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

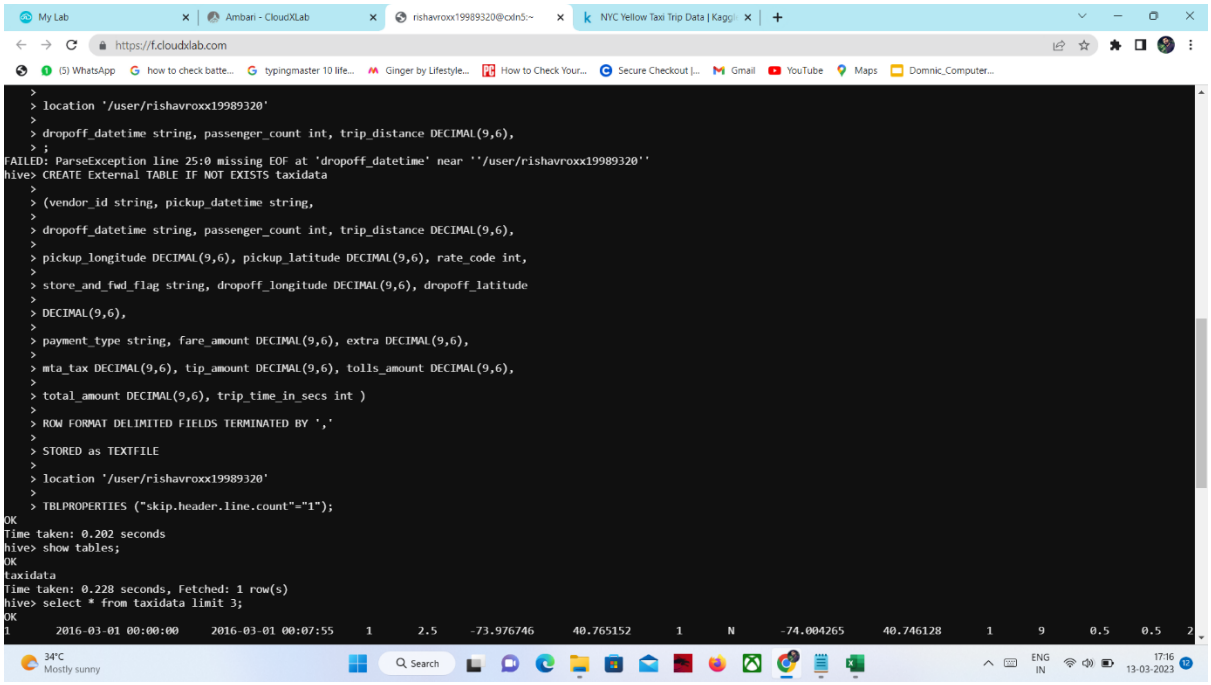
1. vendor_id string,
2. pickup_datetime string,
3. dropoff_datetime string,
4. passenger_count int,
5. trip_distance DECIMAL(9,6),
6. pickup_longitude DECIMAL(9,6),
7. pickup_latitude DECIMAL(9,6),
8. rate_code int,
9. store_and_fwd_flag string,
10. dropoff_longitude DECIMAL(9,6),
11. dropoff_latitude 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:

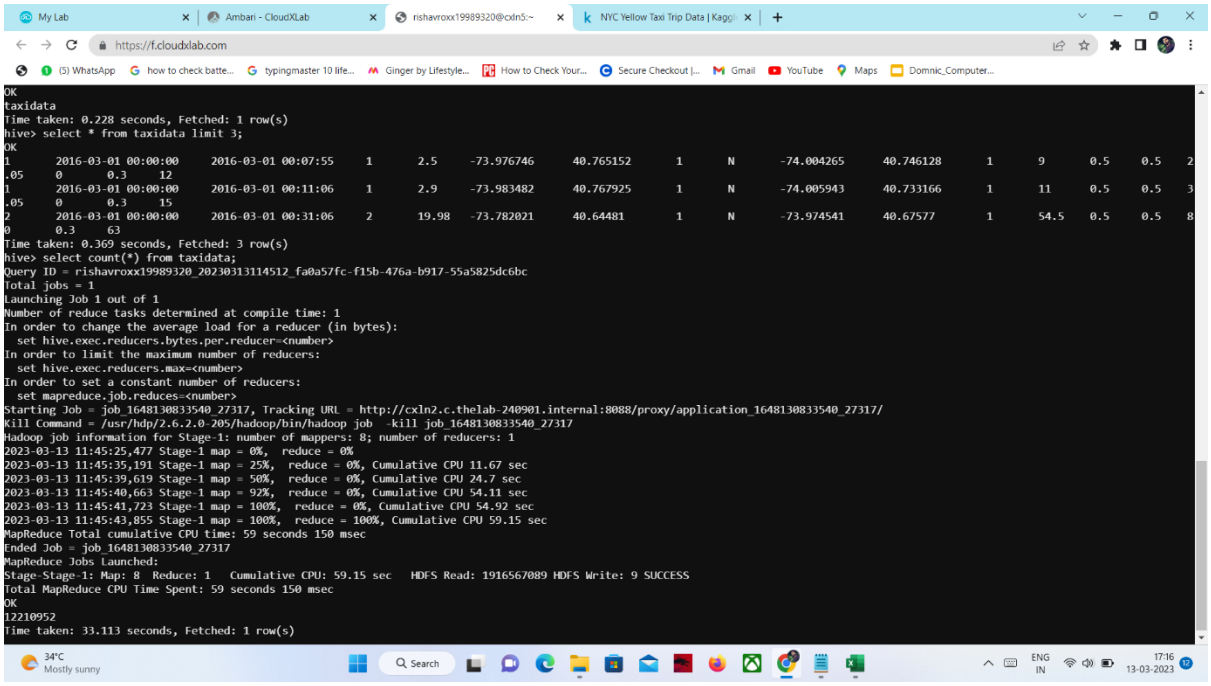
1. What is the total Number of trips (equal to the number of rows)?
2. What is the total revenue generated by all the trips? The fare is stored in the column total_amount.
3. What fraction of the total is paid for tolls? The toll is stored in tolls_amount.
4. What fraction of it is driver tips? The tip 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:-

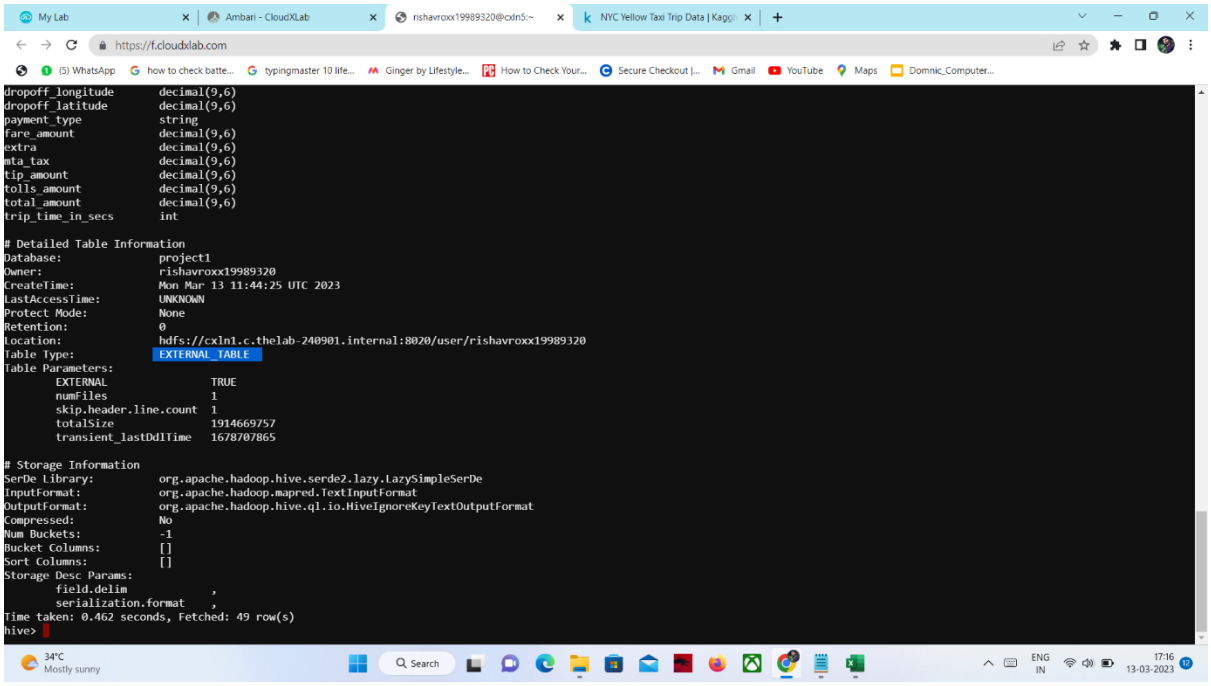


Q.2) Finding no of Records:



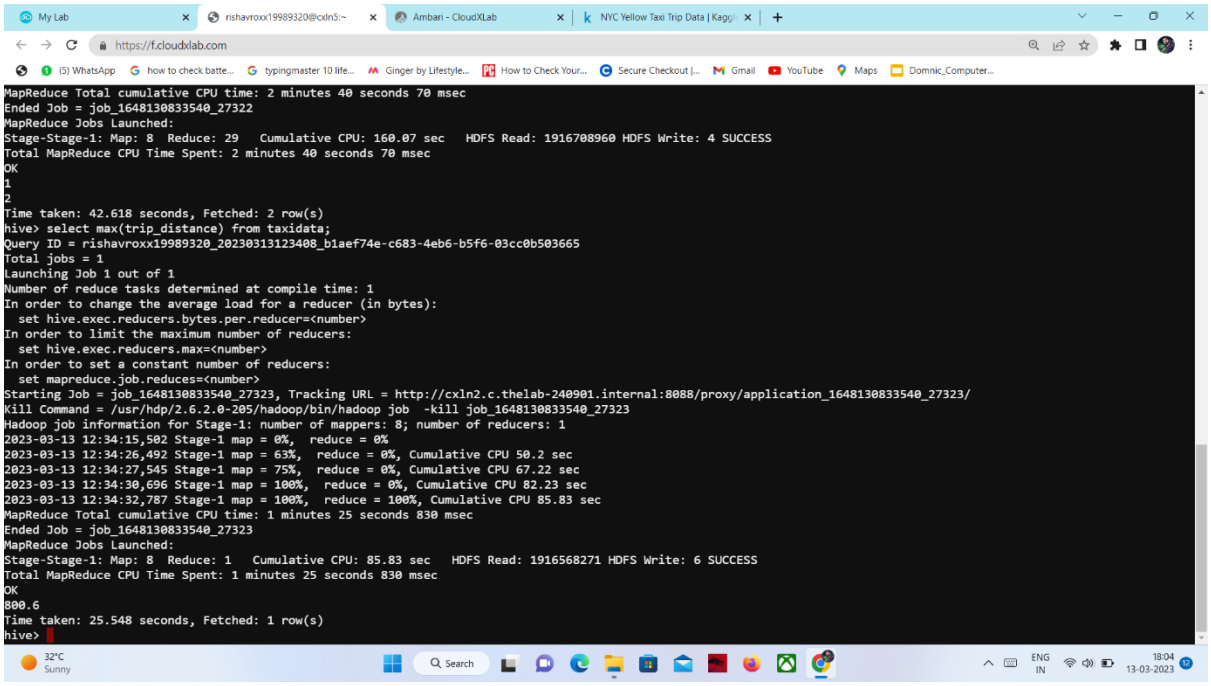
```
OK
taxidata
Time taken: 0.228 seconds, Fetched: 1 row(s)
hive> select * from taxidata limit 3;
