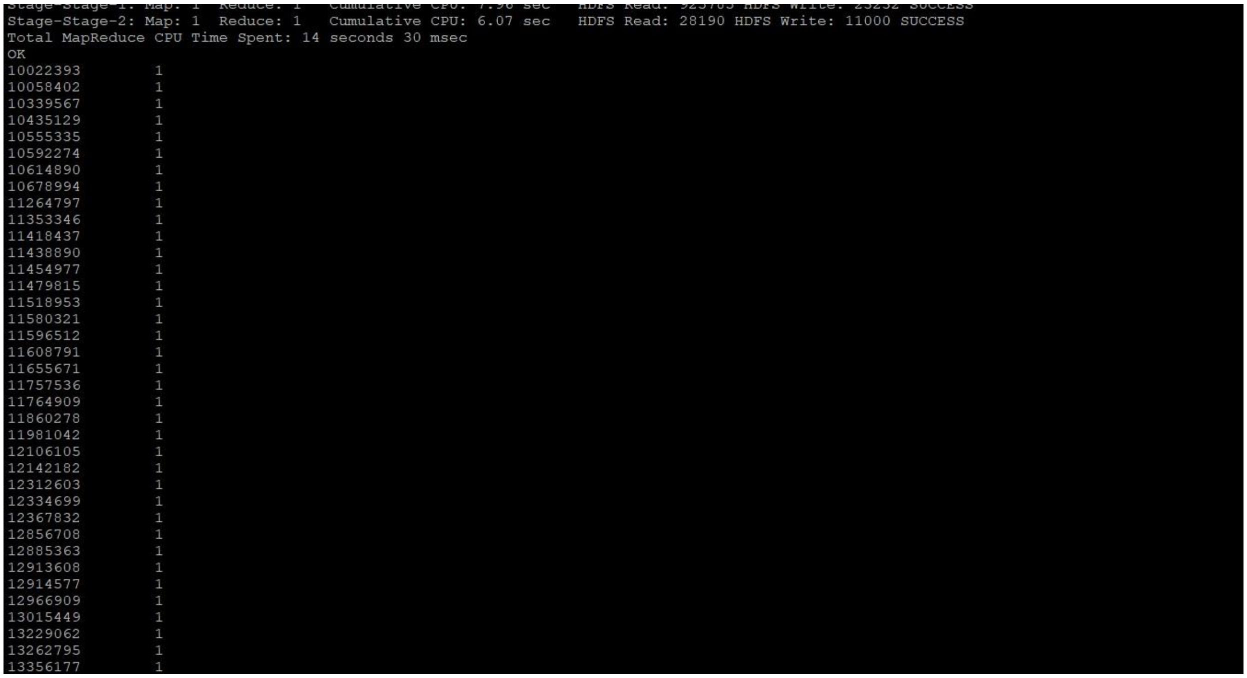
Queries

<Hive Query for Task 5>

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

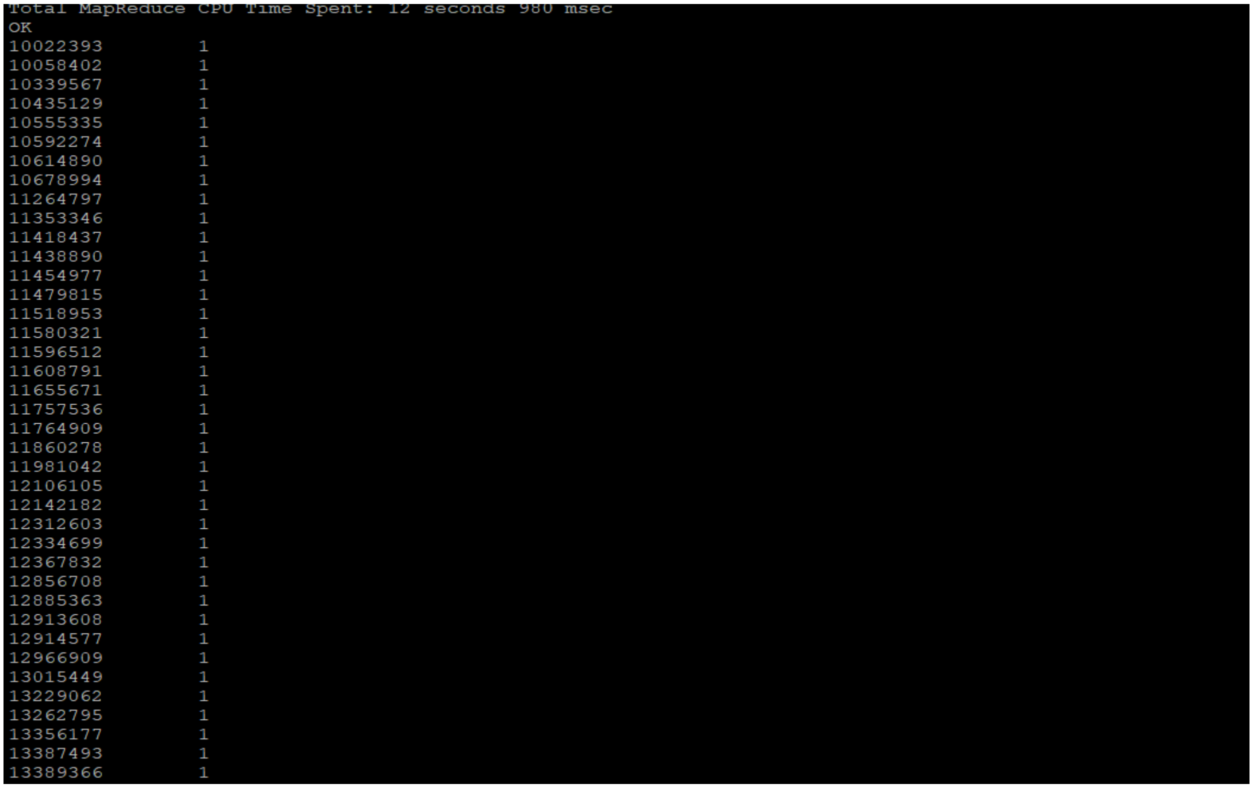
<Screenshot after executing Query>



```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.96 sec   HDFS Read: 925705 HDFS Write: 23232 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 6.07 sec   HDFS Read: 28190 HDFS Write: 11000 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 30 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
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
```

<Hive Query for Task 6>

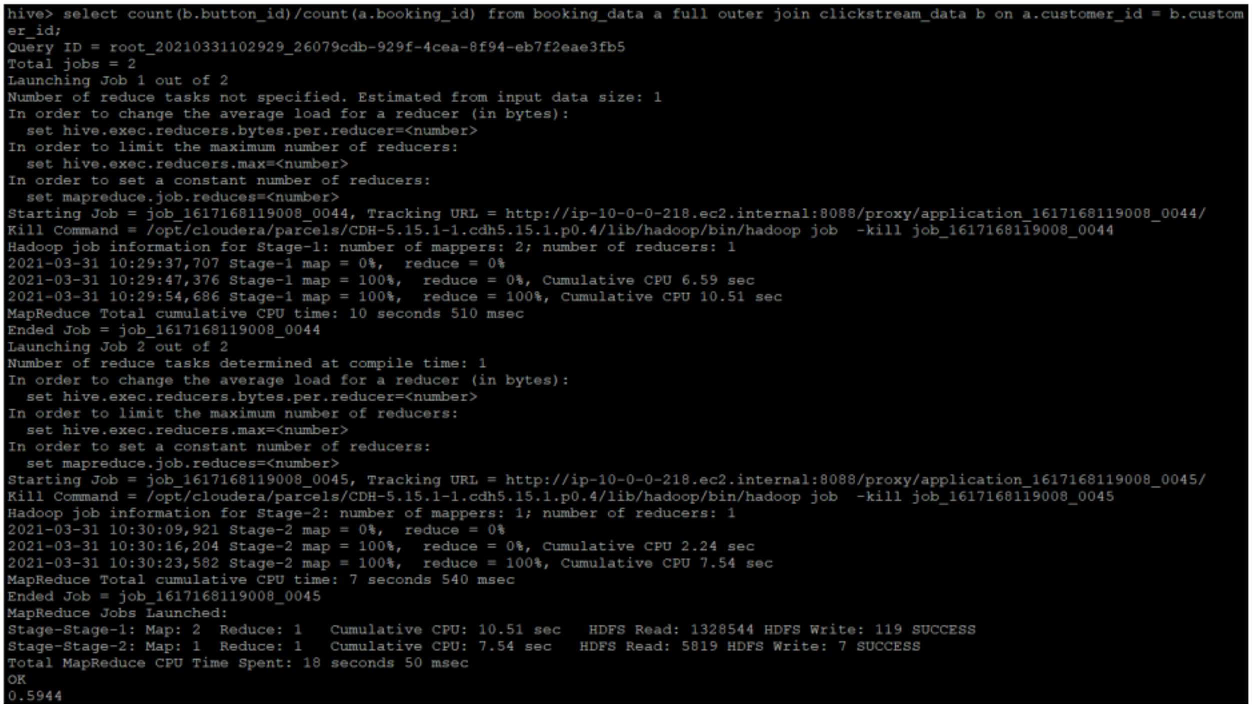
select customer_id ,count(DISTINCT booking_id) from booking_data group by customer_id order by customer_id asc;



```
Total MapReduce CPU Time Spent: 12 seconds 980 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
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>

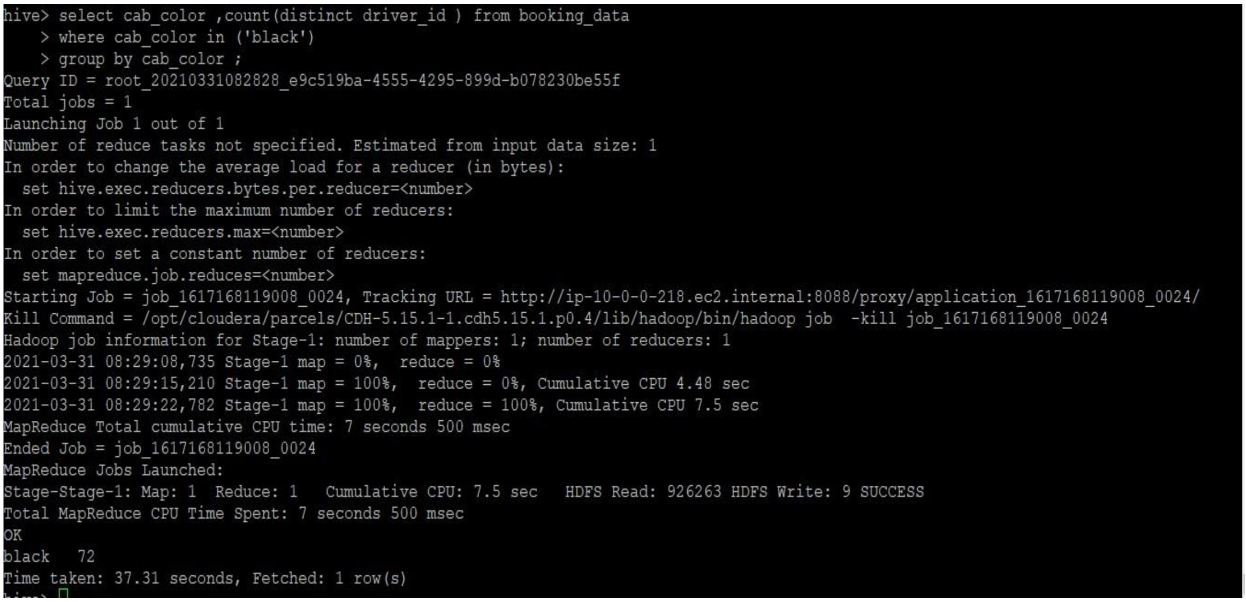
select count(b.button_id)/count(a.booking_id) from booking_data a full outer join clickstream_data b on a.customer_id = b.customer_id;



```
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.custom
er_id;
Query ID = root_20210331102929_26079cdb-929f-4cea-8f94-eb7f2eae3fb5
