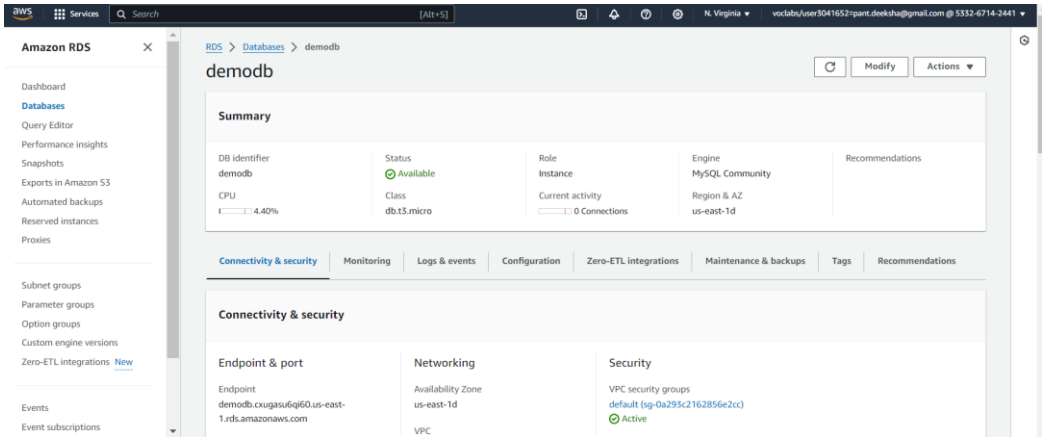
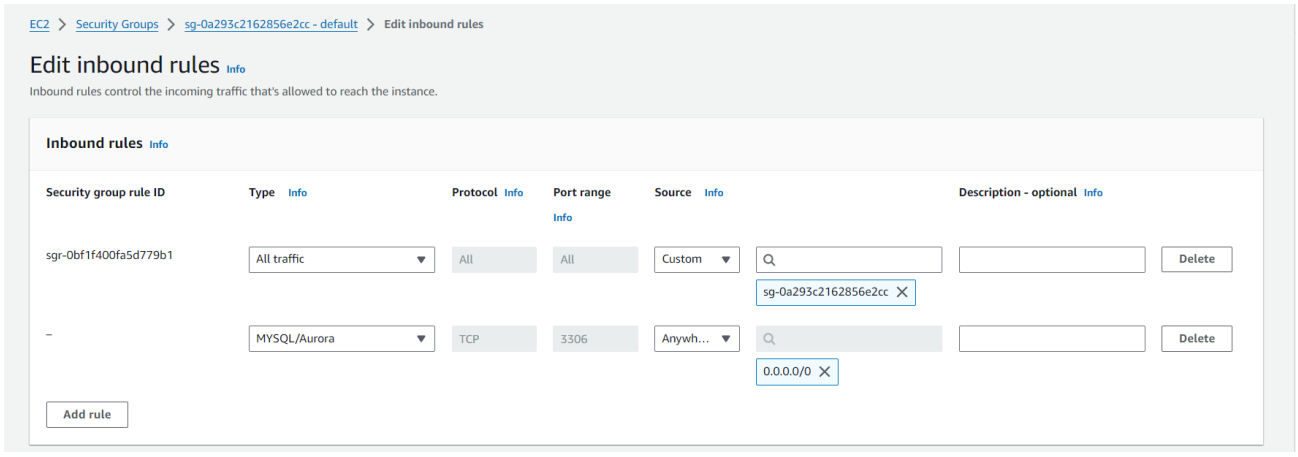
Task 1.
Create an RDS instance in AWS account and upload the data to the RDS instance.
Since the dataset is huge, uploaded the data from only two files
(i.e. yellow_tripdata_2017-01.csv & yellow_tripdata_2017-02.csv) from the dataset.
Created an appropriate schema before uploading the data sets to AWS RDS