OK
1       2016-03-01 00:00:00     2016-03-01 00:07:55     1       2.5     -73.976746      40.765152       1       N       -74.004265      40.746128       1       9       0.5     0.5     2
.05     0       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       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       0.3     63
Time taken: 0.369 seconds, Fetched: 3 row(s)
hive> select count(*) from taxidata;
Query ID = rishavroxx19989320_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=<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_27317, Tracking URL = http://cxln2.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 11:45:25,477 Stage-1 map = 0%,  reduce = 0%
2023-03-13 11:45:35,191 Stage-1 map = 25%,  reduce = 0%, Cumulative CPU 11.67 sec
2023-03-13 11:45:39,619 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 24.7 sec
2023-03-13 11:45:40,663 Stage-1 map = 92%,  reduce = 0%, Cumulative CPU 54.11 sec
2023-03-13 11:45:41,723 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 54.92 sec
2023-03-13 11:45:43,855 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: -

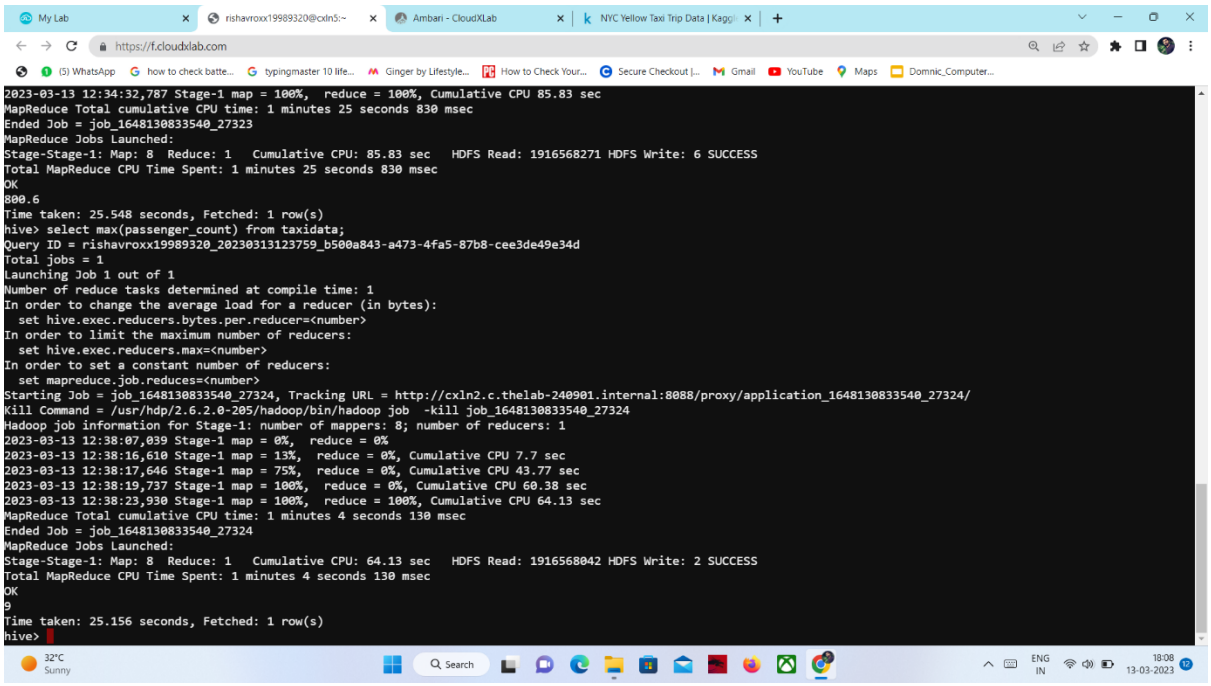


```
dropoff_longitude       decimal(9,6)
