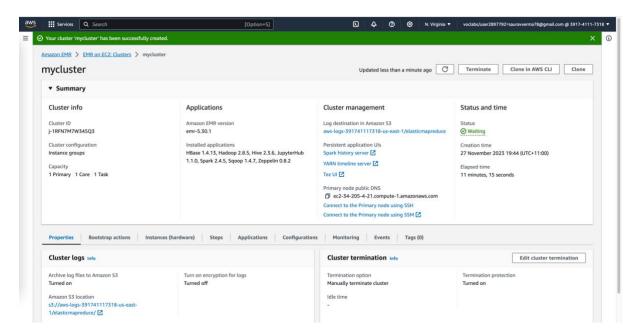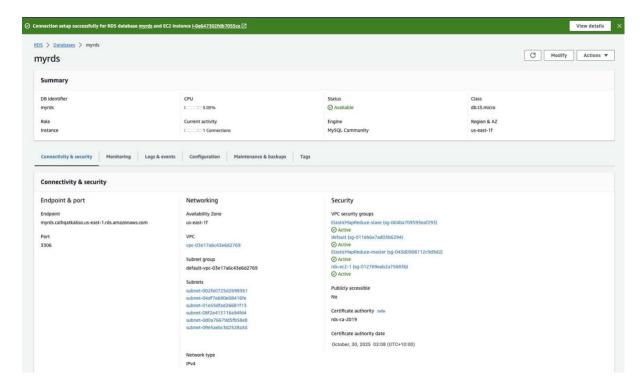# Task 1

Create an RDS instance in your AWS account and upload the data to the RDS instance.
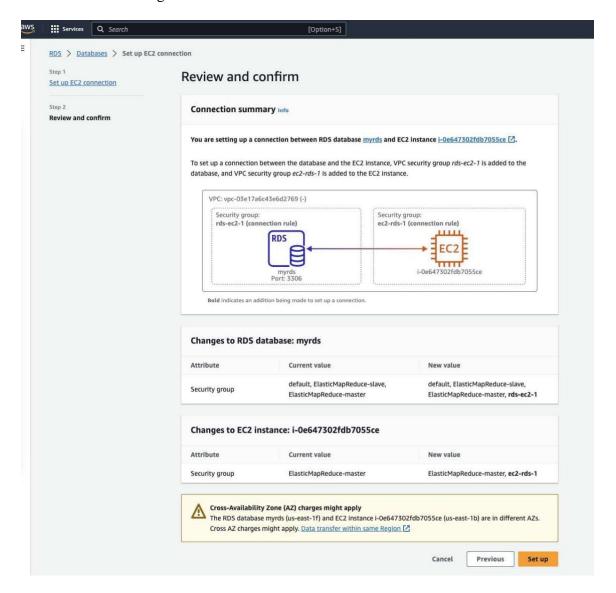
1. EMR EC2 cluster with required applications Hadoop, HBase, Sqoop etc.

2. RDS instance created.

3. RDS to EC2 linkage.

4. Login to EC2 instance and download database files using commands:

```
ssh -i ~/Downloads/mr-assignment.pem ec2-user@ec2-35-172-218-206.compute-1.amazonaws.com
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
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-03.csv
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-04.csv
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-05.csv
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-06.csv
```