Created an RDS instance in AWS account

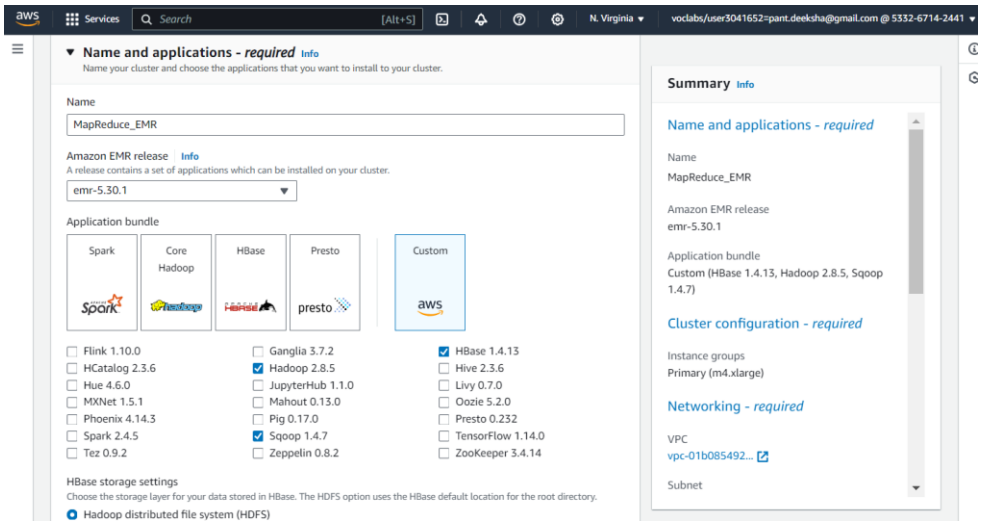


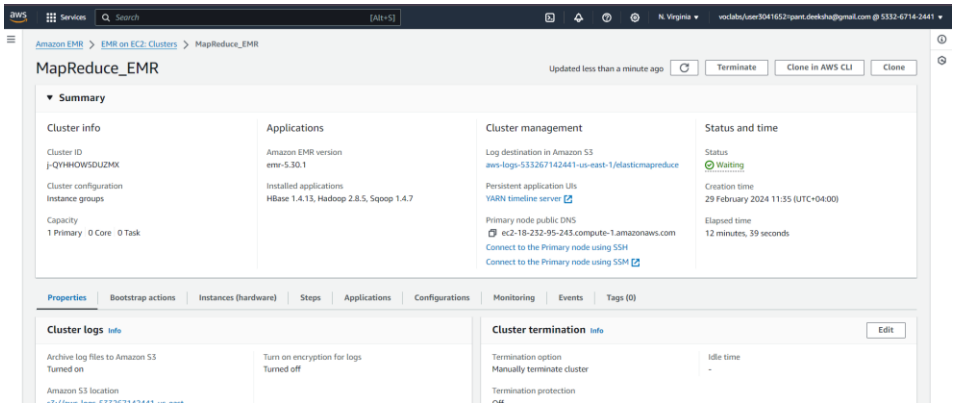
Database Name: demodb
User: admin

Updated the IP address under the Security Groups to access the RDS instance.
Provided access to the 3306 port to the required IP address.



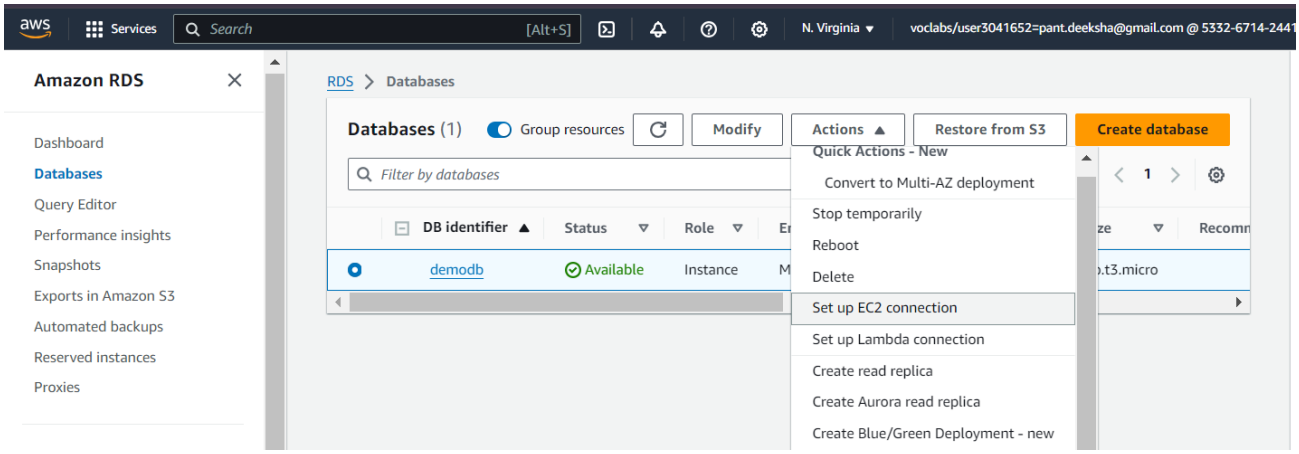
Created EMR cluster & configured Security group->inbound rules

