Logic For Final Submission

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

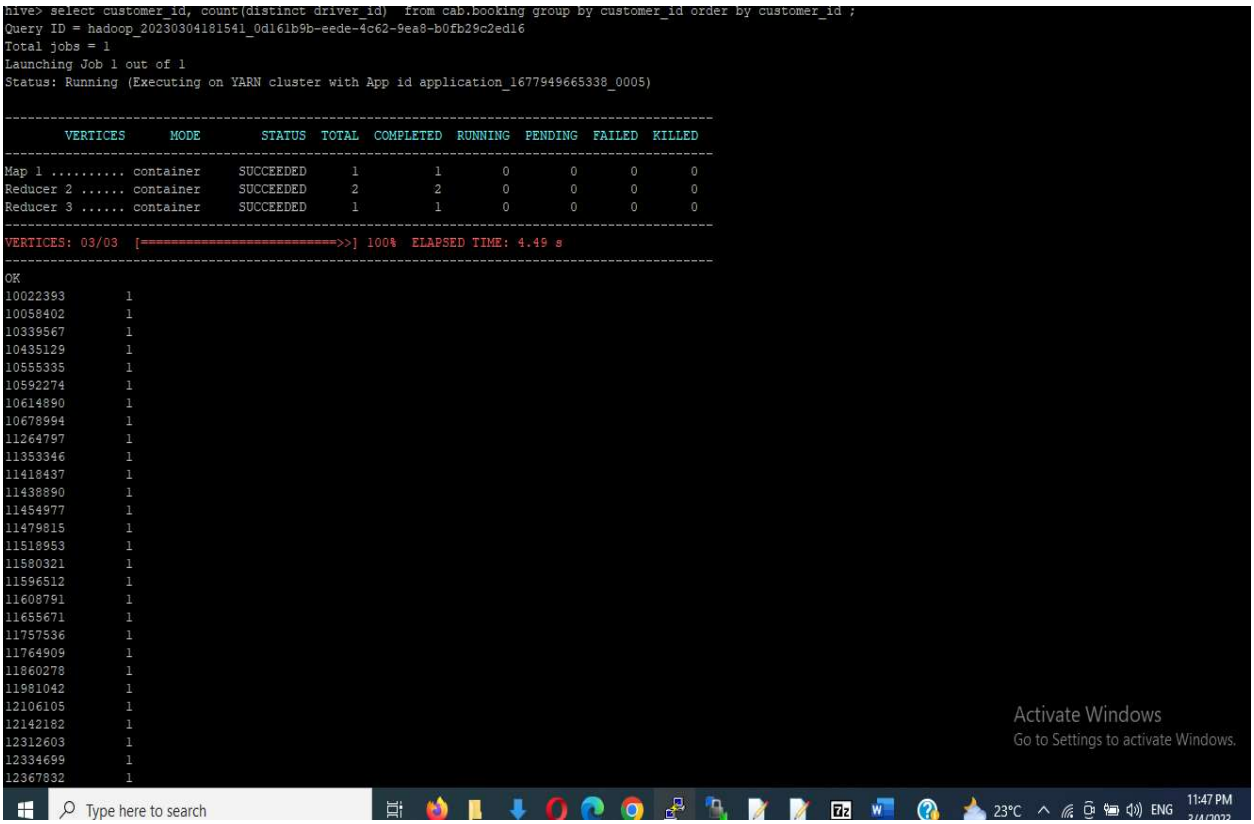
Task 5: Calculate the total number of different drivers for each customer.

```
select customer_id, count(distinct driver_id)  from cab.booking group by customer_id order by customer_id ;
```

Query Explanation

Fetch the customer_id column, count based on the driver_id column from the bookings table, and group the data based on customer_id.

Screenshot after executing Query



Task 6: Calculate the total rides taken by each customer.

Select customer_id, count(*) as rides from cab.booking group by customer_id order by customer_id;

Query Explanation

Fetch the customer_id column, count based on the pickup_lat column from the bookings table, and group the data based on customer_id.