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_1617168119008_0044, Tracking URL = http://ip-10-0-0-218.ec2.internal:8088/proxy/application_1617168119008_0044/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1617168119008_0044
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2021-03-31 10:29:37,707 Stage-1 map = 0%,  reduce = 0%
2021-03-31 10:29:47,376 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.59 sec
2021-03-31 10:29:54,686 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.51 sec
MapReduce Total cumulative CPU time: 10 seconds 510 msec
Ended Job = job_1617168119008_0044
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_1617168119008_0045, Tracking URL = http://ip-10-0-0-218.ec2.internal:8088/proxy/application_1617168119008_0045/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1617168119008_0045
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-03-31 10:30:09,921 Stage-2 map = 0%,  reduce = 0%
2021-03-31 10:30:16,204 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.24 sec
2021-03-31 10:30:23,582 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 7.54 sec
MapReduce Total cumulative CPU time: 7 seconds 540 msec
Ended Job = job_1617168119008_0045
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 10.51 sec   HDFS Read: 1328544 HDFS Write: 119 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 7.54 sec   HDFS Read: 5819 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 50 msec
OK
0.5944
```

<Hive Query for Task 8>

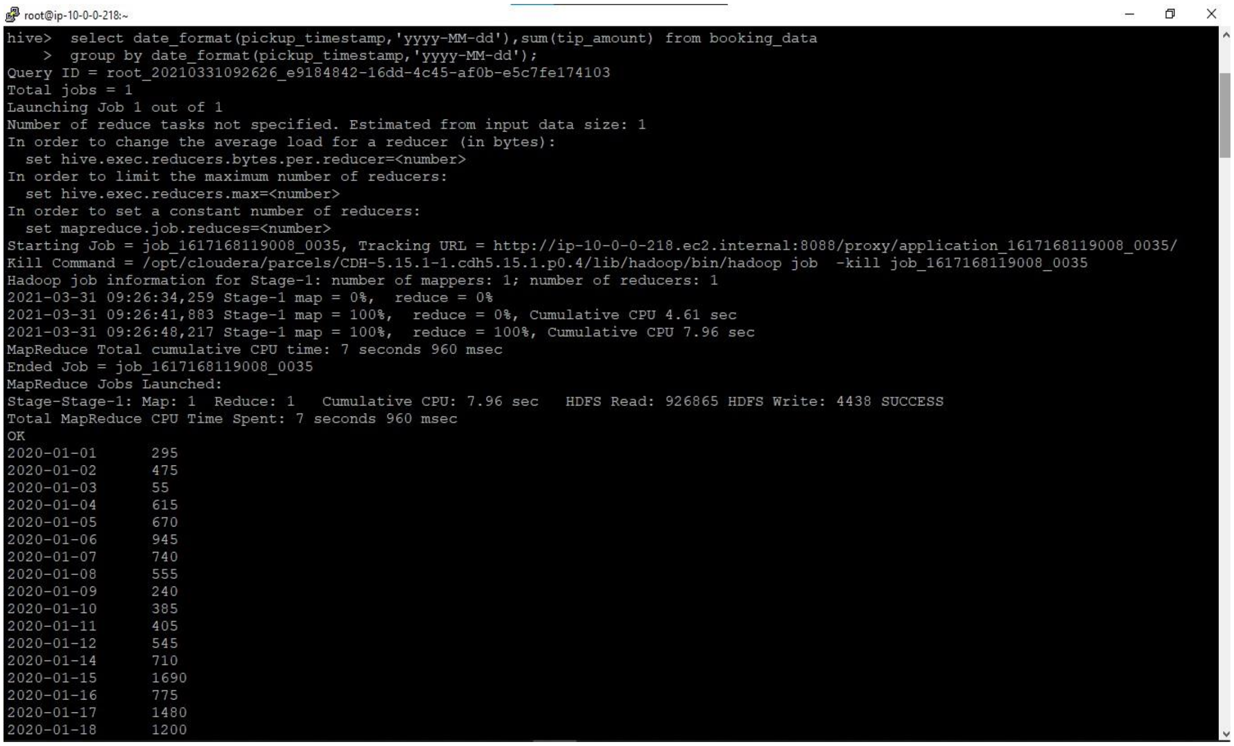
select cab_color ,count(distinct driver_id) from booking_data
where cab_color in ('black')
group by cab_color ;



```
hive> select cab_color ,count(distinct driver_id ) from booking_data
    > where cab_color in ('black')
    > group by cab_color ;
Query ID = root_20210331082828_e9c519ba-4555-4295-899d-b078230be55f
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_1617168119008_0024, Tracking URL = http://ip-10-0-0-218.ec2.internal:8088/proxy/application_1617168119008_0024/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1617168119008_0024
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-31 08:29:08,735 Stage-1 map = 0%,  reduce = 0%
2021-03-31 08:29:15,210 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.48 sec
2021-03-31 08:29:22,782 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.5 sec
MapReduce Total cumulative CPU time: 7 seconds 500 msec
Ended Job = job_1617168119008_0024
