

Project 1 (Big-Data)

Querying using Hive on Yellow Taxi data

<https://www.kaggle.com/datasets/elemento/nyc-yellow-taxi-trip-data>

Problem statement:

In this case study, we are giving a real-world example of how to use HIVE on top of the HADOOP for different exploratory data analysis. In here, we have a predefined dataset (2018_Yellow_Taxi_Trip_Data.csv) having more than 15 columns and more than 100000 records in it. The dataset has different attributes like

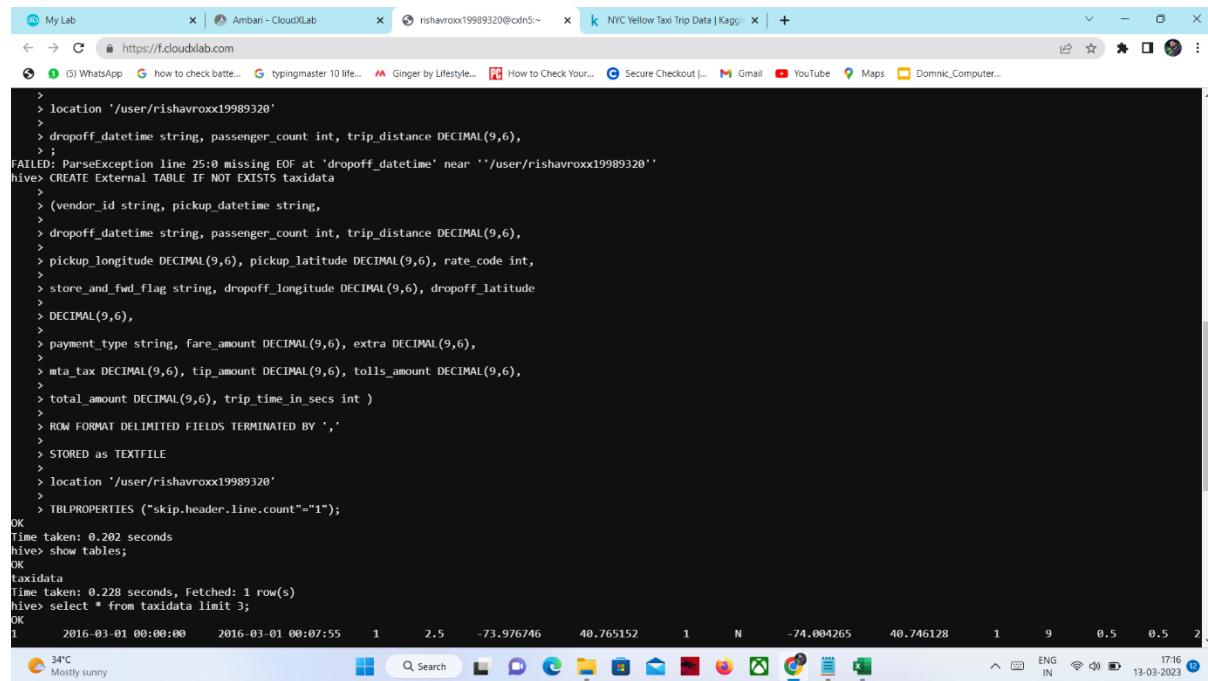
1. vendor_id string,
2. pickup_datetime string,
3. dropoff_datetime string,
4. passenger_count int,
5. trip_distance DECIMAL(9,6),
6. pickup_longitude DECIMAL(9,6),
7. pickup_latitude DECIMAL(9,6),
8. rate_code int,
9. store_and_fwd_flag string,
10. dropoff_longitude DECIMAL(9,6),
11. dropoff_latitude DECIMAL(9,6),
12. payment_type string,
13. fare_amount DECIMAL(9,6),
14. extra DECIMAL(9,6),
15. mta_tax DECIMAL(9,6),
16. tip_amount DECIMAL(9,6),
17. tolls_amount DECIMAL(9,6),
18. total_amount DECIMAL(9,6),
19. trip_time_in_secs int

Perform taxi trip analysis by solving the questions below:

1. What is the total Number of trips (equal to the number of rows)?
2. What is the total revenue generated by all the trips? The fare is stored in the column total_amount.
3. What fraction of the total is paid for tolls? The toll is stored in tolls_amount.
4. What fraction of it is driver tips? The tip is stored in tip_amount.
5. What is the average trip amount?
6. What is the average distance of the trips? Distance is stored in the column trip_distance.

7. How many different payment types are used?
8. For each payment type, display the following details:
 - Average fare generated
 - Average tip
 - Average tax – tax is stored in column mta_tax
9. On average which hour of the day generates the highest revenue?

Q1) Creating table:-



The screenshot shows a Windows desktop with a terminal window open. The terminal window displays the following command and its execution:

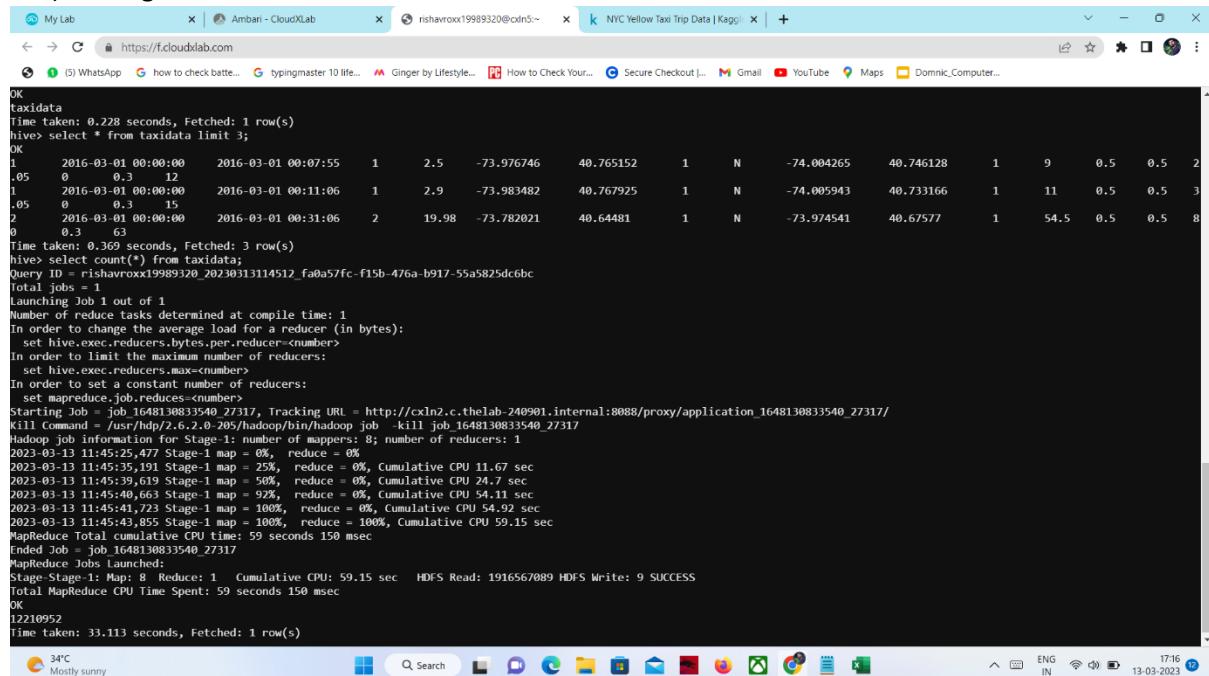
```

My Lab x | Ambari - CloudXLab x | rishavroxx19989320@oxn5:~ x | k NYC Yellow Taxi Trip Data | Kaggle x | +
< → C https://f.cloudxlab.com
(5) WhatsApp G how to check batte... G typingmaster 10 life... M Ginger by Lifestyle... P How to Check Your... Secure Checkout ... Gmail YouTube Maps Dominic_Computer...
> location '/user/rishavroxx19989320'
> dropoff_datetime string, passenger_count int, trip_distance DECIMAL(9,6),
> ;
FAILED: ParseException line 25:0 missing EOF at 'dropoff_datetime' near ''/user/rishavroxx19989320''
hive> CREATE External TABLE IF NOT EXISTS taxidata
> (vendor_id string, pickup_datetime string,
> dropoff_datetime string, passenger_count int, trip_distance DECIMAL(9,6),
> pickup_longitude DECIMAL(9,6), pickup_latitude DECIMAL(9,6), rate_code int,
> store_and_fwd_flag string, dropoff_longitude DECIMAL(9,6), dropoff_latitude
> DECIMAL(9,6),
> payment_type string, fare_amount DECIMAL(9,6), extra DECIMAL(9,6),
> mta_tax DECIMAL(9,6), tip_amount DECIMAL(9,6), tolls_amount DECIMAL(9,6),
> total_amount DECIMAL(9,6), trip_time_in_secs int )
> ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
> STORED as TEXTFILE
> location '/user/rishavroxx19989320'
> TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.202 seconds
hive> show tables;
OK
taxidata
Time taken: 0.228 seconds, Fetched: 1 row(s)
hive> select * from taxidata limit 3;
OK
1 2016-03-01 00:00:00 2016-03-01 00:07:55 1 2.5 -73.976746 40.765152 1 N -74.004265 40.746128 1 9 0.5 0.5 2

