1.4.1 Command

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table train --target-dir /user/hadoop/mlctest/train --username student -P -m 1

The system will prompt for password and please provide **STUDENT123**

1.4.2 Scoop import

```
21/10/14 16:37:18 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/fcb287fe185f36cf1c1eb5771cc26bfa/train.jar
21/10/14 16:37:18 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/10/14 16:37:18 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/10/14 16:37:18 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/10/14 16:37:18 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/10/14 16:37:18 INFO mapreduce.ImportJobBase: Beginning import of train
21/10/14 16:37:18 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/10/14 16:37:19 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
21/10/14 16:37:19 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-75-0.ec2.internal/172.31.75.0:8032
21/10/14 16:37:19 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-75-0.ec2.internal/172.31.75.0:10200
21/10/14 16:37:26 INFO db.DBInputFormat: Using read commited transaction isolation
21/10/14 16:37:26 INFO mapreduce.JobSubmitter: number of splits:1
21/10/14 16:37:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634228715435_0004
21/10/14 16:37:26 INFO conf.Configuration: resource-types.xml not found
21/10/14 16:37:26 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/10/14 16:37:26 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
21/10/14 16:37:26 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
21/10/14 16:37:27 INFO impl.YarnClientImpl: Submitted application application_1634228715435_0004
21/10/14 16:37:27 INFO mapreduce.Job: The url to track the job: http://ip-172-31-75-0.ec2.internal:20888/proxy/application_1634228715435_0004/
21/10/14 16:37:27 INFO mapreduce.Job: Running job: job_1634228715435_0004
21/10/14 16:37:33 INFO mapreduce.Job: Job job_1634228715435_0004 running in uber mode : false
21/10/14 16:37:33 INFO mapreduce.Job:  map 0% reduce 0%
21/10/14 16:37:38 INFO mapreduce.Job:  map 100% reduce 0%
21/10/14 16:37:38 INFO mapreduce.Job: Job job_1634228715435_0004 completed successfully
21/10/14 16:37:38 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=228459
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=2421599
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=305760
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=3185
                Total vcore-milliseconds taken by all map tasks=3185
                Total megabyte-milliseconds taken by all map tasks=9784320
        Map-Reduce Framework
                Map input records=74645
                Map output records=74645
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=72
                CPU time spent (ms)=2210
                Physical memory (bytes) snapshot=397111296
                Virtual memory (bytes) snapshot=4635516928
                Total committed heap usage (bytes)=320864256
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=2421599
21/10/14 16:37:38 INFO mapreduce.ImportJobBase: Transferred 2.3094 MB in 18.9104 seconds (125.0552 KB/sec)
21/10/14 16:37:38 INFO mapreduce.ImportJobBase: Retrieved 74645 records.
[hadoop@ip-172-31-75-0 ~]$
```

1.4.3 HDFS storage

The following command will list the HDFS contents
hadoop fs -ls mlctest/train

```
[hadoop@ip-172-31-75-0 ~]$ hadoop fs -ls mlctest/train
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2021-10-14 16:37 mlctest/train/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup    2421599 2021-10-14 16:37 mlctest/train/part-m-00000
[hadoop@ip-172-31-75-0 ~]$ beeline -u jdbc:hive2://localhost:10000/default -n hadoop
Connecting to jdbc:hive2://localhost:10000/default
Connected to: Apache Hive (version 2.3.7-amzn-4)
Driver: Hive JDBC (version 2.3.7-amzn-4)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.3.7-amzn-4 by Apache Hive
0: jdbc:hive2://localhost:10000/default> create database mlctest;
INFO  : Compiling command(queryId=hive_20211014163853_fa7e34b9-c95b-4d65-be62-a0ce007b402d): create database mlctest
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014163853_fa7e34b9-c95b-4d65-be62-a0ce007b402d); Time taken: 0.589 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014163853_fa7e34b9-c95b-4d65-be62-a0ce007b402d): create database mlctest
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211014163853_fa7e34b9-c95b-4d65-be62-a0ce007b402d); Time taken: 0.193 seconds
INFO  : OK
No rows affected (0.935 seconds)
0: jdbc:hive2://localhost:10000/default> use mlctest;
INFO  : Compiling command(queryId=hive_20211014163903_f09f44a8-a843-4c0f-8578-35c18dfdd161): use mlctest
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014163903_f09f44a8-a843-4c0f-8578-35c18dfdd161); Time taken: 0.021 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014163903_f09f44a8-a843-4c0f-8578-35c18dfdd161): use mlctest
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211014163903_f09f44a8-a843-4c0f-8578-35c18dfdd161); Time taken: 0.01 seconds
INFO  : OK
No rows affected (0.055 seconds)
0: jdbc:hive2://localhost:10000/default>
```

2 HIVE: Database, Tables Creation and Loading data into HIVE tables

Before we create the HIVE tables, we need to create the database mlctest and issue a command to use that database.
To create database and table, we need to utilize **beeline** which is a HIVE client and uses JDBC.
beeline -u jdbc:hive2://localhost:10000/default -n hadoop
create database mlctest;
use mlctest;


```
[hadoop@ip-172-31-75-0 ~]$ hadoop fs -ls mlctest/train
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2021-10-14 16:37 mlctest/train/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup    2421599 2021-10-14 16:37 mlctest/train/part-m-00000
[hadoop@ip-172-31-75-0 ~]$ beeline -u jdbc:hive2://localhost:10000/default -n hadoop
Connecting to jdbc:hive2://localhost:10000/default
Connected to: Apache Hive (version 2.3.7-amzn-4)
Driver: Hive JDBC (version 2.3.7-amzn-4)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.3.7-amzn-4 by Apache Hive
0: jdbc:hive2://localhost:10000/default> create database mlctest;
INFO  : Compiling command(queryId=hive_20211014163853_fa7e34b9-c95b-4d65-be62-a0ce007b402d): create database mlctest
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014163853_fa7e34b9-c95b-4d65-be62-a0ce007b402d); Time taken: 0.589 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014163853_fa7e34b9-c95b-4d65-be62-a0ce007b402d): create database mlctest
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211014163853_fa7e34b9-c95b-4d65-be62-a0ce007b402d); Time taken: 0.193 seconds
INFO  : OK
No rows affected (0.935 seconds)
0: jdbc:hive2://localhost:10000/default> use mlctest;
INFO  : Compiling command(queryId=hive_20211014163903_f09f44a8-a843-4c0f-8578-35c18dfdd161): use mlctest
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014163903_f09f44a8-a843-4c0f-8578-35c18dfdd161); Time taken: 0.021 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014163903_f09f44a8-a843-4c0f-8578-35c18dfdd161): use mlctest
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211014163903_f09f44a8-a843-4c0f-8578-35c18dfdd161); Time taken: 0.01 seconds
INFO  : OK
No rows affected (0.055 seconds)
0: jdbc:hive2://localhost:10000/default>
```

2.1 HIVE Tables Creation

2.1.1 Table Creation Commands

The following commands are used to create the following HIVE external tables

- app_events_external
- train_external
- brand_device_external
- events_external
- app_labels_external
- label_categories_external

app_events_external

create external table if not exists app_events_external (event_id int, app_id string, is_installed int, is_active int) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;

train_external

create external table if not exists train_external (device_id string, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;

brand_device_external

create external table if not exists brand_device_external (device_id string, phone_brand string, device_model string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;

events_external

create external table if not exists events_external (event_id int, device_id string, event_time timestamp, latitude float, longitude float) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;

app_labels_external

create external table if not exists app_labels_external (app_id string, label_id int) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1");

label_categories_external

create external table if not exists label_categories_external (label_id int, category string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1");

