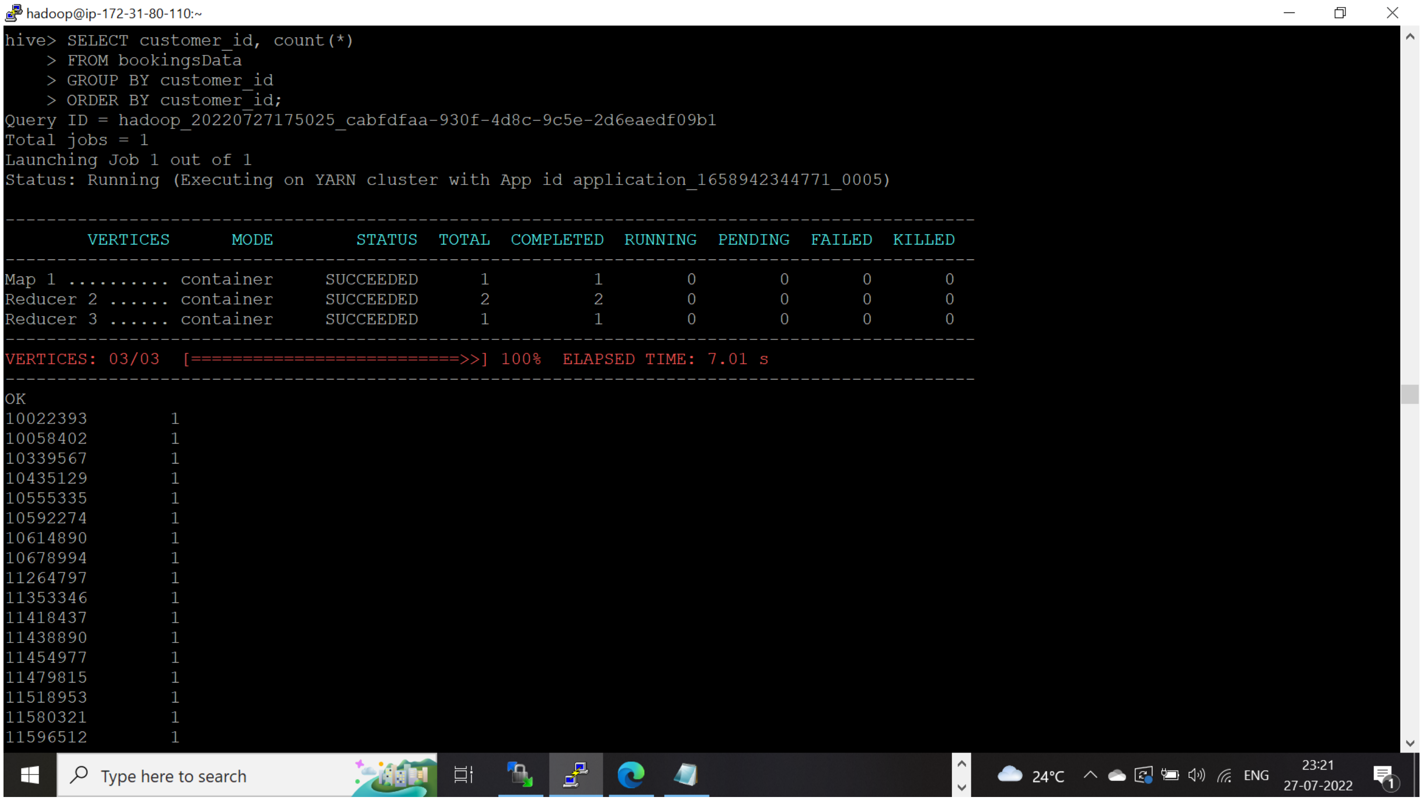
Logic For Final Submission

1. **Hive Query for Task 5:**
   Calculate the total number of different drivers for each customer.
   Here total number of different drivers for each customer is calculated by grouping customer_id and sorted customer_id in ascending order.

```
SELECT customer_id, count(*)
FROM bookingsData
GROUP BY customer_id
ORDER BY customer_id;
```

Screenshot after executing Query:



