Logic For Final Submission

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

** While loading the data please add the following code
Load data inpath ‘’ into table ‘’ -> changes as per the query
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','

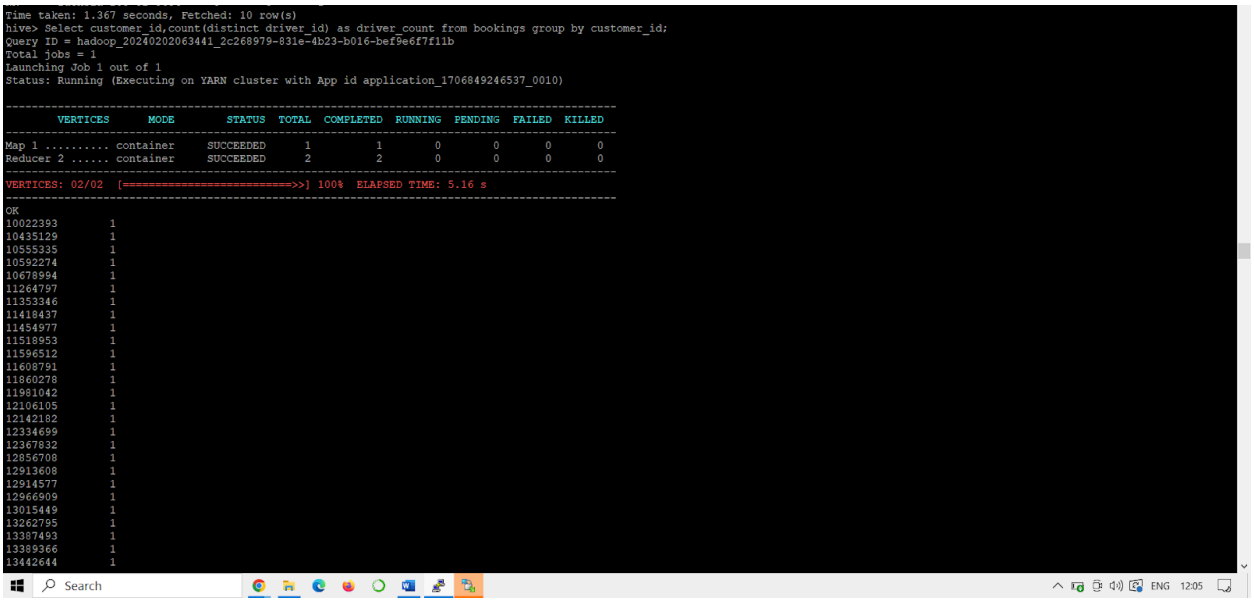
<u><Hive Query for Task 5></u>

<u>Task 5: Calculate the total number of different drivers for each customer.</u>

- For This task we need to aggregate the numbers of driver per each customer.
- We use the hive Bookings table.
- We use the Count(distinct driver_id) to count the number of unique drivers who have been booked and aggregate this at customer_id level
- **The query:**

```
Select customer_id,count(distinct driver_id)
From bookings
Group by customer_id
```

- **<Screenshot after executing Query>**



```
Time taken: 1.367 seconds, Fetched: 10 row(s)
hive> Select customer_id,count(distinct driver_id) as driver_count from bookings group by customer_id;
Query ID = hadoop_20240202063441_2c268979-831e-4b23-b016-bef9e6f7f11b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_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: 5.16 s
--------------------------------------------------------------------------------
OK
10022393        1
10435129        1
10555335        1
10592274        1
10678994        1
11264797        1
11353346        1
11418437        1
11454977        1
11518953        1
11596512        1
11608791        1
11860278        1
11981042        1
12106105        1
12142182        1
12334699        1
12367832        1
12856708        1
12913608        1
12914577        1
12966909        1
13015449        1
13262795        1
13387493        1
13389366        1
13442644        1
```

