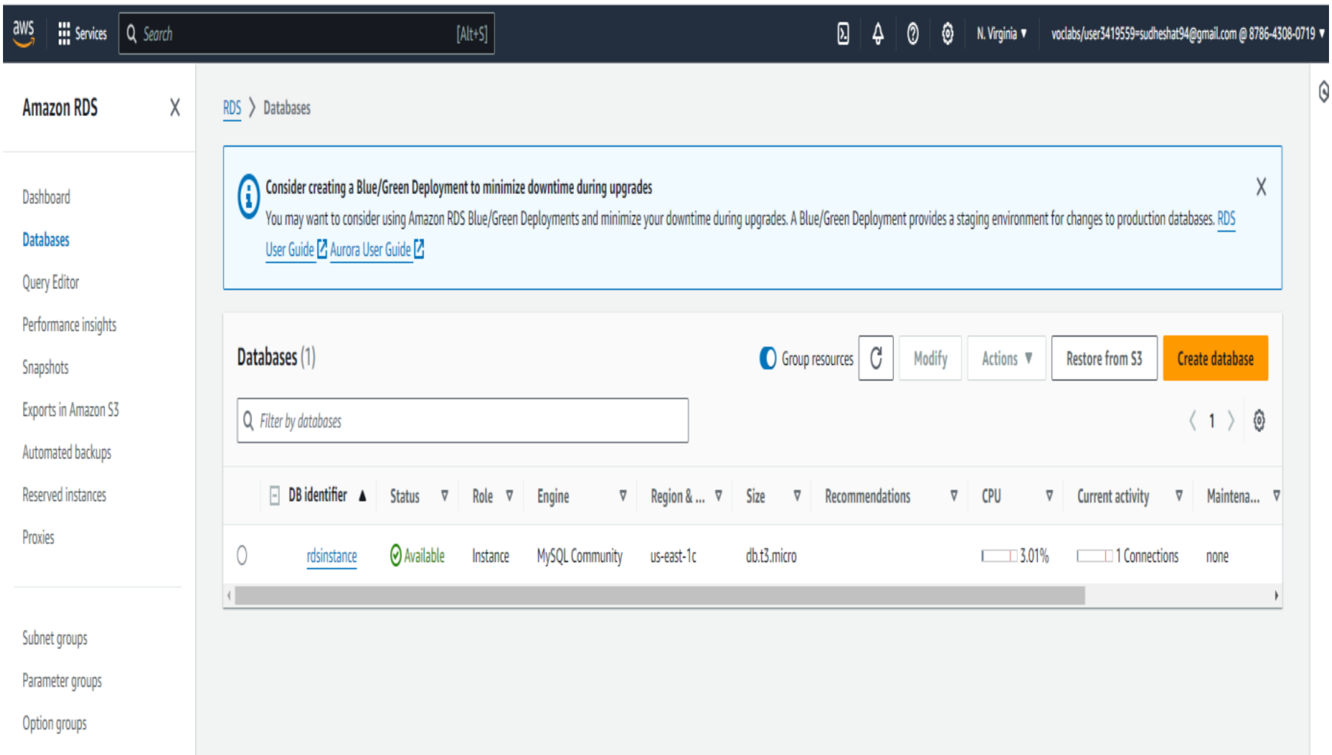
MapReduce Programing Assignment

Taks 1:

Go to AWS console and search for RDS and the click on create cluster. Select the options as per our requirements and click on create. The below RDS cluster with Database DemoDB is created.



