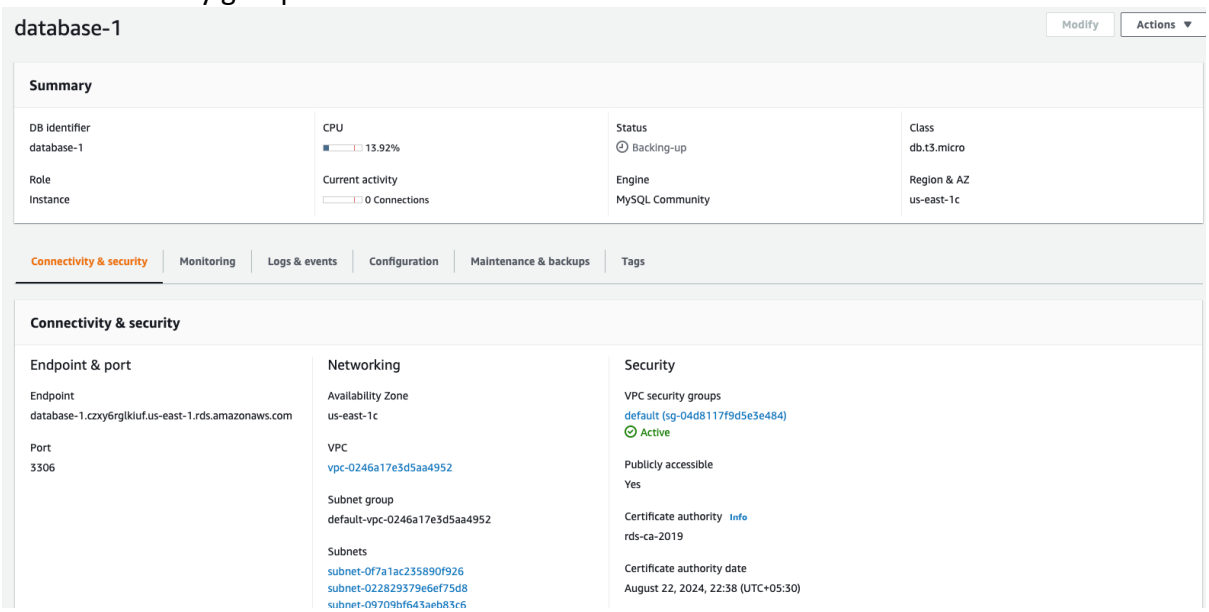
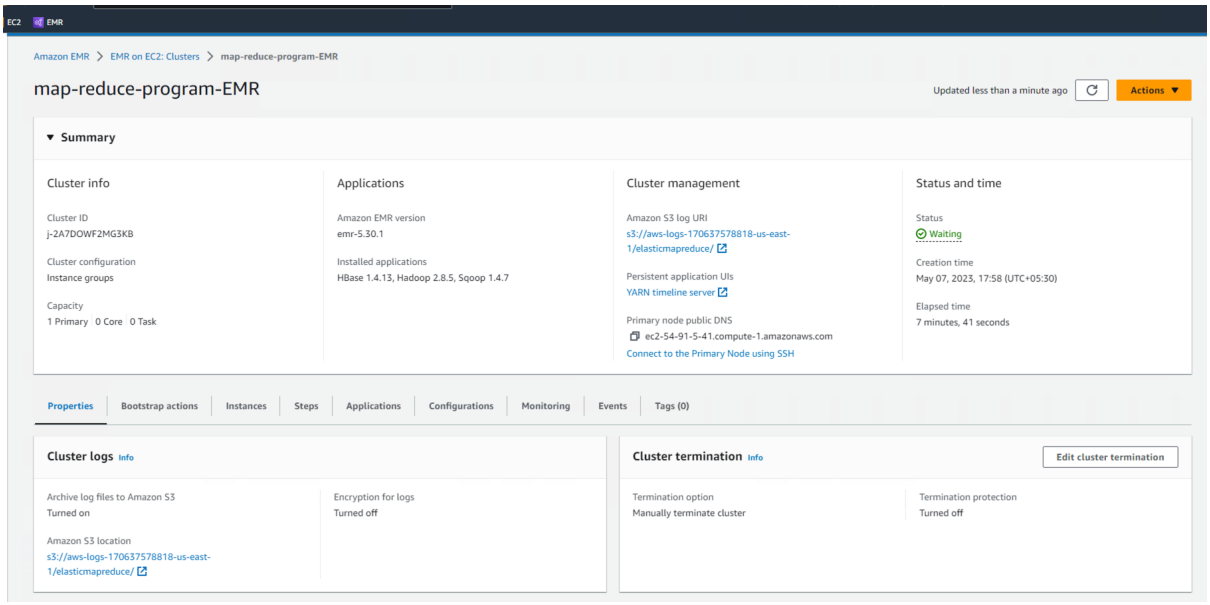
Creation of RDS instance, DB, Table creation in RDS instance and to upload the csv file into the RDS instance. And creation of EMR cluster.

