

1. Creation of RDS instance in AWS

The screenshot shows the AWS Management Console interface for Amazon RDS. The left sidebar contains the navigation menu with options like Dashboard, Databases, Query Editor, Performance insights, Snapshots, Exports in Amazon S3, Automated backups, Reserved instances, Proxies, Subnet groups, Parameter groups, Option groups, and Custom engine versions. The main content area is titled 'Databases (1)' and includes a 'Group resources' toggle, 'Modify', 'Actions', 'Restore from S3', and 'Create database' buttons. A search bar is present with the text 'Filter by databases'. Below this is a table listing the database instances.

DB identifier	Status	Role	Engine	Region & AZ	Size
mapreduce-assignment	Available	Instance	MySQL Community	us-east-1b	db.t

At the bottom of the console, there is a footer with 'CloudShell', 'Feedback', and copyright information: '© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences'.

aws

Services

Search

[Alt+S]

N. Virginia

voclabs/user3206019-darshuj26@gmail.com @ 6703-9503-1243

Amazon RDS

Dashboard

Databases

Query Editor

Performance insights

Snapshots

Exports in Amazon S3

Automated backups

Reserved instances

Proxies

Subnet groups

Parameter groups

Option groups

Custom engine versions

CloudShell

Feedback

RDS > Databases > mapreduce-assignment

mapreduce-assignment

Summary

DB Identifier
mapreduce-assignment
CPU
4.19%

Status
Available
Class
db.t3.micro

Role
Instance
Current activity
0
Connections

Engine
MySQL
Community
Region & AZ
us-east-1b

Recommendations

Connectivity & securityMonitoringLogs & eventsConfigurationZero-ETL integrations

Connectivity & security

Endpoint & port

Networking

Security

Endpoint
mapreduce-assignment.cmuimgzbr8mc.us-east-1.rds.amazonaws.com
Port
3306

Availability Zone
us-east-1b
VPC
vpc-02395df4c29aaf0f5
Subnet group
default-vpc-02395df4c29aaf0f5
Subnets
subnet-09230259f2505a4b9
subnet-01967024df5ea2c43
subnet-0ba2c21c434bcd9aa
subnet-076b55fa8da24fe9e
subnet-0938cb16c0924ce0b
subnet-004f5281ca8dc8338
Network type
IPv4

VPC security groups
default (sg-0e31844a7cc53df24)
Active
Publicly accessible
No
Certificate authority
rds-ca-rsa2048-g1
Certificate authority date
May 26, 2061, 05:04 (UTC+05:30)
DB instance certificate expiration date
May 10, 2025, 21:01 (UTC+05:30)

Connected compute resources (0)

Filter by compute resources

No connected compute resources

Set up EC2 connectionSet up Lambda connection

Proxies (0)

Filter by proxies

No proxies

Create proxy

Security group rules (2)

Filter by Security group rules

Security group	Type	Rule
default (sg-0e31844a7cc53df24)	EC2 Security Group - Inbound	sg-0e31844a7cc53df24
default (sg-0e31844a7cc53df24)	CIDR/IP - Outbound	0.0.0.0/0

Replication (1)

Filter by Replication

DB identifier	Role	Region & AZ	Replication source	Replication state	Lag
mapreduce-assignment	Instance	us-east-1b	-	-	-

2. Creation of EMR instance with bundled applications such as Hadoop, Hbase, Hue, Hive, Sqoop.

The screenshot displays the AWS Management Console interface for an Amazon EMR mapReduce Assignment. The top navigation bar includes the AWS logo, a search bar, and user information. The breadcrumb trail shows the path: Amazon EMR > EMR on EC2: Clusters > mapReduce Assignment. The main heading is "mapReduce Assignment", with a timestamp "Updated less than a minute ago" and action buttons: "Terminate", "Clone in AWS CLI", and "Clone".

The "Summary" section is divided into four columns:

- Cluster info:** Cluster ID is j-3RFJOSAI95WU7. Cluster configuration includes Instance groups. Capacity is 1 Primary, 0 Core, and 0 Task.
- Applications:** Amazon EMR version is emr-6.10.0. Installed applications include HBase 2.4.15, Hadoop 3.3.3, Hive 3.1.3, Hue 4.10.0, and Sqoop 1.4.7.
- Cluster management:** Log destination in Amazon S3 is "Logging not configured". Persistent application UIs include "YARN timeline server" and "Tez UI". Primary node public DNS is ec2-3-89-85-109.compute-1.amazonaws.com. Links for "Connect to the Primary node using SSH" and "Connect to the Primary node using" are provided.
- Status and time:** Status is "Waiting". Creation time is May 10, 2024, 20:25 (UTC+05:30). Elapsed time is 17 minutes, 43 seconds.

