

1. RDS Instance created on AWS :

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

DB identifier	CPU	Status	Class
mapr-assignment-database	3.39%	Available	db.t3.micro

Additional details include:

Role	Current activity	Engine	Region & AZ
Instance	0 Connections	MySQL Community	us-east-1f

The 'Connectivity & security' section shows the endpoint 'mapr-assignment-database.ctpdcmr9bof2.us-east-1.rds.amazonaws.com', the availability zone 'us-east-1f', and the VPC 'vpc-0f5ddde5998f6d1c4'. The security group is 'default (sg-055c97699ef5b6830)' and is 'Active'. The instance is 'Publicly accessible'.

2. EMR Instance created with Application as Hbase, Hadoop, Sqoop etc.

The screenshot displays the AWS Management Console for an Amazon EMR cluster. The cluster is named 'mapr_assignment' and is in the 'Waiting' state. The summary section shows the following details:

Cluster info	Applications	Cluster management	Status and time
<p>Cluster ID: j-3TO55Z1XQJIR</p> <p>Cluster configuration: Instance groups</p> <p>Capacity: 1 Primary 0 Core 0 Task</p>	<p>Amazon EMR version: emr-5.30.1</p> <p>Installed applications: HBase 1.4.13, Hadoop 2.8.5, Hive 2.3.6, Hue 4.6.0, Pig 0.17.0, Spark 2.4.5, Sqoop 1.4.7</p>	<p>Log destination in Amazon S3: Logging not configured</p> <p>Persistent application UIs: Spark History Server, YARN timeline server, Tez UI</p> <p>Primary node public DNS: ec2-44-203-238-44.compute-1.amazonaws.com</p> <p>Connect to the Primary Node using SSH</p>	<p>Status: Waiting</p> <p>Creation time: August 06, 2023, 22:03 (UTC+05:30)</p> <p>Elapsed time: 14 minutes</p>

The 'Cluster logs' section shows 'Archive log files to Amazon S3' and 'Encryption for logs'. The 'Cluster termination' section shows 'Termination option' and 'Termination protection'.

3. Open EMR Instance on Terminal, Create a new Directory and Download NYC data file in new directory using below command:

```
// to make new directory
mkdir mapr_assignment
```

```
// move to new directory
cd mapr_assignment
```

```
//download NYC data files
```

```
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-0{1..2}.csv
```

[illegible]

```
Script Editor  File  Edit  View  Script  Font  Format  Window  Help  zoom  Sun 6 Aug 10:21 PM
Downloads — ec2-user@ip-172-31-13-99:~/mapr_assignment — ssh -i mac-academy-key.pem ec2-user@44.203.238.44 — 204x55

  _ _ _ _ _
 _/ _ _ _ _ \
 _/ _ _ _ _ \
 _/ _ _ _ _ \

Amazon Linux 2 AMI

https://aws.amazon.com/amazon-linux-2/
86 package(s) needed for security, out of 154 available
Run "sudo yum update" to apply all updates.
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory

EEEEEEEEEEEEEEEEEEEE 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 R::::R R::::R
E::::EEEEEEEEEE M::::::::M M::M::M M::M::M R::RRRRR::::R
E::::::::::::::::::::E M::::::::M M::M::M M::M::M R:::::::::R
E::::EEEEEEEEEE M::::::::M M::M::M M::M::M R::RRRRR::::R
E::::E M::::::::M M::M M::M::M R::::R R::::R
E::::E EEEEE M::::::::M M::M M::M::M R::::R R::::R
EE::::::::EEEEEEEE::::E M::::::::M M::M::M R::::R R::::R
E::::::::::::::::::::E M::::::::M M::M::M RR::::::::R R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM RRRRRRR RRRRRR

[ec2-user@ip-172-31-13-99 ~]$ pwd
/home/ec2-user
[ec2-user@ip-172-31-13-99 ~]$ ls
[ec2-user@ip-172-31-13-99 ~]$ mkdir mapr_assignment
[ec2-user@ip-172-31-13-99 ~]$ cd mapr_assignment/
[ec2-user@ip-172-31-13-99 mapr_assignment]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-0(1..2).csv
--2023-08-06 16:50:18-- 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)... 3.5.3.22, 16.182.97.57, 52.216.49.161, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)[3.5.3.22]: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 37.4MB/s in 25s

2023-08-06 16:50:35 (35.1 MB/s) - 'yellow_tripdata_2017-01.csv' saved [914029540/914029540]

--2023-08-06 16:50:35-- https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
Reusing existing connection to nyc-tlc-upgrad.s3.amazonaws.com:443.
HTTP request sent, awaiting response... 200 OK
Length: 863487050 (823M) [text/csv]
Saving to: 'yellow_tripdata_2017-02.csv'

100%[=====] 863,487,050 34.3MB/s in 22s

2023-08-06 16:50:58 (36.8 MB/s) - 'yellow_tripdata_2017-02.csv' saved [863487050/863487050]

FINISHED --2023-08-06 16:50:58--
Total wall clock time: 48s
Downloaded: 2 files, 1.7G in 47s (35.9 MB/s)
[ec2-user@ip-172-31-13-99 mapr_assignment]$
```