```

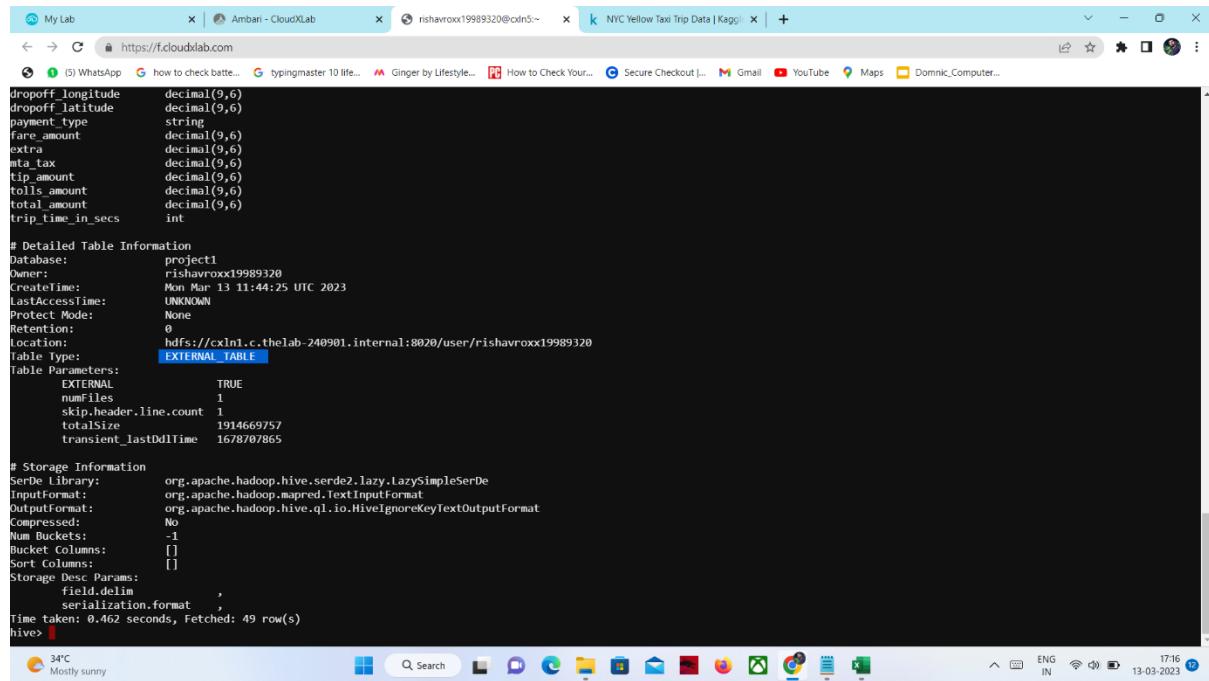
The terminal window is part of a larger desktop environment. At the bottom, there's a taskbar with various icons including a weather widget showing "34°C Mostly sunny". The system tray shows the date and time as "13-03-2023 17:16".

Q.2) Finding no of Records:



```
OK
taxidata
Time taken: 0.228 seconds, Fetched: 1 row(s)
hive> select * from taxidata limit 3;
OK
1 2016-03-01 00:00:00 2016-03-01 00:07:55 1 2.5 -73.976746 40.765152 1 N -74.004265 40.746128 1 9 0.5 0.5 2
.05 0 .3 12
1 2016-03-01 00:00:00 2016-03-01 00:11:06 1 2.9 -73.983482 40.767925 1 N -74.005943 40.733166 1 11 0.5 0.5 3
.05 0 .3 15
2 2016-03-01 00:00:00 2016-03-01 00:31:06 2 19.98 -73.782021 40.64481 1 N -73.974541 40.67577 1 54.5 0.5 0.5 8
0 .3 63
Time taken: 0.369 seconds, Fetched: 3 row(s)
hive> select count(*) from taxidata;
Query ID = rishavrox19989320_20230313114512_fa0a57fc-f15b-476a-b917-55a5825dc6bc
Total jobs = 1
Launching Job 1 out of 1
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=<n>number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<n>number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<n>number>
Starting Job = job_1648130833540_27317, Tracking URL = http://cxln1.c.thelab-240901.internal:8088/proxy/application_1648130833540_27317/
Kill Command = /usr/hdp/2.6.2.0-205/hadoop/bin/hadoop job -kill job_1648130833540_27317
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 11:49:25,477 Stage-1 map = 0%, reduce = 0%
2023-03-13 11:49:35,191 Stage-1 map = 25%, reduce = 0%, Cumulative CPU 11.67 sec
2023-03-13 11:49:39,619 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 24.7 sec
2023-03-13 11:49:40,663 Stage-1 map = 92%, reduce = 0%, Cumulative CPU 54.11 sec
2023-03-13 11:49:41,723 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 54.92 sec
2023-03-13 11:49:43,855 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 59.15 sec
MapReduce Total cumulative CPU time: 59 seconds 150 msec
Ended Job = job_1648130833540_27317
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 59.15 sec HDFS Read: 1916567089 HDFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 59 seconds 150 msec
OK
12210952
Time taken: 33.113 seconds, Fetched: 1 row(s)
```

