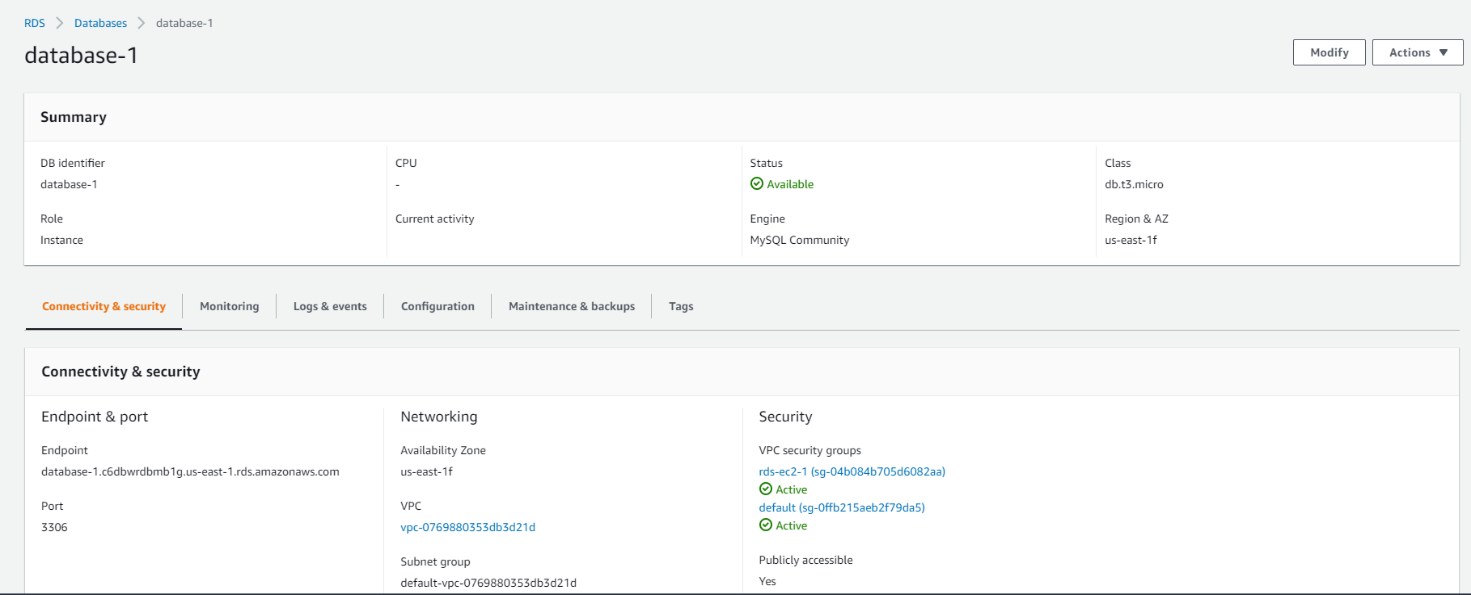
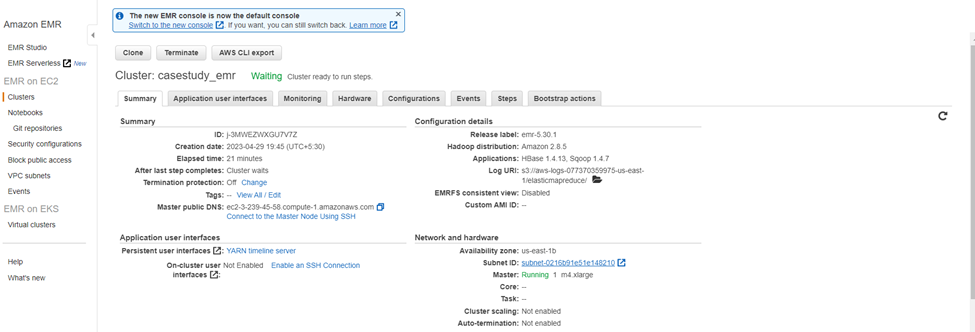
**DATA INGESTION TASKS**

**Task- 1**

* Create an RDS instance in your AWS account and upload the data to the RDS instance



* Create an EMR Cluster in your AWS account and setting up an EC2 instance connection between EMR cluster and the RDS Instance.



