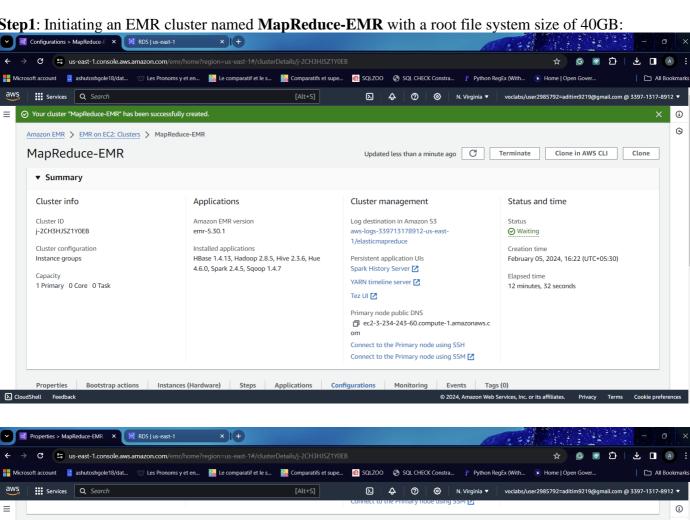
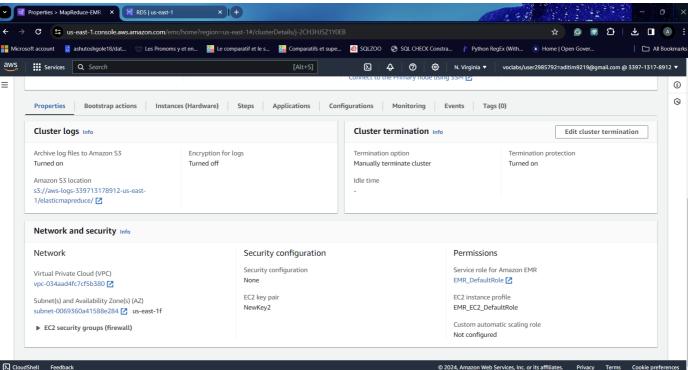
# Task 1:

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.

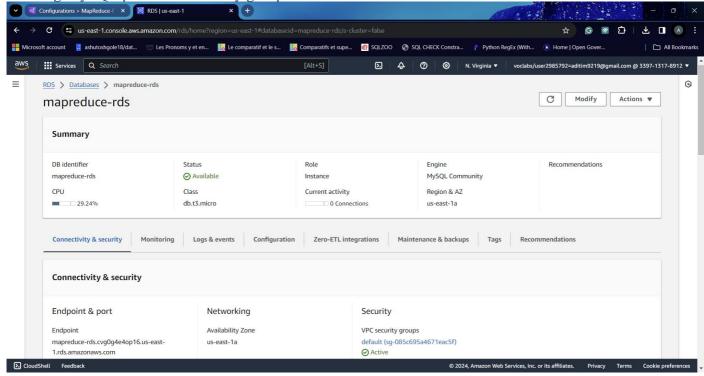
**Step1**: Initiating an EMR cluster named **MapReduce-EMR** with a root file system size of 40GB:





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:

```
hadoop@ip-172-31-70-33:~
                                                                                           X
🛂 login as: hadoop
  Authenticating with public key "NewKey2"
Last login: Mon Feb 5 11:18:34 2024
                     Amazon Linux 2 AMI
https://aws.amazon.com/amazon-linux-2/
94 package(s) needed for security, out of 161 available
Run "sudo yum update" to apply all updates.
EEEEEEEEEEEEEEEEE MMMMMMM
                                        EE:::::EEEEEEEEE:::E M:::::::M
                                      M:::::::: M R:::::RRRRRR:::::R
 E::::E
              EEEEE M:::::::M
                                     R::::R
 E::::E
                     M:::::::M::::M
                                                               R::::R
                                    M:::M:::::M
                                                   R:::R
 E::::EEEEEEEEE
                     \texttt{M} \colon \colon \colon \colon \colon \texttt{M} \ \texttt{M} \colon \colon \colon \texttt{M} \ \texttt{M} \colon \colon \colon \texttt{M} \ \texttt{M} \colon \colon \colon \texttt{M}
                                                    R:::RRRRRR::::R
                                                    R:::RRRRRR::::R
                                         M:::::M
 E::::EEEEEEEEE
                     M:::::M
                               M:::::M
 E::::E
                     M:::::M
                                M:::M
                                          M:::::M
                                                    R:::R
                                                               R::::R
  E::::E
               EEEEE M:::::M
                                 MMM
                                          M:::::M
                                                    R:::R
                                                               R::::R
EE:::::EEEEEEEE::::E M:::::M
                                          M:::::M
M:::::M RR::::R
EEEEEEEEEEEEEEEEEE MMMMMMM
                                         MMMMMMM RRRRRRR
                                                               RRRRRR
[hadoop@ip-172-31-70-33 ~]$ sudo yum update
Loaded plugins: extras suggestions, langpacks, priorities, update-motd
                                                                            00:00:00
amzn2-core
                                                                   | 3.6 kB
14 packages excluded due to repository priority protections
Resolving Dependencies
 -> Running transaction check
```