Note : Above command will download first 2 files sequentially.

4. Connect to RDS on EMR instance through terminal using below command:

```
//Connect RDS on EMR
```

```
mysql -h mapr-assignment-database.ctpdcmc9bof2.us-east-1.rds.amazonaws.com -P 3306 -u
admin -p
```

```
//to list database
show databases;
```

```
//to create database
create database <database_name>
```

```
//to go to db
use <database_name>;
```

```
//to list tables
show tables;
```

```
Terminal Shell Edit View Window Help
Downloads — ec2-user@ip-172-31-13-99:~/mapr_assignment — ssh -i mac-academy-key.pem ec2-user@44.203.238.44 — 204x55

EEEEEEEEEEEEEEEEEEEE MMMMMMMM RRRRRR RRRRRR

[ec2-user@ip-172-31-13-99 ~]$ pwd
/home/ec2-user
[ec2-user@ip-172-31-13-99 ~]$ ls
[ec2-user@ip-172-31-13-99 ~]$ mkdir mapr_assignment
[ec2-user@ip-172-31-13-99 ~]$ cd mapr_assignment/
[ec2-user@ip-172-31-13-99 mapr_assignment]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
--2023-08-06 16:58:10-- 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)... 3.5.3.22, 16.182.97.57, 52.216.49.161, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|3.5.3.22|: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 37.4MB/s in 25s

2023-08-06 16:58:35 (35.1 MB/s) - 'yellow_tripdata_2017-01.csv' saved [914029540/914029540]

--2023-08-06 16:58:35-- https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
Reusing existing connection to nyc-tlc-upgrad.s3.amazonaws.com:443.
HTTP request sent, awaiting response... 200 OK
Length: 863487050 (823M) [text/csv]
Saving to: 'yellow_tripdata_2017-02.csv'

100%[=====] 863,487,050 34.3MB/s in 22s

2023-08-06 16:58:58 (36.8 MB/s) - 'yellow_tripdata_2017-02.csv' saved [863487050/863487050]

FINISHED --2023-08-06 16:58:58--
Total wall clock time: 48s
Downloaded: 2 files, 1.70 in 47s (35.9 MB/s)
[ec2-user@ip-172-31-13-99 mapr_assignment]$ mysql -h mapr-assignment-database.ctdcmc9bof2.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 77.
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)]> show databases;
+-----+
| Database |
+-----+
| information_schema |
| mapr |
| mysql |
| performance_schema |
| sys |
+-----+
5 rows in set (0.03 sec)

MySQL [(none)]>
```

