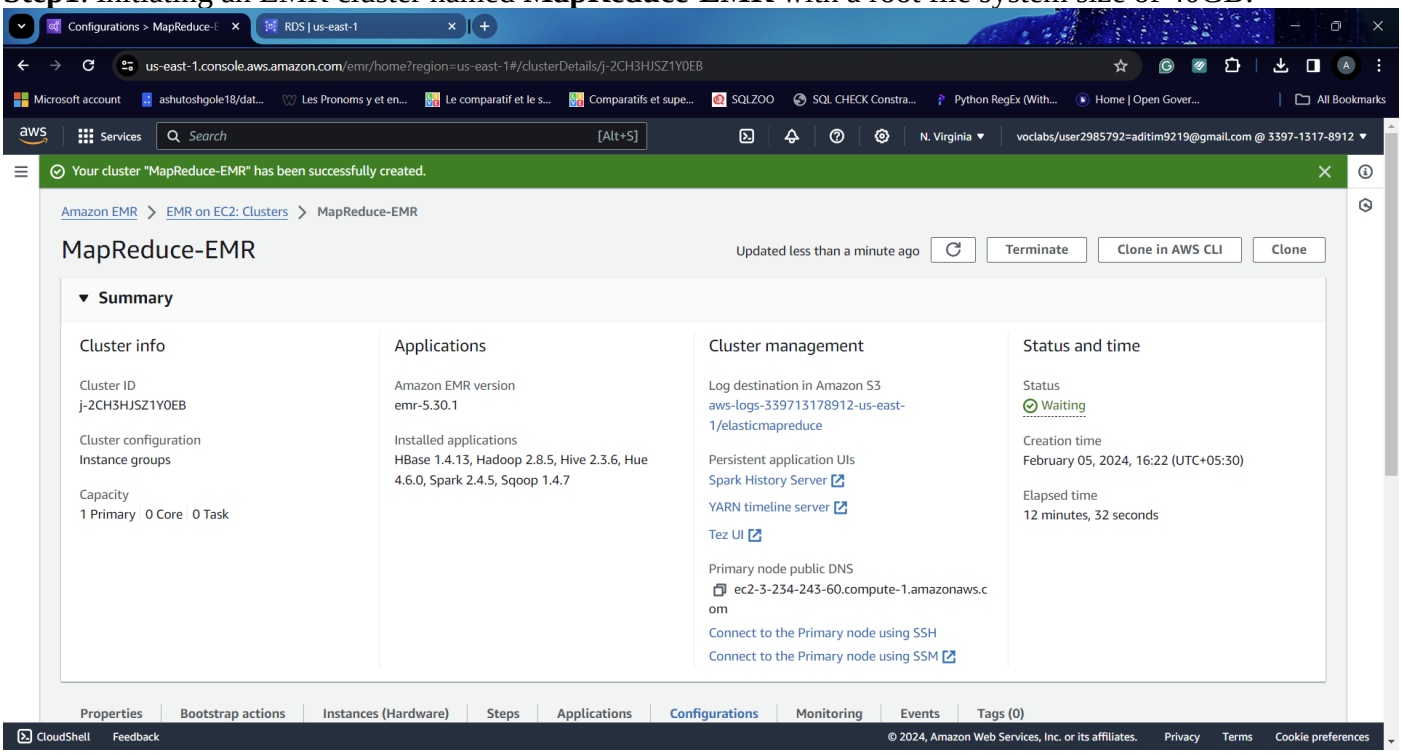
