



MapReduce - Programming Assignment

AMR_RDS_Task



BY
Dharan R
Sanjana Dutta
Punith Vadla

AWS RDS on EMR Cluster

```

hadoop@ip-172-31-81-233:~$ ssh
login as: hadoop
Authenticating with public key "imported-openssh-key"
Last login: Sat Apr 15 18:01:45 2023

    _ _ _ _ _
   /   /   /   \
  /___/___/___\
   |   |   |   |
   |___|___|___|

Amazon Linux 2 AMI

https://aws.amazon.com/amazon-linux-2/
102 package(s) needed for security, out of 168 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMM      MMMMMMMM RRRRRRRRRRRRRRRR
E::::::::::::::::::::E M:::::M      M:::::M R:::::::::::::R
EE::::::::::::::::::E M:::::M      M:::::M R:::::RRRRR:::::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:::::RR
E:::::EEEEEEEEEE M:::::M M:::::M      M:::::M R:::::RRRRR:::::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::::::::::::::::::E M:::::M      M:::::M R:::::R      R:::::R
E::::::::::::::::::E M:::::M      M:::::M RR:::::R      R:::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM      MMMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-81-233 ~]$ mkdir /home/hadoop/dataset
[hadoop@ip-172-31-81-233 ~]$ ls
dataset
[hadoop@ip-172-31-81-233 ~]$ cd /home/hadoop/dataset
[hadoop@ip-172-31-81-233 dataset]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
--2023-04-15 18:06:03-- 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)... 52.217.101.12, 52.217.165.17, 52.217.167.33, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com) [52.217.101.12]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 914029540 (872M) [text/csv]
Saving to: 'yellow_tripdata_2017-01.csv'

100%[=====]
2023-04-15 18:06:27 (37.3 MB/s) - 'yellow_tripdata_2017-01.csv' saved [914029540/914029540]

[hadoop@ip-172-31-81-233 dataset]$

```

```
[hadoop@ip-172-31-81-233 dataset]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
--2023-04-15 18:07:43-- https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 3.5.16.186, 3.5.20.112, 52.217.13.132, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|3.5.16.186|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 863487050 (823M) [text/csv]
Saving to: 'yellow_tripdata_2017-02.csv'

100%[=====]

2023-04-15 18:08:05 (38.0 MB/s) - 'yellow_tripdata_2017-02.csv' saved [863487050/863487050]

[hadoop@ip-172-31-81-233 dataset]$ ls
yellow_tripdata_2017-01.csv yellow_tripdata_2017-02.csv
[hadoop@ip-172-31-81-233 dataset]$
```

RDS > Databases > nyc-taxi

nyc-taxi

Modify Actions

Summary

DB identifier	nyc-taxi
Role	
Instance	

```
[root@ip-172-31-90-29 ~]# mysql -h nyc-taxi.cdenutupqcb3.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 199
Server version: 8.0.32 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)]>
```

Connectivity & security

Endpoint & port Endpoint nyc-taxi.cdenutupqcb3.us-east-1.rds.amazonaws.com Port 3306	Networking Availability Zone us-east-1b VPC vpc-08bbcac340b4a0217 Subnet group	Security VPC security groups rds-ec2-3 (sg-0788b61c0a9235719) Active default (sg-0df1c654ac92908b4) Active Publicly accessible
---	--	---

mysql -h nyc-taxi.cqwodvedrqu4.us-east-1.rds.amazonaws.com -P 3306 -u admin -p

show databases;

create database assignmentdb;

use assignmentdb;

show tables;

```
MySQL [(none)]> show databases;
```

```
+-----+  
| Database           |  
+-----+  
| information_schema |  
| mysql              |  
| performance_schema |  
| sys                |  
+-----+
```

```
4 rows in set (0.03 sec)
```

```
MySQL [(none)]> CREATE DATABASE yellowtaxi;
```

```
Query OK, 1 row affected (0.01 sec)
```

```
MySQL [(none)]> USE yellowtaxi;
```

```
Database changed
```

```
MySQL [yellowtaxi]> CREATE TABLE tripdata
```

```
-> (  
-> VendorID varchar(10),  
-> tpep_pickup_datetime datetime,  
-> tpep_dropoff_datetime datetime,  
-> passenger_count INT,  
-> trip_distance FLOAT,  
-> RatecodeID VARCHAR(10),  
-> store_and_fwd_flag VARCHAR(10),  
-> PULocationID VARCHAR(100),  
-> DOLocationID VARCHAR(100),  
-> payment_type VARCHAR (10),  
-> 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,  
-> UniqueID INT NOT NULL AUTO_INCREMENT,  
-> PRIMARY KEY (UniqueID)  
-> );
```

```
Query OK, 0 rows affected (0.13 sec)
```

