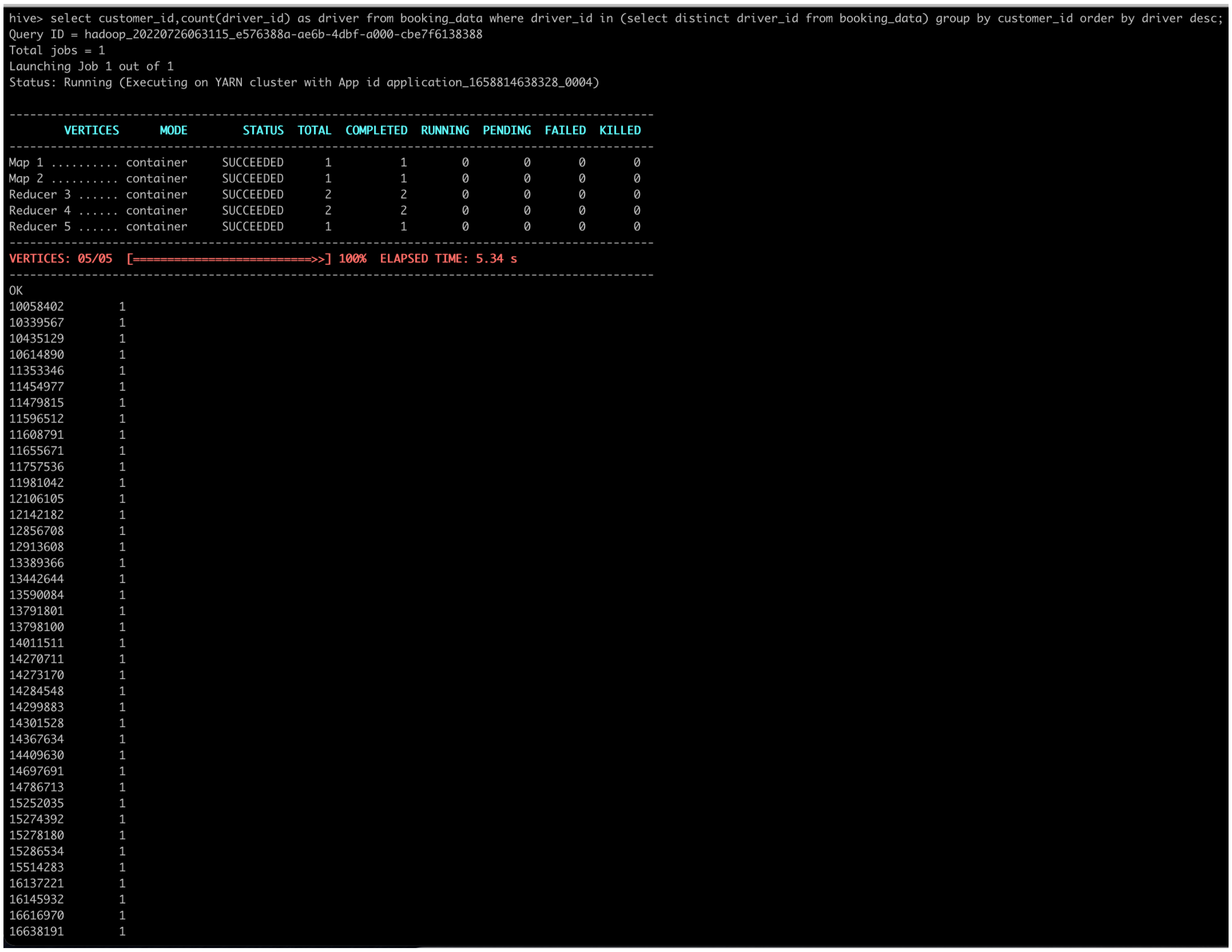
Logic For Final Submission

<Explain the queries, list them and attach screenshots after successful execution of queries>

1. **Calculate the total number of different drivers for each customer.**

→ We have selected 2 columns customer_id and count with driver_id with where clause for unique driver_id on booking_data table and finally grouped based on customer_id to get count for unique driver for each customer_id.

select customer_id,count(driver_id) as driver from booking_data where driver_id in (select distinct driver_id from booking_data) group by customer_id order by driver desc;

