Capstone Project – Mid submission

(HQL Tasks)

1. **Creation of Hive Tables.**

   Below are the queries used for creation and insertion of data for 6 tables :

   a. **Train table**

```
create table train
(
 device_id  string,
 gender  string,
 age  int,
 group_train string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';

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

   b. **Brand Device Table**

```
create table brand_device
(
 device_id  string,
 phone_brand  string,
 device_model  string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';

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

   c. **Events Table**

```
create table events
(
 event_id  int,
 device_id  string,
 event_timestamp  timestamp ,
 longitude string,
```

```
 latitude string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';

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

d. **App Events Table**

```
create table app_events
(
 event_id  int,
 app_id  string,
 is_installed  int,
 is_active  int
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';

load  data inpath '/user/hadoop/mlctest/app_events/' into table app_events;
```

e. **App Labels Table**

```
create table app_labels
(
 app_id  string,
 label_id  string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
tblproperties("skip.header.line.count"="1");

load  data  inpath '/home/hadoop/app_labels_new.txt' into table app_labels;
```

f. **Label Categories Table**

```
create table label_categories
(
 label_id  string,
 category  string
```

```
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
tblproperties("skip.header.line.count"="1");

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

2. HQL Tasks Result :

Before performing the HQL tasks, the table *'brand_device'* had duplicate values. All the duplicate rows were removed using the following HQL query.

*insert overwrite table brand_device select distinct * from brand_device;*

The above statement will overwrite the existing table replacing it with the distinct rows.

The following devices had two brand names for the same device ID. Those were removed explicitly using the below mentioned command.

"-3004353610608670000
-5269721363279120000
-6590454305031520000
-7059081542575370000
-7297178577997110000
1186608308763910000
5245428108336910000"

HQL Statement :
*insert overwrite table brand_device select * from brand_device where device_id not in ('-3004353610608670000','-5269721363279120000', '-6590454305031520000', '-7059081542575370000', '-7297178577997110000', '1186608308763910000', '5245428108336910000') ;*

Below are the output and Hive query for all the HQL tasks :

a. The 10 most popular brands and the percentage of the respective Male and Female owners of these brands [Handle the device id duplicates from brand_device table].

   Query :

with brand_count as (select phone_brand,count(*) brand_cnt from brand_device a inner join train b on a.device_id = b.device_id group by phone_brand order by brand_cnt desc limit 10)
select b.phone_brand,
brand_gender.gender,b.brand_cnt,brand_gender.gender_count,
*(brand_gender.gender_count/b.brand_cnt) * 100 gender_percentage from (*
select phone_brand, gender, count(*) gender_count from
(select a.device_id, a.gender, b.phone_brand from train a inner join brand_device b on a.device_id= b.device_id) temp group by phone_brand, gender
) brand_gender inner join brand_count b on brand_gender.phone_brand = b.phone_brand
order by brand_cnt desc, gender_count desc;