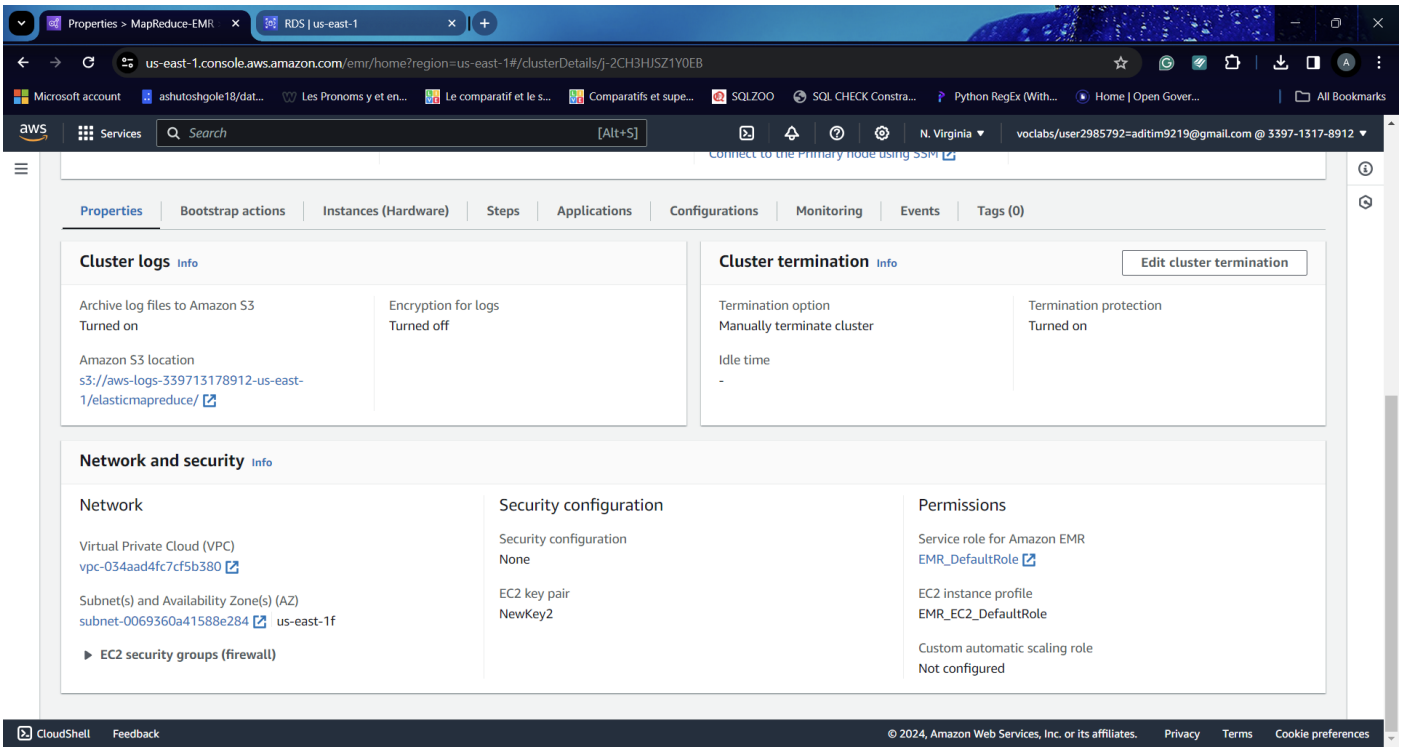
Task 1:

