Task 2

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

1. Login to EMR instance and copy mysql-connector in sqoop library folder.

```
ssh -i ~/Downloads/mr-assignment.pem hadoop@ec2-35-172-218-206.compute-1.amazonaws.com
wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
tar -xvf mysql-connector-java-8.0.25.tar.gz
cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/
```

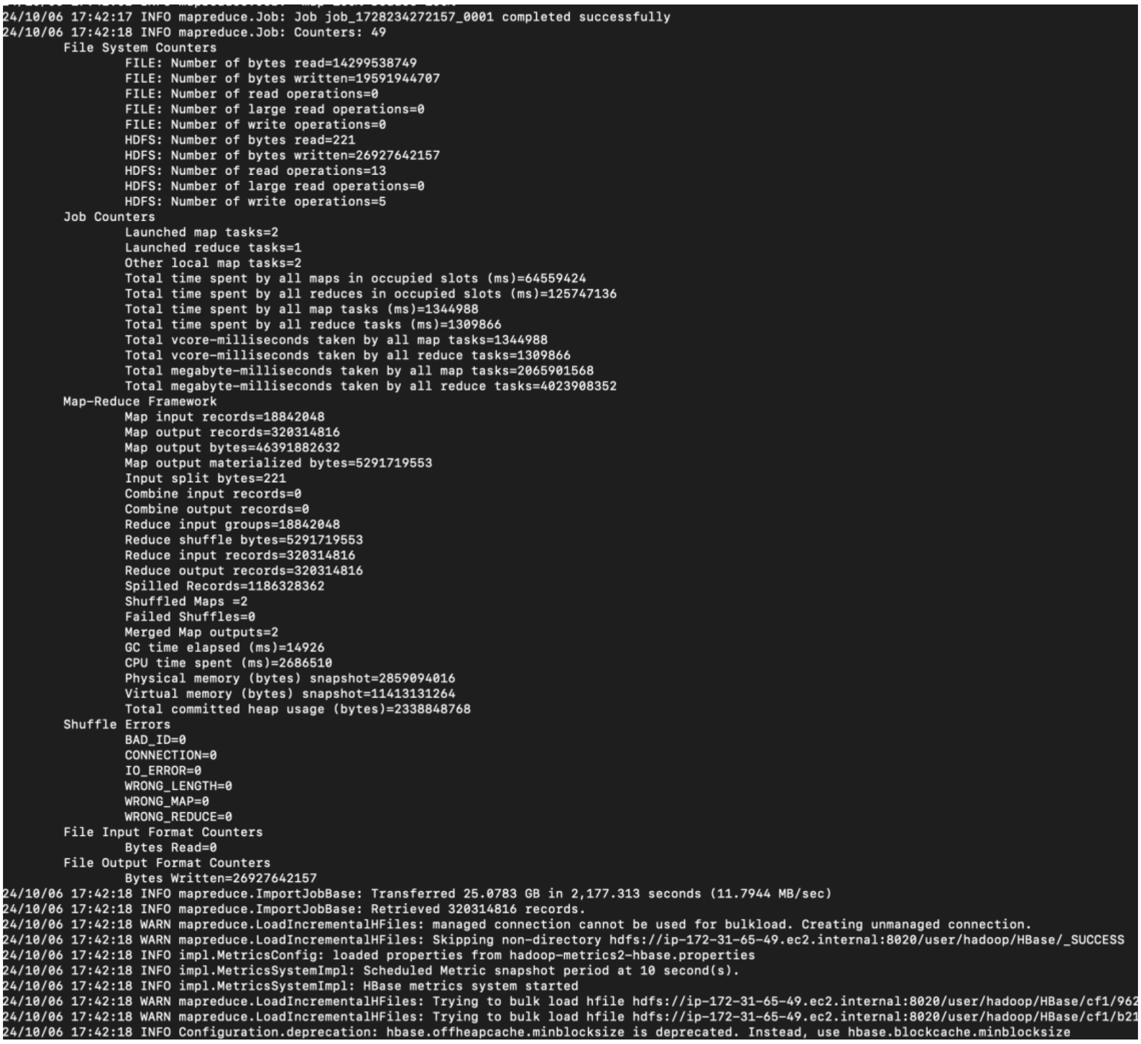
2. Ingest data from MySQL Database into HBase using Sqoop

```
sqoop import \
--connect "jdbc:mysql://mr-assignment-db.cj4l0tk6qknz.us-east-1.rds.amazonaws.com/taxi_trip_db" \
--username admin \
--password bucc1234 \
--table TripRecords \
--target-dir /user/hadoop/HBase \
--hbase-table trip_records_hbase --column-family cf1 --hbase-create-table \
--hbase-row-key tpep_pickup_datetime,tpep_dropoff_datetime \
--hbase-bulkload \
--split-by VendorID
```

Here:

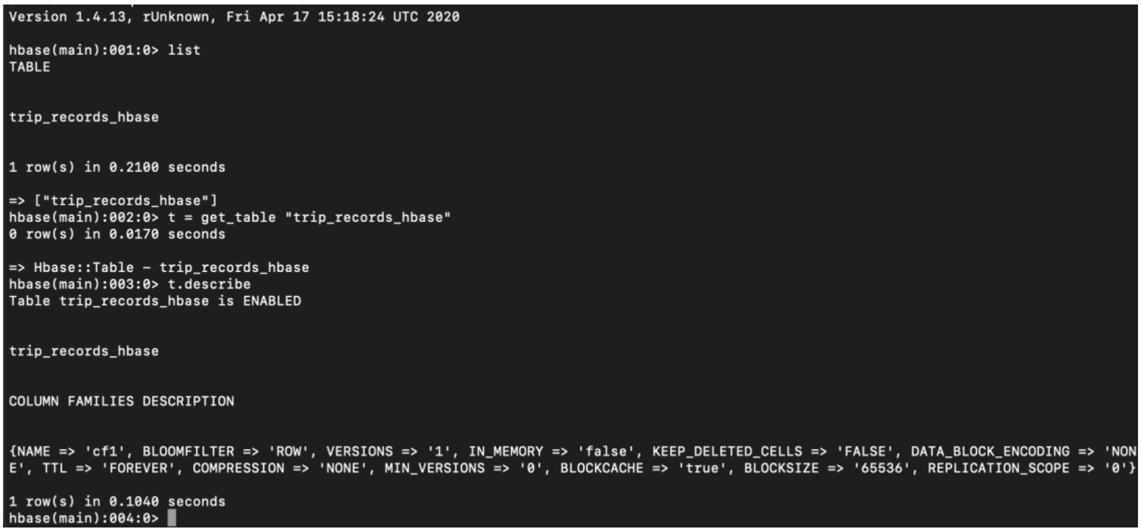
- sqoop import: Imports data from the RDS MySQL database:table “ taxi_trip_db: TripRecords” into HBase table trip_records_hbase.
- --connect is the JDBC string for the RDS MySQL database
- --target-dir is the path where this databse will be created
- --column-family cf1 as column family for the hbase table
- --hbase-create-table: creates an Hbase table as part of the import (here)
- --hbase-row-key : composite key using the two columns that will uniquely identify the Hbase data row “tpep_pickup_datetime,tpep_dropoff_datetime”
- --split-by : We use VendorID to split the data into multiple HBase regions.
- --hbase-bulkload : For fast upload of the data into hbase table

3. The output shows successful data ingestion to HBase.



```
24/10/06 17:42:17 INFO mapreduce.Job: Job job_1728234272157_0001 completed successfully
24/10/06 17:42:18 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=14299538749
                FILE: Number of bytes written=19591944707
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=221
                HDFS: Number of bytes written=26927642157
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=5
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=64559424
                Total time spent by all reduces in occupied slots (ms)=125747136
                Total time spent by all map tasks (ms)=1344988
                Total time spent by all reduce tasks (ms)=1309866
                Total vcore-milliseconds taken by all map tasks=1344988
                Total vcore-milliseconds taken by all reduce tasks=1309866
                Total megabyte-milliseconds taken by all map tasks=2065901568
                Total megabyte-milliseconds taken by all reduce tasks=4023908352
        Map-Reduce Framework
                Map input records=18842048
                Map output records=320314816
                Map output bytes=46391882632
                Map output materialized bytes=5291719553
                Input split bytes=221
                Combine input records=0
                Combine output records=0
                Reduce input groups=18842048
                Reduce shuffle bytes=5291719553
                Reduce input records=320314816
                Reduce output records=320314816
                Spilled Records=1186328362
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=14926
                CPU time spent (ms)=2686510
                Physical memory (bytes) snapshot=2859094016
                Virtual memory (bytes) snapshot=11413131264
                Total committed heap usage (bytes)=2338848768
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=26927642157
24/10/06 17:42:18 INFO mapreduce.ImportJobBase: Transferred 25.0783 GB in 2,177.313 seconds (11.7944 MB/sec)
24/10/06 17:42:18 INFO mapreduce.ImportJobBase: Retrieved 320314816 records.
24/10/06 17:42:18 WARN mapreduce.LoadIncrementalHFiles: managed connection cannot be used for bulkload. Creating unmanaged connection.
24/10/06 17:42:18 WARN mapreduce.LoadIncrementalHFiles: Skipping non-directory hdfs://ip-172-31-65-49.ec2.internal:8020/user/hadoop/HBase/_SUCCESS
24/10/06 17:42:18 INFO impl.MetricsConfig: loaded properties from hadoop-metrics2-hbase.properties
24/10/06 17:42:18 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
24/10/06 17:42:18 INFO impl.MetricsSystemImpl: HBase metrics system started
24/10/06 17:42:18 WARN mapreduce.LoadIncrementalHFiles: Trying to bulk load hfile hdfs://ip-172-31-65-49.ec2.internal:8020/user/hadoop/HBase/cf1/962
24/10/06 17:42:18 WARN mapreduce.LoadIncrementalHFiles: Trying to bulk load hfile hdfs://ip-172-31-65-49.ec2.internal:8020/user/hadoop/HBase/cf1/b21
24/10/06 17:42:18 INFO Configuration.deprecation: hbase.offheapcache.minblocksize is deprecated. Instead, use hbase.blockcache.minblocksize
```

4. Use HBase shell to verify the HBase table creation.



```
Version 1.4.13, rUnknown, Fri Apr 17 15:18:24 UTC 2020

hbase(main):001:0> list
TABLE


trip_records_hbase


1 row(s) in 0.2100 seconds

=> ["trip_records_hbase"]
hbase(main):002:0> t = get_table "trip_records_hbase"
0 row(s) in 0.0170 seconds

=> Hbase::Table - trip_records_hbase
hbase(main):003:0> t.describe
Table trip_records_hbase is ENABLED


trip_records_hbase


COLUMN FAMILIES DESCRIPTION


{NAME => 'cf1', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NON
E', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}

1 row(s) in 0.1040 seconds
hbase(main):004:0> 
```