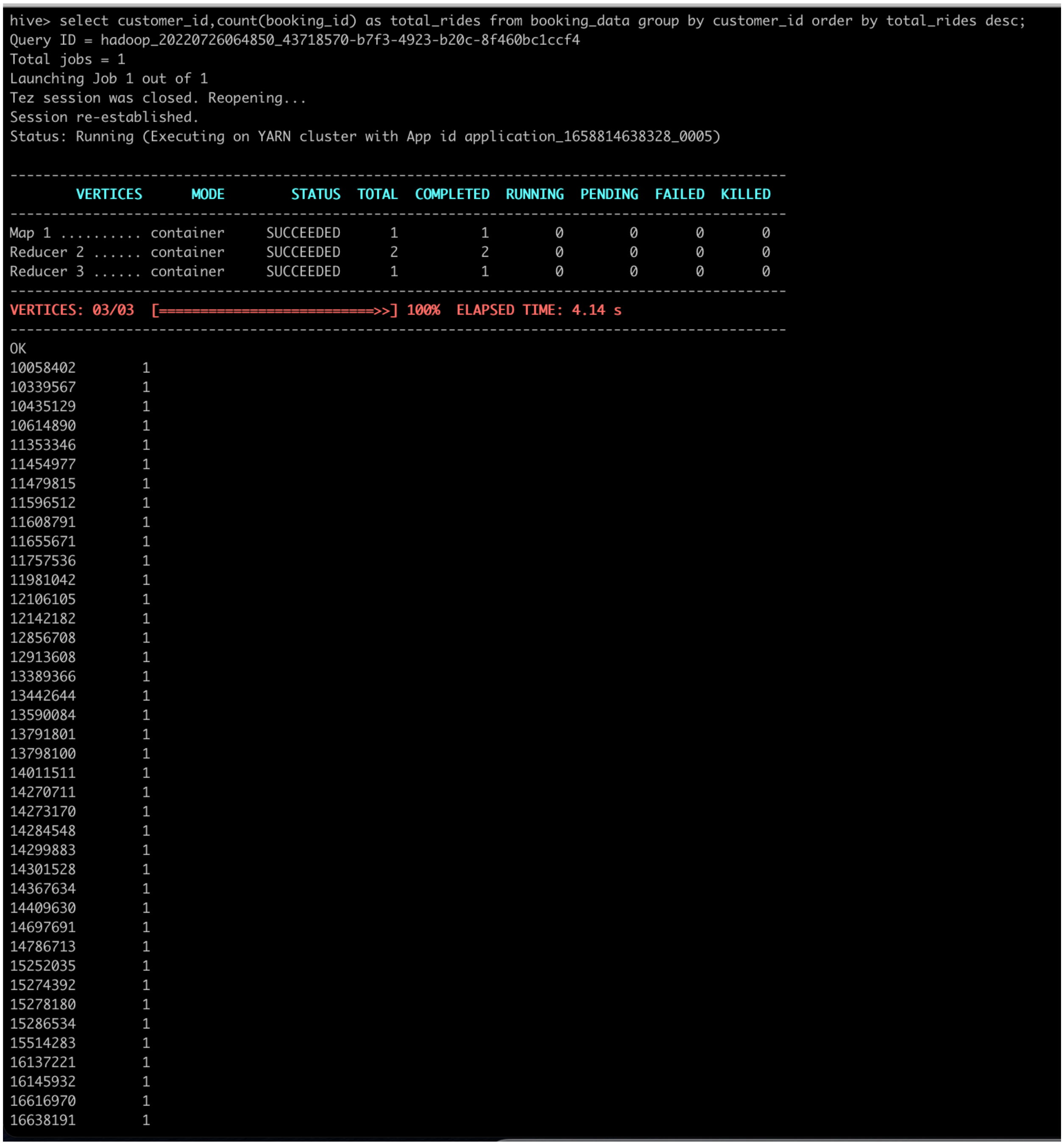


```
hive> select customer_id,count(driver_id) as driver from booking_data where driver_id in (select distinct driver_id from booking_data) group by customer_id order by driver desc;
Query ID = hadoop_20220726063115_e576388a-ae6b-4dbf-a000-cbe7f6138388
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658814638328_0004)

----------------------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 5.34 s
----------------------------------------------------------------------------------------------
OK
10058402        1
10339567        1
10435129        1
10614890        1
11353346        1
11454977        1
11479815        1
11596512        1
11608791        1
11655671        1
11757536        1
11981042        1
12106105        1
12142182        1
12856708        1
12913608        1
13389366        1
13442644        1
13590084        1
13791801        1
13798100        1
14011511        1
14270711        1
14273170        1
14284548        1
14299883        1
14301528        1
14367634        1
14409630        1
14697691        1
14786713        1
15252035        1
15274392        1
15278180        1
15286534        1
15514283        1
16137221        1
16145932        1
16616970        1
16638191        1
```