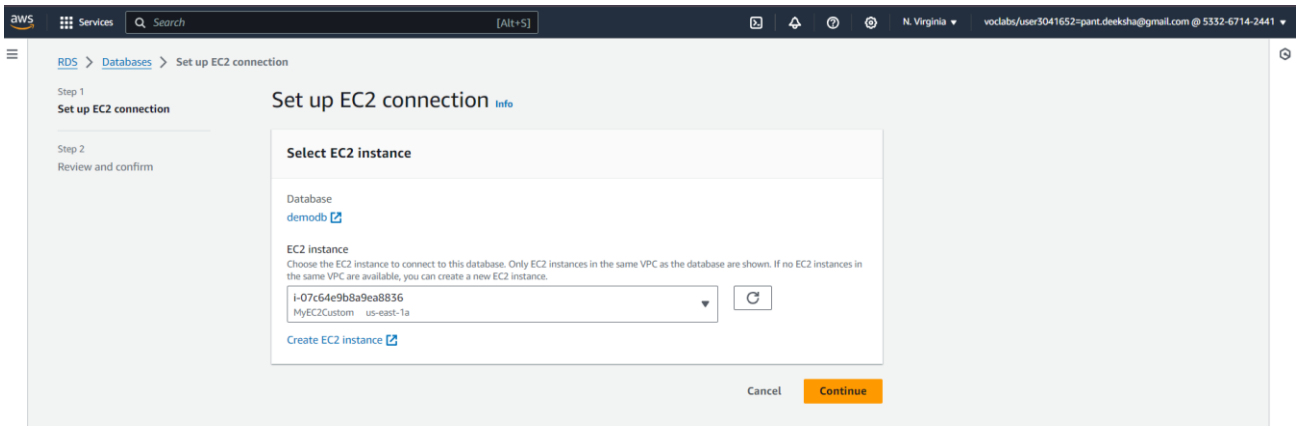


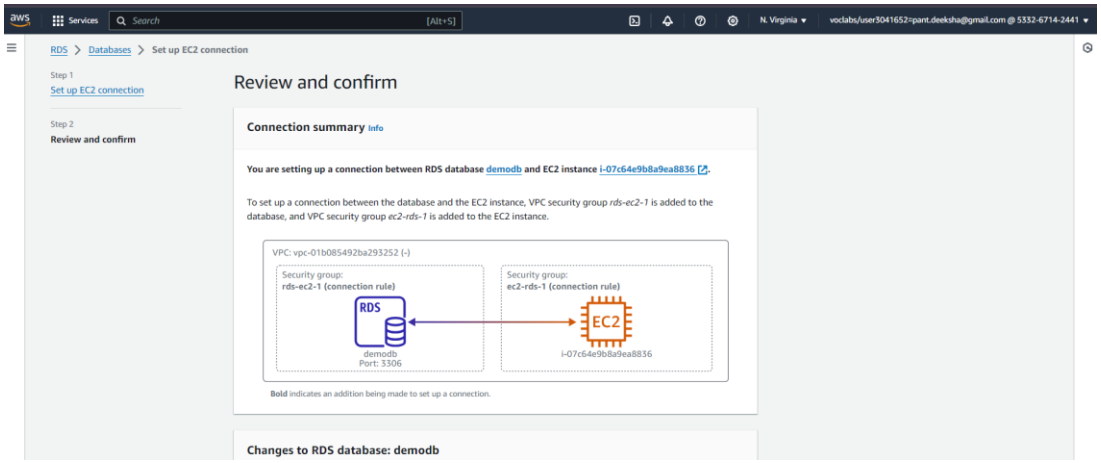


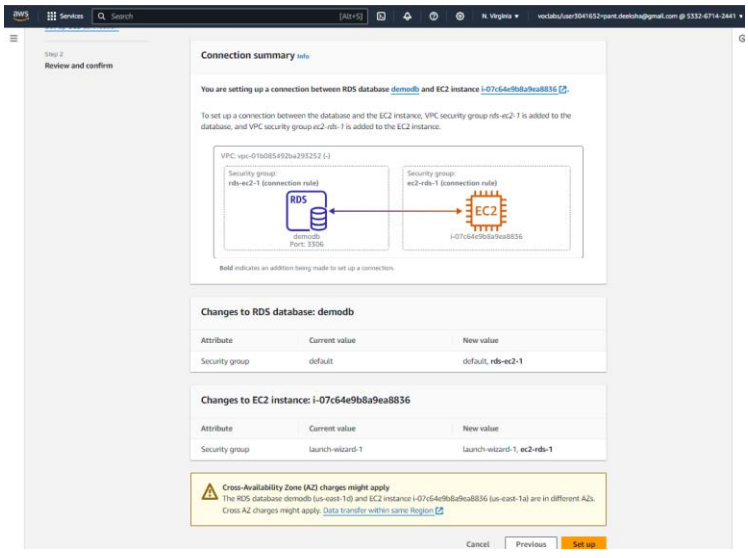
Primary Node DNS: ec2-18-232-95-243.compute-1.amazonaws.com

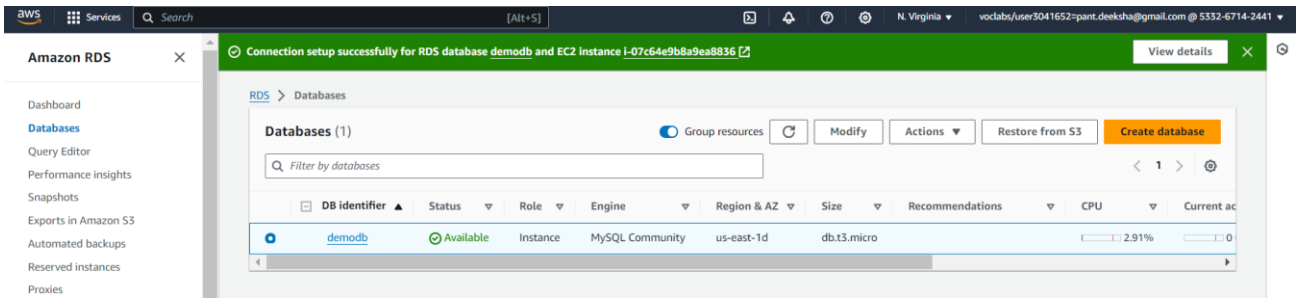
Connected RDS instance to EC2



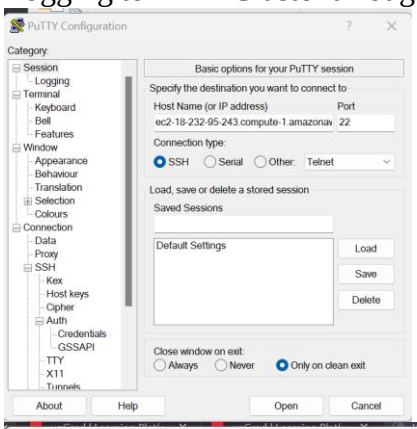

Services
Search
[Alt+S]
N. Virginia
voclabs/user3041652=pant.deeksha@gmail.com @ 5332-6714-2441
RDS > Databases > Set up EC2 connection
Step 1
Set up EC2 connection
Step 2
Review and confirm
Set up EC2 connection Info
Select EC2 instance
Database
demodb
EC2 instance
Choose the EC2 instance to connect 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-07c64e9b8a9ea8836
MyEC2Custom us-east-1a
Create EC2 instance
Cancel
Continue