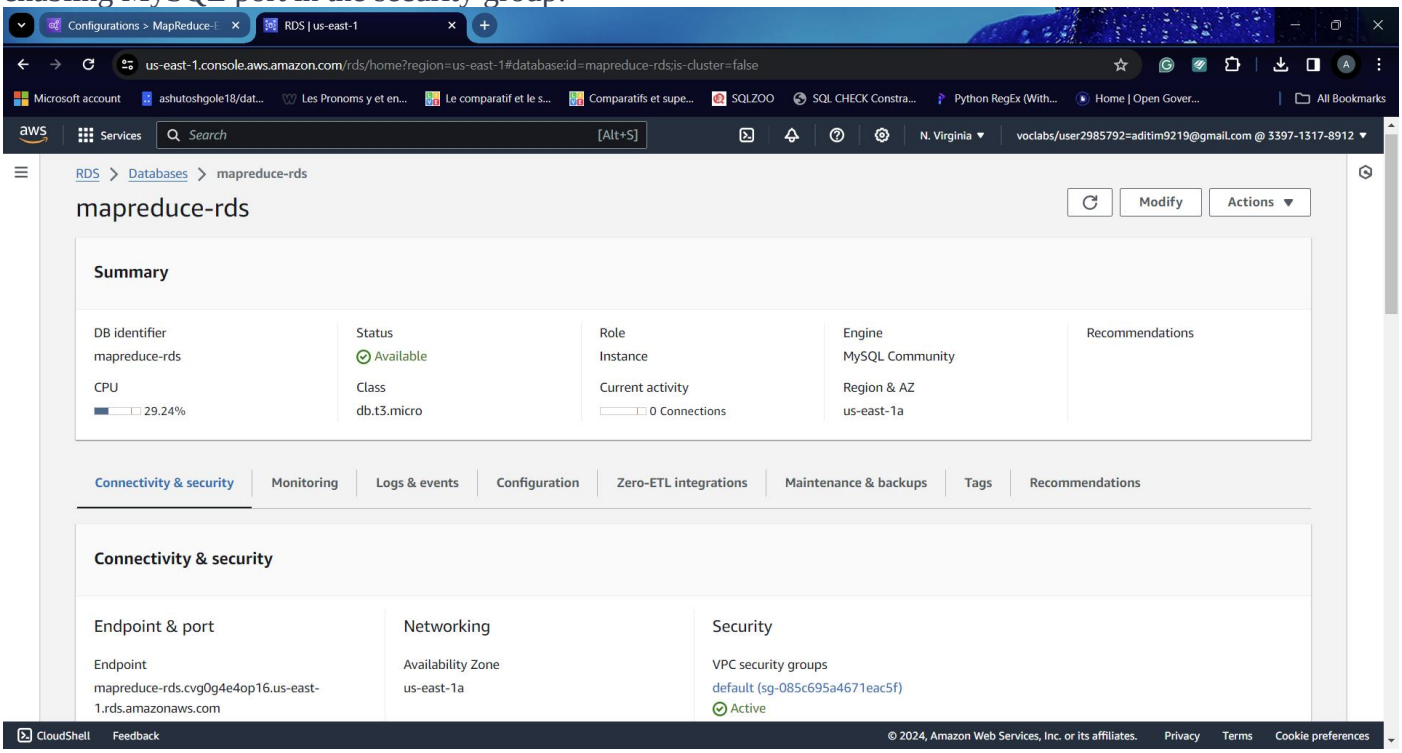
Setting up an RDS instance, establishing a database, creating tables within the RDS instance, and importing CSV files into the RDS instance. Additionally, initiating the creation of an EMR cluster.