* Creating SQL Table to store data
* CREATE TABLE IF NOT EXISTS yellow\_trip

(

VendorID INT,

tpep\_pickup\_datetime TIMESTAMP,

tpep\_dropoff\_datetime TIMESTAMP,

passenger\_count INT,

trip\_distance FLOAT,

RatecodeID INT,

store\_and\_fwd\_flag VARCHAR(10),

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,

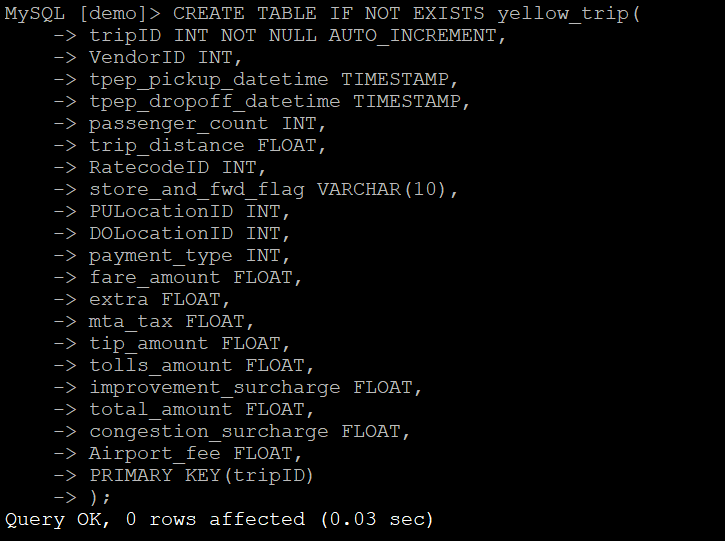
congestion\_surcharge FLOAT,

Airport\_fee FLOAT,

tripID INT NOT NULL AUTO\_INCREMENT,

PRIMARY KEY (tripID)

);



* Load data from csv files to SQL Table:

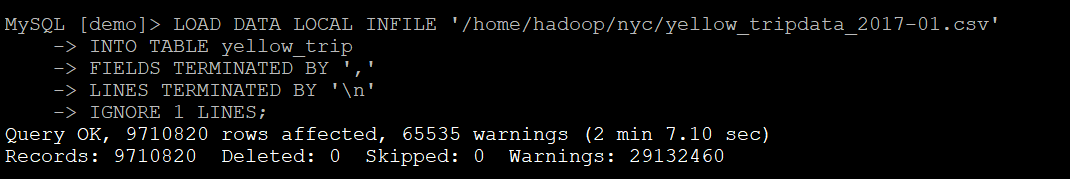
LOAD DATA LOCAL INFILE '/home/hadoop/nyc/yellow\_tripdata\_2017-01.csv'

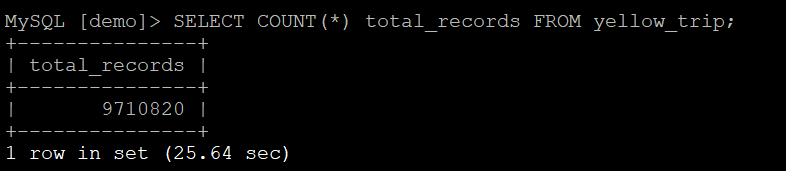
INTO TABLE yellow\_trip

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;





* Load yellow\_tripdata\_2017-02.csv

LOAD DATA LOCAL INFILE '/home/hadoop/nyc/yellow\_tripdata\_2017-02.csv'

