Task 1

Create an RDS instance in your AWS account and upload the data to the RDS instance

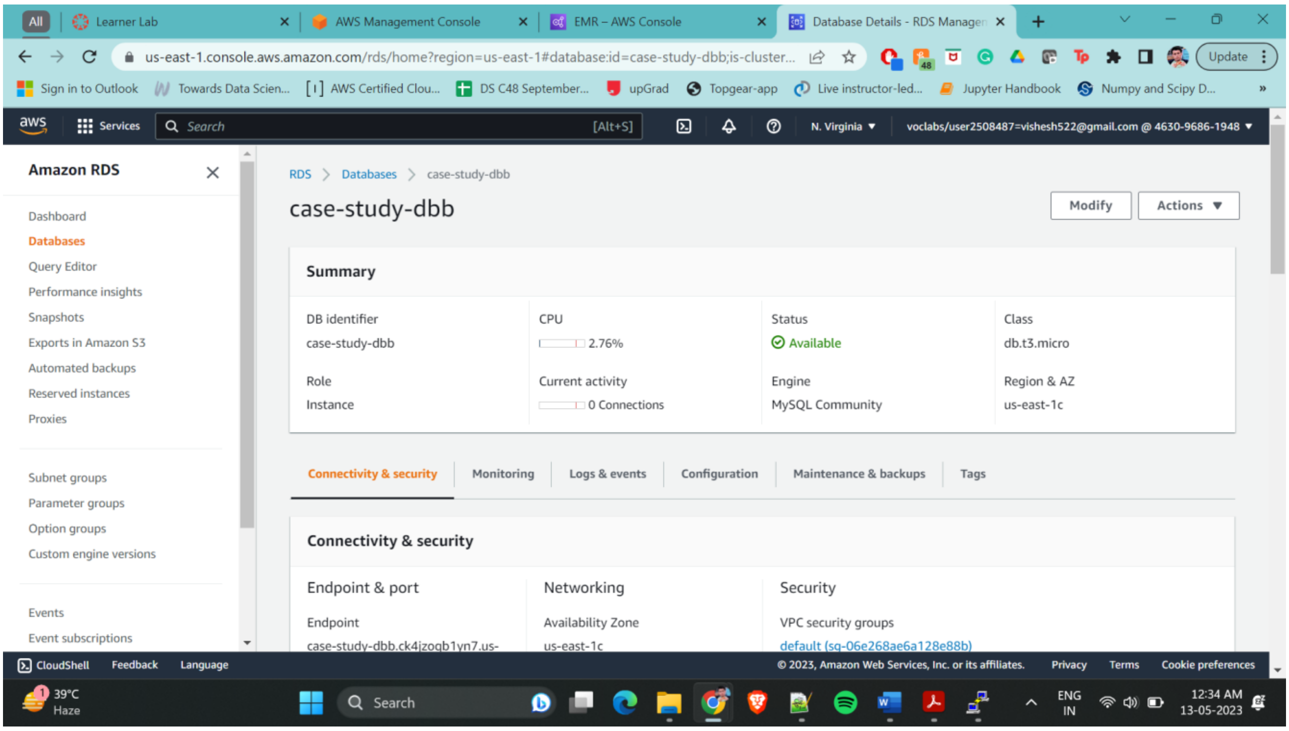
1) We have created RDS instance from Learner Lab and Launched EMR cluster.

2) Downloaded the required files for this task - **yellow_tripdata_2017-01.csv** & **yellow_tripdata_2017-02.csv**

3) Connect RDS instance from EMR and create table and load records from csv files.

