Hive analysis

PS: External tables are created, and data is stored in parquet format, so they are faster in accessing

<u>Loading data into external tables from files loaded from RDS</u>

Table: app_events

Query to create the external table:

```
create external table app_events_stg3(
   event_id bigint,
   app_id bigint,
   is_installed int,
   is_active int
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';
```

Script to load the data into the external table:

load data inpath '/home/hadoop/capstonetelcom/stage/app_events2' into table app_events_stg3;

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> create table app_events_stg2(
    >     event_id bigint,
    >     app_id bigint,
    >     is_installed int,
    >     is_active int
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n';
OK
Time taken: 1.979 seconds
hive> load data inpath '/home/hadoop/capstonetelcom/stage/app_events2' into app_events_stg2
    > ;
FAILED: ParseException line 1:70 missing TABLE at 'app_events_stg2' near '<EOF>'
hive> load data inpath '/home/hadoop/capstonetelcom/stage/app_events2' into table app_events_stg2;
Loading data to table default.app_events_stg2
OK
Time taken: 1.745 seconds
hive> select count(*) from app_events_stg2
    > ;
Query ID = hadoop_20221227151708_7cc5facd-0db7-4033-b03b-8295f902bd8d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672149773056_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     10         10        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 13.12 s
--------------------------------------------------------------------------------
OK
32473067
Time taken: 16.81 seconds, Fetched: 1 row(s)
hive>
```

Query to convert the external table to table in parquet format:

create table app_events3

stored as parquet

as

select event_id,app_id,is_installed,is_active from app_events_stg3;

```
hive> create table app_events
    > stored as parquet
    > as
    > select event_id,app_id,is_installed,is_active from app_events_stg2;
Query ID = hadoop_20221227152056_02da390c-fc67-405a-8b39-8b77160e94cb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672149773056_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     10         10        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 21.21 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://ha-nn-uri/user/hive/warehouse/app_events
OK
Time taken: 24.775 seconds
hive> select count(*) from app_events;
OK
32473067
Time taken: 0.486 seconds, Fetched: 1 row(s)
```

Table: brand_device

Query to create the external table:

create external table brand_device_stg3(

device_id bigint,

phone_brand string,

device_model string

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n';

Script to load the data into the external table:

load data inpath '/home/hadoop/capstonetelcom/stage/brand_device' into table brand_device_stg3;

```
hive> create table brand_device_stg(
    >     device_id bigint,
    >     phone_brand string,
    >     device_model string
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n';
OK
Time taken: 1.402 seconds
hive> load data inpath '/home/hadoop/capstonetelcom/stage/brand_device' into table brand_device_stg;
Loading data to table default.brand_device_stg
OK
Time taken: 2.045 seconds
hive> select count(*) from brand_device_stg;
Query ID = hadoop_20221227152941_c97fc454-1eb8-49b9-bcfd-4756e9921387
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672149773056_0009)

--------------------------------------------------------------------------------
        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
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.94 s
--------------------------------------------------------------------------------
OK
187245
Time taken: 9.362 seconds, Fetched: 1 row(s)
```

Query to convert the external table to table in parquet format:

create table brand_device3

stored as parquet

as

select device_id,phone_brand,device_model from brand_device_stg3;

```
Time taken: 1.364 seconds
hive> create table brand_device
    > stored as parquet
    > as
    > select device_id,phone_brand,device_model from brand_device_stg;
Query ID = hadoop_20221227153056_cda9c532-8a63-4471-b0a9-fc31cb6d04a1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672149773056_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 4.41 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ha-nn-uri/user/hive/warehouse/brand_device
OK
Time taken: 7.762 seconds
hive> select count(*) from brand_device;
OK
187245
Time taken: 0.402 seconds, Fetched: 1 row(s)
hive>
```

Table: events

Query to create the external table:

create external table events_stg3 (

event_id bigint,

device_id bigint,

event_timestamp timestamp,

longitude decimal(10,2),

latitude decimal(10,2)

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n';

Script to load the data into the external table:

load data inpath '/home/hadoop/capstonetelcom/stage/events' into table events_stg3;

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> create table events_stg (
    >     event_id bigint,
    >      device_id bigint,
    >       event_timestamp timestamp,
    >     longitude decimal(10,2),
    >     latitude decimal(10,2)
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n';
OK
Time taken: 1.448 seconds
hive> load data inpath '/home/hadoop/capstonetelcom/stage/events' into table events_stg;
Loading data to table default.events_stg
OK
Time taken: 2.209 seconds
hive> select count(*) from events_stg;
Query ID = hadoop_20221227155656_2c20a630-56d3-4066-b2c2-7281f2aad14f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672149773056_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     10         10        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 10.95 s
--------------------------------------------------------------------------------
OK
3252950
Time taken: 15.102 seconds, Fetched: 1 row(s)
```

