Task 2: Using Sqoop commands to ingest the data from RDS into the HBase Table.

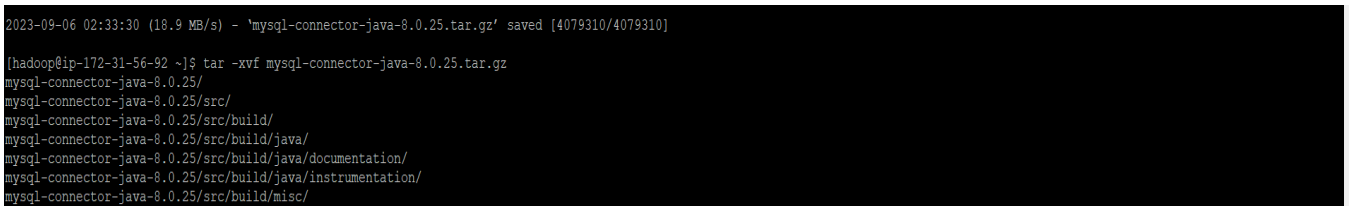
1. First, we log in into the EMR instance and complete the initial steps of setup.

- Now we run the following command to install the MySQL connector jar file

wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz

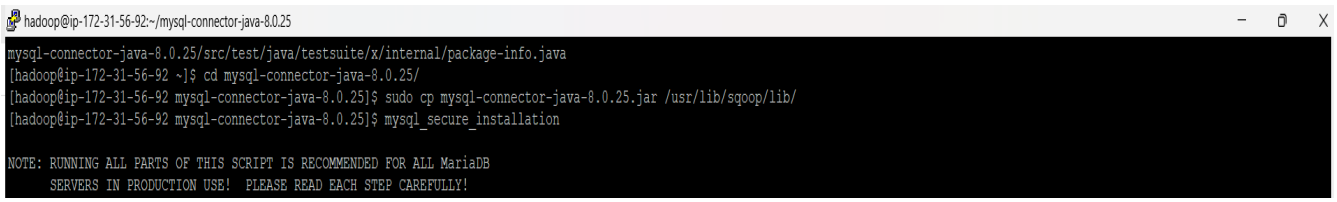
- Now, we run the following step to extract the MySQL connector tar file

  tar -xvf mysql-connector-java-8.0.25.tar.gz



- Now, we to go to the MySQL Connector directory created in the previous step and
  then copy it to the Sqoop library to complete the installation.

  cd mysql-connector-java-8.0.25/
  sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/



2. Having now installed the MySQL Connector. Now, we set up MySQL on EMR cluster and proceed

3. We then run the following command to ingest data from mySQL RDS to HBase table;

   Note: --hbase-create-table : creates an HBase table if it does not exist

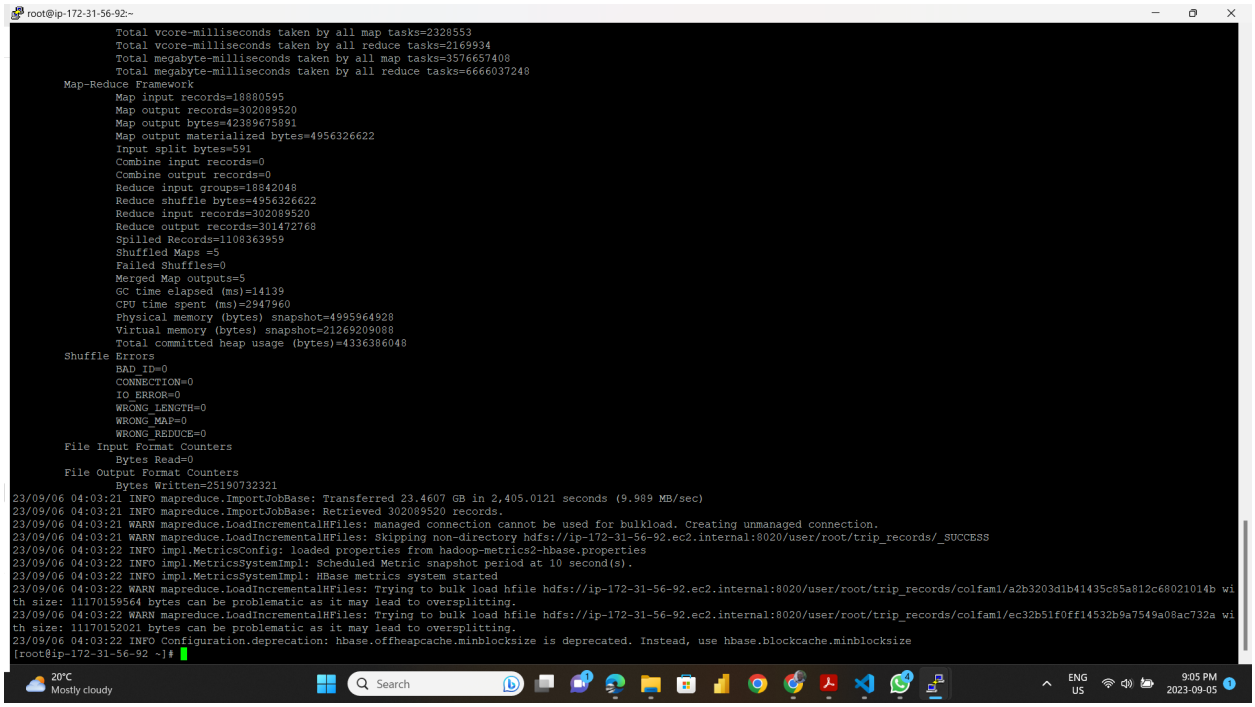
```
sqoop import --connect
jdbc:mysql://dbinstance.ca1depqqnwki.us-east-1.rds.amazonaws.com/yellow_taxi \
--username admin \
--password admin123 \
--table trip_records \
--hbase-table trip_records_hbase \
--column-family colfam1 \
--hbase-create-table \
--hbase-row-key tpep_pickup_datetime, tpep_dropoff_datetime \
--hbase-bulkload \
--split-by payment_type
```