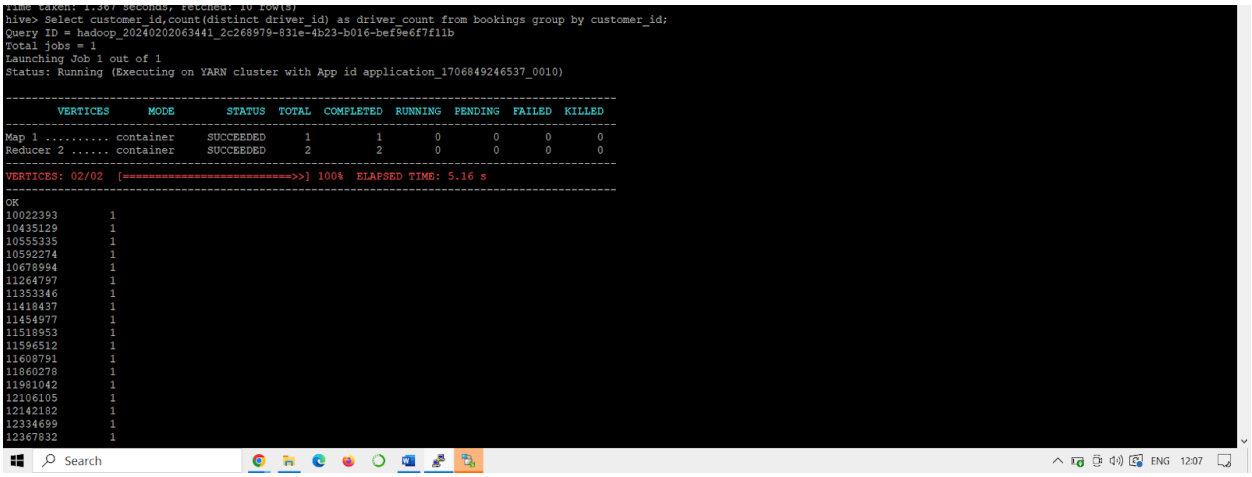
<Hive Query for Task 6>

Task 6: Calculate the total rides taken by each customer

- For this task we are required to calculate the aggregate the total number of rides booked by each customer
- We use the bookings hive table for the same
- We aggregate the rides by counting unique bookings using the booking is as "count(distinct booking_id)" and aggregate it by grouping it at customer level
- The query :

```
Select customer_id, count(distinct booking_id) as  booking_count
From bookings
Group by customer_id
```

- **<Screenshot after executing Query>**



```
hive> Select customer_id,count(distinct driver_id) as driver_count from bookings group by customer_id;
Query ID = hadoop_20240202063441_2c268979-831e-4b23-b016-bef9e6f7f11b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_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: 5.16 s
--------------------------------------------------------------------------------------------------
OK
10022393        1
10435129        1
10555335        1
10592274        1
10678994        1
11264797        1
11353346        1
11418437        1
11454977        1
11518953        1
11596512        1
11608791        1
11860278        1
11981042        1
12106105        1
12142182        1
12334699        1
12367832        1
```

<Hive Query for Task 7>

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.

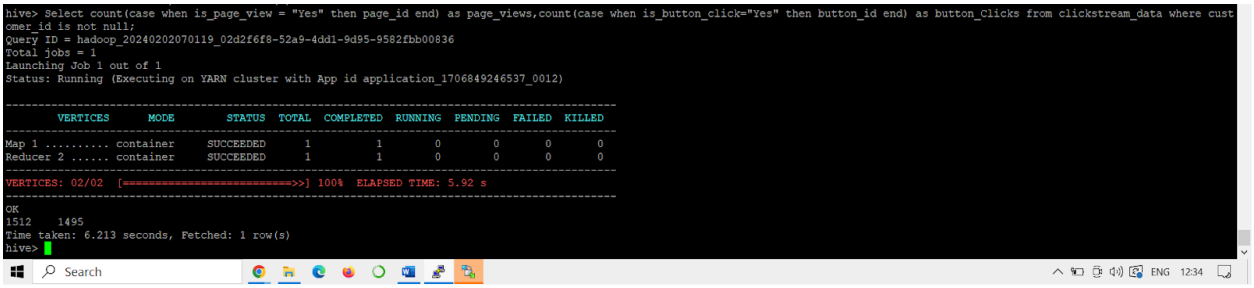
- We need to find the conversion ratio-> we use the clickstream_data table
- The conversion ratio is defined as the total number of button clicks divided by the total number of page_views
- We aggregate the total number of clicks using case when within the count -> count the button_id only when the is_button_click="Yes" in the row
- We aggregate the total number of page_views using case when within the count -> count the page_id only when the is_page_view ="Yes" in the row
- Then we divide the above two points to find the conversion ratio
- The Query :

```
Select count(case when is_button_click="Yes" then button_id end)/count(case when
is_page_view = "Yes" then page_id end)  as conversion_ratio
from clickstream_data
where customer_id is not null;
```