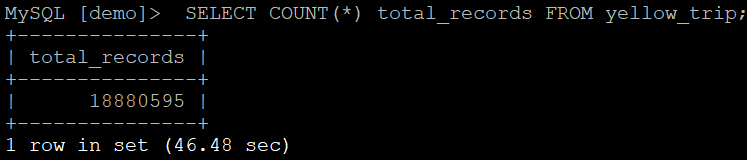
INTO TABLE yellow\_trip

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

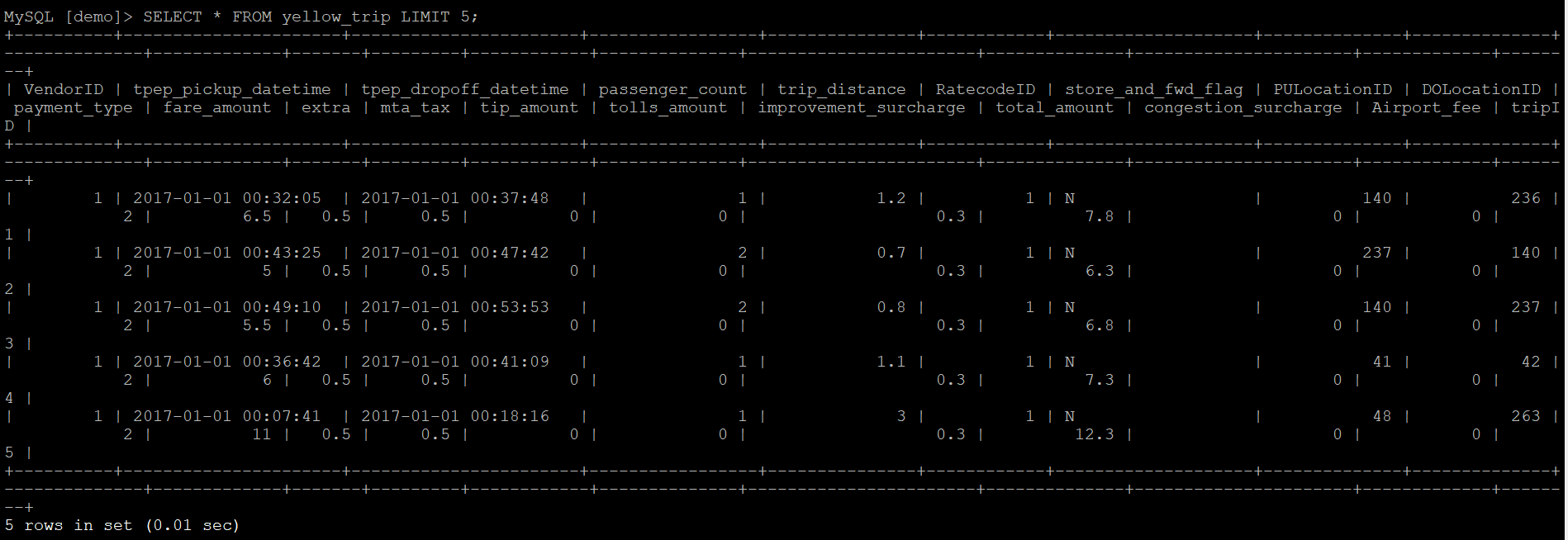
IGNORE 1 LINES;

* Checking the Records in table after loading the data
* There is total 18880595 records are present in table.



* Select 5 rows from table by using this SQL query:

SELECT \* FROM yellow\_trip LIMIT 5;



**TASK- 2**

Using Sqoop command to ingest the data from RDS into the HBase Table.

sqoop import --connect jdbc: mysql://database-1.c6dbwrdbmb1g.us-east-1.rds.amazonaws.com/demo \

--username admin \

--password mysql123 \

--table yellow\_trip \

--hbase-table yellow\_trip\_hb \

--hbase-row-key tripID \

--column-family yellow\_trip\_columns --hbase-create-table -m 4;

