

TASK 01

Create an RDS instance in your AWS account and upload the data from two files (yellow_tripdata_2017-01.csv & yellow_tripdata_2017-02.csv) from the dataset. Make sure to create an appropriate schema for the data sets before uploading them to RDS.

A. Create a RDS instance in AWS

The screenshot shows the AWS Management Console for an RDS instance named 'mrdata01'. The instance is in the 'Available' state. The summary section displays the following details:

Summary	Status	Role	Engine	Recommendations
DB identifier: mrdata01	Available	Instance	MySQL Community	
CPU: 4.58%	Class: db.t4g.micro	Current activity: 0 Connections	Region & AZ: us-east-1b	

The 'Connectivity & security' tab is selected, showing the following details:

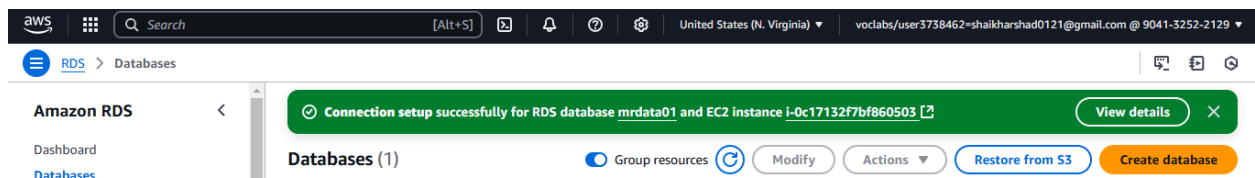
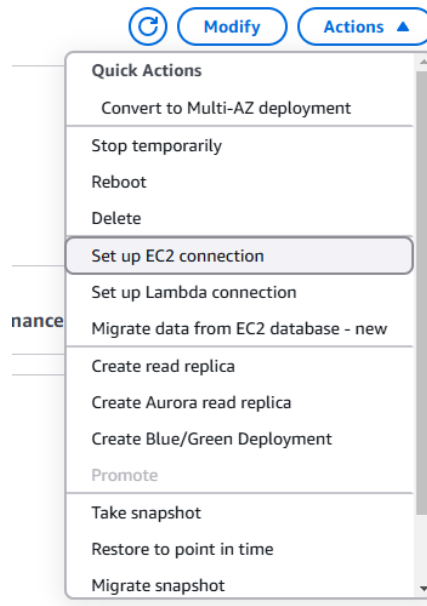
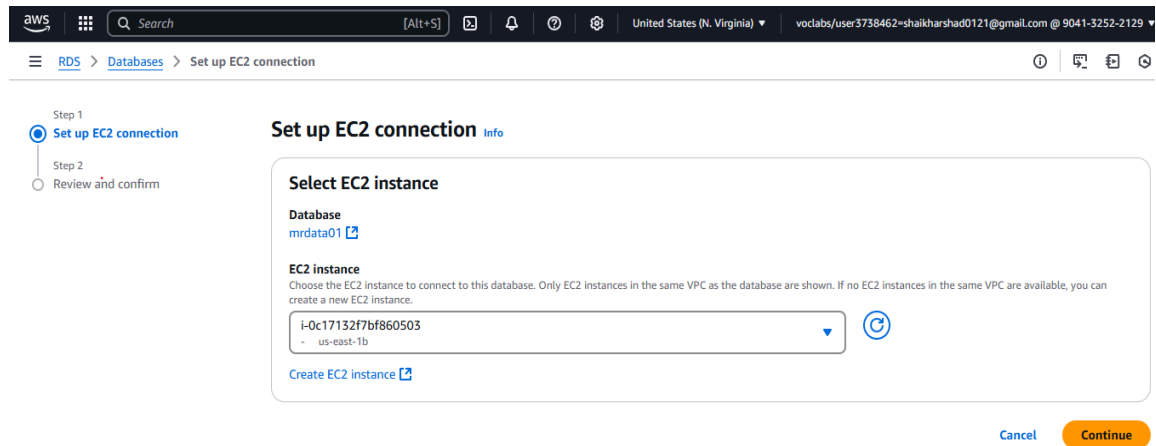
Connectivity & security	Networking	Security
Endpoint & port: Endpoint: mrdata01.cske577dgh1r.us-east-1.rds.amazonaws.com, Port: 3306	Availability Zone: us-east-1b, VPC: project-vpc (vpc-020acbc60c2ba109f)	VPC security groups: default (sg-03bc69d1091601933), VPC security groups: Active, Publicly accessible: Yes

B. Create EMR instance with Hadoop, Apache Sqoop and Apache HBase

The screenshot shows the AWS Management Console for an EMR instance named 'MRAssignmnetCluster02'. The instance is in the 'Waiting' state. The summary section displays the following details:

Summary	Applications	Cluster management	Status and time
Cluster info: Cluster ID: j-3EW74S18C9UOU, Cluster ARN: arn:aws:elasticmapreduce:us-east-1:904132522129:cluster/j-3EW74S18C9UOU, Cluster configuration: Instance groups, Capacity: 1 Primary 0 Core 0 Task	Amazon EMR version: emr-6.15.0, Installed applications: HBase 2.4.17, Hadoop 3.3.6, Sqoop 1.4.7	Log destination in Amazon S3: aws-logs-904132522129-us-east-1/elasticmapreduce, Persistent application Uls: YARN timeline server, Primary node public DNS: ec2-107-22-138-211.compute-1.amazonaws.com, Connect to the Primary node using SSH, Connect to the Primary node using SSM	Status: Waiting, Creation time: February 02, 2025, 15:48 (UTC+05:30), Elapsed time: 4 minutes, 35 seconds

C. After creating a RDS instance, Click Action → Set up EC2 connection. Also make sure to have MySQL added in the security groups.



D. Login to the terminal using putty.

```

hadoop@ip-10-0-25-185:~
login as: hadoop
Authenticating with public key "imported-openssh-key"

#
##### Amazon Linux 2
#####
##### AL2 End of Life is 2026-06-30.
#####
##### 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/

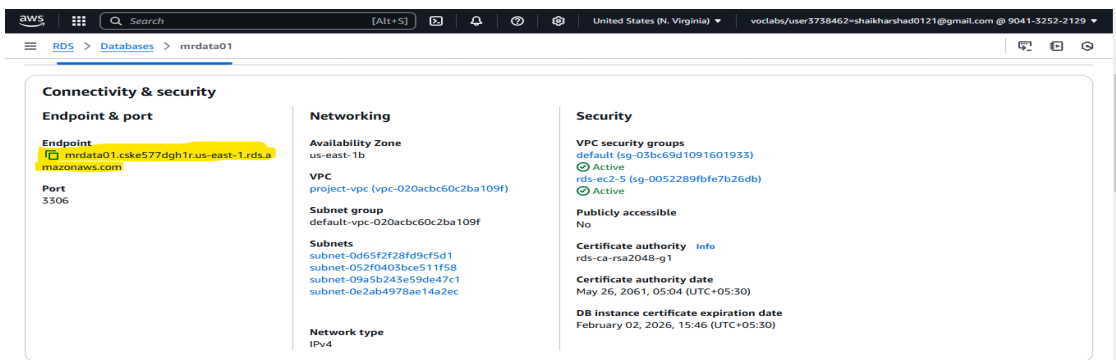
EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M M::::::::M R:::::::::R
EE::::::::EEEEEEEE::::E M::::::::M M::::::::M R:::::::::R
E::::E EEEEE M::::::::M M::::::::M RR::::R R::::R
E::::E M::::::::M M::M::M M::M::M R::R R::::R
E::::::::EEEEEEEE M::::::::M M::M::M M::M::M R::RRRRRR:::R
E::::::::::::E M::M::M M::M::M M::M::M R:::::::::RR
E::::::::EEEEEEEE M::M::M M::M::M M::M::M R::RRRRRR:::R
E::::E M::M::M M::M M::M::M R::R R::::R
E::::E EEEEE M::M::M MMM M::M::M R::R R::::R
EE::::::::EEEEEEEE::::E M::M::M M::M::M R::R R::::R
E::::::::::::E M::M::M M::M::M RR::::R R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRR RRRRRR

[hadoop@ip-10-0-25-185 ~]$

```

E. Use the following syntax to connect to MySql:

`mysql -h RDS Endpoint -P 3306 -u admin -p`



```

[hadoop@ip-10-0-25-185 ~]$ mysql -h mrdata01.cske577dgh1r.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 27
Server version: 8.0.40 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)]>

```

F. Create the database and the table schema

```
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
```

```
MySQL [(none)]> CREATE DATABASE ride_data;  
Query OK, 1 row affected (0.01 sec)
```

```
MySQL [(none)]> USE ride_data;  
Database changed
```

```
MySQL [ride_data]> CREATE TABLE ride_log (  
-> VendorID INT,  
-> pickup_datetime VARCHAR(50),  
-> extra FLOAT,  
-> mta_tax FLOAT,  
-> 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 [ride_data]> █
```

G. Get both the dataset.

```
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
```

```
MySQL [ride_data]> exit;  
Bye  
[hadoop@ip-10-0-25-185 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv  
--2025-02-02 10:30:15-- 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)... 16.15.177.5, 3.5.29.41, 52.216.213.129, ...  
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|16.15.177.5|: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 36.3MB/s in 24s  
  
2025-02-02 10:30:40 (36.2 MB/s) - 'yellow_tripdata_2017-01.csv' saved [914029540/914029540]  
  
[hadoop@ip-10-0-25-185 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv  
--2025-02-02 10:30:43-- 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)... 3.5.7.144, 52.216.53.57, 52.216.216.1, ...  
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|3.5.7.144|: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 39.5MB/s in 22s  
  
2025-02-02 10:31:05 (38.0 MB/s) - 'yellow_tripdata_2017-02.csv' saved [863487050/863487050]  
[hadoop@ip-10-0-25-185 ~]$ █
```

