

# Amazon RDS

## Data Ingestion Task

The codes and methods we used to import data from two csv files to Amazon RDS using EMR are present in this file.

Files we used:-

1. [https://nyc-tlc-upgrad.s3.amazonaws.com/yellow\\_tripdata\\_2017-01.csv](https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv)
2. [https://nyc-tlc-upgrad.s3.amazonaws.com/yellow\\_tripdata\\_2017-02.csv](https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv)

### Step 1: Login in EMR

Code:

```
ssh -i demoKey07.pem hadoop@ec2-3-237-71-254.compute-1.amazonaws.com
```

```
(base) apple@Apples-MacBook-Pro ~ % cd Desktop
(base) apple@Apples-MacBook-Pro Desktop % ssh -i demoKey07.pem hadoop@ec2-3-237-71-254.compute-1.amazonaws.com
Last login: Thu Dec 15 10:01:42 2022 from 49.36.182.104

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

https://aws.amazon.com/amazon-linux-2/
24 package(s) needed for security, out of 46 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 RRRRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M M::::::::M R:::::::::R
EE::::::::::::::::::::E M::::::::M M::::::::M R:::::::::R
E::::E EEEEE M::::::::M M::::::::M RR:::R R:::R
E::::E M::::::::M M:::M M:::M M:::M R:::R R:::R
E::::EEEEEEEE M:::M M:::M M:::M M:::M R:::RRRRR:::R
E::::::::::::E M:::M M:::M M:::M M:::M R:::RRRRRR::RR
E::::EEEEEEEE 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 MMM M:::M R:::R R:::R
EE::::::::::::E M:::M 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-172-31-5-184 ~]$
```

### Step 2: Uploading raw data csv files to EMR

Created a dictionary to upload raw csv files

Code:

```
mkdir rawDataFiles
```

```
/home/hadoop
[hadoop@ip-172-31-5-184 ~]$ mkdir rawDataFiles
[hadoop@ip-172-31-5-184 ~]$ ls
rawDataFiles
[hadoop@ip-172-31-5-184 ~]$
```

Upload first file from local to EMR

**Code:**

```
scp -i demoKey07.pem yellow_tripdata_2017-01.csv  
hadoop@ec2-3-237-71-254.compute-1.amazonaws.com:/home/hadoop/rawDataFiles/
```

```
(base) apple@Apples-MacBook-Pro ~ % cd Desktop  
(base) apple@Apples-MacBook-Pro Desktop % scp -i demoKey07.pem yellow_tripdata_2017-01.csv hadoop@ec2-3-237-71-254.compute-1.amazonaws.com:/home/hadoop/rawDataFiles/  
/etc/profile.d/lang.sh: line 19: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory  
yellow_tripdata_2017-01.csv 100% 872MB 6.5MB/s 02:13  
(base) apple@Apples-MacBook-Pro Desktop %
```

Upload second file from local to EMR

**Code:**

```
scp -i demoKey07.pem yellow_tripdata_2017-02.csv  
hadoop@ec2-3-218-250-160.compute-1.amazonaws.com:/home/hadoop/rawDataFiles/
```

```
(base) apple@Apples-MacBook-Pro Desktop % scp -i demoKey07.pem yellow_tripdata_2017-02.csv hadoop@ec2-3-218-250-160.compute-1.amazonaws.com:/home/hadoop/rawDataFiles/  
/etc/profile.d/lang.sh: line 19: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory  
yellow_tripdata_2017-02.csv 100% 823MB 7.1MB/s 01:56  
(base) apple@Apples-MacBook-Pro Desktop %
```

Both the files are now available on EMR Directory

```
[hadoop@ip-172-31-5-184 rawDataFiles]$ ls  
yellow_tripdata_2017-01.csv yellow_tripdata_2017-02.csv  
[hadoop@ip-172-31-5-184 rawDataFiles]$ rm -r yellow_tripdata_20
```