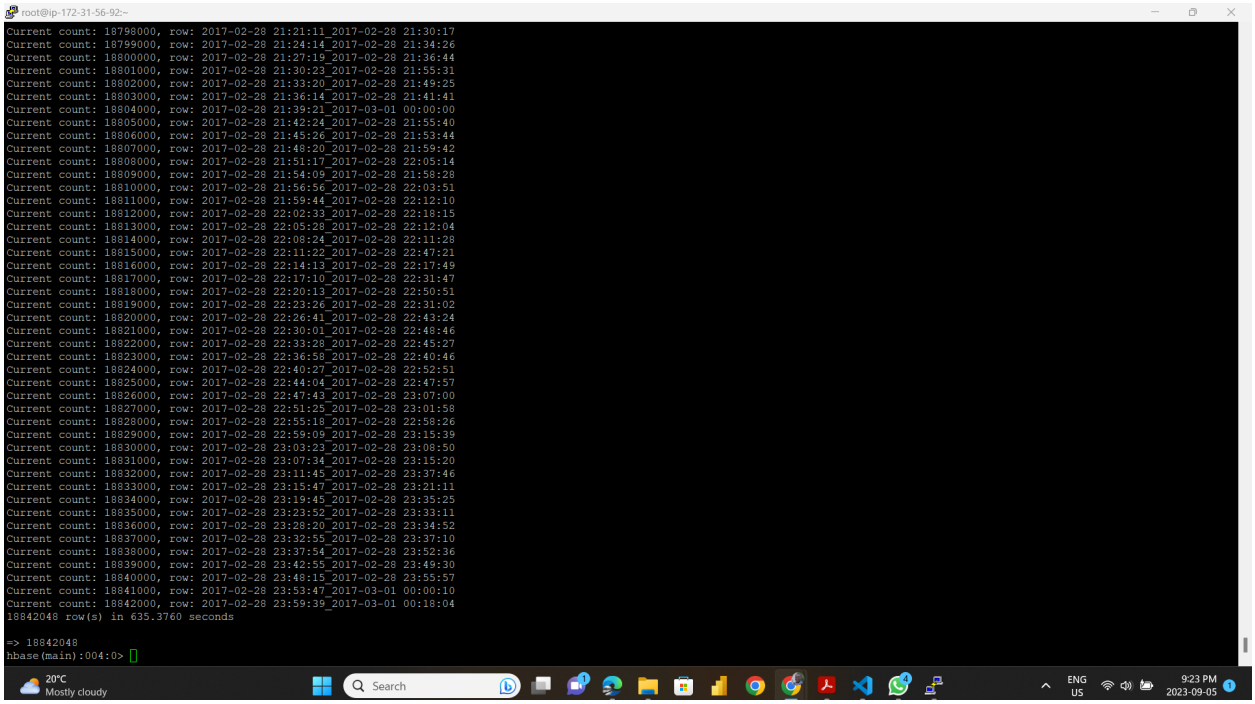
4. Code explanation :