Step 1: Created RDS instance with MySQL and publicly accessible instance with MySQL port allowed security group.



Step2: Created EMR cluster with the name map-reduce-program-EMR with root file system size with 40GB



Step 3: Logged in to EMR cluster and created a directory named ‘mapreduce-assignment’ under /.

Commands used:

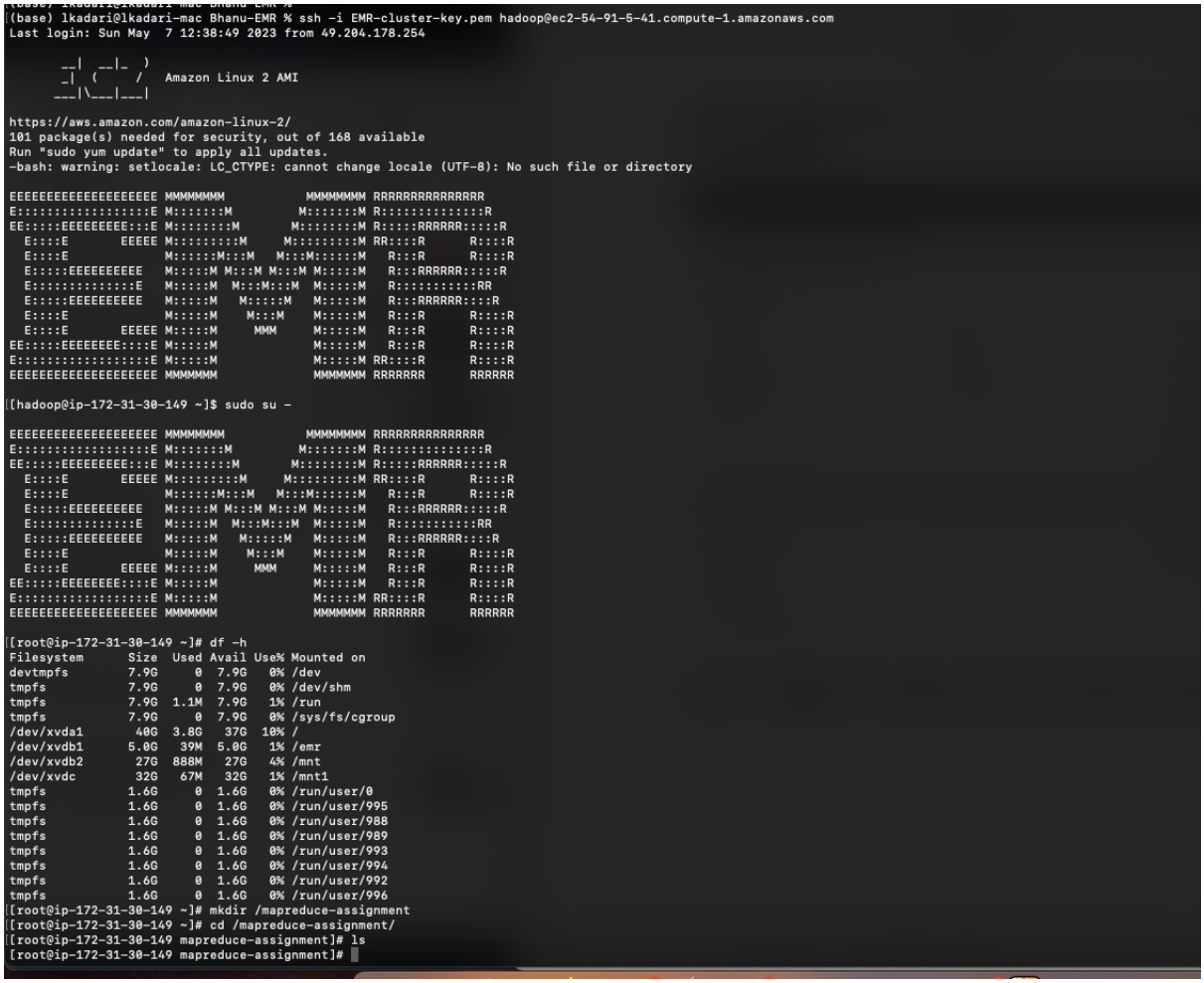
ssh -i EMR-cluster-key.pem hadoop@ec2-54-91-5-41.compute-1.amazonaws.com # to login to EMR cluster