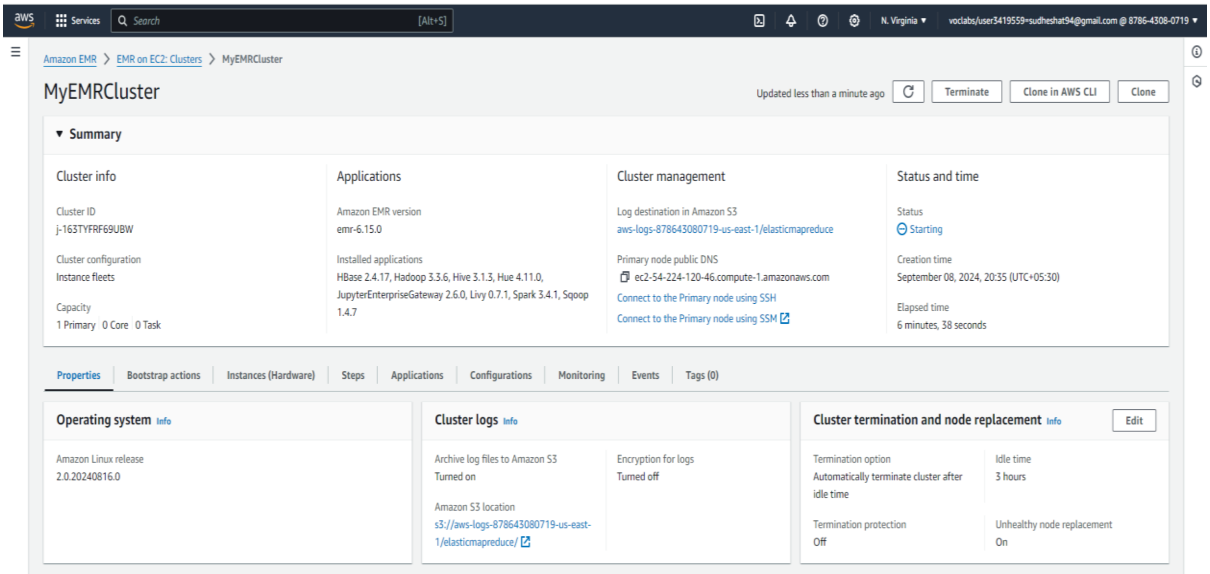
Consider creating a Blue/Green Deployment to minimize downtime during upgrades

You may want to consider using Amazon RDS Blue/Green Deployments and minimize your downtime during upgrades. A Blue/Green Deployment provides a staging environment for changes to production databases. RDS User Guide Aurora User Guide

Databases (1)

	DB identifier ▲	Status	Role	Engine	Region & ...	Size	Recommendations	CPU	Current activity	Maintena...
○	rdsinstance	✓ Available	Instance	MySQL Community	us-east-1c	db.t3.micro		3.01%	1 Connections	none

Then went back to Console and searched for EMR. Created EMR cluster as per requirements. Below EMR cluster was created.



MyEMRCluster

▾ Summary

Cluster info

Cluster ID
j-163TYFRF69UBW

Cluster configuration
Instance fleets

Capacity
1 Primary | 0 Core | 0 Task

Applications

Amazon EMR version
emr-6.15.0

Installed applications
HBase 2.4.17, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, Livy 0.7.1, Spark 3.4.1, Sqoop 1.4.7

Cluster management

Log destination in Amazon S3
aws-logs-878643080719-us-east-1/elasticmapreduce

Primary node public DNS
ec2-54-224-120-46.compute-1.amazonaws.com
Connect to the Primary node using SSH
Connect to the Primary node using SSM

Status and time

Status
Starting

Creation time
September 08, 2024, 20:35 (UTC+05:30)