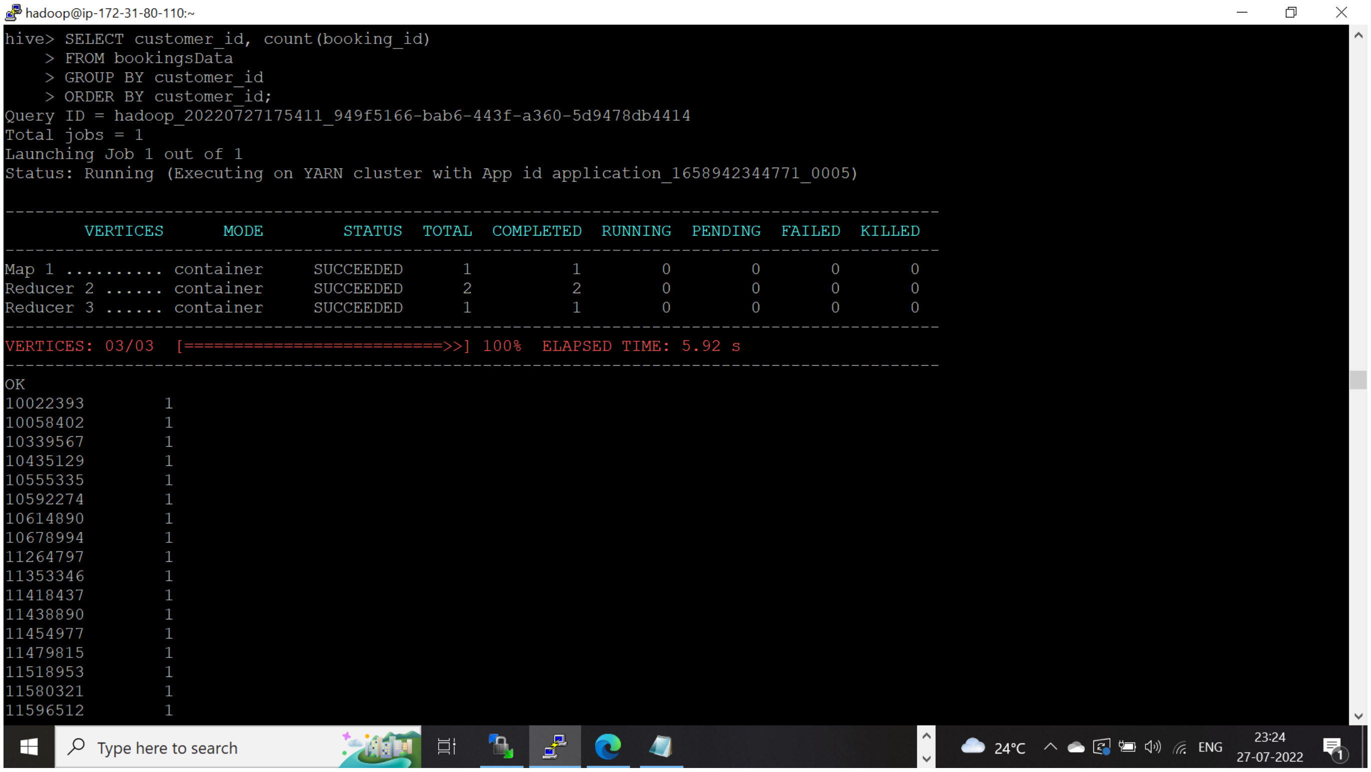
```
hive> SELECT customer_id, count(*)
    > FROM bookingsData
    > GROUP BY customer_id
    > ORDER BY customer_id;
Query ID = hadoop_20220727175025_cabfdfaa-930f-4d8c-9c5e-2d6eaedf09b1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658942344771_0005)

----------------------------------------------------------------------------------------------
        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: 7.01 s
----------------------------------------------------------------------------------------------
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
```

2. **Hive Query for Task 6:**
   Calculate the total rides taken by each customer.
   Here total rides are calculated by considering count of booking_id and grouped by customer_id. Finally sorted the result based on customer_id in ascending order.

```
SELECT customer_id,
count(booking_id)
FROM bookingsData
GROUP BY customer_id
ORDER BY customer_id;
```

Screenshot after executing Query:



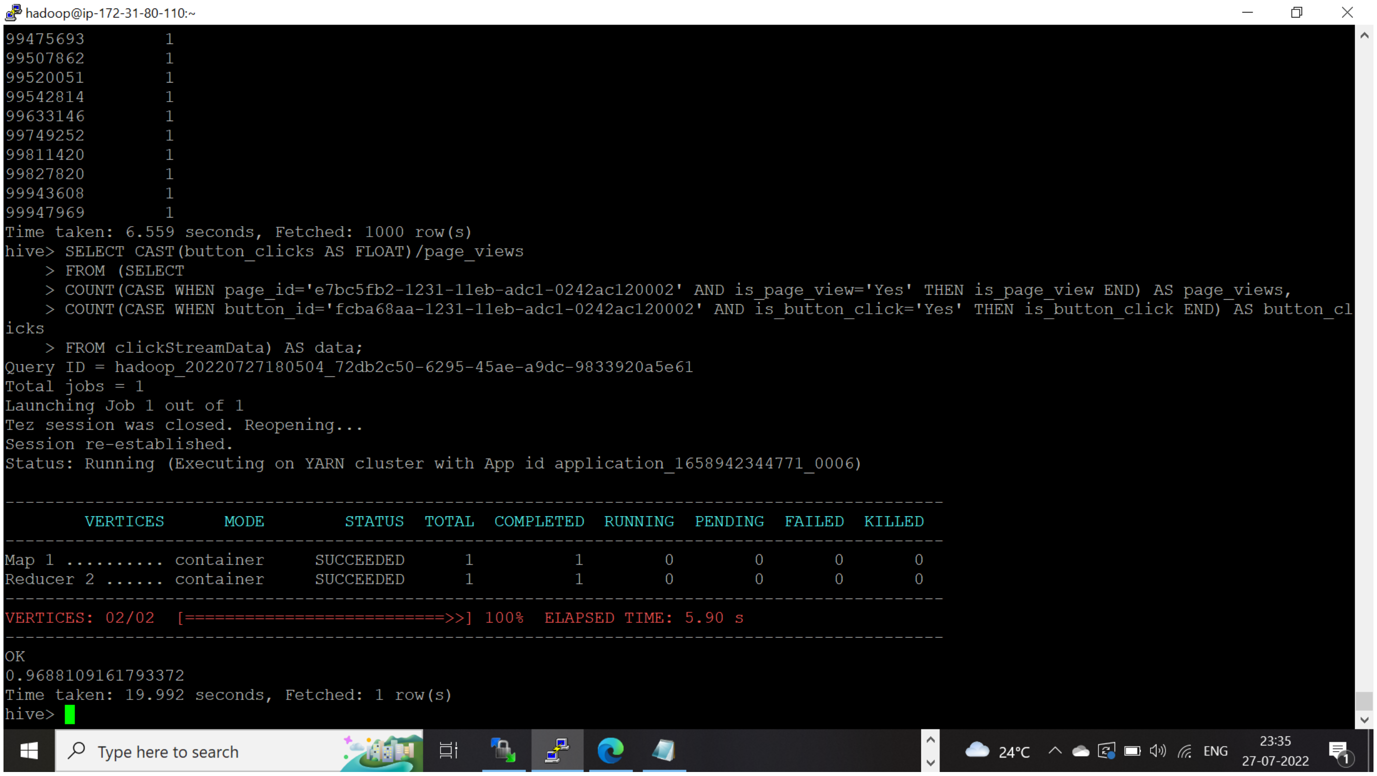
```
hive> SELECT customer_id, count(booking_id)
    > FROM bookingsData
    > GROUP BY customer_id
    > ORDER BY customer_id;
Query ID = hadoop_20220727175411_949f5166-bab6-443f-a360-5d9478db4414
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658942344771_0005)

----------------------------------------------------------------------------------------------
        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: 5.92 s
----------------------------------------------------------------------------------------------
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
```

3. **Hive Query for 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.
   Here total count of button clicks and page views are calculated based on given button_id and page_id using Case statements. Finally conversion ratio is calculated as button_clicks/page_views.

```
SELECT CAST(button_clicks AS FLOAT)/page_views
FROM (SELECT
COUNT(CASE WHEN page_id='e7bc5fb2-1231-11eb-
adc1-0242ac120002' AND is_page_view='Yes' THEN
is_page_view END) AS page_views,
COUNT(CASE WHEN button_id='fcba68aa-1231-11eb-
adc1-0242ac120002' AND is_button_click='Yes'
THEN is_button_click END) AS button_clicks
FROM clickStreamData) AS data;
```

