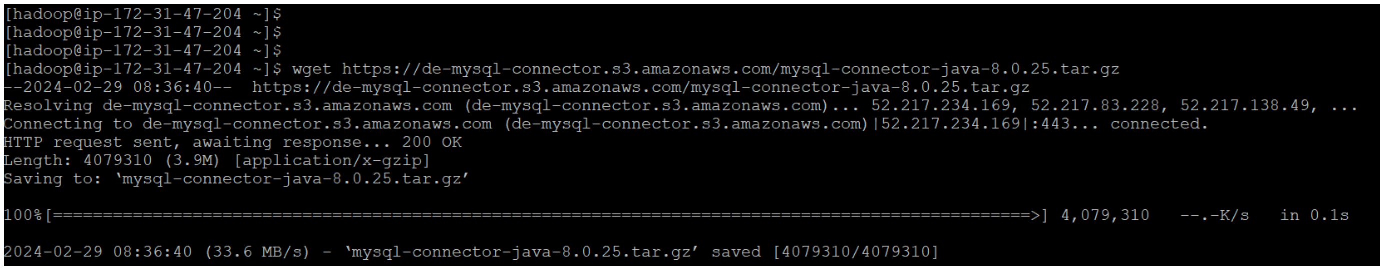
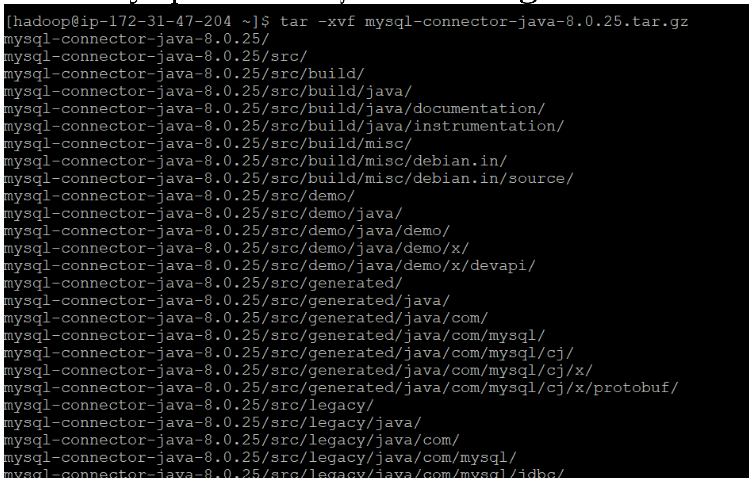
Task 2. Use Sqoop command to ingest the data from RDS into the HBase Table

Logged into EMR cluster
Executed the following command to install MySQL connector jar file
wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz



```
[hadoop@ip-172-31-47-204 ~]$
[hadoop@ip-172-31-47-204 ~]$
[hadoop@ip-172-31-47-204 ~]$
[hadoop@ip-172-31-47-204 ~]$ wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
--2024-02-29 08:36:40--  https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
Resolving de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)... 52.217.234.169, 52.217.83.228, 52.217.138.49, ...
Connecting to de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)|52.217.234.169|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4079310 (3.9M) [application/x-gzip]
Saving to: ‘mysql-connector-java-8.0.25.tar.gz’

100%[==========================================================================================>] 4,079,310   --.-K/s   in 0.1s

2024-02-29 08:36:40 (33.6 MB/s) - ‘mysql-connector-java-8.0.25.tar.gz’ saved [4079310/4079310]
```

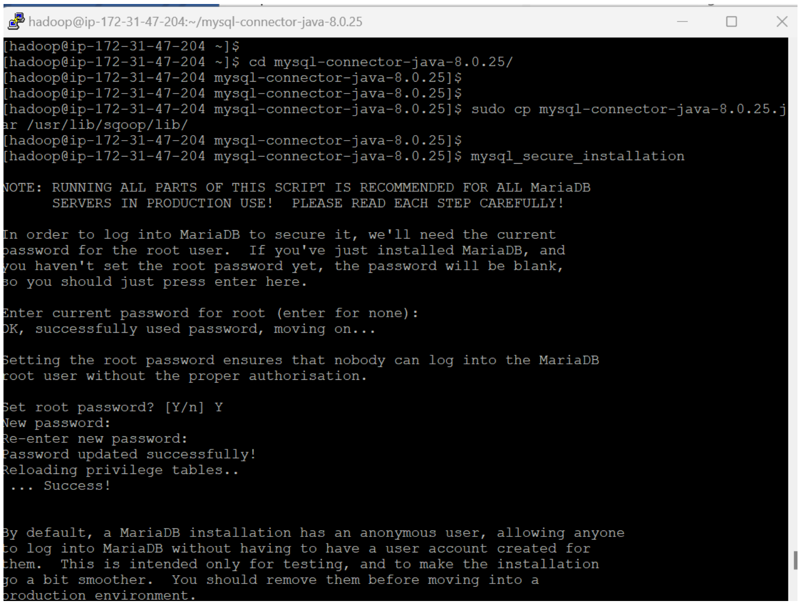
Extracted MySQL connector tar file:
tar -xvf mysql-connector-java-8.0.25.tar.gz



```
[hadoop@ip-172-31-47-204 ~]$ tar -xvf mysql-connector-java-8.0.25.tar.gz
mysql-connector-java-8.0.25/
mysql-connector-java-8.0.25/src/
mysql-connector-java-8.0.25/src/build/
mysql-connector-java-8.0.25/src/build/java/
mysql-connector-java-8.0.25/src/build/java/documentation/
mysql-connector-java-8.0.25/src/build/java/instrumentation/
mysql-connector-java-8.0.25/src/build/misc/
mysql-connector-java-8.0.25/src/build/misc/debian.in/
mysql-connector-java-8.0.25/src/build/misc/debian.in/source/
mysql-connector-java-8.0.25/src/demo/
mysql-connector-java-8.0.25/src/demo/java/
mysql-connector-java-8.0.25/src/demo/java/demo/
mysql-connector-java-8.0.25/src/demo/java/demo/x/
mysql-connector-java-8.0.25/src/demo/java/demo/x/devapi/
mysql-connector-java-8.0.25/src/generated/
mysql-connector-java-8.0.25/src/generated/java/
mysql-connector-java-8.0.25/src/generated/java/com/
mysql-connector-java-8.0.25/src/generated/java/com/mysql/
mysql-connector-java-8.0.25/src/generated/java/com/mysql/cj/
mysql-connector-java-8.0.25/src/generated/java/com/mysql/cj/x/
mysql-connector-java-8.0.25/src/generated/java/com/mysql/cj/x/protobuf/
mysql-connector-java-8.0.25/src/legacy/
mysql-connector-java-8.0.25/src/legacy/java/
mysql-connector-java-8.0.25/src/legacy/java/com/
mysql-connector-java-8.0.25/src/legacy/java/com/mysql/
mysql-connector-java-8.0.25/src/legacy/java/com/mysql/jdbc/
```