4

add column with
auto increment at
end which will act
as rowkey in
hbase

create table trip_data

(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(255),

PULocationID INT,

DOLocationID INT,

```

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),
congestion_surcharge float(10,2),
airport_fee float(10,2),
UniqueID INT NOT NULL AUTO_INCREMENT,
PRIMARY KEY(ID)
);

```

```

Query OK, 0 rows affected (0.13 sec)

MySQL [yellowtaxi]> DESC tripdata;
+-----+-----+-----+-----+-----+-----+
| Field | Type | Null | Key | Default | Extra |
+-----+-----+-----+-----+-----+-----+
| VendorID | varchar(10) | YES | | NULL | |
| tpep_pickup_datetime | datetime | YES | | NULL | |
| tpep_dropoff_datetime | datetime | YES | | NULL | |
| passenger_count | int | YES | | NULL | |
| trip_distance | float | YES | | NULL | |
| RatecodeID | varchar(10) | YES | | NULL | |
| store_and_fwd_flag | varchar(10) | YES | | NULL | |
| PULocationID | varchar(100) | YES | | NULL | |
| DOLocationID | varchar(100) | YES | | NULL | |
| payment_type | varchar(10) | YES | | NULL | |
| fare_amount | float | YES | | NULL | |
| extra | float | YES | | NULL | |
| mta_tax | float | YES | | NULL | |
| tip_amount | float | YES | | NULL | |
| tolls_amount | float | YES | | NULL | |
| improvement_surcharge | float | YES | | NULL | |
| total_amount | float | YES | | NULL | |
| Congestion_surcharge | float | YES | | NULL | |
| Airport_fee | float | YES | | NULL | |
| UniqueID | int | NO | PRI | NULL | auto_increment |
+-----+-----+-----+-----+-----+-----+
20 rows in set (0.04 sec)

MySQL [yellowtaxi]> LOAD DATA LOCAL INFILE '/home/hadoop/dataset/yellow_tripdata_2017-01.csv'
-> INTO TABLE tripdata
-> FIELDS TERMINATED BY ','
-> LINES TERMINATED BY '\n'
-> IGNORE 1 LINES;

Query OK, 9710820 rows affected, 65535 warnings (3 min 10.59 sec)
Records: 9710820 Deleted: 0 Skipped: 0 Warnings: 29132460

MySQL [yellowtaxi]> LOAD DATA LOCAL INFILE '/home/hadoop/dataset/yellow_tripdata_2017-02.csv'
-> INTO TABLE tripdata
-> FIELDS TERMINATED BY ','
-> LINES TERMINATED BY '\n'
-> IGNORE 1 LINES;

Query OK, 9169775 rows affected, 65535 warnings (3 min 0.75 sec)
Records: 9169775 Deleted: 0 Skipped: 0 Warnings: 27509325

```

5

Path where dataset is stored in first step

```

LOAD DATA LOCAL INFILE '/home/hadoop/dataset/yellow_tripdata_2017-01.csv'
INTO TABLE trip_data
FIELDS TERMINATED BY ','

```

Activat
Go to Set

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

show tables;

6

```
MySQL [yellowtaxi]> SELECT * FROM tripdata
-> LIMIT 5;
```

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	tip_amount	tolls_amount	improvement_surcharge	total_amount	Congestion_surcharge	Airport_fee	UniqueID
1	2017-01-01 00:32:05	2017-01-01 00:37:48	1	1.2	1	N	140	236	1	2									
1	2017-01-01 00:43:25	2017-01-01 00:47:42	2	0.7	1	N	237	140	2	1									
1	2017-01-01 00:49:10	2017-01-01 00:53:53	2	0.8	1	N	140	237	3	1									
1	2017-01-01 00:36:42	2017-01-01 00:41:09	1	1.1	1	N	41	42	4	1									
1	2017-01-01 00:07:41	2017-01-01 00:18:16	1	3	1	N	48	263	5	1									

5 rows in set (0.01 sec)

select * from tripdata limit 5;

select COUNT(*) from tripdata;

LOAD DATA LOCAL INFILE '/home/hadoop/dataset/yellow_tripdata_2017-02.csv'

INTO TABLE trip_data

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

select COUNT(*) from trip_data;