sudo su - # To switch user to root

df -h # to check the sizes of file system to use that file system to download Datasets.

mkdir /mapreduce-assignment # to create directory under / filesystem since it is having enough space as shown in the below image.

cd /mapreduce-assignment/ # to change the directory.



Step 4: Downloading the datasets in to ‘mapreduce-assignment’ directory.

Commands to download the Datasets in to ‘/mapreduce-assignment’ directory.

wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-03.csv
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-04.csv
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-05.csv
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-06.csv

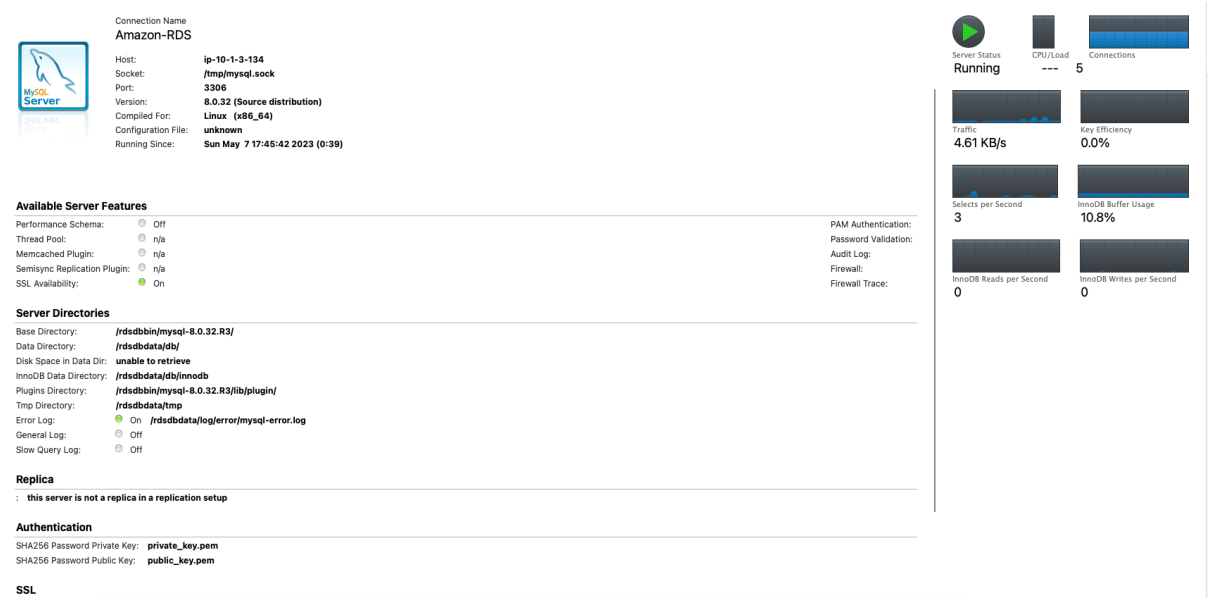
Command:
ls -lrt # command used to list all the files in directory in a sorted manner.

