

## Task 1: Setting up the AWS environment and loading data

Creating an RDS instance in my AWS account and uploading the data to the RDS instance.

### 1. RDS instance creation in AWS

emr-database

Summary

DB identifier

emr-database

CPU

3.42%

Status

Available

Class

db.t4g.micro

Role

Instance

Current activity

0 Connections

Engine

MySQL Community

Region & AZ

us-east-1f

Recommendations

Connectivity & security

Monitoring

Logs & events

Configuration

Zero-ETL integrations

Maintenance & backups

Data migrations - new

Tags

Recommendations

Connectivity & security

Endpoint & port

Endpoint

emr-database.c1o6d30u73d-us-east-1.elb.amazonaws.com

Networking

Availability Zone

us-east-1f

Security

VPC security groups

default for subnets in us-east-1f

### 2. EMR creation,some applications selected include: Apache Sqoop, Apache Hbase, Hadoop

Your cluster "EMR-Instance" has been successfully created.

EMR-Instance

Updated less than a minute ago

Terminate

Clone in AWS CLI

Clone

▼ Summary

Cluster info

Cluster ID

j-2ZPWEF37TLIQA

Cluster configuration

Instance groups

Capacity

1 Primary | 1 Core | 1 Task

Applications

Amazon EMR version

emr-7.4.0

Installed applications

Hadoop 3.4.0, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5.2, Sqoop 1.4.7

Cluster management

Log destination in Amazon S3

aws-logs-891377221307-us-east-1/elasticmapreduce

Persistent application UIs

Spark History Server

YARN timeline server

Tez UI

Primary node public DNS

ec2-34-207-98-69.compute-1.amazonaws.com

Connect to the Primary node using SSH

Connect to the Primary node using SSM

Status and time

Status

Waiting

Creation time

November 30, 2024, 17:46 (UTC+05:30)

Elapsed time

6 minutes, 33 seconds

### 3. Connecting RDS with the EMR instance: -

We configure security group by editing inbound rules

sg-01fd0f3e64d7be90a - ElasticMapReduce-master

Details

Security group name

ElasticMapReduce-master

Security group ID

sg-01fd0f3e64d7be90a

Description

Master group for Elastic MapReduce created on 2024-10-30T16:52:05.614Z

VPC ID

vpc-0155a506d2cb34af6

Owner

891377221307

Inbound rules count

9 Permission entries

Outbound rules count

1 Permission entry

Inbound rules

Outbound rules

Sharing - new

VPC associations - new

Tags

Inbound rules (9)

Search

1

	Name	Security group rule ID	IP version	Type	Protocol	Port range
<input type="checkbox"/>	-	sgr-0804e42e756a1ba85	-	Custom TCP	TCP	8443
<input type="checkbox"/>	-	sgr-08d46cf5c33b8382f	IPv4	SSH	TCP	22
<input type="checkbox"/>	-	sgr-000c59e81a4893eab	-	All ICMP - IPv4	ICMP	All
<input type="checkbox"/>	-	sgr-0fbe3d6fe4bcb4a67	-	All UDP	UDP	0 - 65535
<input type="checkbox"/>	-	sgr-0426d63082750a8dc	IPv4	MySQL/Aurora	TCP	3306

- Then we click on 'Action' button on RDS menu and then 'Set up EC2 connection'

## Set up EC2 connection [Info](#)

### Select EC2 instance

#### Database

[emr-database](#) [↗](#)

#### 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-0262add576b9b2c6d  
- us-east-1b



[Create EC2 instance](#) [↗](#)

[Cancel](#)

[Continue](#)

## Review and confirm

### Connection summary [Info](#)

You are setting up a connection between RDS database [emr-database](#) and EC2 instance [i-0262add576b9b2c6d](#) [↗](#).

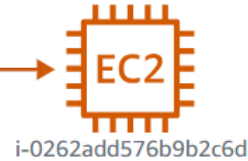
To set up a connection between the database and the EC2 instance, VPC security group **rds-ec2-2** is added to the database, and VPC security group **ec2-rds-2** is added to the EC2 instance.

VPC: vpc-0155a506d2cb34af6 (-)

Security group:  
**rds-ec2-2 (connection rule)**



Security group:  
**ec2-rds-2 (connection rule)**



Bold indicates an addition being made to set up a connection.

### Changes to RDS database: emr-database

Attribute	Current value	New value
Security group	default	default, <b>rds-ec2-2</b>

### Changes to EC2 instance: i-0262add576b9b2c6d

Attribute	Current value	New value
Security group	ElasticMapReduce-master	ElasticMapReduce-master, <b>ec2-rds-2</b>

#### Cross-Availability Zone (AZ) charges might apply

The RDS database emr-database (us-east-1f) and EC2 instance i-0262add576b9b2c6d (us-east-1b) are in different AZs. Cross AZ charges might apply. [Data transfer within same Region](#) [↗](#)

[Cancel](#)

[Previous](#)

[Set up](#)

#### 4. We then log into RDS through EMR instance using the command

'mysql -h emr-upgrad-task.c1e04g20st3d.us-east-1.rds.amazonaws.com -P 3306 -u admin -p'

```
[hadoop@ip-172-31-32-112 ~]$ mysql -h emr-database.c1e04g20st3d.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 28
Server version: 8.0.39 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)]>
```

#### 5. We then create database using code below:

> create database yellow\_taxi;

```
MySQL [(none)]> create database yellow_taxi;
Query OK, 1 row affected (0.007 sec)

MySQL [(none)]> show databases;
+-----+
| Database |
+-----+
| information_schema |
| mysql |
| performance_schema |
| sys |
| yellow_taxi |
+-----+
5 rows in set (0.004 sec)
```

We then create table using code below;

> use yellow\_taxi;

> create table trip\_records (VendorID INT, tpep\_pickup\_datetime VARCHAR(255), tpep\_dropo\_datetime VARCHAR(255), Passenger\_count INT, Trip\_distance FLOAT, RatecodeID INT, store\_and\_fwd\_ag 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 );

```
MySQL [(none)]> use yellow_taxi;
Database changed
MySQL [yellow_taxi]> create table trip_records (VendorID INT, tpep_pickup_datetime VARCHAR(255), tpep_dropo_datetime VARCHAR(255), Passenger_count INT, Trip_distance FLOAT, RatecodeID INT, store_and_fwd_ag 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 );
Query OK, 0 rows affected (0.047 sec)
```

## 6. Downloading required csv files from internet in local using command

```
` 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`
```

```
[hadoop@ip-172-31-32-112 ~]$ wget "https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-06.csv"
--2024-11-30 12:37:48-- https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-06.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 54.230.150.101
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com) | 54.230.150.101 |:443
HTTP request sent, awaiting response... 200 OK
Length: 910028408 (868M) [text/csv]
Saving to: 'yellow_tripdata_2017-06.csv'

yellow_tripdata_2017-06.csv                                100%[=====]
2024-11-30 12:38:16 (31.4 MB/s) - 'yellow_tripdata_2017-06.csv' saved [910028408]
```

## 7. To load data in mysql table we have to login and then run sql command:

```
LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv' INTO TABLE trip_records
FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
```

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

```
SELECT COUNT(*) FROM taxi_records.trip_log;
```

```
MySQL [yellow_taxi]> SELECT COUNT(*) FROM trip_records;
+-----+
| COUNT(*) |
+-----+
| 18880595 |
+-----+
1 row in set (52.842 sec)
```