Q3. Info about Table:-

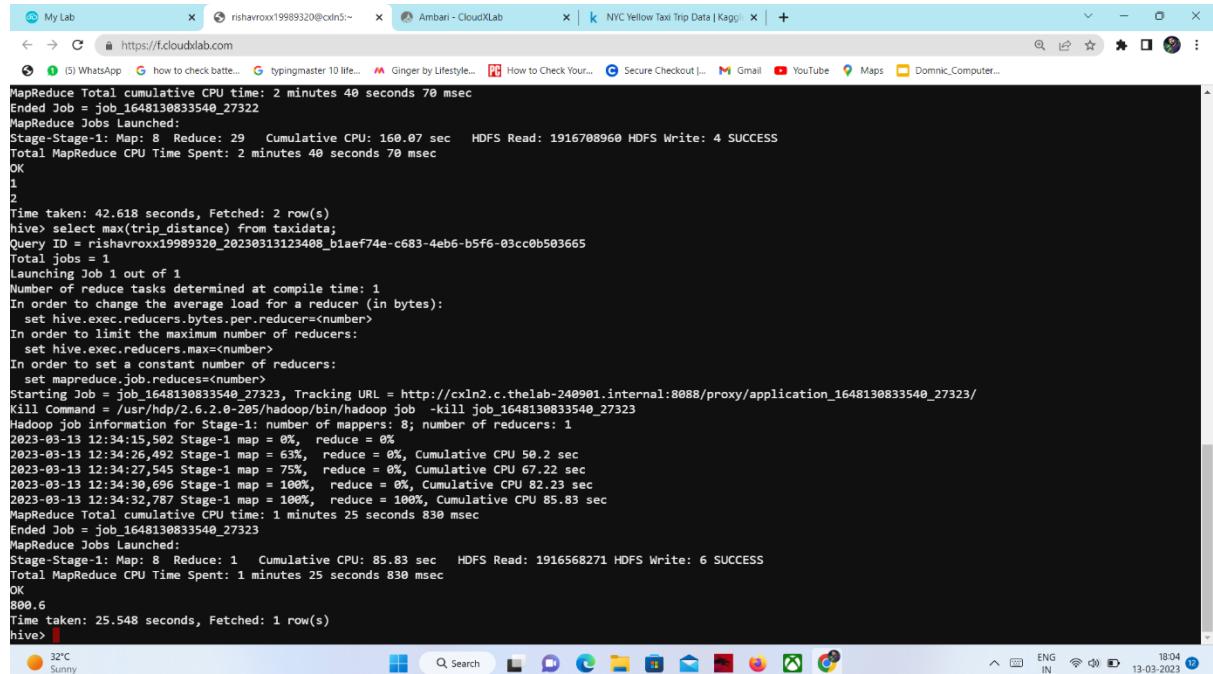


```
dropoff_longitude decimal(9,6)
dropoff_latitude decimal(9,6)
payment_type string
fare_amount decimal(9,6)
extra decimal(9,6)
mta_tax decimal(9,6)
tip_amount decimal(9,6)
tolls_amount decimal(9,6)
total_amount decimal(9,6)
trip_time_in_secs int

# Detailed Table Information
Database: project1
Owner: rishavrox19989320
Create time: Mon Mar 13 11:44:25 UTC 2023
Last access time: UNKNOWN
Protect Mode: None
Retention: 0
Location: hdfs://cxln1.c.thelab-240901.internal:8020/user/rishavrox19989320
Table Type: EXTERNAL_TABLE
Table Parameters:
  EXTERNAL          TRUE
  numFiles          1
  skip.header.line.count 1
  totalSize        1914669757
  transient_lastDdlTime 1678707865

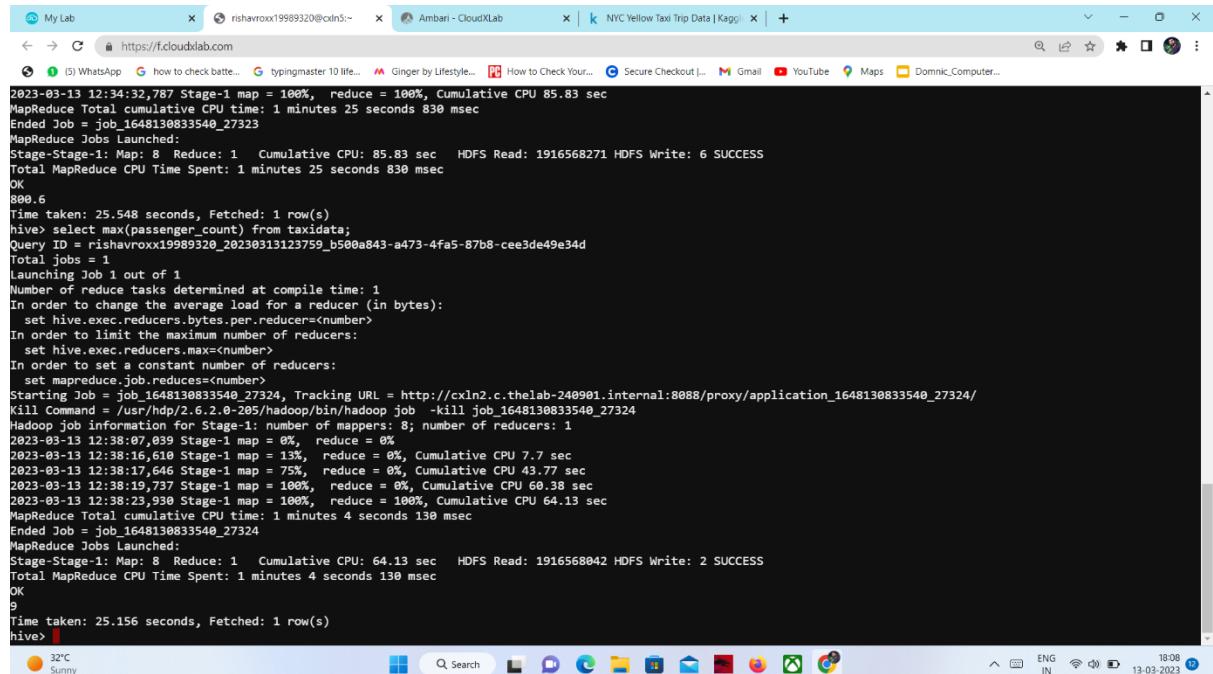
# Storage Information
SerDe Library: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat: org.apache.hadoop.mapred.TextInputFormat
OutputFormat: org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat
Compressed: No
Num Buckets: -1
Bucket Columns: []
Sort Columns: []
Storage Desc Params:
  field.delim   ,
  serialization.format  ,
Time taken: 0.462 seconds, Fetched: 49 row(s)
hive>
```

Q4) Listing the Distinct Vendors: -



```
MapReduce Total cumulative CPU time: 2 minutes 48 seconds 70 msec
Ended Job = job_1648130833540_27322
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 29 Cumulative CPU: 160.07 sec HDFS Read: 1916708960 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 48 seconds 70 msec
OK
1
2
Time taken: 42.618 seconds, Fetched: 2 row(s)
hive> select max(trip_distance) from taxidata;
Query ID = rishavrox19989320_20230313123408_b1aef74e-c683-4eb6-b5f6-03cc0b503665
Total jobs = 1
Launching Job 1 out of 1
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_1648130833540_27323, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27323
Kill Command = /usr/hdp/2.6.2.0-205/hadoop/bin/hadoop job -kill job_1648130833540_27323
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 12:34:15,502 Stage-1 map = 0%, reduce = 0%
2023-03-13 12:34:26,492 Stage-1 map = 63%, reduce = 0%, Cumulative CPU 50.2 sec
2023-03-13 12:34:27,545 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 67.22 sec
2023-03-13 12:34:30,699 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 82.23 sec
2023-03-13 12:34:32,787 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 85.83 sec
MapReduce Total cumulative CPU time: 1 minutes 25 seconds 830 msec
Ended Job = job_1648130833540_27323
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 85.83 sec HDFS Read: 1916568271 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 25 seconds 830 msec
OK
800.6
Time taken: 25.548 seconds, Fetched: 1 row(s)
hive>
```

Q5) Maximum Passengers hold by a vendor:-



```
2023-03-13 12:34:32,787 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 85.83 sec
MapReduce Total cumulative CPU time: 1 minutes 25 seconds 830 msec
Ended Job = job_1648130833540_27323
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 85.83 sec HDFS Read: 1916568271 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 25 seconds 830 msec
OK
800.6
Time taken: 25.548 seconds, Fetched: 1 row(s)
hive> select max(passenger_count) from taxidata;
Query ID = rishavrox19989320_20230313123759_b500843-a473-4fa5-87b8-cee3de49e34d
Total jobs = 1
Launching Job 1 out of 1
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_1648130833540_27324, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27324
Kill Command = /usr/hdp/2.6.2.0-205/hadoop/bin/hadoop job -kill job_1648130833540_27324
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 12:38:07,039 Stage-1 map = 0%, reduce = 0%
2023-03-13 12:38:16,610 Stage-1 map = 13%, reduce = 0%, Cumulative CPU 7.7 sec
2023-03-13 12:38:17,640 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 43.77 sec
2023-03-13 12:38:19,737 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 60.38 sec
2023-03-13 12:38:23,930 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 64.13 sec
MapReduce Total cumulative CPU time: 1 minutes 4 seconds 130 msec
Ended Job = job_1648130833540_27324
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 64.13 sec HDFS Read: 1916568042 HDFS Write: 2 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 4 seconds 130 msec
OK
9
Time taken: 25.156 seconds, Fetched: 1 row(s)
hive>
```

