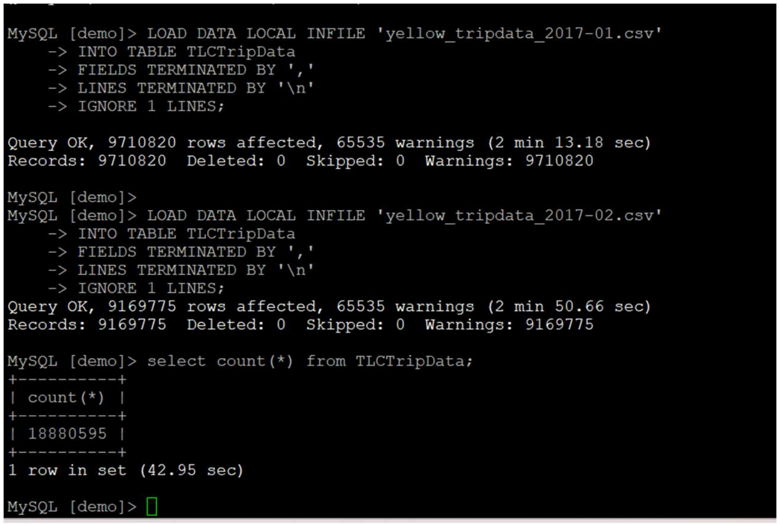
MapReduce - Programming Assignment

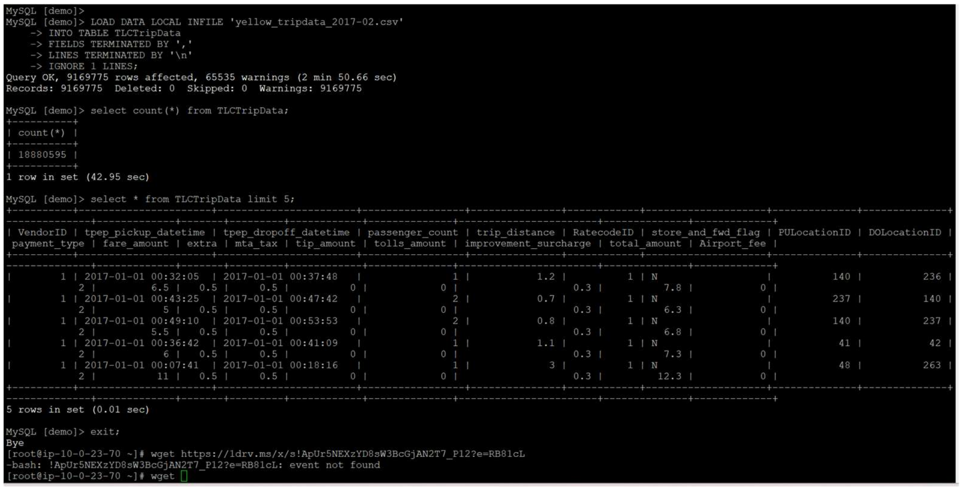
Task 2. Use Sqoop command to ingest the data from RDS into the HBase Table.

The following steps were followed for data ingestion from RDS into HBase table

1) Load the data to RDS instance
Following screen short for reference from previous step (i.e. Task 1)



```
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]>
```



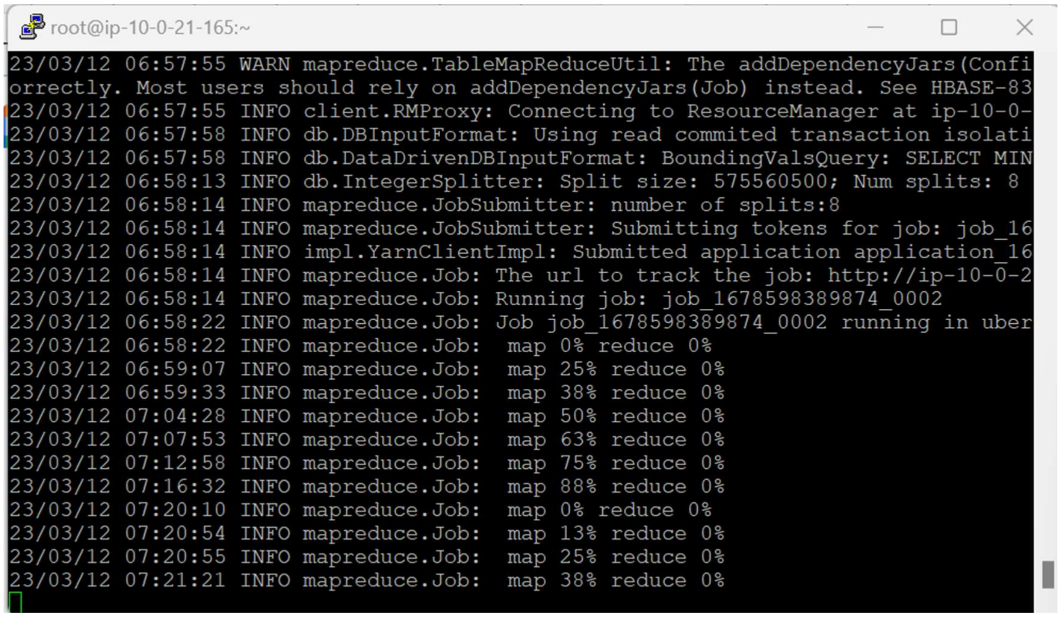
```
MySQL [demo]>
MySQL [demo]> LOAD DATA LOCAL INFILE 'yellow_tripdata_2017-02.csv'
    -> INTO TABLE TLCTripData
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (2 min 50.66 sec)
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 9169775

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

MySQL [demo]> select * from TLCTripData 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 | Airport_fee |
+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+
--------------+-------------+-------+---------+------------+--------------+-----------------------+--------------+-------------+
|        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 |
|        1 | 2017-01-01 00:43:25  | 2017-01-01 00:47:42   |               2 |           0.7 |          1 | N                  |          237 |          140 |
            2 |           5 |   0.5 |     0.5 |          0 |            0 |                   0.3 |          6.3 |           0 |
|        1 | 2017-01-01 00:49:10  | 2017-01-01 00:53:53   |               2 |           0.8 |          1 | N                  |          140 |          237 |
            2 |         5.5 |   0.5 |     0.5 |          0 |            0 |                   0.3 |          6.8 |           0 |
|        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 |
|        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 |
+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+
--------------+-------------+-------+---------+------------+--------------+-----------------------+--------------+-------------+
5 rows in set (0.01 sec)

MySQL [demo]> exit;
Bye
[root@ip-10-0-23-70 ~]# wget https://1drv.ms/x/s!ApUr5NEXzYD8sW3BcGjAN2T7_P12?e=RB81cL
-bash: !ApUr5NEXzYD8sW3BcGjAN2T7_P12?e=RB81cL: event not found
[root@ip-10-0-23-70 ~]# wget
```