## Step 2: Login in to RDS using MySQL CLI

**Code:**

```
sudo mysql --local-infile=1 -h  
mapreduceassignment.cjipwhexa9kv.us-east-1.rds.amazonaws.com -u admin -p
```

```
[hadoop@ip-172-31-5-184 rawDataFiles]$ sudo mysql --local-infile=1 -h mapreduceassignment.cjipwhexa9kv.us-east-1.rds.amazonaws.com -u admin -p  
Enter password:  
Welcome to the MariaDB monitor. Commands end with ; or \g.  
Your MySQL connection id is 62  
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)]>
```

## Step 2: Creating Database and table in RDS

Create new database "MapReduceAssignment"

**Code:**

```
create database MapReduceAssignment;
```

```
[MySQL [(none)]]> show databases;
+-----+
| Database |
+-----+
| information_schema |
| mysql |
| performance_schema |
| sys |
+-----+
4 rows in set (0.01 sec)

[MySQL [(none)]]> create database MapReduceAssignment;
Query OK, 1 row affected (0.02 sec)
```

Use create database

**Code:**

```
Use MapReduceAssignment;
```

```
[MySQL [(none)]]> use MapReduceAssignment;
Database changed
MySQL [MapReduceAssignment]> █
```

Create table command

**Code:**

```
CREATE TABLE trip_record(
  VendorID          INTEGER
  ,tpep_pickup_datetime DATETIME
  ,tpep_dropoff_datetime DATETIME
  ,passenger_count  INTEGER
  ,trip_distance     NUMERIC(5,2)
  ,RatecodeID       INTEGER
  ,store_and_fwd_flag VARCHAR(1)
  ,PULocationID     INTEGER
  ,DOLocationID     INTEGER
  ,payment_type      INTEGER
  ,fare_amount       NUMERIC(4,1)
  ,extra             NUMERIC(3,1)
  ,mta_tax           NUMERIC(3,1)
  ,tip_amount        NUMERIC(5,2)
  ,tolls_amount      NUMERIC(4,2)
  ,improvement_surcharge NUMERIC(3,1)
```

```
,total_amount      NUMERIC(5,2)
,congestion_surcharge NUMERIC(4,1)
,airport_fee        NUMERIC(4,1));
```

```
MySQL [MapReduceAssignment]> CREATE TABLE trip_record(
-> VendorID          INTEGER
-> ,tpep_pickup_datetime DATETIME
-> ,tpep_dropoff_datetime DATETIME
-> ,passenger_count   INTEGER
-> ,trip_distance     NUMERIC(5,2)
-> ,RatecodeID        INTEGER
-> ,store_and_fwd_flag VARCHAR(1)
-> ,PULocationID      INTEGER
-> ,DOLocationID      INTEGER
-> ,payment_type       INTEGER
-> ,fare_amount        NUMERIC(4,1)
-> ,extra              NUMERIC(3,1)
-> ,mta_tax            NUMERIC(3,1)
-> ,tip_amount         NUMERIC(5,2)
-> ,tolls_amount       NUMERIC(4,2)
-> ,improvement_surcharge NUMERIC(3,1)
-> ,total_amount       NUMERIC(5,2)
-> ,congestion_surcharge NUMERIC(4,1)
-> ,airport_fee        NUMERIC(4,1)
[ -> );
Query OK, 0 rows affected (0.04 sec)
```

Importing data from first file

**Code:**

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

```
MySQL [MapReduceAssignment]> LOAD DATA LOCAL INFILE '/home/hadoop/rawDataFiles/yellow_tripdata_2017-01.csv' INTO TABLE trip_record FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9718820 rows affected, 6535 warnings (1 min 48.68 sec)
Records: 9718820 Deleted: 0 Skipped: 0 Warnings: 19424237
```

Checking whether data is loaded or not

**Code:**

```
select * from trip_record LIMIT 10;
```

```
MySQL [MapReduceAssignment]> select * from trip_record LIMIT 10;
```