4) Loading the data to RDS

RDS Created:



case-study-dbb

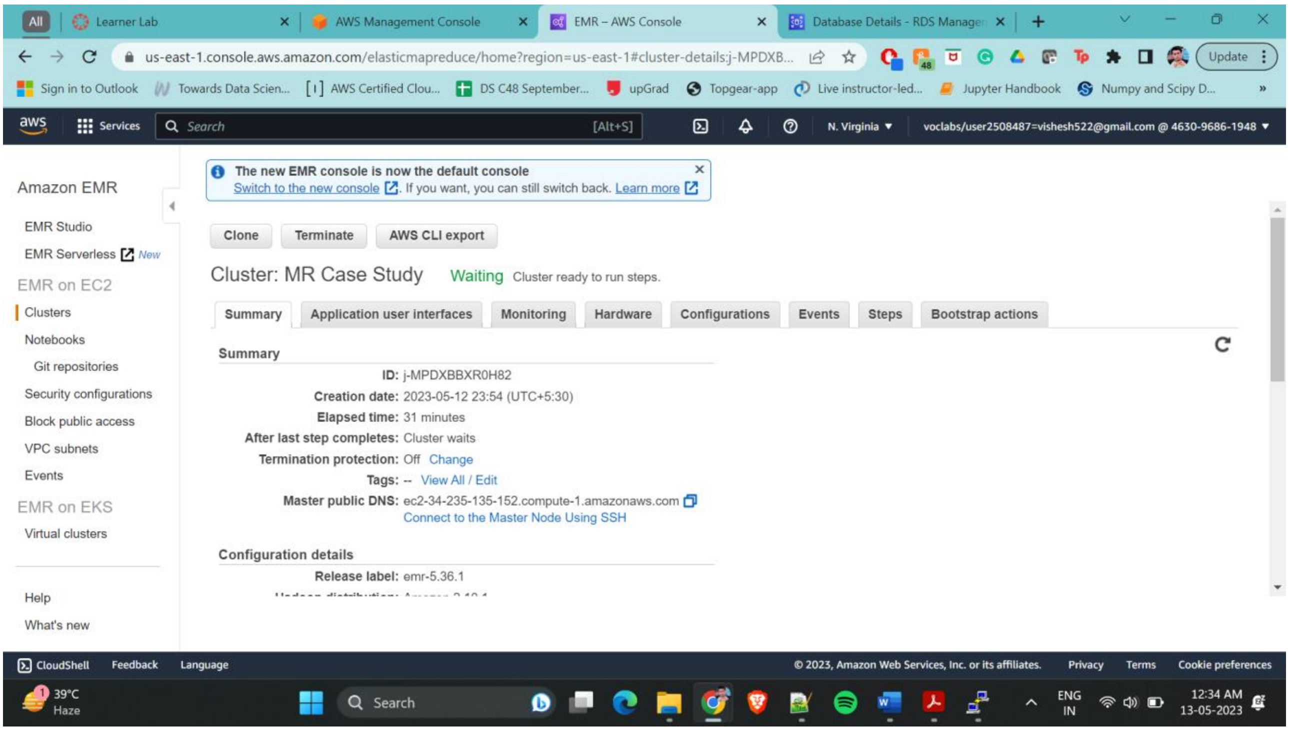
Summary

DB identifier	CPU	Status	Class
case-study-dbb	2.76%	Available	db.t3.micro
Role	**Current activity**	**Engine**	**Region & AZ**
Instance	0 Connections	MySQL Community	us-east-1c

Connectivity & security

Endpoint & port	Networking	Security
Endpoint	Availability Zone	VPC security groups
case-study-dbb.ck4jzogb1yn7.us-	us-east-1c	default (sg-06e268ae6a128e88b)

EMR Cluster:



The new EMR console is now the default console
Switch to the new console. If you want, you can still switch back. Learn more

Cluster: MR Case Study Waiting Cluster ready to run steps.