dropoff_latitude        decimal(9,6)
payment_type            string
fare_amount             decimal(9,6)
extra                   decimal(9,6)
mta_tax                 decimal(9,6)
tip_amount              decimal(9,6)
tolls_amount            decimal(9,6)
total_amount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

# Detailed Table Information
Database:               project1
Owner:                  rishavroxx19989320
CreateTime:             Mon Mar 13 11:44:25 UTC 2023
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://cxln1.c.thelab-240901.internal:8020/user/rishavroxx19989320
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.ql.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>
```

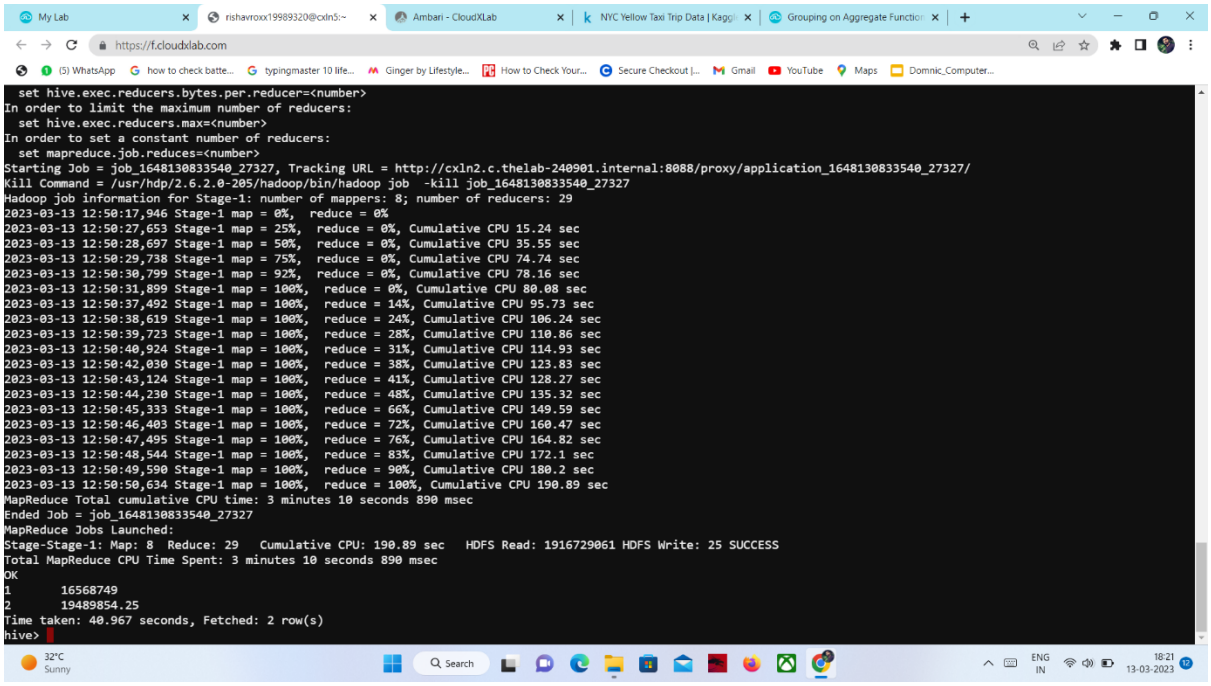
Q4) Listing the Distinct Vendors: -



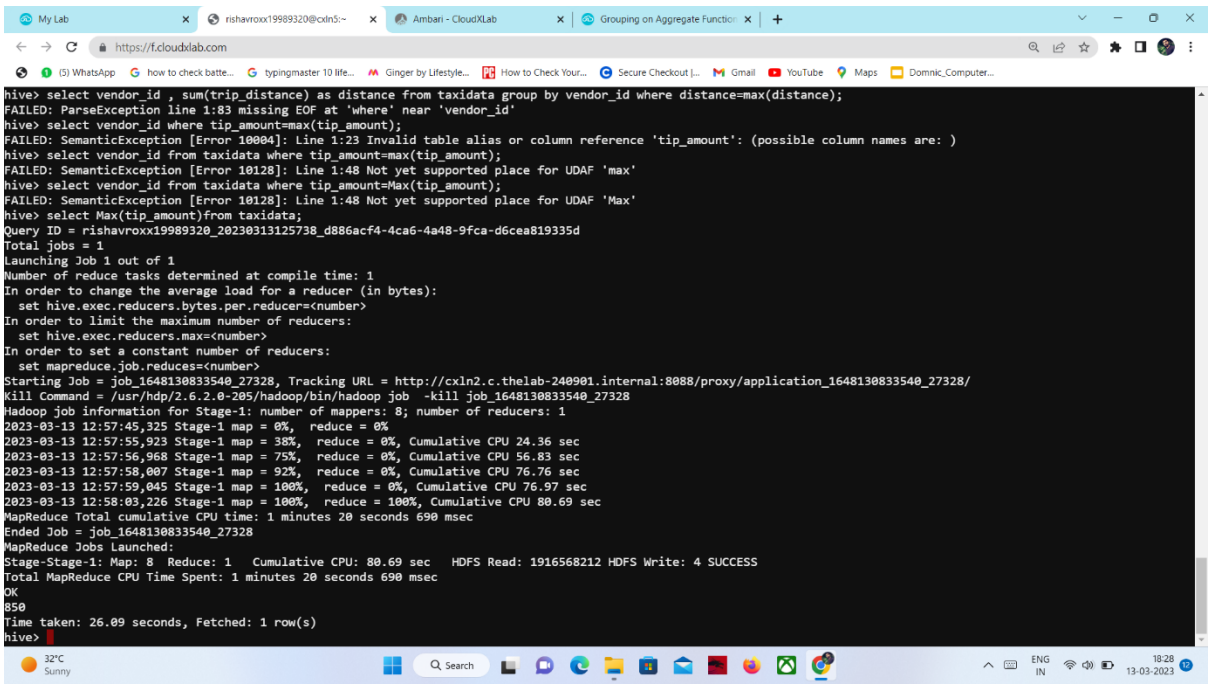
Q5) Maximum Passengers hold by a vendor:-



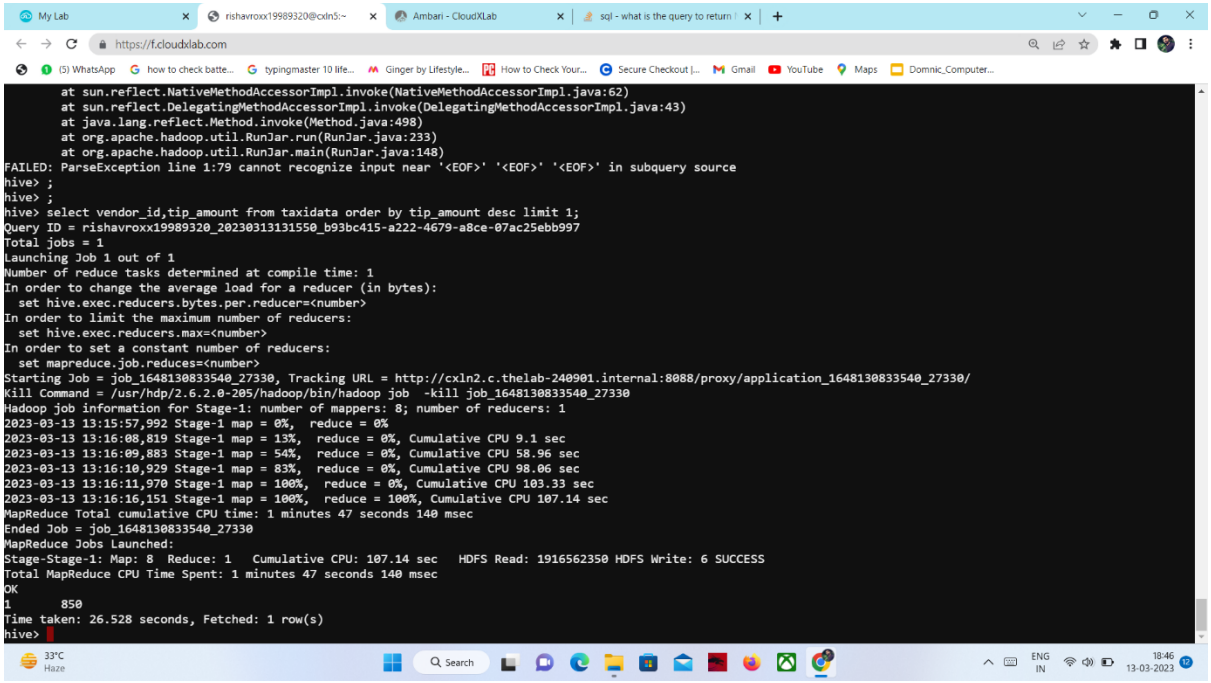
Q6) How many times each vendor carry their passengers: -



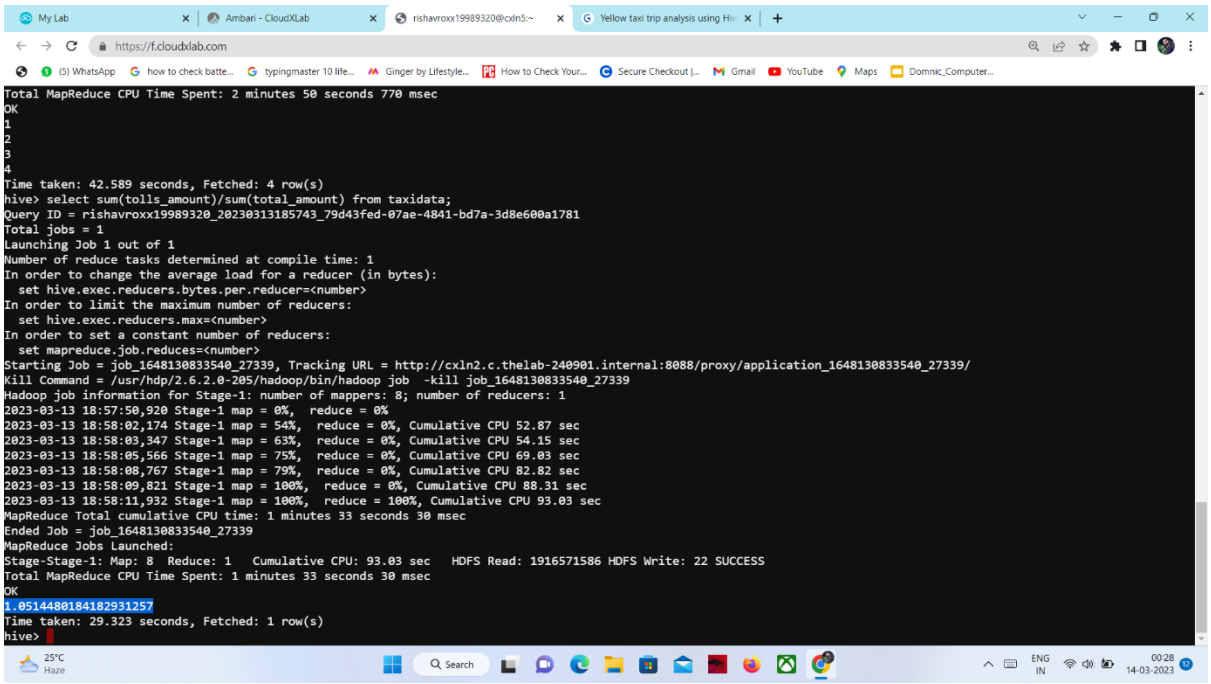
Q7) Maximum tip to the vendor ever: -



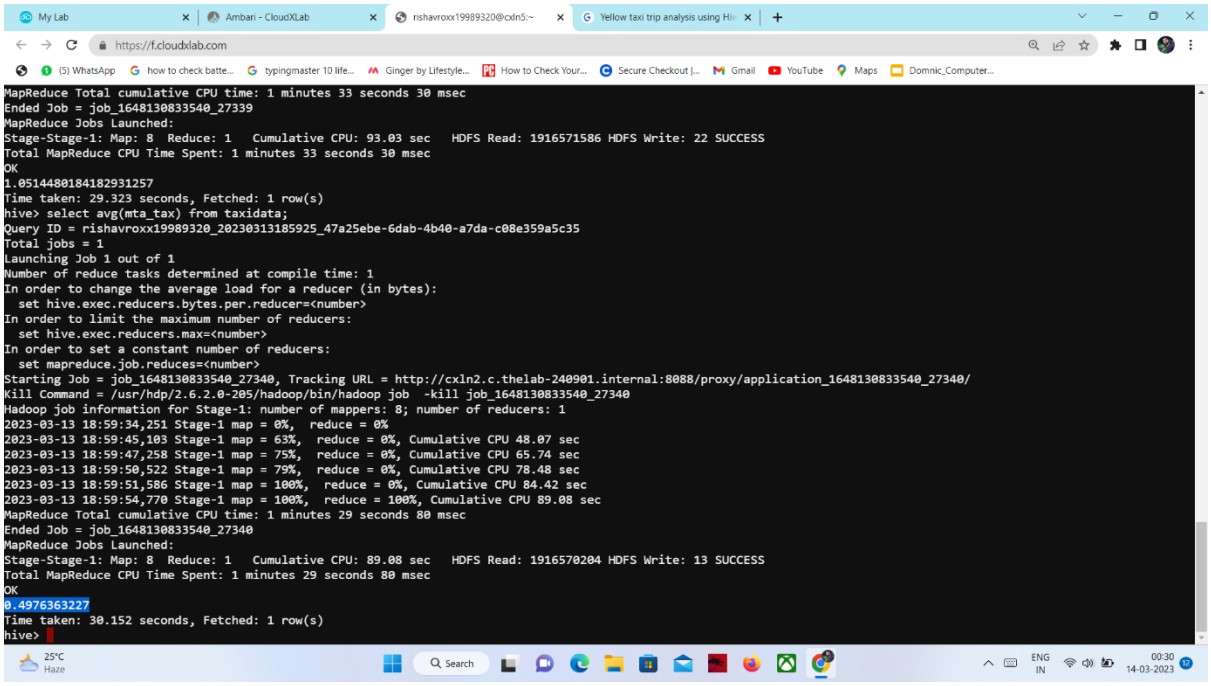
Q8) which vendor have been given maximum tip and how much:-