Step 5: Connecting RDS instance from EMR cluster and MySQLWorkbench

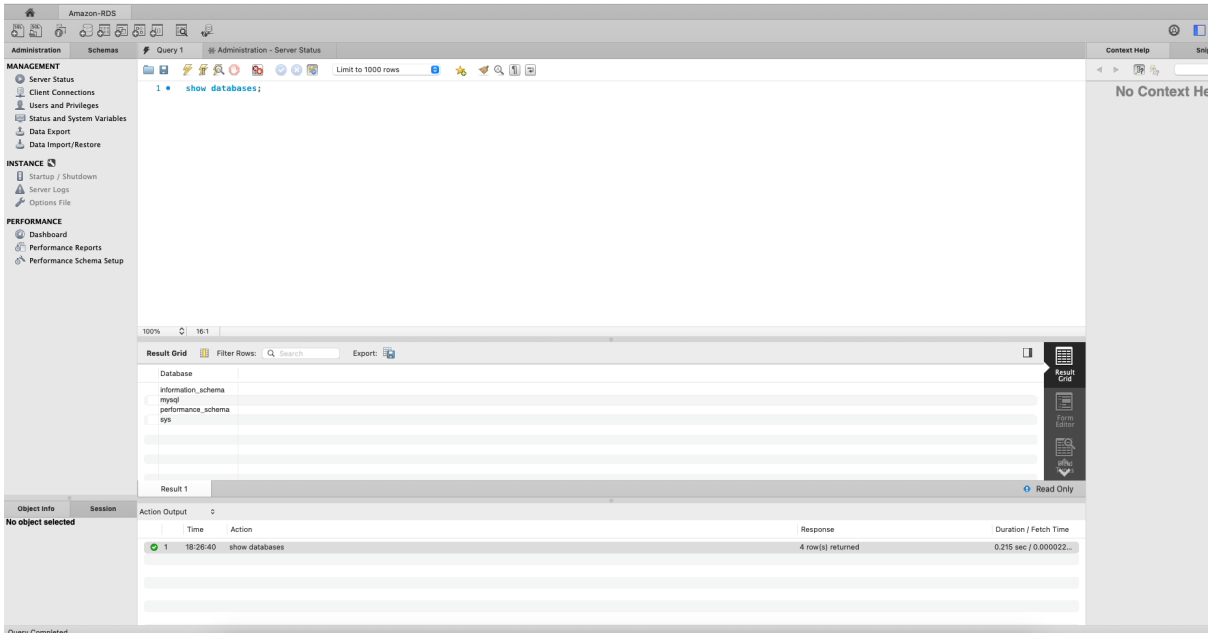
Commands:

mysql -h database-1.czxy6rglkiuf.us-east-1.rds.amazonaws.com -P 3306 -u admin -p # to connect RDS instance from EMR cluster.

show databses; # to see existing databases.


