Logic For Final Submission

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

Printing two columns customer_id and count of different drivers taken. Group by is done on CustomerId as we need query for each customer. The result is sorted in ascending order.
> SELECT customer_id, count(DISTINCT driver_id) FROM bookings_data GROUP BY customer_id ORDER BY customer_id ASC;
We can see the output below

```
hive> SELECT customer_id, count(DISTINCT driver_id) FROM bookings_data GROUP BY customer_id ORDER BY customer_id ASC;
Query ID = hadoop_20220123184004_57ca58fa-2d31-47c5-9da6-8fefef140d64
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1642920539798_0041)

--------------------------------------------------------------------------------
        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: 4.63 s
--------------------------------------------------------------------------------
OK
NULL    0
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
11981042        1
12106105        1
12142182        1
12312603        1
12334699        1
12367832        1
12856708        1
12885363        1
12913608        1
12914577        1
12966909        1
13015449        1
13229062        1
```

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

Same as above, we print customer_id and count of total bookings. Group by is done on customer_id, which is then sorted.

SELECT customer_id, COUNT(DISTINCT booking_id) FROM bookings_data GROUP BY customer_id ORDER BY customer_id ASC;

```
hive> SELECT customer_id, COUNT(DISTINCT booking_id) FROM bookings_data GROUP BY customer_id ORDER BY customer_id ASC;
Query ID = hadoop_20220123184656_cfa2832c-5c7f-4d33-9812-052ea44a2afb
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1642920539798_0042)

--------------------------------------------------------------------------------
        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: 3.96 s
--------------------------------------------------------------------------------
OK
NULL    1
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
11981042        1
12106105        1
12142182        1
12312603        1
12334699        1
12367832        1
12856708        1
12885363        1
12913608        1
12914577        1
12966909        1
13015449        1
13229062        1
13262795        1
13356177        1
13387493        1
13389366        1
13442644        1
13500355        1
13590084        1
13791801        1
13798100        1
14011511        1
14143225        1
14236627        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.

Find the total visits made by each customer on the booking page

This query will show all the customers who have opened booking page with counts.
This is achieved with use of WHERE clause on page_id.

```
select customer_id, COUNT(page_id) from click_stream_data where
page_id = 'e7bc5fb2-1231-11eb-adc1-0242ac120002' GROUP BY customer_id
ORDER BY customer_id ASC;
```

total 'Book Now' button presses

This query will show all the customers who have pressed book_now button.
This is achieved with use of where clause on button_id

```
select customer_id, COUNT(button_id) from click_stream_data where
button_id = 'fcba68aa-1231-11eb-adc1-0242ac120002' GROUP BY
customer_id ORDER BY customer_id ASC;
```

Conversion ratio

From the output of the above queries shows that each customers has opened booking page and pressed book_now max 1 time, we can find number of book button pressed, divided by home page visit. Following query does the same.

```
select x.result / y.result from (select count(customer_id) as result
from click_stream_data where page_id = 'e7bc5fb2-1231-11eb-adc1-
0242ac120002') y join (select count(customer_id) as result from
click_stream_data where button_id = 'fcba68aa-1231-11eb-adc1-
0242ac120002') x on 1=1;
```

SCREENSHOTS OF THE ABOVE QUERIES

Find the total visits made by each customer on the booking page

```
hive> select customer_id, COUNT(page_id) from click_stream_data where page_id = 'e7bc5fb2-1231-11eb-adc
1-0242ac120002' GROUP BY customer_id ORDER BY customer_id ASC;
Query ID = hadoop_20220123193554_9c91bd39-7778-4ec0-b2ce-bb0d4ea0fe51
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1642920539798_0044)

--------------------------------------------------------------------------------
        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: 4.64 s
--------------------------------------------------------------------------------
OK
10168879        1
10276292        1
10405598        1
10463231        1
10707209        1
10917583        1
10985972        1
11234701        1
11372759        1
11439057        1
11459135        1
11617260        1
11702141        1
11970941        1
12252116        1
12275339        1
12388855        1
12609914        1
12635200        1
12648576        1
12731678        1
13014916        1
13042136        1
13066424        1
13125118        1
13172005        1
13219572        1
13222167        1
13288349        1
13593893        1
13785948        1
13867614        1
13948107        1
14004235        1
14111800        1
14147392        1
14171711        1
14197474        1
14281485        1
14329925        1
```

total ‘Book Now’ button presses

```
hive> select customer_id, COUNT(button_id) from click_stream_data where button_id = 'fcba68aa-1231-11eb
-adc1-0242ac120002' GROUP BY customer_id ORDER BY customer_id ASC;
Query ID = hadoop_20220123193657_509a4d4b-c37d-4529-b855-4ac50828806c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1642920539798_0044)

--------------------------------------------------------------------------------
        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
10097931        1
10276292        1
10303507        1
10318382        1
10405598        1
10697432        1
10800309        1
11037726        1
11235483        1
11439057        1
11651952        1
11970941        1
11980742        1
11988474        1
12089943        1
12269901        1
12452446        1
12635200        1
12636650        1
```

Conversion ratio = 0.9852

