

1. Downloading files to EMR using wget and listing them.

```
[root@ip-172-31-58-150 ~]# wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
--2023-04-26 04:30:43-- 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)... 52.217.72.100, 52.217.206.65, 52.217.233.137, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|52.217.72.100|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 914029540 (872M) [text/csv]
Saving to: 'yellow_tripdata_2017-01.csv'

100%[=====>] 914,029,540 38.2MB/s in 25s

2023-04-26 04:31:08 (34.5 MB/s) - 'yellow_tripdata_2017-01.csv' saved [914029540/914029540]

[root@ip-172-31-58-150 ~]# wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
--2023-04-26 04:31:17-- 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)... 54.231.227.241, 54.231.228.57, 54.231.230.1, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|54.231.227.241|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 863487050 (823M) [text/csv]
Saving to: 'yellow_tripdata_2017-02.csv'

100%[=====>] 863,487,050 33.7MB/s in 25s

2023-04-26 04:31:42 (33.2 MB/s) - 'yellow_tripdata_2017-02.csv' saved [863487050/863487050]

[hadoop@ip-172-31-58-150 ~]$ ls *.csv
yellow_tripdata_2017-01.csv  yellow_tripdata_2017-03.csv
yellow_tripdata_2017-02.csv  yellow_tripdata_2017-04.csv
[hadoop@ip-172-31-58-150 ~]$
```

2. Logging in to RDS and selecting data base.

```
mysql -h nyc-tlc-mr-db.cv4nmpnzjxab.us-east-1.rds.amazonaws.com -P 3306 -u admin -p
use taximrdb;
```

```
[root@ip-172-31-58-150 ~]# mysql -h nyc-tlc-mr-rds.cv4nmpnzjxab.us-east-1.r
ds.amazonaws.com -P 3306 -u admin -p
Enter password:
Welcome to the MariaDB monitor. Commands end with ; or \g.
Your MySQL connection id is 17
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)]> use taximrdb;
Database changed
```

3. Creating table and adding autoincrement column.

```
CREATE TABLE yellow_taxi_trips (VendorID INT, tpep_pickup_datetime DateTime,
tpep_dropoff_datetime DateTime, passenger_count INT, trip_distance FLOAT, RatecodeID
INT, store_and_fwd_flag CHAR(1), PULocationID INT, DOLocationID INT, payment_type INT,
fare_amount FLOAT, mta_tax FLOAT, extra FLOAT, tip_amount FLOAT, tolls_amount FLOAT,
improvement_surcharge FLOAT, total_amount FLOAT, congestion_surcharge FLOAT,
airport_fee FLOAT);
```

```
ALTER TABLE yellow_taxi_trips ADD (trip_id INT PRIMARY KEY AUTO_INCREMENT);
```

```
MySQL [taximrdb]> CREATE TABLE yellow_taxi_trips (VendorID INT, tpep_pickup_datetime DateTime, tpep_dropoff_datetime DateTime, passenger_count INT, trip_distance FLOAT, RatecodeID INT, store_and_fwd_flag CHAR(1), PULocationID INT, DOLocationID INT, payment_type INT, fare_amount FLOAT, mta_tax FLOAT, extra FLOAT, tip_amount FLOAT, tolls_amount FLOAT, improvement_surcharge FLOAT, total_amount FLOAT, congestion_surcharge FLOAT, airport_fee FLOAT);
Query OK, 0 rows affected (0.03 sec)

MySQL [taximrdb]> DESCRIBE yellow_taxi_trips;
+-----+-----+-----+-----+-----+-----+
| Field | Type | Null | Key | Default | Extra |
+-----+-----+-----+-----+-----+-----+
| VendorID | int | 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 | int | YES | | NULL | |
| store_and_fwd_flag | char(1) | YES | | NULL | |
| PULocationID | int | YES | | NULL | |
| DOLocationID | int | YES | | NULL | |
| payment_type | int | YES | | NULL | |
| fare_amount | float | YES | | NULL | |
| mta_tax | float | YES | | NULL | |
| extra | 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 | |
+-----+-----+-----+-----+-----+-----+
19 rows in set (0.00 sec)

MySQL [taximrdb]> ALTER TABLE yellow_taxi_trips ADD (trip_id INT PRIMARY KEY AUTO_INCREMENT);
Query OK, 0 rows affected (0.05 sec)
Records: 0 Duplicates: 0 Warnings: 0
```

4. Loading data from the files

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

```
MySQL [taximrdb]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv' INTO TABLE yellow_taxi_trips FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9710820 rows affected, 65535 warnings (1 min 51.40 sec)
Records: 9710820 Deleted: 0 Skipped: 0 Warnings: 29132460
MySQL [taximrdb]>
```

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

```
MySQL [taximrdb]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv' INTO TABLE yellow_taxi_trips FIELDS TERMINATED BY ',' LINES
TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (2 min 41.39 sec)
Records: 9169775 Deleted: 0 Skipped: 0 Warnings: 27509325
MySQL [taximrdb]>
```

5. Checking count and data

```
MySQL [taximrdb]> SELECT COUNT(*) FROM yellow_taxi_trips;
+-----+
| COUNT(*) |
+-----+
| 18880595 |
+-----+
1 row in set (37.40 sec)

MySQL [taximrdb]>
```

