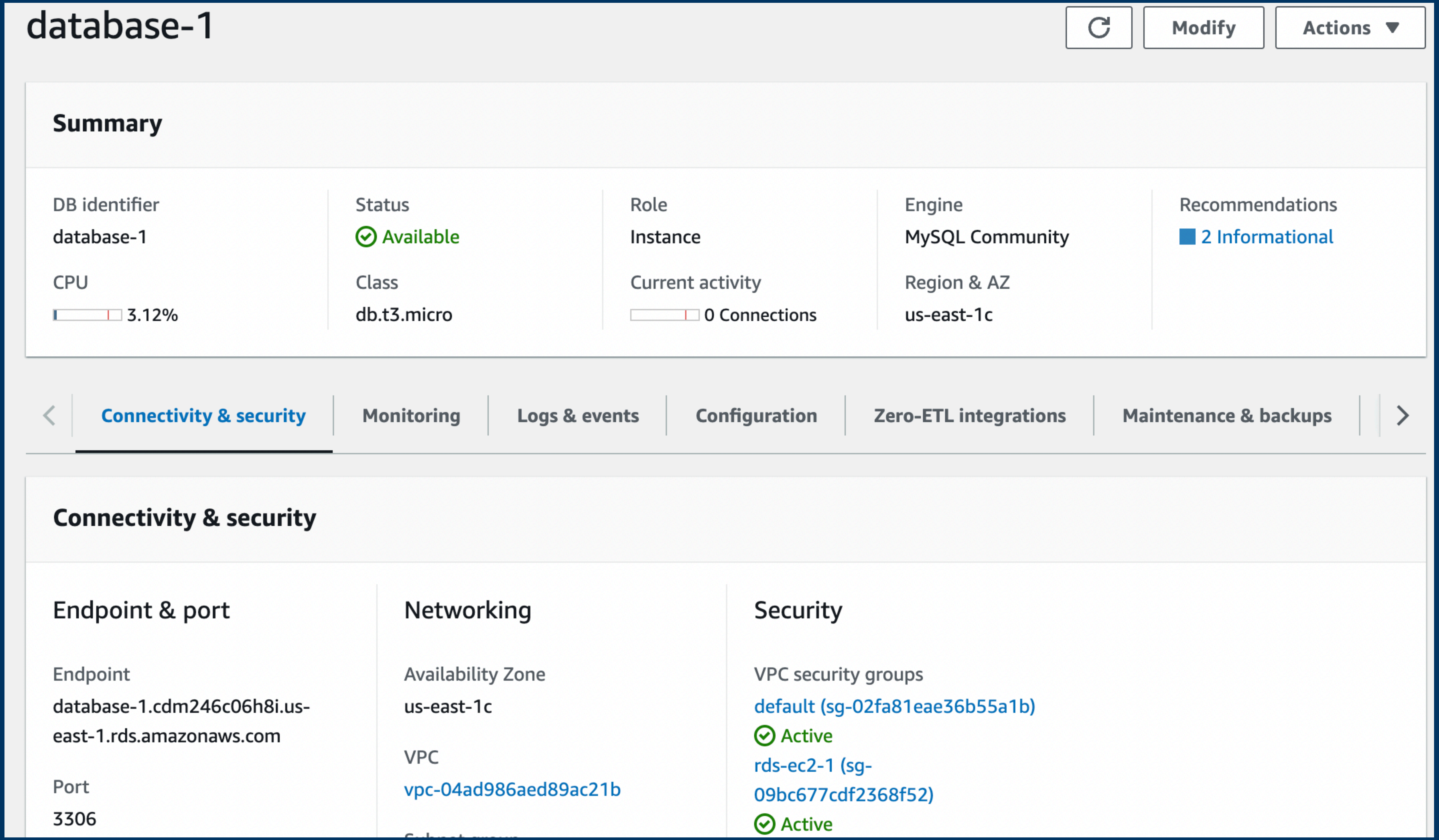
Map Reduce Assignment

AWS RDS

Siddhartha Ghoshal , 5th March 2024

Step 1 : Create an RDS instance in your AWS account and upload the data to the RDS instance.

