MapReduce Assignment

Data Ingestion

Siddhartha Ghoshal , 5 March 2024

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

Step 1 : login to EMR cluster

ssh -i sidKeyPair.pem hadoop@ec2-18-205-114-255.compute-1.amazonaws.com

Step 2 : Data ingest from RDS to Hbase table and create the table if not present already.

```
--- importing to hbase

sqoop import \
--connect jdbc:mysql://mapreduce-assignment.cncycg4u4cya.us-east-1.rds.amazonaws.com:3306/TLC \
--driver com.mysql.jdbc.Driver \
--username admin \
--password Apple123 \
--table TLC_Record_Data \
--hbase-create-table \
--hbase-table TLC_Record_Data \
--column-family records \
--hbase-bulkload \
--hbase-row-key VendorID -m 1
```

Step 3 : verifying table

hbase(main):005:0> describe 'TLC_Record_Data'
Table TLC_Record_Data is ENABLED
TLC_Record_Data
COLUMN FAMILIES DESCRIPTION
{NAME => 'records', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE',
DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', M
IN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.1900 seconds

**** Only one column family - records*

Step 4 : verifying data

```
hbase(main):013:0* scan 'TLC_Record_Data', {LIMIT => 4}
ROW                                    COLUMN+CELL
 1                                     column=records:DOLocationID, timestamp=1709356560164, value=238
 1                                     column=records:PULocationID, timestamp=1709356560164, value=74
 1                                     column=records:RatecodeID, timestamp=1709356560164, value=1
 1                                     column=records:airport_fee, timestamp=1709356560164, value=0.00
 1                                     column=records:congestion_surcharge, timestamp=1709356560164, value=0.00
 1                                     column=records:extra, timestamp=1709356560164, value=1.00
 1                                     column=records:fare_amount, timestamp=1709356560164, value=13.00
 1                                     column=records:improvement_surcharge, timestamp=1709356560164, value=0.30
 1                                     column=records:mta_tax, timestamp=1709356560164, value=0.50
 1                                     column=records:passenger_count, timestamp=1709356560164, value=1
 1                                     column=records:payment_type, timestamp=1709356560164, value=1
 1                                     column=records:store_and_fwd_flag, timestamp=1709356560164, value=N
 1                                     column=records:tip_amount, timestamp=1709356560164, value=0.00
 1                                     column=records:tolls_amount, timestamp=1709356560164, value=0.00
 1                                     column=records:total_amount, timestamp=1709356560164, value=14.80
 1                                     column=records:tpep_dropoff_datetime, timestamp=1709356560164, value=2017-01-31 18:59:17.0
 1                                     column=records:tpep_pickup_datetime, timestamp=1709356560164, value=2017-01-31 18:42:04.0
 1                                     column=records:trip_distance, timestamp=1709356560164, value=2.30
 2                                     column=records:DOLocationID, timestamp=1709356560164, value=238
 2                                     column=records:PULocationID, timestamp=1709356560164, value=162
 2                                     column=records:RatecodeID, timestamp=1709356560164, value=1
 2                                     column=records:airport_fee, timestamp=1709356560164, value=0.00
 2                                     column=records:congestion_surcharge, timestamp=1709356560164, value=0.00
 2                                     column=records:extra, timestamp=1709356560164, value=1.00
 2                                     column=records:fare_amount, timestamp=1709356560164, value=20.50
 2                                     column=records:improvement_surcharge, timestamp=1709356560164, value=0.30
 2                                     column=records:mta_tax, timestamp=1709356560164, value=0.50
 2                                     column=records:passenger_count, timestamp=1709356560164, value=1
 2                                     column=records:payment_type, timestamp=1709356560164, value=1
 2                                     column=records:store_and_fwd_flag, timestamp=1709356560164, value=N
 2                                     column=records:tip_amount, timestamp=1709356560164, value=4.46
 2                                     column=records:tolls_amount, timestamp=1709356560164, value=0.00
 2                                     column=records:total_amount, timestamp=1709356560164, value=26.76
 2                                     column=records:tpep_dropoff_datetime, timestamp=1709356560164, value=2017-01-31 19:15:00.0
 2                                     column=records:tpep_pickup_datetime, timestamp=1709356560164, value=2017-01-31 18:44:42.0
 2                                     column=records:trip_distance, timestamp=1709356560164, value=4.32
```

