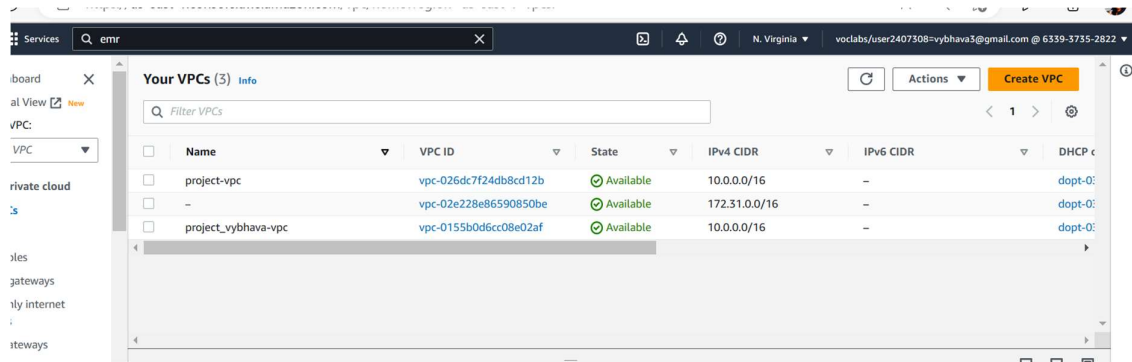


MapReduce - Programming Assignment

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

Since the dataset is huge, you need to upload the data from only two files (i.e. yellow_tripdata_2017-01.csv & yellow_tripdata_2017-02.csv) from the dataset.

1. CREATE VPC (Amazon virtual private cloud)



2. CREATE EMR Instance (Creation of amazon elastic map reduce instance)

[Clone](#) [Terminate](#) [AWS CLI export](#)

Cluster: My cluster_6 Waiting Cluster ready to run steps.

[Summary](#) [Application user interfaces](#) [Monitoring](#) [Hardware](#) [Configurations](#) [Events](#) [Steps](#) [Bootstrap actions](#)

Summary

ID: j-2N1PQ382Z8KN3
Creation date: 2023-03-12 10:39 (UTC+5:30)
Elapsed time: 1 hour, 52 minutes
After last step completes: Cluster waits
Termination protection: Off [Change](#)
Tags: -- [View All](#) / [Edit](#)
Master public DNS: ec2-18-213-4-69.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

Configuration details

Release label: emr-5.30.1
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.6.0, HBase 1.4.13, Sqoop 1.4.7
Log URI: s3://aws-logs-633937352822-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces [YARN timeline server](#), [Tez UI](#)
On-cluster user interfaces [Not Enabled](#) [Enable an SSH Connection](#)
interfaces [YARN](#)

Network and hardware

Availability zone: us-east-1b
Subnet ID: [subnet-017e02c51d01ea9ba](#)
Master: Running 1 m4.xlarge
Core: --
Task: --
Cluster scaling: Not enabled

MapReduce - Programming Assignment

```
login as: hadoop
Authenticating with public key "keyppk"
Last login: Sun Mar 12 07:16:49 2023

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

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEEE MMMMMMMM          MMMMMMMM RRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M          M::::::::M R:::::::::R
EE::::::::EEEEEEEEEE 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 M::M          R::::RRRRRR::::R
E::::::::::::::::E M::::::::M M::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          M::::::::M R::::R          R::::R
EE::::::::EEEEEEEEEE E M::::::::M          M::::::::M R::::R          R::::R
E::::::::::::::::::::E M::::::::M          M::::::::M RR::::R          R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMMM          MMMMMMMM RRRRRRR          RRRRRR

[hadoop@ip-10-0-21-165 ~]$ sudo -i

EEEEEEEEEEEEEEEEEEEE MMMMMMMM          MMMMMMMM RRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M          M::::::::M R:::::::::R
EE::::::::EEEEEEEEEE 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 M::M          R::::RRRRRR::::R
E::::::::::::::::E M::::::::M M::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          M::::::::M R::::R          R::::R
EE::::::::EEEEEEEEEE E M::::::::M          M::::::::M R::::R          R::::R
E::::::::::::::::::::E M::::::::M          M::::::::M RR::::R          R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMMM          MMMMMMMM RRRRRRR          RRRRRR

[root@ip-10-0-21-165 ~]# hbase shell