Changed the directory to the MySQL Connector directory and copied to the Sqoop library as follows:

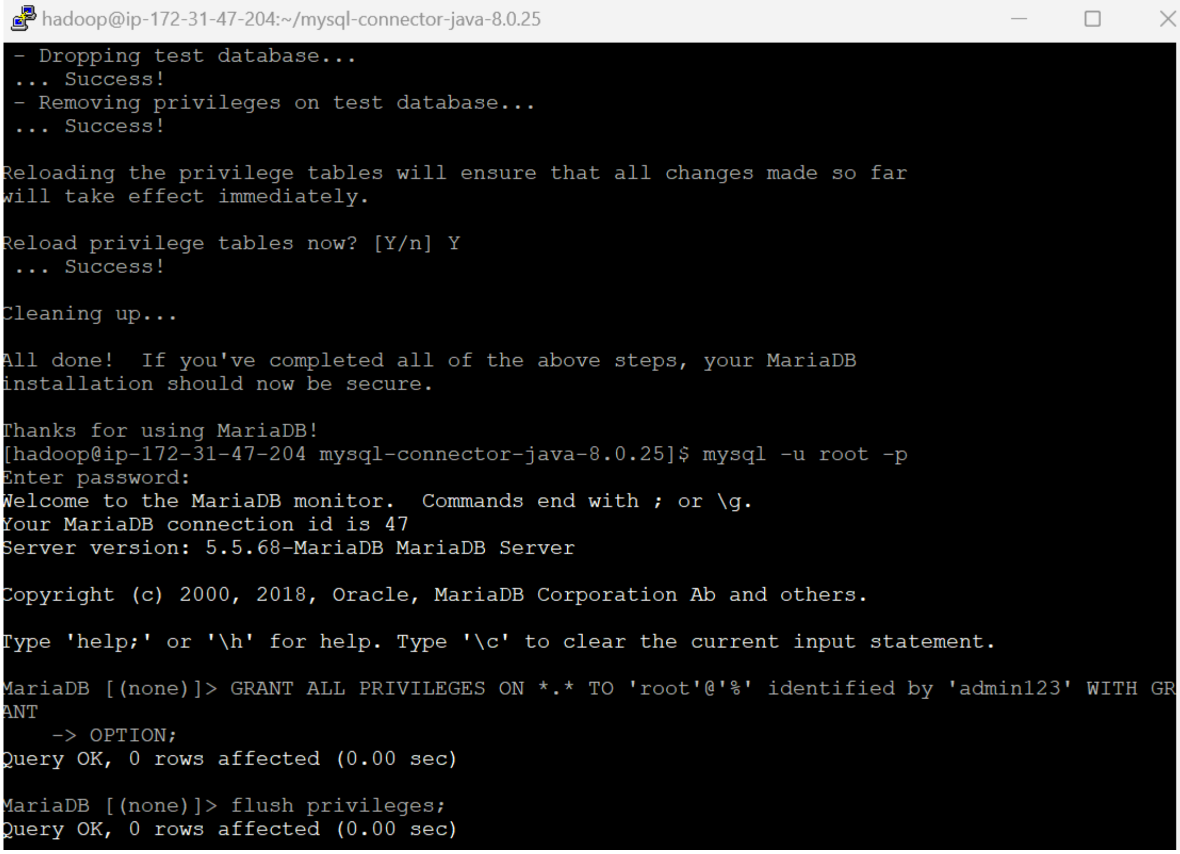
- cd mysql-connector-java-8.0.25/
- sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/



hadoop@ip-172-31-47-204:~/mysql-connector-java-8.0.25

```
[hadoop@ip-172-31-47-204 ~]$
[hadoop@ip-172-31-47-204 ~]$ cd mysql-connector-java-8.0.25/
[hadoop@ip-172-31-47-204 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-47-204 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-47-204 mysql-connector-java-8.0.25]$ sudo cp mysql-connector-java-8.0.25.j
ar /usr/lib/sqoop/lib/
[hadoop@ip-172-31-47-204 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-47-204 mysql-connector-java-8.0.25]$ mysql_secure_installation

NOTE: RUNNING ALL PARTS OF THIS SCRIPT IS RECOMMENDED FOR ALL MariaDB
      SERVERS IN PRODUCTION USE!  PLEASE READ EACH STEP CAREFULLY!

In order to log into MariaDB to secure it, we'll need the current
password for the root user.  If you've just installed MariaDB, and
you haven't set the root password yet, the password will be blank,
so you should just press enter here.

Enter current password for root (enter for none):
OK, successfully used password, moving on...

Setting the root password ensures that nobody can log into the MariaDB
root user without the proper authorisation.

Set root password? [Y/n] Y
New password:
Re-enter new password:
Password updated successfully!
Reloading privilege tables..
 ... Success!


By default, a MariaDB installation has an anonymous user, allowing anyone
to log into MariaDB without having to have a user account created for
them.  This is intended only for testing, and to make the installation
go a bit smoother.  You should remove them before moving into a
production environment.
```

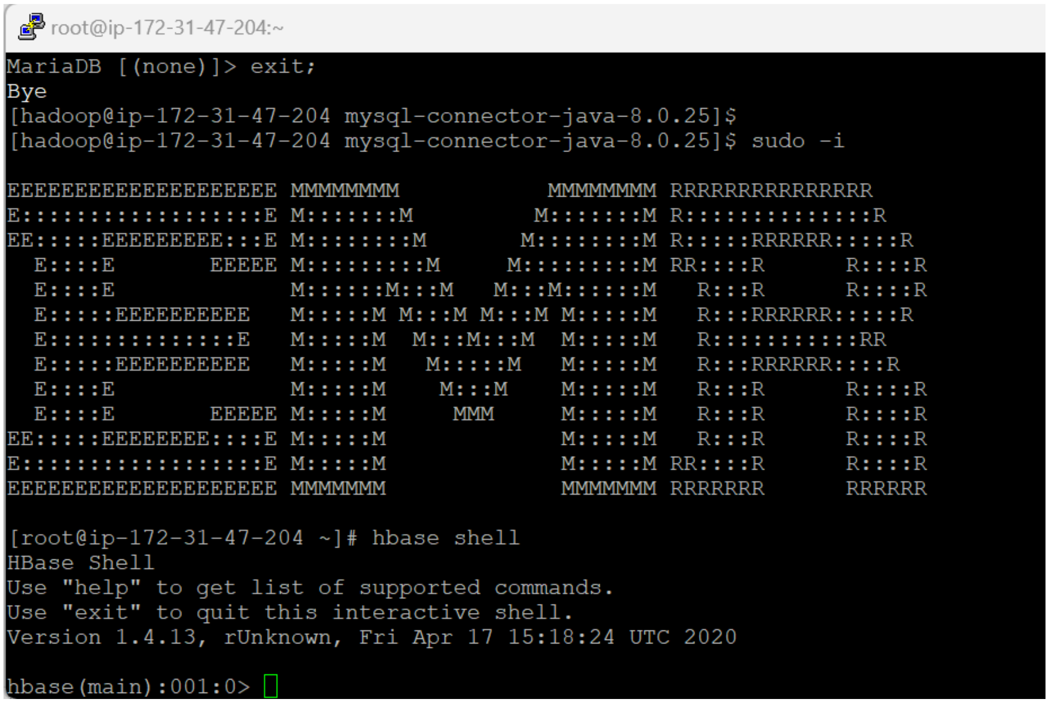
Installed & setup MySQL Connector on EMR cluster:



```
 - Dropping test database...
 ... Success!
 - Removing privileges on test database...
 ... Success!

Reloading the privilege tables will ensure that all changes made so far
will take effect immediately.

Reload privilege tables now? [Y/n] Y
 ... Success!

Cleaning up...

All done!  If you've completed all of the above steps, your MariaDB
installation should now be secure.

Thanks for using MariaDB!
[hadoop@ip-172-31-47-204 mysql-connector-java-8.0.25]$ mysql -u root -p
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MariaDB connection id is 47
Server version: 5.5.68-MariaDB MariaDB Server

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [(none)]> GRANT ALL PRIVILEGES ON *.* TO 'root'@'%' identified by 'admin123' WITH GR
ANT
    -> OPTION;
Query OK, 0 rows affected (0.00 sec)

MariaDB [(none)]> flush privileges;
Query OK, 0 rows affected (0.00 sec)
```