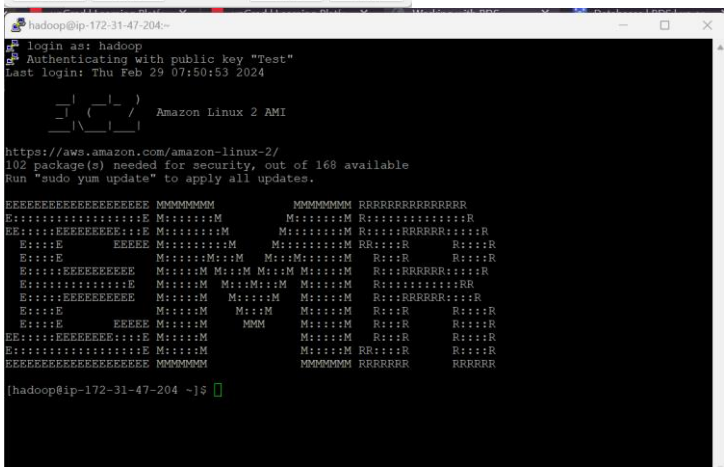

Services
Search
[Alt+S]
N. Virginia
voclabs/user3041652=pant.deeksha@gmail.com @ 5332-6714-2441
RDS > Databases > Set up EC2 connection
Step 1
Set up EC2 connection
Step 2
Review and confirm
Review and confirm
Connection summary Info
You are setting up a connection between RDS database demodb and EC2 instance i-07c64e9b8a9ea8836.
To set up a connection between the database and the EC2 instance, VPC security group rds-ec2-1 is added to the database, and VPC security group ec2-rds-1 is added to the EC2 instance.
VPC: vpc-01b085492ba293252 (-)
Security group:
rds-ec2-1 (connection rule)
RDS
demodb
Port: 3306
Security group:
ec2-rds-1 (connection rule)
EC2
i-07c64e9b8a9ea8836
Bold indicates an addition being made to set up a connection.
Changes to RDS database: demodb


Connection summary Info
You are setting up a connection between RDS database demodb and EC2 instance i-07c64e9b8a9ea8836.
To set up a connection between the database and the EC2 instance, VPC security group rds-ec2-1 is added to the database, and VPC security group ec2-rds-1 is added to the EC2 instance.
RDS
EC2
Changes to RDS database: demodb
Attribute
Current value
New value
Security group
default
default, rds-ec2-1
Changes to EC2 instance: i-07c64e9b8a9ea8836
Attribute
Current value
New value
Security group
launch-wizard-1
launch-wizard-1, ec2-rds-1
Cross-Availability Zone (AZ) charges might apply
The RDS database demodb (us-east-1d) and EC2 instance i-07c64e9b8a9ea8836 (us-east-1a) are in different AZs.
Cross AZ charges might apply. Data transfer within same Region
Cancel
Previous
Set up


Services
Search
[Alt+S]
N. Virginia
voclabs/user3041652=pant.deeksha@gmail.com @ 5332-6714-2441
Amazon RDS
Dashboard
Databases
Query Editor
Performance insights
Snapshots
Exports in Amazon S3
Automated backups
Reserved instances
Proxies
Connection setup successfully for RDS database demodb and EC2 instance i-07c64e9b8a9ea8836
View details
RDS > Databases
Databases (1)
Group resources
Modify
Actions
Restore from S3
Create database
Filter by databases
DB identifier
Status
Role
Engine
Region & AZ
Size
Recommendations
CPU
demodb
Available
Instance
MySQL Community
us-east-1d
db.t3.micro
2.91%