2.1.2 Table Creation Screenshots

```
0: jdbc:hive2://localhost:10000/default> create external table if not exists app_events_external (event_id int, app_id string, is_installed int, is_active int) row format delimited fields terminated by ","
," lines terminated by "\n" stored as textfile;
INFO  : Compiling command(queryId=hive_20211014163936_286bf1e5-9569-47dc-97c1-8b5baeb095e2): create external table if not exists app_events_external (event_id int, app_id string, is_installed int, is_active int) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014163936_286bf1e5-9569-47dc-97c1-8b5baeb095e2); Time taken: 0.101 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014163936_286bf1e5-9569-47dc-97c1-8b5baeb095e2): create external table if not exists app_events_external (event_id int, app_id string, is_installed int, is_active int) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211014163936_286bf1e5-9569-47dc-97c1-8b5baeb095e2); Time taken: 0.3 seconds
INFO  : OK
No rows affected (0.417 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists train_external (device_id string, gender string, age int, group_train string) row format delimited fields terminated by "," l
ines terminated by "\n" stored as textfile;
INFO  : Compiling command(queryId=hive_20211014163950_f17d7938-ede2-499c-a726-260150cd37d0): create external table if not exists train_external (device_id string, gender string, age int, group_train str
ing) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014163950_f17d7938-ede2-499c-a726-260150cd37d0); Time taken: 0.021 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014163950_f17d7938-ede2-499c-a726-260150cd37d0): create external table if not exists train_external (device_id string, gender string, age int, group_train str
ing) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211014163950_f17d7938-ede2-499c-a726-260150cd37d0); Time taken: 0.035 seconds
INFO  : OK
No rows affected (0.076 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists brand_device_external (device_id string, phone_brand string, device_model string) row format delimited fields terminated by "
," lines terminated by "\n" stored as textfile;
INFO  : Compiling command(queryId=hive_20211014164003_23434659-a8a4-4e38-bd1a-5e7c022d1c4c): create external table if not exists brand_device_external (device_id string, phone_brand string, device_model
 string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164003_23434659-a8a4-4e38-bd1a-5e7c022d1c4c); Time taken: 0.021 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164003_23434659-a8a4-4e38-bd1a-5e7c022d1c4c): create external table if not exists brand_device_external (device_id string, phone_brand string, device_model
 string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211014164003_23434659-a8a4-4e38-bd1a-5e7c022d1c4c); Time taken: 0.058 seconds
INFO  : OK
No rows affected (0.1 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists events_external (event_id int, device_id string, event_time timestamp, latitude float, longitude float) row format delimited
fields terminated by "," lines terminated by "\n" stored as textfile;
INFO  : Compiling command(queryId=hive_20211014164011_278fe935-84b1-4194-b719-855680234db7): create external table if not exists events_external (event_id int, device_id string, event_time timestamp, la
titude float, longitude float) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164011_278fe935-84b1-4194-b719-855680234db7); Time taken: 0.019 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164011_278fe935-84b1-4194-b719-855680234db7): create external table if not exists events_external (event_id int, device_id string, event_time timestamp, la
titude float, longitude float) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211014164011_278fe935-84b1-4194-b719-855680234db7); Time taken: 0.038 seconds
INFO  : OK
No rows affected (0.07 seconds)
0: jdbc:hive2://localhost:10000/default>
```

```
0: jdbc:hive2://localhost:10000/default> create external table if not exists app_lables_external (app_id string, label_id int) row format delimited fields terminated by "," lines terminated by "\n" stor
ed as textfile TBLPROPERTIES("skip.header.line.count"="1");
INFO  : Compiling command(queryId=hive_20211014164036_866904cb-dd00-4a4e-a3e0-6a6a1035782d): create external table if not exists app_lables_external (app_id string, label_id int) row format delimited fi
elds terminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1")
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164036_866904cb-dd00-4a4e-a3e0-6a6a1035782d); Time taken: 0.02 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164036_866904cb-dd00-4a4e-a3e0-6a6a1035782d): create external table if not exists app_lables_external (app_id string, label_id int) row format delimited fi
elds terminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1")
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211014164036_866904cb-dd00-4a4e-a3e0-6a6a1035782d); Time taken: 0.032 seconds
INFO  : OK
No rows affected (0.063 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists label_categories_external (label_id int, category string) row format delimited fields terminated by "," lines terminated by "
\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1");
INFO  : Compiling command(queryId=hive_20211014164043_4ebb7307-5ee9-45ca-9e60-a1d20e1bcbb1): create external table if not exists label_categories_external (label_id int, category string) row format deli
mited fields terminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1")
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164043_4ebb7307-5ee9-45ca-9e60-a1d20e1bcbb1); Time taken: 0.018 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164043_4ebb7307-5ee9-45ca-9e60-a1d20e1bcbb1): create external table if not exists label_categories_external (label_id int, category string) row format deli
mited fields terminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1")
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20211014164043_4ebb7307-5ee9-45ca-9e60-a1d20e1bcbb1); Time taken: 0.041 seconds
INFO  : OK
No rows affected (0.07 seconds)
0: jdbc:hive2://localhost:10000/default>
```

```
0: jdbc:hive2://localhost:10000/default> load data inpath '/user/hadoop/mlctest/app_events' into table app_events_external;
INFO  : Compiling command(queryId=hive_20211014164654_b6320d29-e322-4671-b23a-fadf299dcc77): load data inpath '/user/hadoop/mlctest/app_events' into table app_events_external
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164654_b6320d29-e322-4671-b23a-fadf299dcc77); Time taken: 0.035 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164654_b6320d29-e322-4671-b23a-fadf299dcc77): load data inpath '/user/hadoop/mlctest/app_events' into table app_events_external
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mlctest.app_events_external from hdfs://ip-172-31-75-0.ec2.internal:8020/user/hadoop/mlctest/app_events
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20211014164654_b6320d29-e322-4671-b23a-fadf299dcc77); Time taken: 0.407 seconds
INFO  : OK
No rows affected (0.452 seconds)
0: jdbc:hive2://localhost:10000/default> select * from app_events_external limit 5;
INFO  : Compiling command(queryId=hive_20211014164707_1497d466-3044-4a1c-ae62-4bf15f24619c): select * from app_events_external limit 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:app_events_external.event_id, type:int, comment:null), FieldSchema(name:app_events_external.app_id, type:string, comment:null), Field
Schema(name:app_events_external.is_installed, type:int, comment:null), FieldSchema(name:app_events_external.is_active, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164707_1497d466-3044-4a1c-ae62-4bf15f24619c); Time taken: 0.155 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164707_1497d466-3044-4a1c-ae62-4bf15f24619c): select * from app_events_external limit 5
INFO  : Completed executing command(queryId=hive_20211014164707_1497d466-3044-4a1c-ae62-4bf15f24619c); Time taken: 0.001 seconds
INFO  : OK
+-----------------------------+---------------------------+---------------------------------+-------------------------------+
| app_events_external.event_id  | app_events_external.app_id  | app_events_external.is_installed  | app_events_external.is_active  |
+-----------------------------+---------------------------+---------------------------------+-------------------------------+
| 2                           | 5927333115845830913       | 1                               | 1                             |
| 2                           | -5720078949152207372      | 1                               | 0                             |
| 2                           | -1633887856876571208      | 1                               | 0                             |
| 2                           | -653184325010919369       | 1                               | 1                             |
| 2                           | 8693964245073640147       | 1                               | 1                             |
+-----------------------------+---------------------------+---------------------------------+-------------------------------+
5 rows selected (0.19 seconds)
0: jdbc:hive2://localhost:10000/default>
```

2.2 Loading data from HDFS to HIVE Tables and validation

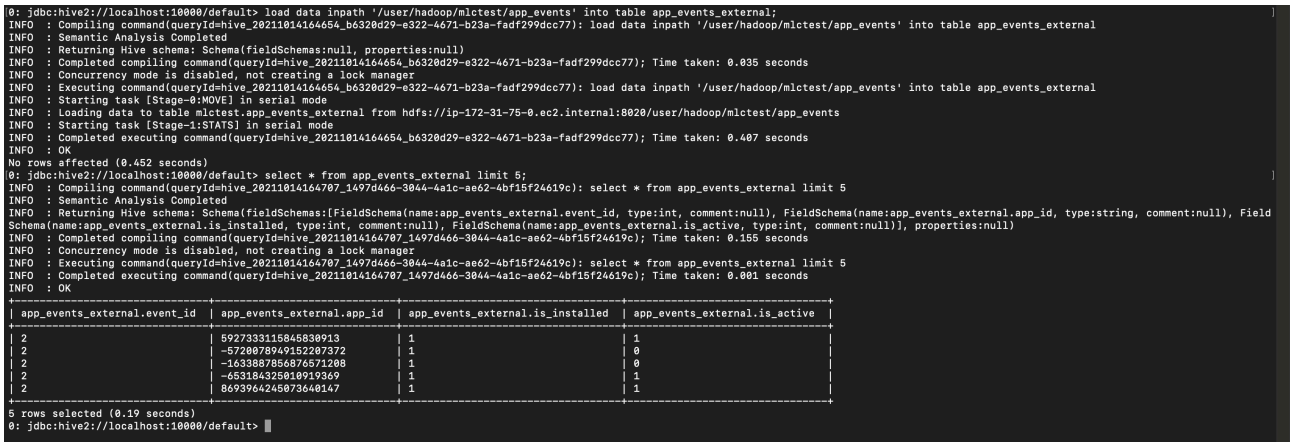
2.2.1 HIVE Table: app_events_external

The following command will load the data from HDFS and store in the HIVE table

load data inpath '/user/17adoop/mlctest/app_events' into table app_events_external;

The following command will check the loaded data from HIVE table

*select * from app_events_external limit 5;*



```
0: jdbc:hive2://localhost:10000/default> load data inpath '/user/hadoop/mlctest/app_events' into table app_events_external;
INFO  : Compiling command(queryId=hive_20211014164654_b6320d29-e322-4671-b23a-fadf299dcc77): load data inpath '/user/hadoop/mlctest/app_events' into table app_events_external
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164654_b6320d29-e322-4671-b23a-fadf299dcc77); Time taken: 0.035 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164654_b6320d29-e322-4671-b23a-fadf299dcc77): load data inpath '/user/hadoop/mlctest/app_events' into table app_events_external
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mlctest.app_events_external from hdfs://ip-172-31-75-0.ec2.internal:8020/user/hadoop/mlctest/app_events
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20211014164654_b6320d29-e322-4671-b23a-fadf299dcc77); Time taken: 0.407 seconds
INFO  : OK
No rows affected (0.452 seconds)
0: jdbc:hive2://localhost:10000/default> select * from app_events_external limit 5;
INFO  : Compiling command(queryId=hive_20211014164707_1497d466-3044-4a1c-ae62-4bf15f24619c): select * from app_events_external limit 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:app_events_external.event_id, type:int, comment:null), FieldSchema(name:app_events_external.app_id, type:string, comment:null), Field
Schema(name:app_events_external.is_installed, type:int, comment:null), FieldSchema(name:app_events_external.is_active, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164707_1497d466-3044-4a1c-ae62-4bf15f24619c); Time taken: 0.155 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164707_1497d466-3044-4a1c-ae62-4bf15f24619c): select * from app_events_external limit 5
INFO  : Completed executing command(queryId=hive_20211014164707_1497d466-3044-4a1c-ae62-4bf15f24619c); Time taken: 0.001 seconds
INFO  : OK
+-------------------------------+---------------------------+---------------------------------+--------------------------------+
| app_events_external.event_id  | app_events_external.app_id  | app_events_external.is_installed  | app_events_external.is_active  |
+-------------------------------+---------------------------+---------------------------------+--------------------------------+
| 2                             | 5927333115845830913       | 1                               | 1                              |
| 2                             | -5720078949152207372      | 1                               | 0                              |
| 2                             | -1633887856876571208      | 1                               | 0                              |
| 2                             | -6531843250109193 69      | 1                               | 1                              |
| 2                             | 8693964245073640147       | 1                               | 1                              |
+-------------------------------+---------------------------+---------------------------------+--------------------------------+
5 rows selected (0.19 seconds)
0: jdbc:hive2://localhost:10000/default>
```

