

1. Creation of RDS instance called database-1

database-1

ModifyActions

Summary

DB identifier database-1	CPU <div>3.47%</div>	Status Available	Class db.t3.micro
Role Instance	Current activity <div>0 Connections</div>	Engine MySQL Community	Region & AZ us-east-1d

Connectivity & security

Monitoring

Logs & events

Configuration

Zero-ETL integrations

Maintenance & backups

Tags

Connectivity & security

Endpoint & port

Endpoint
database-1.c4b5zbadcthl.us-east-1.rds.amazonaws.com

Port
3306

Networking

Availability Zone
us-east-1d

VPC
default (vpc-0af2eb19fdf81d919)

Subnet group

Security

VPC security groups
rds-ec2-11 (sg-0fd84a5e587524efc)
Adding
default (sg-060995c4707e4527e)
Active

Publicly accessible

2. Creation of EMR with Hadoop, Hbase and Scoop

Amazon EMR > EMR on EC2: Clusters > project3

Updated 8 minutes ago

TerminateClone in AWS CLIClone

project3

Summary

Cluster info Cluster ID j-15EMIAU49T2AJ Cluster configuration Instance groups Capacity 1 Primary 0 Core 0 Task	Applications Amazon EMR version emr-6.10.1 Installed applications HBase 2.4.15, Hadoop 3.3.3, Sqoop 1.4.7	Cluster management Log destination in Amazon S3 aws-logs-384930780055-us-east-1/elasticmapreduce Persistent application UIs YARN timeline server Primary node public DNS ec2-52-202-125-37.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 December 06, 2023, 23:05 (UTC+05:30) Elapsed time 14 minutes, 23 seconds
---	--	---	--

3. Connect RDS with EMR

RDS > Databases > Set up EC2 connection

Step 1
Set up EC2 connection

Step 2
Review and confirm

Set up EC2 connection

Select EC2 instance

Database
prodata

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.

Choose an EC2 instance

i-015eaabba804ee1bd
us-east-1c

CancelContinue

Security group: **rds-ec2-6 (connection rule)**

RDS
prodata
Port: 3306

Security group: **ec2-rds-6 (connection rule)**

EC2
i-015eaabba804ee1bd

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

Changes to RDS database: prodata

Attribute	Current value	New value
Security group	default	default, rds-ec2-6

Changes to EC2 instance: i-015eaabba804ee1bd

Attribute	Current value	New value
Security group	ElasticMapReduce-master	ElasticMapReduce-master, ec2-rds-6

⚠ Cross-Availability Zone (AZ) charges might apply

The RDS database prodata (us-east-1b) and EC2 instance i-015eaabba804ee1bd (us-east-1c) are in different AZs. Cross AZ charges might apply. [Data transfer within same Region](#)

Cancel
Previous
Set up

4. Connect emr instance through PuTTY

PuTTY Configuration
✕

Category:

- Session
 - Logging
- Terminal
 - Keyboard
 - Bell
 - Features
- Window
 - Appearance
 - Behaviour
 - Translation
- Selection
 - Colours
- Connection
 - Data
 - Proxy
 - SSH
 - Serial
 - Telnet
 - Rlogin
 - SUPDUP

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address)

Port

!-34-201-169-68.compute-1.amazonaws.com

22

Connection type:

☒ SSH
☐ Serial
☐ Other: Telnet

Load, save or delete a stored session

Saved Sessions

Default Settings

Load
Save
Delete

Close window on exit

☐ Always
☐ Never
☒ Only on clean exit

About
Open
Cancel

5. Login to MYSQL using RDS endpoint

```
hadoop@ip-172-31-36-96:~  
login as: hadoop  
Authenticating with public key "imported-openssh-key"  
Last login: Wed Dec 6 17:47:57 2023  
  
#  
~\##### Amazon Linux 2  
~~\#####  
~~\###| AL2 End of Life is 2025-06-30.  
~~\#/ V~'-'>  
~~  
~~  
~~  
~..-./  
_/_/m/'-./  
  
A newer version of Amazon Linux is available!  
Amazon Linux 2023, GA and supported until 2028-03-15.  
https://aws.amazon.com/linux/amazon-linux-2023/  
  
26 package(s) needed for security, out of 35 available  
Run "sudo yum update" to apply all updates.  
  
EEEEEEEEEEEEEEEEEEEE MMMMMMMM MMMMMMMM RRRRRRRRRRRRRRRR  
E::::::::::::::::::::E M::::::::M M::::::::M R:::::::::R  
EE::::::::EEEEEEEE::::E M::::::::M M::::::::M R::::RRRRRR::::R  
E:::E EEEEE M::::::::M M::::::::M RR:::R R:::R  
E:::E M::::M::M M::M::M R::R R:::R  
E:::EEEEEEEE M::::M M::M M::M M:::M R::RRRRRR::::R  
E::::::::::::E M::::M M::M::M M:::M R:::::::::RR  
E:::EEEEEEEE M::::M M:::M M:::M R::RRRRRR::::R  
E:::E M::::M M::M M:::M R::R R:::R  
E:::E EEEEE M::::M MMM M:::M R::R R:::R  
EE::::::::EEEEEEEE::::E M::::M M:::M R::R R:::R  
E::::::::::::E M::::M M:::M RR:::R R:::R  
EEEEEEEEEEEEEEEEEEEE MMMMMMMM MMMMMMMM RRRRRRR RRRRRR  
  
[hadoop@ip-172-31-36-96 ~]$ mysql -h database-1.c4b5zbadctl.us-east-1.rds.amaz  
naws.com -P 3306 -u admin -p  
Enter password:  
Welcome to the MariaDB monitor. Commands end with ; or \g.  
Your MySQL connection id is 23  
Server version: 8.0.33 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)]> █
```

6. Creation of Database “taxi_records” and table “trip_log”

```
CREATE DATABASE taxi_records;
```

```
USE taxi_records;
```

```
CREATE TABLE trip_log (  
  VendorID INT,  
  tpep_pickup_datetime VARCHAR(50),  
  tpep_dropoff_datetime VARCHAR(50),  
  Passenger_count INT,
```

```

Trip_distance FLOAT,
RateCodeID INT,
store_and_fwd_flag VARCHAR(2),
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 taxi_records;
Database changed
MySQL [taxi_records]> CREATE TABLE trip_log
-> (
-> VendorID INT,
-> tpep_pickup_datetime VARCHAR(50),
-> tpep_dropoff_datetime VARCHAR(50),
-> Passenger_count INT,
-> Trip_distance FLOAT,
-> RatecodeID INT,
-> store_and_fwd_flag VARCHAR(2),
-> 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 (0.04 sec)

MySQL [taxi_records]> █

```

7. Downloading required csv files from internet in local using command

wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv

```
[hadoop@ip-172-31-36-96 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
--2023-12-06 17:56:17-- 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)... 3.5.29.253, 16.182.34.177, 52.216.37.17, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|3.5.29.253|: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 31.5MB/s in 33s

2023-12-06 17:56:50 (26.7 MB/s) - 'yellow_tripdata_2017-01.csv' saved [914029540/914029540]

[hadoop@ip-172-31-36-96 ~]$
```

wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv

```
[hadoop@ip-172-31-36-96 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
--2023-12-06 17:57:35-- 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)... 3.5.28.181, 16.182.67.9, 52.216.152.164, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|3.5.28.181|: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 26.8MB/s in 31s

2023-12-06 17:58:06 (26.4 MB/s) - 'yellow_tripdata_2017-02.csv' saved [863487050/863487050]

[hadoop@ip-172-31-36-96 ~]$
```

8. 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_log
FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
```

```
[hadoop@ip-172-31-36-96 ~]$ mysql -h database-1.c4b5zbdcchl.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 26
Server version: 8.0.33 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)]> use taxi_records;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MySQL [taxi_records]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv' INTO TABLE trip_log
-> FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9710820 rows affected, 65535 warnings (2 min 47.94 sec)
Records: 9710820 Deleted: 0 Skipped: 0 Warnings: 9710820

MySQL [taxi_records]>
```

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

```
MySQL [taxi_records]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv' INTO TABLE trip_log
-> FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (2 min 49.17 sec)
Records: 9169775 Deleted: 0 Skipped: 0 Warnings: 9169775

MySQL [taxi_records]>
```

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

```
MySQL [taxi_records]> SELECT COUNT(*) FROM taxi_records.trip_log;
+-----+
| COUNT(*) |
+-----+
| 18880595 |
+-----+
1 row in set (51.84 sec)

MySQL [taxi_records]> █
```