H. Login to MySQL again and write the following syntax to load the data:

```
LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv'
```

```
INTO TABLE ride_log
```

```
FIELDS TERMINATED BY ','
```

```
ENCLOSED BY ''
```

```
LINES TERMINATED BY '\n'
IGNORE 1 LINES;
```

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

Once this syntax run the data is loaded into our RDS table.

```
[hadoop@ip-10-0-25-185 ~]$ mysql -h mrdatab01.cske577dghlr.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: 8.0.40 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 ride_data;
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 [ride_data]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv'
-> INTO TABLE ride_log
-> FIELDS TERMINATED BY ','
-> ENCLOSED BY '"'
-> LINES TERMINATED BY '\n'
-> IGNORE 1 LINES;
Query OK, 9710820 rows affected, 65535 warnings (2 min 10.64 sec)
Records: 9710820  Deleted: 0  Skipped: 0  Warnings: 9710820

MySQL [ride_data]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv'
-> INTO TABLE ride_log
-> FIELDS TERMINATED BY ','
-> ENCLOSED BY '"'
-> LINES TERMINATED BY '\n'
-> IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (1 min 56.54 sec)
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 9169775
```

Records: 9169775 Deleted: 0 Skipped: 0 Warnings: 9169775

MySQL [ride_data]> SELECT COUNT(*) FROM ride_log;

```
+-----+
| COUNT(*) |
+-----+
| 18880595 |
+-----+
1 row in set (54.52 sec)
```

MySQL [ride_data]>

MySQL [ride_data]> SELECT * FROM ride_log LIMIT 10;

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| VendorID | pickup_datetime | dropoff_datetime | Passenger_count | Trip_distance | RatecodeID | store_and_fwd_flag | PULocationID | DOLocationID | payment_type |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 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 | 237 | 140 | 2 |
| 1 | 2017-01-01 00:43:25 | 2017-01-01 00:47:42 | 2 | 0.7 | 1 | N | 140 | 237 | 2 |
| 5 | 0.5 | 0.5 | 0 | 0 | 0.3 | 6.3 | 0 | 140 | 237 | 2 |
| 1 | 2017-01-01 00:49:10 | 2017-01-01 00:53:53 | 2 | 0.8 | 1 | N | 140 | 237 | 2 |
| 5.5 | 0.5 | 0.5 | 0 | 0 | 0.3 | 6.8 | 0 | 41 | 42 | 2 |
| 1 | 2017-01-01 00:36:42 | 2017-01-01 00:41:09 | 1 | 1.1 | 1 | N | 41 | 42 | 2 |
| 6 | 0.5 | 0.5 | 0 | 0 | 0.3 | 7.3 | 0 | 48 | 263 | 2 |
| 1 | 2017-01-01 00:07:41 | 2017-01-01 00:18:16 | 1 | 3 | 1 | N | 48 | 263 | 2 |
| 11 | 0.5 | 0.5 | 0 | 0 | 0.3 | 12.3 | 0 | 236 | 262 | 2 |
| 1 | 2017-01-01 00:20:52 | 2017-01-01 00:24:59 | 2 | 0.7 | 1 | N | 236 | 262 | 2 |
| 5 | 0.5 | 0.5 | 0 | 0 | 0.3 | 6.3 | 0 | 236 | 238 | 1 |
| 1 | 2017-01-01 00:33:49 | 2017-01-01 00:42:38 | 2 | 1.6 | 1 | N | 236 | 238 | 1 |
| 8 | 0.5 | 0.5 | 1.85 | 0 | 0.3 | 11.15 | 0 | 238 | 239 | 1 |
| 1 | 2017-01-01 00:48:22 | 2017-01-01 00:52:15 | 2 | 0.6 | 1 | N | 238 | 239 | 1 |
| 5 | 0.5 | 0.5 | 1.25 | 0 | 0.3 | 7.55 | 0 | 239 | 48 | 1 |
| 1 | 2017-01-01 00:57:12 | 2017-01-01 01:06:28 | 2 | 1 | 1 | N | 239 | 48 | 1 |
| 7.5 | 0.5 | 0.5 | 1.75 | 0 | 0.3 | 10.55 | 0 | 246 | 48 | 2 |
| 1 | 2017-01-01 00:10:25 | 2017-01-01 00:29:06 | 1 | 1 | 1 | N | 246 | 48 | 2 |
| 12 | 0.5 | 0.5 | 0 | 0 | 0.3 | 13.3 | 0 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
10 rows in set (0.08 sec)
```

MySQL [ride_data]> █