Task : Bulk import data from next two files in the dataset on your EMR cluster to your HBase Table using the relevant codes

***** Downloading files ********

wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-03.csv

wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-04.csv

***** copy to hapoop cluster ********

hadoop fs -put yellow_tripdata_2017-03.csv /user/hadoop/

hadoop fs -put yellow_tripdata_2017-04.csv /user/hadoop/

******** create base table *********

create 'TLC_Record_Data', 'records'

--- running bulk import command ----

```
hbase org.apache.hadoop.hbase.mapreduce.ImportTsv \
-Dimporttsv.separator=, \
-Dimporttsv.columns="HBASE_ROW_KEY,records:tpep_pickup_datetime,records:tpep_dropoff_datetime,records:passenger_count,records:trip_distance,records:RatecodeID,records:s
e_and_fwd_flag,records:PULocationID,records:DOLocationID,records:payment_type,records:fare_amount,records:extra,records:mta_tax,records:tip_amount,records:tolls_amount,reco
improvement_surcharge,records:total_amount,records:congestion_surcharge,records:airport_fee" \
-Dimporttsv.skip.bad.lines=true \
TLC_Record_Data \
/user/hadoop/yellow_tripdata_2017-03.csv
```

Checking Logs :

— — checking log ---

```
hadoop@ip-172-31-92-49 hadoop]$ tail -f /var/log/hbase/hbase.log
2024-03-02 05:46:03,274 INFO  [main] mapreduce.Job:  map 34% reduce 0%
2024-03-02 05:46:09,324 INFO  [main] mapreduce.Job:  map 35% reduce 0%
2024-03-02 05:46:15,344 INFO  [main] mapreduce.Job:  map 36% reduce 0%
2024-03-02 05:46:21,369 INFO  [main] mapreduce.Job:  map 37% reduce 0%
2024-03-02 05:46:27,400 INFO  [main] mapreduce.Job:  map 38% reduce 0%
2024-03-02 05:46:34,424 INFO  [main] mapreduce.Job:  map 39% reduce 0%
2024-03-02 05:46:40,444 INFO  [main] mapreduce.Job:  map 40% reduce 0%
2024-03-02 05:46:46,464 INFO  [main] mapreduce.Job:  map 41% reduce 0%
2024-03-02 05:46:51,481 INFO  [main] mapreduce.Job:  map 42% reduce 0%
2024-03-02 05:46:57,499 INFO  [main] mapreduce.Job:  map 43% reduce 0%
2024-03-02 05:47:03,522 INFO  [main] mapreduce.Job:  map 44% reduce 0%
2024-03-02 05:47:09,551 INFO  [main] mapreduce.Job:  map 45% reduce 0%
2024-03-02 05:47:15,570 INFO  [main] mapreduce.Job:  map 46% reduce 0%
2024-03-02 05:47:21,594 INFO  [main] mapreduce.Job:  map 47% reduce 0%
2024-03-02 05:47:27,624 INFO  [main] mapreduce.Job:  map 48% reduce 0%
2024-03-02 05:47:33,663 INFO  [main] mapreduce.Job:  map 49% reduce 0%
2024-03-02 05:47:37,683 INFO  [main] mapreduce.Job:  map 50% reduce 0%
2024-03-02 05:47:54,755 INFO  [main] mapreduce.Job:  map 51% reduce 0%
2024-03-02 05:48:00,772 INFO  [main] mapreduce.Job:  map 52% reduce 0%
2024-03-02 05:48:06,794 INFO  [main] mapreduce.Job:  map 53% reduce 0%
2024-03-02 05:48:15,856 INFO  [main] mapreduce.Job:  map 54% reduce 0%
2024-03-02 05:48:18,869 INFO  [main] mapreduce.Job:  map 55% reduce 0%
2024-03-02 05:48:24,901 INFO  [main] mapreduce.Job:  map 56% reduce 0%
2024-03-02 05:48:33,934 INFO  [main] mapreduce.Job:  map 57% reduce 0%
2024-03-02 05:48:39,964 INFO  [main] mapreduce.Job:  map 58% reduce 0%
2024-03-02 05:48:45,981 INFO  [main] mapreduce.Job:  map 59% reduce 0%
2024-03-02 05:48:51,999 INFO  [main] mapreduce.Job:  map 60% reduce 0%
2024-03-02 05:48:58,015 INFO  [main] mapreduce.Job:  map 61% reduce 0%
2024-03-02 05:49:04,032 INFO  [main] mapreduce.Job:  map 62% reduce 0%