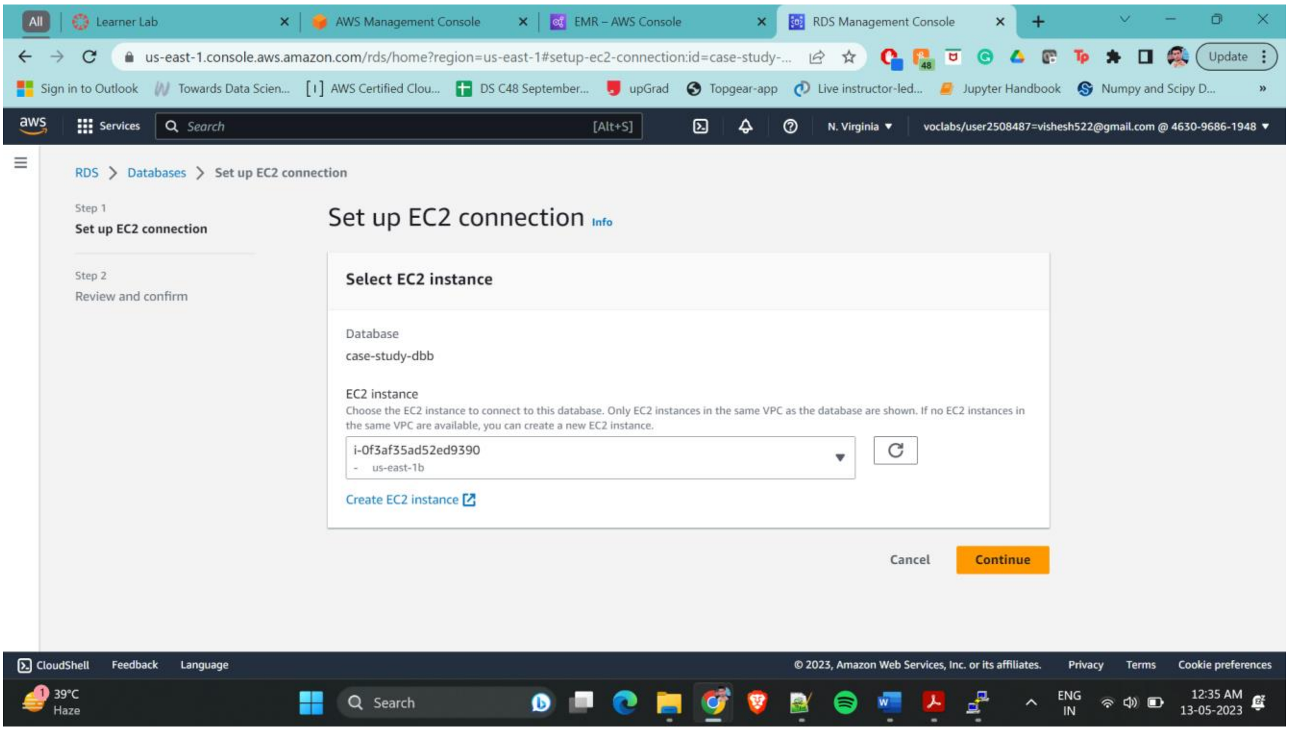
Summary

ID: j-MPDXBBXR0H82
Creation date: 2023-05-12 23:54 (UTC+5:30)
Elapsed time: 31 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-34-235-135-152.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.36.1

Connecting RDS with EMR EC2 instance:



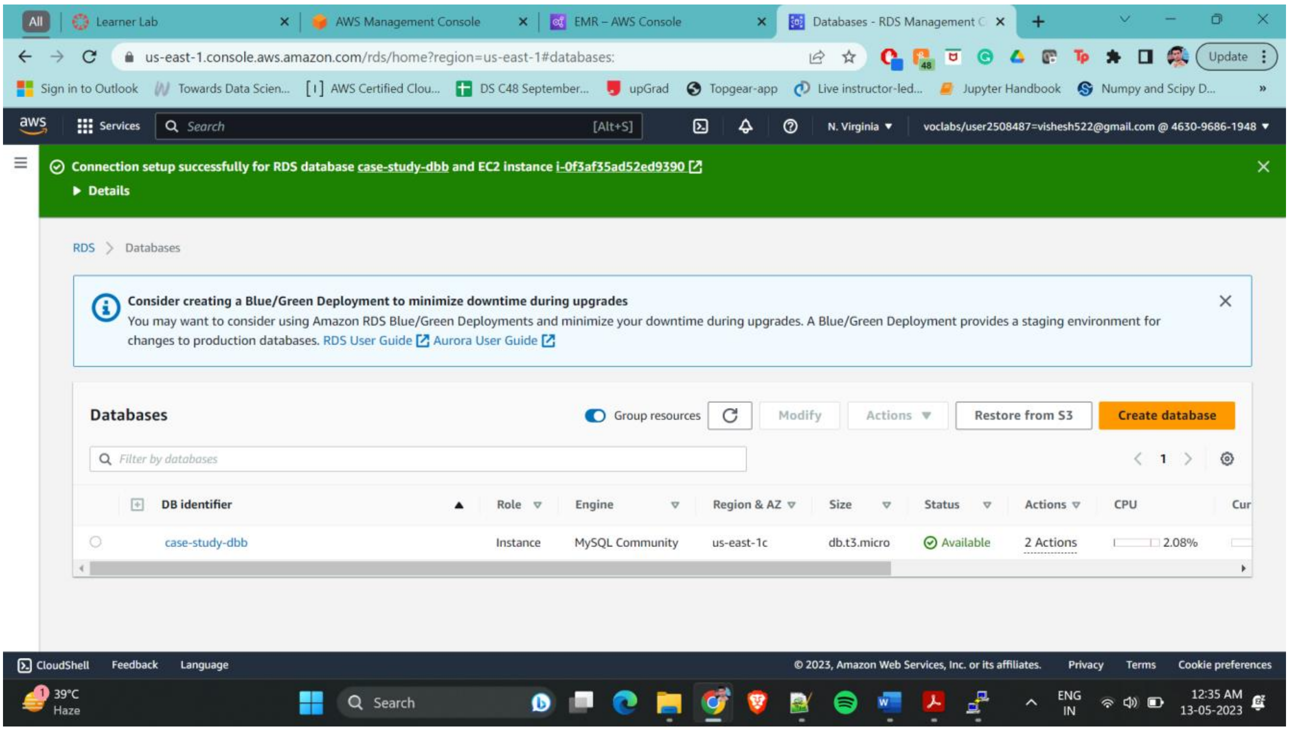
Set up EC2 connection

Select EC2 instance

Database
case-study-dbb

EC2 instance
Choose the EC2 instance to connect to this database. Only EC2 instances in the same VPC as the database are shown. If no EC2 instances in the same VPC are available, you can create a new EC2 instance.

i-0f3af35ad52ed9390
- us-east-1b



Connection setup successfully for RDS database case-study-dbb and EC2 instance i-0f3af35ad52ed9390

RDS > Databases

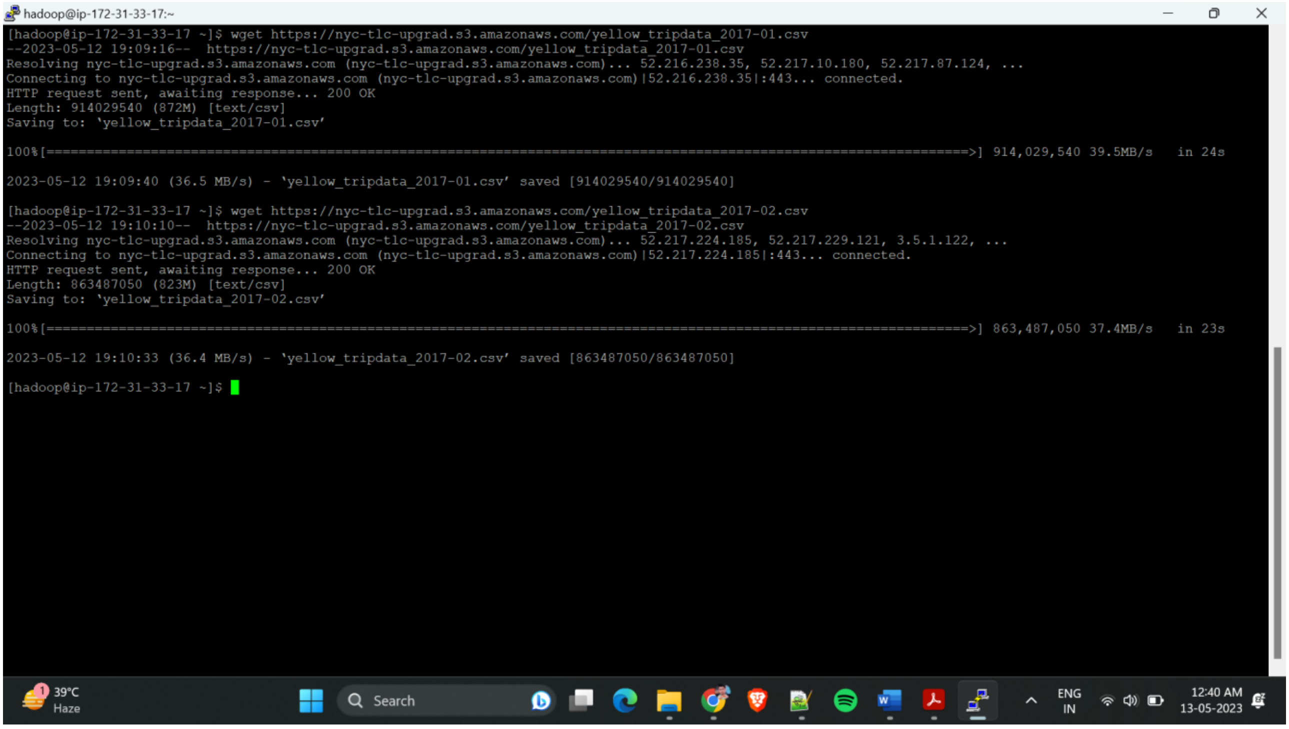
Consider creating a Blue/Green Deployment to minimize downtime during upgrades
You may want to consider using Amazon RDS Blue/Green Deployments and minimize your downtime during upgrades. A Blue/Green Deployment provides a staging environment for changes to production databases. RDS User Guide Aurora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