Screenshot after executing Query

```
hive> SELECT customer_id, COUNT(*) AS rides
    > FROM cab. booking
    > GROUP BY customer_id
    > ORDER BY customer_id;
Query ID = hadoop_20230304183301_flebbc2b-d20e-4f90-a6ae-a6b7ddf332ab
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_1677949665338_0006)

--------------------------------------------------------------------------------
        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: 5.05 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
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.

select sum(case when button_id = 'fcba68aa-1231-11eb-adc1-0242ac120002' and is_button_click ='Yes' THEN 1 ELSE 0 END)/sum(case when page_id = 'e7bc5fb2-1231-

11eb-adc1-0242ac120002' and is_page_view = 'Yes' then 1 else 0 end) as conversion_ratio
from cab.clickstream;

Query Explanation

Used CASE to get the total visits made by each customer on the booking page using the button id 'fcba68aa-1231-11eb-adc1-0242ac120002' and sum up the total visits and used CASE to get the total Book Now button presses with button_id 'e7bc5fb2-1231-11eb-adc1-0242ac120002' and sum up the total button clicks. Divided the total visits and total button clicks and convert the value to decimal datatype with 4 decimal places.

Screenshot after executing Query

```
hive> create table cab.clickstream(customer_id string,
    > app_version string,
    > os_version string,
    > lat decimal(10,2), lon decimal(10,2), page_id string,
    > button_id string, is_button_click string,
    > is_page_view string, is_scroll_up string, is_scroll_down string, date_timestamp string)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > LOCATION '/user/capstone/clickstream/';
OK
Time taken: 0.287 seconds
hive> select sum(case when button_id = 'fcba68aa-1231-11eb-adc1-0242ac120002' and is_button_click ='Yes' THEN
adc1-0242ac120002' and is_page_view = 'Yes' then 1 else 0 end) as conversion_ratio
    > from cab.clickstream;
Query ID = hadoop_20230314103035_80847d82-f741-41c4-99ba-11805aa4fd48
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1678787986655_0002)

--------------------------------------------------------------------------------
        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: 4.99 s
--------------------------------------------------------------------------------
OK
0.9688109161793372
Time taken: 8.44 seconds, Fetched: 1 row(s)
```

Task 8: Calculate the count of all trips done on black cabs.

Select count(*) from cab.booking where cab_color = 'black';

Query Explanation

Fetch the records from the bookings table with cab_color 'Black' and count the data based on the booking_id column.

Screenshot after executing Query

```
hive> Select count(*) from cab.booking where cab_color = 'black';
Query ID = hadoop_20230304184606_4f9a0887-2410-41cd-8ff6-6b58d2fbf58b
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_1677949665338_0007)

--------------------------------------------------------------------------------
        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: 4.42 s
--------------------------------------------------------------------------------
OK
72
Time taken: 9.366 seconds, Fetched: 1 row(s)
```


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

```
Select DATE(pickup_timestamp)as bookingdate, cast(sum(tip_amount)as
decimal(10,0))as tips from cab.booking
group by DATE(pickup_timestamp)
order by bookingdate;
```

Query Explanation

Fetch the pickup date total of tip_amount from booking tables and group the data based on pickup_date to aggregate the data datewise and sorted it based on date.

Screenshot after executing Query

```
hive> Select DATE(pickup_timestamp)as bookingdate, cast(sum(tip_amount)as
    > decimal(10,0))as tips from cab.booking group by DATE(pickup_timestamp) order by bookingdate;
Query ID = hadoop_20230304185532_92658e17-cc5e-4567-a48b-199f8279476d
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_1677949665338_0008)