5. Create New table with name NYC_TRIPS

create table **NYC_TRIPS**

```
(
  VendorID INT,
  tpep_pickup_datetime DATETIME,
  tpep_dropoff_datetime DATETIME,
  passenger_count INT,
  trip_distance FLOAT(10,2),
  RatecodeID INT,
  store_and_fwd_flag VARCHAR(1),
  PULocationID VARCHAR(50),
  DOLocationID VARCHAR(50),
  payment_type INT,
  fare_amount FLOAT(10,2),
  extra FLOAT(10,2),
  mta_tax FLOAT(10,2),
  tip_amount FLOAT(10,2),
  tolls_amount FLOAT(10,2),
  improvement_surcharge FLOAT(10,2),
  total_amount FLOAT(10,2),
  Airport_fee FLOAT(10,2)
);
```

```
[ec2-user@ip-172-31-13-99 mapr_assignment]$ mysql -h mapr-assignment-database.ctdcm9bof2.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 77
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)]> show databases;
+-----+
| Database |
+-----+
| information_schema |
| mapr      |
| mysql     |
| performance_schema |
| sys       |
+-----+
5 rows in set (0.03 sec)

MySQL [(none)]> drop database mapr;
Query OK, 1 row affected (1.55 sec)

MySQL [(none)]> create database maprdb;
Query OK, 1 row affected (0.01 sec)

MySQL [(none)]> use maprdb;
Database changed
MySQL [maprdb]> show tables;
Empty set (0.00 sec)

MySQL [maprdb]> create table NYC_TRIPS
(
  VendorID INT,
  tpep_pickup_datetime DATETIME,
  tpep_dropoff_datetime DATETIME,
  passenger_count INT,
  trip_distance FLOAT(10,2),
  RatecodeID INT,
  store_and_fwd_flag VARCHAR(1),
  PULocationID VARCHAR(50),
  DOLocationID VARCHAR(50),
  payment_type INT,
  fare_amount FLOAT(10,2),
  extra FLOAT(10,2),
  mta_tax FLOAT(10,2),
  tip_amount FLOAT(10,2),
  tolls_amount FLOAT(10,2),
  improvement_surcharge FLOAT(10,2),
  total_amount FLOAT(10,2),
  Airport_fee FLOAT(10,2)
);
Query OK, 0 rows affected, 9 warnings (0.07 sec)
```

6. Load Data from File 1 and File 2 to RDS Database

LOAD DATA LOCAL INFILE '/home/ec2-user/mapr_assignment/yellow_tripdata_2017-01.csv'
INTO TABLE NYC_TRIPS FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;

LOAD DATA LOCAL INFILE '/home/ec2-user/mapr_assignment/yellow_tripdata_2017-02.csv'
INTO TABLE NYC_TRIPS FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;

//check data in database

Select * from NYC_TRIPS LIMIT 10;