Q6) How many times each vendor carry their passengers: -

```

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_1648130833540_27327, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27327
Kill Command = /usr/hdp/2.6.2.0-285/hadoop/bin/hadoop job -kill job_1648130833540_27327
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 29
2023-03-13 12:50:17,944 Stage-1 map = 0%, reduce = 0%
2023-03-13 12:50:27,653 Stage-1 map = 25%, reduce = 0%, Cumulative CPU 15.24 sec
2023-03-13 12:50:28,697 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 35.55 sec
2023-03-13 12:50:29,738 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 74.74 sec
2023-03-13 12:50:30,799 Stage-1 map = 92%, reduce = 0%, Cumulative CPU 78.16 sec
2023-03-13 12:50:31,899 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 88.08 sec
2023-03-13 12:50:37,492 Stage-1 map = 100%, reduce = 14%, Cumulative CPU 95.73 sec
2023-03-13 12:50:38,619 Stage-1 map = 100%, reduce = 24%, Cumulative CPU 106.24 sec
2023-03-13 12:50:38,723 Stage-1 map = 100%, reduce = 28%, Cumulative CPU 118.86 sec
2023-03-13 12:50:40,924 Stage-1 map = 100%, reduce = 31%, Cumulative CPU 114.93 sec
2023-03-13 12:50:42,030 Stage-1 map = 100%, reduce = 38%, Cumulative CPU 123.83 sec
2023-03-13 12:50:43,124 Stage-1 map = 100%, reduce = 41%, Cumulative CPU 128.27 sec
2023-03-13 12:50:44,230 Stage-1 map = 100%, reduce = 48%, Cumulative CPU 135.32 sec
2023-03-13 12:50:45,333 Stage-1 map = 100%, reduce = 66%, Cumulative CPU 149.59 sec
2023-03-13 12:50:46,403 Stage-1 map = 100%, reduce = 72%, Cumulative CPU 160.47 sec
2023-03-13 12:50:47,495 Stage-1 map = 100%, reduce = 76%, Cumulative CPU 164.82 sec
2023-03-13 12:50:48,544 Stage-1 map = 100%, reduce = 83%, Cumulative CPU 172.1 sec
2023-03-13 12:50:49,594 Stage-1 map = 100%, reduce = 90%, Cumulative CPU 188.2 sec
2023-03-13 12:50:50,634 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 190.89 sec
MapReduce Total cumulative CPU time: 3 minutes 10 seconds 890 msec
Ended Job = job_1648130833540_27327
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 29 Cumulative CPU: 190.89 sec HDFS Read: 1916729061 HDFS Write: 25 SUCCESS
Total MapReduce CPU Time Spent: 3 minutes 10 seconds 890 msec
OK
1      16568749
2      19489854.25
Time taken: 40.967 seconds, Fetched: 2 row(s)
hive>
```

32°C Sunny 18:21 13-03-2023

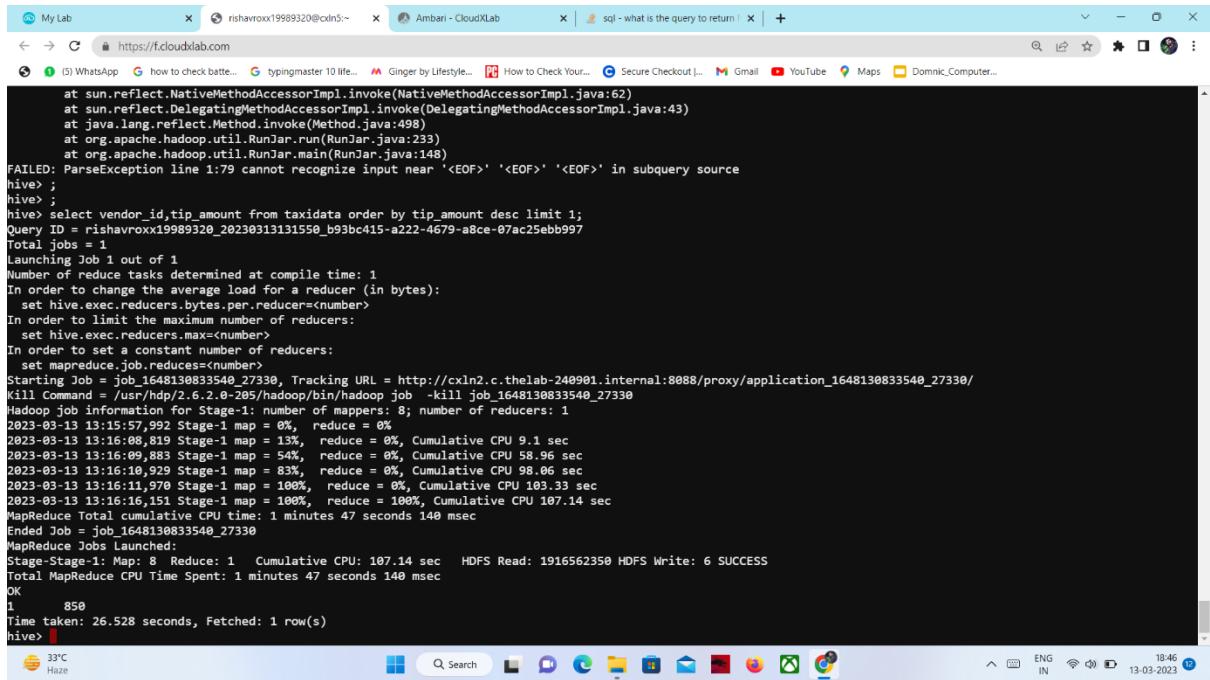
Q7) Maximum tip to the vendor ever: -

```

hive> select vendor_id , sum(trip_distance) as distance from taxidata group by vendor_id where distance=max(distance);
FAILED: ParseException line:1:83 missing EOF at 'where' near 'vendor_id'
hive> select vendor_id where tip_amount=max(tip_amount);
FAILED: SemanticException [Error 10004]: Line 1:23 Invalid table alias or column reference 'tip_amount': (possible column names are: )
hive> select vendor_id from taxidata where tip_amount=max(tip_amount);
FAILED: SemanticException [Error 10128]: Line 1:48 Not yet supported place for UDAF 'max'
hive> select vendor_id from taxidata where tip_amount=Max(tip_amount);
FAILED: SemanticException [Error 10128]: Line 1:48 Not yet supported place for UDAF 'Max'
hive> select Max(tip_amount)from taxidata;
Query ID = rishavroxx19989320_20230313125738_d886acf4-4ca6-4a48-9fca-d6cea819335d
Total jobs = 1
Launching Job 1 out of 1
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_1648130833540_27328, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27328
Kill Command = /usr/hdp/2.6.2.0-285/hadoop/bin/hadoop job -kill job_1648130833540_27328
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 12:57:45,325 Stage-1 map = 0%, reduce = 0%
2023-03-13 12:57:55,923 Stage-1 map = 38%, reduce = 0%, Cumulative CPU 24.36 sec
2023-03-13 12:57:56,968 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 56.83 sec
2023-03-13 12:57:59,007 Stage-1 map = 92%, reduce = 0%, Cumulative CPU 76.76 sec
2023-03-13 12:57:59,045 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 76.97 sec
2023-03-13 12:58:03,228 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 88.69 sec
MapReduce Total cumulative CPU time: 1 minutes 20 seconds 690 msec
Ended Job = job_1648130833540_27328
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 80.69 sec HDFS Read: 1916568212 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 20 seconds 690 msec
OK
850
Time taken: 26.09 seconds, Fetched: 1 row(s)
hive>
```