SELECT trip_id, VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, RatecodeID, store_and_fwd_flag, PULocationID, DOLocationID, payment_type, fare_amount, mta_tax, extra, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge, airport_fee from yellow_taxi_trips LIMIT 5;

```
root@ip-172-31-58-150:/home/hadoop
MySQL [taximrdb]> SELECT trip_id, VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, RatecodeID, store_and_fwd_flag, PULocationID, DOLocationID, payment
_type, fare_amount, mta_tax, extra, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge, airport_fee from yellow_taxi_trips LIMIT 5;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| trip_id | VendorID | tpep_pickup_datetime | tpep_dropoff_datetime | passenger_count | trip_distance | RatecodeID | store_and_fwd_flag | PULocationID | DOLocationID | payment_type | fare_a | mta_tax | extra | tip_amount | tolls_amount | improvement_surcharge | total_amount | congestion_surcharge | airport_fee |
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | 1 | 2017-01-01 00:32:05 | 2017-01-01 00:37:48 | 1 | 1.2 | 1 | N | 140 | 236 | 2 | 6.5 | 0.5 | 0.5 | 0 | 0 | 0.3 | 7.8 | 0 | 0 |
| 2 | 1 | 2017-01-01 00:43:25 | 2017-01-01 00:47:42 | 2 | 6.3 | 1 | N | 237 | 140 | 2 | 5 | 0.5 | 0.5 | 0 | 0 | 0.3 | 6.3 | 0 | 0 |
| 3 | 1 | 2017-01-01 00:49:10 | 2017-01-01 00:53:53 | 2 | 6.8 | 1 | N | 140 | 237 | 2 | 5.5 | 0.5 | 0.5 | 0 | 0 | 0.3 | 6.8 | 0 | 0 |
| 4 | 1 | 2017-01-01 00:36:42 | 2017-01-01 00:41:09 | 1 | 1.1 | 1 | N | 41 | 42 | 2 | 6 | 0.5 | 0.5 | 0 | 0 | 0.3 | 7.3 | 0 | 0 |
| 5 | 1 | 2017-01-01 00:07:41 | 2017-01-01 00:18:16 | 1 | 3 | 1 | N | 48 | 263 | 2 | 11 | 0.5 | 0.5 | 0 | 0 | 0.3 | 12.3 | 0 | 0 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
5 rows in set (0.01 sec)

MySQL [taximrdb]> SELECT trip_id, VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, RatecodeID, store_and_fwd_flag, PULocationID, DOLocationID, payment
_type from yellow_taxi_trips LIMIT 5;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| trip_id | VendorID | tpep_pickup_datetime | tpep_dropoff_datetime | passenger_count | trip_distance | RatecodeID | store_and_fwd_flag | PULocationID | DOLocationID | payment_type |
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | 1 | 2017-01-01 00:32:05 | 2017-01-01 00:37:48 | 1 | 1.2 | 1 | N | 140 | 236 | 2 |
| 2 | 1 | 2017-01-01 00:43:25 | 2017-01-01 00:47:42 | 2 | 6.3 | 1 | N | 237 | 140 | 2 |
| 3 | 1 | 2017-01-01 00:49:10 | 2017-01-01 00:53:53 | 2 | 6.8 | 1 | N | 140 | 237 | 2 |
| 4 | 1 | 2017-01-01 00:36:42 | 2017-01-01 00:41:09 | 1 | 1.1 | 1 | N | 41 | 42 | 2 |
| 5 | 1 | 2017-01-01 00:07:41 | 2017-01-01 00:18:16 | 1 | 3 | 1 | N | 48 | 263 | 2 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
5 rows in set (0.00 sec)

MySQL [taximrdb]> SELECT trip_id, fare_amount, mta_tax, extra, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge, airport_fee from yellow_taxi_trips LIMIT
5;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| trip_id | fare_amount | mta_tax | extra | tip_amount | tolls_amount | improvement_surcharge | total_amount | congestion_surcharge | airport_fee |
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | 6.5 | 0.5 | 0.5 | 0 | 0 | 0.3 | 7.8 | 0 | 0 |
| 2 | 5 | 0.5 | 0.5 | 0 | 0 | 0.3 | 6.3 | 0 | 0 |
| 3 | 5.5 | 0.5 | 0.5 | 0 | 0 | 0.3 | 6.8 | 0 | 0 |
| 4 | 6 | 0.5 | 0.5 | 0 | 0 | 0.3 | 7.3 | 0 | 0 |
| 5 | 11 | 0.5 | 0.5 | 0 | 0 | 0.3 | 12.3 | 0 | 0 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
5 rows in set (0.00 sec)
```

6. Row count in file – including headers

```
MySQL [taximrdb]> exit;
Bye
[root@ip-172-31-58-150 hadoop]# wc -l yellow_tripdata_2017-01.csv
9710821 yellow_tripdata_2017-01.csv
[root@ip-172-31-58-150 hadoop]# wc -l yellow_tripdata_2017-02.csv
9169776 yellow_tripdata_2017-02.csv
[root@ip-172-31-58-150 hadoop]#
```