

## Queries

### Hive Query for Task 5

Calculate the total number of different drivers for each customer.

Query :-

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

Screenshot after executing Query

Screenshot:-

```
hive> select customer_id, count(distinct driver_id) from cab.booking group by customer_id order by customer_id ;
Query ID = hadoop_20230304181541_0d161b9b-eede-4c62-9ea8-b0fb29c2ed16
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1677949665338_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: 4.49 s
OK
10022393      1
10058402      1
10339567      1
10435129      1
10555335      1
10592274      1
10614890      1
10678894      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
```

### Hive Query for Task 6

Calculate the total rides taken by each customer.

Query:-

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

Screenshot after executing Query

Screenshot

```
hive> SELECT customer_id, COUNT(*) AS rides
> FROM cab. booking
> GROUP BY customer_id
> ORDER BY customer_id;
Query ID = hadoop_20230304183301_f1ebbc2b-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
```

## 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.**

Query:-

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;

## Screenshot after executing Query

### Screenshot

```
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)
```

## Hive Query for Task 8

Calculate the count of all trips done on black cabs.

Query:-

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

## Screenshot after executing Query

### Screenshot

```
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)
```

Activate Windows  
Go to Settings to activate Windows.

## Hive Query for Task 9

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

Query:-

**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;**

Screenshot after executing Query

Screenshot

```
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
```

## 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.

### Query

```
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;
```

### Screenshot after executing Query

#### Screenshot

```
hive> 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 ID = hadoop_20230304190440_1f7295c9-2804-4f48-921d-03eac59718a3
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_0009)

-----
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: 5.84 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: 12.046 seconds, Fetched: 10 row(s)
hive>
```

Activate Windows  
Go to Settings to activate Windows.

## Hive Query for Task 11

Calculate the count of total iOS users.

### Query

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

### Screenshot after executing Query

#### Screenshot

```
hive>
>
> SELECT COUNT(*)
> from cab.clickstream
> where os_version = 'iOS';
Query ID = hadoop_20230314103242_5796bcfd-clb5-46de-blfb3-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>
```