Screenshot after executing Query:



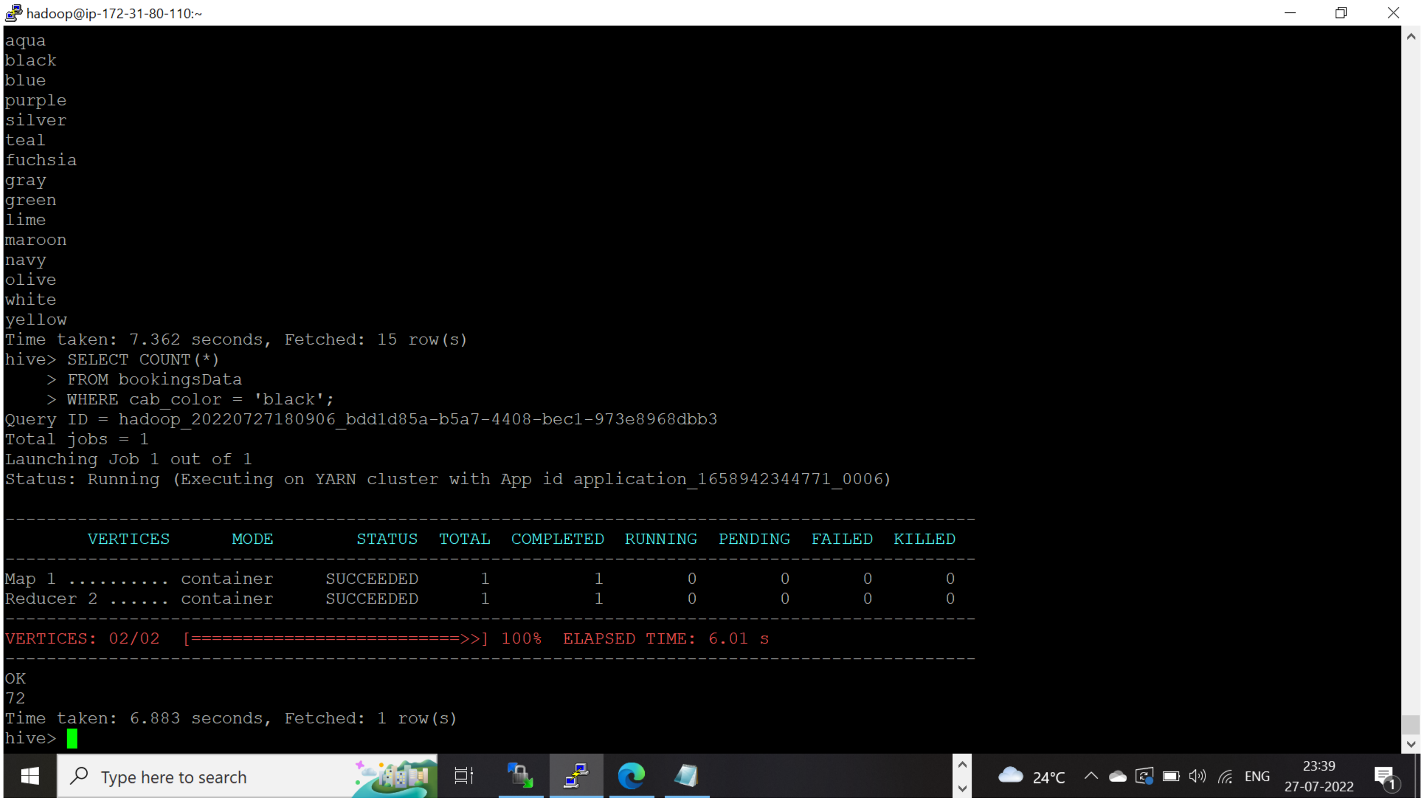
```
99475693        1
99507862        1
99520051        1
99542814        1
99633146        1
99749252        1
99811420        1
99827820        1
99943608        1
99947969        1
Time taken: 6.559 seconds, Fetched: 1000 row(s)
hive> SELECT CAST(button_clicks AS FLOAT)/page_views
    > FROM (SELECT
    > COUNT(CASE WHEN page_id='e7bc5fb2-1231-11eb-adc1-0242ac120002' AND is_page_view='Yes' THEN is_page_view END) AS page_views,
    > COUNT(CASE WHEN button_id='fcba68aa-1231-11eb-adc1-0242ac120002' AND is_button_click='Yes' THEN is_button_click END) AS button_cl
icks
    > FROM clickStreamData) AS data;
Query ID = hadoop_20220727180504_72db2c50-6295-45ae-a9dc-9833920a5e61
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_1658942344771_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: 5.90 s
----------------------------------------------------------------------------------------------
OK
0.9688109161793372
Time taken: 19.992 seconds, Fetched: 1 row(s)
hive>
```