## **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

```
hadoop@ip-172-31-70-33:~
                                                                  X
EE:::::EEEEEEEEE:::E M:::::::M
                                  M:::::::M R:::::RRRRRR:::::R
             EEEEE M:::::::M
 E::::E
                                 E::::E
                                M:::M:::::M
                                              R:::R
                                                        R::::R
 E::::EEEEEEEEE
                  M:::::M M::::M M::::M
                                              R:::RRRRRR::::R
                  M:::::M M:::M:::M M:::::M
                                              R::::::::RR
                   M:::::M
 E::::EEEEEEEEE
                            M:::::M
                                     M:::::M
                                              R:::RRRRRR::::R
 E::::E
                   M:::::M
                             M:::M
                                     M:::::M
                                              R:::R
                                                        R::::R
         EEEEE M:::::M
 E::::E
                             MMM
                                     M:::::M
                                              R:::R
                                                        R::::R
EE:::::EEEEEEEE::::E M:::::M
                                     M:::::M
                                              R:::R
                                                        R::::R
M:::::M RR::::R
                                                        R::::R
EEEEEEEEEEEEEEEEE MMMMMMM
                                     MMMMMM RRRRRRR
                                                        RRRRRR
[hadoop@ip-172-31-70-33 ~]$ mysql -h mapreduce-rds.cvg0g4e4op16.us-east-1.rds.am
azonaws.com -P 3306 -u admin -p
Enter password:
Welcome to the MariaDB monitor. Commands end with ; or \g.
Your MySQL connection id is 32
Server version: 8.0.35 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)]>
```

#### 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

```
hadoop@ip-172-31-70-33:~
                                                                                          X
[hadoop@ip-172-31-70-33 ~]$ mysql -h mapreduce-rds.cvg0g4e4op16.us-east-1.rds.am
azonaws.com -P 3306 -u admin -p
Enter password:
Welcome to the MariaDB monitor. Commands end with ; or \g.
Your MySQL connection id is 32
Server version: 8.0.35 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)]> create database nyc taxi;
Query OK, 1 row affected (0.00 sec)
MySQL [(none)]> show databases;
 Database
 information schema
 mysql
 nyc_taxi
 performance schema
 rows in set (0.00 sec)
MySQL [(none)]> use nyc taxi;
Database changed
MySQL [nyc_taxi]> show tables;
Empty set (0.00 \text{ sec})
MySQL [nyc taxi]>
```

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:~
                                                                                           X
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
 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;

## To check data in database, use command:

Select \* from NYC\_TRIPS LIMIT 10; (To view data in table)

```
A hadoop@ip-172-31-70-33:~
                                                                                                                          MySQL [nyc_taxi]> Select * from NYC_TRIPS 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_amount | tolls_amount |
improvement_surcharge | total_amount | Airport_fee |
                                                                                  0.50
                                                                                                0.00 |
                                                                                                               0.00 |
                                                                                                0.00
                                                                                                               0.00 |
         | 237
                                                                        0.50
                                                    2 |
                                                               11.00 | 0.50 |
                                                                                                               0.00 |
                               12.30
                                                                        2 | 0.50 |
                                                                                                0.00 |
                                                                                                               0.00 |
                                                    2 1
                                                                                                               0.00 |
                                                                5.00 |
                                                                                  0.50 |
                                                                                                               0.00 |
         238
                                                                        0.50
                                                               12.00 |
                                                                                  0.50
                                                                                                               0.00
                                                                                                0.00 |
         246
                        48
 rows in set (0.02 sec)
```

## Select count(\*) from NYC\_TRIPS; (To get count of number of rows added)