2) Exit form RDS and load the table data in to the hbase.
Created the hbase table : nyc_yellow_taxi, set the column-family as TripDetails. Copied the data from the RDS TLCTripData table

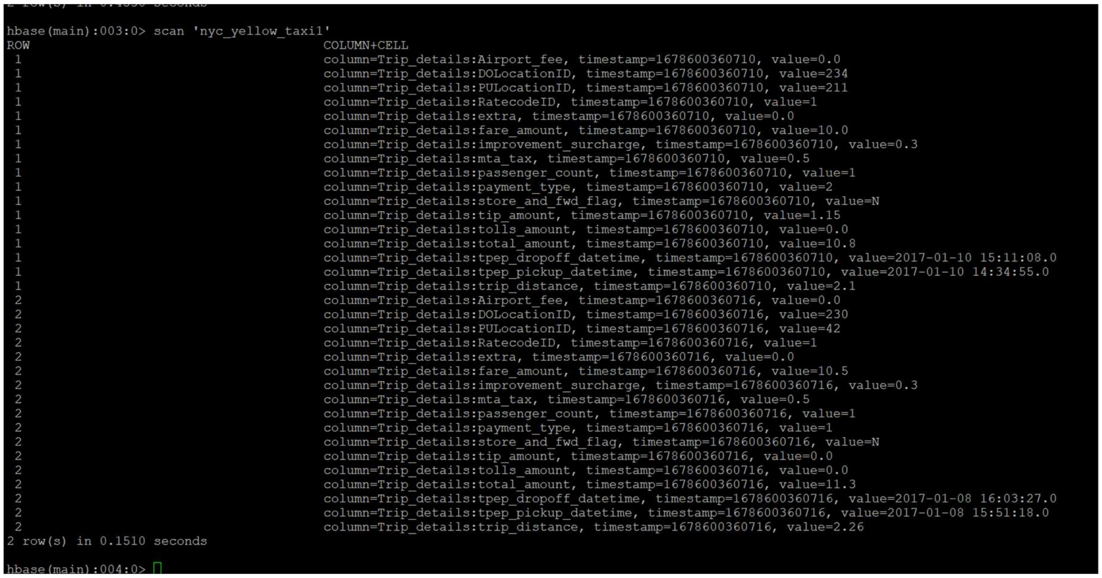
Sqoop import command:

```
sqoop import \
--connect "jdbc:mysql://assignment-db.crnreri2hsgn.us-east-1.rds.amazonaws.com:3306/demo" \
--username admin --password admin123 \
--table TLCTripData \
--columns
"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" \
--hbase-create-table \
--hbase-table nyc_yellow_taxi \
--column-family Trip_details \
--hbase-row-key VendorID,tpep_pickup_datetime,tpep_dropoff_datetime \
--split-by tpep_dropoff_datetime \
-m 8
```