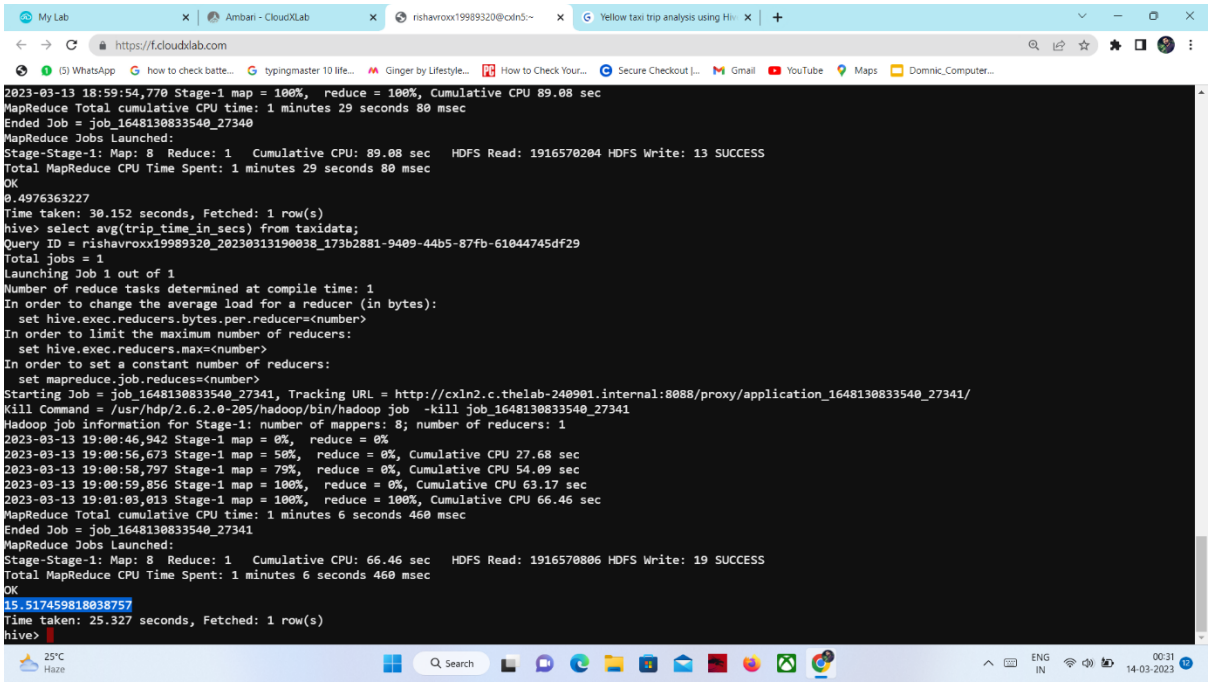
Q9) Fraction to tolls amount w.r.t to total amount:-



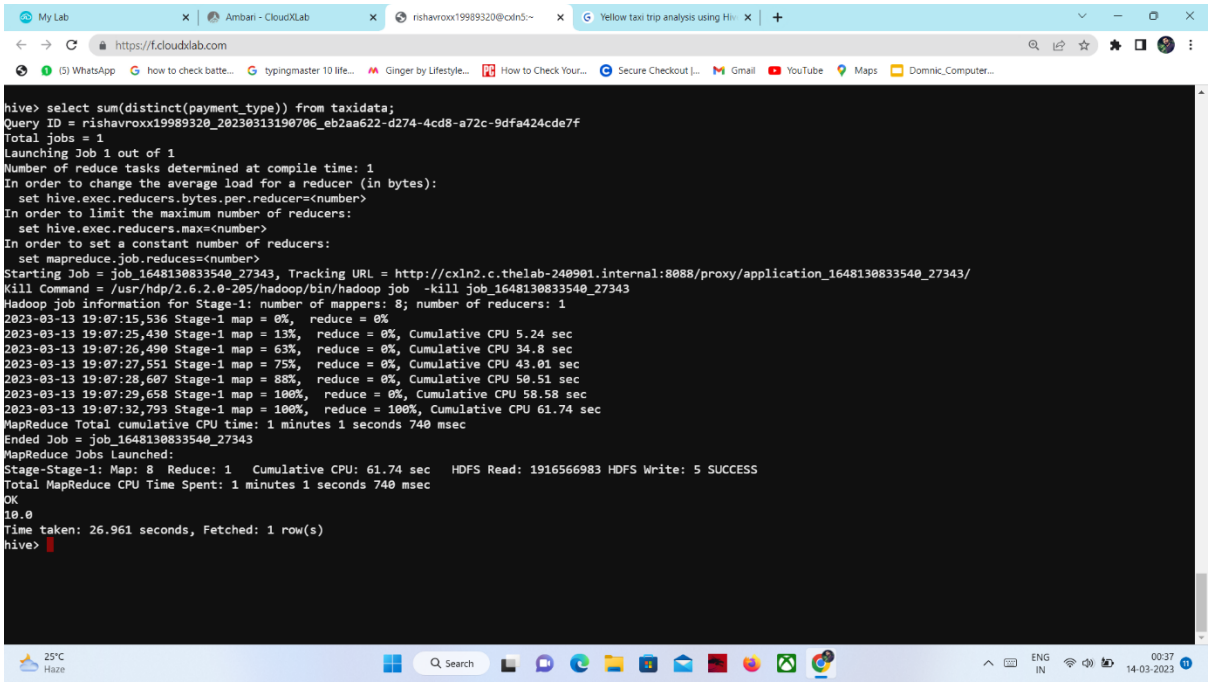
Q10) On an average how much mta_tax has been given:-



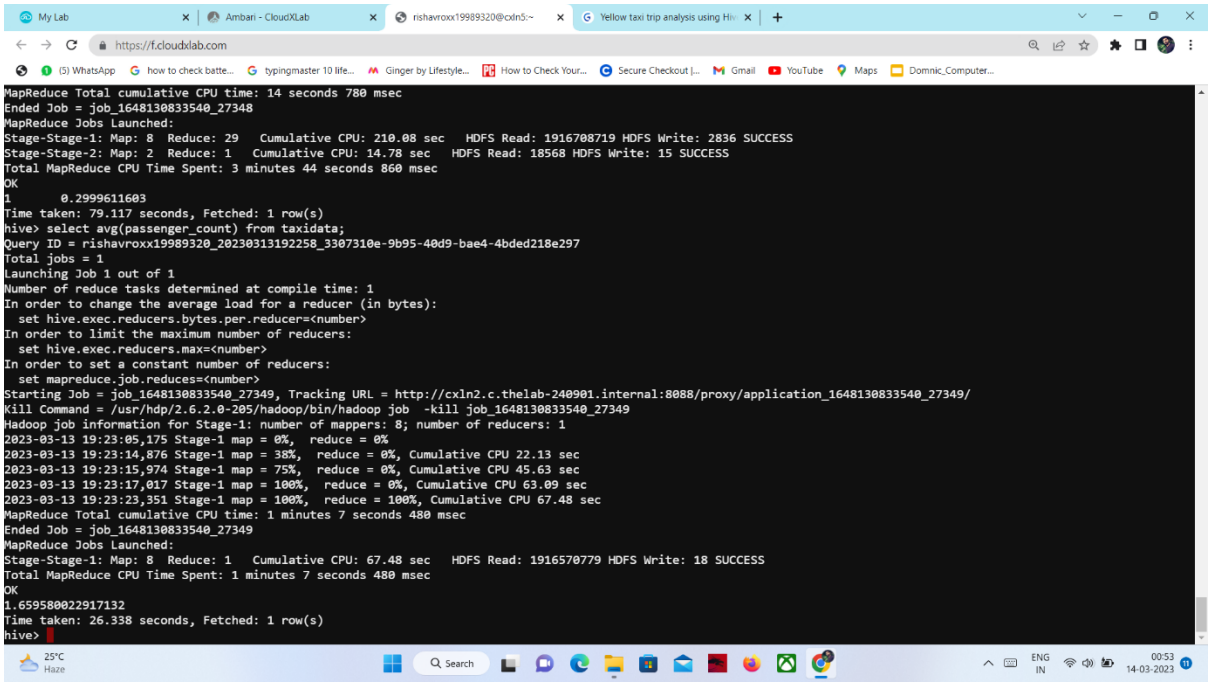
Q11) Average trip time :-



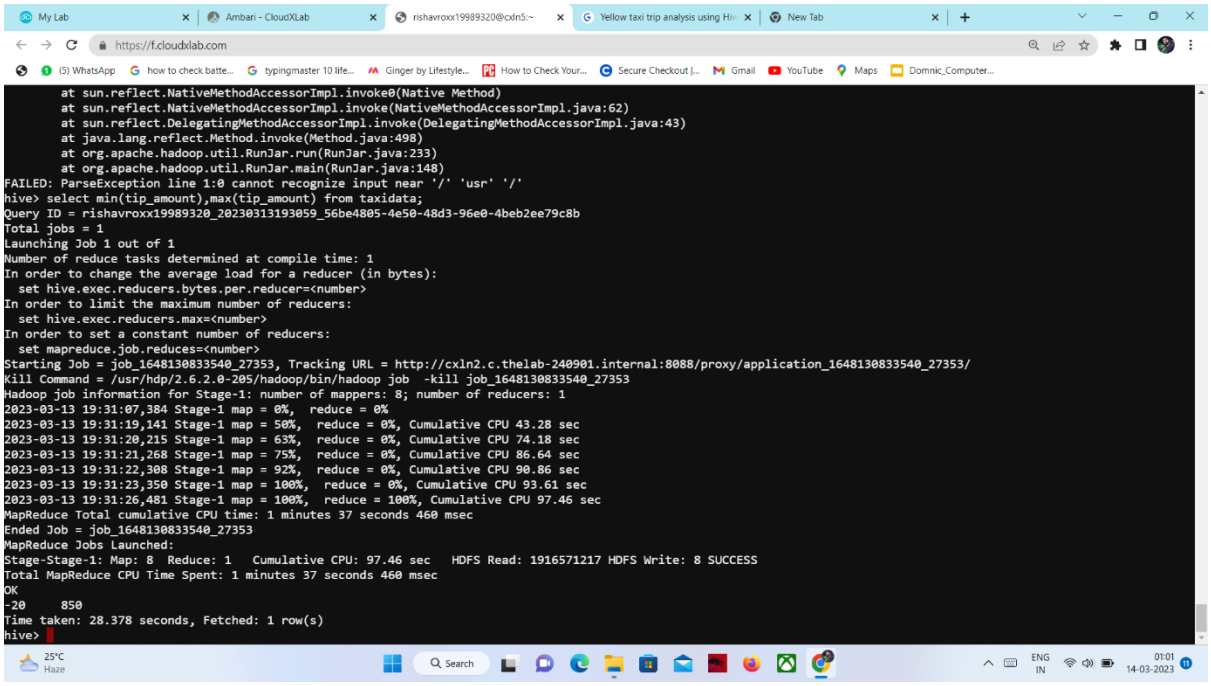
Q12) How many different payment types are used:-



Q13) Average passenger count:

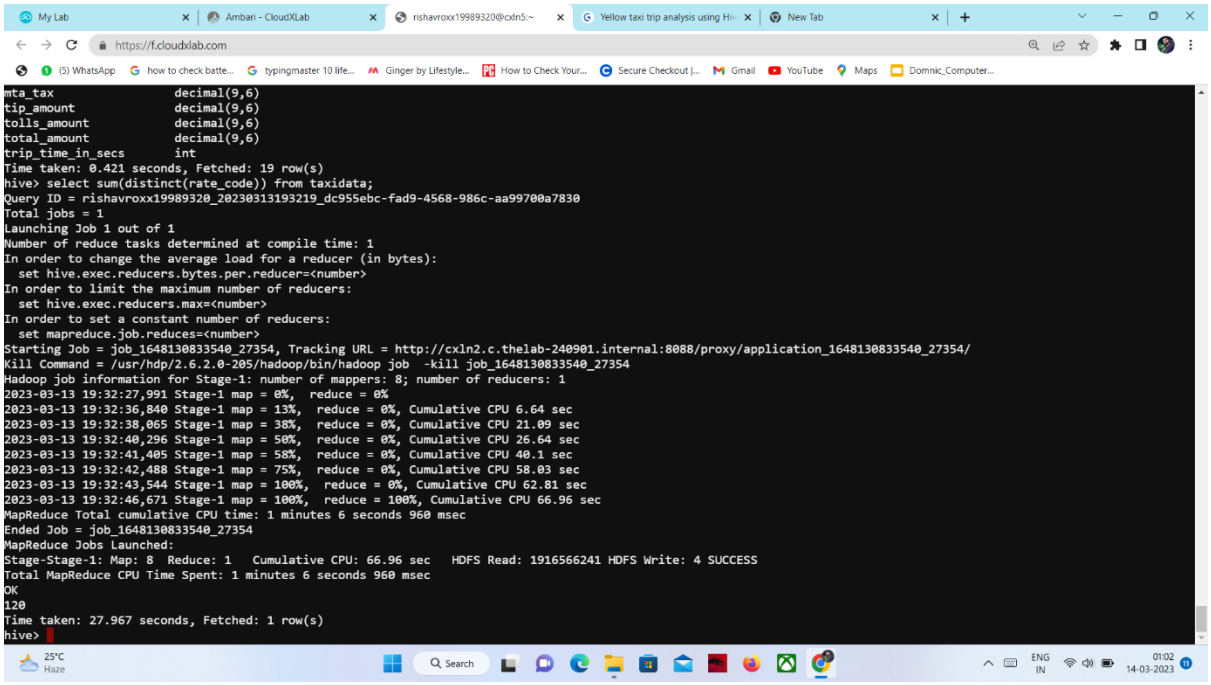


Q14) Maximum time spent by passenger:-



```
        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 = rishavroxx19989320_20230313193059_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/2.6.2.0-205/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 = 50%,  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 19:31:22,308 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 = rishavroxx19989320_20230313193219_dc955ebc-fad9-4568-986c-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/2.6.2.0-205/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 19:32:27,991 Stage-1 map = 0%,  reduce = 0%
2023-03-13 19:32:36,840 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 19:32:40,296 Stage-1 map = 50%,  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>
```