Elapsed time
6 minutes, 38 seconds

Operating system Info

Amazon Linux release
2.0.20240816.0

Cluster logs Info

Archive log files to Amazon S3
Turned on

Encryption for logs
Turned off

Amazon S3 location
s3://aws-logs-878643080719-us-east-1/elasticmapreduce/

Cluster termination and node replacement Info

Termination option
Automatically terminate cluster after idle time

Idle time
3 hours

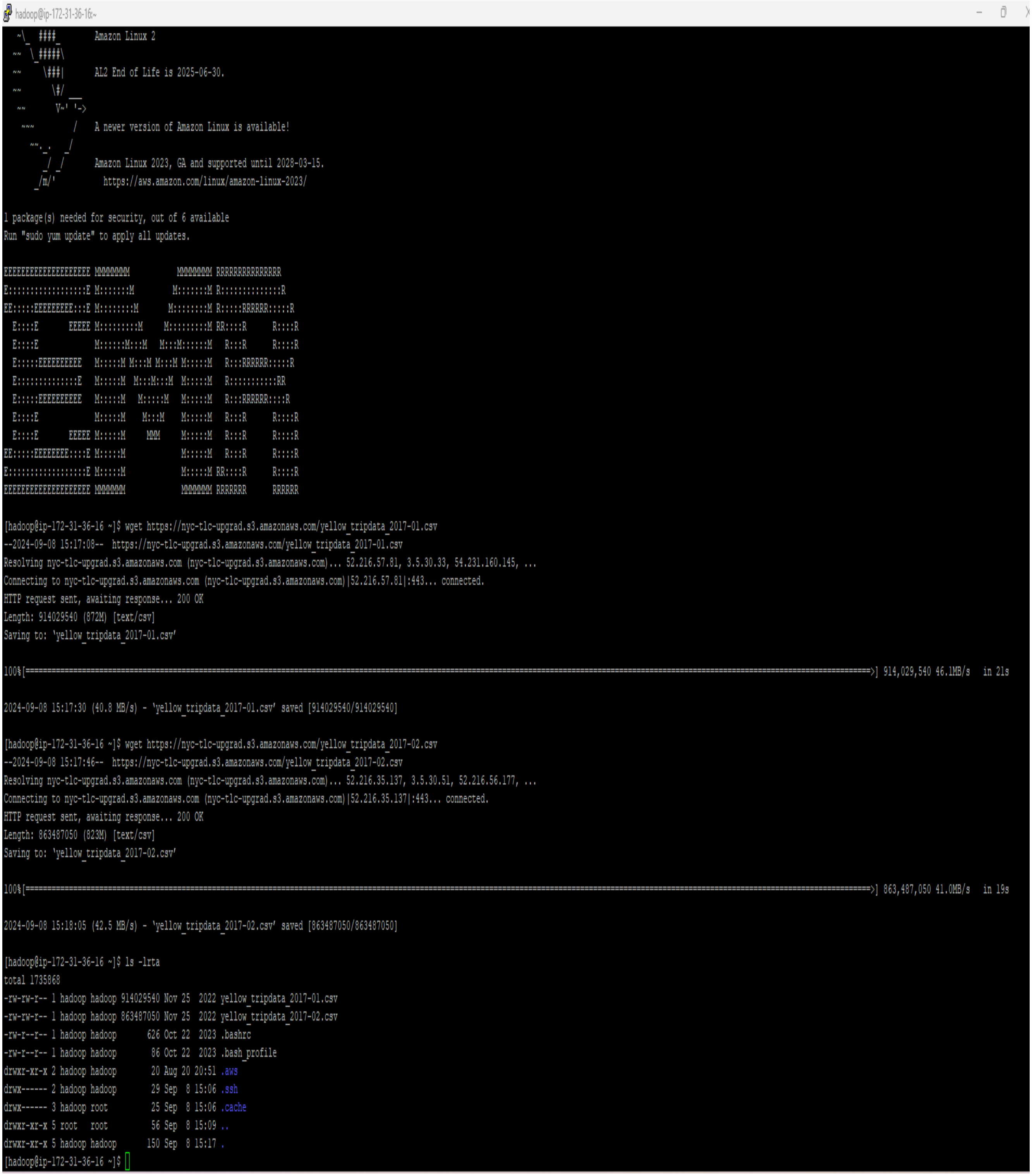
Termination protection
Off

Unhealthy node replacement
On

Open the EMR cluster using putty from local machine. And download below 2 files using wget commands.

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



```
  ~\_  ####_        Amazon Linux 2
 ~~  \_#####\
 ~~     \###|       AL2 End of Life is 2025-06-30.
 ~~       \#/ ___
  ~~       V~' '->
   ~~~         /    A newer version of Amazon Linux is available!
     ~~._.   _/
        _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
      _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

1 package(s) needed for security, out of 6 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  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   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-36-16 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
--2024-09-08 15:17:08--  https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 52.216.57.81, 3.5.30.33, 54.231.160.145, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|52.216.57.81|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 914029540 (872M) [text/csv]
Saving to: 'yellow_tripdata_2017-01.csv'

100%[=====================================================================================================================================================================================================>] 914,029,540 46.1MB/s   in 21s

2024-09-08 15:17:30 (40.8 MB/s) - 'yellow_tripdata_2017-01.csv' saved [914029540/914029540]

[hadoop@ip-172-31-36-16 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
--2024-09-08 15:17:46--  https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 52.216.35.137, 3.5.30.51, 52.216.56.177, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|52.216.35.137|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 863487050 (823M) [text/csv]
Saving to: 'yellow_tripdata_2017-02.csv'

100%[=====================================================================================================================================================================================================>] 863,487,050 41.0MB/s   in 19s

