INGESTION TASK:

Data ingestion can be done by various ways:

- Initially we have downloaded all 6 input files and pushed them to hdfs.

```
[hadoop@ip-172-31-66-221 ~]$ ls
mysql-connector-java-8.0.25         yellow_tripdata_2017-01.csv  yellow_tripdata_2017-03.csv  yellow_tripdata_2017-05.csv
mysql-connector-java-8.0.25.tar.gz  yellow_tripdata_2017-02.csv  yellow_tripdata_2017-04.csv  yellow_tripdata_2017-06.csv
[hadoop@ip-172-31-66-221 ~]$ hadoop fs -put /home/hadoop/yellow_tripdata_2017-01.csv /user/root/yellow_trip01
[hadoop@ip-172-31-66-221 ~]$ hadoop fs -put /home/hadoop/yellow_tripdata_2017-02.csv /user/root/yellow_trip02
[hadoop@ip-172-31-66-221 ~]$ hadoop fs -put /home/hadoop/yellow_tripdata_2017-03.csv /user/root/yellow_trip03
[hadoop@ip-172-31-66-221 ~]$ hadoop fs -put /home/hadoop/yellow_tripdata_2017-04.csv /user/root/yellow_trip04
[hadoop@ip-172-31-66-221 ~]$ hadoop fs -put /home/hadoop/yellow_tripdata_2017-05.csv /user/root/yellow_trip05
[hadoop@ip-172-31-66-221 ~]$ hadoop fs -put /home/hadoop/yellow_tripdata_2017-06.csv /user/root/yellow_trip06
[hadoop@ip-172-31-66-221 ~]$ hadoop fs -ls /user/root
Found 6 items
-rw-r--r--   1 hadoop hadoop  914029540 2023-02-14 15:00 /user/root/yellow_trip01
-rw-r--r--   1 hadoop hadoop  863487050 2023-02-14 15:00 /user/root/yellow_trip02
-rw-r--r--   1 hadoop hadoop  969809025 2023-02-14 15:01 /user/root/yellow_trip03
-rw-r--r--   1 hadoop hadoop  946349441 2023-02-14 15:01 /user/root/yellow_trip04
-rw-r--r--   1 hadoop hadoop  951965526 2023-02-14 15:01 /user/root/yellow_trip05
-rw-r--r--   1 hadoop hadoop  910028408 2023-02-14 15:01 /user/root/yellow_trip06
[hadoop@ip-172-31-66-221 ~]$
```

Here we have ingested data into hbase using 2 methods:

1. **Data ingestion using sqoop command.**
   - Created a hbase table “yellowTaxi”with column family as in the screenshot below.

```
hbase(main):002:0> describe "yellowTaxi"
Table yellowTaxi is ENABLED
yellowTaxi
COLUMN FAMILIES DESCRIPTION
{NAME => 'location', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DE
VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'payment_info', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEE
MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'
{NAME => 'trip_info', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_D
_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
3 row(s) in 0.0310 seconds
```

   - Dataset 1 and 2 which are loaded in RDS has been ingested to hbase table using sqoop command.

     For column family vendor

```
sqoop import \
--connect "jdbc:mysql:// yellowtaxi.ch1mu6dvhmli.us-east-
1.rds.amazonaws.com:3306/yellow" \
--username user –password 123 \
--table taxi \
--columns id,VendorID \
--hbase-create-table \
--hbase-table yellowTaxi \
--column-family vendor \
```

--hbase-row-key id

For column family trip_info

```
sqoop import \
--connect "jdbc:mysql:// yellowtaxi.ch1mu6dvhmli.us-east-1.rds.amazonaws.com:
3306/yellow" \
--username user –password 123 \
--table taxi \
--columns id, tpep_pickup_datetime, tpep_dropoff_datetime,
passenger_count,trip_distance, RatecodeID,store_and_fwd_flag \
--hbase-table yellowTaxi \
--column-family trip_info \
--hbase-row-key id
```

Same sqoop command for location and payment_info.

```
                Total megabyte-milliseconds taken by all map tasks=1921929216
        Map-Reduce Framework
                Map input records=18880595
                Map output records=18880595
                Input split bytes=438
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=9243
                CPU time spent (ms)=227760
                Physical memory (bytes) snapshot=2641354752
                Virtual memory (bytes) snapshot=13443190784
                Total committed heap usage (bytes)=2408579072
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
23/02/14 16:09:37 INFO mapreduce.ImportJobBase: Transferred 0 bytes in 336.9497 seconds (0 bytes/sec)
23/02/14 16:09:37 INFO mapreduce.ImportJobBase: Retrieved 18880595 records.
```