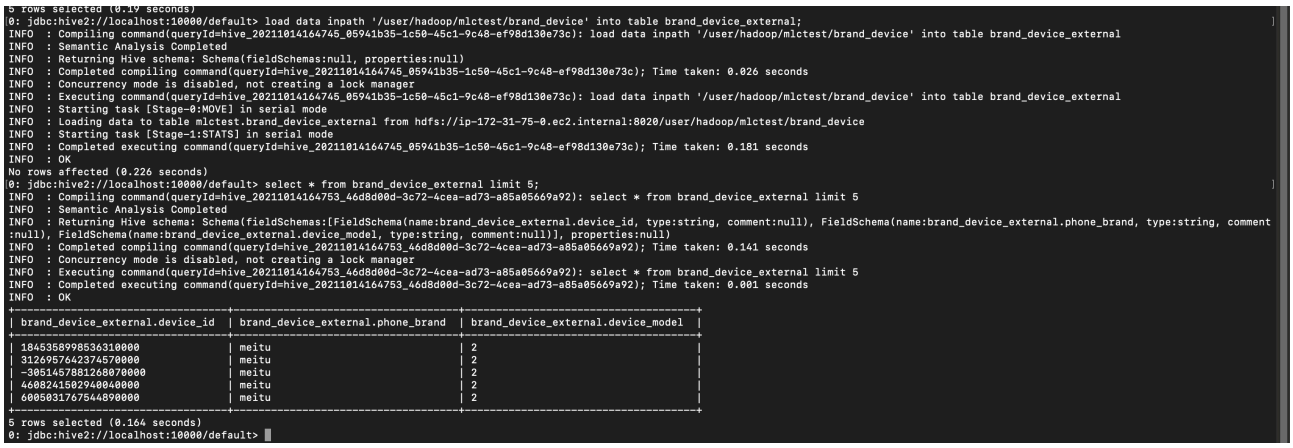
2.2.2 HIVE Table: brand_device_external

The following command will load the data from HDFS and store in the HIVE table

load data inpath '/user/hadoop/mlctest/brand_device' into table brand_device_external;

The following command will check the loaded data from HIVE table

*select * from brand_device_external limit 5;*



```
5 rows selected (0.19 seconds)
0: jdbc:hive2://localhost:10000/default> load data inpath '/user/hadoop/mlctest/brand_device' into table brand_device_external;
INFO  : Compiling command(queryId=hive_20211014164745_05941b35-1c50-45c1-9c48-ef98d130e73c): load data inpath '/user/hadoop/mlctest/brand_device' into table brand_device_external
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164745_05941b35-1c50-45c1-9c48-ef98d130e73c); Time taken: 0.026 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164745_05941b35-1c50-45c1-9c48-ef98d130e73c): load data inpath '/user/hadoop/mlctest/brand_device' into table brand_device_external
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mlctest.brand_device_external from hdfs://ip-172-31-75-0.ec2.internal:8020/user/hadoop/mlctest/brand_device
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20211014164745_05941b35-1c50-45c1-9c48-ef98d130e73c); Time taken: 0.181 seconds
INFO  : OK
No rows affected (0.226 seconds)
0: jdbc:hive2://localhost:10000/default> select * from brand_device_external limit 5;
INFO  : Compiling command(queryId=hive_20211014164753_46d8d00d-3c72-4cea-ad73-a85a05669a92): select * from brand_device_external limit 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:brand_device_external.device_id, type:string, comment:null), FieldSchema(name:brand_device_external.phone_brand, type:string, comment
:null), FieldSchema(name:brand_device_external.device_model, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164753_46d8d00d-3c72-4cea-ad73-a85a05669a92); Time taken: 0.141 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164753_46d8d00d-3c72-4cea-ad73-a85a05669a92): select * from brand_device_external limit 5
INFO  : Completed executing command(queryId=hive_20211014164753_46d8d00d-3c72-4cea-ad73-a85a05669a92); Time taken: 0.001 seconds
INFO  : OK
+-------------------------------+---------------------------------+---------------------------------+
| brand_device_external.device_id  | brand_device_external.phone_brand  | brand_device_external.device_model  |
+-------------------------------+---------------------------------+---------------------------------+
| 1845358998536310000           | meitu                           | 2                               |
| 3126957642374570000           | meitu                           | 2                               |
| -3051457881268070000          | meitu                           | 2                               |
| 460824150294040000            | meitu                           | 2                               |
| 600503176754490000            | meitu                           | 2                               |
+-------------------------------+---------------------------------+---------------------------------+
5 rows selected (0.164 seconds)
0: jdbc:hive2://localhost:10000/default>
```

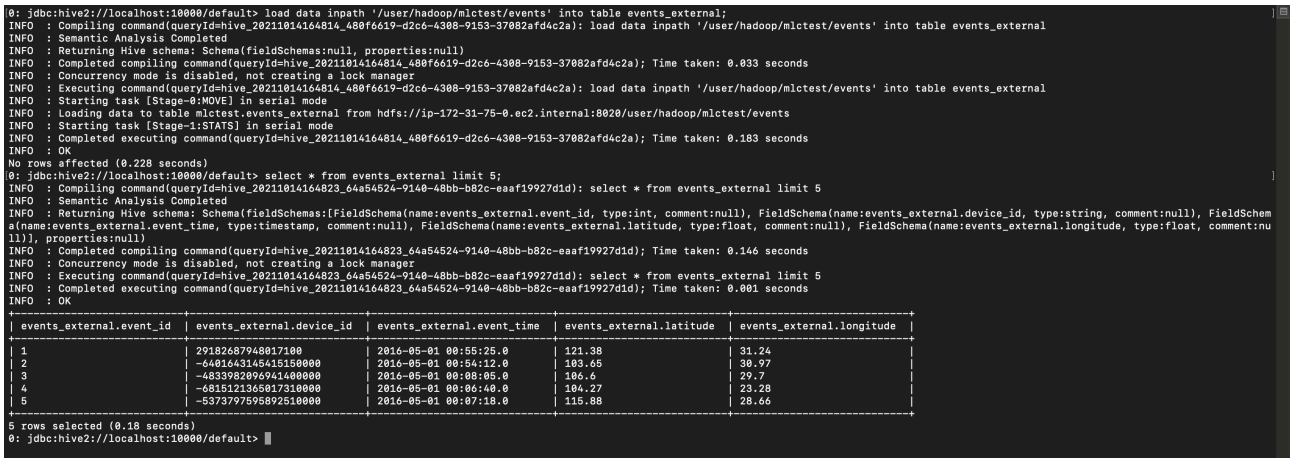
2.2.3 HIVE Table: events_external

The following command will load the data from HDFS and store in the HIVE table

load data inpath '/user/hadoop/mlctest/events' into table events_external;

The following command will check the loaded data from HIVE table

*select * from events_external limit 5;*



```
0: jdbc:hive2://localhost:10000/default> load data inpath '/user/hadoop/mlctest/events' into table events_external;
INFO  : Compiling command(queryId=hive_20211014164814_480f6619-d2c6-4308-9153-37082afd4c2a): load data inpath '/user/hadoop/mlctest/events' into table events_external
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164814_480f6619-d2c6-4308-9153-37082afd4c2a); Time taken: 0.033 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164814_480f6619-d2c6-4308-9153-37082afd4c2a): load data inpath '/user/hadoop/mlctest/events' into table events_external
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mlctest.events_external from hdfs://ip-172-31-75-0.ec2.internal:8020/user/hadoop/mlctest/events
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20211014164814_480f6619-d2c6-4308-9153-37082afd4c2a); Time taken: 0.183 seconds
INFO  : OK
No rows affected (0.228 seconds)
0: jdbc:hive2://localhost:10000/default> select * from events_external limit 5;
INFO  : Compiling command(queryId=hive_20211014164823_64a54524-9140-48bb-b82c-eaaf19927d1d): select * from events_external limit 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:events_external.event_id, type:int, comment:null), FieldSchema(name:events_external.device_id, type:string, comment:null), FieldSchem
a(name:events_external.event_time, type:timestamp, comment:null), FieldSchema(name:events_external.latitude, type:float, comment:null), FieldSchema(name:events_external.longitude, type:float, comment:nu
ll)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164823_64a54524-9140-48bb-b82c-eaaf19927d1d); Time taken: 0.146 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164823_64a54524-9140-48bb-b82c-eaaf19927d1d): select * from events_external limit 5
INFO  : Completed executing command(queryId=hive_20211014164823_64a54524-9140-48bb-b82c-eaaf19927d1d); Time taken: 0.001 seconds
INFO  : OK
+--------------------------+-------------------------+---------------------------+--------------------------+--------------------------+
| events_external.event_id | events_external.device_id | events_external.event_time | events_external.latitude | events_external.longitude |
+--------------------------+-------------------------+---------------------------+--------------------------+--------------------------+
| 1                        | 29182687948017100       | 2016-05-01 00:55:25.0     | 121.38                   | 31.24                    |
| 2                        | -64016431454151680000   | 2016-05-01 00:54:12.0     | 103.65                   | 30.97                    |
| 3                        | -48339820969414000000   | 2016-05-01 00:08:05.0     | 106.6                    | 29.7                     |
| 4                        | -68151213650173100000   | 2016-05-01 00:06:40.0     | 104.27                   | 23.28                    |
| 5                        | -53737975958925100000   | 2016-05-01 00:07:18.0     | 115.88                   | 28.66                    |
+--------------------------+-------------------------+---------------------------+--------------------------+--------------------------+
5 rows selected (0.18 seconds)
0: jdbc:hive2://localhost:10000/default>
```