2024-09-08 15:18:05 (42.5 MB/s) - 'yellow_tripdata_2017-02.csv' saved [863487050/863487050]

[hadoop@ip-172-31-36-16 ~]$ ls -lrta
total 1735868
-rw-rw-r-- 1 hadoop hadoop 914029540 Nov 25  2022 yellow_tripdata_2017-01.csv
-rw-rw-r-- 1 hadoop hadoop 863487050 Nov 25  2022 yellow_tripdata_2017-02.csv
-rw-r--r-- 1 hadoop hadoop       626 Oct 22  2023 .bashrc
-rw-r--r-- 1 hadoop hadoop        86 Oct 22  2023 .bash_profile
drwxr-xr-x 2 hadoop hadoop        20 Aug 20 20:51 .aws
drwx------ 2 hadoop hadoop        29 Sep  8 15:06 .ssh
drwx------ 3 hadoop root          25 Sep  8 15:06 .cache
drwxr-xr-x 5 root   root          56 Sep  8 15:09 ..
drwxr-xr-x 5 hadoop hadoop       150 Sep  8 15:17 .
[hadoop@ip-172-31-36-16 ~]$ 
```

Connect to RDS instance using below command.

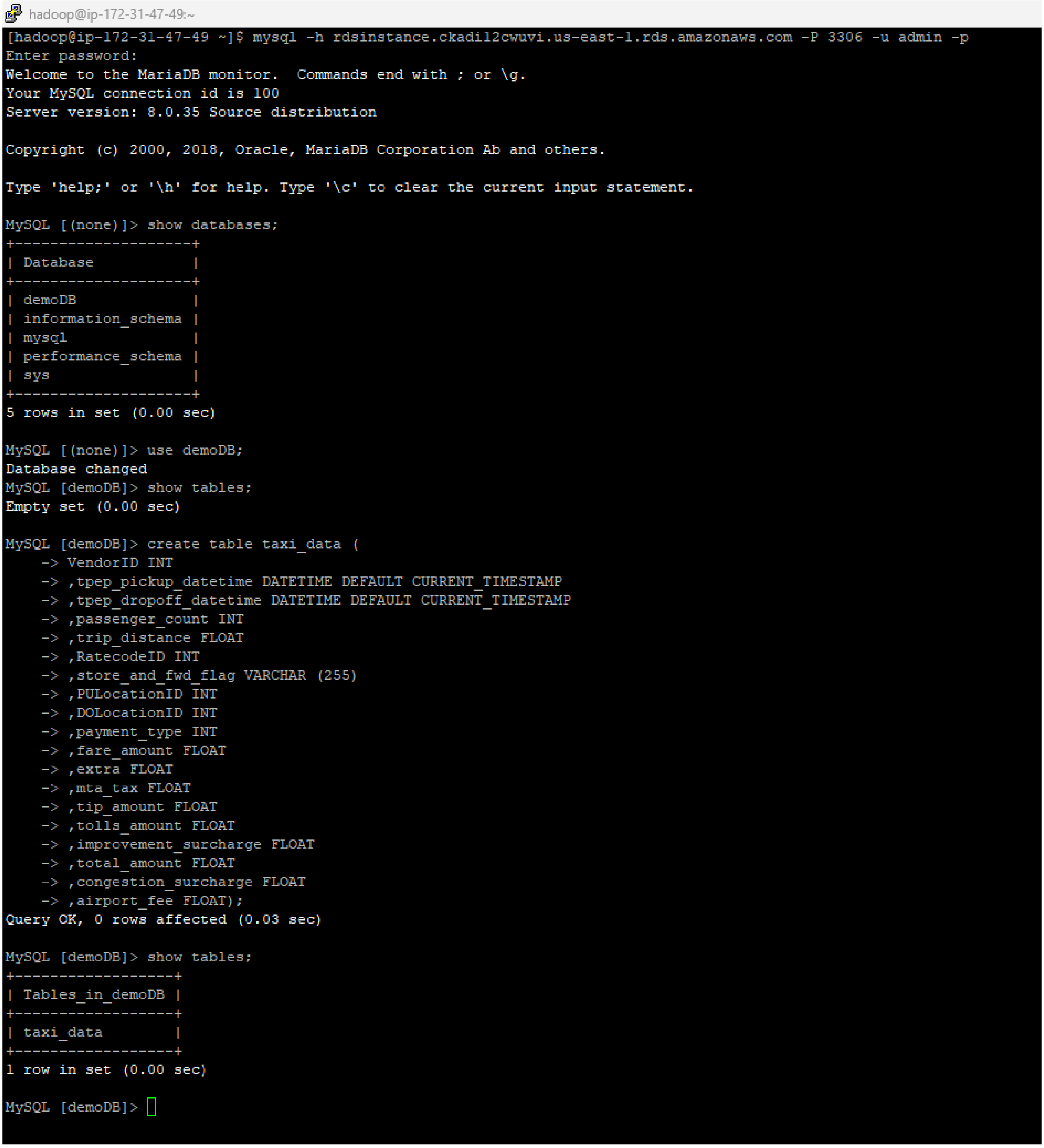
```
mysql -h rdsinstance.ckadi12cwuvi.us-east-1.rds.amazonaws.com -P 3306 -u admin -p
```

And give the password.

Use demoDB database and run below create table query to create the table.