Query to convert the external table to table in parquet format:

create table events3

stored as parquet

as

select event_id, device_id, event_timestamp, longitude, latitude from events_stg3;

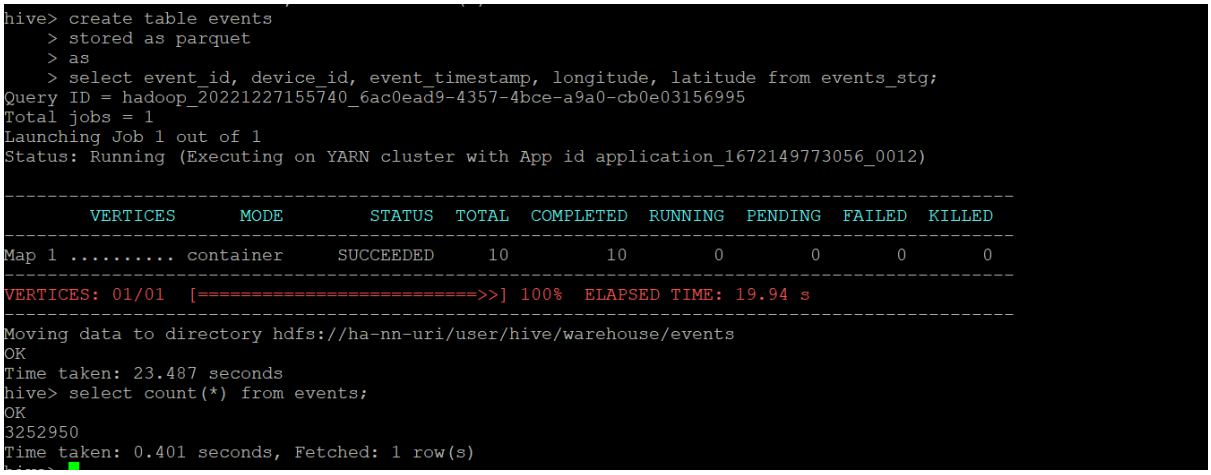


Table: train

Query to create the external table:

```
create external table train_stg3 (
    device_id bigint,
     gender string,
    age int,
    group_name string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';
```

Script to load the data into the external table:

```
load data inpath '/home/hadoop/capstonetelcom/stage/train' into table train_stg3;
```

```
hive> create table train_stg (
    >      device_id bigint,
    >       gender string,
    >      age int,
    >      group_name string
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n';
OK
Time taken: 1.395 seconds
hive> load data inpath '/home/hadoop/capstonetelcom/stage/train' into table train_stg;
Loading data to table default.train_stg
OK
Time taken: 1.865 seconds
hive> select count(*) from train_stg;
Query ID = hadoop_20221227160452_0a4962ca-e85a-42eb-8073-ac368a944c42
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672149773056_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
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.98 s
--------------------------------------------------------------------------------
OK
74645
Time taken: 8.647 seconds, Fetched: 1 row(s)
hive>
```

Query to convert the external table to table in parquet format:

create table train3

stored as parquet

as

select device_id, gender, age, group_name from train_stg3;

```
Time taken: 8.647 seconds, Fetched: 1 row(s)
hive> create table train
    > stored as parquet
    > as
    > select device_id, gender, age, group_name from train_stg;
Query ID = hadoop_20221227160529_f9a4121f-6208-4cc6-b008-1cfe11b0dd03
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672149773056_0014)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 4.46 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ha-nn-uri/user/hive/warehouse/train
OK
Time taken: 8.432 seconds
hive> select count(*) from train;
OK
74645
Time taken: 0.368 seconds, Fetched: 1 row(s)
hive>
```

Loading data into tables from files loaded from S3

File: label_categories.csv

Query to create the external table:

```
create external table label_categories_stg3(
   label_id bigint,
   category string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';
```

Script to load the data into the external table:

```
load data local inpath '/home/hadoop/capstonetelcom/stage/labelcategories/label_categories.csv' into table label_categories_stg3;
```

Query to convert the external table to table in parquet format:

```
create table label_categories3
stored as parquet
as
select label_id,category from label_categories_stg3;
```

```
hive> select count(*) from label_categories_stg3;
Query ID = hadoop_20221230203408_3851e806-6e84-4beb-94f2-ab2653b9e5a9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672392463972_0052)

--------------------------------------------------------------------------------
        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
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.67 s
--------------------------------------------------------------------------------
OK
931
Time taken: 6.662 seconds, Fetched: 1 row(s)
hive> create table label_categories3
    > stored as parquet
    > as
    > select label_id,category from label_categories_stg3;
Query ID = hadoop_20221230203528_aee70e85-72c3-47b8-8c8d-2c427e756c33
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672392463972_0052)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.49 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-15-164.us-east-2.compute.internal:8020/user/hive/warehouse/label_categories3
OK
Time taken: 7.583 seconds
hive> select count(*) from label_categories3;
OK
931
Time taken: 0.399 seconds, Fetched: 1 row(s)
```