Accessed HBase shell on EMR using PuTTY

- Switched to 'root' user using the **sudo -i** command
  sudo -i
- Access the HBase shell by using the "hbase shell" command
  hbase shell



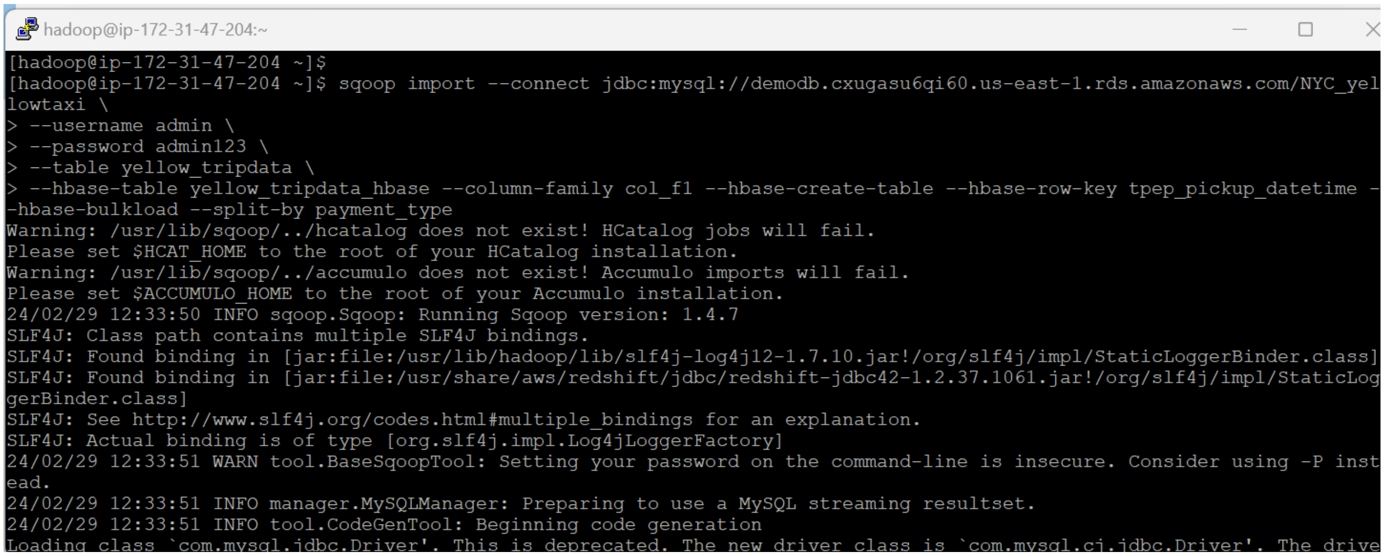
```
MariaDB [(none)]> exit;
Bye
[hadoop@ip-172-31-47-204 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-47-204 mysql-connector-java-8.0.25]$ sudo -i

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[root@ip-172-31-47-204 ~]# hbase shell
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>
```

Ingested data from RDS to HBase table using following command:

```
sqoop import --connect jdbc:mysql://demodb.cxugasu6qi60.us-east-1.rds.amazonaws.com/NYC_yellowtaxi \
--username admin \
--password admin123 \
--table yellow_tripdata \
--hbase-table yellow_tripdata_hbase \
--column-family col_f1 --hbase-create-table --hbase-row-key tpep_pickup_datetime --hbase-bulkload
--split-by payment_type
```

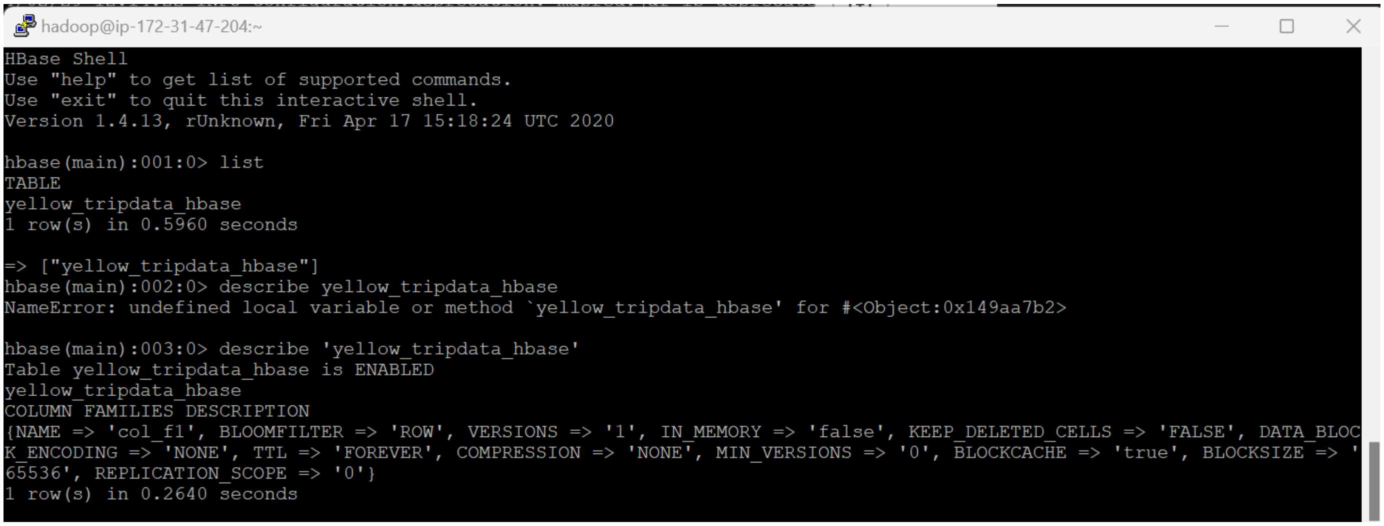
This command transfers data from RDS table "yellow_tripdata" to an HBase table "yellow_tripdata_HBase".