2. **Calculate the total rides taken by each customer.**

→ We have selected customer_id and count with booking_id to get total rides for each customer from booking_data table and order the output in descending based on count of booking_id.

select customer_id,count(booking_id) as total_rides from booking_data group by customer_id order by total_rides desc;



```
hive> select customer_id,count(booking_id) as total_rides from booking_data group by customer_id order by total_rides desc;
Query ID = hadoop_20220726064850_43718570-b7f3-4923-b20c-8f460bc1ccf4
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_1658814638328_0005)

----------------------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.14 s
----------------------------------------------------------------------------------------------
OK
10058402        1
10339567        1
10435129        1
10614890        1
11353346        1
11454977        1
11479815        1
11596512        1
11608791        1
11655671        1
11757536        1
11981042        1
12106105        1
12142182        1
12856708        1
12913608        1
13389366        1
13442644        1
13590084        1
13791801        1
13798100        1
14011511        1
14270711        1
14273170        1
14284548        1
14299883        1
14301528        1
14367634        1
14409630        1
14697691        1
14786713        1
15252035        1
15274392        1
15278180        1
15286534        1
15514283        1
16137221        1
16145932        1
16616970        1
16638191        1
```

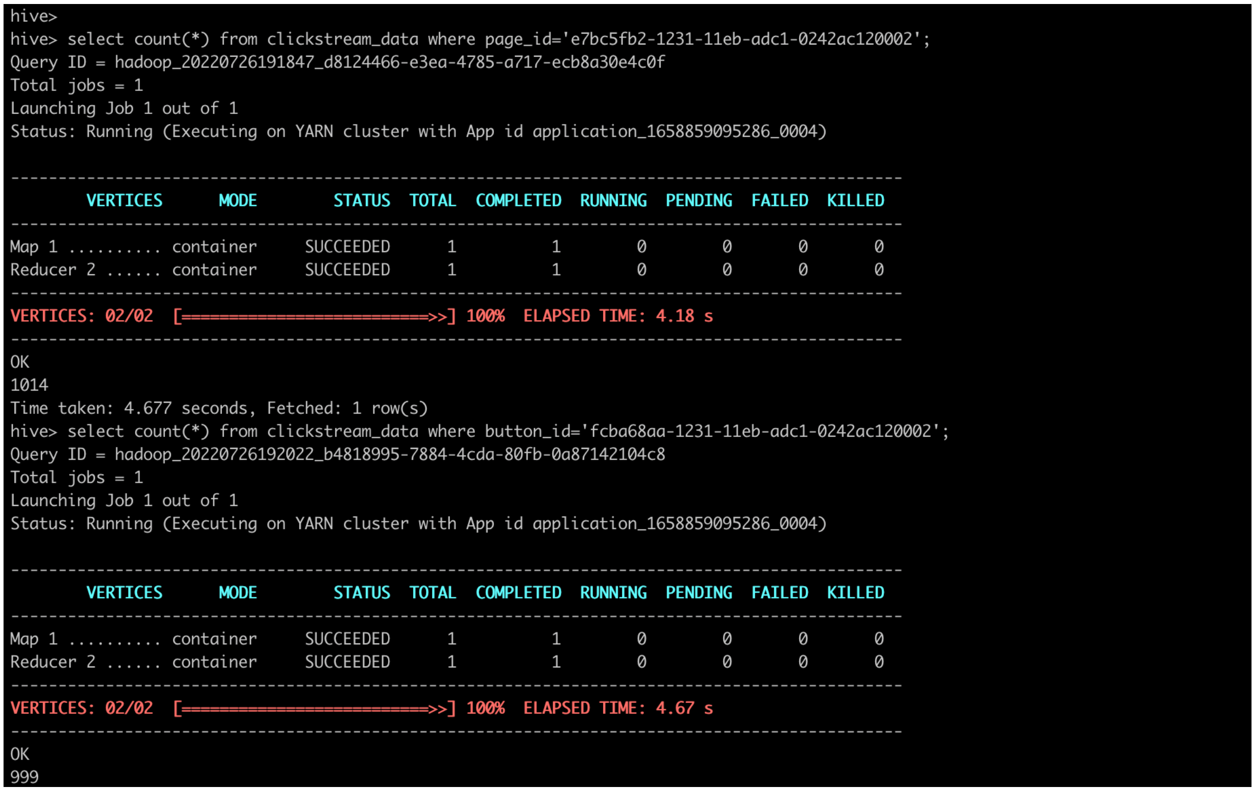
3. Find the total visits made by each customer on the booking page and the total ‘Book Now’ button presses. This can show the conversion ratio.