4. **Hive Query for Task 8:**
   Calculate the count of all trips done on black cabs.
   Here count of all trips is calculated by applying filter as cab_color='black'

```
SELECT COUNT(*)
FROM bookingsData
WHERE cab_color = 'black';
```

Screenshot after executing Query:



```
aqua
black
blue
purple
silver
teal
fuchsia
gray
green
lime
maroon
navy
olive
white
yellow
Time taken: 7.362 seconds, Fetched: 15 row(s)
hive> SELECT COUNT(*)
    > FROM bookingsData
    > WHERE cab_color = 'black';
Query ID = hadoop_20220727180906_bdd1d85a-b5a7-4408-bec1-973e8968dbb3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658942344771_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: 6.01 s
----------------------------------------------------------------------------------------------
OK
72
Time taken: 6.883 seconds, Fetched: 1 row(s)
hive>
```

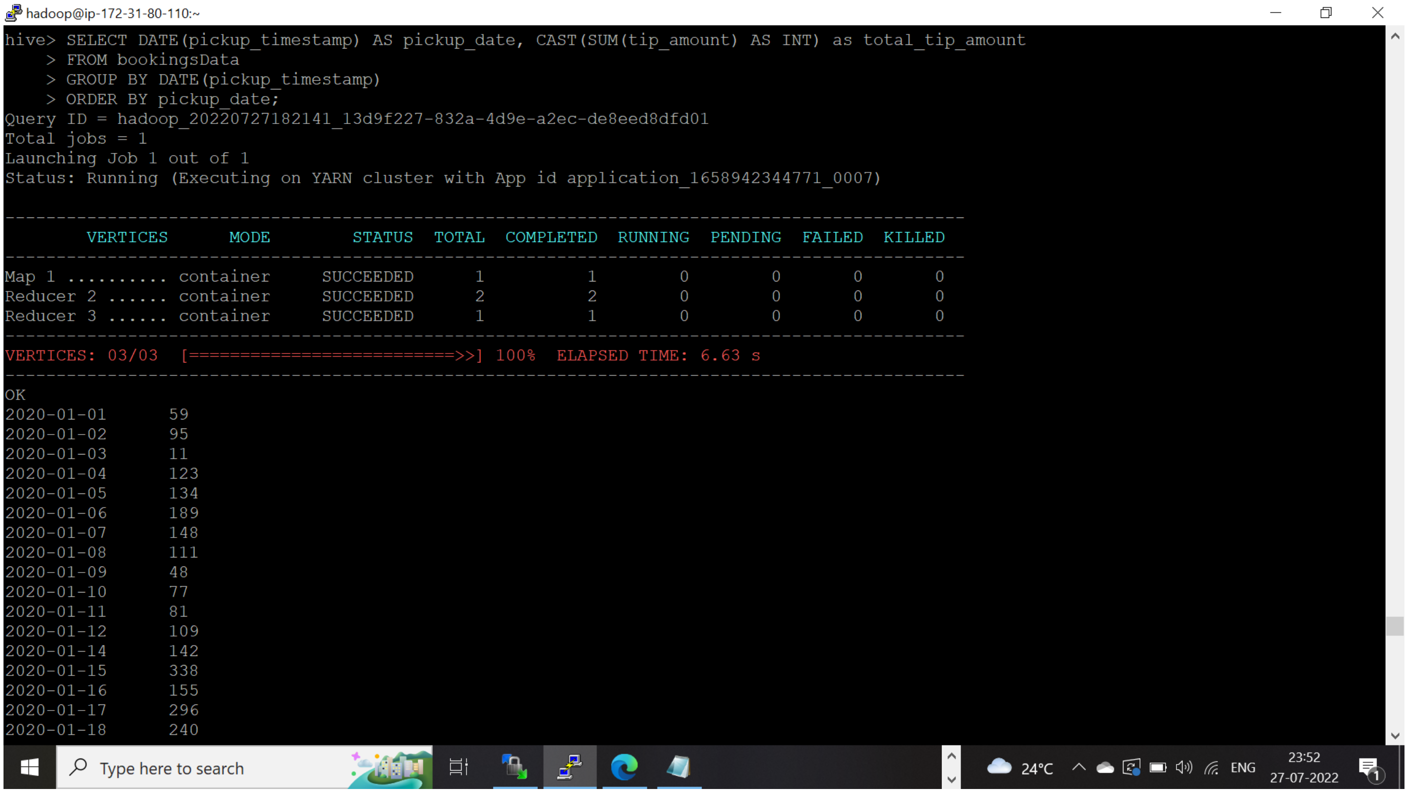
5. Hive Query for Task 9:

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

Here date is extracted from pickup_timestamp. Date wise sum of tip_amount is calculated. Finally the results are sorted based on extracted date in ascending order.

```
SELECT DATE(pickup_timestamp) AS pickup_date,
CAST(SUM(tip_amount) AS INT) as
total_tip_amount
FROM bookingsData
GROUP BY DATE(pickup_timestamp)
ORDER BY pickup_date;
```

Screenshot after executing Query:



hadoop@ip-172-31-80-110:~

```
hive> SELECT DATE(pickup_timestamp) AS pickup_date, CAST(SUM(tip_amount) AS INT) as total_tip_amount
    > FROM bookingsData
    > GROUP BY DATE(pickup_timestamp)
    > ORDER BY pickup_date;
Query ID = hadoop_20220727182141_13d9f227-832a-4d9e-a2ec-de8eed8dfd01
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658942344771_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 6.63 s
----------------------------------------------------------------------------------------------
OK
2020-01-01      59
2020-01-02      95
2020-01-03      11
2020-01-04      123
2020-01-05      134
2020-01-06      189
2020-01-07      148
2020-01-08      111
2020-01-09      48
2020-01-10      77
2020-01-11      81
2020-01-12      109
2020-01-14      142
2020-01-15      338
2020-01-16      155
2020-01-17      296
2020-01-18      240
```