Databases

us-east-1.console.aws.amazon.com/rds/home?region=us-east-1#databases:
N. Virginia
voclabs/user2508487=vishesh522@gmail.com @ 4630-9686-1948
Group resources
Modify
Actions
Restore from S3
Create database
Filter by databases

DB identifier	Role	Engine	Region & AZ	Size	Status	Actions	CPU
case-study-dbb	Instance	MySQL Community	us-east-1c	db.t3.micro	Available	2 Actions	2.08%

Connecting EMR instance with PuTTy and then downloading **yellow_tripdata_2017-01.csv** & **yellow_tripdata_2017-02.csv** :

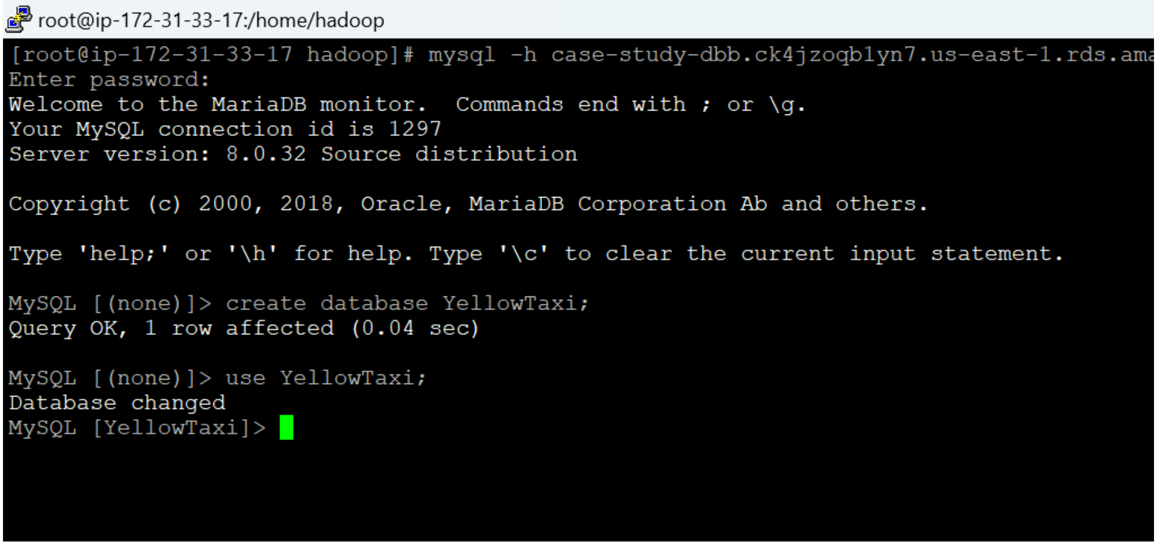


```
[hadoop@ip-172-31-33-17 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
--2023-05-12 19:09:16--  https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 52.216.238.35, 52.217.10.180, 52.217.87.124, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|52.216.238.35|: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 39.5MB/s   in 24s

2023-05-12 19:09:40 (36.5 MB/s) - ‘yellow_tripdata_2017-01.csv’ saved [914029540/914029540]

[hadoop@ip-172-31-33-17 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
--2023-05-12 19:10:10--  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)... 52.217.224.185, 52.217.229.121, 3.5.1.122, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|52.217.224.185|: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 37.4MB/s   in 23s

2023-05-12 19:10:33 (36.4 MB/s) - ‘yellow_tripdata_2017-02.csv’ saved [863487050/863487050]

[hadoop@ip-172-31-33-17 ~]$ 
```