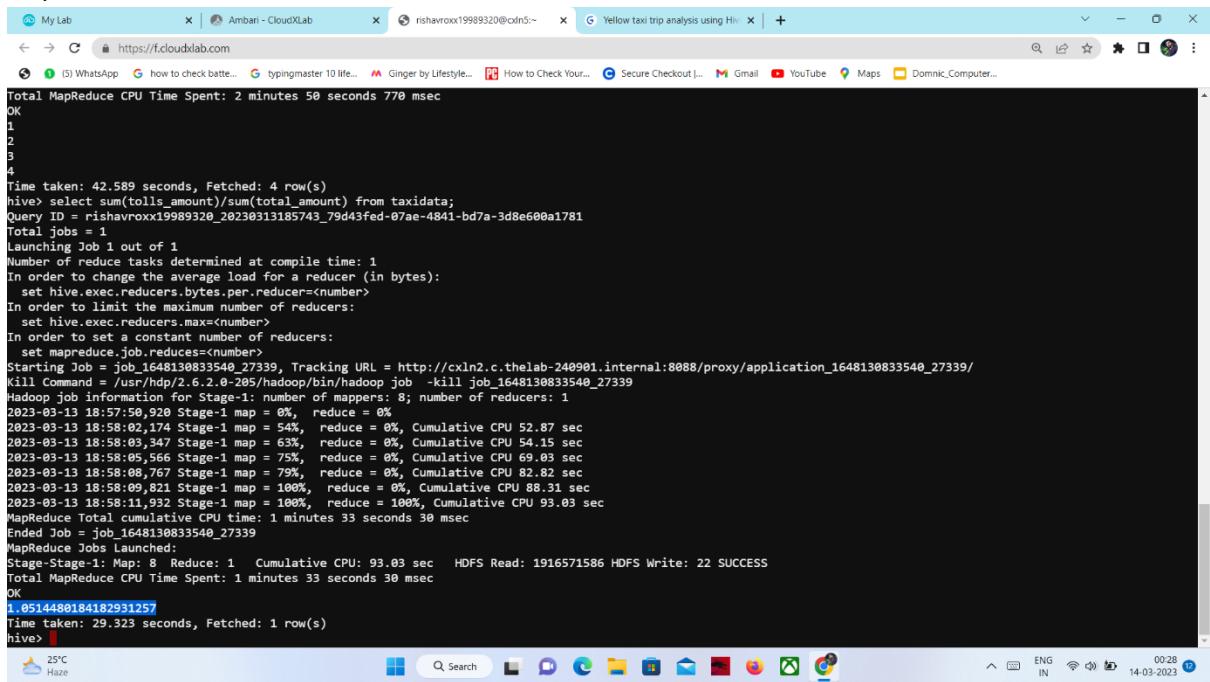
32°C Sunny 18:28 13-03-2023

Q8) which vendor have been given maximum tip and how much:-



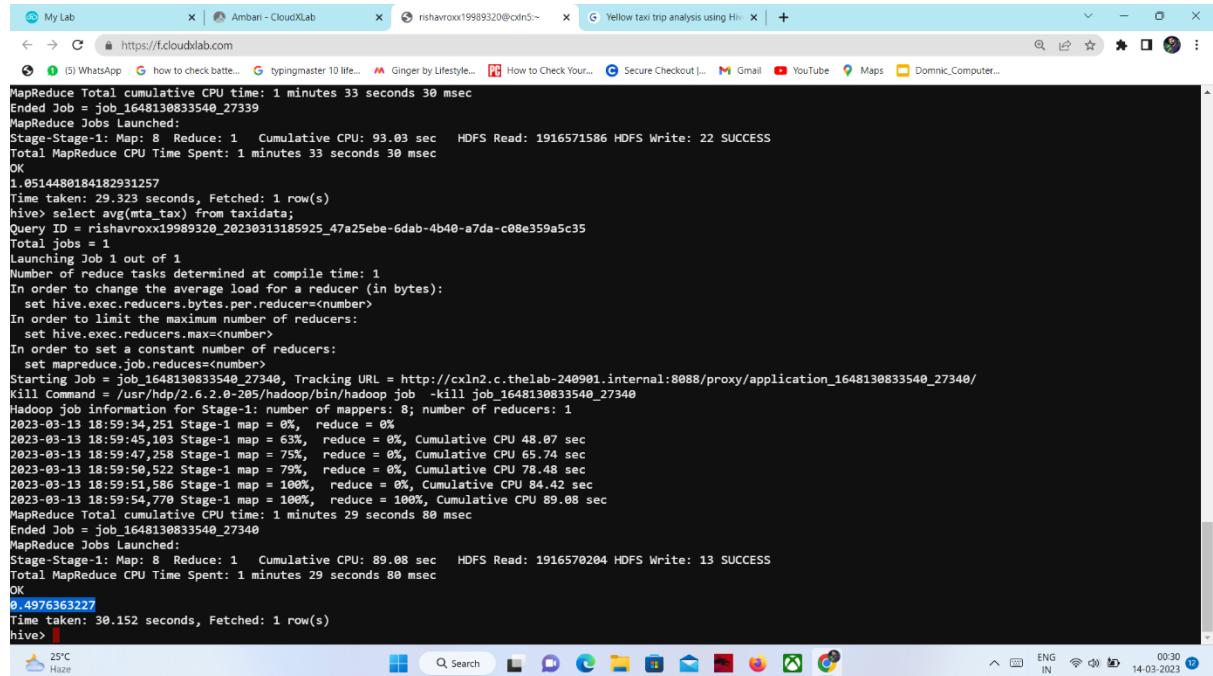
```
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:233)
at org.apache.hadoop.util.RunJar.main(RunJar.java:148)
FAILED: ParseException line 1:79 cannot recognize input near '<EOF>' '<EOF>' '<EOF>' in subquery source
hive> ;
hive> select vendor_id,tip_amount from taxidata order by tip_amount desc limit 1;
Query ID = rishavrox19989320_20230313131550_b93bc415-a222-4679-a8ce-07ac25ebb997
Total jobs = 1
Launching Job 1 out of 1
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=<n>number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<n>number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<n>number>
Starting Job = job_1648130833540_27338, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27338
Kill Command = /usr/hdp/2.6.2.0-285/hadoop/bin/hadoop job -kill job_1648130833540_27338
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 13:15:57,992 Stage-1 map = 0%, reduce = 0%
2023-03-13 13:16:08,819 Stage-1 map = 13%, reduce = 0%, Cumulative CPU 9.1 sec
2023-03-13 13:16:09,883 Stage-1 map = 54%, reduce = 0%, Cumulative CPU 58.96 sec
2023-03-13 13:16:10,929 Stage-1 map = 83%, reduce = 0%, Cumulative CPU 98.06 sec
2023-03-13 13:16:11,970 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 103.33 sec
2023-03-13 13:16:16,151 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 107.14 sec
MapReduce Total cumulative CPU time: 1 minutes 47 seconds 140 msec
Ended Job = job_1648130833540_27338
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 107.14 sec HDFS Read: 1916562350 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 47 seconds 140 msec
OK
1          850
Time taken: 26.528 seconds, Fetched: 1 row(s)
hive>
```