```
create table taxi_data (
VendorID INT
,tpep_pickup_datetime DATETIME DEFAULT CURRENT_TIMESTAMP
,tpep_dropoff_datetime DATETIME DEFAULT CURRENT_TIMESTAMP
,passenger_count INT
,trip_distance FLOAT
,RatecodeID INT
,store_and_fwd_flag VARCHAR (255)
,PULocationID INT
,DOLocationID INT
,payment_type INT
,fare_amount FLOAT
,extra FLOAT
,mta_tax FLOAT
,tip_amount FLOAT
,tolls_amount FLOAT
,improvement_surcharge FLOAT
,total_amount FLOAT
,congestion_surcharge FLOAT
,airport_fee FLOAT);
```

Taxi_data table is created.



```
[hadoop@ip-172-31-47-49 ~]$ mysql -h rdsinstance.ckadil2cwuvi.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 100
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)]> show databases;
+--------------------+
| Database           |
+--------------------+
| demoDB             |
| information_schema |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
5 rows in set (0.00 sec)

MySQL [(none)]> use demoDB;
Database changed
MySQL [demoDB]> show tables;
Empty set (0.00 sec)

MySQL [demoDB]> create table taxi_data (
    -> VendorID INT
    -> ,tpep_pickup_datetime DATETIME DEFAULT CURRENT_TIMESTAMP
    -> ,tpep_dropoff_datetime DATETIME DEFAULT CURRENT_TIMESTAMP
    -> ,passenger_count INT
    -> ,trip_distance FLOAT
    -> ,RatecodeID INT
    -> ,store_and_fwd_flag VARCHAR (255)
    -> ,PULocationID INT
    -> ,DOLocationID INT
    -> ,payment_type INT
    -> ,fare_amount FLOAT
    -> ,extra FLOAT
    -> ,mta_tax FLOAT
    -> ,tip_amount FLOAT
    -> ,tolls_amount FLOAT
    -> ,improvement_surcharge FLOAT
    -> ,total_amount FLOAT
    -> ,congestion_surcharge FLOAT
    -> ,airport_fee FLOAT);
Query OK, 0 rows affected (0.03 sec)

MySQL [demoDB]> show tables;
+------------------+
| Tables_in_demoDB |
+------------------+
| taxi_data        |
+------------------+
1 row in set (0.00 sec)