Step1: Initiating an EMR cluster named **MapReduce-EMR** with a root file system size of 40GB:

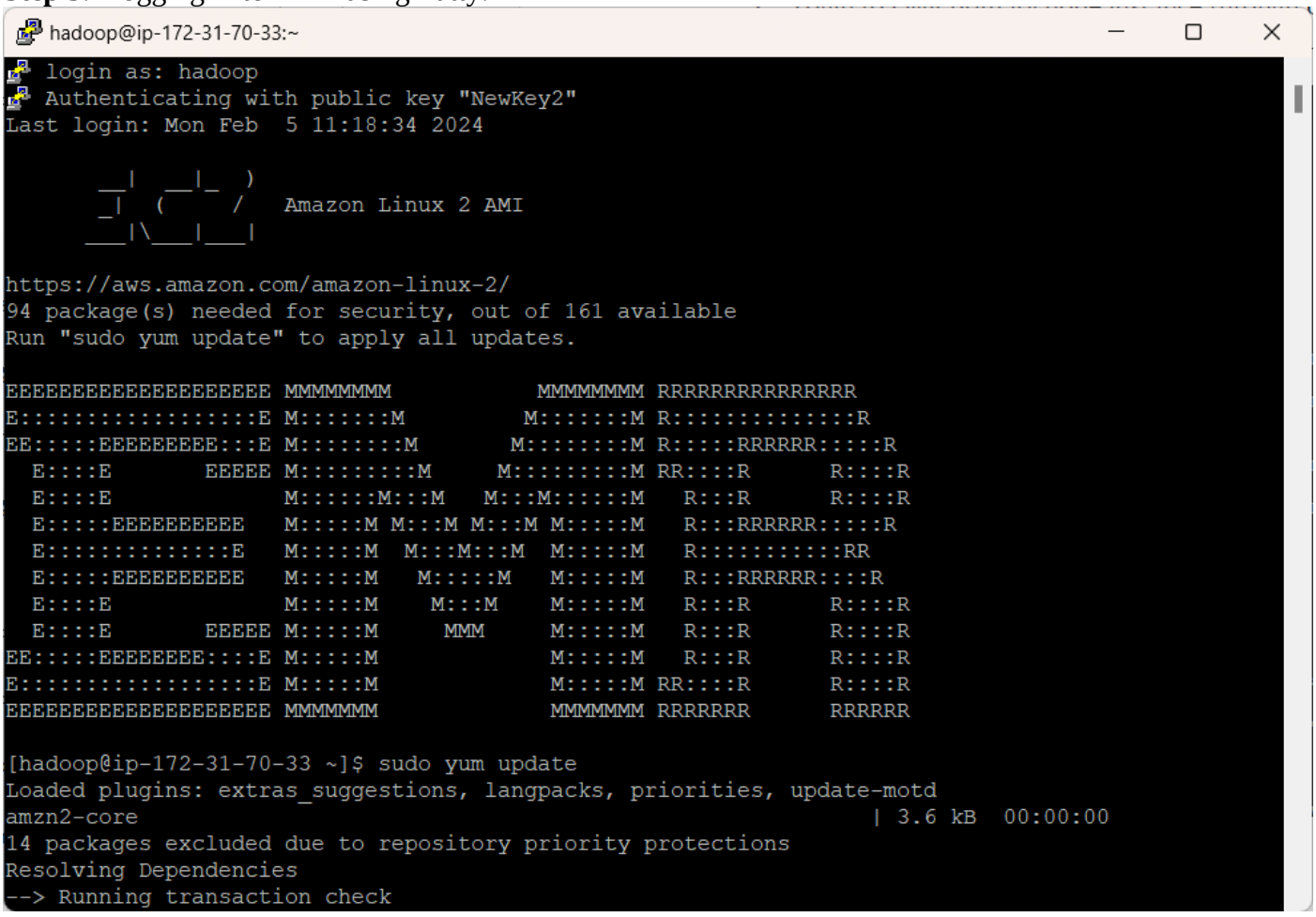




Step 2: Creating an RDS instance **mapreduce-rds** with MySQL, making it publicly accessible and enabling MySQL port in the security group:



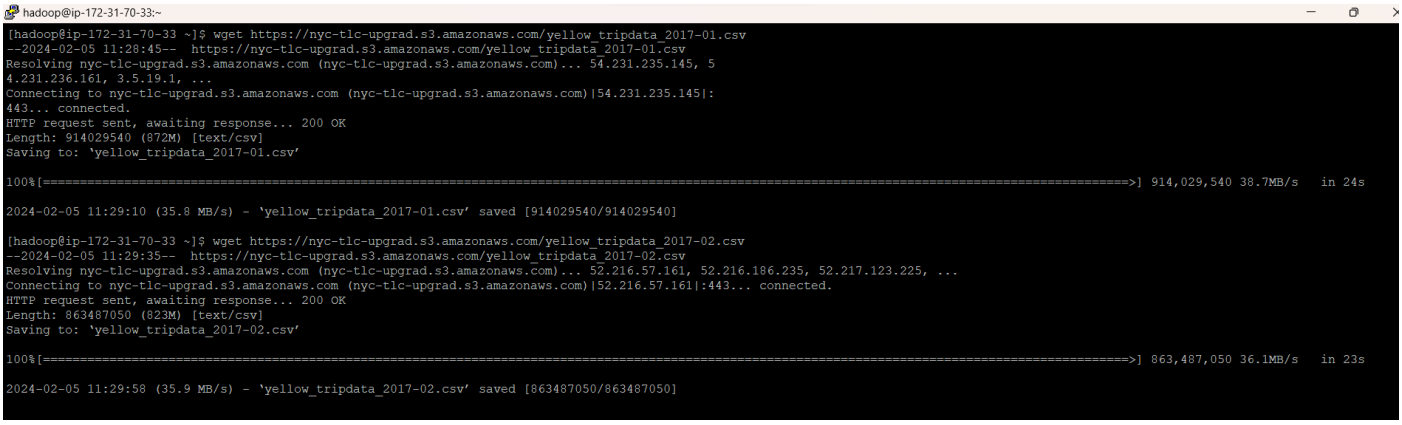
Step 3: Logging in to EMR using Putty:



Step 4: Fetching necessary CSV files from the internet and saving them locally:

Use command:

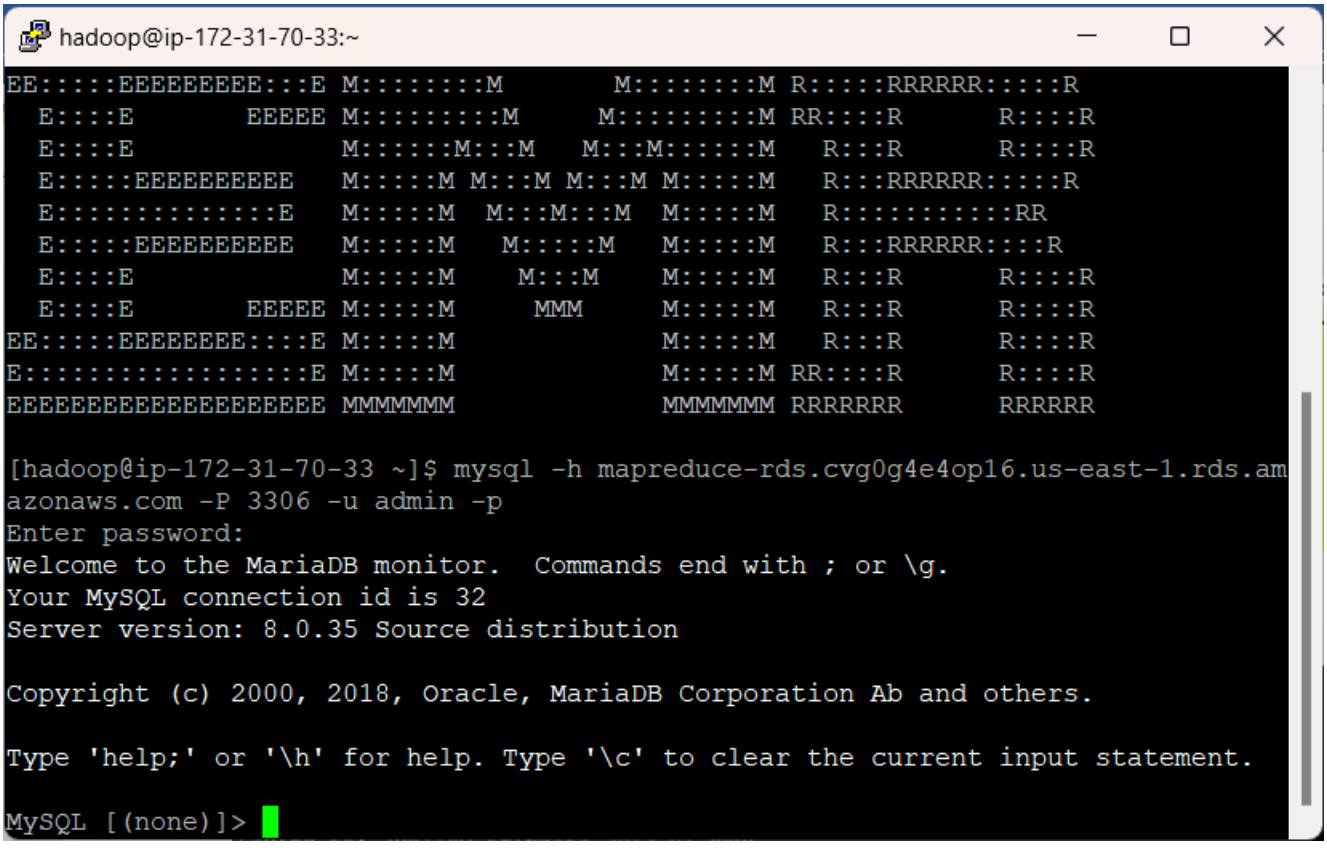
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv



Step 5: Establishing a connection between an EMR cluster and an RDS instance, along with executing MySQLWorkbench commands:

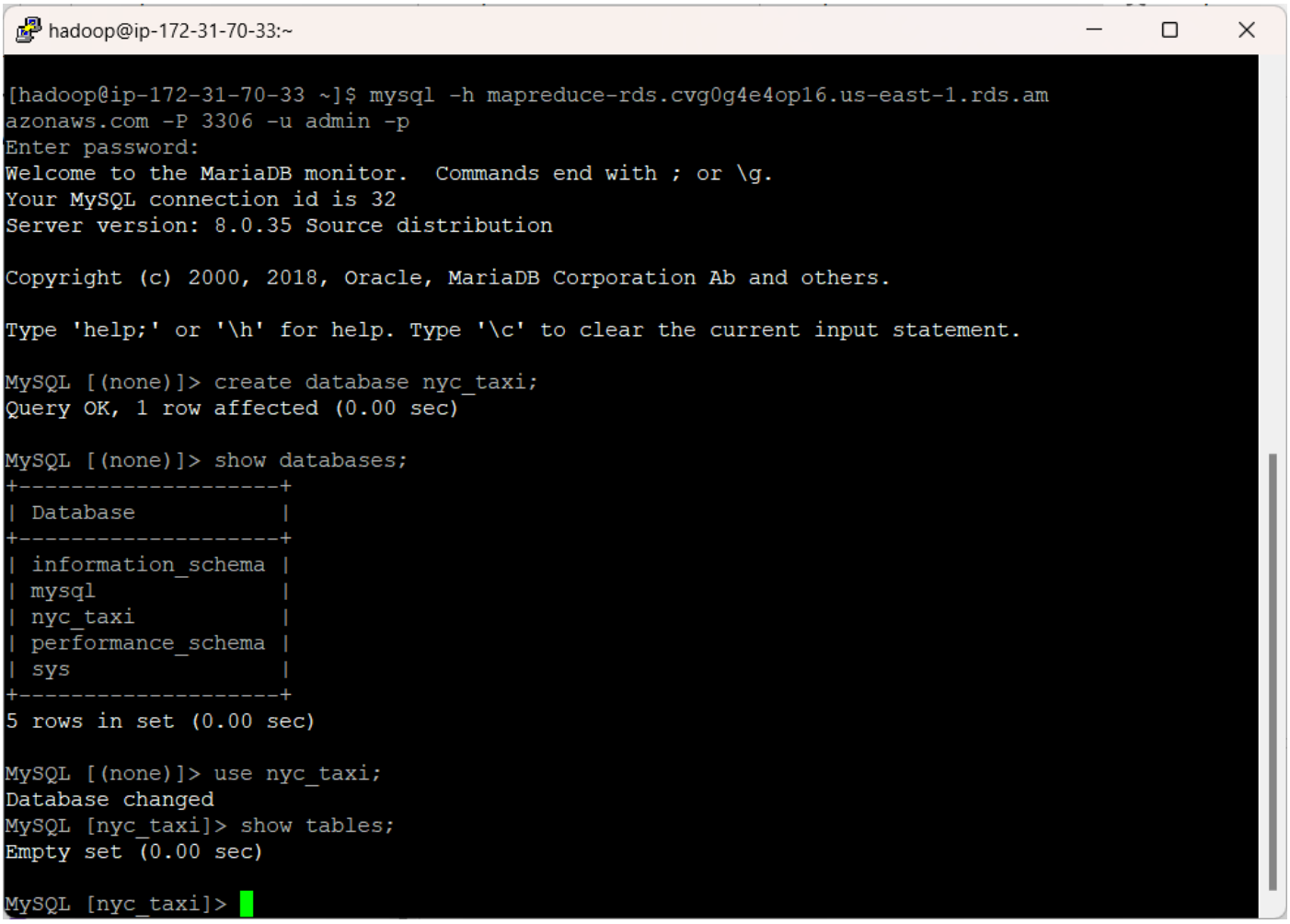
Use command:

mysql -h mapreduce-rds.cvg0g4e4op16.us-east-1.rds.amazonaws.com -P 3306 -u admin -p