Files to be loaded - *yellow_tripdata_2017-01.csv & yellow_tripdata_2017-02.csv*



database-1

Summary

DB identifier	Status	Role	Engine	Recommendations
database-1	✓ Available	Instance	MySQL Community	2 Informational
CPU	Class	Current activity	Region & AZ	
3.12%	db.t3.micro	0 Connections	us-east-1c	

Connectivity & security

Endpoint & port

Endpoint
database-1.cdm246c06h8i.us-east-1.rds.amazonaws.com

Port
3306

Networking

Availability Zone
us-east-1c

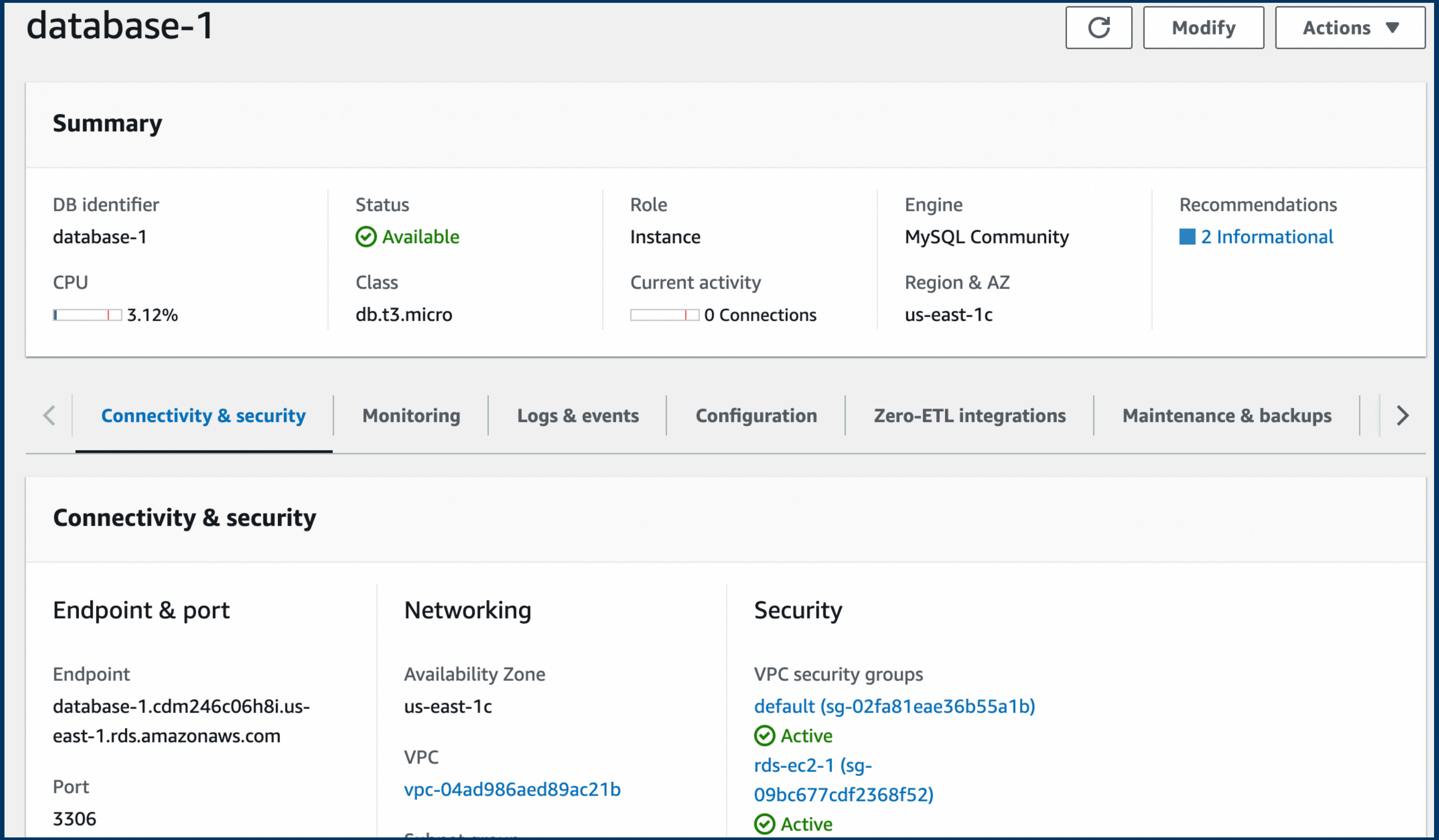
VPC
vpc-04ad986aed89ac21b

Security

VPC security groups
default (sg-02fa81eae36b55a1b)
✓ Active
rds-ec2-1 (sg-09bc677cdf2368f52)
✓ Active

Step 2 : Connect to RDS instance

```
mysql -h database-1.cdm246c06h8i.us-east-1.rds.amazonaws.com -P 3306 -u admin -p
```



database-1

Summary

DB identifier
database-1

CPU
3.12%

Status
✓ Available

Class
db.t3.micro

Role
Instance

Current activity
0 Connections

Engine
MySQL Community

Region & AZ
us-east-1c

Recommendations
🟦 2 Informational

Connectivity & security

Endpoint & port

Endpoint
database-1.cdm246c06h8i.us-east-1.rds.amazonaws.com

Port
3306

Networking

Availability Zone
us-east-1c

VPC
vpc-04ad986aed89ac21b

Security

VPC security groups
default (sg-02fa81eae36b55a1b)
✓ Active
rds-ec2-1 (sg-09bc677cdf2368f52)
✓ Active

