Logic For Final Submission

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

<Hive Query for Task 5>

```
select customer_id ,count( DISTINCT driver_id) from booking_data group by customer_id order by customer_id asc;
```

The task was to get the total drivers assigned to each customer, the query will be executed over the booking_data table, grouping them by the customer_id which is unique for customer and order by the customer id. The query will 2 column information one is the customer_id and other is count of DISTINCT driver_id who was allocated in the booking. Note: the ordering is done in ascending order

<Execution Screen >

```
hive> select customer_id ,count( DISTINCT driver_id) from booking_data group by customer_id order by customer_id asc;
Query ID = ec2-user_20220301073131_f746409f-212f-45d0-9564-c744c64994a9
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1646119500330_0001, Tracking URL = http://ip-10-0-0-212.ec2.internal:8088/proxy/application_1646119500330_0001/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1646119500330_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-01 07:31:55,877 Stage-1 map = 0%,  reduce = 0%
2022-03-01 07:32:02,310 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.82 sec
2022-03-01 07:32:08,602 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.53 sec
MapReduce Total cumulative CPU time: 5 seconds 530 msec
Ended Job = job_1646119500330_0001
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1646119500330_0002, Tracking URL = http://ip-10-0-0-212.ec2.internal:8088/proxy/application_1646119500330_0002/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1646119500330_0002
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-03-01 07:32:23,381 Stage-2 map = 0%,  reduce = 0%
2022-03-01 07:32:32,166 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.91 sec
2022-03-01 07:32:41,601 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.77 sec
MapReduce Total cumulative CPU time: 4 seconds 770 msec
Ended Job = job_1646119500330_0002
```

<Screenshot after executing Query>

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.92 sec   HDFS Read: 193168 HDFS Write: 23250 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.61 sec   HDFS Read: 28232 HDFS Write: 11005 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 530 msec
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
13262795        1
13356177        1
13387493        1
13389366        1
13442644        1
13500355        1
```

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

<Hive Query for Task 6>

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

The task was to get the total riders taken to each customer, the query will be executed over the booking_data table, grouping them by the customer_id which is unique for customer and order by the customer id. The query will 2 column information one is the customer_id and other is count of DISTINCT booking_id of the booking. Note: the ordering is done in ascending order

<Execution Screen >

```
hive> select customer_id ,count( DISTINCT booking_id) from booking_data group by customer_id order by customer_id asc;
Query ID = ec2-user_20220301073434_61777524-43f0-451a-8e13-a13ba215169b
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1646119500330_0003, Tracking URL = http://ip-10-0-0-212.ec2.internal:8088/proxy/application_1646119500330_0003/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1646119500330_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-01 07:34:17,449 Stage-1 map = 0%,  reduce = 0%
2022-03-01 07:34:22,671 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.67 sec
2022-03-01 07:34:28,939 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.55 sec
MapReduce Total cumulative CPU time: 5 seconds 550 msec
Ended Job = job_1646119500330_0003
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1646119500330_0004, Tracking URL = http://ip-10-0-0-212.ec2.internal:8088/proxy/application_1646119500330_0004/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1646119500330_0004
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-03-01 07:34:44,816 Stage-2 map = 0%,  reduce = 0%
2022-03-01 07:34:50,085 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.67 sec
2022-03-01 07:34:56,344 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.28 sec
MapReduce Total cumulative CPU time: 5 seconds 280 msec
Ended Job = job_1646119500330_0004
```

<Screenshot after executing Query>

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.93 sec   HDFS Read: 193232 HDFS Write: 23250 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.61 sec   HDFS Read: 28234 HDFS Write: 11005 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 540 msec
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
```

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.

<Hive Query for Task 7>

```
select count(b.button_id)/count(a.booking_id) from booking_data a full outer join
clickstream_data b on a.customer_id = b.customer_id;
```

The task was to get the hit ratio/ booking ratio from the number of visit done by the customer, the query will be join of both booking_data and clickstream_data for getting the number of hit/successful conversion. The query is about getting the outer join of both clickstream_data and booking_data and match with the customer_id which is unique in both and there by the query will return division of total hit in booking_data to the total visit data in clickstream_data