```
hive> select x.result / y.result from (select count(customer_id) as result from click_stream_data where
 page_id = 'e7bc5fb2-1231-11eb-adc1-0242ac120002') y join (select count(customer_id) as result from cli
ck_stream_data where button_id = 'fcba68aa-1231-11eb-adc1-0242ac120002') x on 1=1;
Warning: Map Join MAPJOIN[21][bigTable=?] in task 'Reducer 2' is a cross product
Query ID = hadoop_20220123193742_90956543-08ab-4a6d-919f-2e35c0c2af63
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1642920539798_0044)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 3 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 5.47 s
--------------------------------------------------------------------------------
OK
0.985207100591716
```

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

We find black cabs by using WHERE clause on cab_colour. We also group by cab_color to count different driver_id.

SELECT count(distinct driver_id) FROM bookings_data WHERE cab_color IN ('black') GROUP BY cab_color;

Screenshot:

```
hive> SELECT count(distinct driver_id) FROM bookings_data WHERE cab_color IN ('black') GROUP BY cab_col
or;
Query ID = hadoop_20220123194052_57103c85-d78f-4122-9653-bb8597e8ece9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1642920539798_0044)

--------------------------------------------------------------------------------
        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      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.10 s
--------------------------------------------------------------------------------
OK
72
Time taken: 4.9 seconds, Fetched: 1 row(s)
```

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

We use SUM method over GROUP BY to find SUM of given INT field (here: tip_amount). Group by is done on pickup_timestamp.

QUERY:
SELECT DATE_FORMAT(pickup_timestamp,'yyyy-MM-dd'), SUM(tip_amount) FROM bookings_data GROUP BY date_format(pickup_timestamp, 'yyyy-MM-dd') ORDER BY MIN(pickup_timestamp) ASC;

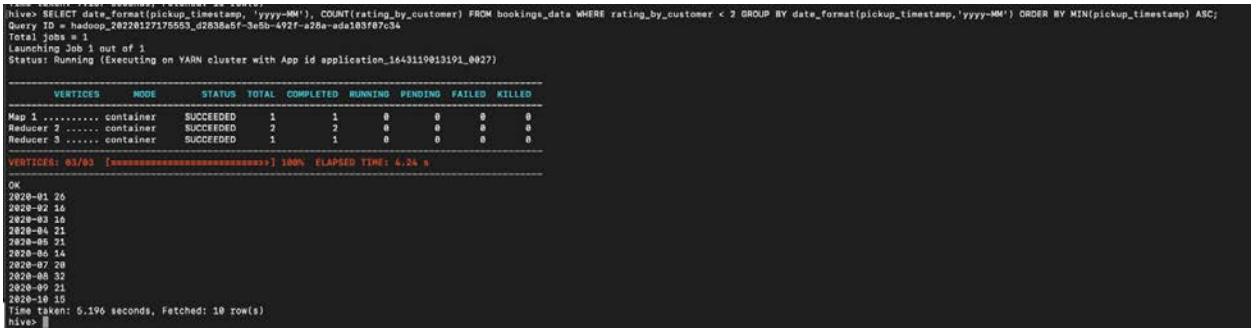
```
hive> SELECT DATE_FORMAT(pickup_timestamp,'yyyy-MM-dd'), SUM(tip_amount) FROM bookings_data GROUP BY date_format(pickup_timestamp, 'yyyy-MM-dd') ORDER BY MIN(pickup_timestamp) ASC;
Query ID = hadoop_20220127175747_31229f6d-ab07-40f1-b2ec-b69815fa7664
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1643119013191_0027)

--------------------------------------------------------------------------------
        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: 4.38 s
--------------------------------------------------------------------------------
OK
NULL    NULL
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
2020-01-28      567
2020-01-29      123
2020-01-30      112
2020-01-31      256
2020-02-01      317
2020-02-02      338
2020-02-03      191
2020-02-04      258
2020-02-05      212
2020-02-06      154
2020-02-07      91
2020-02-08      270
2020-02-09      266
2020-02-10      115
2020-02-11      3
2020-02-12      252
2020-02-13      147
2020-02-15      108
2020-02-16      133
2020-02-17      519
2020-02-18      120
2020-02-19      33
2020-02-20      105
2020-02-21      34
2020-02-22      233
2020-02-23      61
2020-02-24      292
2020-02-25      29
2020-02-26      78
2020-02-27      108
```

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

To find rating lesser than 2, we use WHERE clause. Which is grouped for each month-year as each month uniquely identifies with month and year values.

SELECT date_format(pickup_timestamp, 'yyyy-MM'), COUNT(rating_by_customer) FROM bookings_data WHERE rating_by_customer < 2 GROUP BY date_format(pickup_timestamp,'yyyy-MM') ORDER BY MIN(pickup_timestamp) ASC;



<Hive Query for Task 11>
Task 11: Calculate the count of total iOS users.

We use data from click_stream_data as there can be customers who haven't booked any cabs. We use WHERE clause on os_version to filter the 'iOS' lines, which is grouped by to find the total count.

SELECT count(distinct customer_id) FROM click_stream_data WHERE os_version in ('iOS') GROUP BY os_version;

```
hive> SELECT count(distinct customer_id) FROM click_stream_data WHERE os_version in ('iOS')
[    > GROUP BY os_version;
Query ID = hadoop_20220123195611_f4c7058b-6726-483d-ad49-9e1f3dab84a4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1642920539798_0044)

--------------------------------------------------------------------------------
        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      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.27 s
--------------------------------------------------------------------------------
OK
1515
Time taken: 5.006 seconds, Fetched: 1 row(s)
hive>
```