Step 2 : create database and table

```
/* drop database*/
drop database if exists TLC;

/* create database*/
create database TLC;

/* use database*/
use TLC;
```

```
drop table if exists TLC_Record_Data;

create table TLC_Record_Data
(
VendorID integer,
tpep_pickup_datetime timestamp,
tpep_dropoff_datetime timestamp,
passenger_count integer,
trip_distance decimal(18,2),
RatecodeID integer,
store_and_fwd_flag char(1),
PULocationID integer,
DOLocationID integer,
payment_type integer,
fare_amount decimal(18,2),
extra  decimal(18,2),
mta_tax decimal(18,2),
tip_amount decimal(18,2),
tolls_amount decimal(18,2),
improvement_surcharge decimal(18,2),
total_amount decimal(18,2),
congestion_surcharge decimal(18,2),
airport_fee decimal(18,2),
);
```

Step 3 (approach 1) : Load data to RDS table from local using pandas df and sql alchemy

```
#python script to load using pandas

import pandas as pd
from sqlalchemy import create_engine
import ssl
from urllib.request import urlopen

# RDS configurations
RDS_ENDPOINT = 'database-1.cdm246c06h8i.us-east-1.rds.amazonaws.com'
RDS_PORT = '3306'
RDS_DATABASE = 'TLC'
RDS_USER = 'admin'
RDS_PASSWORD = 'Apple123'

# URLs of the CSV files
CSV_URL_1 = 'https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv'
CSV_URL_2 = 'https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv'

# Initialize SQLAlchemy engine for RDS
print("Establishing connection")
rds_engine = create_engine(f'mysql+pymysql://{RDS_USER}:{RDS_PASSWORD}@{RDS_ENDPOINT}:{RDS_PORT}/{RDS_DATABASE}')
print("Connection established")


def load_data_to_rds(csv_url, table_name):
    # Load data from CSV URL directly into RDS MySQL
    ctx = ssl.create_default_context()
    ctx.check_hostname = False
    ctx.verify_mode = ssl.CERT_NONE

    print("Data Frame Read")

    df = pd.read_csv(urlopen(csv_url, context=ctx), skiprows=1)

    print("Data Load begins")

    df.to_sql(name=table_name, con=rds_engine, if_exists='replace', index=False)

    print("Data Load fineshed for {}".format(CSV_URL_1))
if __name__ == "__main__":
    # Load data from first CSV URL to RDS
    load_data_to_rds(CSV_URL_1, 'TLC_Record_Data_Stg_1')

    # Load data from second CSV URL to RDS
    load_data_to_rds(CSV_URL_2, 'TLC_Record_Data_Stg_2')


insert into TLC_Record_Data select * from TLC_Record_Data_Stg_1

insert into TLC_Record_Data select * from TLC_Record_Data_Stg_2
```

Verify result :