2024-03-02 05:49:10,048 INFO  [main] mapreduce.Job:  map 63% reduce 0%
2024-03-02 05:49:16,064 INFO  [main] mapreduce.Job:  map 64% reduce 0%
2024-03-02 05:49:22,082 INFO  [main] mapreduce.Job:  map 65% reduce 0%
2024-03-02 05:49:28,098 INFO  [main] mapreduce.Job:  map 66% reduce 0%
2024-03-02 05:49:34,115 INFO  [main] mapreduce.Job:  map 67% reduce 0%
2024-03-02 05:49:40,139 INFO  [main] mapreduce.Job:  map 68% reduce 0%
2024-03-02 05:49:46,156 INFO  [main] mapreduce.Job:  map 69% reduce 0%
2024-03-02 05:49:52,174 INFO  [main] mapreduce.Job:  map 70% reduce 0%
2024-03-02 05:49:58,192 INFO  [main] mapreduce.Job:  map 71% reduce 0%
2024-03-02 05:50:04,217 INFO  [main] mapreduce.Job:  map 72% reduce 0%
2024-03-02 05:50:10,235 INFO  [main] mapreduce.Job:  map 73% reduce 0%
2024-03-02 05:50:16,261 INFO  [main] mapreduce.Job:  map 74% reduce 0%
2024-03-02 05:50:22,285 INFO  [main] mapreduce.Job:  map 75% reduce 0%
2024-03-02 05:50:39,372 INFO  [main] mapreduce.Job:  map 76% reduce 0%
2024-03-02 05:50:43,396 INFO  [main] mapreduce.Job:  map 78% reduce 0%
2024-03-02 05:50:49,424 INFO  [main] mapreduce.Job:  map 80% reduce 0%
2024-03-02 05:50:55,451 INFO  [main] mapreduce.Job:  map 82% reduce 0%
2024-03-02 05:51:01,477 INFO  [main] mapreduce.Job:  map 84% reduce 0%
2024-03-02 05:51:07,523 INFO  [main] mapreduce.Job:  map 86% reduce 0%
2024-03-02 05:51:13,555 INFO  [main] mapreduce.Job:  map 88% reduce 0%
2024-03-02 05:51:19,583 INFO  [main] mapreduce.Job:  map 90% reduce 0%
2024-03-02 05:51:21,593 INFO  [main] mapreduce.Job:  map 91% reduce 0%
2024-03-02 05:51:33,639 INFO  [main] mapreduce.Job:  map 92% reduce 0%
2024-03-02 05:51:45,674 INFO  [main] mapreduce.Job:  map 93% reduce 0%
2024-03-02 05:51:58,719 INFO  [main] mapreduce.Job:  map 94% reduce 0%
2024-03-02 05:52:10,762 INFO  [main] mapreduce.Job:  map 95% reduce 0%
2024-03-02 05:52:22,793 INFO  [main] mapreduce.Job:  map 96% reduce 0%
2024-03-02 05:52:34,827 INFO  [main] mapreduce.Job:  map 97% reduce 0%
2024-03-02 05:52:46,860 INFO  [main] mapreduce.Job:  map 98% reduce 0%
2024-03-02 05:52:58,901 INFO  [main] mapreduce.Job:  map 99% reduce 0%
2024-03-02 05:53:10,934 INFO  [main] mapreduce.Job:  map 100% reduce 0%
2024-03-02 05:53:13,945 INFO  [main] mapreduce.Job: Job job_1709354164340_0004 completed successfully
2024-03-02 05:53:14,096 INFO  [main] mapreduce.Job:  bytes written=0
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=9
                Data-local map tasks=9
                Total time spent by all maps in occupied slots (ms)=63031584
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=1313158
                Total vcore-milliseconds taken by all map tasks=1313158
                Total megabyte-milliseconds taken by all map tasks=2017010688
        Map-Reduce Framework
                Map input records=10295442
                Map output records=10295442
                Input split bytes=1160
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=4905
                CPU time spent (ms)=260900
                Physical memory (bytes) snapshot=4720021504
                Virtual memory (bytes) snapshot=26629623808
                Total committed heap usage (bytes)=3927441408
        ImportTsv
                Bad Lines=0
        File Input Format Counters
                Bytes Read=970267777
        File Output Format Counters
                Bytes Written=0
```