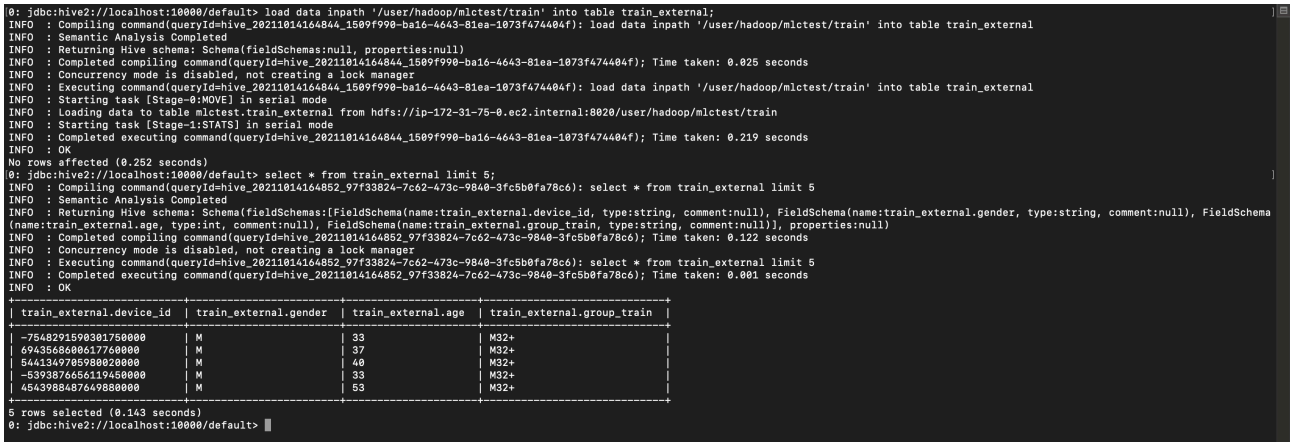
2.2.4 HIVE Table: train_external

The following command will load the data from HDFS and store in the HIVE table

load data inpath '/user/hadoop/mlctest/train' into table train_external;

The following command will check the loaded data from HIVE table

*select * from train_external limit 5;*



```
0: jdbc:hive2://localhost:10000/default> load data inpath '/user/hadoop/mlctest/train' into table train_external;
INFO  : Compiling command(queryId=hive_20211014164844_1509f990-ba16-4643-81ea-1073f474404f): load data inpath '/user/hadoop/mlctest/train' into table train_external
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164844_1509f990-ba16-4643-81ea-1073f474404f); Time taken: 0.025 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164844_1509f990-ba16-4643-81ea-1073f474404f): load data inpath '/user/hadoop/mlctest/train' into table train_external
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mlctest.train_external from hdfs://ip-172-31-75-0.ec2.internal:8020/user/hadoop/mlctest/train
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20211014164844_1509f990-ba16-4643-81ea-1073f474404f); Time taken: 0.219 seconds
INFO  : OK
No rows affected (0.252 seconds)
0: jdbc:hive2://localhost:10000/default> select * from train_external limit 5;
INFO  : Compiling command(queryId=hive_20211014164852_97f33824-7c62-473c-9840-3fc5b0fa78c6): select * from train_external limit 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:train_external.device_id, type:string, comment:null), FieldSchema(name:train_external.gender, type:string, comment:null), FieldSchema
(name:train_external.age, type:int, comment:null), FieldSchema(name:train_external.group_train, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014164852_97f33824-7c62-473c-9840-3fc5b0fa78c6); Time taken: 0.122 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014164852_97f33824-7c62-473c-9840-3fc5b0fa78c6): select * from train_external limit 5
INFO  : Completed executing command(queryId=hive_20211014164852_97f33824-7c62-473c-9840-3fc5b0fa78c6); Time taken: 0.001 seconds
INFO  : OK
+--------------------------+-------------------------+-------------------------+-----------------------------+
| train_external.device_id | train_external.gender | train_external.age | train_external.group_train |
+--------------------------+-------------------------+-------------------------+-----------------------------+
| -754829159030175000      | M                       | 33                      | M32+                        |
| 6943568608617760000      | M                       | 37                      | M32+                        |
| 5441349705980280000      | M                       | 40                      | M32+                        |
| -5393876656119450000     | M                       | 33                      | M32+                        |
| 454398487649880000       | M                       | 53                      | M32+                        |
+--------------------------+-------------------------+-------------------------+-----------------------------+
5 rows selected (0.143 seconds)
0: jdbc:hive2://localhost:10000/default>
```

2.3 Loading data from CSV to HIVE Tables and validation

2.3.1 HIVE Table: app_labels_external

The following command will load the data from local path which was downloaded from S3 and store in the HIVE table

load data local inpath '/home/hadoop/app_labels_new.txt' into table app_lables_external;

The following command will check the loaded data from HIVE table

*select * from app_lables_external limit 5;*

```
6 rows selected (0.04 seconds)
0: jdbc:hive2://localhost:10000/default> load data local inpath '/home/hadoop/app_labels_new.txt' into table app_labels_external;
INFO  : Compiling command(queryId=hive_20211014165818_ccbc5792-4ef4-4d02-a0a8-b8e067708696): load data local inpath '/home/hadoop/app_labels_new.txt' into table app_labels_external
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014165818_ccbc5792-4ef4-4d02-a0a8-b8e067708696); Time taken: 0.019 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014165818_ccbc5792-4ef4-4d02-a0a8-b8e067708696): load data local inpath '/home/hadoop/app_labels_new.txt' into table app_labels_external
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mlctest.app_labels_external from file:/home/hadoop/app_labels_new.txt
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20211014165818_ccbc5792-4ef4-4d02-a0a8-b8e067708696); Time taken: 0.303 seconds
INFO  : OK
No rows affected (0.333 seconds)
0: jdbc:hive2://localhost:10000/default> select * from app_lables_external limit 5;
Error: Error while compiling statement: FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'app_lables_external' (state=42S02,code=10001)
0: jdbc:hive2://localhost:10000/default> select * from app_labels_external limit 5;
INFO  : Compiling command(queryId=hive_20211014165840_4135a9af-d715-43ce-b594-8aec368a8b70): select * from app_labels_external limit 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:app_labels_external.app_id, type:string, comment:null), FieldSchema(name:app_labels_external.label_id, type:int, comment:null)], prop
erties:null)
INFO  : Completed compiling command(queryId=hive_20211014165840_4135a9af-d715-43ce-b594-8aec368a8b70); Time taken: 0.104 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014165840_4135a9af-d715-43ce-b594-8aec368a8b70): select * from app_labels_external limit 5
INFO  : Completed executing command(queryId=hive_20211014165840_4135a9af-d715-43ce-b594-8aec368a8b70); Time taken: 0.001 seconds
INFO  : OK
+-----------------------------+-------------------------------+
| app_labels_external.app_id  | app_labels_external.label_id  |
+-----------------------------+-------------------------------+
| 7324884708820027918         | 251                           |
| -4494216993218550286        | 251                           |
| 6058196446775239644         | 406                           |
| 6058196446775239644         | 407                           |
| 8694625920731541625         | 406                           |
+-----------------------------+-------------------------------+
5 rows selected (0.124 seconds)
0: jdbc:hive2://localhost:10000/default>
```

2.3.2 HIVE Table: label_categories_external

The following command will load the data from local path which was downloaded from S3 and store in the HIVE table

load data local inpath '/home/hadoop/label_categories.csv' into table label_categories_external;

The following command will check the loaded data from HIVE table

*select * from label_categories_external limit 5;*

```
0: jdbc:hive2://localhost:10000/default> load data local inpath '/home/hadoop/label_categories.csv' into table label_categories_external;
INFO  : Compiling command(queryId=hive_20211014165900_b9a029d3-ac9a-4fd1-b626-e8d0ddb1cd2d): load data local inpath '/home/hadoop/label_categories.csv' into table label_categories_external
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20211014165900_b9a029d3-ac9a-4fd1-b626-e8d0ddb1cd2d); Time taken: 0.018 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014165900_b9a029d3-ac9a-4fd1-b626-e8d0ddb1cd2d): load data local inpath '/home/hadoop/label_categories.csv' into table label_categories_external
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mlctest.label_categories_external from file:/home/hadoop/label_categories.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20211014165900_b9a029d3-ac9a-4fd1-b626-e8d0ddb1cd2d); Time taken: 0.193 seconds
INFO  : OK
No rows affected (0.22 seconds)
0: jdbc:hive2://localhost:10000/default> select * from label_categories_external limit 5;
INFO  : Compiling command(queryId=hive_20211014165906_eb9250f8-8524-4a6b-a19d-5bc609adecd6): select * from label_categories_external limit 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:label_categories_external.label_id, type:int, comment:null), FieldSchema(name:label_categories_external.category, type:string, commen
t:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014165906_eb9250f8-8524-4a6b-a19d-5bc609adecd6); Time taken: 0.106 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014165906_eb9250f8-8524-4a6b-a19d-5bc609adecd6): select * from label_categories_external limit 5
INFO  : Completed executing command(queryId=hive_20211014165906_eb9250f8-8524-4a6b-a19d-5bc609adecd6); Time taken: 0.001 seconds
INFO  : OK
+-----------------------------------+-----------------------------------+
| label_categories_external.label_id  | label_categories_external.category  |
+-----------------------------------+-----------------------------------+
| 1                                 |                                   |
| 2                                 | game-game type                    |
| 3                                 | game-Game themes                  |
| 4                                 | game-Art Style                    |
| 5                                 | game-Leisure time                 |
+-----------------------------------+-----------------------------------+
5 rows selected (0.129 seconds)
0: jdbc:hive2://localhost:10000/default>
```