<Execution Screen >

```
hive> select count(b.button_id)/count(a.booking_id) from booking_data a full outer join clickstream_data b on a.customer_id = b.customer_id;
Query ID = ec2-user_20220301073737_3f91e9cd-7811-42fc-a044-917b053c1c35
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1646119500330_0005, Tracking URL = http://ip-10-0-0-212.ec2.internal:8088/proxy/application_1646119500330_0005/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1646119500330_0005
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2022-03-01 07:37:57,072 Stage-1 map = 0%,  reduce = 0%
2022-03-01 07:38:05,557 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.93 sec
2022-03-01 07:38:11,766 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.73 sec
MapReduce Total cumulative CPU time: 8 seconds 730 msec
Ended Job = job_1646119500330_0005
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1646119500330_0006, Tracking URL = http://ip-10-0-0-212.ec2.internal:8088/proxy/application_1646119500330_0006/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1646119500330_0006
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-03-01 07:38:24,963 Stage-2 map = 0%,  reduce = 0%
2022-03-01 07:38:32,308 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.92 sec
2022-03-01 07:38:39,504 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.99 sec
MapReduce Total cumulative CPU time: 4 seconds 990 msec
Ended Job = job_1646119500330_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 8.73 sec   HDFS Read: 599352 HDFS Write: 119 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.99 sec   HDFS Read: 5843 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 720 msec
```

<Screenshot after executing Query>

```
MapReduce Total cumulative CPU time: 4 seconds 990 msec
Ended Job = job_1646119500330_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 8.73 sec   HDFS Read: 599352 HDFS Write: 119 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.99 sec   HDFS Read: 5843 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 720 msec
OK
2.998001998001998
Time taken: 60.537 seconds, Fetched: 1 row(s)
```

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

<Hive Query for Task 8>

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

The task was to count the trips done on the black cabs, for this task the query will be over the booking_data, where the cab_color is ‘black’ and the groupby cab_color for the query out. The query will finally give the cab_color and count of the distinct driver_id in the groupby option

<Execution Screen >

```
hive> select cab_color ,count(distinct driver_id ) from booking_data where cab_color in ('black') group by cab_color ;
Query ID = ec2-user_20220301074242_26b21f57-545a-4654-bb67-2edd7b79822a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1646119500330_0007, Tracking URL = http://ip-10-0-0-212.ec2.internal:8088/proxy/application_1646119500330_0007/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1646119500330_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-01 07:42:43,373 Stage-1 map = 0%,  reduce = 0%
2022-03-01 07:42:49,565 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.25 sec
2022-03-01 07:42:56,891 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.03 sec
MapReduce Total cumulative CPU time: 6 seconds 30 msec
Ended Job = job_1646119500330_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.03 sec   HDFS Read: 193878 HDFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 30 msec
```

<Screenshot after executing Query>

```
MapReduce Total cumulative CPU time: 6 seconds 30 msec
Ended Job = job_1646119500330_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.03 sec   HDFS Read: 193878 HDFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 30 msec
OK
black   72
Time taken: 28.219 seconds, Fetched: 1 row(s)
```

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

<Hive Query for Task 9>

```
select date_format(pickup_timestamp,'yyyy-MM-dd'),sum(tip_amount) from booking_data group by date_format(pickup_timestamp,'yyyy-MM-dd');
```

The task was to total tips in the trips date-wise, for this task the query will be over the booking_data, grouped by the date_format column contents, the final table will provide the data in the yyyy-mm-dd format and the sum of the tip_amount given in that day.

<Execution Screen >

```
hive> select date_format(pickup_timestamp,'yyyy-MM-dd'),sum(tip_amount) from booking_data group by date_format(pickup_timestamp,'yyyy-MM-dd');
Query ID = ec2-user_20220301074646_e3717bd1-cff0-4631-9677-e8c8e671dd01
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1646119500330_0008, Tracking URL = http://ip-10-0-0-212.ec2.internal:8088/proxy/application_1646119500330_0008/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1646119500330_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-01 07:47:01,540 Stage-1 map = 0%,  reduce = 0%
2022-03-01 07:47:07,731 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.67 sec
2022-03-01 07:47:13,966 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.25 sec
MapReduce Total cumulative CPU time: 5 seconds 250 msec
Ended Job = job_1646119500330_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.25 sec   HDFS Read: 194456 HDFS Write: 9 SUCCESS
```

<Screenshot after executing Query>

```
Total MapReduce CPU Time Spent: 7 seconds 960 msec
OK
2020-01-01      295
2020-01-02      475
2020-01-03      55
2020-01-04      615
2020-01-05      670
2020-01-06      945
2020-01-07      740
2020-01-08      555
2020-01-09      240
2020-01-10      385
2020-01-11      405
2020-01-12      545
2020-01-14      710
2020-01-15      1690
2020-01-16      775
```