Use command:

create database <database_name>; : To create new database
show databases; : To view list of databases
use <database_name>; : To go to the database
show tables; : To view list of tables in a database



To create new table in database **nyc_taxi**, use command:

```
create table NYC_TRIPS
(
VendorID INT,
tpep_pickup_datetime DATETIME,
tpep_dropoff_datetime DATETIME,
passenger_count INT,
trip_distance FLOAT(10,2),
RatecodeID INT,
store_and_fwd_flag VARCHAR(1),
PULocationID VARCHAR(50),
DOLocationID VARCHAR(50),
payment_type INT,
fare_amount FLOAT(10,2),
extra FLOAT(10,2),
mta_tax FLOAT(10,2),
tip_amount FLOAT(10,2),
tolls_amount FLOAT(10,2),
improvement_surcharge FLOAT(10,2),
total_amount FLOAT(10,2),
Airport_fee FLOAT(10,2)
);
```

```
hadoop@ip-172-31-70-33:~

MySQL [nyc_taxi]> create table NYC_TRIPS
    -> (
    -> VendorID INT,
    -> tpep_pickup_datetime DATETIME,
    -> tpep_dropoff_datetime DATETIME,
    -> passenger_count INT,
    -> trip_distance FLOAT(10,2),
    -> RatecodeID INT,
    -> store_and_fwd_flag VARCHAR(1),
    -> PULocationID VARCHAR(50),
    -> DOLocationID VARCHAR(50),
    -> payment_type INT,
    -> fare_amount FLOAT(10,2),
    -> extra FLOAT(10,2),
    -> mta_tax FLOAT(10,2),
    -> tip_amount FLOAT(10,2),
    -> tolls_amount FLOAT(10,2),
    -> improvement_surcharge FLOAT(10,2),
    -> total_amount FLOAT(10,2),
    -> Airport_fee FLOAT(10,2)
    -> );
Query OK, 0 rows affected, 9 warnings (0.02 sec)

MySQL [nyc_taxi]> show tables;
+--------------------+
| Tables_in_nyc_taxi |
+--------------------+
| NYC_TRIPS          |
+--------------------+
1 row in set (0.00 sec)

MySQL [nyc_taxi]>
```