File: app_labels_new.txt

Query to create the external table:

```
create external table app_labels_stg3(
   app_id bigint,
   label_id bigint
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';
```

Script to load the data into the external table:

```
load data local inpath '/home/hadoop/capstonetelcom/stage/applables/app_labels_new.txt' into table app_labels_stg3;
```

Query to convert the external table to table in parquet format:

```
create table app_labels3
stored as parquet
as
select app_id,label_id from app_labels_stg3;
```

```
hive> select count(*) from app_labels_stg3;
Query ID = hadoop_20221230202910_fcb0257d-0a0f-4b49-9f22-b092099087ad
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672392463972_0052)

--------------------------------------------------------------------------------
        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
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.41 s
--------------------------------------------------------------------------------
OK
459944
Time taken: 12.685 seconds, Fetched: 1 row(s)
hive> create table app_labels3
    > stored as parquet
    > as
    > select app_id,label_id from app_labels_stg3;
Query ID = hadoop_20221230203002_8d17cd33-961d-4c25-9662-7f052ffc7663
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672392463972_0052)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 7.94 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-15-164.us-east-2.compute.internal:8020/user/hive/warehouse/app_labels3
OK
Time taken: 11.308 seconds
hive> select count(*) from app_labels3;
OK
459944
Time taken: 0.293 seconds, Fetched: 1 row(s)
```

Hive Analytics report

1. 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: select phone_brand, ((malecount/totalcount)*100) as male_owner_percentage,

((femalecount/totalcount)*100) as female_owner_percentage

from

(select count(a.device_id) as cnt1, a.phone_brand, count(case when gender=='M' then 1 end) as malecount, count(case when gender=='F' then 1 end) as femalecount, count(gender) as totalcount

from

(Select device_id, count(1) cnt from brand_device3 group by device_id having cnt=1) c,

brand_device3 a,

train3 b

where

a.device_id=b.device_id and

a.device_id=c.device_id

group by a.phone_brand

order by cnt1 desc

limit 10) n;

```
Query ID = hadoop_20221230184233_52b3488f-dffb-4c28-a39d-4a38a2d6b114
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672392463972_0045)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 5 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 6 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 06/06  [==========================>>] 100%  ELAPSED TIME: 26.15 s
----------------------------------------------------------------------------------------------
OK
phone_brand     male_owner_percentage   female_owner_percentage
Xiaomi  65.78611980071834       34.21388019928166
samsung 60.24794600938967       39.75205399061033
Huawei  67.2497871352272        32.75021286477281
OPPO    55.569049271339345      44.43095072866065
vivo    52.95584045584045       47.04415954415954
Meizu   72.30637934713036       27.693620652869637
Coolpad 67.58786422349054       32.41213577650946
lenovo  66.80312616300708       33.19687383699293
Gionee  64.26024955436719       35.7397504456328
HTC     68.44708209693373       31.55291790306627
Time taken: 27.27 seconds, Fetched: 10 row(s)
hive>
```

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

Query:

```
with cte as (
  select
    b.gender,
    a.phone_brand,
    count(a.device_id) as c,
    dense_rank() over (partition by b.gender order by count(a.device_id) desc) as dr
  from
    (Select device_id, count(1) cnt from brand_device3 group by device_id having cnt=1) c,
    brand_device3 a,
    train3 b
  where
    a.device_id=b.device_id and
    a.device_id=c.device_id
  group by b.gender, a.phone_brand
)
select *
from cte
where dr <= 10
order by gender, c desc;
```

```
VERTICES: 07/07  [==========================>>] 100%  ELAPSED TIME: 20.3
--------------------------------------------------------------------------------
OK
cte.gender      cte.phone_brand cte.c    cte.dr
F       Xiaomi  5906    1
F       samsung 5419    2
F       Huawei  4231    3
F       vivo    2642    4
F       OPPO    2561    5
F       Meizu   1298    6
F       Coolpad 1079    7
F       lenovo  892     8
F       Gionee  401     9
F       HTC     319     10
M       Xiaomi  11356   1
M       Huawei  8688    2
M       samsung 8213    3
M       Meizu   3389    4
M       OPPO    3203    5
M       vivo    2974    6
M       Coolpad 2250    7
M       lenovo  1795    8
M       Gionee  721     9
M       HTC     692     10
Time taken: 22.511 seconds, Fetched: 20 row(s)
```

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

Query: select male as male_count, female as female_count, (male/total)*100 as male_percentage, (female/total)*100 as female_percentage