→Here we are first finding the count of data with button_id fcba68aa-1231-11eb-adc1-0242ac120002 from the clickstream_data tablen which comes 999.
Next we are finding the count of data with page_id e7bc5fb2-1231-11eb-adc1-0242ac120002 from the clickstream_data table which comes 1014.
Ratio = 999/1014 = 0.985

select b.count(customer_id)/a.count(customer_id) from clickstream_data as a join clickstream_data as b on a.customer_id=b.customer_id where a.page_id='e7bc5fb2-1231-11eb-adc1-0242ac120002' and b.button_id='fcba68aa-1231-11eb-adc1-0242ac120002';

OR

(select count(*) from clickstream_data where button_id='fcba68aa-1231-11eb-adc1-0242ac120002')/(select count(*) from clickstream_data where page_id='e7bc5fb2-1231-11eb-adc1-0242ac120002');

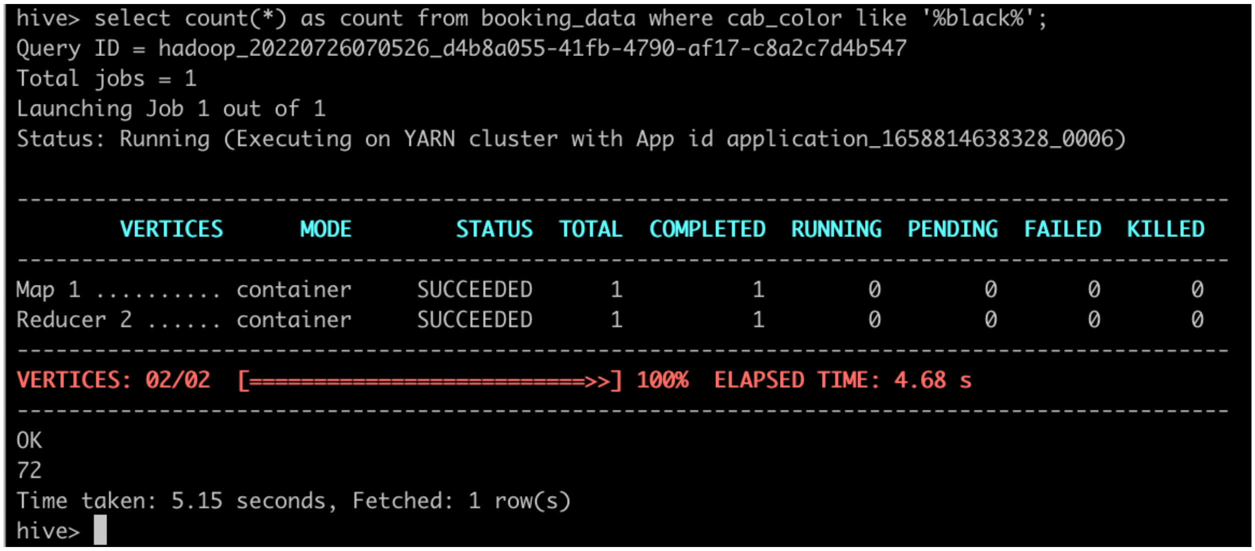


```
hive>
hive> select count(*) from clickstream_data where page_id='e7bc5fb2-1231-11eb-adc1-0242ac120002';
Query ID = hadoop_20220726191847_d8124466-e3ea-4785-a717-ecb8a30e4c0f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658859095286_0004)

----------------------------------------------------------------------------------------------
        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: 4.18 s
----------------------------------------------------------------------------------------------
OK
1014
Time taken: 4.677 seconds, Fetched: 1 row(s)
hive> select count(*) from clickstream_data where button_id='fcba68aa-1231-11eb-adc1-0242ac120002';
Query ID = hadoop_20220726192022_b4818995-7884-4cda-80fb-0a87142104c8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658859095286_0004)

----------------------------------------------------------------------------------------------
        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: 4.67 s
----------------------------------------------------------------------------------------------
OK
999
```

4. **Calculate the count of all trips done on black cabs.**

→ Here we have simply counted all the data from booking_data table with cab_color black. Also used like instead of equals to match black irrespective of case.

select count(*) as count from booking_data where cab_color like '%black%';