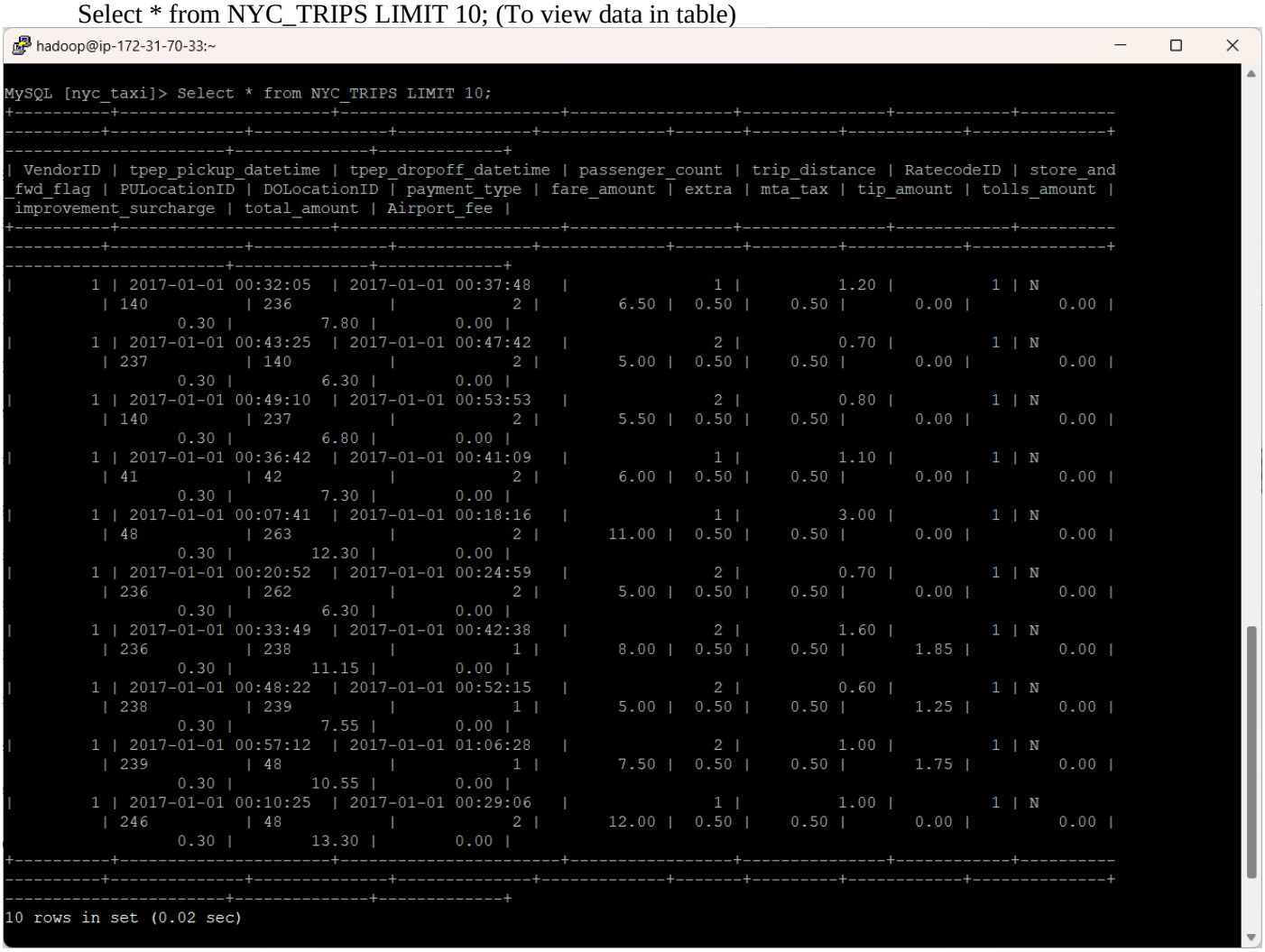
Step 6: Within the RDS instance, importing the datasets 'yellow_tripdata_2017-01.csv' and 'yellow_tripdata_2017-02.csv' into the pre-existing '**NYC_TRIPS**' table within the '**nyc_taxi**' database in MySQL:
Use commands:
LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv' INTO TABLE NYC_TRIPS FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;

LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv' INTO TABLE NYC_TRIPS FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;

```
hadoop@ip-172-31-70-33:~
    -> );
Query OK, 0 rows affected, 9 warnings (0.02 sec)

MySQL [nyc_taxi]> show tables;
+--------------------+
| Tables_in_nyc_taxi |
+--------------------+
| NYC_TRIPS          |
+--------------------+
1 row in set (0.00 sec)

MySQL [nyc_taxi]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv' INTO TABLE N
YC_TRIPS FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9710820 rows affected, 65535 warnings (2 min 8.57 sec)
Records: 9710820  Deleted: 0  Skipped: 0  Warnings: 9710820
MySQL [nyc_taxi]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv' INTO TABLE NYC_TRIPS FIELDS TE
RMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (2 min 10.11 sec)
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 9169775

MySQL [nyc_taxi]>
```

To check data in database, use command:

Select * from NYC_TRIPS LIMIT 10; (To view data in table)



Select count(*) from NYC_TRIPS; (To get count of number of rows added)