- Successfully got data transferred to hbase table.

```
hbase(main):003:0> scan "yellowTaxi",{LIMIT=>1}
ROW                                     COLUMN+CELL
 1                                      column=location:DOLocationID, timestamp=1676391656564, value=41
 1                                      column=location:PULocationID, timestamp=1676391656564, value=4
 1                                      column=payment_info:airport_fee, timestamp=1676394827688, value=0.0
 1                                      column=payment_info:congestion_surcharge, timestamp=1676394827688, value=0.0
 1                                      column=payment_info:extra, timestamp=1676394827688, value=0.500
 1                                      column=payment_info:fare_amount, timestamp=1676394827688, value=27.000
 1                                      column=payment_info:improvement_surcharge, timestamp=1676394827688, value=0.300
 1                                      column=payment_info:mta_tax, timestamp=1676394827688, value=0.500
 1                                      column=payment_info:payment_type, timestamp=1676394827688, value=2
 1                                      column=payment_info:tip_amount, timestamp=1676394827688, value=0.000
 1                                      column=payment_info:tolls_amount, timestamp=1676394827688, value=0.0
 1                                      column=payment_info:total_amount, timestamp=1676394827688, value=28.300
 1                                      column=trip_info:RatecodeID, timestamp=1676390966337, value=1
 1                                      column=trip_info:passenger_count, timestamp=1676389704510, value=1
 1                                      column=trip_info:store_and_fwd_flag, timestamp=1676390966337, value=N
 1                                      column=trip_info:tpep_dropoff_datetime, timestamp=1676389704510, value=2017-01-01 02:41:00.0
 1                                      column=trip_info:tpep_pickup_datetime, timestamp=1676389704510, value=2017-01-01 02:11:42.0
 1                                      column=trip_info:trip_distance, timestamp=1676389704510, value=7.5
 1                                      column=vendor:VendorID, timestamp=1676389091405, value=1
1 row(s) in 0.1020 seconds
```

2. **Bulk import data from EMR cluster to Hbase using batch insert. (batch insert code is in batch_ingest.py)**

```
=> ["yellowtaxi"]
hbase(main):003:0> get 'yellowtaxi','18880596'
COLUMN                                          CELL
 Vendor:VendorID                                timestamp=1676619367950, value=1
 location:DOLocationID                          timestamp=1676619367950, value=42
 location:PULocationID                          timestamp=1676619367950, value=231
 payment_info:airport_fee                       timestamp=1676619367950, value=
 payment_info:congestion_surcharge              timestamp=1676619367950, value=
 payment_info:extra                             timestamp=1676619367950, value=0.5
 payment_info:fare_amount                       timestamp=1676619367950, value=30.5
 payment_info:improvement_surcharge             timestamp=1676619367950, value=0.3
 payment_info:mta_tax                           timestamp=1676619367950, value=0.5
 payment_info:payment_type                      timestamp=1676619367950, value=1
 payment_info:tip_amount                        timestamp=1676619367950, value=6.0
 payment_info:tolls_amount                      timestamp=1676619367950, value=0.0
 payment_info:total_amount                      timestamp=1676619367950, value=37.8
 trip_info:RatecodeID                           timestamp=1676619367950, value=1
 trip_info:passenger_count                      timestamp=1676619367950, value=1
 trip_info:store_and_fwd_flag                   timestamp=1676619367950, value=N
 trip_info:tpep_dropoff_datetime                timestamp=1676619367950, value=2017-03-01 00:59:21
 trip_info:tpep_pickup_datetime                 timestamp=1676619367950, value=2017-03-01 00:38:16
 trip_info:trip_distance                        timestamp=1676619367950, value=10.5
1 row(s) in 0.1210 seconds

hbase(main):004:0> get 'yellowtaxi','18880597'
COLUMN                                          CELL
 Vendor:VendorID                                timestamp=1676619367950, value=1
 location:DOLocationID                          timestamp=1676619367950, value=262
 location:PULocationID                          timestamp=1676619367950, value=239
 payment_info:airport_fee                       timestamp=1676619367950, value=
 payment_info:congestion_surcharge              timestamp=1676619367950, value=
 payment_info:extra                             timestamp=1676619367950, value=0.5
 payment_info:fare_amount                       timestamp=1676619367950, value=7.5
 payment_info:improvement_surcharge             timestamp=1676619367950, value=0.3
 payment_info:mta_tax                           timestamp=1676619367950, value=0.5
 payment_info:payment_type                      timestamp=1676619367950, value=1
 payment_info:tip_amount                        timestamp=1676619367950, value=1.75
 payment_info:tolls_amount                      timestamp=1676619367950, value=0.0
 payment_info:total_amount                      timestamp=1676619367950, value=10.55
 trip_info:RatecodeID                           timestamp=1676619367950, value=1
 trip_info:passenger_count                      timestamp=1676619367950, value=1
 trip_info:store_and_fwd_flag                   timestamp=1676619367950, value=N
 trip_info:tpep_dropoff_datetime                timestamp=1676619367950, value=2017-03-01 00:31:36
 trip_info:tpep_pickup_datetime                 timestamp=1676619367950, value=2017-03-01 00:25:01
 trip_info:trip_distance                        timestamp=1676619367950, value=1.4
1 row(s) in 0.0430 seconds
```