VendorID	trip_pickup_datetime	trip_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	DOLocationID	payment_type	fare_amount	extra	mta_tax	tip_amount
1	2017-01-01 00:32:05	2017-01-01 00:37:48	1	1.20	1	N	140	236	2	6.5	0.5	0.5	0.0
1	2017-01-01 00:43:25	2017-01-01 00:47:42	2	0.70	1	N	237	140	2	5.0	0.5	0.5	0.0
1	2017-01-01 00:49:10	2017-01-01 00:53:53	2	0.80	1	N	140	237	2	5.5	0.5	0.5	0.0
1	2017-01-01 00:36:42	2017-01-01 00:41:09	1	1.10	1	N	41	42	2	6.0	0.5	0.5	0.0
1	2017-01-01 00:07:41	2017-01-01 00:10:16	1	3.00	1	N	48	263	2	11.0	0.5	0.5	0.0
1	2017-01-01 00:33:49	2017-01-01 00:42:38	2	0.70	1	N	236	262	2	5.0	0.5	0.5	0.0
1	2017-01-01 00:28:52	2017-01-01 00:24:59	2	0.60	1	N	236	238	1	8.0	0.5	0.5	1.8
1	2017-01-01 00:48:22	2017-01-01 00:52:15	2	0.60	1	N	238	239	1	5.0	0.5	0.5	1.2
1	2017-01-01 00:57:12	2017-01-01 01:06:28	2	1.00	1	N	239	48	1	7.5	0.5	0.5	1.7
1	2017-01-01 00:10:25	2017-01-01 00:29:06	1	1.00	1	N	246	48	2	12.0	0.5	0.5	0.0

```
10 rows in set (0.00 sec)
```

Importing data from second file

**Code:**

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

```
MySQL [MapReduceAssignment]> LOAD DATA LOCAL INFILE '/home/hadoop/rawDataFiles/yellow_tripdata_2017-02.csv' INTO TABLE trip_record FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9169775 rows affected, 68535 warnings (1 min 45.87 sec)
Records: 9169775 Deleted: 0 Skipped: 0 Warnings: 18342814
```

Checking whether data is loaded or not

```
MySQL [MapReduceAssignment]> select * from trip_record LIMIT 10;
```

VendorID	trip_pickup_datetime	trip_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	DOLocationID	payment_type	fare_amount	extra	mta_tax	tip_amount
1	2017-01-01 00:32:05	2017-01-01 00:37:48	1	1.20	1	N	140	236	2	6.5	0.5	0.5	0.0
1	2017-01-01 00:43:25	2017-01-01 00:47:42	2	0.70	1	N	237	140	2	5.0	0.5	0.5	0.0
1	2017-01-01 00:49:10	2017-01-01 00:53:53	2	0.80	1	N	140	237	2	5.5	0.5	0.5	0.0
1	2017-01-01 00:36:42	2017-01-01 00:41:09	1	1.10	1	N	41	42	2	6.0	0.5	0.5	0.0
1	2017-01-01 00:07:41	2017-01-01 00:10:16	1	3.00	1	N	48	263	2	11.0	0.5	0.5	0.0
1	2017-01-01 00:33:49	2017-01-01 00:42:38	2	0.70	1	N	236	262	2	5.0	0.5	0.5	0.0
1	2017-01-01 00:28:52	2017-01-01 00:24:59	2	0.60	1	N	236	238	1	8.0	0.5	0.5	1.8
1	2017-01-01 00:48:22	2017-01-01 00:52:15	2	0.60	1	N	238	239	1	5.0	0.5	0.5	1.2
1	2017-01-01 00:57:12	2017-01-01 01:06:28	2	1.00	1	N	239	48	1	7.5	0.5	0.5	1.7
1	2017-01-01 00:10:25	2017-01-01 00:29:06	1	1.00	1	N	246	48	2	12.0	0.5	0.5	0.0

```
10 rows in set (0.00 sec)
```

Total row in RDS:

```
[MySQL [MapReduceAssignment]> SELECT COUNT(*) FROM trip_record;
```

COUNT(*)
18880595

```
1 row in set (36.53 sec)
```

**Submitted By:**

Member 1: Nikhil Dhiman (nikhildhiman3644@gmail.com)

Member 2: Rohan kulkarni (rohan.kulkarni951@gmail.com)

Member 3: Mahesh Kumar Patra (maheshkumarpatra5@gmail.com)