--------------------------------------------------------------------------------
        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: 4.92 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
2020-01-20      210
2020-01-21      5
2020-01-23      148
2020-01-24      472
2020-01-25      98
2020-01-26      209
2020-01-27      231
```

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

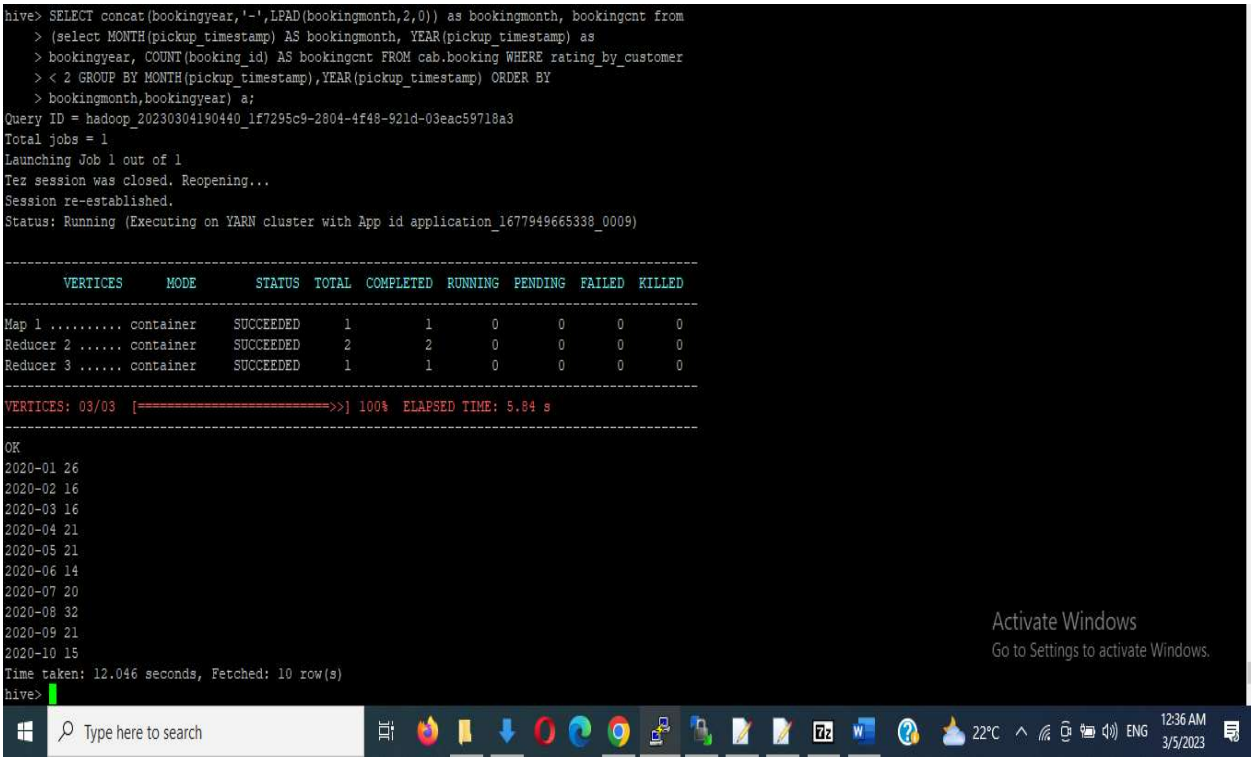
```
SELECT concat(bookingyear,'-',LPAD(bookingmonth,2,0)) as bookingmonth, bookingcnt
from(select MONTH(pickup_timestamp) AS bookingmonth, YEAR(pickup_timestamp) as
bookingyear, COUNT(booking_id) AS bookingcnt FROM cab.booking WHERE
rating_by_customer < 2
GROUP BY MONTH(pickup_timestamp),YEAR(pickup_timestamp)
ORDER BY bookingmonth,bookingyear) a;
```

Query Explanation

Used subquery to fetch the month and year from pickup_timestamp column and count of bookings based on booking id from the bookings table if the rating_by_customer column has a value less than 2. Then Grouped the data based on the month and year of pickup_timestamp and sorted based on month and year.

Fetch year and month and concatenate them and used LPAD function to left-pads 0 with month, to a certain length 2 and count of bookings from the subquery.

Screenshot after executing Query



Task 11: Calculate the count of total iOS users.

```
SELECT COUNT(*)
from cab.clickstream
where os_version = 'iOS';
```

Query Explanation

Fetch the count of ios users data from the clickstream table if OS_version column has the value ‘IOS’ and used UCASE to avoid the letter case mismatch.

Screenshot after executing Query

```
hive>
    >
    > SELECT COUNT(*)
    > from cab.clickstream
    > where os_version = 'iOS';
Query ID = hadoop_20230314103242_5796bcfd-c1b5-46de-b1f3-cc4b5697d919
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1678787986655_0002)

--------------------------------------------------------------------------------
        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: 3.87 s
--------------------------------------------------------------------------------
OK
1515
Time taken: 4.484 seconds, Fetched: 1 row(s)
hive>
```