Screenshots for reference:



```
23/03/12 06:57:55 WARN mapreduce.TableMapReduceUtil: The addDependencyJars(Confi
orrectly. Most users should rely on addDependencyJars(Job) instead. See HBASE-83
23/03/12 06:57:55 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-
23/03/12 06:57:58 INFO db.DBInputFormat: Using read commited transaction isolati
23/03/12 06:57:58 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN
23/03/12 06:58:13 INFO db.IntegerSplitter: Split size: 575560500; Num splits: 8
23/03/12 06:58:14 INFO mapreduce.JobSubmitter: number of splits:8
23/03/12 06:58:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
23/03/12 06:58:14 INFO impl.YarnClientImpl: Submitted application application_16
23/03/12 06:58:14 INFO mapreduce.Job: The url to track the job: http://ip-10-0-2
23/03/12 06:58:14 INFO mapreduce.Job: Running job: job_1678598389874_0002
23/03/12 06:58:22 INFO mapreduce.Job: Job job_1678598389874_0002 running in uber
23/03/12 06:58:22 INFO mapreduce.Job:  map 0% reduce 0%
23/03/12 06:59:07 INFO mapreduce.Job:  map 25% reduce 0%
23/03/12 06:59:33 INFO mapreduce.Job:  map 38% reduce 0%
23/03/12 07:04:28 INFO mapreduce.Job:  map 50% reduce 0%
23/03/12 07:07:53 INFO mapreduce.Job:  map 63% reduce 0%
23/03/12 07:12:58 INFO mapreduce.Job:  map 75% reduce 0%
23/03/12 07:16:32 INFO mapreduce.Job:  map 88% reduce 0%
23/03/12 07:20:10 INFO mapreduce.Job:  map 0% reduce 0%
23/03/12 07:20:54 INFO mapreduce.Job:  map 13% reduce 0%
23/03/12 07:20:55 INFO mapreduce.Job:  map 25% reduce 0%
23/03/12 07:21:21 INFO mapreduce.Job:  map 38% reduce 0%
```



```
hbase(main):003:0> scan 'nyc_yellow_taxi1'
ROW                                      COLUMN+CELL
 1                                       column=Trip_details:Airport_fee, timestamp=1678600360710, value=0.0
 1                                       column=Trip_details:DOLocationID, timestamp=1678600360710, value=234
 1                                       column=Trip_details:PULocationID, timestamp=1678600360710, value=211
 1                                       column=Trip_details:RatecodeID, timestamp=1678600360710, value=1
 1                                       column=Trip_details:extra, timestamp=1678600360710, value=0.0
 1                                       column=Trip_details:fare_amount, timestamp=1678600360710, value=10.0
 1                                       column=Trip_details:improvement_surcharge, timestamp=1678600360710, value=0.3
 1                                       column=Trip_details:mta_tax, timestamp=1678600360710, value=0.5
 1                                       column=Trip_details:passenger_count, timestamp=1678600360710, value=1
 1                                       column=Trip_details:payment_type, timestamp=1678600360710, value=2
 1                                       column=Trip_details:store_and_fwd_flag, timestamp=1678600360710, value=N
 1                                       column=Trip_details:tip_amount, timestamp=1678600360710, value=1.15
 1                                       column=Trip_details:tolls_amount, timestamp=1678600360710, value=0.0
 1                                       column=Trip_details:total_amount, timestamp=1678600360710, value=10.8
 1                                       column=Trip_details:tpep_dropoff_datetime, timestamp=1678600360710, value=2017-01-10 15:11:08.0
 1                                       column=Trip_details:tpep_pickup_datetime, timestamp=1678600360710, value=2017-01-10 14:34:55.0
 1                                       column=Trip_details:trip_distance, timestamp=1678600360710, value=2.1
 2                                       column=Trip_details:Airport_fee, timestamp=1678600360716, value=0.0
 2                                       column=Trip_details:DOLocationID, timestamp=1678600360716, value=230
 2                                       column=Trip_details:PULocationID, timestamp=1678600360716, value=42
 2                                       column=Trip_details:RatecodeID, timestamp=1678600360716, value=1
 2                                       column=Trip_details:extra, timestamp=1678600360716, value=0.0
 2                                       column=Trip_details:fare_amount, timestamp=1678600360716, value=10.5
 2                                       column=Trip_details:improvement_surcharge, timestamp=1678600360716, value=0.3
 2                                       column=Trip_details:mta_tax, timestamp=1678600360716, value=0.5
 2                                       column=Trip_details:passenger_count, timestamp=1678600360716, value=1
 2                                       column=Trip_details:payment_type, timestamp=1678600360716, value=1
 2                                       column=Trip_details:store_and_fwd_flag, timestamp=1678600360716, value=N
 2                                       column=Trip_details:tip_amount, timestamp=1678600360716, value=0.0
 2                                       column=Trip_details:tolls_amount, timestamp=1678600360716, value=0.0
 2                                       column=Trip_details:total_amount, timestamp=1678600360716, value=11.3
 2                                       column=Trip_details:tpep_dropoff_datetime, timestamp=1678600360716, value=2017-01-08 16:03:27.0
 2                                       column=Trip_details:tpep_pickup_datetime, timestamp=1678600360716, value=2017-01-08 15:51:18.0
 2                                       column=Trip_details:trip_distance, timestamp=1678600360716, value=2.26
2 row(s) in 0.1510 seconds

hbase(main):004:0>
```

Assignment Submitted by:

Susil Patro, Vybhava P, Vivek Agrawal