Connecting RDS with EMR Instance:

Hostname: case-study-dbb.ck4jzoqb1yn7.us-east-1.rds.amazonaws.com

mysql -h case-study-dbb.ck4jzoqb1yn7.us-east-1.rds.amazonaws.com -P 3306 -u admin -p



```
[root@ip-172-31-33-17 hadoop]# mysql -h case-study-dbb.ck4jzoqb1yn7.us-east-1.rds.ama
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 1297
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)]> create database YellowTaxi;
Query OK, 1 row affected (0.04 sec)

MySQL [(none)]> use YellowTaxi;
Database changed
MySQL [YellowTaxi]>
```

create database YellowTaxi;

user YellowTaxi;

Creating table--

CREATE TABLE taxi

(

vendorID INT,

tpep_pickup_datetime DATETIME,

tpep_dropoff_datetime DATETIME,

passenger_count INT,

trip_distance DOUBLE,

puLocationID INT,

doLocationID INT,

rateCodeID INT,

store_and_fwd_flag VARCHAR(255),

payment_type INT,

fare_amount DOUBLE,

extra DOUBLE,

mta_tax DOUBLE,

improvement_surcharge DOUBLE,

tip_amount DOUBLE,

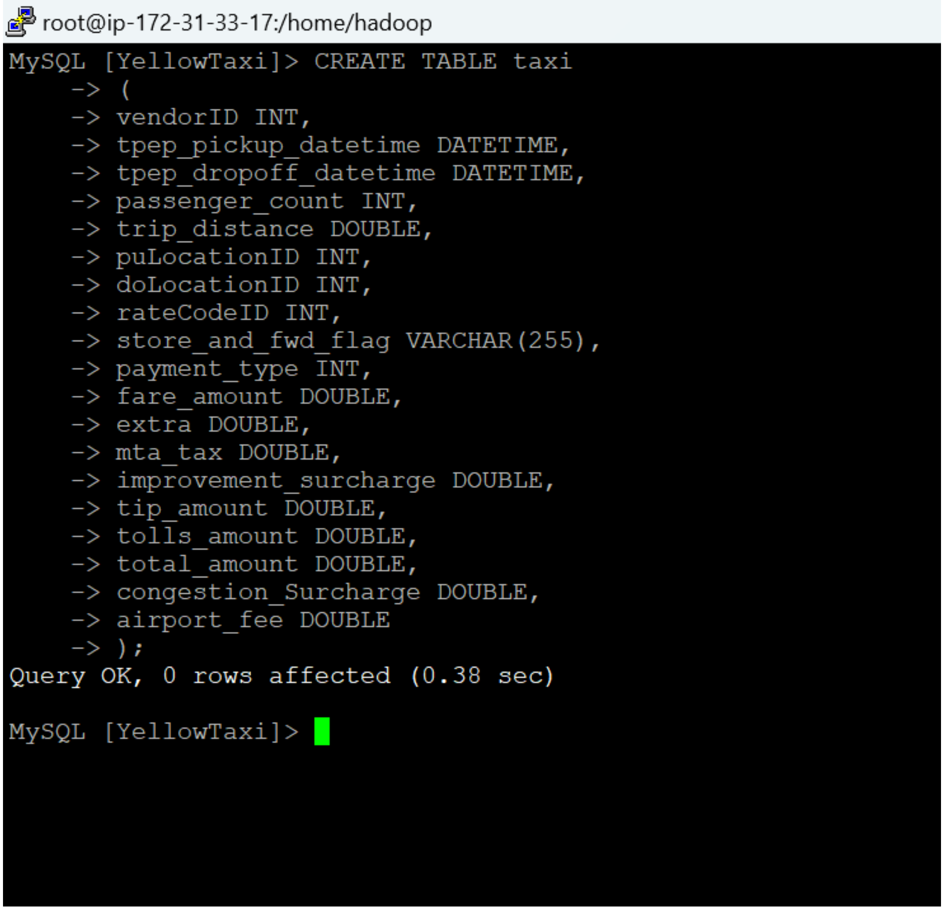
tolls_amount DOUBLE,

total_amount DOUBLE,

congestion_Surcharge DOUBLE,

airport_fee DOUBLE

);



root@ip-172-31-33-17:/home/hadoop

```
MySQL [YellowTaxi]> CREATE TABLE taxi
    -> (
    -> vendorID INT,
    -> tpep_pickup_datetime DATETIME,
    -> tpep_dropoff_datetime DATETIME,
    -> passenger_count INT,