- **<Screenshot after executing Query>**

** The below query is to find the total number of page_views and button_clicks separately

```
Select count(case when is_page_view = "Yes" then page_id end) as page_views ,
count(case when is_button_click="Yes" then button_id end) as button_click
from clickstream_data
where customer_id is not null;
```



```
hive> Select count(case when is_page_view = "Yes" then page_id end) as page_views,count(case when is_button_click="Yes" then button_id end) as button_Clicks from clickstream_data where cust
omer_id is not null;
Query ID = hadoop_20240202070119_02d2f6f8-52a9-4dd1-9d95-9582fbb00836
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0012)

--------------------------------------------------------------------------------
        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: 5.92 s
--------------------------------------------------------------------------------
OK
1512    1495
Time taken: 6.213 seconds, Fetched: 1 row(s)
hive>
```

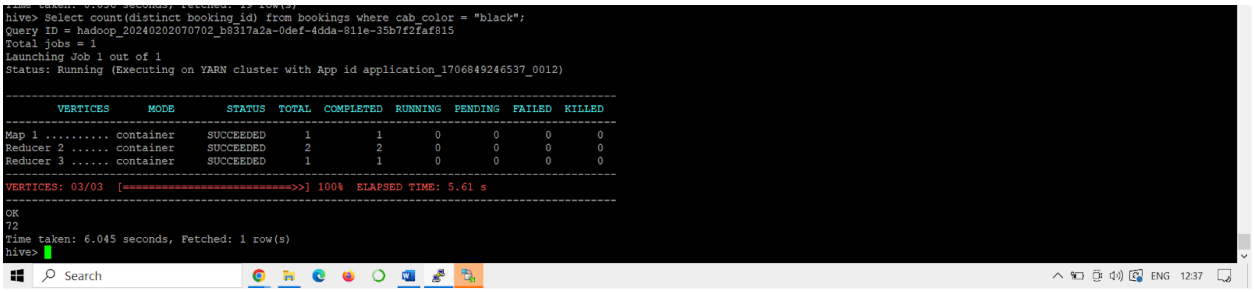
<Hive Query for Task 8>

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

- In this task we need caluclate the total trips done only on black cabs
- We will use the bookings Hive table
- We will first filter out the rows which have cab_color='black' using the where clause
- The aggregate the booking_id using count function
- The query:

```
Select count(distinct booking_id)
from bookings
where cab_Color='black'
```

- **<Screenshot after executing Query>**



```
hive> Select count(distinct booking_id) from bookings where cab_color = "black";
Query ID = hadoop_20240202070702_b8317a2a-0def-4dda-811e-35b7f2faf815
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0012)

--------------------------------------------------------------------------------
        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: 5.61 s
--------------------------------------------------------------------------------
OK
72
Time taken: 6.045 seconds, Fetched: 1 row(s)
hive>
```