--connect: JDBC connection string for the RDS(MySQL) database
--username: username to use to connect to the Database
--password: password to use to connect to the Database
--table: name of the source (MySQL) table
--hbase-table: name of the target (HBase) table
--column-family: name of the column family in HBase table
--hbase-create-table: creates HBase table (in case not exists)
--hbase-row-key: column/s of source table to be used as key in HBase table
--hbase-bulkload: Enables bulk loading
--split-by: column used to create the split while importing the data into the cluster



```
[hadoop@ip-172-31-47-204 ~]$
[hadoop@ip-172-31-47-204 ~]$ sqoop import --connect jdbc:mysql://demodb.cxugasu6qi60.us-east-1.rds.amazonaws.com/NYC_yel
lowtaxi \
> --username admin \
> --password admin123 \
> --table yellow_tripdata \
> --hbase-table yellow_tripdata_hbase --column-family col_f1 --hbase-create-table --hbase-row-key tpep_pickup_datetime -
-hbase-bulkload --split-by payment_type
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
24/02/29 12:33:50 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLog
gerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/02/29 12:33:51 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P inst
ead.
24/02/29 12:33:51 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/02/29 12:33:51 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The drive
```

Checked the HBase table and its statistics created by Sqoop command in HBase shell:



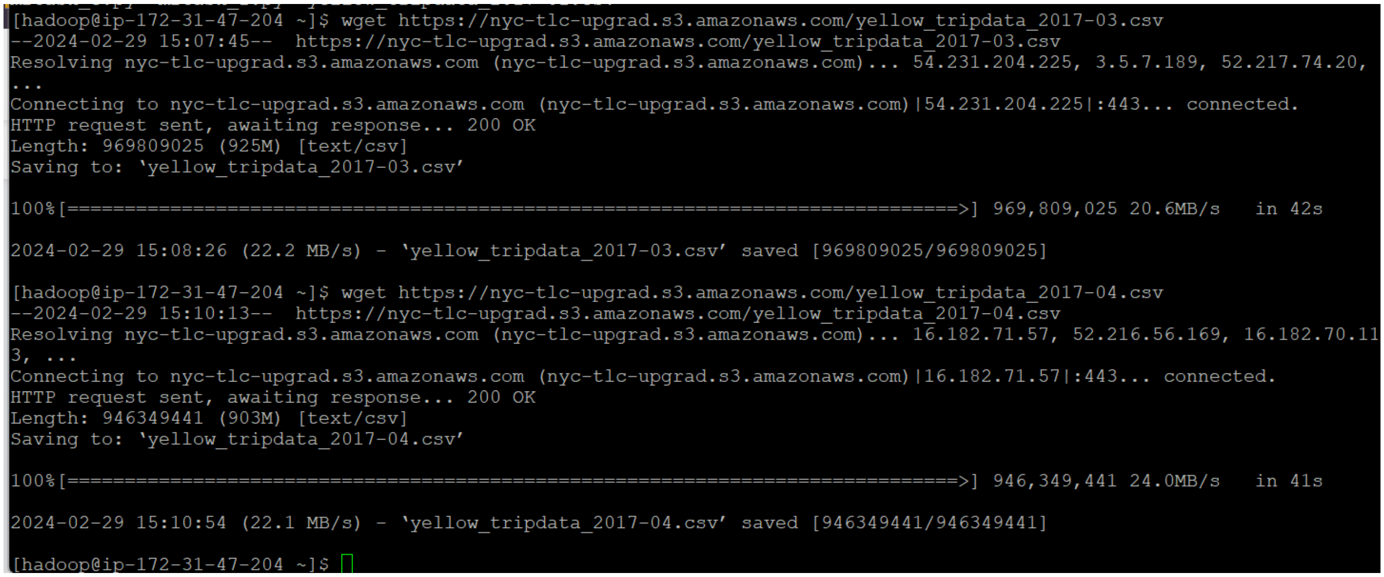
```
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.13, rUnknown, Fri Apr 17 15:18:24 UTC 2020

hbase(main):001:0> list
TABLE
yellow_tripdata_hbase
1 row(s) in 0.5960 seconds

=> ["yellow_tripdata_hbase"]
hbase(main):002:0> describe yellow_tripdata_hbase
NameError: undefined local variable or method `yellow_tripdata_hbase' for #<Object:0x149aa7b2>

hbase(main):003:0> describe 'yellow_tripdata_hbase'
Table yellow_tripdata_hbase is ENABLED
yellow_tripdata_hbase
COLUMN FAMILIES DESCRIPTION
{NAME => 'col_f1', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOC
K_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '
65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.2640 seconds
```

****

Task 3. Bulk import data from next two files in the dataset on your EMR cluster to your HBase Table using the relevant codes.
Note: For the above task 3, you just need to import data from the subsequent 2 csv files (*i.e.* yellow_tripdata_2017-03.csv & yellow_tripdata_2017-04.csv) on your EMR cluster.

Downloaded the necessary data on local file system.

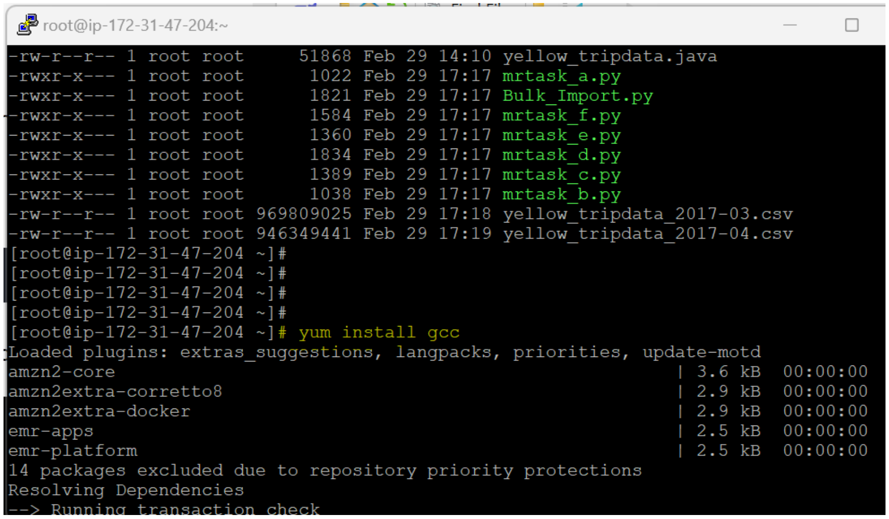
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-03.csv
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-04.csv



```
[hadoop@ip-172-31-47-204 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-03.csv
--2024-02-29 15:07:45--  https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-03.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 54.231.204.225, 3.5.7.189, 52.217.74.20,
...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|54.231.204.225|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 969809025 (925M) [text/csv]
Saving to: ‘yellow_tripdata_2017-03.csv’

100%[=============================================================================>] 969,809,025 20.6MB/s   in 42s

2024-02-29 15:08:26 (22.2 MB/s) - ‘yellow_tripdata_2017-03.csv’ saved [969809025/969809025]

[hadoop@ip-172-31-47-204 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-04.csv
--2024-02-29 15:10:13--  https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-04.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 16.182.71.57, 52.216.56.169, 16.182.70.11
3, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|16.182.71.57|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 946349441 (903M) [text/csv]
Saving to: ‘yellow_tripdata_2017-04.csv’

100%[=============================================================================>] 946,349,441 24.0MB/s   in 41s

2024-02-29 15:10:54 (22.1 MB/s) - ‘yellow_tripdata_2017-04.csv’ saved [946349441/946349441]

[hadoop@ip-172-31-47-204 ~]$
```