```

3. CREATE RDS DATABASE (Creation of RDS database)

RDS > Databases > assignment-db

assignment-db

Modify Actions ▼

Summary

DB identifier assignment-db	CPU <div><div></div>4.41%</div>	Status ✔ Available	Class db.t2.micro
Role Instance	Current activity <div><div></div>0 Connections</div>	Engine MySQL Community	Region & AZ us-east-1a

Connectivity & security

Monitoring

Logs & events

Configuration

Maintenance & backups

Tags

Connectivity & security

Endpoint & port	Networking	Security
Endpoint assignment-db.cnreri2hsgn.us-east-1.rds.amazonaws.com	Availability Zone us-east-1a	VPC security groups rds-ec2-8 (sg-078ffa331822b9337) ✔ Active
Port 3306	VPC project-vpc (vpc-026dc7f24db8cd12b)	ElasticMapReduce-master (sg-09288cc1ac4505475) ✔ Active

4. DOWNLOADING THE DATA /FILES TO EMR CLUSTER:

Code:

- 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

5. CONNECTING TO RDS FROM EMR CLUSTER

Code:

```
mysql -h assignment-db.cnreri2hsgn.us-east-1.rds.amazonaws.com -P 3306 -u admin -p
```

6. CREATING THE DATA BASE DEMO AND TABLE TLCTRIPDATA IN RDS.

```
create database demo;  
  
use demo;  
  
CREATE TABLE TLCTripData  
(VendorID int,  
tpep_pickup_datetime datetime,  
tpep_dropoff_datetime datetime,  
passenger_count int,  
trip_distance float,  
RatecodeID int,
```

store_and_fwd_flag char,
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);

```
root@ip-10-0-21-165:~  
Enter password:  
Welcome to the MariaDB monitor.  Commands end with ; or \g.  
Your MySQL connection id is 17  
Server version: 8.0.28 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)]> create database demo;  
Query OK, 1 row affected (0.01 sec)  
  
MySQL [(none)]> use demo;  
Database changed  
MySQL [demo]> CREATE TABLE TLCTripData  
-> (VendorID int,  
-> tpep_pickup_datetime datetime,  
-> tpep_dropoff_datetime datetime,  
-> passenger_count int,  
-> trip_distance float,  
-> RatecodeID int,  
-> store_and_fwd_flag char,  
-> 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 (0.03 sec)  
  
MySQL [demo]> LOAD DATA LOCAL INFILE 'yellow_tripdata_2017-01.csv'  
-> INTO TABLE TLCTripData  
-> FIELDS TERMINATED BY ','  
-> LINES TERMINATED BY '\n'  
-> IGNORE 1 LINES;  
Query OK, 9710820 rows affected, 65535 warnings (2 min 15.09 sec)  
Records: 9710820 Deleted: 0 Skipped: 0 Warnings: 9710820  
  
MySQL [demo]> █
```

7. LOADING THE DATA TO RDS TABLE : TLCTRIPDATA

LOAD DATA LOCAL INFILE 'yellow_tripdata_2017-01.csv'

INTO TABLE TLCTripData

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

LOAD DATA LOCAL INFILE 'yellow_tripdata_2017-02.csv'

INTO TABLE TLCTripData

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

```
MySQL [demo]> LOAD DATA LOCAL INFILE 'yellow_tripdata_2017-01.csv'
-> INTO TABLE TLCTripData
-> FIELDS TERMINATED BY ','
-> LINES TERMINATED BY '\n'
-> IGNORE 1 LINES;

Query OK, 9710820 rows affected, 65535 warnings (2 min 13.18 sec)
Records: 9710820 Deleted: 0 Skipped: 0 Warnings: 9710820

MySQL [demo]>
MySQL [demo]> LOAD DATA LOCAL INFILE 'yellow_tripdata_2017-02.csv'
-> INTO TABLE TLCTripData
-> FIELDS TERMINATED BY ','
-> LINES TERMINATED BY '\n'
-> IGNORE 1 LINES;

Query OK, 9169775 rows affected, 65535 warnings (2 min 50.66 sec)
Records: 9169775 Deleted: 0 Skipped: 0 Warnings: 9169775

MySQL [demo]> select count(*) from TLCTripData;
+-----+
| count(*) |
+-----+
| 18880595 |
+-----+
1 row in set (42.95 sec)

MySQL [demo]> █
```

Assignment Submitted by:
Susil Patro, Vybhava P, Vivek Agrawal