Logic For Final Submission

Validate count of rows in all Three Tables

clickstream

```
hive> select count(*) from clickstream;
Query ID = ec2-user_20221009192151_48515473-4ffa-4294-ba88-2b1179a17165
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665331741275
_0018)

Map 1: 0/1      Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 0/1      Reducer 2: 0/1
Map 1: 1/1      Reducer 2: 0(+1)/1
Map 1: 1/1      Reducer 2: 1/1
OK
3001
Time taken: 7.325 seconds, Fetched: 1 row(s)
```

booking Table

```
hive> select count(*) from booking;
Query ID = ec2-user_20221009191918_617c4312-6511-46eb-9c44-860ddc7ae76d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665331741275
_0018)

Map 1: 0/1      Reducer 2: 0/1
Map 1: 0/1      Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 1/1      Reducer 2: 0(+1)/1
Map 1: 1/1      Reducer 2: 1/1
OK
1000
Time taken: 9.778 seconds, Fetched: 1 row(s)
```

aggregate_datewise table

```
hive> select count(*) from aggregate_datewise;
Query ID = ec2-user_20221009192247_ea700d2c-67ab-4970-b3e9-d8a5efab0507
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665331741275
_0018)

Map 1: 0/1      Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 1/1      Reducer 2: 0(+1)/1
Map 1: 1/1      Reducer 2: 1/1
OK
290
Time taken: 5.354 seconds, Fetched: 1 row(s)
```

Now we will move on to tasks

Task 5: Calculate the total number of different drivers for each customer.

We will use bookings table since it has information about drivers and the customers, we will count distinct driver_id and group by to get total_number

```
hive> select customer_id ,count( DISTINCT driver_id) from booking
    > group by customer_id
    > order by customer_id asc;
Query ID = ec2-user_20221009193047_d14508b2-2779-47ac-97f9-bd457797a9b5
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_166533174127
_0019)

Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 0(+1)/2      Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 1(+1)/2      Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 1/1
OK
10022393        1
10058402        1
10339567        1
10435129        1
```

```
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 1/1
OK
10022393        1
10058402        1
10339567        1
10435129        1
10555335        1
10592274        1
10614890        1
10678994        1
11264797        1
11353346        1
11418437        1
11438890        1
11454977        1
11479815        1
11518953        1
11580321        1
11596512        1
11608791        1
11655671        1
11757536        1
11764909        1
```

Task 6: Calculate the total rides taken by each customer
To get total rides, we will do count aggregate on booking_id and then group by

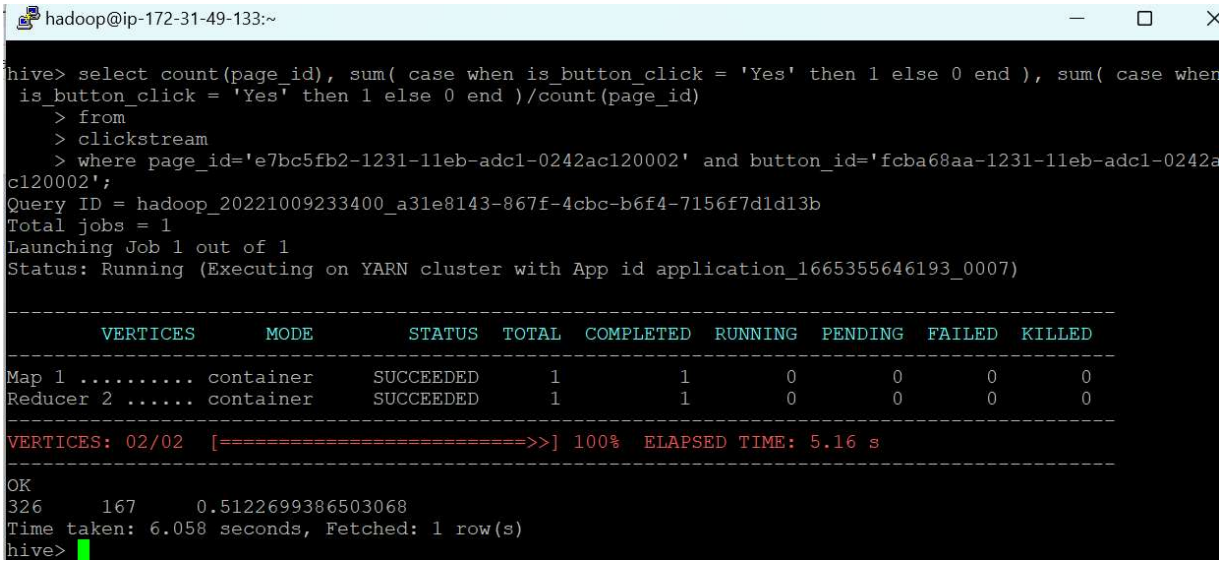
```
hive> select customer_id ,count( DISTINCT booking_id) from booking
    > group by customer_id
    > order by customer_id asc;
Query ID = ec2-user_20221009193635_105825f5-4937-40b1-82c0-a6ef564b14ac
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1665331741275
_0020)

Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 0(+1)/2      Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 1(+1)/2      Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 1/1
OK
10022393        1
10058402        1
```