Screenshot :

```
hive> with brand_count as (select phone_brand,count(*) brand_cnt from brand_devi
ce a inner join train b on a.device_id = b.device_id group by phone_brand order
by brand_cnt desc limit 10 )
    > select b.phone_brand, brand_gender.gender,b.brand_cnt,brand_gender.gender_
count, (brand_gender.gender_count/b.brand_cnt ) * 100 gender_percentage from (
    > select phone_brand, gender, count(*) gender_count from
    > (select a.device_id, a.gender, b.phone_brand from train a inner join brand
_device b on a.device_id= b.device_id ) temp group by phone_brand, gender
    > ) brand_gender inner join brand_count b on brand_gender.phone_brand = b.ph
one_brand
    > order by brand_cnt desc, gender_count desc;
No Stats for default@train, Columns: device_id, gender
No Stats for default@brand_device, Columns: device_id, phone_brand
No Stats for default@brand_device, Columns: device_id, phone_brand
No Stats for default@train, Columns: device_id
Query ID = hadoop_20210111122820_0b3fba24-002c-4a0c-a0c5-78b13c118eba
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610340234319
_0010)
```

```
b.phone_brand   brand_gender.gender     b.brand_cnt     brand_gender.gender_count       gender_percentage
Xiaomi  M       17299   11381   65.78993005376034
Xiaomi  F       17299   5918    34.21006994623967
samsung M       13669   8238    60.26775916306973
samsung F       13669   5431    39.73224083693028
Huawei  M       12960   8716    67.25308641975309
Huawei  F       12960   4244    32.74691358024691
OPPO    M       5783    3212    55.54210617326647
OPPO    F       5783    2571    44.45789382673353
vivo    M       5637    2986    52.9714387085329
vivo    F       5637    2651    47.02856129146709
Meizu   M       4698    3396    72.2860791826309
Meizu   F       4698    1302    27.713920817369093
Coolpad M       3339    2260    67.68493560946392
Coolpad F       3339    1079    32.315064390536094
lenovo  M       2691    1798    66.81531029357116
lenovo  F       2691    893     33.18468970642884
Gionee  M       1123    721     64.20302760463046
Gionee  F       1123    402     35.79697239536955
HTC     M       1013    693     68.41066140177689
HTC     F       1013    320     31.589338598223097
Time taken: 32.549 seconds, Fetched: 20 row(s)
```

b. The 10 most popular brands for Male and Female? [Handle the device id duplicates from the brand_device data set.]

Query:

(select phone_brand,'Female' gender, count(*) gender_count from

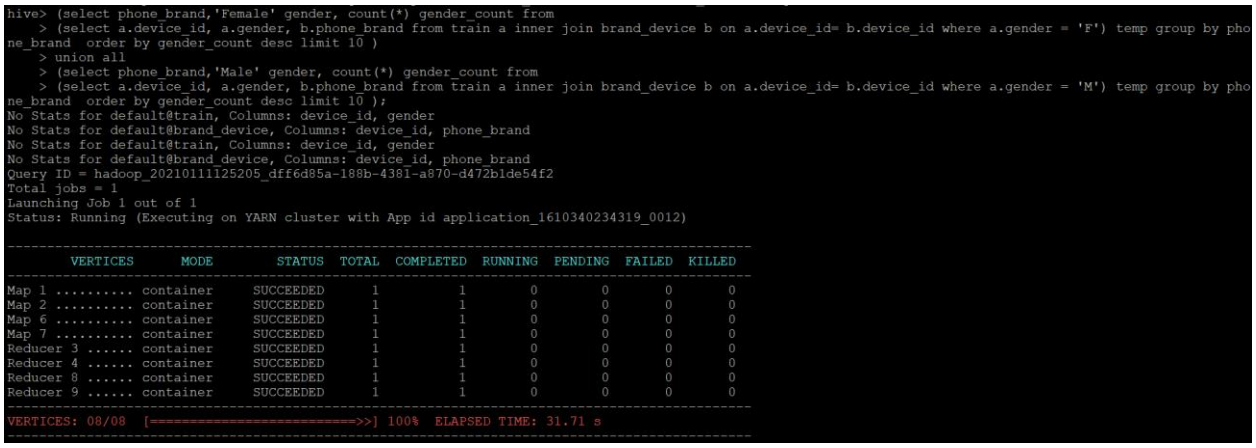
(select a.device_id, a.gender, b.phone_brand from train a inner join brand_device b on a.device_id= b.device_id where a.gender = 'F') temp group by phone_brand order by gender_count desc limit 10)