Select count(*) from NYC_TRIPS;

```
Terminal Shell Edit View Window Help
Downloads — ec2-user@ip-172-31-13-99:~/mapr_assignment — ssh -i mac-academy-key.pem ec2-user@44.203.238.44 — 204x55

| NYC_TRIPS |
+-----+
1 row in set (0.01 sec)

MySQL [maprdb]> select * from NYC_TRIPS;
Empty set (0.02 sec)

MySQL [maprdb]> LOAD DATA LOCAL INFILE '/home/ec2-user/mapr_assignment/yellow_tripdata_2017-01.csv' INTO TABLE NYC_TRIPS FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9718820 rows affected, 65535 warnings (2 min 15.86 sec)
Records: 9718820 Deleted: 0 Skipped: 0 Warnings: 9718820

MySQL [maprdb]> LOAD DATA LOCAL INFILE '/home/ec2-user/mapr_assignment/yellow_tripdata_2017-02.csv' INTO TABLE NYC_TRIPS FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (2 min 25.22 sec)
Records: 9169775 Deleted: 0 Skipped: 0 Warnings: 9169775

MySQL [maprdb]> select * from NYC_TRIPS 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 |
+-----+
| 1 | 2017-01-01 00:32:05 | 2017-01-01 00:37:48 | 1 | 1.20 | 1 | N | 140 | 236 | 2 | 6.50 | 0.50 | 0.50 | |
| 0.00 | 0.00 | 0.30 | 7.80 | 0.00 | 1.20 | 1 | N | 140 | 236 | 2 | 6.50 | 0.50 | 0.50 |
| 1 | 2017-01-01 00:43:25 | 2017-01-01 00:47:42 | 2 | 0.70 | 1 | N | 237 | 140 | 2 | 5.00 | 0.50 | 0.50 |
| 0.00 | 0.00 | 0.30 | 6.30 | 0.00 | 0.70 | 1 | N | 237 | 140 | 2 | 5.00 | 0.50 | 0.50 |
| 1 | 2017-01-01 00:49:10 | 2017-01-01 00:53:53 | 2 | 0.80 | 1 | N | 140 | 237 | 2 | 5.50 | 0.50 | 0.50 |
| 0.00 | 0.00 | 0.30 | 6.80 | 0.00 | 0.80 | 1 | N | 140 | 237 | 2 | 5.50 | 0.50 | 0.50 |
| 1 | 2017-01-01 00:36:42 | 2017-01-01 00:41:09 | 1 | 1.10 | 1 | N | 41 | 42 | 2 | 6.00 | 0.50 | 0.50 |
| 0.00 | 0.00 | 0.30 | 7.30 | 0.00 | 1.10 | 1 | N | 41 | 42 | 2 | 6.00 | 0.50 | 0.50 |
| 1 | 2017-01-01 00:07:41 | 2017-01-01 00:18:16 | 1 | 3.00 | 1 | N | 48 | 263 | 2 | 11.00 | 0.50 | 0.50 |
| 0.00 | 0.00 | 0.30 | 12.30 | 0.00 | 3.00 | 1 | N | 48 | 263 | 2 | 11.00 | 0.50 | 0.50 |
| 1 | 2017-01-01 00:20:52 | 2017-01-01 00:24:59 | 2 | 0.70 | 1 | N | 236 | 262 | 2 | 5.00 | 0.50 | 0.50 |
| 0.00 | 0.00 | 0.30 | 6.30 | 0.00 | 0.70 | 1 | N | 236 | 262 | 2 | 5.00 | 0.50 | 0.50 |
| 1 | 2017-01-01 00:33:49 | 2017-01-01 00:42:38 | 2 | 1.60 | 1 | N | 236 | 238 | 1 | 8.00 | 0.50 | 0.50 |
| 1.05 | 0.00 | 0.30 | 11.15 | 0.00 | 1.60 | 1 | N | 236 | 238 | 1 | 8.00 | 0.50 | 0.50 |
| 1 | 2017-01-01 00:48:22 | 2017-01-01 00:52:15 | 2 | 0.60 | 1 | N | 238 | 239 | 1 | 5.00 | 0.50 | 0.50 |
| 1.25 | 0.00 | 0.30 | 7.55 | 0.00 | 0.60 | 1 | N | 238 | 239 | 1 | 5.00 | 0.50 | 0.50 |
| 1 | 2017-01-01 00:57:12 | 2017-01-01 01:06:28 | 2 | 1.00 | 1 | N | 239 | 48 | 1 | 7.50 | 0.50 | 0.50 |
| 1.75 | 0.00 | 0.30 | 10.55 | 0.00 | 1.00 | 1 | N | 239 | 48 | 1 | 7.50 | 0.50 | 0.50 |
| 1 | 2017-01-01 00:10:25 | 2017-01-01 00:29:06 | 1 | 1.00 | 1 | N | 246 | 48 | 2 | 12.00 | 0.50 | 0.50 |
| 0.00 | 0.00 | 0.30 | 13.30 | 0.00 | 1.00 | 1 | N | 246 | 48 | 2 | 12.00 | 0.50 | 0.50 |
+-----+
10 rows in set (0.01 sec)

MySQL [maprdb]> select count(*) from NYC_TRIPS;
+-----+
| count(*) |
+-----+
| 18880595 |
+-----+
1 row in set (38.54 sec)

MySQL [maprdb]>
```