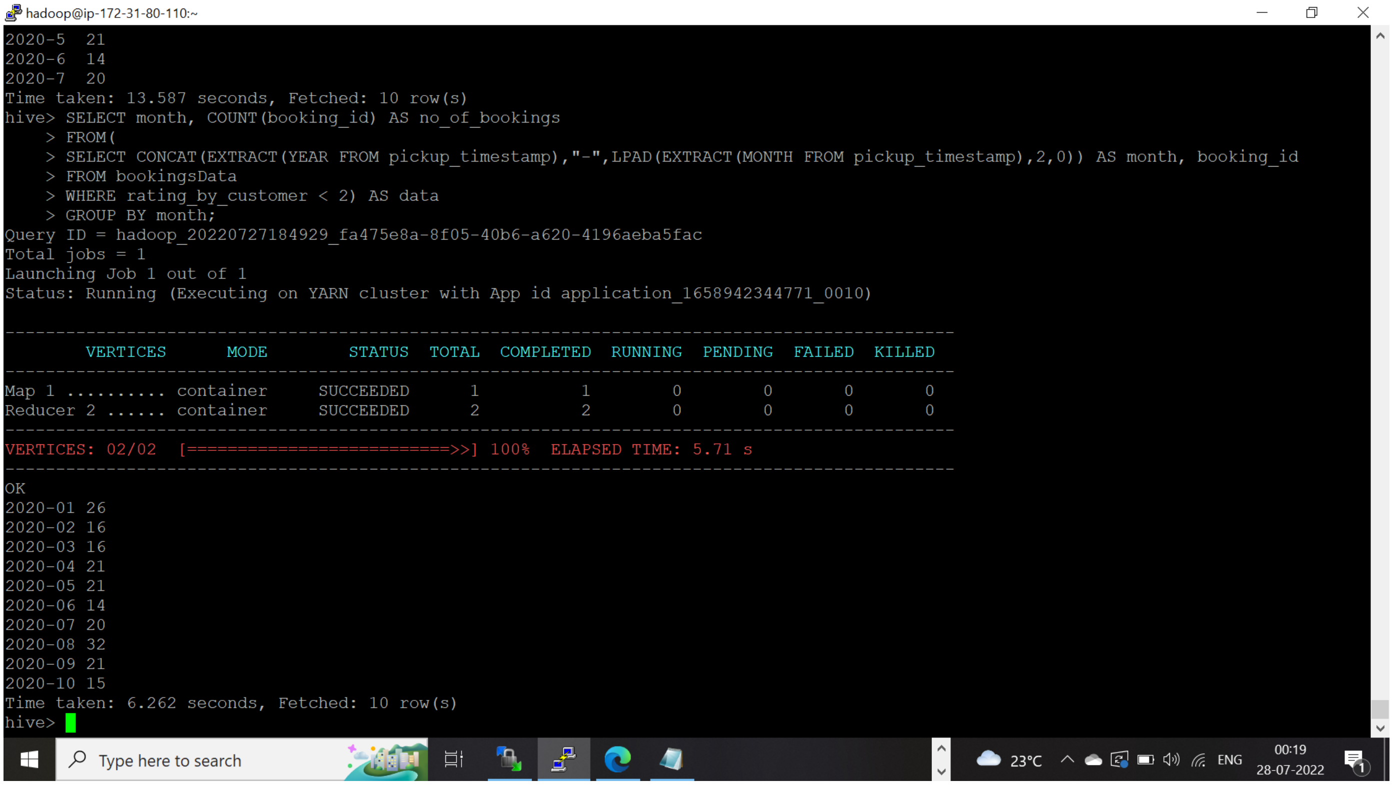
6. Hive Query for Task 10:

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

Here year and month are extracted from pickup_timestamp and month wise count of all bookings is calculated from number of booking_id. Finally applied the filter where rating_by_customer is less than 2.

```
SELECT month, COUNT(booking_id) AS
no_of_bookings
FROM(
SELECT CONCAT(EXTRACT(YEAR FROM
pickup_timestamp),"-",LPAD(EXTRACT(MONTH
FROM pickup_timestamp),2,0)) AS month,
booking_id
FROM bookingsData
WHERE rating_by_customer < 2) AS data
GROUP BY month;
```

Screenshot after executing Query:



```
2020-5  21
2020-6  14
2020-7  20
Time taken: 13.587 seconds, Fetched: 10 row(s)
hive> SELECT month, COUNT(booking_id) AS no_of_bookings
    > FROM(
    > SELECT CONCAT(EXTRACT(YEAR FROM pickup_timestamp),"-",LPAD(EXTRACT(MONTH FROM pickup_timestamp),2,0)) AS month, booking_id
    > FROM bookingsData
    > WHERE rating_by_customer < 2) AS data
    > GROUP BY month;
Query ID = hadoop_20220727184929_fa475e8a-8f05-40b6-a620-4196aeba5fac
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658942344771_0010)

----------------------------------------------------------------------------------------------
        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: 5.71 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: 6.262 seconds, Fetched: 10 row(s)
hive>
```