Checking Logs :

— — checking log ---

```
hadoop@ip-172-31-92-49 hadoop]$ tail -f /var/log/hbase/hbase.log
2024-03-02 05:46:03,274 INFO  [main] mapreduce.Job:  map 34% reduce 0%
2024-03-02 05:46:09,324 INFO  [main] mapreduce.Job:  map 35% reduce 0%
2024-03-02 05:46:15,344 INFO  [main] mapreduce.Job:  map 36% reduce 0%
2024-03-02 05:46:21,369 INFO  [main] mapreduce.Job:  map 37% reduce 0%
2024-03-02 05:46:27,400 INFO  [main] mapreduce.Job:  map 38% reduce 0%
2024-03-02 05:46:34,424 INFO  [main] mapreduce.Job:  map 39% reduce 0%
2024-03-02 05:46:40,444 INFO  [main] mapreduce.Job:  map 40% reduce 0%
2024-03-02 05:46:46,464 INFO  [main] mapreduce.Job:  map 41% reduce 0%
2024-03-02 05:46:51,481 INFO  [main] mapreduce.Job:  map 42% reduce 0%
2024-03-02 05:46:57,499 INFO  [main] mapreduce.Job:  map 43% reduce 0%
2024-03-02 05:47:03,522 INFO  [main] mapreduce.Job:  map 44% reduce 0%
2024-03-02 05:47:09,551 INFO  [main] mapreduce.Job:  map 45% reduce 0%
2024-03-02 05:47:15,570 INFO  [main] mapreduce.Job:  map 46% reduce 0%
2024-03-02 05:47:21,594 INFO  [main] mapreduce.Job:  map 47% reduce 0%
2024-03-02 05:47:27,624 INFO  [main] mapreduce.Job:  map 48% reduce 0%
2024-03-02 05:47:33,663 INFO  [main] mapreduce.Job:  map 49% reduce 0%
2024-03-02 05:47:37,683 INFO  [main] mapreduce.Job:  map 50% reduce 0%
2024-03-02 05:47:54,755 INFO  [main] mapreduce.Job:  map 51% reduce 0%
2024-03-02 05:48:00,772 INFO  [main] mapreduce.Job:  map 52% reduce 0%
2024-03-02 05:48:06,794 INFO  [main] mapreduce.Job:  map 53% reduce 0%
2024-03-02 05:48:15,856 INFO  [main] mapreduce.Job:  map 54% reduce 0%
2024-03-02 05:48:18,869 INFO  [main] mapreduce.Job:  map 55% reduce 0%
2024-03-02 05:48:24,901 INFO  [main] mapreduce.Job:  map 56% reduce 0%
2024-03-02 05:48:33,934 INFO  [main] mapreduce.Job:  map 57% reduce 0%
2024-03-02 05:48:39,964 INFO  [main] mapreduce.Job:  map 58% reduce 0%
2024-03-02 05:48:45,981 INFO  [main] mapreduce.Job:  map 59% reduce 0%
2024-03-02 05:48:51,999 INFO  [main] mapreduce.Job:  map 60% reduce 0%
2024-03-02 05:48:58,015 INFO  [main] mapreduce.Job:  map 61% reduce 0%
2024-03-02 05:49:04,032 INFO  [main] mapreduce.Job:  map 62% reduce 0%
2024-03-02 05:49:10,048 INFO  [main] mapreduce.Job:  map 63% reduce 0%
2024-03-02 05:49:16,064 INFO  [main] mapreduce.Job:  map 64% reduce 0%
2024-03-02 05:49:22,082 INFO  [main] mapreduce.Job:  map 65% reduce 0%
2024-03-02 05:49:28,098 INFO  [main] mapreduce.Job:  map 66% reduce 0%
2024-03-02 05:49:34,115 INFO  [main] mapreduce.Job:  map 67% reduce 0%
2024-03-02 05:49:40,139 INFO  [main] mapreduce.Job:  map 68% reduce 0%
2024-03-02 05:49:46,156 INFO  [main] mapreduce.Job:  map 69% reduce 0%
2024-03-02 05:49:52,174 INFO  [main] mapreduce.Job:  map 70% reduce 0%
2024-03-02 05:49:58,192 INFO  [main] mapreduce.Job:  map 71% reduce 0%