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.5 sec   HDFS Read: 926263 HDFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 500 msec
OK
black   72
Time taken: 37.31 seconds, Fetched: 1 row(s)
hive>
```

<Hive Query for Task 9>

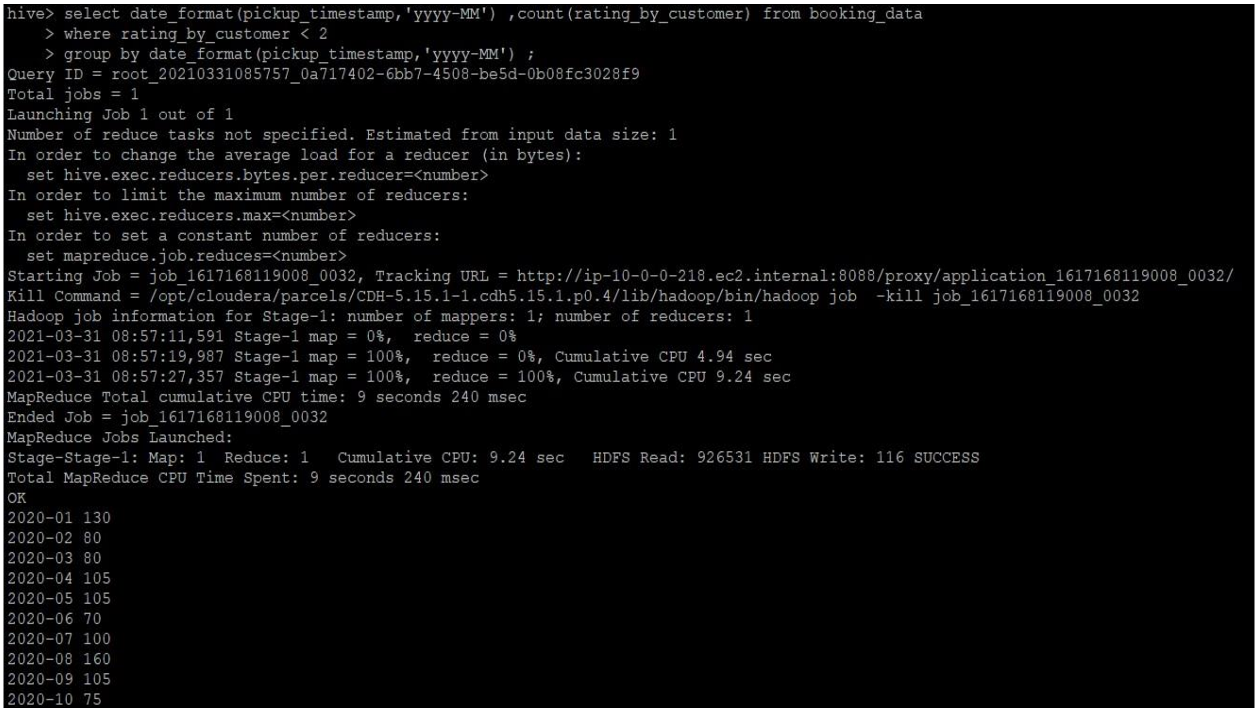
select date_format(pickup_timestamp,'yyyy-MM-dd'),sum(tip_amount) from booking_data
group by date_format(pickup_timestamp,'yyyy-MM-dd');



```
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 = root_20210331092626_e9184842-16dd-4c45-af0b-e5c7fe174103
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_1617168119008_0035, Tracking URL = http://ip-10-0-0-218.ec2.internal:8088/proxy/application_1617168119008_0035/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1617168119008_0035
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-31 09:26:34,259 Stage-1 map = 0%,  reduce = 0%
2021-03-31 09:26:41,883 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.61 sec
2021-03-31 09:26:48,217 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.96 sec
MapReduce Total cumulative CPU time: 7 seconds 960 msec
Ended Job = job_1617168119008_0035
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.96 sec   HDFS Read: 926865 HDFS Write: 4438 SUCCESS