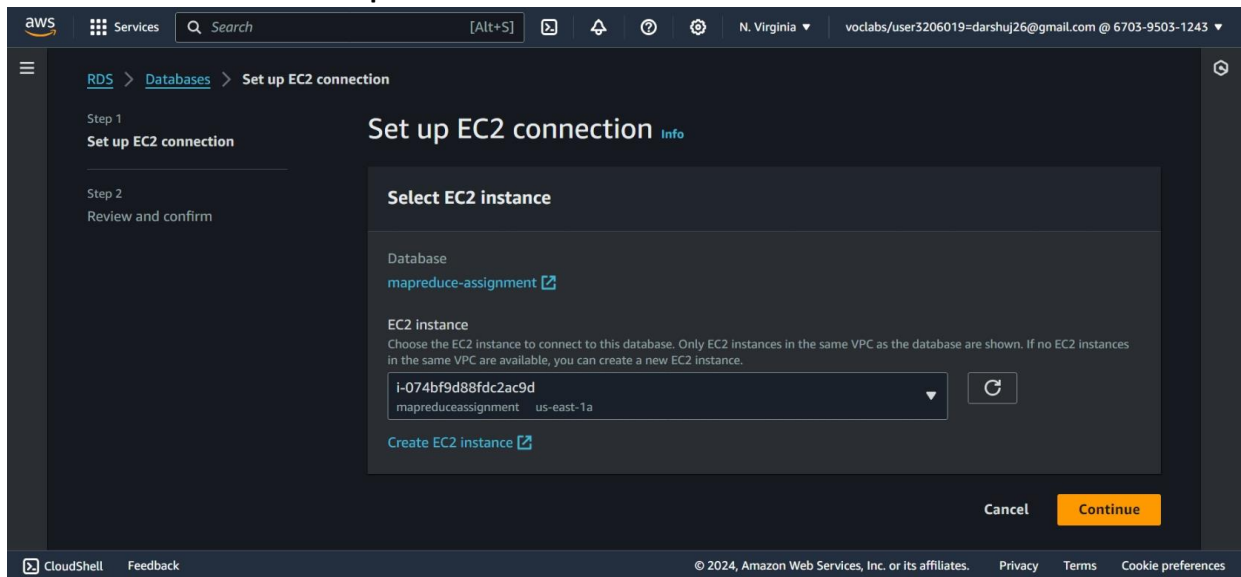
The bottom section contains tabs for "Properties", "Bootstrap actions", "Instances (Hardware)", "Steps", "Applications", "Configurations", "Monitoring", "Events", and "Tags (0)". The "Properties" tab is active, showing three main sections:

- Operating system:** Amazon Linux release 2.0.20240329.0.
- Cluster logs:** Archive log files to Amazon S3 is "Turned off". Encryption for logs is "Turned off".
- Cluster termination and node replacement:** Termination option is "Automatically terminate cluster after idle time". Idle time is 3 hours. Termination protection is "Off". Unhealthy node replacement is "On".

The "Network and security" section is also visible, containing three sub-sections:

- Network:** Virtual Private Cloud (VPC) is vpc-0e69174daeb4f6d89. Subnet(s) and Availability Zone(s) (AZ) is subnet-06d27a03b82db1ae6 in us-east-1a. A link for "EC2 security groups (firewall)" is provided.
- Security configuration:** Security configuration is "None". EC2 key pair is akashDell.
- Permissions:** Service role for Amazon EMR is EMR_DefaultRole. EC2 instance profile is EMR_EC2_DefaultRole. Custom automatic scaling role is EMR_AutoScaling_DefaultRole.

3. To connect RDS with EMR instance, we have to click on “Action” button on RDS instance menu and then “Set up an EC2 Instance”.



4. Login to RDS through EMR instance using command: ``mysql -h demodb.cluster-cjd8k4oc9drm.ap-south-1.rds.amazonaws.com -P 3306 -u admin -p``

```
hadoop@ip-172-31-14-82:~  
[hadoop@ip-172-31-14-82 ~]$ mysql -h mapreducedb.cjd8k4oc9drm.ap-south-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 20  
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 |  
| mysql |  
| performance_schema |  
| sys |  
+-----+  
4 rows in set (0.01 sec)  
  
MySQL [(none)]>
```

5. Create Database “taxi_records”

```
`create database taxi_records'
```

6. Create table "trip_log"

```
CREATE TABLE trip_log (  
    trip_id INT,  
    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  
);
```

```
hadoop@ip-172-31-14-82:~  
+-----+  
| information_schema |  
| mysql              |  
| performance_schema |  
| sys                |  
+-----+  
4 rows in set (0.01 sec)  
  
MySQL [(none)]> create database taxi_records  
-> ;  
Query OK, 1 row affected (0.01 sec)  
  
MySQL [(none)]> use taxi_records  
Database changed  
MySQL [taxi_records]>  
MySQL [taxi_records]> CREATE TABLE trip_log (  
->     trip_id INT ,  
->     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,
```

7. 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
```

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;
```

```
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;
```

```
hadoop@ip-172-31-14-82:~  
[hadoop@ip-172-31-14-82 ~]$ mysql -h mapreducedb.cjd8k4oc9drm.ap-south-1.rds.ama  
zonaws.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: 8.0.33 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 28.03 sec)  
Records: 9710820 Deleted: 0 Skipped: 0 Warnings: 46808699  
  
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 44.88 sec)  
Records: 9169775 Deleted: 0 Skipped: 0 Warnings: 44172593
```

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

hadoop@ip-172-31-14-82:~

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 28.03 sec)

Records: 9710820 Deleted: 0 Skipped: 0 Warnings: 46808699

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 44.88 sec)

Records: 9169775 Deleted: 0 Skipped: 0 Warnings: 44172593

MySQL [taxi_records]> SELECT COUNT(*) FROM taxi_records.trip_log;

+-----+
| COUNT(*) |

+-----+
| 18880595 |

+-----+

1 row in set (47.62 sec)

MySQL [taxi_records]>