5. Remaining Steps:

- List all the downloaded files
- Login to MySQL server
  - Create a new Database
  - Create a new Table with schema similar to the dataset
  - Load data from the csv files into the database.
  - Check if the data is present or not.

```
[ec2-user@ip-172-31-67-221 ~]$ ls -l
total 5425468
-rw-rw-r-- 1 ec2-user ec2-user 914029540 Nov 25  2022 yellow_tripdata_2017-01.csv
-rw-rw-r-- 1 ec2-user ec2-user 863487050 Nov 25  2022 yellow_tripdata_2017-02.csv
-rw-rw-r-- 1 ec2-user ec2-user 969809025 Nov 25  2022 yellow_tripdata_2017-03.csv
-rw-rw-r-- 1 ec2-user ec2-user 946349441 Nov 25  2022 yellow_tripdata_2017-04.csv
-rw-rw-r-- 1 ec2-user ec2-user 951965526 Nov 25  2022 yellow_tripdata_2017-05.csv
-rw-rw-r-- 1 ec2-user ec2-user 910028408 Nov 25  2022 yellow_tripdata_2017-06.csv
[ec2-user@ip-172-31-67-221 ~]$ mysql -h mr-assignment-db.cj4l0tk6qknz.us-east-1.rds.amazonaws.com -u admin -p
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 607
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 taxi_trip_db;
Query OK, 1 row affected (0.05 sec)

MySQL [(none)]> use taxi_trip_db;
Database changed
MySQL [taxi_trip_db]> create table TripRecords (
    ->     VendorID int NOT NULL,
    ->     tpep_pickup_datetime datetime NOT NULL,
    ->     tpep_dropoff_datetime datetime NOT NULL,
    ->     Passenger_count int NOT NULL,
    ->     Trip_distance float NOT NULL,
    ->     RateCodeID int NOT NULL,
    ->     Store_and_fwd_flag varchar(1) NOT NULL,
    ->     PULocationID int NOT NULL,
    ->     DOLocationID int NOT NULL,
    ->     Payment_type int NOT NULL,
    ->     Fare_amount float NOT NULL,
    ->     Extra float NOT NULL,
    ->     MTA_tax float NOT NULL,
    ->     Tip_amount float NOT NULL,
    ->     Tolls_amount float NOT NULL,
    ->     Improvement_surcharge float NOT NULL,
    ->     Total_amount float NOT NULL,
    ->     Congestion_Surcharge float NOT NULL,
    ->     Airport_fee float NOT NULL,
    ->     PRIMARY KEY (tpep_pickup_datetime, tpep_dropoff_datetime)
    -> );
Query OK, 0 rows affected (0.11 sec)

MySQL [taxi_trip_db]> load data local infile '/home/ec2-user/yellow_tripdata_2017-01.csv'
    -> into table trip_records
    -> fields terminated by','
    -> lines terminated by'\n'
    -> ignore 1 lines;
ERROR 1146 (42S02): Table 'taxi_trip_db.trip_records' doesn't exist
MySQL [taxi_trip_db]> load data local infile '/home/ec2-user/yellow_tripdata_2017-01.csv'
    -> into table TripRecords
    -> fields terminated by','
    -> lines terminated by'\n'
    -> ignore 1 lines;
Query OK, 9691163 rows affected, 65535 warnings (2 min 29.43 sec)
Records: 9710820  Deleted: 0  Skipped: 19657  Warnings: 19441297

MySQL [taxi_trip_db]> load data local infile '/home/ec2-user/yellow_tripdata_2017-02.csv'
    -> into table TripRecords
    -> fields terminated by','
    -> lines terminated by'\n'
    -> ignore 1 lines;
Query OK, 9150885 rows affected, 65535 warnings (2 min 9.66 sec)
Records: 9169775  Deleted: 0  Skipped: 18890  Warnings: 18358440
```

```
MySQL [taxi_trip_db]> load data local infile '/home/ec2-user/yellow_tripdata_2017-01.csv'
    -> into table TripRecords
    -> fields terminated by','
    -> lines terminated by'\n'
    -> ignore 1 lines;
Query OK, 9691163 rows affected, 65535 warnings (2 min 29.43 sec)
Records: 9710820  Deleted: 0  Skipped: 19657  Warnings: 19441297

MySQL [taxi_trip_db]> load data local infile '/home/ec2-user/yellow_tripdata_2017-02.csv'
    -> into table TripRecords
    -> fields terminated by','
    -> lines terminated by'\n'
    -> ignore 1 lines;
Query OK, 9150885 rows affected, 65535 warnings (2 min 9.66 sec)
Records: 9169775  Deleted: 0  Skipped: 18890  Warnings: 18358440

MySQL [taxi_trip_db]> select count(*) from TripRecords;
+----------+
| count(*) |
+----------+
| 18842048 |
+----------+
1 row in set (55.62 sec)

MySQL [taxi_trip_db]> select * from TripRecords 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	Congestion_Surcharge	Airport_fee
2	2017-01-01 00:00:00	2017-01-01 00:00:00	1	0.02	2	N	249	234	2	52	0	0.5	0	0	0.3	52.8	0	0
1	2017-01-01 00:00:02	2017-01-01 00:03:50	1	0.5	1	N	48	48	2	4	0.5	0.5	0	0	0.3	5.3	0	0
2	2017-01-01 00:00:02	2017-01-01 00:39:22	4	7.75	1	N	186	36	1	22	0.5	0.5	4.66	0	0.3	27.96	0	0
1	2017-01-01 00:00:03	2017-01-01 00:06:58	1	0.8	1	N	162	161	1	6	0.5	0.5	1.45	0	0.3	8.75	0	0
2	2017-01-01 00:00:05	2017-01-01 00:05:04	5	1.76	1	N	140	74	2	7	0.5	0.5	0	0	0.3	8.3	0	0
1	2017-01-01 00:00:05	2017-01-01 00:08:33	2	0.9	1	N	48	50	1	7	0.5	0.5	0	0	0.3	8.3	0	0
2	2017-01-01 00:00:05	2017-01-01 00:15:36	1	8.47	1	N	138	262	1	24	0.5	0.5	7.71	5.54	0.3	38.55	0	0
1	2017-01-01 00:00:06	2017-01-01 00:08:53	1	0.9	1	N	234	186	1	7	0.5	0.5	2.05	0	0.3	10.35	0	0
2	2017-01-01 00:00:06	2017-01-01 00:09:30	4	2.43	1	N	141	107	1	9.5	0.5	0.5	2.7	0	0.3	13.5	0	0
1	2017-01-01 00:00:06	2017-01-01 00:11:56	2	2.4	1	N	142	236	2	10.5	0.5	0.5	0	0	0.3	11.8	0	0

```
10 rows in set (0.02 sec)
```