2024-03-02 05:50:04,217 INFO  [main] mapreduce.Job:  map 72% reduce 0%
2024-03-02 05:50:10,235 INFO  [main] mapreduce.Job:  map 73% reduce 0%
2024-03-02 05:50:16,261 INFO  [main] mapreduce.Job:  map 74% reduce 0%
2024-03-02 05:50:22,285 INFO  [main] mapreduce.Job:  map 75% reduce 0%
2024-03-02 05:50:39,372 INFO  [main] mapreduce.Job:  map 76% reduce 0%
2024-03-02 05:50:43,396 INFO  [main] mapreduce.Job:  map 78% reduce 0%
2024-03-02 05:50:49,424 INFO  [main] mapreduce.Job:  map 80% reduce 0%
2024-03-02 05:50:55,451 INFO  [main] mapreduce.Job:  map 82% reduce 0%
2024-03-02 05:51:01,477 INFO  [main] mapreduce.Job:  map 84% reduce 0%
2024-03-02 05:51:07,523 INFO  [main] mapreduce.Job:  map 86% reduce 0%
2024-03-02 05:51:13,555 INFO  [main] mapreduce.Job:  map 88% reduce 0%
2024-03-02 05:51:19,583 INFO  [main] mapreduce.Job:  map 90% reduce 0%
2024-03-02 05:51:21,593 INFO  [main] mapreduce.Job:  map 91% reduce 0%
2024-03-02 05:51:33,639 INFO  [main] mapreduce.Job:  map 92% reduce 0%
2024-03-02 05:51:45,674 INFO  [main] mapreduce.Job:  map 93% reduce 0%
2024-03-02 05:51:58,719 INFO  [main] mapreduce.Job:  map 94% reduce 0%
2024-03-02 05:52:10,762 INFO  [main] mapreduce.Job:  map 95% reduce 0%
2024-03-02 05:52:22,793 INFO  [main] mapreduce.Job:  map 96% reduce 0%
2024-03-02 05:52:34,827 INFO  [main] mapreduce.Job:  map 97% reduce 0%
2024-03-02 05:52:46,860 INFO  [main] mapreduce.Job:  map 98% reduce 0%
2024-03-02 05:52:58,901 INFO  [main] mapreduce.Job:  map 99% reduce 0%
2024-03-02 05:53:10,934 INFO  [main] mapreduce.Job:  map 100% reduce 0%
2024-03-02 05:53:13,945 INFO  [main] mapreduce.Job: Job job_1709354164340_0004 completed successfully
2024-03-02 05:53:14,096 INFO  [main] mapreduce.Job:  bytes written=0
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=9
                Data-local map tasks=9
                Total time spent by all maps in occupied slots (ms)=63031584
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=1313158
                Total vcore-milliseconds taken by all map tasks=1313158
                Total megabyte-milliseconds taken by all map tasks=2017010688
        Map-Reduce Framework
                Map input records=10295442
                Map output records=10295442
                Input split bytes=1160
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=4905
                CPU time spent (ms)=260900
                Physical memory (bytes) snapshot=4720021504
                Virtual memory (bytes) snapshot=26629623808
                Total committed heap usage (bytes)=3927441408
        ImportTsv
                Bad Lines=0
        File Input Format Counters
                Bytes Read=970267777
        File Output Format Counters
                Bytes Written=0
```