3. HQL Tasks

3.1 Top 10 most popular brands and respective % for Male and Female in it ?

3.1.1 Query Execution

```
SELECT b.phone_brand AS Phone_Brand,
Count(*) AS Total,
Sum(CASE t.gender
WHEN 'M' THEN 1
ELSE 0
end) * 100 / Count(*) AS male_pct,
Sum(CASE t.gender
WHEN 'F' THEN 1
ELSE 0
end) * 100 / Count(*) AS female_pct
FROM (SELECT *
FROM train_external) t
JOIN (SELECT DISTINCT( device_id ),
phone_brand
FROM brand_device_external) b
ON t.device_id = b.device_id
GROUP BY b.phone_brand
ORDER BY total DESC
LIMIT 10;
```

```
+-----------------------------------+-----------------------------------+
5 rows selected (0.129 seconds)
0: jdbc:hive2://localhost:10000/default> SELECT b.phone_brand AS Phone_Brand,
. . . . . . . . . . . . . . . . . . . . .>        Count(*)           AS Total,
. . . . . . . . . . . . . . . . . . . . .>        Sum(CASE t.gender
. . . . . . . . . . . . . . . . . . . . .>             WHEN 'M' THEN 1
. . . . . . . . . . . . . . . . . . . . .>             ELSE 0
. . . . . . . . . . . . . . . . . . . . .>            end) * 100 / Count(*) AS male_pct,
. . . . . . . . . . . . . . . . . . . . .>        Sum(CASE t.gender
. . . . . . . . . . . . . . . . . . . . .>             WHEN 'F' THEN 1
. . . . . . . . . . . . . . . . . . . . .>             ELSE 0
. . . . . . . . . . . . . . . . . . . . .>            end) * 100 / Count(*) AS female_pct
. . . . . . . . . . . . . . . . . . . . .> FROM   (SELECT *
. . . . . . . . . . . . . . . . . . . . .>         FROM   train_external) t
. . . . . . . . . . . . . . . . . . . . .>        JOIN (SELECT DISTINCT( device_id ),
. . . . . . . . . . . . . . . . . . . . .>                     phone_brand
. . . . . . . . . . . . . . . . . . . . .>              FROM   brand_device_external) b
. . . . . . . . . . . . . . . . . . . . .>          ON t.device_id = b.device_id
. . . . . . . . . . . . . . . . . . . . .> GROUP  BY b.phone_brand
. . . . . . . . . . . . . . . . . . . . .> ORDER  BY total DESC
. . . . . . . . . . . . . . . . . . . . .> LIMIT  10;
INFO  : Compiling command(queryId=hive_20211014165929_f190fb3a-4454-40f6-8b86-c78b31b5a63c): SELECT b.phone_brand AS Phone_Brand,
Count(*)           AS Total,
Sum(CASE t.gender
WHEN 'M' THEN 1
ELSE 0
end) * 100 / Count(*) AS male_pct,
Sum(CASE t.gender
WHEN 'F' THEN 1
ELSE 0
end) * 100 / Count(*) AS female_pct
FROM   (SELECT *
FROM   train_external) t
JOIN (SELECT DISTINCT( device_id ),
phone_brand
FROM   brand_device_external) b
ON t.device_id = b.device_id
GROUP  BY b.phone_brand
```


```
+-----------------------------------+-----------------------------------+
5 rows selected (0.129 seconds)
0: jdbc:hive2://localhost:10000/default> SELECT b.phone_brand AS Phone_Brand,
. . . . . . . . . . . . . . . . . . . . .>        Count(*)           AS Total,
. . . . . . . . . . . . . . . . . . . . .>        Sum(CASE t.gender
. . . . . . . . . . . . . . . . . . . . .>             WHEN 'M' THEN 1
. . . . . . . . . . . . . . . . . . . . .>             ELSE 0
. . . . . . . . . . . . . . . . . . . . .>            end) * 100 / Count(*) AS male_pct,
. . . . . . . . . . . . . . . . . . . . .>        Sum(CASE t.gender
. . . . . . . . . . . . . . . . . . . . .>             WHEN 'F' THEN 1
. . . . . . . . . . . . . . . . . . . . .>             ELSE 0
. . . . . . . . . . . . . . . . . . . . .>            end) * 100 / Count(*) AS female_pct
. . . . . . . . . . . . . . . . . . . . .> FROM   (SELECT *
. . . . . . . . . . . . . . . . . . . . .>         FROM   train_external) t
. . . . . . . . . . . . . . . . . . . . .>        JOIN (SELECT DISTINCT( device_id ),
. . . . . . . . . . . . . . . . . . . . .>                     phone_brand
. . . . . . . . . . . . . . . . . . . . .>              FROM   brand_device_external) b
. . . . . . . . . . . . . . . . . . . . .>          ON t.device_id = b.device_id
. . . . . . . . . . . . . . . . . . . . .> GROUP  BY b.phone_brand
. . . . . . . . . . . . . . . . . . . . .> ORDER  BY total DESC
. . . . . . . . . . . . . . . . . . . . .> LIMIT  10;
INFO  : Compiling command(queryId=hive_20211014165929_f190fb3a-4454-40f6-8b86-c78b31b5a63c): SELECT b.phone_brand AS Phone_Brand,
Count(*)           AS Total,
Sum(CASE t.gender
WHEN 'M' THEN 1
ELSE 0
end) * 100 / Count(*) AS male_pct,
Sum(CASE t.gender
WHEN 'F' THEN 1
ELSE 0
end) * 100 / Count(*) AS female_pct
FROM   (SELECT *
FROM   train_external) t
JOIN (SELECT DISTINCT( device_id ),
phone_brand
FROM   brand_device_external) b
ON t.device_id = b.device_id
GROUP  BY b.phone_brand
```

3.2 Top 10 most popular brands for Male and Female ?

3.2.1 Query Execution – Popular Brands for Male

```
SELECT b.phone_brand as Phone_Brand,
Count(*) AS Total,
t.gender as Gender
FROM (SELECT *
FROM train_external
WHERE gender = 'M') t
JOIN (SELECT DISTINCT( device_id ),
phone_brand
FROM brand_device_external) b
ON t.device_id = b.device_id
GROUP BY b.phone_brand, t.gender
ORDER BY total DESC
LIMIT 10;
```

```
INFO  : Query ID = hive_20211014165929_f190fb3a-4454-40f6-8b86-c78b31b5a63c
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: SELECT b.phone_brand AS Phone_Brand,
Co...10(Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Status: Running (Executing on YARN cluster with App id application_1634228715435_0005)

INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/2  Reducer 4: 0/2  Reducer 5: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0/1      Reducer 3: 0/2  Reducer 4: 0/2  Reducer 5: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/2  Reducer 4: 0/2  Reducer 5: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/2  Reducer 4: 0/2  Reducer 5: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+2)/2      Reducer 4: 0/2  Reducer 5: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1(+1)/2      Reducer 4: 0(+1)/2      Reducer 5: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 2/2  Reducer 4: 0(+2)/2      Reducer 5: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 2/2  Reducer 4: 2/2  Reducer 5: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 2/2  Reducer 4: 2/2  Reducer 5: 1/1
INFO  : Completed executing command(queryId=hive_20211014165929_f190fb3a-4454-40f6-8b86-c78b31b5a63c); Time taken: 14.644 seconds
INFO  : OK
+--------------+--------+--------------------+--------------------+
| phone_brand  | total  |      male_pct      |     female_pct     |
+--------------+--------+--------------------+--------------------+
| Xiaomi       | 17300  | 65.79190751445087  | 34.20809248554913  |
| samsung      | 13669  | 60.26775916306972  | 39.73224083693028  |
| Huawei       | 12960  | 67.25308641975309  | 32.74691358024691  |
| OPPO         | 5783   | 55.54210617326647  | 44.45789382673353  |
| vivo         | 5637   | 52.97143870853291  | 47.02856129146709  |
| Meizu        | 4699   | 72.29197701638647  | 27.708022983613535 |
| Coolpad      | 3339   | 67.6849356094639   | 32.31506439053609  |
| lenovo       | 2691   | 66.81531029357116  | 33.184689706428834 |
| Gionee       | 1123   | 64.20302760463045  | 35.796972395369544 |
| HTC          | 1013   | 68.4186614017769   | 31.5893385982231   |
+--------------+--------+--------------------+--------------------+
10 rows selected (15.32 seconds)
0: jdbc:hive2://localhost:10000/default>
```

```
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:phone_brand, type:string, comment:null), FieldSchema(name:total, type:bigint, comment:null), FieldSchema(name:gender, type:string, co
mment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014170024_a1c996d8-d000-43a4-873b-9122dc2efe5c); Time taken: 0.193 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014170024_a1c996d8-d000-43a4-873b-9122dc2efe5c): SELECT b.phone_brand as Phone_Brand,
Count(*) AS Total,
t.gender as Gender
FROM (SELECT *
FROM train_external
WHERE gender = 'M') t
JOIN (SELECT DISTINCT( device_id ),
phone_brand
FROM brand_device_external) b
ON t.device_id = b.device_id
GROUP BY b.phone_brand, t.gender
ORDER BY total DESC
LIMIT 10
INFO  : Query ID = hive_20211014170024_a1c996d8-d000-43a4-873b-9122dc2efe5c
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: SELECT b.phone_brand as Phone_Brand,
Co...10(Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Status: Running (Executing on YARN cluster with App id application_1634228715435_0005)

INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/2  Reducer 4: 0/2  Reducer 5: 0/1
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/2  Reducer 4: 0/2  Reducer 5: 0/1
INFO  : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/2  Reducer 4: 0/2  Reducer 5: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+2)/2      Reducer 4: 0/2  Reducer 5: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 2/2  Reducer 4: 0(+2)/2      Reducer 5: 0/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 2/2  Reducer 4: 2/2  Reducer 5: 0(+1)/1
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 2/2  Reducer 4: 2/2  Reducer 5: 1/1
INFO  : Completed executing command(queryId=hive_20211014170024_a1c996d8-d000-43a4-873b-9122dc2efe5c); Time taken: 6.035 seconds
INFO  : OK
+--------------+--------+---------+
| phone_brand  | total  | gender  |
+--------------+--------+---------+
| Xiaomi       | 11382  | M       |
| Huawei       | 8716   | M       |
| samsung      | 8238   | M       |
| Meizu        | 3397   | M       |
| OPPO         | 3212   | M       |
| vivo         | 2986   | M       |
| Coolpad      | 2260   | M       |
| lenovo       | 1798   | M       |
| Gionee       | 721    | M       |
| HTC          | 693    | M       |
+--------------+--------+---------+
10 rows selected (6.25 seconds)
0: jdbc:hive2://localhost:10000/default>
```