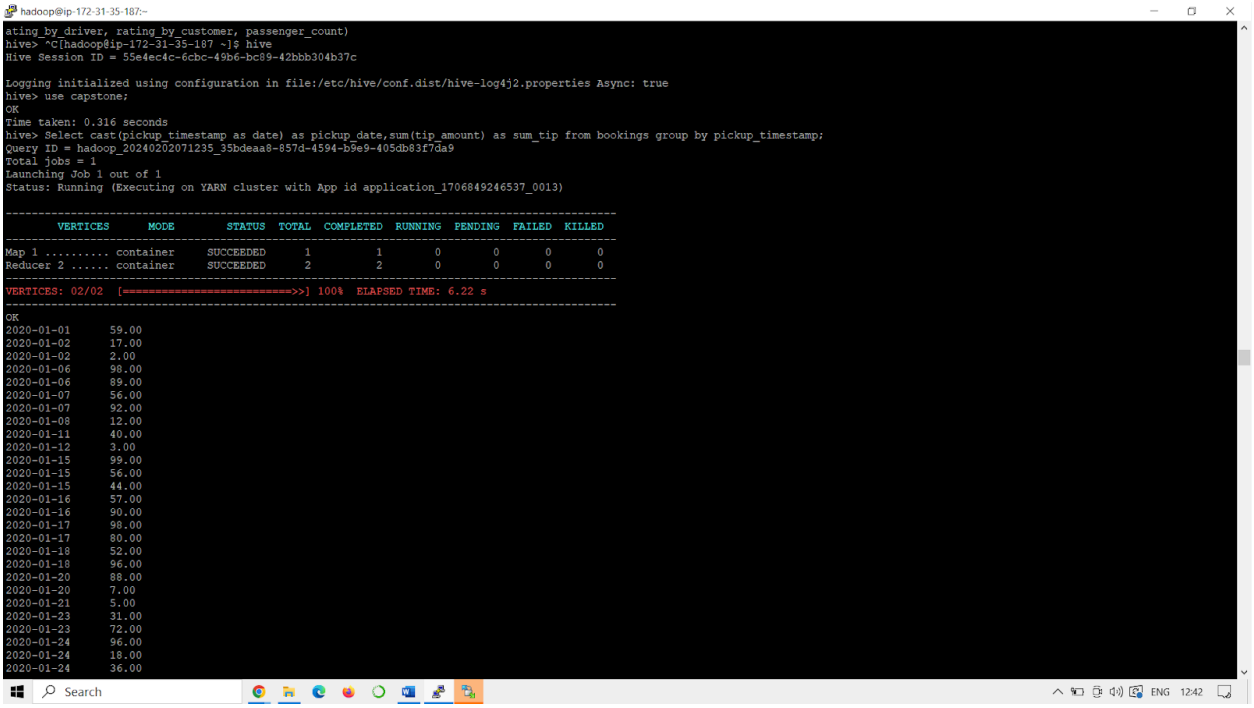
<Hive Query for Task 9>

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

- For this task we need to calculate the aggregate sum of the tips received by drivers on each date.
- We use the bookings table for the same
- We aggregate the sum of the tips by using the sum function over column tip_amount: sum(tip_amount) and group the same at date level using the column pickup_timestamp
- Since the pickup_timestamp is a timestamp data type we cast is to date type using cast function cast(pickup_timestamp as date) as pickup_date
- The query :

```
Select cast(pickup_timestamp as date) as pickup_date,sum(tip_amount) as sum_tip
from bookings
group by pickup_timestamp;
```

- **<Screenshot after executing Query>**



```
hadoop@ip-172-31-35-187:~
ating_by_driver, rating_by_customer, passenger_count)
hive> ^C[hadoop@ip-172-31-35-187 ~]$ hive
Hive Session ID = 55e4ec4c-6cbc-49b6-bc89-42bbb304b37c

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> use capstone;
OK
Time taken: 0.316 seconds
hive> Select cast(pickup_timestamp as date) as pickup_date,sum(tip_amount) as sum_tip from bookings group by pickup_timestamp;
Query ID = hadoop_20240202071235_35bdeaa8-857d-4594-b9e9-405db83f7da9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0013)

--------------------------------------------------------------------------------
        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: 6.22 s
--------------------------------------------------------------------------------
OK
2020-01-01      59.00
2020-01-02      17.00
2020-01-02      2.00
2020-01-06      98.00
2020-01-06      89.00
2020-01-07      56.00
2020-01-07      92.00
2020-01-08      12.00
2020-01-11      40.00
2020-01-12      3.00
2020-01-15      99.00
2020-01-15      56.00
2020-01-15      44.00
2020-01-16      57.00
2020-01-16      90.00
2020-01-17      98.00
2020-01-17      80.00
2020-01-18      52.00
2020-01-18      96.00
2020-01-20      88.00
2020-01-20      7.00
2020-01-21      5.00
2020-01-23      31.00
2020-01-23      72.00
2020-01-24      96.00
2020-01-24      18.00
2020-01-24      36.00
```