    -> trip_distance DOUBLE,
    -> puLocationID INT,
    -> doLocationID INT,
    -> rateCodeID INT,
    -> store_and_fwd_flag VARCHAR(255),
    -> payment_type INT,
    -> fare_amount DOUBLE,
    -> extra DOUBLE,
    -> mta_tax DOUBLE,
    -> improvement_surcharge DOUBLE,
    -> tip_amount DOUBLE,
    -> tolls_amount DOUBLE,
    -> total_amount DOUBLE,
    -> congestion_Surcharge DOUBLE,
    -> airport_fee DOUBLE
    -> );
Query OK, 0 rows affected (0.38 sec)

MySQL [YellowTaxi]>
```

Load Data into Above table

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

INTO TABLE taxi

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

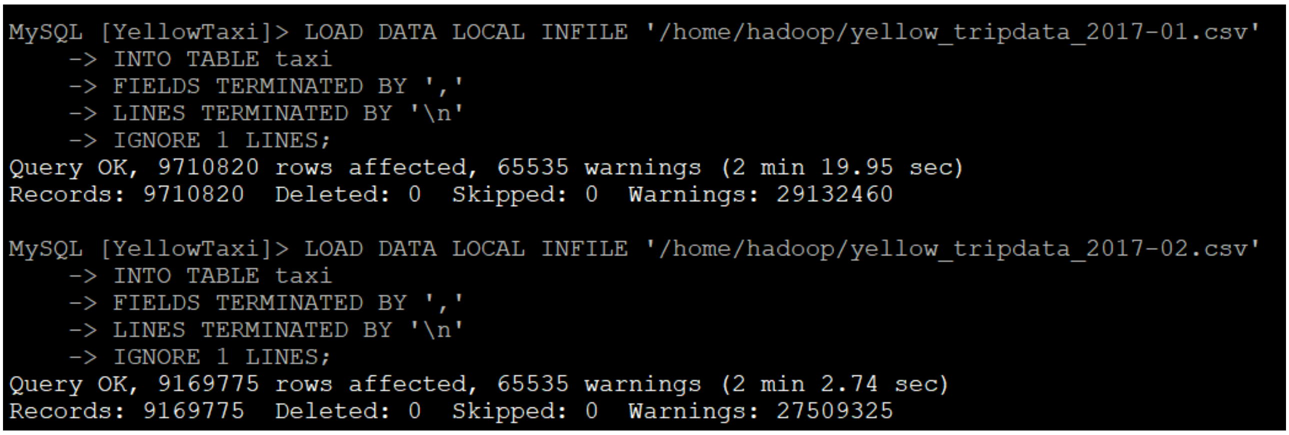
LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv'

INTO TABLE taxi

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;



```
MySQL [YellowTaxi]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv'
    -> INTO TABLE taxi
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 9710820 rows affected, 65535 warnings (2 min 19.95 sec)
Records: 9710820  Deleted: 0  Skipped: 0  Warnings: 29132460

MySQL [YellowTaxi]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv'
    -> INTO TABLE taxi
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (2 min 2.74 sec)
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 27509325
```