3.2.3 Query Execution – Popular Brands for Female

```
SELECT b.phone_brand AS Phone_Brand,
Count(*) AS Total,
t.gender AS Gender
FROM (SELECT *
FROM train_external
WHERE gender = 'F') t
JOIN (SELECT DISTINCT( device_id ),
phone_brand
FROM brand_device_external) b
ON t.device_id = b.device_id
GROUP BY b.phone_brand,
t.gender
ORDER BY total DESC
LIMIT 10;
```

```
0: jdbc:hive2://localhost:10000/default> SELECT b.phone_brand AS Phone_Brand,
. . . . . .> Count(*) AS Total,
. . . . . .> t.gender AS Gender
. . . . . .> FROM (SELECT *
. . . . . .> FROM train_external
. . . . . .> WHERE gender = 'F') t
. . . . . .> JOIN (SELECT DISTINCT( device_id ),
. . . . . .> phone_brand
. . . . . .> FROM brand_device_external) b
. . . . . .> ON t.device_id = b.device_id
. . . . . .> GROUP BY b.phone_brand,
. . . . . .> t.gender
. . . . . .> ORDER BY total DESC
. . . . . .> LIMIT 10;
INFO  : Compiling command(queryId=hive_20211014170102_34983817-024d-4939-94e6-4ba4daec8744): SELECT b.phone_brand AS Phone_Brand,
Count(*) AS Total,
t.gender AS Gender
FROM (SELECT *
FROM train_external
WHERE gender = 'F') t
JOIN (SELECT DISTINCT( device_id ),
phone_brand
FROM brand_device_external) b
ON t.device_id = b.device_id
GROUP BY b.phone_brand,
t.gender
ORDER BY total DESC
LIMIT 10
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:phone_brand, type:string, comment:null), FieldSchema(name:total, type:bigint, comment:null), FieldSchema(name:gender, type:string, co
mment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014170102_34983817-024d-4939-94e6-4ba4daec8744); Time taken: 0.246 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014170102_34983817-024d-4939-94e6-4ba4daec8744): SELECT b.phone_brand AS Phone_Brand,
Count(*) AS Total,
t.gender AS Gender
FROM (SELECT *
FROM train_external
WHERE gender = 'F') t
JOIN (SELECT DISTINCT( device_id ),
phone_brand
FROM brand_device_external) b
ON t.device_id = b.device_id
GROUP BY b.phone_brand,
t.gender
ORDER BY total DESC
LIMIT 10
```

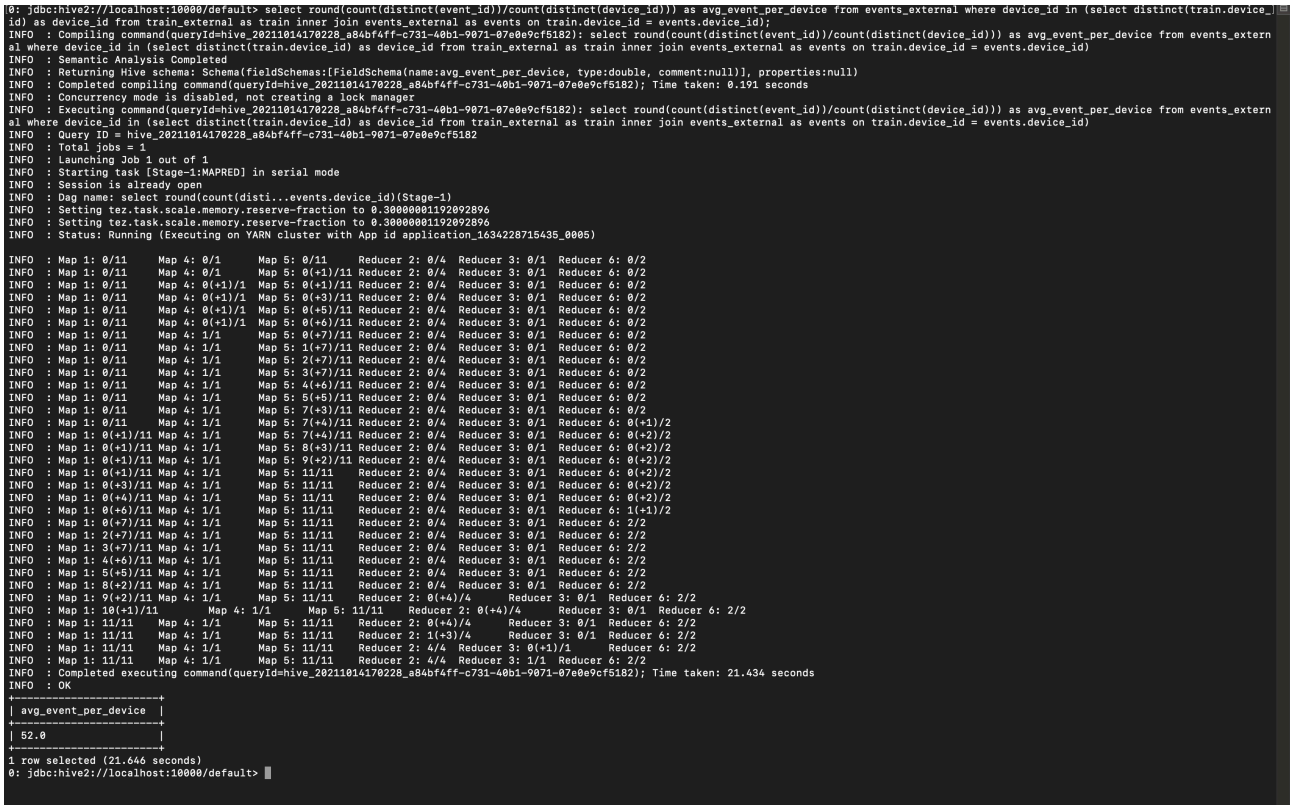
```
0: jdbc:hive2://localhost:10000/default> SELECT SUM(IF(gender = 'M', 1, 0)) AS male_count,
. . . . . .> Round(( SUM(IF(gender = 'M', 1, 0)) / Count(1) ) * 100, 2)
. . . . . .> ||'%' AS male_ratio,
. . . . . .> SUM(IF(gender = 'F', 1, 0)) AS female_count,
. . . . . .> Round(( SUM(IF(gender = 'F', 1, 0)) / Count(1) ) * 100, 2)
. . . . . .> ||'%' AS female_ratio
. . . . . .> FROM train_external;
INFO  : Compiling command(queryId=hive_20211014170141_363a8dd2-f15c-41e5-a8a9-ef87108bcbe7): SELECT SUM(IF(gender = 'M', 1, 0)) AS male_count,
Round(( SUM(IF(gender = 'M', 1, 0)) / Count(1) ) * 100, 2)
||'%' AS male_ratio,
SUM(IF(gender = 'F', 1, 0)) AS female_count,
Round(( SUM(IF(gender = 'F', 1, 0)) / Count(1) ) * 100, 2)
||'%' AS female_ratio
FROM train_external
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:male_count, type:bigint, comment:null), FieldSchema(name:male_ratio, type:string, comment:null), FieldSchema(name:female_count, type:
bigint, comment:null), FieldSchema(name:female_ratio, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014170141_363a8dd2-f15c-41e5-a8a9-ef87108bcbe7); Time taken: 0.118 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014170141_363a8dd2-f15c-41e5-a8a9-ef87108bcbe7): SELECT SUM(IF(gender = 'M', 1, 0)) AS male_count,
Round(( SUM(IF(gender = 'M', 1, 0)) / Count(1) ) * 100, 2)
||'%' AS male_ratio,
SUM(IF(gender = 'F', 1, 0)) AS female_count,
Round(( SUM(IF(gender = 'F', 1, 0)) / Count(1) ) * 100, 2)
||'%' AS female_ratio
FROM train_external
INFO  : Query ID = hive_20211014170141_363a8dd2-f15c-41e5-a8a9-ef87108bcbe7
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: SELECT SUM(IF(gender = 'M',...train_external(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1634228715435_0005)

INFO  : Map 1: 0/1      Reducer 2: 0/1
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/1
INFO  : Map 1: 1/1      Reducer 2: 0(+1)/1
INFO  : Map 1: 1/1      Reducer 2: 1/1
INFO  : Completed executing command(queryId=hive_20211014170141_363a8dd2-f15c-41e5-a8a9-ef87108bcbe7); Time taken: 4.89 seconds
INFO  : OK
+-------------+-------------+---------------+---------------+
| male_count  | male_ratio  | female_count  | female_ratio  |
+-------------+-------------+---------------+---------------+
| 47904       | 64.18%      | 26741         | 35.82%        |
+-------------+-------------+---------------+---------------+
1 row selected (5.025 seconds)
0: jdbc:hive2://localhost:10000/default>
```

3.3 Count and percentage Analysis of the Gender in the train Dataset

3.3.1 Query Execution and Output

```
SELECT SUM(IF(gender = 'M', 1, 0)) AS male_count,
Round(( SUM(IF(gender = 'M', 1, 0)) / Count(1) ) * 100, 2)
ll'%' AS male_ratio,
SUM(IF(gender = 'F', 1, 0)) AS female_count,
Round(( SUM(IF(gender = 'F', 1, 0)) / Count(1) ) * 100, 2)
ll'%' AS female_ratio
FROM train_external;
```



3.4 Top 3 mobile phone brands offering the highest number of models

3.4.1 Query Execution and Output

select phone_brand, count(device_model) as model_count from brand_device_external group by phone_brand order by model_count desc limit 3;

```
0: jdbc:hive2://localhost:10000/default> select phone_brand, count(device_model) as model_count from brand_device_external group by phone_brand order by model_count desc limit 3;
INFO  : Compiling command(queryId=hive_20211014170205_41163fd3-b00c-4d9f-ae96-e7a1a94a8a4f): select phone_brand, count(device_model) as model_count from brand_device_external group by phone_brand order
by model_count desc limit 3
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:phone_brand, type:string, comment:null), FieldSchema(name:model_count, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014170205_41163fd3-b00c-4d9f-ae96-e7a1a94a8a4f); Time taken: 0.092 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014170205_41163fd3-b00c-4d9f-ae96-e7a1a94a8a4f): select phone_brand, count(device_model) as model_count from brand_device_external group by phone_brand order
by model_count desc limit 3
INFO  : Query ID = hive_20211014170205_41163fd3-b00c-4d9f-ae96-e7a1a94a8a4f
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select phone_brand, count(device_model) ...3(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1634228715435_0005)

INFO  : Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/1
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2  Reducer 3: 0/1
INFO  : Map 1: 1/1      Reducer 2: 0(+1)/2      Reducer 3: 0/1
INFO  : Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 0(+1)/1
INFO  : Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 1/1
INFO  : Completed executing command(queryId=hive_20211014170205_41163fd3-b00c-4d9f-ae96-e7a1a94a8a4f); Time taken: 4.488 seconds
INFO  : OK
+--------------+--------------+
| phone_brand  | model_count  |
+--------------+--------------+
| Xiaomi       | 43210        |
| samsung      | 34286        |
| Huawei       | 32564        |
+--------------+--------------+
3 rows selected (4.596 seconds)
0: jdbc:hive2://localhost:10000/default>
```