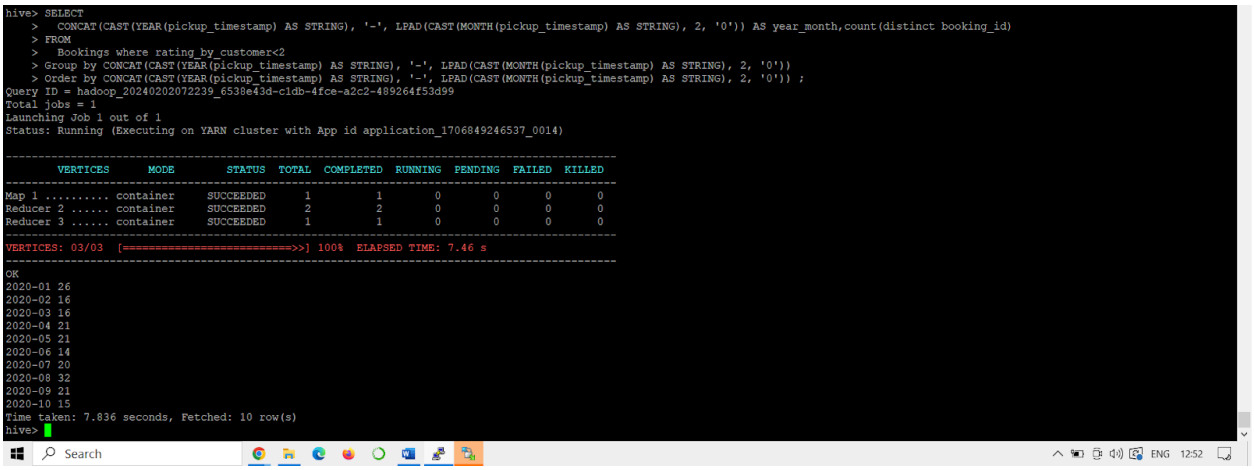
<Hive Query for Task 10>

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 task we need to calculate the total number of bookings which had lower rating than 2 given by the customer every month.
- For this we will use the bookings Hive table
- We need to aggregate the total number of bookings at month level:
- In order to calculate at month level we use cast and concat to generate the month year data using the following “CONCAT(CAST(YEAR(pickup_timestamp) AS STRING), '-', LPAD(CAST(MONTH(pickup_timestamp) AS STRING), 2, '0')) AS year_month”
    - Here we use to find the year using the Year() -> YEAR(pickup_timestamp)
    - We pull the month value using Month() function -> MONTH(pickup_timestamp)
    - We use Lpad(string,2,’0’) to add trailing 0 at the left side of the string to max of 2 values so that we can have month values as 02 for February
    - Then we concatenate the two to get 2022-02 format of data YYYY-MM
- We aggregate the number of bookings using : count(distinct booking_id)
- The query :

```
SELECT
CONCAT(CAST(YEAR(pickup_timestamp) AS STRING), '-',
LPAD(CAST(MONTH(pickup_timestamp) AS STRING), 2, '0')) AS year_month,count(distinct
booking_id)
FROM
  Bookings where rating_by_customer<2
Group by CONCAT(CAST(YEAR(pickup_timestamp) AS STRING), '-',
LPAD(CAST(MONTH(pickup_timestamp) AS STRING), 2, '0'))
Order by CONCAT(CAST(YEAR(pickup_timestamp) AS STRING), '-',
LPAD(CAST(MONTH(pickup_timestamp) AS STRING), 2, '0')) ;
```

- **<Screenshot after executing Query>**



<Hive Query for Task 11>

Task 11: Calculate the count of total iOS users.

- In this task we are required to pull the total count of users who use iOS
- For this we will use clickstream_data
- We will filter out rows using “where OS_version='iOS'”
- We will then simply aggregate the distinct customer_id using “count(distinct customer_id)”

```
Select count(distinct customer_id)
from clickstream_data
where OS_version='iOS';
```

- **<Screenshot after executing Query>**

```
hive> Select count(distinct customer_id) from clickstream_data where OS_version='iOS';
Query ID = hadoop_20240202072459_9cf5f7f6-4bce-4ed6-be3d-57b93a4e199c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706849246537_0014)

--------------------------------------------------------------------------------
        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: 5.44 s
--------------------------------------------------------------------------------
OK
1515
Time taken: 5.764 seconds, Fetched: 1 row(s)
hive>
```