

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.

The screenshot displays the AWS Management Console for an Amazon EMR cluster named 'mycluster'. The cluster is in a 'Waiting' status. The console shows various details including cluster ID, configuration, capacity, and a list of installed applications like HBase, Hadoop, Hive, and Sqoop. It also provides information on cluster management, such as log destinations and persistent application URLs, and status and time details like creation and elapsed time.

Cluster info	Applications	Cluster management	Status and time
Cluster ID j-1RFN7M7W345Q3	Amazon EMR version emr-5.30.1	Log destination in Amazon S3 aws-logs-391741117318-us-east-1/elasticmapreduce	Status Waiting
Cluster configuration	Installed applications HBase 1.4.13, Hadoop 2.8.5, Hive 2.3.6, JupyterHub 1.1.0, Spark 2.4.5, Sqoop 1.4.7, Zeppelin 0.8.2	Persistent application UIs Spark history server YARN timeline server Tez UI	Creation time 27 November 2023 19:44 (UTC+11:00)
Instance groups 1 Primary 1 Core 1 Task		Primary node public DNS ec2-34-205-4-21.compute-1.amazonaws.com Connect to the Primary node using SSH Connect to the Primary node using SSM	Elapsed time 11 minutes, 15 seconds

2. RDS instance created.

The screenshot displays the AWS Management Console for an Amazon RDS instance named 'myrds'. The instance is in an 'Available' status. The console shows details such as DB identifier, CPU usage, role, and engine. It also provides information on connectivity and security, including endpoint, port, VPC, subnet group, and security groups.

DB identifier	CPU	Status	Class
myrds	3.05%	Available	db.t3.micro
Role Instance	Current activity 1 Connections	Engine MySQL Community	Region & AZ us-east-1f

Endpoint & port	Networking	Security
Endpoint myrds.caliqatka6so.us-east-1.rds.amazonaws.com	Availability Zone us-east-1f	VPC security groups ElasticMapReduce-slave (sg-004ba709593eaf293) Active default (sg-011eb6a7ad03b6294) Active ElasticMapReduce-master (sg-043d0908112c9d9d2) Active rds-ec2-1 (sg-012769eab2a75683b) Active
Port 3306	VPC vpc-03e17a6c43e6d2769	Publicly accessible No
	Subnet group default-vpc-03e17a6c43e6d2769	Certificate authority rds-ca-2019
	Subnets subnet-002fe0725d2b98361 subnet-04d7feb90e08416fe subnet-01e55d7d26981f13 subnet-08f2e413115d84e04 subnet-0d0a76679d5fb58e8 subnet-09e5aebc5d2528a3d	Certificate authority date October, 30, 2025 03:08 (UTC+10:00)
	Network type IPv4	

3. RDS to EC2 linkage.

Review and confirm

Connection summary [info](#)

You are setting up a connection between RDS database **myrds** and EC2 instance **i-0e647302fdb7055ce**.

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.

Changes to RDS database: myrds

Attribute	Current value	New value
Security group	default, ElasticMapReduce-slave, ElasticMapReduce-master	default, ElasticMapReduce-slave, ElasticMapReduce-master, rds-ec2-1

Changes to EC2 instance: i-0e647302fdb7055ce

Attribute	Current value	New value
Security group	ElasticMapReduce-master	ElasticMapReduce-master, ec2-rds-1

Cross-Availability Zone (AZ) charges might apply
The RDS database myrds (us-east-1f) and EC2 instance i-0e647302fdb7055ce (us-east-1b) are in different AZs. Cross AZ charges might apply. [Data transfer within same Region](#)

Cancel Previous **Set up**

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, 968163 rows affected, 65536 warnings (2 min 29.43 sec)
Records: 9758828 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, 918846 rows affected, 65536 warnings (2 min 9.46 sec)
Records: 9169776 Deleted: 0 Skipped: 18898 Warnings: 18358448

MySQL [taxi_trip_db]: select count(*) from TripRecords;
+-----+
| count(*) |
+-----+
| 18842848 |
+-----+
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 |
+----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | 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 |
| 1 | 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 |
| 1 | 2 | 2017-01-01 00:00:02 | 2017-01-01 00:39:22 | 1 | 7.75 | 1 | N | 186 | 36 | 1 | 22 | 0.5 | 0.5 | 4.66 | 0 | 0.3 | 27.96 | 0 |
| 1 | 1 | 2017-01-01 00:00:03 | 2017-01-01 00:08:50 | 4 | 0.8 | 1 | N | 162 | 161 | 1 | 4 | 0.5 | 0.5 | 1.45 | 0 | 0.3 | 8.75 | 0 |
| 2 | 2 | 2017-01-01 00:00:05 | 2017-01-01 00:05:04 | 5 | 1.76 | 1 | N | 148 | 74 | 2 | 7 | 0.5 | 0.5 | 0 | 0 | 0.3 | 8.3 | 0 |
| 1 | 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 |
| 2 | 2 | 2017-01-01 00:00:06 | 2017-01-01 00:15:36 | 1 | 8.47 | 1 | N | 138 | 262 | 1 | 24 | 0.5 | 0.5 | 7.71 | 5.64 | 0.3 | 38.55 | 0 |
| 1 | 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 | 2.85 | 0 | 0.3 | 10.35 | 0 |
| 2 | 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 | 13.0 | 0 |
| 1 | 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 | 13.8 | 0 |

```