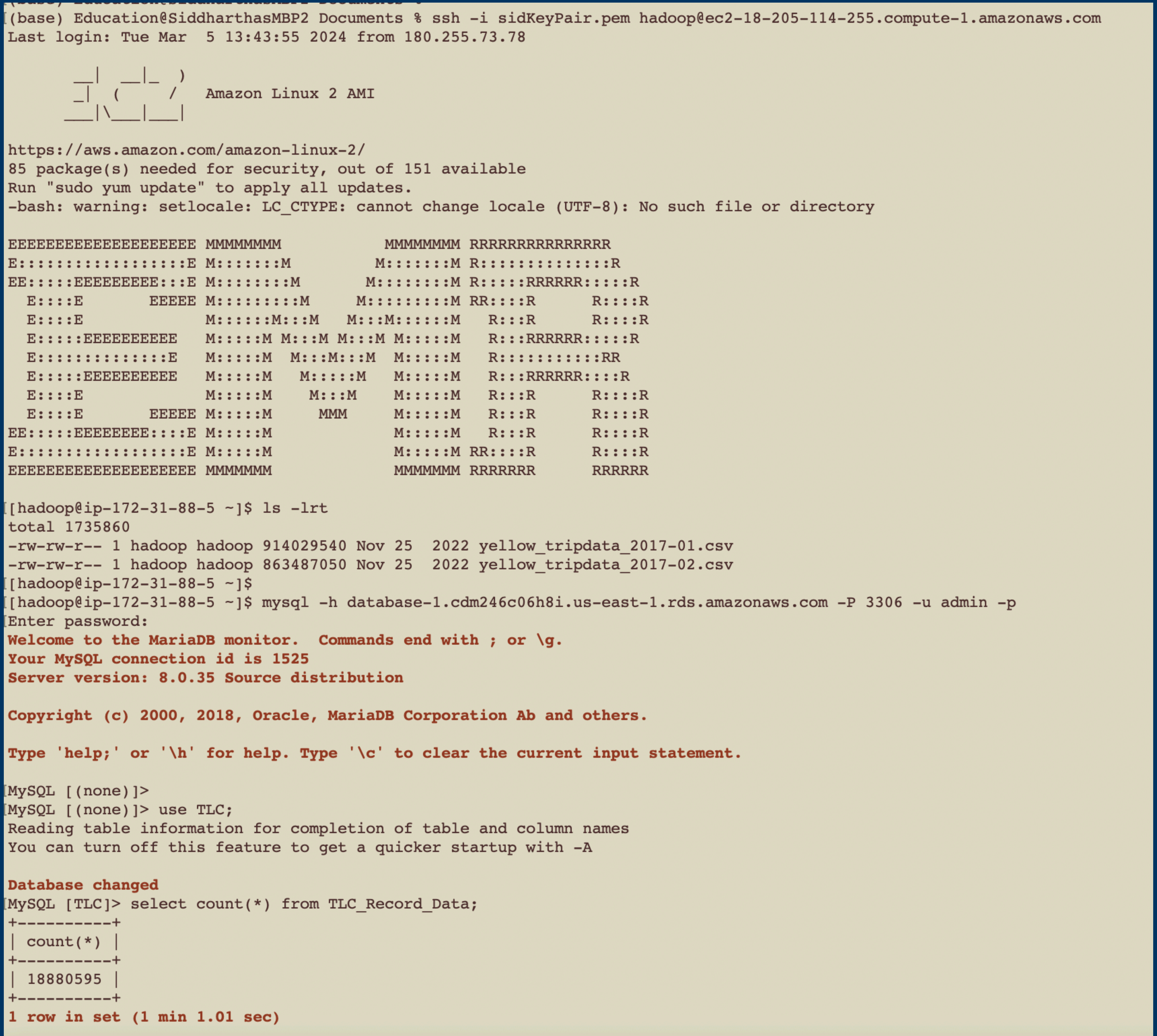
```
select count(*) from TLC_Record_Data

MySQL [TLC]> select count(*) from TLC_Record_Data;

+----------+
| count(*) |
+----------+
| 18880595 |
+----------+
1 row in set (1 min 11.21 sec)
```

Step 3 (approach 2) : Load data to RDS table from EMR cluster

```
-- create a EMR cluster
-- login to EMR cluster

ssh -i sidKeyPair.pem hadoop@ec2-54-197-38-85.compute-1.amazonaws.com

--- download  files

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

--- connect to DB

mysql -h database-1.cdm246c06h8i.us-east-1.rds.amazonaws.com -P 3306 -u admin -p

--- connected

Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 75
Server version: 8.0.35 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 TLC;

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

Query OK, 9710820 rows affected, 65535 warnings (2 min 50.19 sec)
Records: 9710820  Deleted: 0  Skipped: 0  Warnings: 19421640
```

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


Query OK, 9169775 rows affected, 65535 warnings (2 min 41.67 sec)
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 18339550


------ verify records

select count(*) from TLC_Record_Data

MySQL [TLC]> select count(*) from TLC_Record_Data;

+----------+
| count(*) |
+----------+
| 18880595 |
+----------+
1 row in set (1 min 11.21 sec)

[hadoop@ip-172-31-83-75 ~]$ wc -l *
   9710821 yellow_tripdata_2017-01.csv
   9169776 yellow_tripdata_2017-02.csv
  18880597 total

------ records verified ----------
```

Screenshot



```
[(base) Education@SiddharthasMBP2 Documents % ssh -i sidKeyPair.pem hadoop@ec2-18-205-114-255.compute-1.amazonaws.com
Last login: Tue Mar  5 13:43:55 2024 from 180.255.73.78

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
85 package(s) needed for security, out of 151 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 MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::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:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   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 MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[[hadoop@ip-172-31-88-5 ~]$ ls -lrt
total 1735860
-rw-rw-r-- 1 hadoop hadoop 914029540 Nov 25  2022 yellow_tripdata_2017-01.csv
-rw-rw-r-- 1 hadoop hadoop 863487050 Nov 25  2022 yellow_tripdata_2017-02.csv
[[hadoop@ip-172-31-88-5 ~]$
[[hadoop@ip-172-31-88-5 ~]$ mysql -h database-1.cdm246c06h8i.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 1525
Server version: 8.0.35 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)]>
[MySQL [(none)]> use TLC;
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 [TLC]> select count(*) from TLC_Record_Data;
+----------+
| count(*) |
+----------+
| 18880595 |
+----------+
1 row in set (1 min 1.01 sec)
```