MySQL [demoDB]>
```

Loading the data from 2 files: Run below 2 commands to load data to RDS table taxi_data.

LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv'

INTO TABLE taxi_data

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

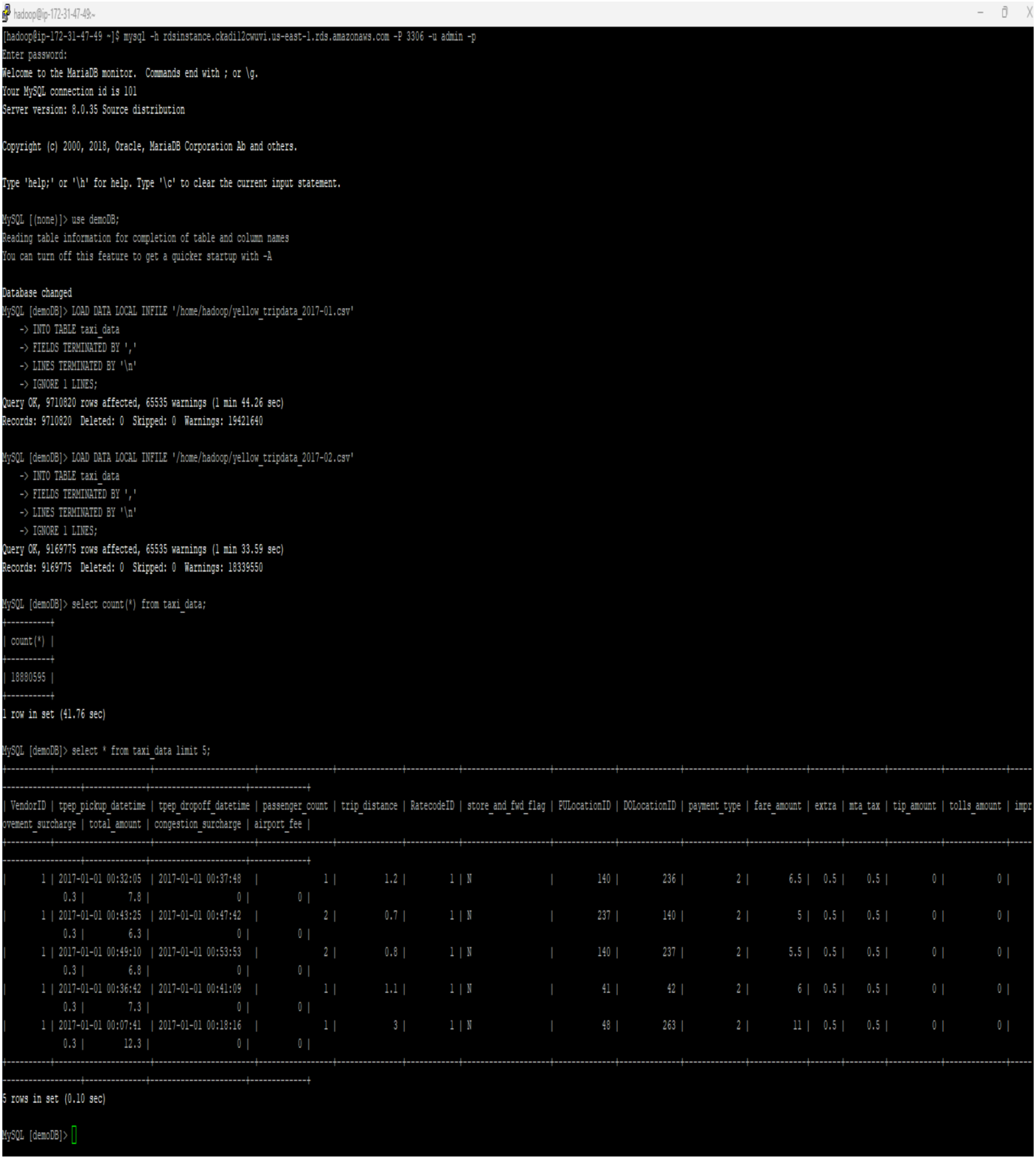
```
LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv'

INTO TABLE taxi_data

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;
```



```
[hadoop@ip-172-31-47-49 ~]$ mysql -h rdsinstance.ckadil2cwuvi.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 101
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 demoDB;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MySQL [demoDB]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv'
    -> INTO TABLE taxi_data
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 9710820 rows affected, 65535 warnings (1 min 44.26 sec)
Records: 9710820  Deleted: 0  Skipped: 0  Warnings: 19421640

MySQL [demoDB]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv'
    -> INTO TABLE taxi_data
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (1 min 33.59 sec)
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 18339550

MySQL [demoDB]> select count(*) from taxi_data;
+----------+
| count(*) |
+----------+
| 18880595 |
+----------+
1 row in set (41.76 sec)

MySQL [demoDB]> select * from taxi_data limit 5;
+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+--------------+-------------+-------+---------+------------+--------------+-----
------------------+--------------+----------------------+-------------+
| 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 | impr
ovement_surcharge | total_amount | congestion_surcharge | 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 |           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 |           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 |           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 |           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 |           0 |
+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+--------------+-------------+-------+---------+------------+--------------+-----
------------------+--------------+----------------------+-------------+
5 rows in set (0.10 sec)

MySQL [demoDB]>
```