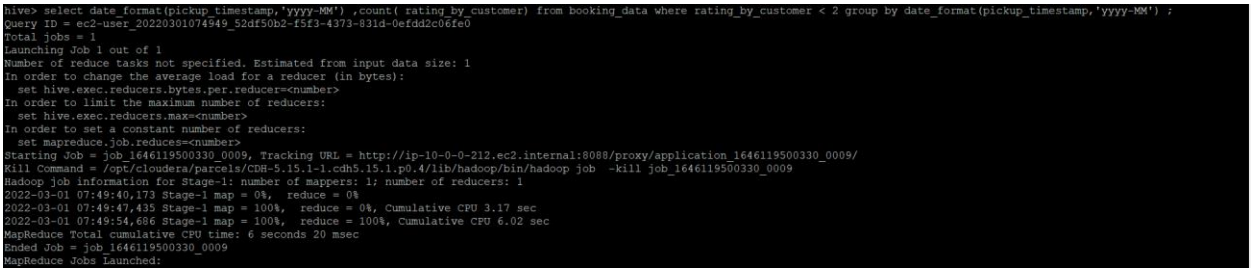
Task 10: Calculate the total count of all the bookings with ratings lower than 2 as given by customers in a particular month.

<Hive Query for Task 10>

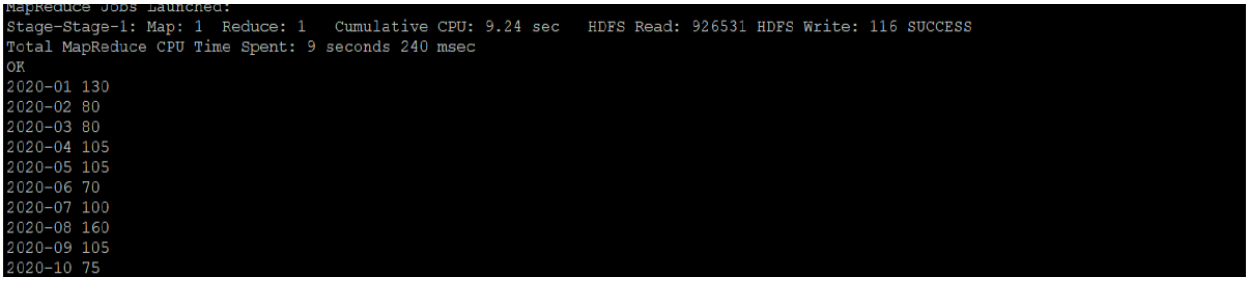
```
select date_format(pickup_timestamp,'yyyy-MM') ,count( rating_by_customer) from booking_data where rating_by_customer < 2 group by date_format(pickup_timestamp,'yyyy-MM') ;
```

The task was to get the customer rating for <2 in a particular month. For this task the query will be over the booking_data, grouped by the date_format(month) column contents where the rating_by_customer < 2 (i.e 1 or 0) , the final table will provide the month in the yyyy-MM format and the count of the 'rating_by_customer' for that month

<Execution Screen >



<Screenshot after executing Query>



Task 11: Calculate the count of total iOS users.

<Hive Query for Task 11>

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

The task is to get the count of iOS user who are using the app for booking the cab, for the query we use the clickstream_data and group the data by os_version where the os_version is iOS. From the queried data, the output will be os_version and count of distinct customer_id user of the platform

<Execution Screen >

```
hive> select os_version ,count(distinct customer_id) from clickstream_data where os_version in ('iOS') group by os_version;
Query ID = ec2-user_20220301075353_d1fbe768-1e42-4138-890e-346f347a2dc0
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1646119500330_0010, Tracking URL = http://ip-10-0-0-212.ec2.internal:8088/proxy/application_1646119500330_0010/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1646119500330_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-01 07:53:19,394 Stage-1 map = 0%,  reduce = 0%
2022-03-01 07:53:27,691 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.02 sec
2022-03-01 07:53:33,898 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.03 sec
MapReduce Total cumulative CPU time: 7 seconds 30 msec
Ended Job = job_1646119500330_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.03 sec   HDFS Read: 407324 HDFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 30 msec
```

<Screenshot after executing Query>

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.03 sec   HDFS Read: 407324 HDFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 30 msec
OK
iOS     1515
Time taken: 29.12 seconds, Fetched: 1 row(s)
```

Conclusion

Hive query for tasks 5-11 are completed and recorded for the final submission of the cab booking project.