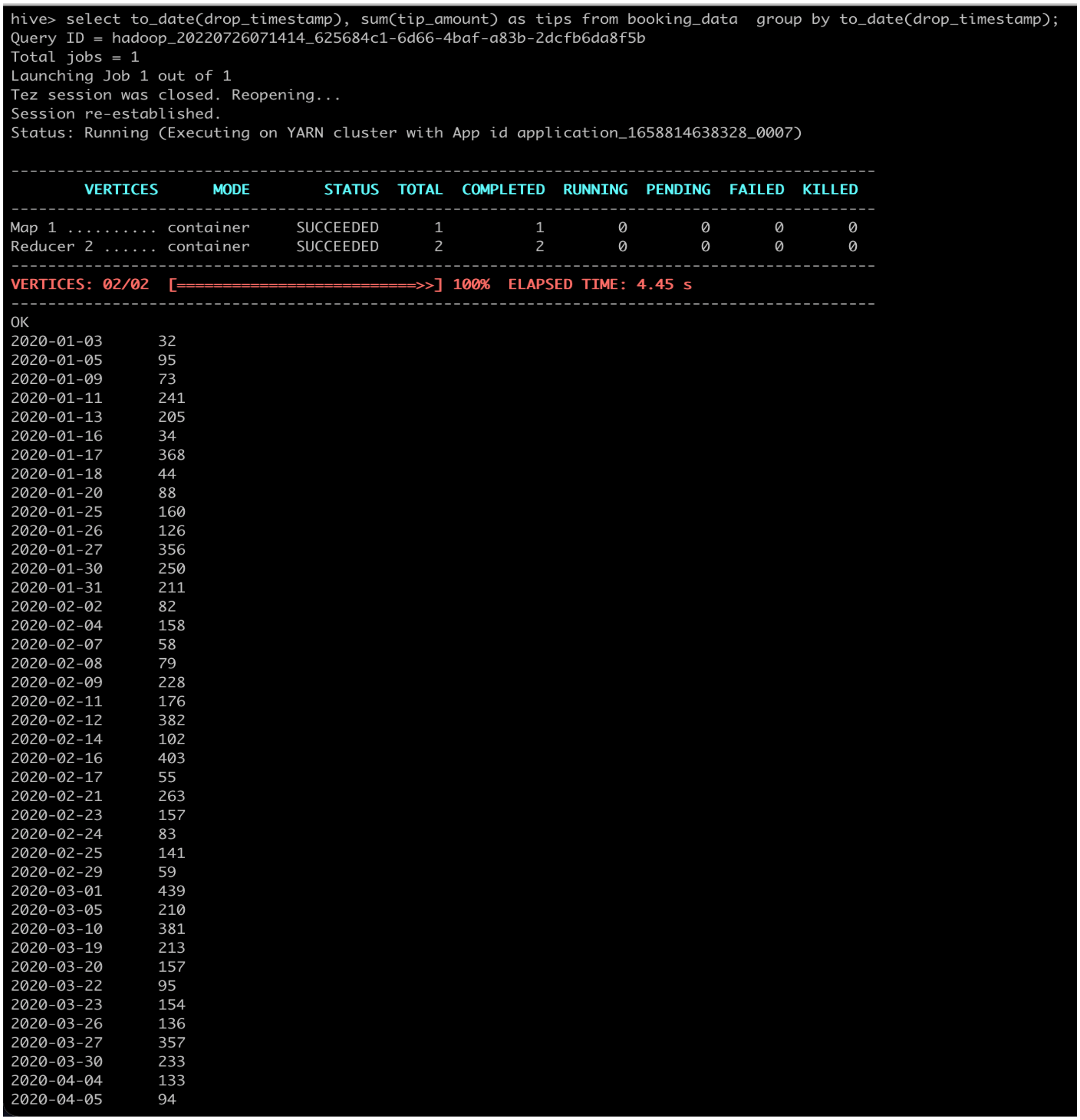


```
hive> select count(*) as count from booking_data where cab_color like '%black%';
Query ID = hadoop_20220726070526_d4b8a055-41fb-4790-af17-c8a2c7d4b547
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658814638328_0006)

----------------------------------------------------------------------------------------------
        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: 4.68 s
----------------------------------------------------------------------------------------------
OK
72
Time taken: 5.15 seconds, Fetched: 1 row(s)
hive> 
```

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

→ Here we have first extracted date from drop_timestamp using to_date function and summed tip_amount from booking_data table and grouped based on date obtained.