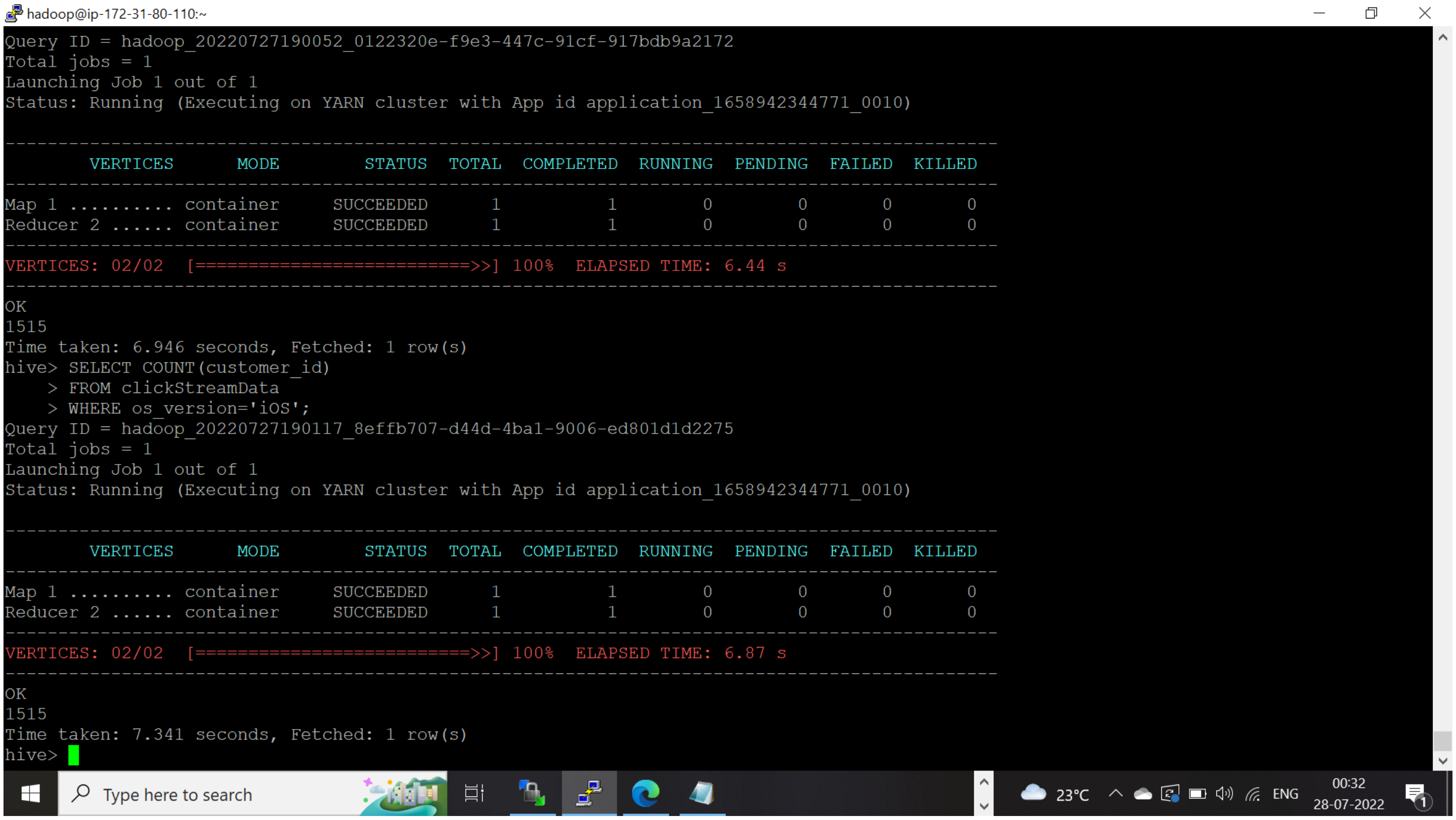
7. Hive Query for Task 11:

Calculate the count of total iOS users.

Here total count of customers is calculated by applying filter as os_version='iOS'.

```
SELECT COUNT(customer_id)
FROM clickStreamData
WHERE os_version='iOS';
```

Screenshot after executing Query:



```
Query ID = hadoop_20220727190052_0122320e-f9e3-447c-91cf-917bdb9a2172
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658942344771_0010)

----------------------------------------------------------------------------------------------
        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: 6.44 s
----------------------------------------------------------------------------------------------
OK
1515
Time taken: 6.946 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(customer_id)
    > FROM clickStreamData
    > WHERE os_version='iOS';
Query ID = hadoop_20220727190117_8effb707-d44d-4ba1-9006-ed801d1d2275
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658942344771_0010)

----------------------------------------------------------------------------------------------
        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: 6.87 s
----------------------------------------------------------------------------------------------
OK
1515
Time taken: 7.341 seconds, Fetched: 1 row(s)
hive>
```