union all

(select phone_brand,'Male' gender, count(*) gender_count from

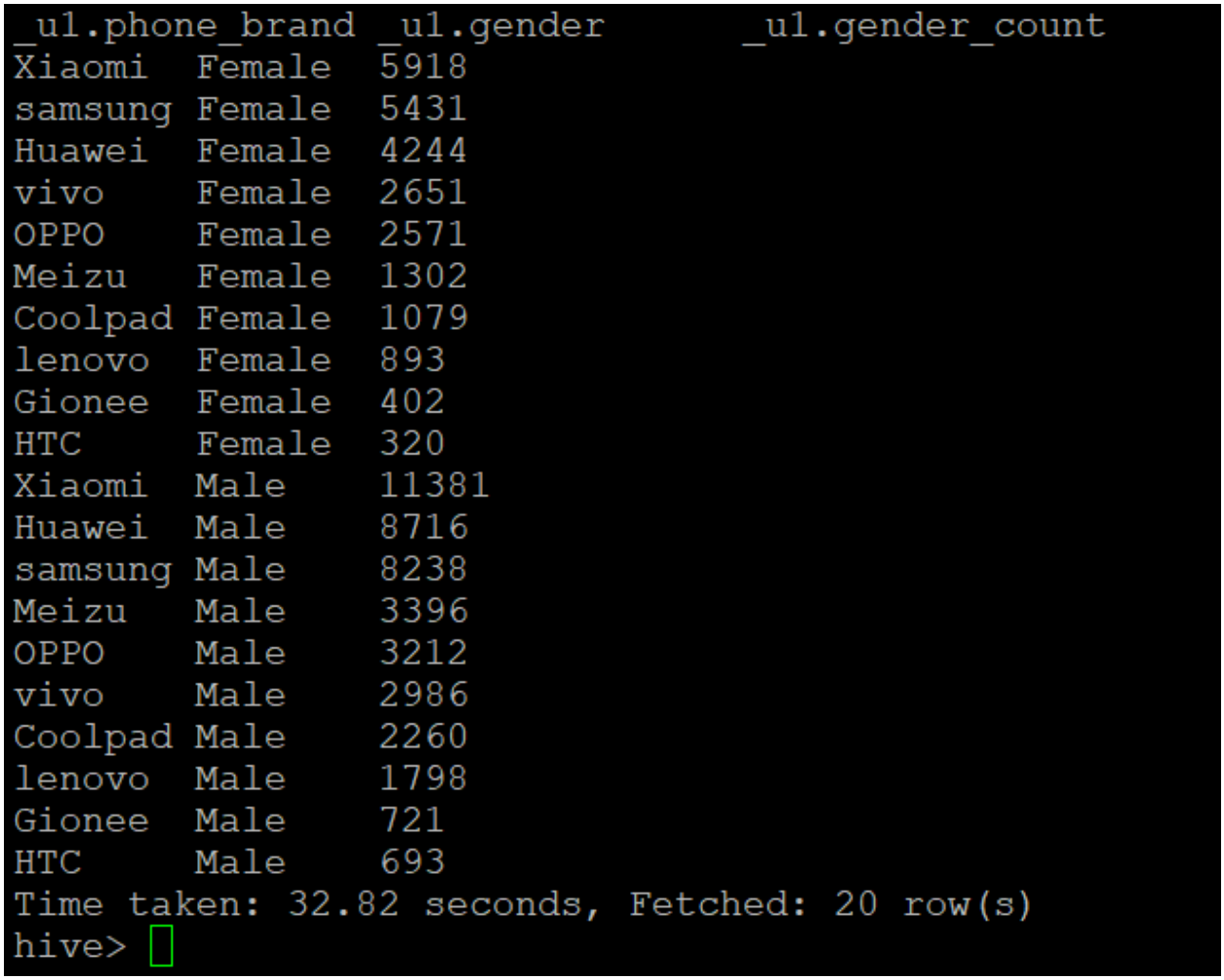
(select a.device_id, a.gender, b.phone_brand from train a inner join brand_device b on a.device_id= b.device_id where a.gender = 'M') temp group by phone_brand order by gender_count desc limit 10);

