Task 1

select customer_id, count(driver_id) from bookings group by customer_id limit 20;

```
    > select customer_id, count(driver_id) from bookings group by customer_id
    > limit 20;
Query ID = hadoop_20220615143228_cd3f9eb5-aa3c-430c-9e51-2b748529bac7
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_1655298450638_0003)

--------------------------------------------------------------------------------
        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: 3.83 s
--------------------------------------------------------------------------------
OK
10022393        1
10555335        1
10592274        1
10678994        1
11264797        1
11418437        1
11438890        1
11518953        1
11580321        1
10058402        1
10339567        1
10435129        1
10614890        1
11353346        1
11454977        1
11479815        1
11596512        1
11608791        1
11655671        1
11757536        1
Time taken: 9.194 seconds, Fetched: 20 row(s)
```

In this task, we use bookings table and we select 2 columns : customer_id and driver_id. We group by customer_id and count the values in column driver_id

task 2

select customer_id, count(*) from bookings group by customer_id limit 20;

```
hive> select customer_id, count(*) from bookings group by customer_id limit 20;
Query ID = hadoop_20220615143808_c251879f-1877-4828-98bc-363e74105d1b
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_1655298450638_0004)

--------------------------------------------------------------------------------
        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: 3.91 s
--------------------------------------------------------------------------------
OK
10022393        1
10555335        1
10592274        1
10678994        1
11264797        1
11418437        1
11438890        1
11518953        1
11580321        1
10058402        1
10339567        1
10435129        1
10614890        1
11353346        1
11454977        1
11479815        1
11596512        1
11608791        1
11655671        1
11757536        1
Time taken: 9.268 seconds, Fetched: 20 row(s)
hive>
```

In this task, we also use bookings table and we also select customer_id column, but this time, we count values in every row then group by customer_id. So we will get each customer.

task 3

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;

```
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 = "e7bc5fb2-12
31-11eb-adc1-0242ac120002" and is_page_view = 'Yes' then 1 end)) as conversion_ratio from clickstream;
Query ID = hadoop_20220615144044_0fea4406-e698-4d9e-af21-241efcb1cf47
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655298450638_0004)

--------------------------------------------------------------------------------
        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: 4.00 s
--------------------------------------------------------------------------------
OK
0.9688109161793372
Time taken: 4.563 seconds, Fetched: 1 row(s)
hive>
```

In this task, first we use clicking_stream table, and we sum up values where button_id = "fcba68aa-1231-11eb-adc1-0242ac120002" and is_button_click = ‘Yes’ and save it. And then we sum up values where page_id = "e7bc5fb2-1231-11eb-adc1-0242ac120002" amd is_page_view = ‘Yes’ and save it again. Finally, we divide the first sum by the second sum to get conversion ratio

task 4

select count(*) from bookings where cab_color = 'black';

```
hive> select count(*) from bookings where cab_color = 'black';
Query ID = hadoop_20220615144228_31c2c28d-8110-49a4-b440-98d20444937a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655298450638_0004)

--------------------------------------------------------------------------------
        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: 4.59 s
--------------------------------------------------------------------------------
OK
72
Time taken: 5.104 seconds, Fetched: 1 row(s)
hive>
```

In this task, we use bookings table and just count every row where cab_color = ‘black’

task 5

select date(pickup_timestamp), sum(tip_amount) from bookings group by date(pickup_timestamp) limit 20;

```
hive> select date(pickup_timestamp), sum(tip_amount) from bookings group by date(pickup_timestamp) limit 20;
Query ID = hadoop_20220615144410_b1a2fae2-f919-4f29-a7ae-3040cf56d11b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655298450638_0004)

--------------------------------------------------------------------------------
        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: 4.46 s
--------------------------------------------------------------------------------
OK
2020-01-03      11
2020-01-05      134
2020-01-09      48
2020-01-11      81
2020-01-16      155
2020-01-17      296
2020-01-18      240
2020-01-20      210
2020-01-01      59
2020-01-02      95
2020-01-04      123
2020-01-06      189
2020-01-07      148
2020-01-08      111
2020-01-10      77
2020-01-12      109
2020-01-14      142
2020-01-15      338
2020-01-21      5
2020-01-23      148
Time taken: 4.909 seconds, Fetched: 20 row(s)
hive>
```

In this task, we use bookings table again. Here, we select 2 columns pickup_timestamp and tip_amount. We sum tip_amount and group by pickup_timestamp to get total tip_amount for each particular date.

task 6

select date_format(pickup_timestamp, 'YYYY-MM') as monthwise, count(*) as total_bookings from bookings where rating_by_customer < 2 group by date_format(pickup_timestamp, 'YYYY-MM') order by monthwise;

```
hive> select date_format(pickup_timestamp, 'YYYY-MM') as monthwise, count(*) as total_bookings from bookings where rating_by_customer < 2 group
t(pickup_timestamp, 'YYYY-MM') order by monthwise;
Query ID = hadoop_20220615144549_0ef08bbb-7860-441c-a2dc-a19bba54b001
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655298450638_0004)

--------------------------------------------------------------------------------
        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: 4.43 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: 4.929 seconds, Fetched: 10 row(s)
hive>
```

In this task, we also use bookings table. We use function date_format with pickup_timestamp column to retrieve the desired format ‘YYYYMM’ and name it as monthwise. Then we count each row of which rating_by_customer is less than 2. Then we group by the date_format we created earlier and finally sort the query in ascending order.

task 7

select count(*) from clickstream where os_version = 'iOS';

```
hive> select count(*) from clickstream where os_version = 'iOS';
Query ID = hadoop_20220615144853_1cbd6fe2-ee98-4044-84b6-8a741efa33ad
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1655298450638_0004)

--------------------------------------------------------------------------------
        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
1515
Time taken: 4.853 seconds, Fetched: 1 row(s)
hive>
```

In this task, we use clicking_stream table and we count every row where os_version = ‘iOS’ to get the number of users who use iOS.