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
2020-01-17      1480
2020-01-18      1200
```

<Hive Query for Task 10>

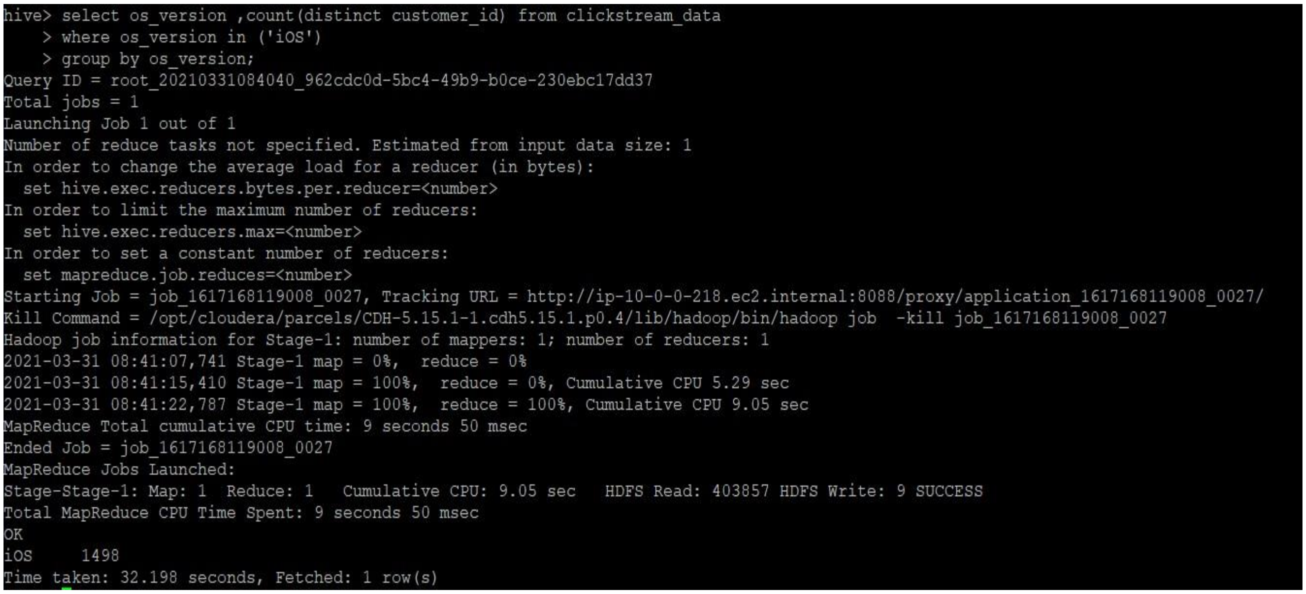
select date_format(pickup_timestamp,'yyyy-MM') ,count(rating_by_customer) from booking_data
where rating_by_customer < 2
group by date_format(pickup_timestamp,'yyyy-MM') ;



```
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 = root_20210331085757_0a717402-6bb7-4508-be5d-0b08fc3028f9
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_1617168119008_0032, Tracking URL = http://ip-10-0-0-218.ec2.internal:8088/proxy/application_1617168119008_0032/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1617168119008_0032
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-31 08:57:11,591 Stage-1 map = 0%,  reduce = 0%
2021-03-31 08:57:19,987 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.94 sec
2021-03-31 08:57:27,357 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.24 sec
MapReduce Total cumulative CPU time: 9 seconds 240 msec
Ended Job = job_1617168119008_0032
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.24 sec   HDFS Read: 926531 HDFS Write: 116 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 240 msec
OK
2020-01 130
2020-02 80
2020-03 80
2020-04 105
2020-05 105
2020-06 70
2020-07 100
2020-08 160
2020-09 105
2020-10 75
```

<Hive Query for Task 11>

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



```
hive> select os_version ,count(distinct customer_id) from clickstream_data
    > where os_version in ('iOS')
    > group by os_version;
Query ID = root_20210331084040_962cdc0d-5bc4-49b9-b0ce-230ebc17dd37
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_1617168119008_0027, Tracking URL = http://ip-10-0-0-218.ec2.internal:8088/proxy/application_1617168119008_0027/
Kill Command = /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/lib/hadoop/bin/hadoop job  -kill job_1617168119008_0027
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-31 08:41:07,741 Stage-1 map = 0%,  reduce = 0%
2021-03-31 08:41:15,410 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.29 sec
2021-03-31 08:41:22,787 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.05 sec
MapReduce Total cumulative CPU time: 9 seconds 50 msec
Ended Job = job_1617168119008_0027
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.05 sec   HDFS Read: 403857 HDFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 50 msec
OK
iOS     1498
Time taken: 32.198 seconds, Fetched: 1 row(s)
```