Q9) Fraction to tolls amount w.r.t to total amount:-



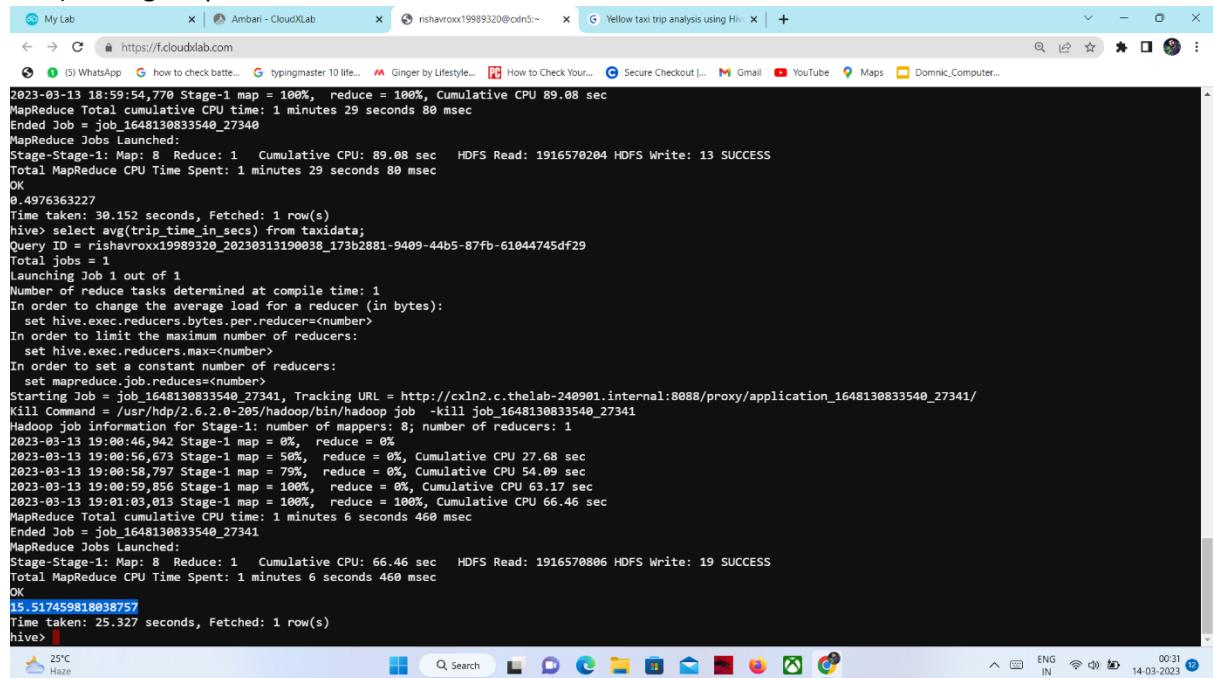
```
Total MapReduce CPU Time Spent: 2 minutes 50 seconds 770 msec
OK
1
2
3
4
Time taken: 42.589 seconds, Fetched: 4 row(s)
hive> select sum(tolls_amount)/sum(total_amount) from taxidata;
Query ID = rishavrox19989320_20230313185743_79d43fed-07ae-4841-bd7a-3d8e600a1781
Total jobs = 1
Launching Job 1 out of 1
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=<n>number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<n>number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<n>number>
Starting Job = job_1648130833540_27339, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27339
Kill Command = /usr/hdp/2.6.2.0-285/hadoop/bin/hadoop job -kill job_1648130833540_27339
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 18:57:50,920 Stage-1 map = 0%, reduce = 0%
2023-03-13 18:58:02,174 Stage-1 map = 54%, reduce = 0%, Cumulative CPU 52.87 sec
2023-03-13 18:58:03,347 Stage-1 map = 63%, reduce = 0%, Cumulative CPU 54.15 sec
2023-03-13 18:58:05,566 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 69.03 sec
2023-03-13 18:58:08,767 Stage-1 map = 79%, reduce = 0%, Cumulative CPU 82.82 sec
2023-03-13 18:58:09,821 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 88.31 sec
2023-03-13 18:58:11,932 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 93.03 sec
MapReduce Total cumulative CPU time: 1 minutes 33 seconds 30 msec
Ended Job = job_1648130833540_27339
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 93.03 sec HDFS Read: 1916571586 HDFS Write: 22 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 33 seconds 30 msec
OK
1.0514480184182931257
Time taken: 29.323 seconds, Fetched: 1 row(s)
hive>
```

Q10) On an average how much mta_tax has been given:-



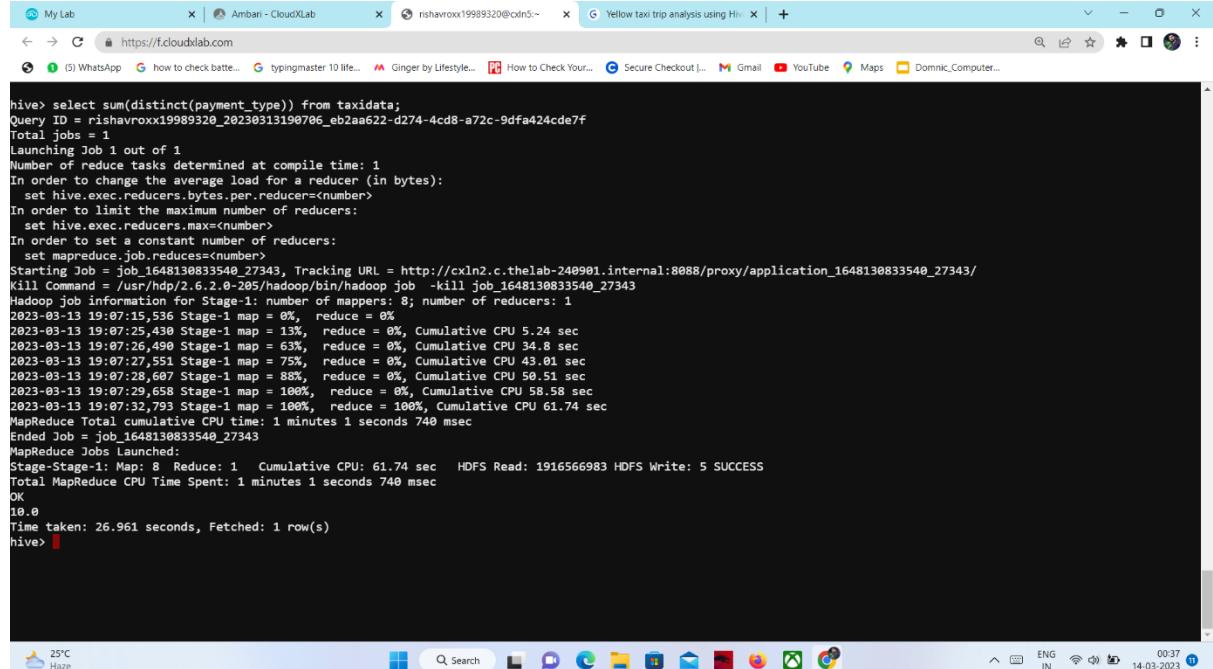
```
MapReduce Total cumulative CPU time: 1 minutes 33 seconds 30 msec
Ended Job = job_1648130833540_27339
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 93.03 sec HDFS Read: 1916571586 HDFS Write: 22 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 33 seconds 30 msec
OK
0.0514480184182931257
Time taken: 29.323 seconds, Fetched: 1 row(s)
hive> select avg(mta_tax) from taxidata;
Query ID = rishavroxx19989320_20230313185925_47a25ebe-6dab-4b40-a7da-c08e359a5c35
Total jobs = 1
Launching Job 1 out of 1
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_1648130833540_27340, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27340/
Kill Command = /usr/hdp/2.6.2.0-285/hadoop/bin/hadoop job -kill job_1648130833540_27340
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 18:59:34,251 Stage-1 map = 0%, reduce = 0%
2023-03-13 18:59:45,103 Stage-1 map = 63%, reduce = 0%, Cumulative CPU 48.07 sec
2023-03-13 18:59:47,258 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 65.74 sec
2023-03-13 18:59:50,522 Stage-1 map = 79%, reduce = 0%, Cumulative CPU 78.48 sec
2023-03-13 18:59:51,580 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 84.42 sec
2023-03-13 18:59:54,770 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 89.08 sec
MapReduce Total cumulative CPU time: 1 minutes 29 seconds 80 msec
Ended Job = job_1648130833540_27340
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 89.08 sec HDFS Read: 1916570204 HDFS Write: 13 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 29 seconds 80 msec
OK
0.4976363227
Time taken: 30.152 seconds, Fetched: 1 row(s)
hive>
```