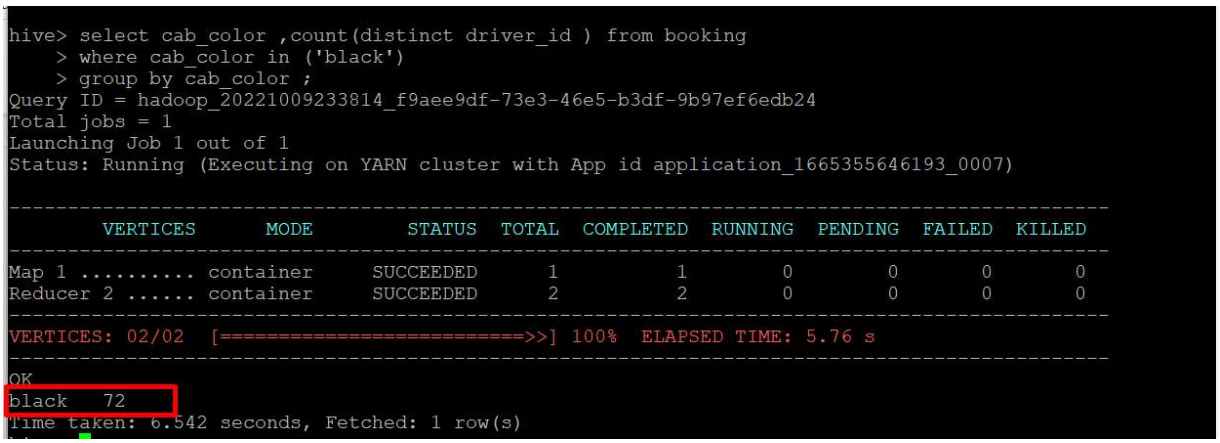
```
Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 1/1
OK
10022393        1
10058402        1
10339567        1
10435129        1
10555335        1
10592274        1
10614890        1
10678994        1
11264797        1
11353346        1
11418437        1
11438890        1
11454977        1
11479815        1
11518953        1
11580321        1
11596512        1
11608791        1
11655671        1
11757536        1
11764909        1
11860278        1
```

Task 7: Find the total visits made by each customer on the booking page and the total 'Book Now' button presses. This can show the conversion ratio.
The booking page id is 'e7bc5fb2-1231-11eb-adc1-0242ac120002'.
The Book Now button id is 'fcba68aa-1231-11eb-adc1-0242ac120002'.
You also need to calculate the conversion ratio as part of this task. Conversion ratio can be calculated as **Total 'Book Now' Button Press/Total Visits** made by customer on the booking page.

For this task, we use clickstream table, in which I do sum of rows that has is_button_click set as Yes and then to get conversion ratio, I divide this sum by count(page_id) that is count of all the pages that user visited.



Task 8: Calculate the count of all trips done on black cabs.
Used booking table to get the cab_color and then did a distinct count on driver_id column with condition to check cab_color is black, to get count of trips done in black cabs



Task 9 Calculate the total amount of tips given date wise to all drivers by customers.

For this task used pickup_timestamp from booking table and only used date format by aggregating on tip_amount column

```
hive> select date_format(pickup_timestamp,'yyyy-MM-dd'),sum(tip_amount) from booking
    > group by date_format(pickup_timestamp,'yyyy-MM-dd');
Query ID = hadoop_20221009234224_39028a90-05d7-4b9f-b46c-f7d1e15a2e61
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665355646193_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.76 s
--------------------------------------------------------------------------------
OK
2020-01-01      59
2020-01-02      95
2020-01-03      11
2020-01-04      123
2020-01-05      134
2020-01-08      111
2020-01-10      77
2020-01-11      81
2020-01-12      109
2020-01-16      155
2020-01-18      240
2020-01-20      210
2020-01-21      5
2020-01-26      209
2020-01-27      231
2020-01-29      123
```

Task 10 Calculate the total count of all the bookings with ratings lower than 2 as given by customers in a particular month.

For this query too, used booking table and did select query on pickup_timestamp, rating_by_customer based on the condition of the rating<2

```
hive> select date_format(pickup_timestamp,'yyyy-MM') ,count( rating_by_customer) from
    > booking
    > where rating_by_customer < 2
    > group by date_format(pickup_timestamp,'yyyy-MM') ;
Query ID = hadoop_20221009234536_1c876f02-c18e-408d-a67c-d548baf3da50
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665355646193_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.66 s
--------------------------------------------------------------------------------
OK
2020-01 26
2020-02 16
2020-03 16
2020-04 21
2020-05 21
2020-06 14
2020-07 20
2020-08 32
2020-09 21
2020-10 15
Time taken: 6.333 seconds, Fetched: 10 row(s)
```

Task 11 Calculate the count of total iOS users.

Used clickstream to get os_version that match iOS and aggregate on customer_id to get total count

```
hive> select os_version ,count(distinct customer_id) from clickstream
    > where os_version in ('iOS')
    > group by os_version;
Query ID = hadoop_20221009234835_2f23002f-9c6f-4eeb-8d92-8cef32487b93
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665355646193_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.48 s
--------------------------------------------------------------------------------
OK
iOS     1515
Time taken: 6.113 seconds, Fetched: 1 row(s)
```