Connection Name
Amazon-RDS
Host: ip-10-1-3-134
Socket: /tmp/mysql.sock
Port: 3306
Version: 8.0.32 (Source distribution)
Compiled For: Linux (x86_64)
Configuration File: unknown
Running Since: Sun May 7 17:45:42 2023 (0:39)
Server Status
Running
CPU/Load

Connections
5
Traffic
4.61 KB/s
Key Efficiency
0.0%
Selects per Second
3
InnoDB Buffer Usage
10.8%
InnoDB Reads per Second
0
InnoDB Writes per Second
0
Available Server Features
Performance Schema: Off
Thread Pool: n/a
Memcached Plugin: n/a
Semisync Replication Plugin: n/a
SSL Availability: On
PAM Authentication:
Password Validation:
Audit Log:
Firewall:
Firewall Trace:
Server Directories
Base Directory: /rdsdbbin/mysql-8.0.32.R3/
Data Directory: /rdsdbdata/db/
Disk Space in Data Dir: unable to retrieve
InnoDB Data Directory: /rdsdbdata/db/innodb
Plugins Directory: /rdsdbbin/mysql-8.0.32.R3/lib/plugin/
Tmp Directory: /rdsdbdata/tmp
Error Log: On /rdsdbdata/log/error/mysql-error.log
General Log: Off
Slow Query Log: Off
Replica
: this server is not a replica in a replication setup
Authentication
SHA256 Password Private Key: private_key.pem
SHA256 Password Public Key: public_key.pem
SSL


Amazon-RDS
Administration
Schemas
Query 1
Administration - Server Status
MANAGEMENT
Server Status
Client Connections
Users and Privileges
Status and System Variables
Data Export
Data Import/Restore
INSTANCE
Startup / Shutdown
Server Logs
Options File
PERFORMANCE
Dashboard
Performance Reports
Performance Schema Setup
Limit to 1000 rows
1 • show databases;
Context Help
No Context He
Result Grid
Filter Rows:
Search
Export:
Database
information_schema
mysql
performance_schema
sys
Result 1
Read Only
Object Info
Session
No object selected
Action Output
Time
Action
Response
Duration / Fetch Time
1 18:26:40 show databases 4 row(s) returned 0.215 sec / 0.000022...


[root@ip-172-31-30-149 mapreduce-assignment]# mysql -h database-1.czxy6rglkiuf.us-east-1.rds.amazonaws.com -P 3306 -u admin -p
Enter password:
Welcome to the MariaDB monitor. Commands end with ; or \g.
Your MySQL connection id is 36
Server version: 8.0.32 Source distribution
Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
MySQL [(none)]> show databases;
+--------------------+
| Database |
+--------------------+
| information_schema |
| mysql |
| performance_schema |
| sys |
+--------------------+
4 rows in set (0.00 sec)
MySQL [(none)]>

Step 6: In RDS instance, creating Database named ‘trip’ and the table named ‘yello’ with all the columns as same as in the given dataset under ‘trip’ database.

Commands:

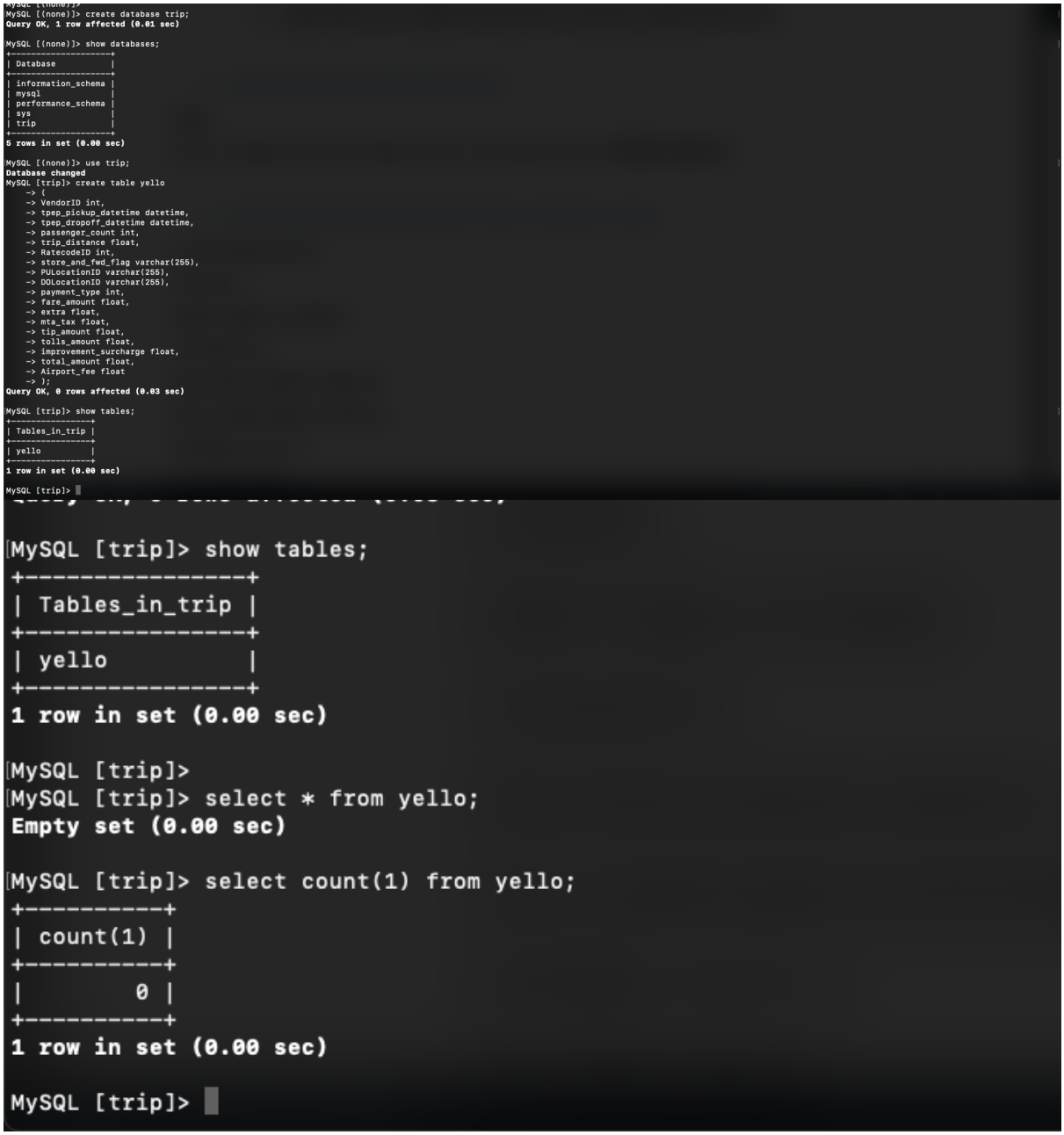
create database trip; # to create database named ‘trip’ in RDS

use trip; # To enter trip database.

Below command is used to create table named ‘yello’ under ‘trip’ with the column names as same as in the datasets.

```
create table yello
(
VendorID int,
tpep_pickup_datetime datetime,
tpep_dropoff_datetime datetime,
passenger_count int,
trip_distance float,
RatecodeID int,
store_and_fwd_flag varchar(255),
PULocationID varchar(255),
DOLocationID varchar(255),
payment_type int,
fare_amount float,
extra float,
mta_tax float,
tip_amount float,
tolls_amount float,
improvement_surcharge float,
total_amount float,
Airport_fee float
);
```

show tables; # to check the existing tables.



Step 7(a): In RDS instance, Loading ‘yellow_tripdata_2017-01.csv’ and ‘yellow_tripdata_2017-02.csv’ datasets [As per the task 1] into the existing table ‘yello’ under ‘trip’ database in MySQL.