3.5 Average number of events per device id

We have approached this task into two ways

- Overall Average events across device
- Average Events per device

3.5.1 Overall Average events across devices

select round(count(distinct(event_id))/count(distinct(device_id))) as avg_event_per_device from events_external where device_id in (select distinct(train.device_id) as device_id from train_external as train inner join events_external as events on train.device_id = events.device_id);

```
0: jdbc:hive2://localhost:10000/default> select round(count(distinct(event_id))/count(distinct(device_id))) as avg_event_per_device from events_external where device_id in (select distinct(train.device_
id) as device_id from train_external as train inner join events_external as events on train.device_id = events.device_id);
INFO  : Compiling command(queryId=hive_20211014170228_a84bf4ff-c731-40b1-9071-07e0e9cf5182): select round(count(distinct(event_id))/count(distinct(device_id))) as avg_event_per_device from events_extern
al where device_id in (select distinct(train.device_id) as device_id from train_external as train inner join events_external as events on train.device_id = events.device_id)
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:avg_event_per_device, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014170228_a84bf4ff-c731-40b1-9071-07e0e9cf5182); Time taken: 0.191 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014170228_a84bf4ff-c731-40b1-9071-07e0e9cf5182): select round(count(distinct(event_id))/count(distinct(device_id))) as avg_event_per_device from events_extern
al where device_id in (select distinct(train.device_id) as device_id from train_external as train inner join events_external as events on train.device_id = events.device_id)
INFO  : Query ID = hive_20211014170228_a84bf4ff-c731-40b1-9071-07e0e9cf5182
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select round(count(disti...events.device_id))(Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Status: Running (Executing on YARN cluster with App id application_1634228715435_0005)

INFO  : Map 1: 0/11     Map 4: 0/1      Map 5: 0/11     Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 0/1      Map 5: 0(+1)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 0(+1)/1  Map 5: 0(+1)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 0(+1)/1  Map 5: 0(+3)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 0(+1)/1  Map 5: 0(+5)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 0(+1)/1  Map 5: 0(+6)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 1/1      Map 5: 0(+7)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 1/1      Map 5: 1(+7)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 1/1      Map 5: 2(+7)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 1/1      Map 5: 3(+7)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 1/1      Map 5: 4(+6)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 1/1      Map 5: 5(+5)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 1/1      Map 5: 7(+3)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0/2
INFO  : Map 1: 0/11     Map 4: 1/1      Map 5: 7(+4)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0(+1)/2
INFO  : Map 1: 0(+1)/11 Map 4: 1/1      Map 5: 7(+4)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0(+2)/2
INFO  : Map 1: 0(+1)/11 Map 4: 1/1      Map 5: 8(+3)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0(+2)/2
INFO  : Map 1: 0(+1)/11 Map 4: 1/1      Map 5: 9(+2)/11 Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0(+2)/2
INFO  : Map 1: 0(+1)/11 Map 4: 1/1      Map 5: 11/11    Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0(+2)/2
INFO  : Map 1: 0(+3)/11 Map 4: 1/1      Map 5: 11/11    Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0(+2)/2
INFO  : Map 1: 0(+4)/11 Map 4: 1/1      Map 5: 11/11    Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 0(+2)/2
INFO  : Map 1: 0(+6)/11 Map 4: 1/1      Map 5: 11/11    Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 1(+1)/2
INFO  : Map 1: 0(+7)/11 Map 4: 1/1      Map 5: 11/11    Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 2/2
INFO  : Map 1: 2(+7)/11 Map 4: 1/1      Map 5: 11/11    Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 2/2
INFO  : Map 1: 3(+7)/11 Map 4: 1/1      Map 5: 11/11    Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 2/2
INFO  : Map 1: 4(+6)/11 Map 4: 1/1      Map 5: 11/11    Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 2/2
INFO  : Map 1: 5(+5)/11 Map 4: 1/1      Map 5: 11/11    Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 2/2
INFO  : Map 1: 8(+2)/11 Map 4: 1/1      Map 5: 11/11    Reducer 2: 0/4  Reducer 3: 0/1  Reducer 6: 2/2
INFO  : Map 1: 9(+2)/11 Map 4: 1/1      Map 5: 11/11    Reducer 2: 0(+4)/4      Reducer 3: 0/1  Reducer 6: 2/2
INFO  : Map 1: 10(+1)/11        Map 4: 1/1      Map 5: 11/11    Reducer 2: 0(+4)/4      Reducer 3: 0/1  Reducer 6: 2/2
INFO  : Map 1: 11/11    Map 4: 1/1      Map 5: 11/11    Reducer 2: 0(+4)/4      Reducer 3: 0/1  Reducer 6: 2/2
INFO  : Map 1: 11/11    Map 4: 1/1      Map 5: 11/11    Reducer 2: 1(+3)/4      Reducer 3: 0/1  Reducer 6: 2/2
INFO  : Map 1: 11/11    Map 4: 1/1      Map 5: 11/11    Reducer 2: 4/4  Reducer 3: 0(+1)/1      Reducer 6: 2/2
INFO  : Map 1: 11/11    Map 4: 1/1      Map 5: 11/11    Reducer 2: 4/4  Reducer 3: 1/1  Reducer 6: 2/2
INFO  : Completed executing command(queryId=hive_20211014170228_a84bf4ff-c731-40b1-9071-07e0e9cf5182); Time taken: 21.434 seconds
INFO  : OK
+----------------------+
| avg_event_per_device |
+----------------------+
| 52.0                 |
+----------------------+
1 row selected (21.646 seconds)
0: jdbc:hive2://localhost:10000/default>
```

```
0: jdbc:hive2://localhost:10000/default> select device_id, count(distinct(event_id)) avg_event_per_device from events_external where device_id in (select distinct(train.device_id) as device_id from trai
n_external as train inner join events_external as events on train.device_id = events.device_id) group by device_id order by avg_event_per_device desc limit 10;
INFO  : Compiling command(queryId=hive_20211014170316_95d8c162-cb0a-4430-8bca-a591f96ff617): select device_id, count(distinct(event_id)) avg_event_per_device from events_external where device_id in (sel
ect distinct(train.device_id) as device_id from train_external as train inner join events_external as events on train.device_id = events.device_id) group by device_id order by avg_event_per_device desc
limit 10
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:device_id, type:string, comment:null), FieldSchema(name:avg_event_per_device, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20211014170316_95d8c162-cb0a-4430-8bca-a591f96ff617); Time taken: 0.218 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014170316_95d8c162-cb0a-4430-8bca-a591f96ff617): select device_id, count(distinct(event_id)) avg_event_per_device from events_external where device_id in (sel
ect distinct(train.device_id) as device_id from train_external as train inner join events_external as events on train.device_id = events.device_id) group by device_id order by avg_event_per_device desc
limit 10
INFO  : Query ID = hive_20211014170316_95d8c162-cb0a-4430-8bca-a591f96ff617
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select device_id, count(distinct(event_...10(Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Status: Running (Executing on YARN cluster with App id application_1634228715435_0005)
```

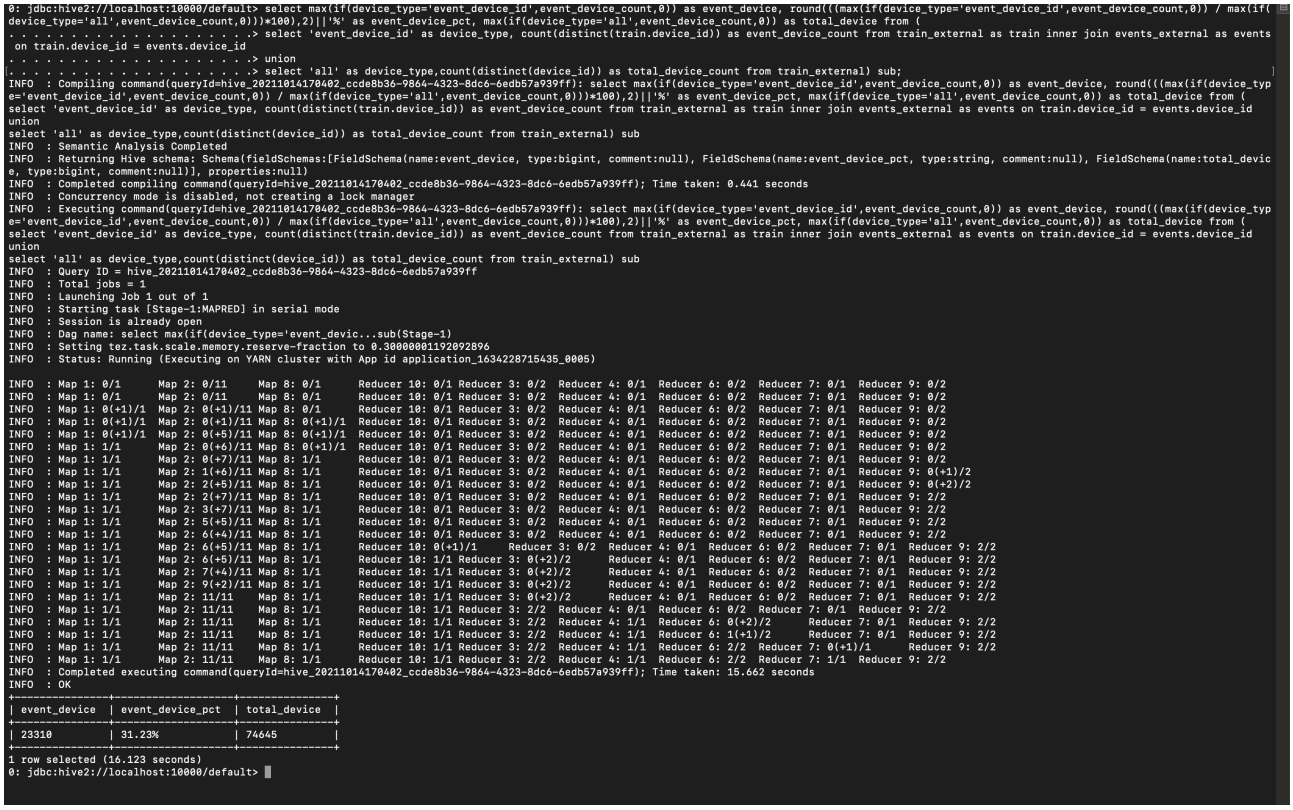
3.5.2 Average events per device

Since we are taking the average per device, the denominator will be always 1 for all the devices. The count will be equal to the average in this case.