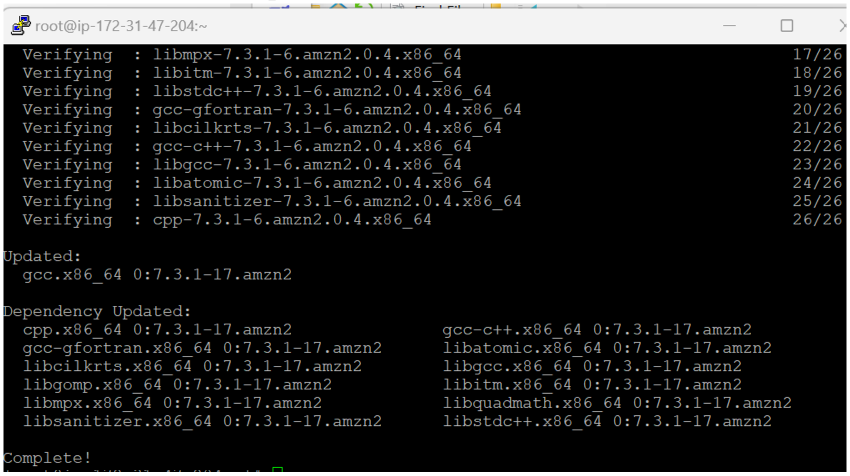
Setting up the environment to install HappyBase (Python-based HBase API) for executing the batch script as follows:

Installing gcc:
yum install gcc



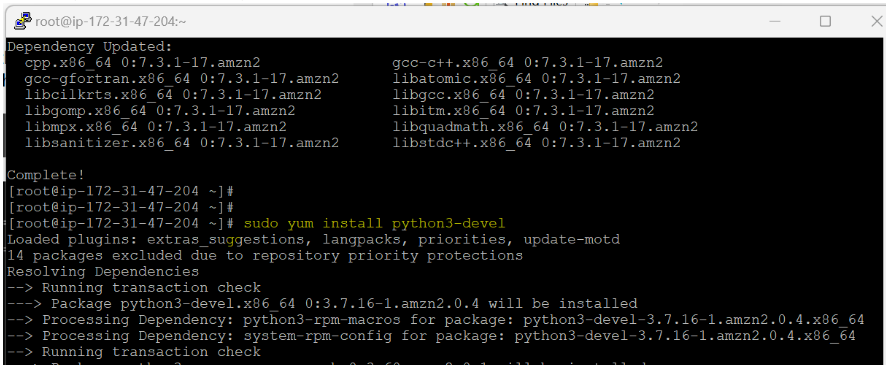
```
-rw-r--r-- 1 root root      51868 Feb 29 14:10 yellow_tripdata.java
-rwxr-x--- 1 root root       1022 Feb 29 17:17 mrtask_a.py
-rwxr-x--- 1 root root       1821 Feb 29 17:17 Bulk_Import.py
-rwxr-x--- 1 root root       1584 Feb 29 17:17 mrtask_f.py
-rwxr-x--- 1 root root       1360 Feb 29 17:17 mrtask_e.py
-rwxr-x--- 1 root root       1834 Feb 29 17:17 mrtask_d.py
-rwxr-x--- 1 root root       1389 Feb 29 17:17 mrtask_c.py
-rwxr-x--- 1 root root       1038 Feb 29 17:17 mrtask_b.py
-rw-r--r-- 1 root root 969809025 Feb 29 17:18 yellow_tripdata_2017-03.csv
-rw-r--r-- 1 root root 946349441 Feb 29 17:19 yellow_tripdata_2017-04.csv
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]# yum install gcc
Loaded plugins: extras_suggestions, langpacks, priorities, update-motd
amzn2-core                                               | 3.6 kB  00:00:00
amzn2extra-corretto8                                     | 2.9 kB  00:00:00
amzn2extra-docker                                        | 2.9 kB  00:00:00
emr-apps                                                 | 2.5 kB  00:00:00
emr-platform                                             | 2.5 kB  00:00:00
14 packages excluded due to repository priority protections
Resolving Dependencies
--> Running transaction check
```

Gcc installation completed



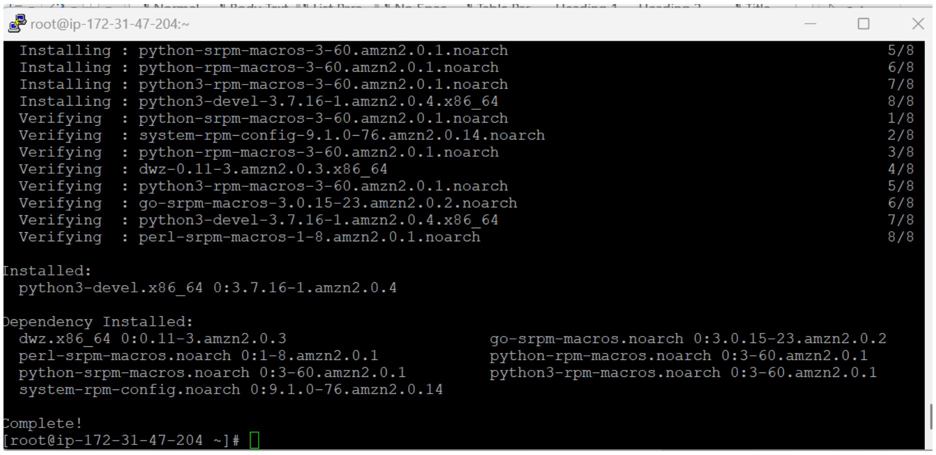
```
  Verifying  : libmpx-7.3.1-6.amzn2.0.4.x86_64                            17/26
  Verifying  : libitm-7.3.1-6.amzn2.0.4.x86_64                            18/26
  Verifying  : libstdc++-7.3.1-6.amzn2.0.4.x86_64                         19/26
  Verifying  : gcc-gfortran-7.3.1-6.amzn2.0.4.x86_64                      20/26
  Verifying  : libcilkrts-7.3.1-6.amzn2.0.4.x86_64                        21/26
  Verifying  : gcc-c++-7.3.1-6.amzn2.0.4.x86_64                           22/26
  Verifying  : libgcc-7.3.1-6.amzn2.0.4.x86_64                            23/26
  Verifying  : libatomic-7.3.1-6.amzn2.0.4.x86_64                         24/26
  Verifying  : libsanitizer-7.3.1-6.amzn2.0.4.x86_64                      25/26
  Verifying  : cpp-7.3.1-6.amzn2.0.4.x86_64                               26/26

Updated:
  gcc.x86_64 0:7.3.1-17.amzn2

Dependency Updated:
  cpp.x86_64 0:7.3.1-17.amzn2                  gcc-c++.x86_64 0:7.3.1-17.amzn2
  gcc-gfortran.x86_64 0:7.3.1-17.amzn2         libatomic.x86_64 0:7.3.1-17.amzn2
  libcilkrts.x86_64 0:7.3.1-17.amzn2           libgcc.x86_64 0:7.3.1-17.amzn2
  libgomp.x86_64 0:7.3.1-17.amzn2              libitm.x86_64 0:7.3.1-17.amzn2
  libmpx.x86_64 0:7.3.1-17.amzn2               libquadmath.x86_64 0:7.3.1-17.amzn2
  libsanitizer.x86_64 0:7.3.1-17.amzn2         libstdc++.x86_64 0:7.3.1-17.amzn2

Complete!
```