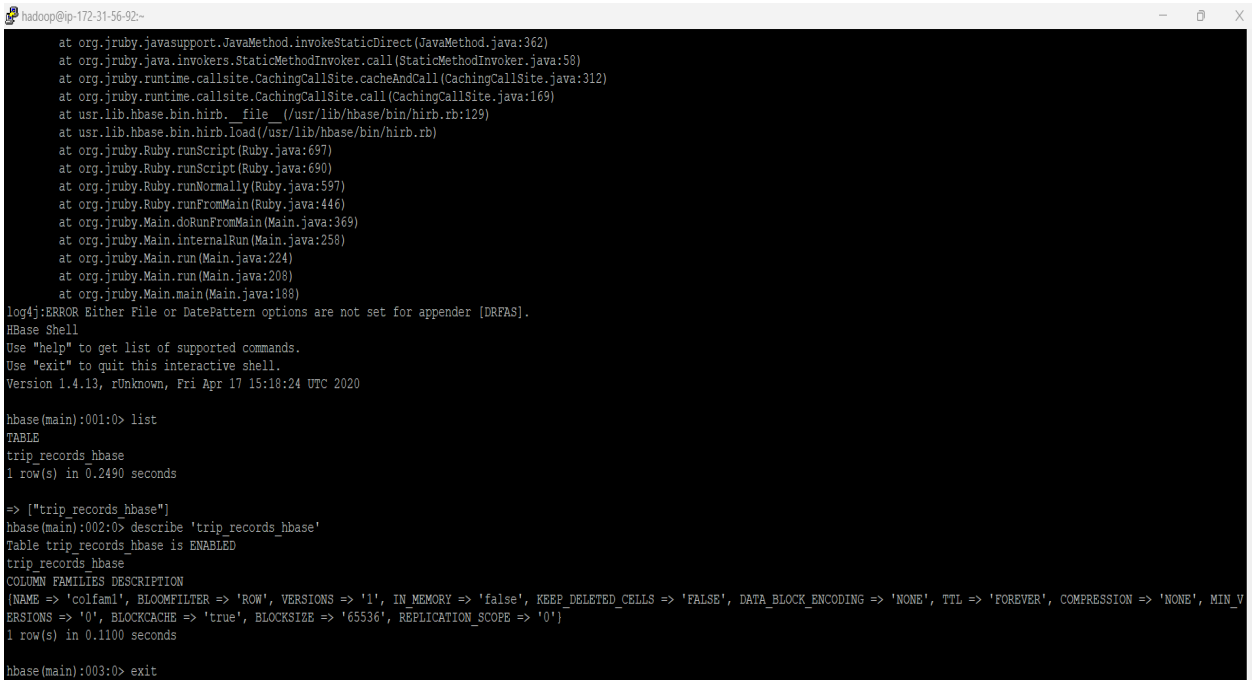
This Sqoop command transfers data from a MySQL database table named trip_records to an HBase table called trip_records_hbase. The function of each option in the command is broken down below;

--split-by : specifies a column from the MySQL table that will be used to split data into multiple HBase regions
--hbase-bulkload : uses HBase bulk load feature for faster data loading
--hbase-row-key : specifies one or more columns from the MySQL table that will be used as the row key in HBase
--hbase-create-table : creates an HBase table if it does not exist
--column-family : specifies the name of the column family in HBase where the imported data will be stored
--hbase-table : specifies the name of the HBase table to import data into
--table : specifies the name of the MySQL table to import data from
--password : specifies the password to use when connecting to the MySQL database
--username : specifies the username to use when connecting to the MySQL database

--connect : specifies the JDBC connection string for the MySQL database






hadoop@ip-172-31-56-92:~
at org.jruby.javasupport.JavaMethod.invokeStaticDirect(JavaMethod.java:362)
at org.jruby.java.invokers.StaticMethodInvoker.call(StaticMethodInvoker.java:58)
at org.jruby.runtime.callsite.CachingCallSite.cacheAndCall(CachingCallSite.java:312)
at org.jruby.runtime.callsite.CachingCallSite.call(CachingCallSite.java:169)
at usr.lib.hbase.bin.hirb.__file__(/usr/lib/hbase/bin/hirb.rb:129)
at usr.lib.hbase.bin.hirb.load(/usr/lib/hbase/bin/hirb.rb)
at org.jruby.Ruby.runScript(Ruby.java:697)
at org.jruby.Ruby.runScript(Ruby.java:690)
at org.jruby.Ruby.runNormally(Ruby.java:597)
at org.jruby.Ruby.runFromMain(Ruby.java:446)
at org.jruby.Main.doRunFromMain(Main.java:369)
at org.jruby.Main.internalRun(Main.java:258)
at org.jruby.Main.run(Main.java:224)
at org.jruby.Main.run(Main.java:208)
at org.jruby.Main.main(Main.java:188)
log4j:ERROR Either File or DatePattern options are not set for appender [DRFAS].
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
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 0.2490 seconds
=> ["trip_records_hbase"]
hbase(main):002:0> describe 'trip_records_hbase'
Table trip_records_hbase is ENABLED
trip_records_hbase
COLUMN FAMILIES DESCRIPTION
{NAME => 'colfam1', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_V
ERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.1100 seconds
hbase(main):003:0> exit