select to_date(drop_timestamp), sum(tip_amount) as tips from booking_data group by to_date(drop_timestamp);



```
hive> select to_date(drop_timestamp), sum(tip_amount) as tips from booking_data  group by to_date(drop_timestamp);
Query ID = hadoop_20220726071414_625684c1-6d66-4baf-a83b-2dcfb6da8f5b
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_1658814638328_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: 4.45 s
----------------------------------------------------------------------------------------------
OK
2020-01-03      32
2020-01-05      95
2020-01-09      73
2020-01-11      241
2020-01-13      205
2020-01-16      34
2020-01-17      368
2020-01-18      44
2020-01-20      88
2020-01-25      160
2020-01-26      126
2020-01-27      356
2020-01-30      250
2020-01-31      211
2020-02-02      82
2020-02-04      158
2020-02-07      58
2020-02-08      79
2020-02-09      228
2020-02-11      176
2020-02-12      382
2020-02-14      102
2020-02-16      403
2020-02-17      55
2020-02-21      263
2020-02-23      157
2020-02-24      83
2020-02-25      141
2020-02-29      59
2020-03-01      439
2020-03-05      210
2020-03-10      381
2020-03-19      213
2020-03-20      157
2020-03-22      95
2020-03-23      154
2020-03-26      136
2020-03-27      357
2020-03-30      233
2020-04-04      133
2020-04-05      94
```

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

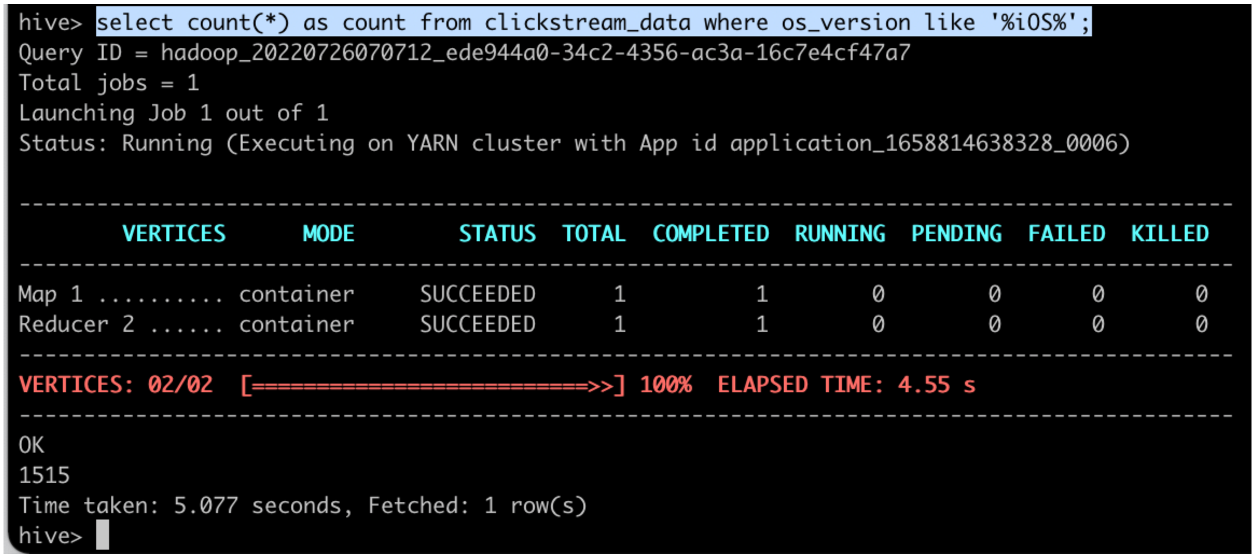
→ Here we have parsed month and year and counted all data with rating_by_customer less than 2 and grouped by year and month.

select year(drop_timestamp) as year,month(drop_timestamp) as month,count(*) from booking_data where rating_by_customer<2 group by year,month;

7. **Calculate the count of total iOS users.**

→ In this query we have simply counted all the data with os_version iOS from clickstream_data. Here also we have used like instead of equal just to avoid any missing data due to case indifference.

select count(*) as count from clickstream_data where os_version like '%iOS%';



```
hive> select count(*) as count from clickstream_data where os_version like '%iOS%';
Query ID = hadoop_20220726070712_ede944a0-34c2-4356-ac3a-16c7e4cf47a7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658814638328_0006)

----------------------------------------------------------------------------------------------
        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: 4.55 s
----------------------------------------------------------------------------------------------
OK
1515
Time taken: 5.077 seconds, Fetched: 1 row(s)
hive> 
```