Installing HappyBase package:
sudo yum install python3-devel



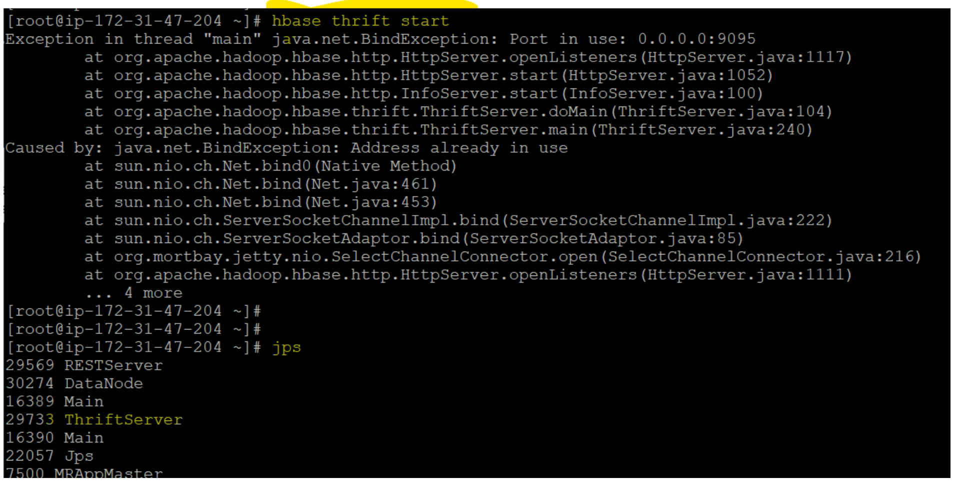
```
Dependency Updated:
  cpp.x86_64 0:7.3.1-17.amzn2                 gcc-c++.x86_64 0:7.3.1-17.amzn2
  gcc-gfortran.x86_64 0:7.3.1-17.amzn2        libatomic.x86_64 0:7.3.1-17.amzn2
  libcilkrts.x86_64 0:7.3.1-17.amzn2          libgcc.x86_64 0:7.3.1-17.amzn2
  libgomp.x86_64 0:7.3.1-17.amzn2             libitm.x86_64 0:7.3.1-17.amzn2
  libmpx.x86_64 0:7.3.1-17.amzn2              libquadmath.x86_64 0:7.3.1-17.amzn2
  libsanitizer.x86_64 0:7.3.1-17.amzn2        libstdc++.x86_64 0:7.3.1-17.amzn2

Complete!
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]# sudo yum install python3-devel
Loaded plugins: extras_suggestions, langpacks, priorities, update-motd
14 packages excluded due to repository priority protections
Resolving Dependencies
--> Running transaction check
---> Package python3-devel.x86_64 0:3.7.16-1.amzn2.0.4 will be installed
--> Processing Dependency: python3-rpm-macros for package: python3-devel-3.7.16-1.amzn2.0.4.x86_64
--> Processing Dependency: system-rpm-config for package: python3-devel-3.7.16-1.amzn2.0.4.x86_64
--> Running transaction check
```

Installation completed



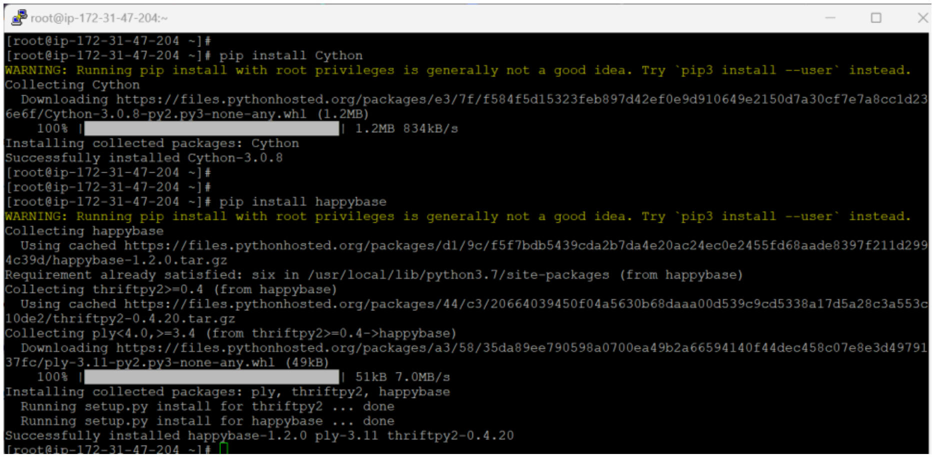
```
  Installing : python-srpm-macros-3-60.amzn2.0.1.noarch                                     5/8
  Installing : python-rpm-macros-3-60.amzn2.0.1.noarch                                      6/8
  Installing : python3-rpm-macros-3-60.amzn2.0.1.noarch                                     7/8
  Installing : python3-devel-3.7.16-1.amzn2.0.4.x86_64                                      8/8
  Verifying  : python-srpm-macros-3-60.amzn2.0.1.noarch                                     1/8
  Verifying  : system-rpm-config-9.1.0-76.amzn2.0.14.noarch                                 2/8
  Verifying  : python-rpm-macros-3-60.amzn2.0.1.noarch                                      3/8
  Verifying  : dwz-0.11-3.amzn2.0.3.x86_64                                                  4/8
  Verifying  : python3-rpm-macros-3-60.amzn2.0.1.noarch                                     5/8
  Verifying  : go-srpm-macros-3.0.15-23.amzn2.0.2.noarch                                    6/8
  Verifying  : python3-devel-3.7.16-1.amzn2.0.4.x86_64                                      7/8
  Verifying  : perl-srpm-macros-1-8.amzn2.0.1.noarch                                        8/8

Installed:
  python3-devel.x86_64 0:3.7.16-1.amzn2.0.4

ependency Installed:
  dwz.x86_64 0:0.11-3.amzn2.0.3                      go-srpm-macros.noarch 0:3.0.15-23.amzn2.0.2
  perl-srpm-macros.noarch 0:1-8.amzn2.0.1            python-rpm-macros.noarch 0:3-60.amzn2.0.1
  python-srpm-macros.noarch 0:3-60.amzn2.0.1         python3-rpm-macros.noarch 0:3-60.amzn2.0.1
  system-rpm-config.noarch 0:9.1.0-76.amzn2.0.14

omplete!
[root@ip-172-31-47-204 ~]#
```

Starting the ThriftServer:
hbase thrift start



```
[root@ip-172-31-47-204 ~]# hbase thrift start
Exception in thread "main" java.net.BindException: Port in use: 0.0.0.0:9095
        at org.apache.hadoop.hbase.http.HttpServer.openListeners(HttpServer.java:1117)
        at org.apache.hadoop.hbase.http.HttpServer.start(HttpServer.java:1052)
        at org.apache.hadoop.hbase.http.InfoServer.start(InfoServer.java:100)
        at org.apache.hadoop.hbase.thrift.ThriftServer.doMain(ThriftServer.java:104)
        at org.apache.hadoop.hbase.thrift.ThriftServer.main(ThriftServer.java:240)
Caused by: java.net.BindException: Address already in use
        at sun.nio.ch.Net.bind0(Native Method)
        at sun.nio.ch.Net.bind(Net.java:461)
        at sun.nio.ch.Net.bind(Net.java:453)
        at sun.nio.ch.ServerSocketChannelImpl.bind(ServerSocketChannelImpl.java:222)
        at sun.nio.ch.ServerSocketAdaptor.bind(ServerSocketAdaptor.java:85)
        at org.mortbay.jetty.nio.SelectChannelConnector.open(SelectChannelConnector.java:216)
        at org.apache.hadoop.hbase.http.HttpServer.openListeners(HttpServer.java:1111)
        ... 4 more
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]# jps
29569 RESTServer
30274 DataNode
16389 Main
29733 ThriftServer
16390 Main
22057 Jps
7500 MRAppMaster
```

Installing Happy base & Cython (Pre-requisite for Happy Base):
pip install Cython
pip install happybase

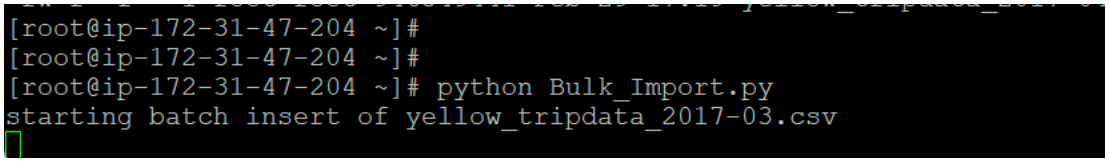


```
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]# pip install Cython
WARNING: Running pip install with root privileges is generally not a good idea. Try `pip3 install --user` instead.
Collecting Cython
  Downloading https://files.pythonhosted.org/packages/e3/7f/f584f5d15323feb897d42ef0e9d910649e2150d7a30cf7e7a8cc1d23
6e6f/Cython-3.0.8-py2.py3-none-any.whl (1.2MB)
    100% |████████████████████████████████| 1.2MB 834kB/s
Installing collected packages: Cython
Successfully installed Cython-3.0.8
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]# pip install happybase
WARNING: Running pip install with root privileges is generally not a good idea. Try `pip3 install --user` instead.
Collecting happybase
  Using cached https://files.pythonhosted.org/packages/d1/9c/f5f7bdb5439cda2b7da4e20ac24ec0e2455fd68aade8397f211d299
4c39d/happybase-1.2.0.tar.gz
Requirement already satisfied: six in /usr/local/lib/python3.7/site-packages (from happybase)
Collecting thriftpy2>=0.4 (from happybase)
  Using cached https://files.pythonhosted.org/packages/44/c3/20664039450f04a5630b68daaa00d539c9cd5338a17d5a28c3a553c
10de2/thriftpy2-0.4.20.tar.gz
Collecting ply<4.0,>=3.4 (from thriftpy2>=0.4->happybase)
  Downloading https://files.pythonhosted.org/packages/a3/58/35da89ee790598a0700ea49b2a66594140f44dec458c07e8e3d49791
37fc/ply-3.11-py2.py3-none-any.whl (49kB)
    100% |████████████████████████████████| 51kB 7.0MB/s
Installing collected packages: ply, thriftpy2, happybase
  Running setup.py install for thriftpy2 ... done
  Running setup.py install for happybase ... done
Successfully installed happybase-1.2.0 ply-3.11 thriftpy2-0.4.20
[root@ip-172-31-47-204 ~]#
```

Developed the Python Script, Bulk_Import.py (later renamed to Batch_Ingest.py to comply with the file naming convention)

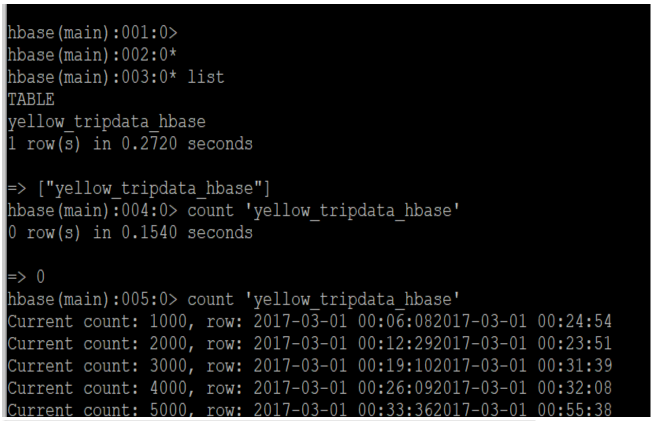
Executed the MR Script for bulk import of 2 files in the dataset to HBase table using following command:

python Bulk_Import.py

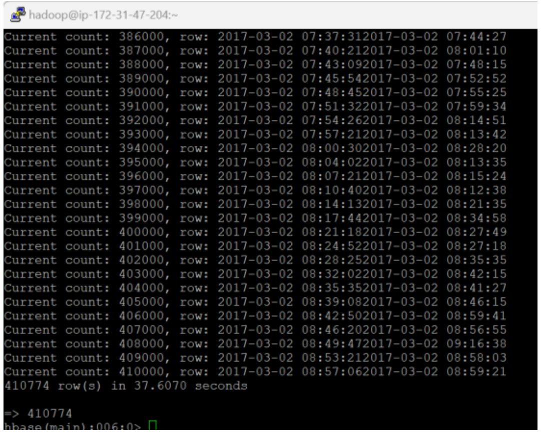


```
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]#
[root@ip-172-31-47-204 ~]# python Bulk_Import.py
starting batch insert of yellow_tripdata_2017-03.csv
```

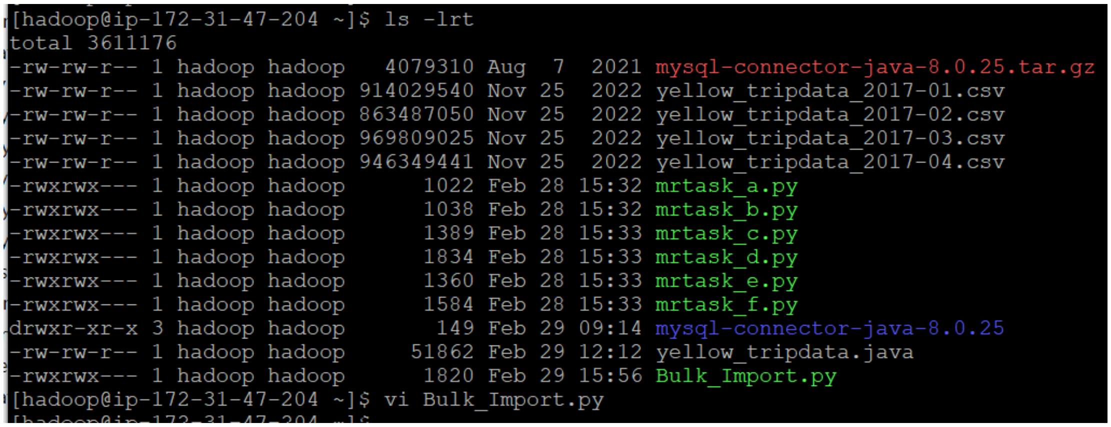
Checked the statistics of the HBase table after bulk import:



```
hbase(main):001:0>
hbase(main):002:0*
hbase(main):003:0* list
TABLE
yellow_tripdata_hbase
1 row(s) in 0.2720 seconds

=> ["yellow_tripdata_hbase"]
hbase(main):004:0> count 'yellow_tripdata_hbase'
0 row(s) in 0.1540 seconds

=> 0
hbase(main):005:0> count 'yellow_tripdata_hbase'
Current count: 1000, row: 2017-03-01 00:06:082017-03-01 00:24:54
Current count: 2000, row: 2017-03-01 00:12:292017-03-01 00:23:51
Current count: 3000, row: 2017-03-01 00:19:102017-03-01 00:31:39
Current count: 4000, row: 2017-03-01 00:26:092017-03-01 00:32:08
Current count: 5000, row: 2017-03-01 00:33:362017-03-01 00:55:38
```



```
Current count: 386000, row: 2017-03-02 07:37:312017-03-02 07:44:27
Current count: 387000, row: 2017-03-02 07:40:212017-03-02 08:01:10
Current count: 388000, row: 2017-03-02 07:43:092017-03-02 07:48:15
Current count: 389000, row: 2017-03-02 07:45:542017-03-02 07:52:52
Current count: 390000, row: 2017-03-02 07:48:452017-03-02 07:55:25
Current count: 391000, row: 2017-03-02 07:51:322017-03-02 07:59:34
Current count: 392000, row: 2017-03-02 07:54:262017-03-02 08:14:51
Current count: 393000, row: 2017-03-02 07:57:212017-03-02 08:13:42
Current count: 394000, row: 2017-03-02 08:00:302017-03-02 08:28:20
Current count: 395000, row: 2017-03-02 08:04:022017-03-02 08:13:35
Current count: 396000, row: 2017-03-02 08:07:212017-03-02 08:15:24
Current count: 397000, row: 2017-03-02 08:10:402017-03-02 08:12:38
Current count: 398000, row: 2017-03-02 08:14:132017-03-02 08:21:35
Current count: 399000, row: 2017-03-02 08:17:442017-03-02 08:34:58
Current count: 400000, row: 2017-03-02 08:21:182017-03-02 08:27:49
Current count: 401000, row: 2017-03-02 08:24:522017-03-02 08:27:18
Current count: 402000, row: 2017-03-02 08:28:252017-03-02 08:35:35
Current count: 403000, row: 2017-03-02 08:32:022017-03-02 08:42:15
Current count: 404000, row: 2017-03-02 08:35:352017-03-02 08:41:27
Current count: 405000, row: 2017-03-02 08:39:082017-03-02 08:46:15
Current count: 406000, row: 2017-03-02 08:42:502017-03-02 08:59:41
Current count: 407000, row: 2017-03-02 08:46:202017-03-02 08:56:55
Current count: 408000, row: 2017-03-02 08:49:472017-03-02 09:16:38
Current count: 409000, row: 2017-03-02 08:53:212017-03-02 08:58:03
Current count: 410000, row: 2017-03-02 08:57:062017-03-02 08:59:21
410774 row(s) in 37.6070 seconds

=> 410774
hbase(main):006:0>
```



```
[hadoop@ip-172-31-47-204 ~]$ ls -lrt
total 3611176
-rw-rw-r-- 1 hadoop hadoop   4079310 Aug  7  2021 mysql-connector-java-8.0.25.tar.gz
-rw-rw-r-- 1 hadoop hadoop 914029540 Nov 25  2022 yellow_tripdata_2017-01.csv
-rw-rw-r-- 1 hadoop hadoop 863487050 Nov 25  2022 yellow_tripdata_2017-02.csv
-rw-rw-r-- 1 hadoop hadoop 969809025 Nov 25  2022 yellow_tripdata_2017-03.csv
-rw-rw-r-- 1 hadoop hadoop 946349441 Nov 25  2022 yellow_tripdata_2017-04.csv
-rwxrwx--- 1 hadoop hadoop      1022 Feb 28 15:32 mrtask_a.py
-rwxrwx--- 1 hadoop hadoop      1038 Feb 28 15:32 mrtask_b.py
-rwxrwx--- 1 hadoop hadoop      1389 Feb 28 15:33 mrtask_c.py
-rwxrwx--- 1 hadoop hadoop      1834 Feb 28 15:33 mrtask_d.py
-rwxrwx--- 1 hadoop hadoop      1360 Feb 28 15:33 mrtask_e.py
-rwxrwx--- 1 hadoop hadoop      1584 Feb 28 15:33 mrtask_f.py
drwxr-xr-x 3 hadoop hadoop       149 Feb 29 09:14 mysql-connector-java-8.0.25
-rw-rw-r-- 1 hadoop hadoop     51862 Feb 29 12:12 yellow_tripdata.java
-rwxrwx--- 1 hadoop hadoop      1820 Feb 29 15:56 Bulk_Import.py
[hadoop@ip-172-31-47-204 ~]$ vi Bulk_Import.py
[hadoop@ip-172-31-47-204 ~]$
```

--X--