Q11) Average trip time :-



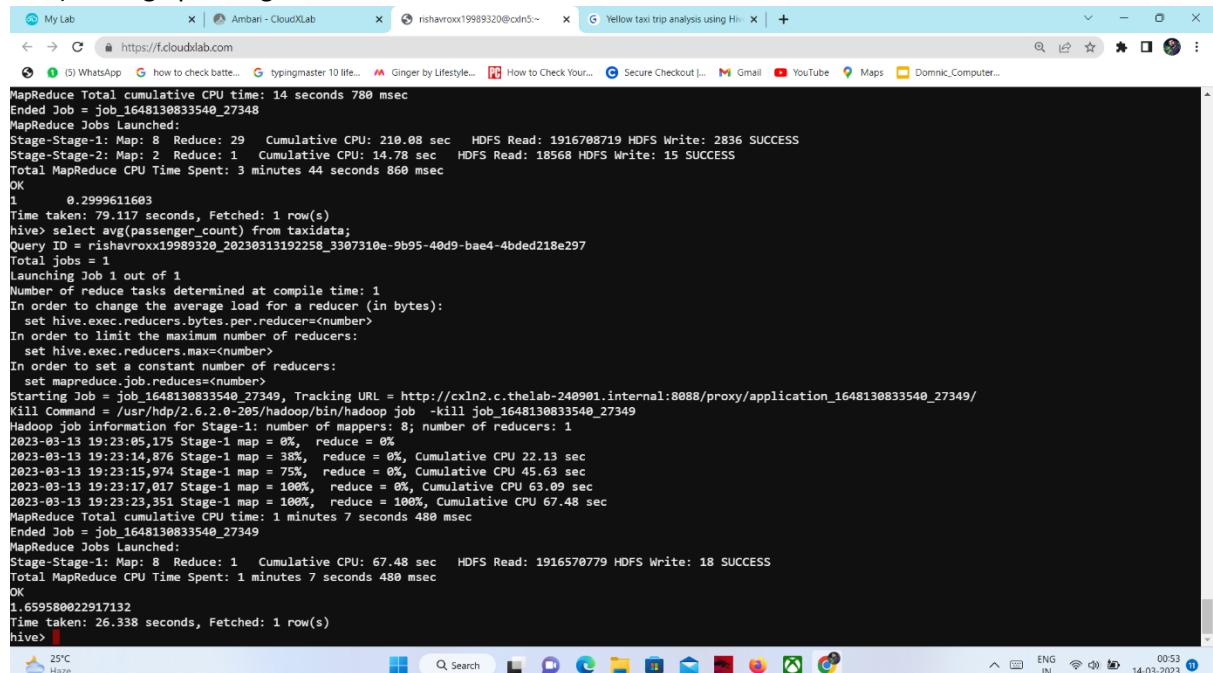
```
2023-03-13 18:59:54,770 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 89.08 sec
MapReduce Total cumulative CPU time: 1 minutes 29 seconds 80 msec
Ended Job = job_1648130833540_27340
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 89.08 sec HDFS Read: 1916570204 HDFS Write: 13 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 29 seconds 80 msec
OK
0.4976363227
Time taken: 30.152 seconds, Fetched: 1 row(s)
hive> select avg(trip_time_in_secs) from taxidata;
Query ID = rishavroxx19989320_20230313190038_173b2881-9409-44b5-87fb-61844745df29
Total jobs = 1
Launching Job 1 out of 1
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_1648130833540_27341, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27341/
Kill Command = /usr/hdp/2.6.2.0-285/hadoop/bin/hadoop job -kill job_1648130833540_27341
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 19:00:46,942 Stage-1 map = 0%, reduce = 0%
2023-03-13 19:00:56,673 Stage-1 map = 56%, reduce = 0%, Cumulative CPU 27.68 sec
2023-03-13 19:00:58,797 Stage-1 map = 79%, reduce = 0%, Cumulative CPU 54.09 sec
2023-03-13 19:00:59,856 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 63.17 sec
2023-03-13 19:01:03,013 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 66.46 sec
MapReduce Total cumulative CPU time: 1 minutes 6 seconds 460 msec
Ended Job = job_1648130833540_27341
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 66.46 sec HDFS Read: 1916570806 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 6 seconds 460 msec
OK
15.517459818038757
Time taken: 25.327 seconds, Fetched: 1 row(s)
hive>
```

Q12) How many different payment types are used:-



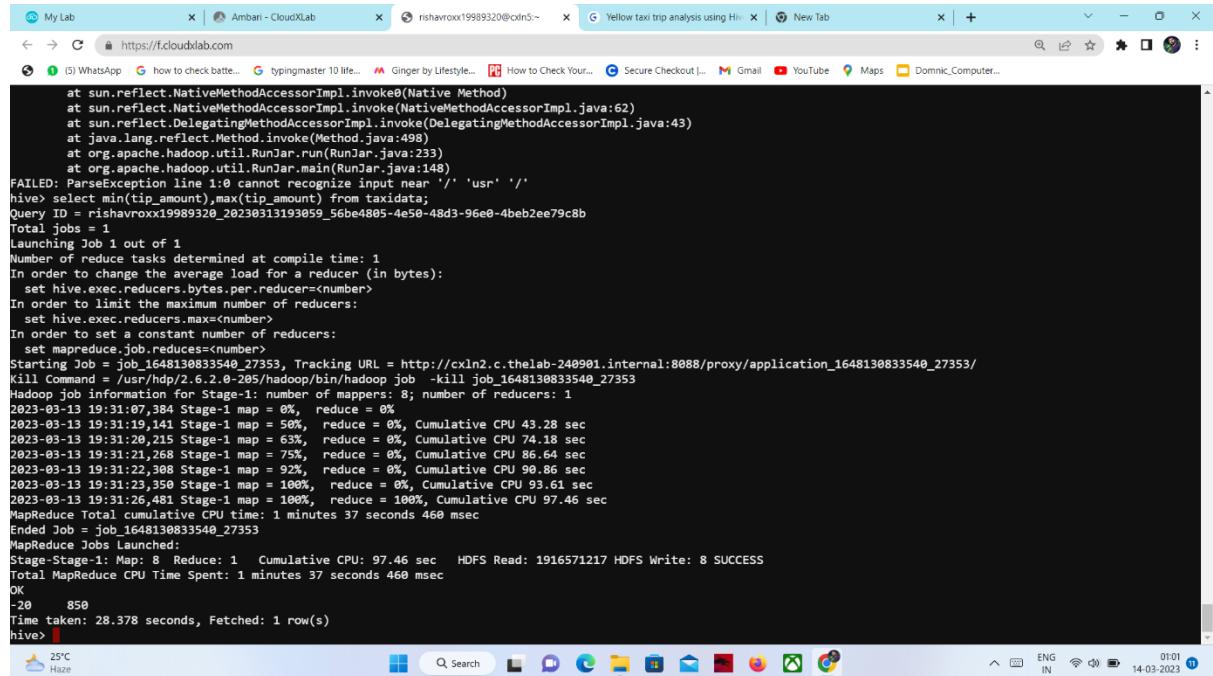
```
hive> select sum(distinct(payment_type)) from taxidata;
Query ID = rishavroxx19989320_20230313190706 Eb2aa622-d274-4cd8-a72c-9dfa424cde7f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set 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_1648130833540_27343, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27343
Kill Command = /usr/hdp/2.6.2.0-285/hadoop/bin/hadoop job -kill job_1648130833540_27343
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 19:07:15,536 Stage-1 map = 0%, reduce = 0%
2023-03-13 19:07:25,436 Stage-1 map = 13%, reduce = 0%, Cumulative CPU 5.24 sec
2023-03-13 19:07:26,490 Stage-1 map = 63%, reduce = 0%, Cumulative CPU 34.8 sec
2023-03-13 19:07:27,551 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 43.01 sec
2023-03-13 19:07:28,607 Stage-1 map = 88%, reduce = 0%, Cumulative CPU 50.51 sec
2023-03-13 19:07:29,658 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 58.58 sec
2023-03-13 19:07:32,793 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 61.74 sec
MapReduce Total cumulative CPU time: 1 minutes 1 seconds 740 msec
Ended Job = job_1648130833540_27343
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 61.74 sec HDFS Read: 1916566983 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 1 seconds 740 msec
OK
10.0
Time taken: 26.961 seconds, Fetched: 1 row(s)
hive>
```