Logging to EMR Cluster through SSH via Putty

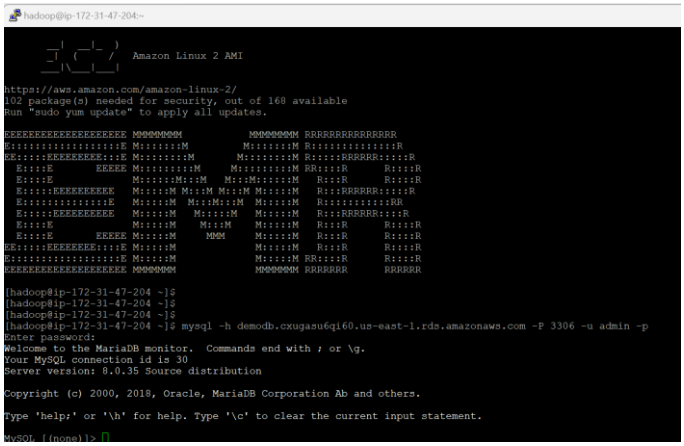




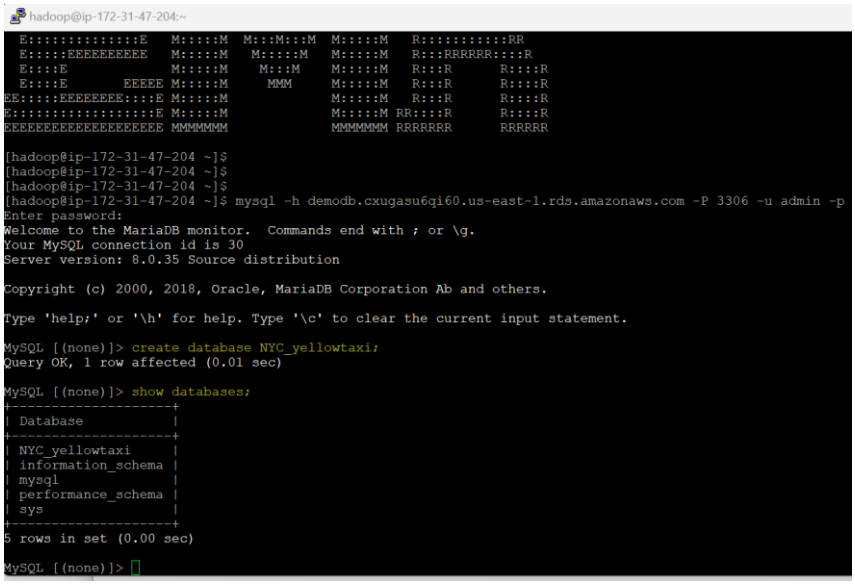
Connecting to RDS:

Command used:
mysql -h demodb.cxugasu6qi60.us-east-1.rds.amazonaws.com -P 3306 -u admin -p

After entering this command, entered the password

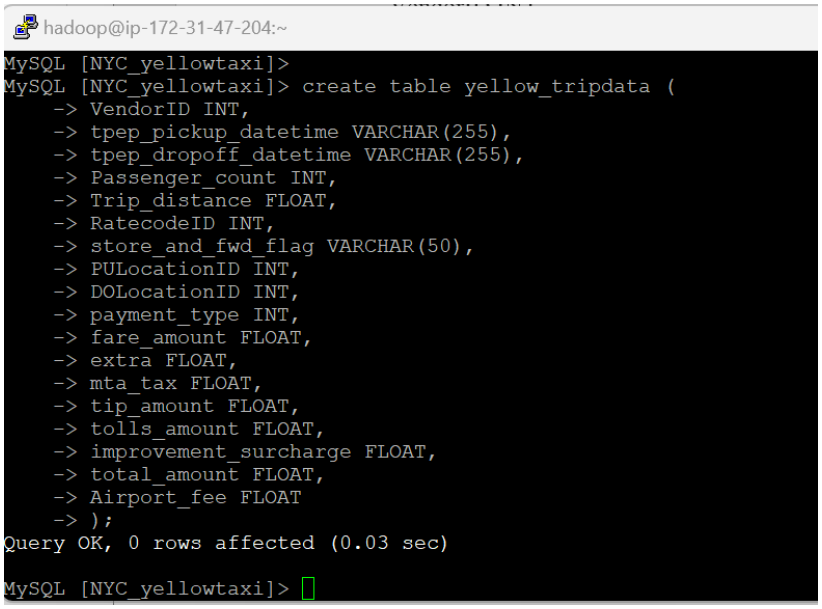


Created a database in RDS:



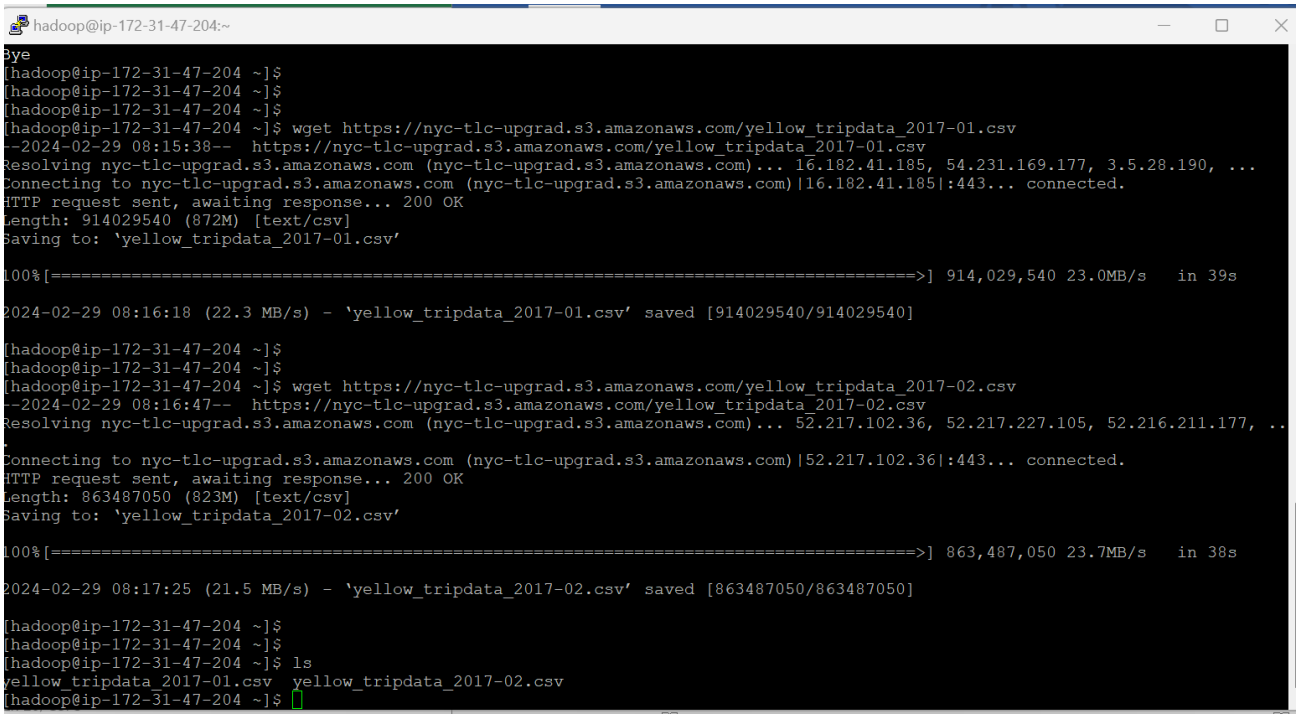
Created a table structure in RDS:

```
create table yellow_tripdata (
VendorID INT,
tpep_pickup_datetime VARCHAR(255),
tpep_dropoff_datetime VARCHAR(255),
Passenger_count INT,
Trip_distance FLOAT,
RatecodeID INT,
store_and_fwd_flag VARCHAR(50),
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,
Airport_fee FLOAT
);
```



Downloaded the necessary data on Hadoop local file system.

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



Loaded data in MySQL table with following commands:

LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv'
INTO TABLE yellow_tripdata
FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'
IGNORE 1 LINES;

```
MySQL [NYC_yellowtaxi]>
MySQL [NYC_yellowtaxi]>
MySQL [NYC_yellowtaxi]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv'
    -> INTO TABLE yellow_tripdata
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 9710820 rows affected, 65535 warnings (2 min 4.06 sec)
Records: 9710820  Deleted: 0  Skipped: 0  Warnings: 9710820
```

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

```
MySQL [NYC_yellowtaxi]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv'
    -> INTO TABLE yellow_tripdata
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;

Query OK, 9169775 rows affected, 65535 warnings (2 min 6.43 sec)
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 9169775
```

Executing commands to make sure that tables have been loaded correctly:

```
MySQL [NYC_yellowtaxi]>
MySQL [NYC_yellowtaxi]> select COUNT(*) from yellow_tripdata;
+----------+
| COUNT(*) |
+----------+
| 18880595 |
+----------+
1 row in set (51.63 sec)

MySQL [NYC_yellowtaxi]> select * from yellow_tripdata limit 2;
+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+
-------------+--------------+-------------+------------+---------+------------+-------------+-----------------------+
--------------+-------------+
| 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 |
+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+
-------------+--------------+-------------+------------+---------+------------+-------------+-----------------------+
--------------+-------------+
|        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 |
+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+
-------------+--------------+-------------+------------+---------+------------+-------------+-----------------------+
--------------+-------------+
2 rows in set (0.07 sec)
```