Screenshot :



```
hive> (select phone_brand,'Female' gender, count(*) gender_count from
    > (select a.device_id, a.gender, b.phone_brand from train a inner join brand_device b on a.device_id= b.device_id where a.gender = 'F') temp group by pho
ne_brand  order by gender_count desc limit 10 )
    > union all
    > (select phone_brand,'Male' gender, count(*) gender_count from
    > (select a.device_id, a.gender, b.phone_brand from train a inner join brand_device b on a.device_id= b.device_id where a.gender = 'M') temp group by pho
ne_brand  order by gender_count desc limit 10 );
No Stats for default@train, Columns: device_id, gender
No Stats for default@brand_device, Columns: device_id, phone_brand
No Stats for default@train, Columns: device_id, gender
No Stats for default@brand_device, Columns: device_id, phone_brand
Query ID = hadoop_20210111125205_dff6d85a-188b-4381-a870-d472b1de54f2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610340234319_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 6 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 7 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 8 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 9 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 08/08  [==========================>>] 100%  ELAPSED TIME: 31.71 s
--------------------------------------------------------------------------------
```



c. The count and percentage analysis of the Gender in the train data set

Query:

WITH gender_part AS (select gender, count(*) gender_count from train group by gender)

select gender , gender_count , gender_count/ (sum(gender_count) over ()) gender_percentage from gender_part;

Screenshot

```
hive> WITH gender_part AS (select gender, count(*) gender_count from train group by gender )
    > select gender , gender_count , gender_count/ (sum(gender_count) over ()) gender_percentage from gender_part;
Query ID = hadoop_20210111131454_a29035c0-a02a-4afd-a5b0-8a1fedb60502
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610340234319_0014)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 7.12 s
--------------------------------------------------------------------------------
```

```
gender  gender_count    gender_percentage
F       26741   0.35824234710965236
M       47904   0.6417576528903477
Time taken: 7.656 seconds, Fetched: 2 row(s)
hive>
```

d. The top mobile phone brands offering the highest number of models [Provide details about the top three brands.]

Query:

select phone_brand, count(*) num_of_models from (

select distinct phone_brand, device_model from brand_device) dist_brand_model

group by phone_brand order by num_of_models desc limit 3

;

Screenshot:

```
hive> select phone_brand, count(*) num_of_models from (
    > select distinct phone_brand, device_model from brand_device ) dist_brand_model
    > group by phone_brand order by num_of_models desc limit 3
    > ;
Query ID = hadoop_20210111132117_2ca7b4b7-130b-477c-9602-debfc6ff1843
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610340234319_0014)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 6.59 s
--------------------------------------------------------------------------------
```

```
phone_brand     num_of_models
lenovo  194
samsung 163
Huawei  145
Time taken: 7.107 seconds, Fetched: 3 row(s)
hive>
```

e. The average number of events per device id [Applicable to the device_id column from the train table, which has at least one associated event in the event table]

Query:

select avg(event_count) avg_device_event_count from

(select a.device_id, count(*) event_count from events a inner join train b on a.device_id = b.device_id group by a.device_id) temp;

Screenshot:

```
hive> select avg(event_count) avg_device_event_count from
    > (select a.device_id, count(*) event_count from events a inner join train b on a.device_id = b.device_id group by a.device_id ) temp;
Query ID = hadoop_20210111142932_eb6bea06-9399-42f0-b8ff-12db5ab57bbf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610340234319_0018)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      4          4        0        0       0       0
Map 4 ..........   container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 50.27 s
----------------------------------------------------------------------------------------------
OK
```

```
avg_device_event_count
52.14920634920635
Time taken: 51.069 seconds, Fetched: 1 row(s)
hive>
```

f. Whether the count and percentage of the device_id column in the train table have corresponding events data available

Query:

with train_count as (select count(*) cnt, 1 dummy from train),

train_event_count as (select count(distinct a.device_id) cnt, 1 dummy from events a inner join train b on a.device_id = b.device_id)

*select a.cnt Train_Count, b.cnt Events_Count , (b.cnt/a.cnt) * 100 Percentage from train_count a inner join train_event_count b on a.dummy= b.dummy;*

Screenshot: