IIT- M Advanced Certificate Program in Machine Learning and Cloud- upGrad Capstone Project

User Demographics Prediction using Telecom dataset

HQL Task Commands

Authors :

Mukul Pahawa

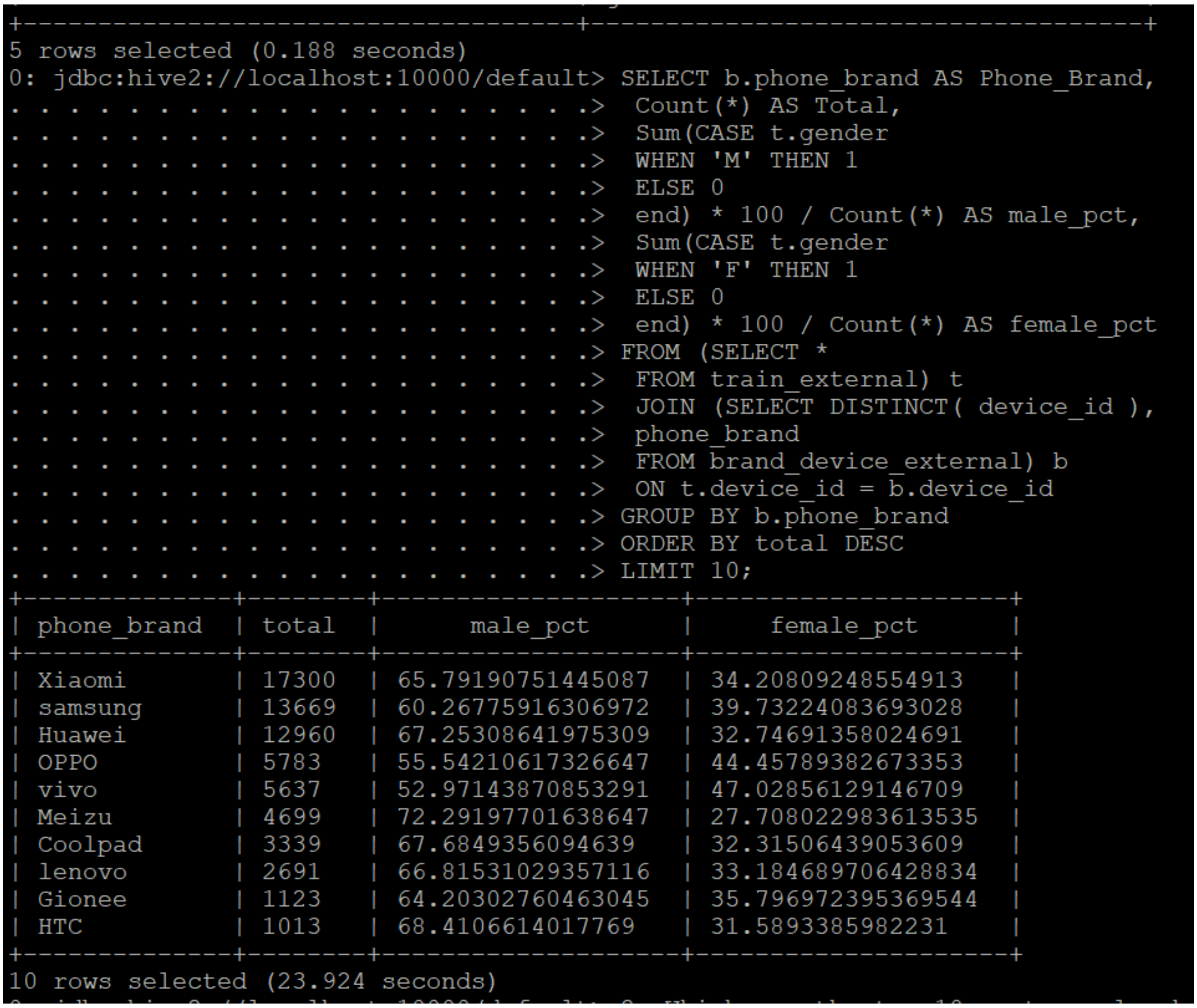
Mitesh

HQL Tasks

===========

1. Which are the top 10 most popular brands and respective % for Male and Female in it ? [Do handle the device_id duplicates from brand_device table]

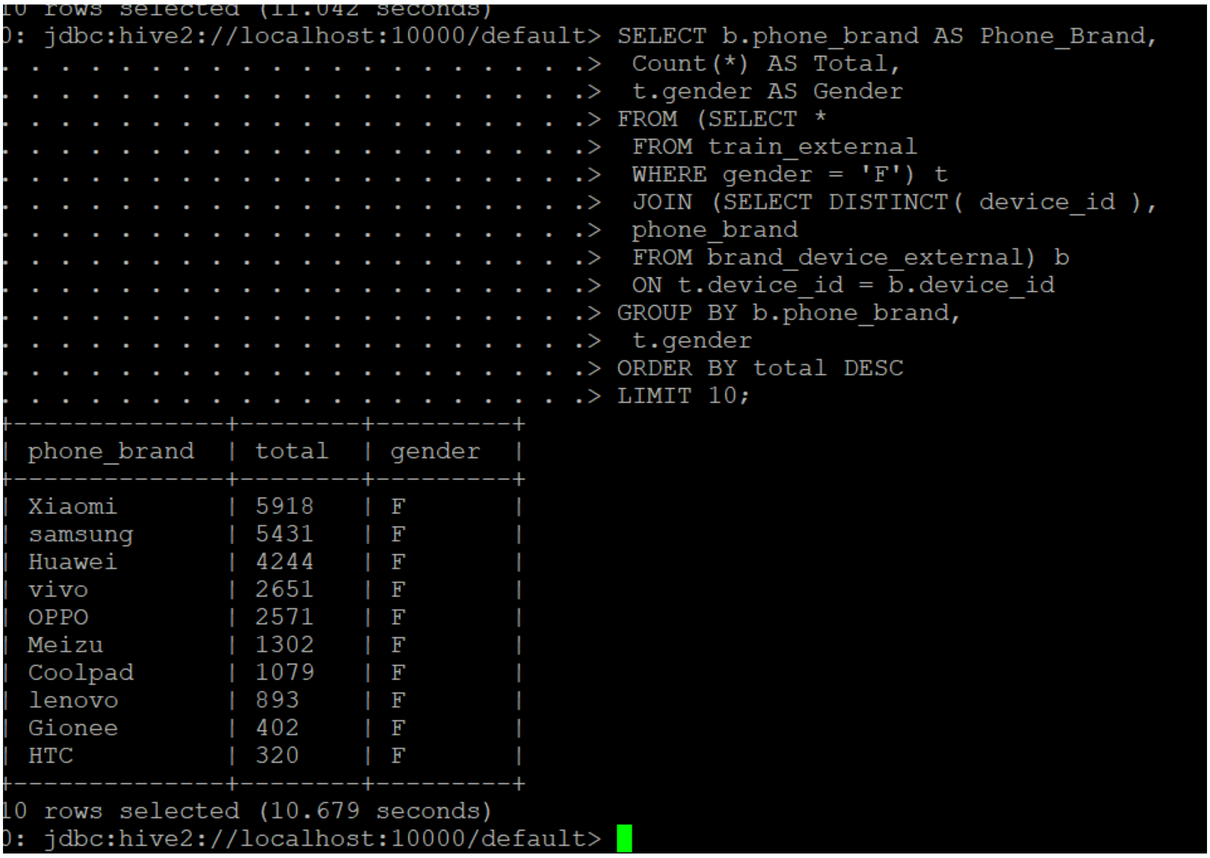
```
SELECT b.phone_brand AS Phone_Brand,
 Count(*) AS Total,
 Sum(CASE t.gender
 WHEN 'M' THEN 1
 ELSE 0
 end) * 100 / Count(*) AS male_pct,
 Sum(CASE t.gender
 WHEN 'F' THEN 1
 ELSE 0
 end) * 100 / Count(*) AS female_pct
FROM (SELECT *
 FROM train_external) t
 JOIN (SELECT DISTINCT( device_id ),
 phone_brand
 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 (0.188 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;
+--------------+--------+---------------------+----------------------+
| 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   | 68.4106614017769    | 31.5893385982231     |
+--------------+--------+---------------------+----------------------+
10 rows selected (23.924 seconds)
```

2. Which are the top 10 most popular brands for Male and Female ? [Do handle the device_id duplicates from brand_device dataset]

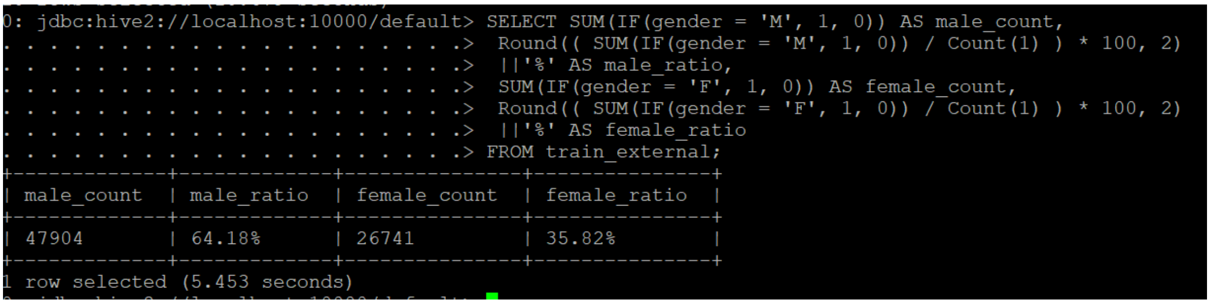
```
SELECT b.phone_brand as Phone_Brand,
 Count(*) AS Total,
 t.gender as Gender
FROM (SELECT *
 FROM train_external
 WHERE gender = 'M') t
 JOIN (SELECT DISTINCT( device_id ),
 phone_brand
 FROM brand_device_external) b
 ON t.device_id = b.device_id
GROUP BY b.phone_brand, t.gender
ORDER BY total DESC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



```
10 rows selected (11.042 seconds)
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;
+--------------+--------+---------+
| phone_brand  | total  | gender  |
+--------------+--------+---------+
| Xiaomi       | 5918   | F       |
| samsung      | 5431   | F       |
| Huawei       | 4244   | F       |
| vivo         | 2651   | F       |
| OPPO         | 2571   | F       |
| Meizu        | 1302   | F       |
| Coolpad      | 1079   | F       |
| lenovo       | 893    | F       |
| Gionee       | 402    | F       |
| HTC          | 320    | F       |
+--------------+--------+---------+
10 rows selected (10.679 seconds)
0: jdbc:hive2://localhost:10000/default>
```

3. Count and percentage Analysis of the Gender in the train Dataset

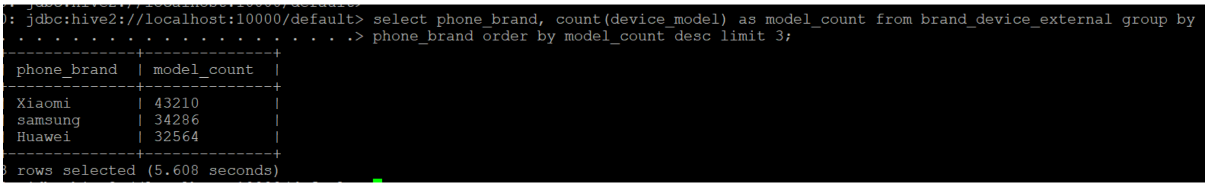
```
SELECT SUM(IF(gender = 'M', 1, 0)) AS male_count,

 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;
```



```
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;
+-------------+-------------+---------------+---------------+
| male_count  | male_ratio  | female_count  | female_ratio  |
+-------------+-------------+---------------+---------------+
| 47904       | 64.18%      | 26741         | 35.82%        |
+-------------+-------------+---------------+---------------+
1 row selected (5.453 seconds)
```

4. Top mobile phone brands offering the highest number of models [Give top three brands]

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;
+--------------+--------------+
| phone_brand  | model_count  |
+--------------+--------------+
| Xiaomi       | 43210        |
| samsung      | 34286        |
| Huawei       | 32564        |
+--------------+--------------+
3 rows selected (5.608 seconds)
```

5. Average number of events per device id [Applicable to device_id from train table which have atleast one associated event in the event table]

5.5.1 Overall Average events across devices

===

SELECT Round(Count(DISTINCT(event_id)) / Count(DISTINCT(device_id))) AS

avg_event_per_device

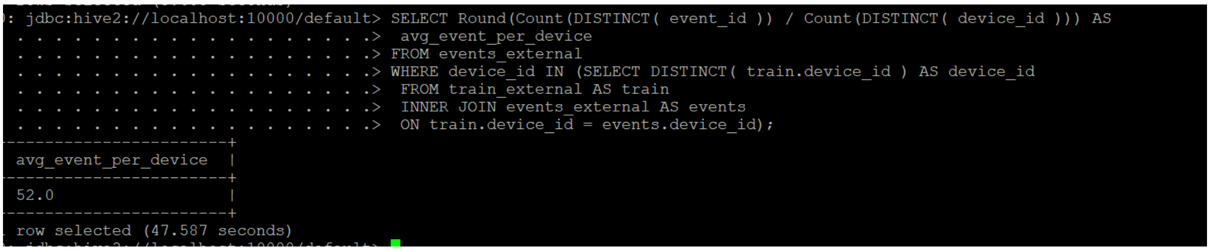
FROM events_external

WHERE device_id IN (SELECT DISTINCT(train.device_id) AS device_id

FROM train_external AS train

INNER JOIN events_external AS events

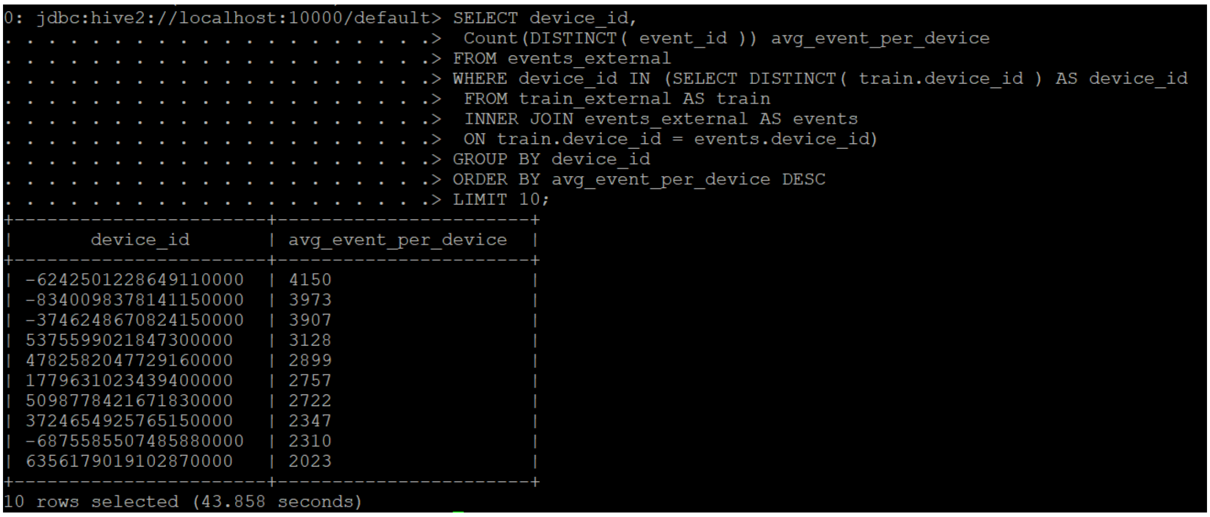
```
 ON train.device_id = events.device_id);
```



```
): 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);
-----------------------+
 avg_event_per_device  |
-----------------------+
 52.0                  |
-----------------------+
 row selected (47.587 seconds)
```

5.5.2 Average events per device

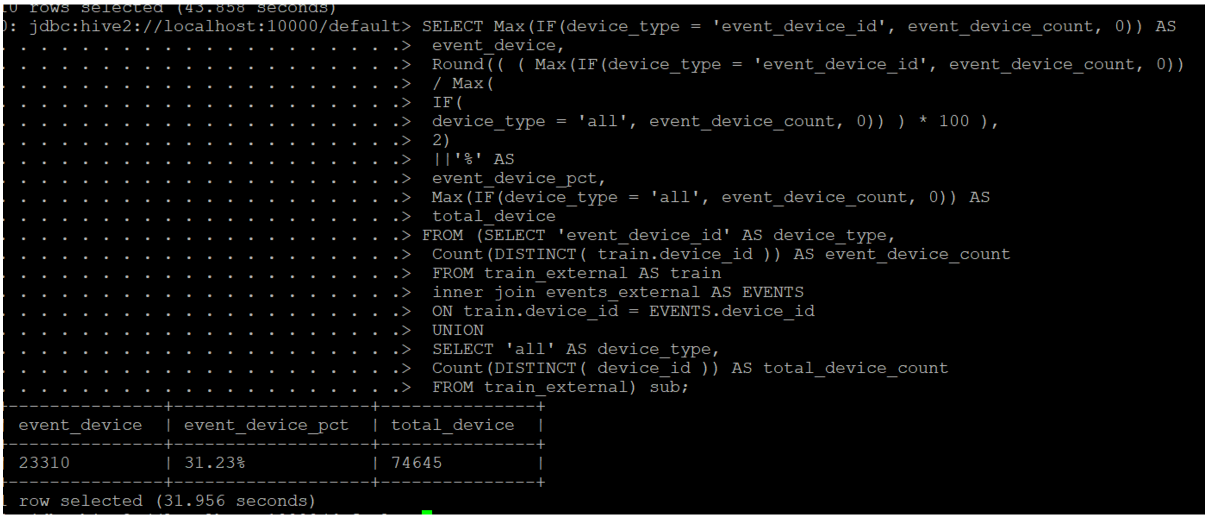
```
SELECT device_id,

 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;
```



```
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 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;
+-----------------------+-----------------------+
|       device_id       | avg_event_per_device  |
+-----------------------+-----------------------+
| -6242501228649110000  | 4150                  |
| -8340098378141150000  | 3973                  |
| -3746248670824150000  | 3907                  |
| 5375599021847300000   | 3128                  |
| 4782582047729160000   | 2899                  |
| 1779631023439400000   | 2757                  |
| 5098778421671830000   | 2722                  |
| 3724654925765150000   | 2347                  |
| -6875585507485880000  | 2310                  |
| 6356179019102870000   | 2023                  |
+-----------------------+-----------------------+
10 rows selected (43.858 seconds)
```

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

```
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;
```



```
10 rows selected (43.858 seconds)
0: jdbc:hive2://localhost:10000/default> 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;
+---------------+--------------------+---------------+
| event_device  | event_device_pct   | total_device  |
+---------------+--------------------+---------------+
| 23310         | 31.23%             | 74645         |
+---------------+--------------------+---------------+
1 row selected (31.956 seconds)
```