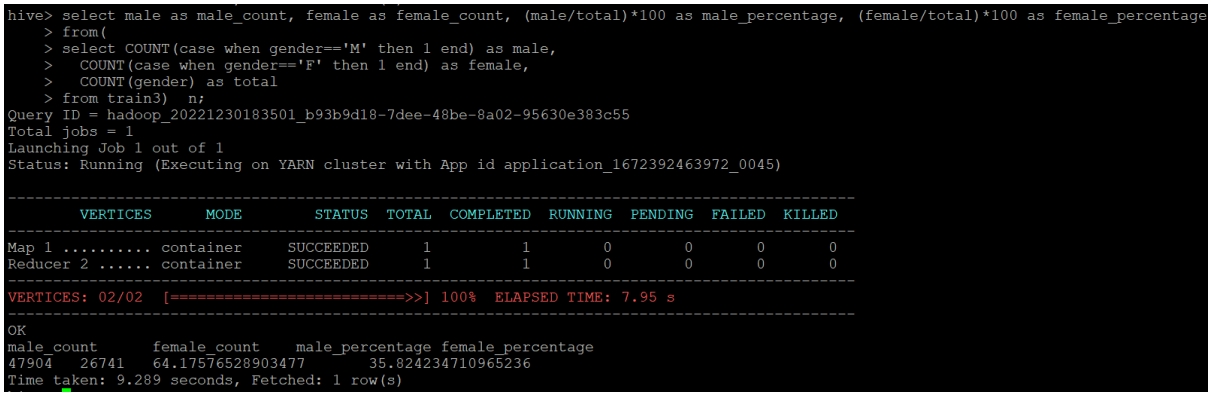
from(

select COUNT(case when gender=='M' then 1 end) as male,

COUNT(case when gender=='F' then 1 end) as female,

COUNT(gender) as total

from train3) n;



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

Query: select count(a.device_id) as device_count, a.phone_brand, count(device_model) as model_count

from

(Select device_id, count(1) cnt from brand_device3 group by device_id having cnt=1) c,

brand_device3 a,

train3 b

where

a.device_id=b.device_id and

a.device_id=c.device_id

group by a.phone_brand

order by device_count desc

limit 3;

```
Query ID = hadoop_20221230191707_6be231e1-b230-4fff-8a90-5bcf37c5953f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672392463972_0047)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 5 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 6 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 06/06  [==========================>>] 100%  ELAPSED TIME: 28.06 s
--------------------------------------------------------------------------------
OK
device_count    a.phone_brand   model_count
17262   Xiaomi  17262
13632   samsung 13632
12919   Huawei  12919
Time taken: 34.742 seconds, Fetched: 3 row(s)
```

5. 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 (sum(event_count)/count(device_id)) as avg_events_per_device

from

(select

a.device_id, count(a.event_id) as event_count

from

events3 a,

train3 b

where

a.device_id = b.device_id

group by a.device_id) as n;

```
Time taken: 30.955 seconds, Fetched: 23310 row(s)
hive> select (sum(event_count)/count(device_id)) as avg_events_per_device
    > from
    > (select
    >     a.device_id, count(a.event_id) as event_count
    > from
    > events3 a,
    > train3 b
    > where
    > a.device_id = b.device_id
    > group by a.device_id) as n;
Query ID = hadoop_20221230160605_4ca3d513-d0e2-412c-9cf6-fa58189f333a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672392463972_0034)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Map 4 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 30.13 s
----------------------------------------------------------------------------------------------
OK
avg_events_per_device
52.14920634920635
Time taken: 31.424 seconds, Fetched: 1 row(s)
```

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

Query: select count(device_id) from train3;

```
hive> (select count(device_id) from train3);
Query ID = hadoop_20221230173054_9075bfba-4e34-4a9c-abb6-fd5f03171991
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672392463972_0041)

--------------------------------------------------------------------------------
        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
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.20 s
--------------------------------------------------------------------------------
OK
_c0
74645
Time taken: 15.139 seconds, Fetched: 1 row(s)
```

Yes, the count and percentage of device id column in train table have events data. Below is the query and screenshot

Query: select

count(b.device_id) as device_count, ((count(b.device_id) * 100.0)/74645) as device_percentage

from

events3 a,

train3 b

where

a.device_id = b.device_id

having count(a.event_id) >0;

```
Time taken: 34.742 seconds, Fetched: 3 row(s)
hive> select
    >     count(b.device_id) as device_count, ((count(b.device_id) * 100.0)/74645) as device_percentage
    > from
    > events3 a,
    > train3 b
    > where
    > a.device_id = b.device_id
    > having count(a.event_id) >0;
Query ID = hadoop_20221230192033_75ab0a97-8c9a-4a04-952a-8880b7c37d55
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1672392463972_0047)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Map 3 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 26.44 s
--------------------------------------------------------------------------------
OK
device_count    device_percentage
1215598 1628.505593
Time taken: 28.69 seconds, Fetched: 1 row(s)
```