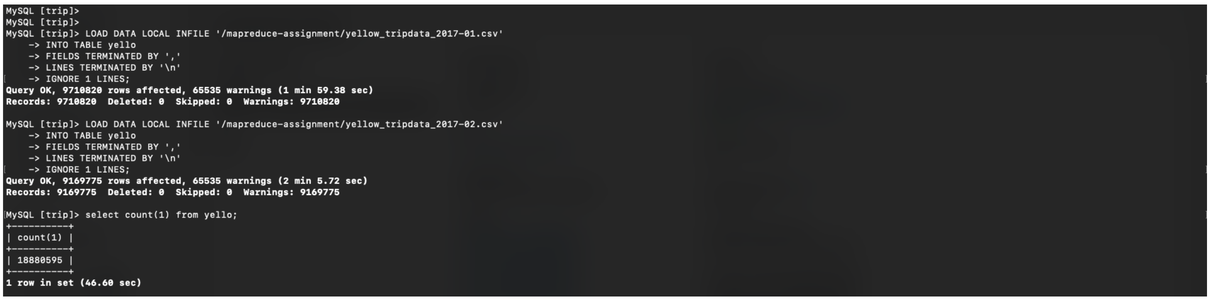
Commands:

Below commands used to upload the data from EMR cluster to RDS instance.

LOAD DATA LOCAL INFILE '/mapreduce-assignment/yellow_tripdata_2017-01.csv'
INTO TABLE yello
FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'
IGNORE 1 LINES;

LOAD DATA LOCAL INFILE '/mapreduce-assignment/yellow_tripdata_2017-02.csv'
INTO TABLE yello
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES;



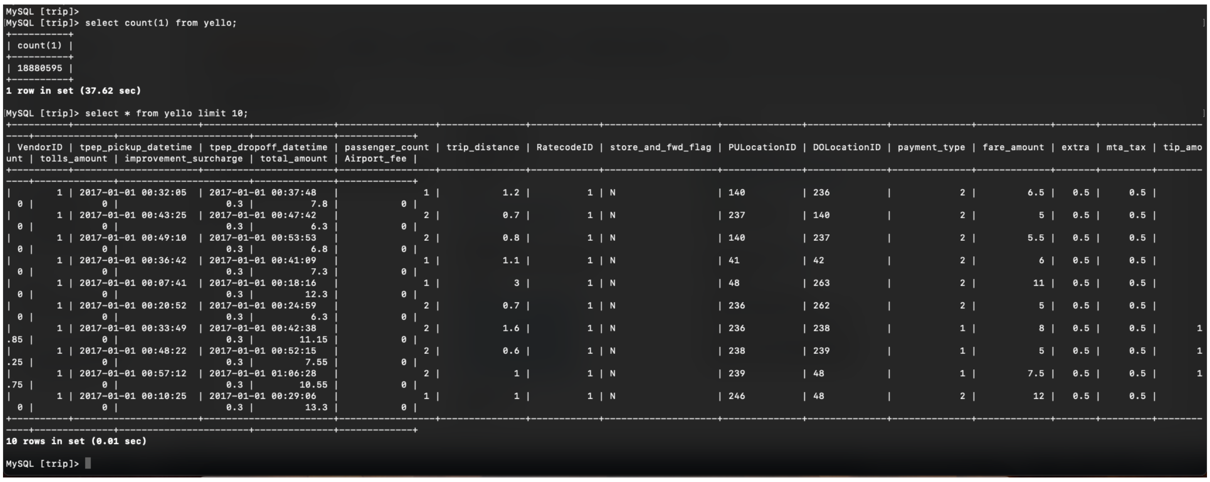
```
MySQL [trip]>
MySQL [trip]>
MySQL [trip]> LOAD DATA LOCAL INFILE '/mapreduce-assignment/yellow_tripdata_2017-01.csv'
    -> INTO TABLE yello
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 9710820 rows affected, 65535 warnings (1 min 59.38 sec)
Records: 9710820  Deleted: 0  Skipped: 0  Warnings: 9710820

MySQL [trip]> LOAD DATA LOCAL INFILE '/mapreduce-assignment/yellow_tripdata_2017-02.csv'
    -> INTO TABLE yello
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (2 min 5.72 sec)
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 9169775

MySQL [trip]> select count(1) from yello;
+----------+
| count(1) |
+----------+
| 18880595 |
+----------+
1 row in set (46.60 sec)
```

Step 7(b): Running count on 'yello' table and viewing the content of the data with limit of 10 in CLI. since it is a huge records.