select device_id, count(distinct(event_id)) avg_event_per_device from events_external where device_id in (select distinct(train.device_id) as device_id from train_external as train inner join events_external as events on train.device_id = events.device_id) group by device_id order by avg_event_per_device desc limit 10;

```
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20211014170316_95d8c162-cb0a-4430-8bca-a591f96ff617): select device_id, count(distinct(event_id)) avg_event_per_device from events_external where device_id in (sel
ect distinct(train.device_id) as device_id from train_external as train inner join events_external as events on train.device_id = events.device_id) group by device_id order by avg_event_per_device desc
limit 10
INFO  : Query ID = hive_20211014170316_95d8c162-cb0a-4430-8bca-a591f96ff617
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select device_id, count(distinct(event_...10(Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Status: Running (Executing on YARN cluster with App id application_1634228715435_0005)

INFO  : Map 1: 0/11     Map 2: 0/1      Map 3: 0/11     Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 0/11     Map 2: 0/1      Map 3: 0/11     Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 0(+3)/11 Map 2: 0/1      Map 3: 0/11     Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 0(+5)/11 Map 2: 0/1      Map 3: 0/11     Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 0(+6)/11 Map 2: 0/1      Map 3: 0/11     Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 0(+7)/11 Map 2: 0/1      Map 3: 0/11     Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 1(+7)/11 Map 2: 0/1      Map 3: 0/11     Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 3(+7)/11 Map 2: 0/1      Map 3: 0/11     Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 4(+6)/11 Map 2: 0/1      Map 3: 0/11     Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 6(+4)/11 Map 2: 0/1      Map 3: 0/11     Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 6(+5)/11 Map 2: 0/1      Map 3: 0/11     Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 8(+3)/11 Map 2: 0(+1)/1  Map 3: 0(+1)/11 Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 8(+3)/11 Map 2: 1/1      Map 3: 0(+2)/11 Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 9(+2)/11 Map 2: 1/1      Map 3: 0(+2)/11 Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 9(+2)/11 Map 2: 1/1      Map 3: 0(+4)/11 Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 10(+1)/11        Map 2: 1/1      Map 3: 0(+5)/11 Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 10(+1)/11        Map 2: 1/1      Map 3: 0(+6)/11 Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 0(+6)/11 Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 1(+6)/11 Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 2(+6)/11 Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 3(+7)/11 Reducer 4: 0/2  Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 5(+6)/11 Reducer 4: 0(+1)/2      Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 7(+4)/11 Reducer 4: 0(+2)/2      Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 8(+3)/11 Reducer 4: 0(+2)/2      Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 10(+1)/11        Reducer 4: 0(+2)/2      Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 11/11    Reducer 4: 0(+2)/2      Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 11/11    Reducer 4: 0(+2)/2      Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 11/11    Reducer 4: 1(+1)/2      Reducer 5: 0/2  Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 11/11    Reducer 4: 2/2  Reducer 5: 0(+2)/2      Reducer 6: 0/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 11/11    Reducer 4: 2/2  Reducer 5: 2/2  Reducer 6: 0(+1)/1
INFO  : Map 1: 11/11    Map 2: 1/1      Map 3: 11/11    Reducer 4: 2/2  Reducer 5: 2/2  Reducer 6: 1/1
INFO  : Completed executing command(queryId=hive_20211014170316_95d8c162-cb0a-4430-8bca-a591f96ff617); Time taken: 20.768 seconds
INFO  : OK
+----------------------+-----------------------+
|      device_id       | avg_event_per_device  |
+----------------------+-----------------------+
| -6242501228649110000 | 4150                  |
| -8340098378141150000 | 3973                  |
| -3746248670824150000 | 3987                  |
| 5375599021847300000  | 3128                  |
| 47825820477291600000 | 2899                  |
| 17796310234394000000 | 2757                  |
| 5098778421671830000  | 2722                  |
| 37246549257651500000 | 2347                  |
| -6875585507485880000 | 2310                  |
| 63561790191028700000 | 2023                  |
+----------------------+-----------------------+
10 rows selected (21.002 seconds)
0: jdbc:hive2://localhost:10000/default>
```

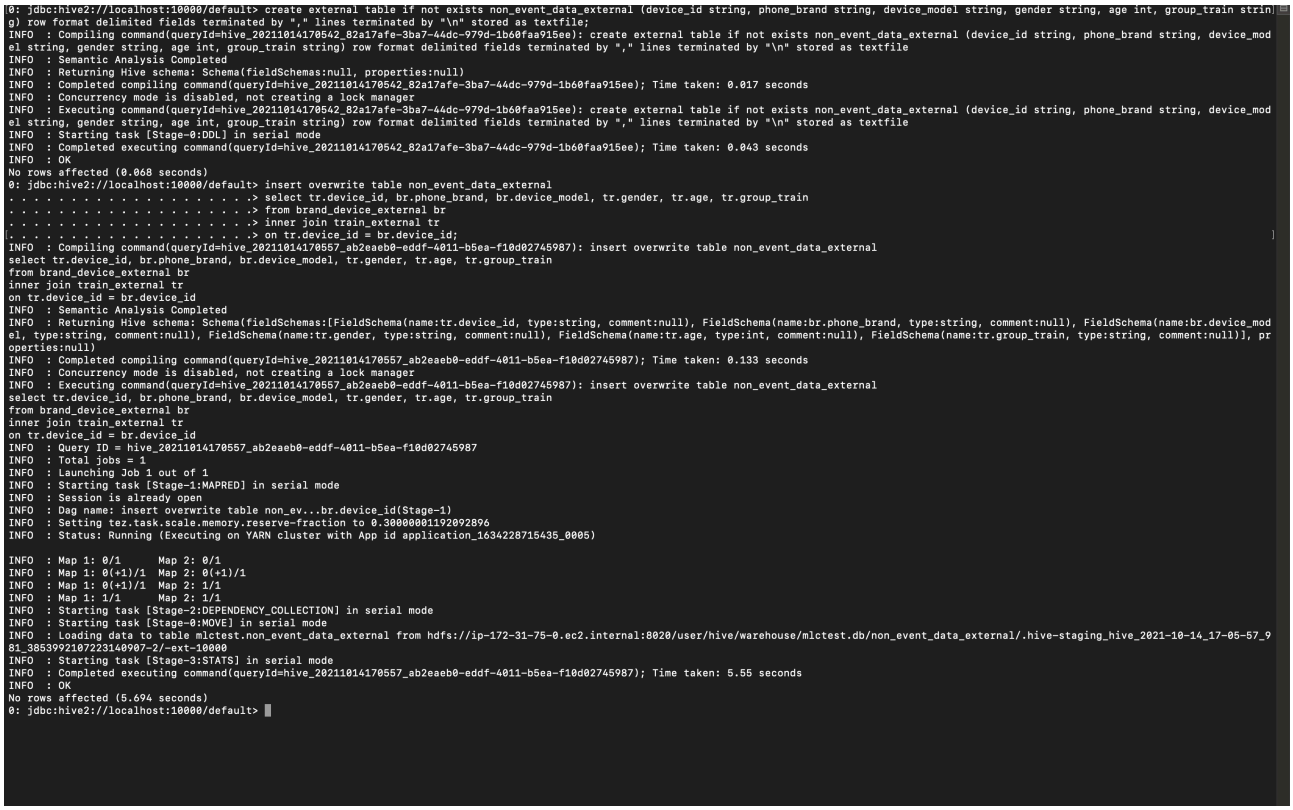


3.6 Count and percentage of device_id in train table have corresponding events data available?

3.6.1 Query Execution

```
select max(if(device_type='event_device_id',event_device_count,0)) as event_device,
round(((max(if(device_type='event_device_id',event_device_count,0)) /
max(if(device_type='all',event_device_count,0)))*100),2)||'%' as event_device_pct,
max(if(device_type='all',event_device_count,0)) as total_device from (
select 'event_device_id' as device_type, count(distinct(train.device_id)) as event_device_count from
train_external as train inner join events_external as events on train.device_id = events.device_id
union
select 'all' as device_type,count(distinct(device_id)) as total_device_count from train_external) sub;
```

3.6.2 Query Output



4. Hive Analysis Report

- Xiaomoi, Samsung and Huawei are the top 3 preferred brands across the gender
- Meizu is a brand that is preferred by Males in the Top 10 brand category
- Vivo is a brand that is preferred by Females in the Top 10 brand category
- Males have a higher % of mobile usage across the 10 top brand categories
- In terms of train dataset provided, the male domination is around 65% and female 35%
- 32% of the device are associated with an event or other
- There is an overall average of 52% events across devices
- Top 3 event count across unique devices
  - 1st Max - 4150 events
  - 2nd Max - 3973 events
  - 3rd Max – 3907 events