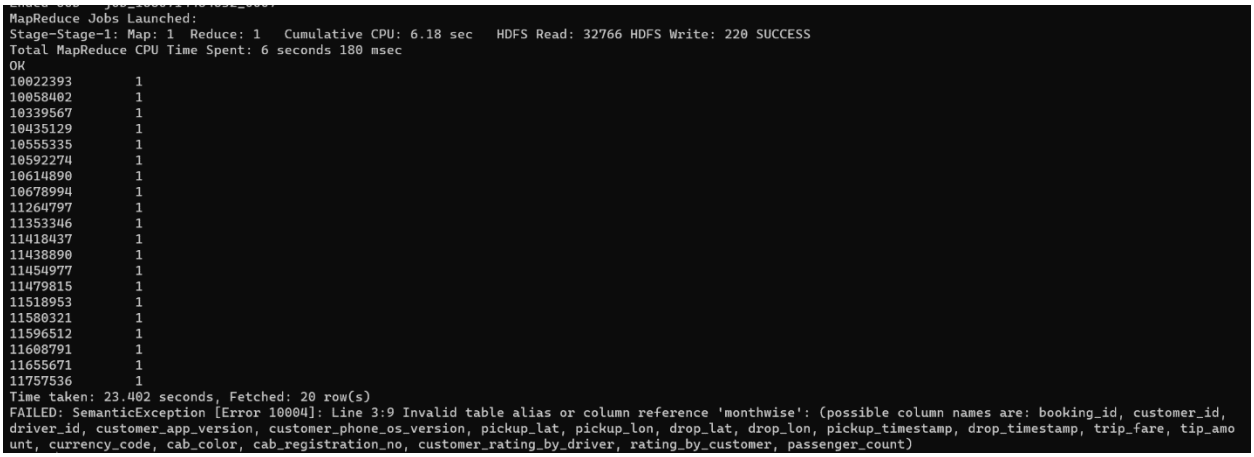
Logic For Final Submission

```
# task 1

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

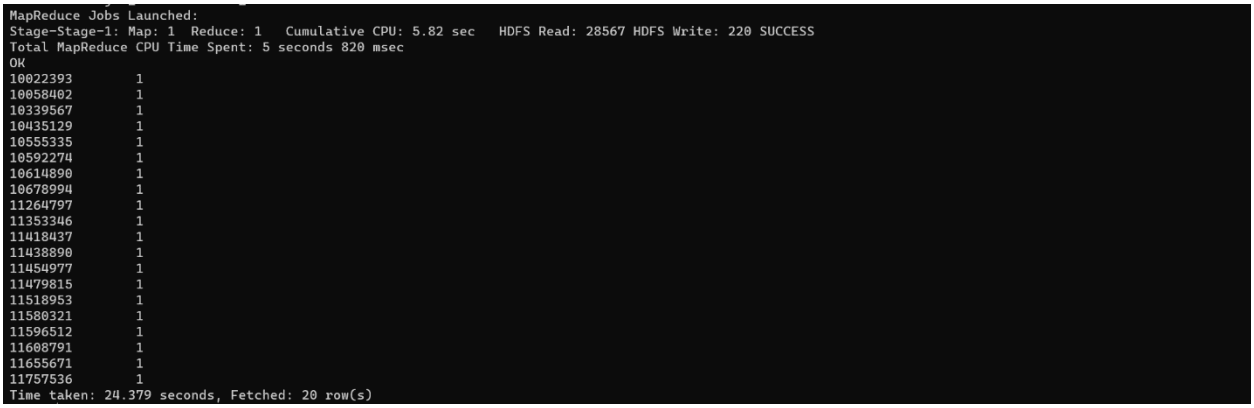


```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.18 sec   HDFS Read: 32766 HDFS Write: 220 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 180 msec
OK
10022393        1
10058402        1
10339567        1
10435129        1
10555335        1
10592274        1
10614890        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
Time taken: 23.402 seconds, Fetched: 20 row(s)
FAILED: SemanticException [Error 10004]: Line 3:9 Invalid table alias or column reference 'monthwise': (possible column names are: booking_id, customer_id,
driver_id, customer_app_version, customer_phone_os_version, pickup_lat, pickup_lon, drop_lat, drop_lon, pickup_timestamp, drop_timestamp, trip_fare, tip_amo
unt, currency_code, cab_color, cab_registration_no, customer_rating_by_driver, rating_by_customer, passenger_count)
```

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;
```



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 clicking_stream;
```

```
MapReduce Total cumulative CPU time: 8 seconds 320 msec
Ended Job = job_1650714454632_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.32 sec   HDFS Read: 26480 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 320 msec
OK
0.9688109161793372
Time taken: 22.487 seconds, Fetched: 1 row(s)
```

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';
```

```
MapReduce Total cumulative CPU time: 7 seconds 80 msec
Ended Job = job_1650714454632_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.08 sec   HDFS Read: 25875 HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 80 msec
OK
72
Time taken: 23.817 seconds, Fetched: 1 row(s)
```

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;
```

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.25 sec   HDFS Read: 37331 HDFS Write: 292 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 250 msec
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
Time taken: 23.606 seconds, Fetched: 20 row(s)
```

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;
```

```
MapReduce Total cumulative CPU time: 4 seconds 310 msec
Ended Job = job_1650714454632_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.73 sec   HDFS Read: 34860 HDFS Write: 356 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.31 sec   HDFS Read: 5375 HDFS Write: 110 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 40 msec
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: 48.28 seconds, Fetched: 10 row(s)
```

In this task, we also use bookings table. We use function date_format with pickup_timestamp column to retrieve the desired format 'YYYY-MM' 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 clicking_stream where os_version = 'iOS';
```

```
MapReduce Total cumulative CPU time: 6 seconds 830 msec
Ended Job = job_1650714454632_0017
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.83 sec   HDFS Read: 20831 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 830 msec
OK
1515
Time taken: 22.442 seconds, Fetched: 1 row(s)
```

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.