Command:

select count(1) from yello; # to check the row count in 'yello' table.

select * from yello limit 10; # To display the first 10 records from the 'yello' table.



```
MySQL [trip]>
MySQL [trip]> select count(1) from yello;
+----------+
| count(1) |
+----------+
| 18880595 |
+----------+
1 row in set (37.62 sec)

MySQL [trip]> select * from yello limit 10;
| VendorID | tpep_pickup_datetime | tpep_dropoff_datetime | passenger_count | trip_distance | RatecodeID | store_and_fwd_flag | PULocationID | DOLocationID | payment_type | fare_amount | extra | mta_tax | tip_amo
unt | tolls_amount | improvement_surcharge | total_amount | Airport_fee |
|        1 | 2017-01-01 00:32:05  | 2017-01-01 00:37:48   |               1 |           1.2 |          1 | N                  | 140          | 236          |            2 |         6.5 |   0.5 |     0.5 |
  0 |            0 |                   0.3 |          7.8 |           0 |
|        1 | 2017-01-01 00:43:25  | 2017-01-01 00:47:42   |               2 |           0.7 |          1 | N                  | 237          | 140          |            2 |           5 |   0.5 |     0.5 |
  0 |            0 |                   0.3 |          6.3 |           0 |
|        1 | 2017-01-01 00:49:10  | 2017-01-01 00:53:53   |               2 |           0.8 |          1 | N                  | 140          | 237          |            2 |         5.5 |   0.5 |     0.5 |
  0 |            0 |                   0.3 |          6.8 |           0 |
|        1 | 2017-01-01 00:36:42  | 2017-01-01 00:41:09   |               1 |           1.1 |          1 | N                  | 41           | 42           |            2 |           6 |   0.5 |     0.5 |
  0 |            0 |                   0.3 |          7.3 |           0 |
|        1 | 2017-01-01 00:07:41  | 2017-01-01 00:18:16   |               1 |             3 |          1 | N                  | 48           | 263          |            2 |          11 |   0.5 |     0.5 |
  0 |            0 |                   0.3 |         12.3 |           0 |
|        1 | 2017-01-01 00:20:52  | 2017-01-01 00:24:59   |               2 |           0.7 |          1 | N                  | 236          | 262          |            2 |           5 |   0.5 |     0.5 |
  0 |            0 |                   0.3 |          6.3 |           0 |
|        1 | 2017-01-01 00:33:49  | 2017-01-01 00:42:38   |               2 |           1.6 |          1 | N                  | 236          | 238          |            1 |           8 |   0.5 |     0.5 |       1
.85 |            0 |                   0.3 |        11.15 |           0 |
|        1 | 2017-01-01 00:48:22  | 2017-01-01 00:52:15   |               2 |           0.6 |          1 | N                  | 238          | 239          |            1 |           5 |   0.5 |     0.5 |       1
.25 |            0 |                   0.3 |         7.55 |           0 |
|        1 | 2017-01-01 00:57:12  | 2017-01-01 01:06:28   |               2 |             1 |          1 | N                  | 239          | 48           |            1 |         7.5 |   0.5 |     0.5 |       1
.75 |            0 |                   0.3 |        10.55 |           0 |
|        1 | 2017-01-01 00:10:25  | 2017-01-01 00:29:06   |               1 |             1 |          1 | N                  | 246          | 48           |            2 |          12 |   0.5 |     0.5 |
  0 |            0 |                   0.3 |         13.3 |           0 |
10 rows in set (0.01 sec)

MySQL [trip]>
```

Step 7(c): Viewing the records in MySQLWorkbench.