Q13) Average passenger count:



```
MapReduce Total cumulative CPU time: 14 seconds 780 msec
Ended Job = job_1648130833540_27348
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 29 Cumulative CPU: 210.08 sec HDFS Read: 1916708719 HDFS Write: 2836 SUCCESS
Stage-Stage-2: Map: 2 Reduce: 1 Cumulative CPU: 14.78 sec HDFS Read: 18568 HDFS Write: 15 SUCCESS
Total MapReduce CPU Time Spent: 3 minutes 44 seconds 860 msec
OK
1.0 0.2999611603
Time taken: 79.117 seconds, Fetched: 1 row(s)
hive> select avg(passenger_count) from taxidata;
Query ID = rishavroxx19989320_20230313192258_3307310e-9b95-40d9-bae4-4bded218e297
Total jobs = 1
Launching Job 1 out of 1
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_1648130833540_27349, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27349
Kill Command = /usr/hdp/2.6.2.0-285/hadoop/bin/hadoop job -kill job_1648130833540_27349
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 19:23:05,175 Stage-1 map = 0%, reduce = 0%
2023-03-13 19:23:14,876 Stage-1 map = 38%, reduce = 0%, Cumulative CPU 22.13 sec
2023-03-13 19:23:15,974 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 45.63 sec
2023-03-13 19:23:17,017 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 63.09 sec
2023-03-13 19:23:23,351 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 67.48 sec
MapReduce Total cumulative CPU time: 1 minutes 7 seconds 480 msec
Ended Job = job_1648130833540_27349
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1 Cumulative CPU: 67.48 sec HDFS Read: 1916570779 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 7 seconds 480 msec
OK
1.659580022917132
Time taken: 26.338 seconds, Fetched: 1 row(s)
hive>
```

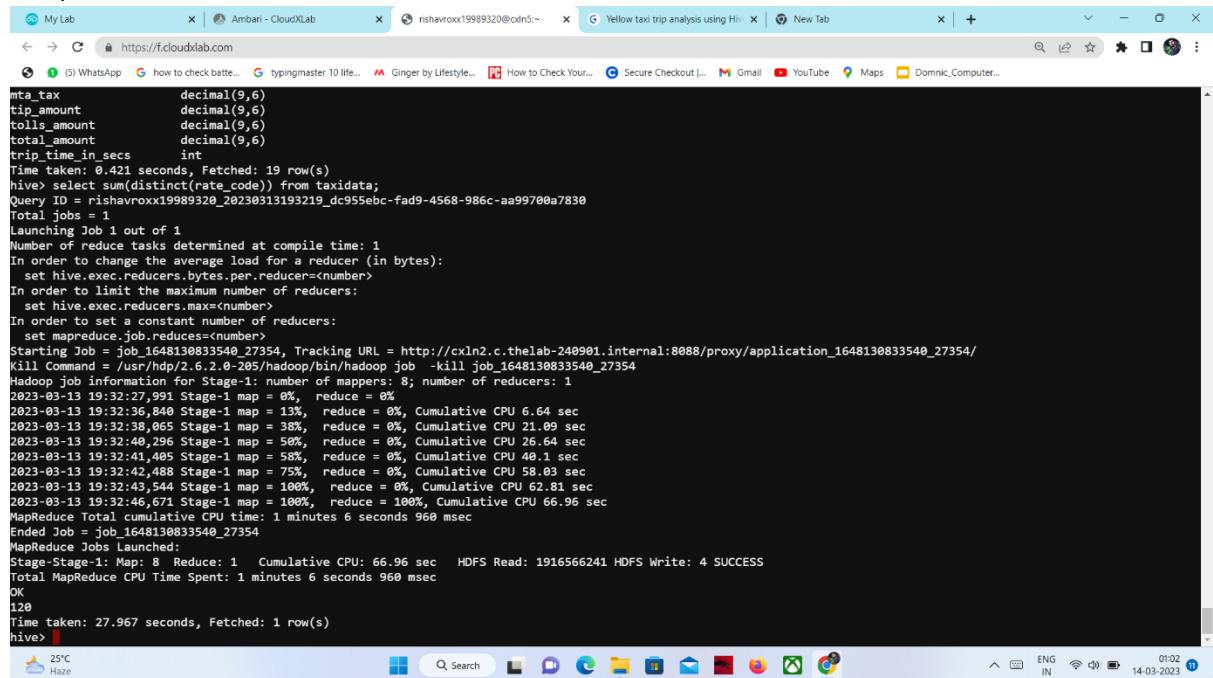
Q14) Maximum time spent by passenger:-



```
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:233)
at org.apache.hadoop.util.RunJar.main(RunJar.java:148)
FAILED: ParseException line 1:0 cannot recognize input near '/ 'usr' '/'

hive> select min(tip_amount),max(tip_amount) from taxidata;
Query ID = rishavrox19989320_20230313193859_56be4805-4e50-48d3-96e0-4beb2ee79c8b
Total jobs = 1
Launching Job 1 out of 1
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_1648130833540_27353, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27353
Kill Command = /usr/hdp/2.6.2.0-285/hadoop/bin/hadoop job -kill job_1648130833540_27353
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 19:31:07,384 Stage-1 map = 0%, reduce = 0%
2023-03-13 19:31:19,141 Stage-1 map = 56%, reduce = 0%, Cumulative CPU 43.28 sec
2023-03-13 19:31:20,215 Stage-1 map = 63%, reduce = 0%, Cumulative CPU 74.18 sec
2023-03-13 19:31:21,268 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 86.64 sec
2023-03-13 19:31:22,304 Stage-1 map = 92%, reduce = 0%, Cumulative CPU 90.86 sec
2023-03-13 19:31:23,350 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 93.61 sec
2023-03-13 19:31:26,481 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 97.46 sec
MapReduce Total cumulative CPU time: 1 minutes 37 seconds 460 msec
Ended Job = job_1648130833540_27353
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1  Cumulative CPU: 97.46 sec  HDFS Read: 1916571217 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 37 seconds 460 msec
OK
-20     850
Time taken: 28.378 seconds, Fetched: 1 row(s)
hive>
```

Q15)Distinct Rate Code:



```
mta_tax      decimal(9,6)
tip_amount    decimal(9,6)
tolls_amount  decimal(9,6)
total_amount  decimal(9,6)
trip_time_in_secs int
Time taken: 0.421 seconds, Fetched: 19 row(s)
hive> select sum(distinct(rate_code)) from taxidata;
Query ID = rishavrox19989320_20230313193219_dc955ebc-fad9-4568-98ec-aa99700a7830
Total jobs = 1
Launching Job 1 out of 1
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_1648130833540_27354, Tracking URL = http://cxln2.c.thelab-240901.internal:8088/proxy/application_1648130833540_27354
Kill Command = /usr/hdp/2.6.2.0-285/hadoop/bin/hadoop job -kill job_1648130833540_27354
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2023-03-13 19:32:27,994 Stage-1 map = 0%, reduce = 0%
2023-03-13 19:32:36,844 Stage-1 map = 13%, reduce = 0%, Cumulative CPU 6.64 sec
2023-03-13 19:32:38,065 Stage-1 map = 38%, reduce = 0%, Cumulative CPU 21.09 sec
2023-03-13 19:32:40,294 Stage-1 map = 56%, reduce = 0%, Cumulative CPU 26.64 sec
2023-03-13 19:32:41,405 Stage-1 map = 58%, reduce = 0%, Cumulative CPU 40.1 sec
2023-03-13 19:32:42,488 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 58.03 sec
2023-03-13 19:32:43,544 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 62.81 sec
2023-03-13 19:32:46,671 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 66.96 sec
MapReduce Total cumulative CPU time: 1 minutes 6 seconds 960 msec
Ended Job = job_1648130833540_27354
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8 Reduce: 1  Cumulative CPU: 66.96 sec  HDFS Read: 1916566241 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 6 seconds 960 msec
OK
120
Time taken: 27.967 seconds, Fetched: 1 row(s)
hive>
```