

# Queries

## <Hive Query for Task 5>

select customer\_id , count(distinct driver\_id) from booking\_data group by customer\_id order by customer\_id asc;

## <Screenshot after executing Query>

```
hive> select customer_id , count(distinct driver_id) from booking_data group by customer_id order by customer_id asc;
Query ID = hadoop_20220802200451_c8733194-f4f6-44a3-b583-a4558bde97bc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659451488083_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: 6.09 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
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
```

## <Hive Query for Task 6>

```
select customer_id ,count(distinct booking_id) from booking_data group by customer_id order
by customer_id asc;
```

## <Screenshot after executing Query>

```
hive> select customer_id ,count(distinct booking_id) from booking_data group by customer_id order by customer_id asc;
Query ID = hadoop_20220802201429_50b45bad-6e74-4fe2-b280-e073d5e5b3bd
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659451488083_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
Reducer 3 ..... container      SUCCEEDED      1          1          0          0          0          0
-----
VERTICES: 03/03 [=====>>] 100% ELAPSED TIME: 5.48 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
11475815      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
```

## <Hive Query for Task 7>

```
select (sum(case when button_id = "fcba68aa-1231-11eb-adc1-0242ac120002" and
is_button_click = 'Yes' then 1 end) / sum(case when page_id = "e7bc5fb2-1231-11eb-adc1-
0242ac120002" and is_page_view = 'Yes' then 1 end)) as conversion_ratio from
clickstream_data;
```

## <Screenshot after executing Query>

```
hive> select (sum(case when button_id = "fcba68aa-1231-11eb-adc1-0242ac120002" and is_button_click = 'Yes' then 1 end) / sum(case when page_id = "e7bc5fb2-1231-11eb-adc1-0242ac120002" and is_page_view = 'Yes' then 1 end)) as conversion_ratio from clickstream_data;
Query ID = hadoop_20220802214751_92c68a23-9b0c-4686-a7ec-d30adbdc2b89
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659451488083_0011)

-----
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: 5.34 s
-----
OK
0.9688109161793372
Time taken: 6.361 seconds, Fetched: 1 row(s)
hive>
```

```
hive> select (sum(case when button_id = "fcba68aa-1231-11eb-adc1-0242ac120002" and is_button_click = 'Yes' then 1 end) / sum(case when page_id = "e7bc5fb2-1231-11eb-adc1-0242ac120002" and is_page_view = 'Yes' then 1 end)) as conversion_ratio from clickstream_data;
Query ID = hadoop_20220802214751_92c68a23-9b0c-4686-a7ec-d30adbdc2b89
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659451488083_0011)

-----
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: 5.34 s
-----
OK
0.9688109161793372
Time taken: 6.361 seconds, Fetched: 1 row(s)
hive>
```

### <Hive Query for Task 8>

```
select cab_color ,count(distinct driver_id ) from booking_data where cab_color in ('black') group by cab_color ;
```

### <Screenshot after executing Query>

```
hive> select cab_color ,count(distinct driver_id ) from booking_data where cab_color in ('black') group by cab_color ;
Query ID = hadoop_20220802214946_24c05ce2-ef24-4291-b006-fd62ae636ca7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659451488083_0011)

-----
      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: 6.10 s
-----
OK
black      72
Time taken: 6.748 seconds, Fetched: 1 row(s)
hive>
```

### <Hive Query for Task 9>

```
select pickup_date, sum(tip_amount) from booking_data group by pickup_date order by pickup_date asc;
```

### <Screenshot after executing Query>

```
hive> select pickup_date, sum(tip_amount) from booking_data group by pickup_date order by pickup_date asc;
Query ID = hadoop_20220803083826_72cb3603-c8db-4e05-8dcb-b15833784a13
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659509066141_0011)

-----
      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: 10.13 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
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
```

### <Hive Query for Task 10>

```
select date_format(pickup_timestamp,'yyyy-MM') ,count( rating_by_customer) from booking_data
where rating_by_customer < 2 group by date_format(pickup_timestamp,'yyyy-MM') ;
```

### <Screenshot after executing Query>

```
hive> select date_format(pickup_timestamp,'yyyy-MM') ,count( rating_by_customer) from booking_data where rating_by_customer < 2 group by date_format(pickup_timestamp,'yyyy-MM') ;
Query ID = hadoop_20220802215651_6fa0e984-4ccf-4945-b300-c753b6e3468b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659451488083_0011)

-----
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: 5.67 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: 6.166 seconds, Fetched: 10 row(s)
hive>
```

### <Hive Query for Task 11>

```
select os_version ,count(distinct customer_id) from clickstream_data where os_version in ('iOS') group
by os_version;
```

### <Screenshot after executing Query>

```
hive> select os_version ,count(distinct customer_id) from clickstream_data where os_version in ('iOS') group by os_version;
Query ID = hadoop_20220802215957_733380e7-d47b-482e-83d1-13b0cacbd519
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659451488083_0011)

-----
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: 7.37 s
-----
OK
iOS      1515
Time taken: 7.86 seconds, Fetched: 1 row(s)
```