Checking Data :

scan 'TLC_Record_Data', {STARTROW => '1', ENDROW => '2'}

hbase(main):019:0> scan 'TLC_Record_Data', {STARTROW => '1', ENDROW => '4'}

```
ROW                                    COLUMN+CELL
 1                                     column=records:DOLocationID, timestamp=1709358115092, value=151
 1                                     column=records:PULocationID, timestamp=1709358115092, value=151
 1                                     column=records:RatecodeID, timestamp=1709358115092, value=1
 1                                     column=records:airport_fee, timestamp=1709358115092, value=
 1                                     column=records:congestion_surcharge, timestamp=1709358115092, value=
 1                                     column=records:extra, timestamp=1709358115092, value=0.5
 1                                     column=records:fare_amount, timestamp=1709358115092, value=4.5
 1                                     column=records:improvement_surcharge, timestamp=1709358115092, value=0.3
 1                                     column=records:mta_tax, timestamp=1709358115092, value=0.5
 1                                     column=records:passenger_count, timestamp=1709358115092, value=1
 1                                     column=records:payment_type, timestamp=1709358115092, value=1
 1                                     column=records:store_and_fwd_flag, timestamp=1709358115092, value=N
 1                                     column=records:tip_amount, timestamp=1709358115092, value=1.15
 1                                     column=records:tolls_amount, timestamp=1709358115092, value=0.0
 1                                     column=records:total_amount, timestamp=1709358115092, value=6.95
 1                                     column=records:tpep_dropoff_datetime, timestamp=1709358115092, value=2017-03-29 22:50:30
 1                                     column=records:tpep_pickup_datetime, timestamp=1709358115092, value=2017-03-29 22:47:31
 1                                     column=records:trip_distance, timestamp=1709358115092, value=0.6
 2                                     column=records:DOLocationID, timestamp=1709358115092, value=68
 2                                     column=records:PULocationID, timestamp=1709358115092, value=107
 2                                     column=records:RatecodeID, timestamp=1709358115092, value=1
 2                                     column=records:airport_fee, timestamp=1709358115092, value=
 2                                     column=records:congestion_surcharge, timestamp=1709358115092, value=
 2                                     column=records:extra, timestamp=1709358115092, value=0.5
 2                                     column=records:fare_amount, timestamp=1709358115092, value=6.0
 2                                     column=records:improvement_surcharge, timestamp=1709358115092, value=0.3
 2                                     column=records:mta_tax, timestamp=1709358115092, value=0.5
 2                                     column=records:passenger_count, timestamp=1709358115092, value=1
 2                                     column=records:payment_type, timestamp=1709358115092, value=1
 2                                     column=records:store_and_fwd_flag, timestamp=1709358115092, value=N
 2                                     column=records:tip_amount, timestamp=1709358115092, value=1.82
 2                                     column=records:tolls_amount, timestamp=1709358115092, value=0.0
 2                                     column=records:total_amount, timestamp=1709358115092, value=9.12
 2                                     column=records:tpep_dropoff_datetime, timestamp=1709358115092, value=2017-03-29 23:02:39
 2                                     column=records:tpep_pickup_datetime, timestamp=1709358115092, value=2017-03-29 22:57:02
 2                                     column=records:trip_distance, timestamp=1709358115092, value=1.07
```