Task 2:
Loading data from an RDS table into an HBase table.

Step 1: Access the EMR instance and retrieve the MySQL connector while logged in as the root user:
Use command:

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

```
root@ip-172-31-76-129:~
[root@ip-172-31-76-129 ~]# wget https://de-mysql-connector.s3.amazonaws.com/mysql-connecto
r-java-8.0.25.tar.gz
--2024-02-05 12:54:03--  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.
216.38.73, 52.216.139.99, 52.216.210.233, ...
Connecting to de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)|52
.216.38.73|: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 0.04s

2024-02-05 12:54:04 (99.4 MB/s) - 'mysql-connector-java-8.0.25.tar.gz' saved [4079310/4079
310]

[root@ip-172-31-76-129 ~]# 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/
```

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

```
root@ip-172-31-76-129:~/mysql-connector-java-8.0.25
mysql-connector-java-8.0.25/src/test/java/testsuite/x/devapi/TableInsertTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/devapi/TableSelectTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/devapi/TableTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/devapi/TableUpdateTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/devapi/TransactionTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/devapi/package-info.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/InternalXBaseTestCase.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/MysqlxSessionTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/XProtocolAsyncTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/XProtocolAuthTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/XProtocolTest.java
mysql-connector-java-8.0.25/src/test/java/testsuite/x/internal/package-info.java
[root@ip-172-31-76-129 ~]# cd mysql-connector-java-8.0.25/
[root@ip-172-31-76-129 mysql-connector-java-8.0.25]# sudo cp mysql-connector-java-8.0.25.j
ar /usr/lib/sqoop/lib/
[root@ip-172-31-76-129 mysql-connector-java-8.0.25]#
```

Step 2: Verifying connectivity between the RDS and EMR cluster via JDBC, and listing all databases within the RDS. Use command:
sqoop-list-databases --connect jdbc:mysql://mapreduce-rds.cvg0g4e4op16.us-east-1.rds.amazonaws.com:3306/ --username admin --password admin123

```
hadoop@ip-172-31-66-76:~
  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

[hadoop@ip-172-31-66-76 ~]$ sqoop-list-databases --connect jdbc:mysql://mapreduc
e-rds.cvg0g4e4op16.us-east-1.rds.amazonaws.com:3306/ --username admin --password
 admin123
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
24/02/05 16:29:07 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/
org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1
.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.jar!/
org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
24/02/05 16:29:07 WARN tool.BaseSqoopTool: Setting your password on the command-
line is insecure. Consider using -P instead.
24/02/05 16:29:07 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
mysql
information_schema
performance_schema
sys
nyc_taxi
[hadoop@ip-172-31-66-76 ~]$
```

Step 3: Ingesting data from RDS to HBase table:
Use Command:

```
sqoop import \
--connect "jdbc:mysql://mapreduce-rds.cvg0g4e4op16.us-east-1.rds.amazonaws.com:3306/nyc_taxi" \
--username admin --password admin123 \
--table NYC_TRIPS \
--columns
"VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,Passenger_count,Trip_distance,RateCodeID,Store_and_f
wd_flag,PULocationID,DOLocationID,Payment_type,Fare_amount,Extra,MTA_tax,Tip_amount,Tolls_amount,Im
provement_surcharge,Total_amount" \
--hbase-create-table \
--hbase-table NYC_TRIPS_hbase \
--column-family Trip_details \
--hbase-row-key VendorID,tpep_pickup_datetime,tpep_dropoff_datetime \
-m 1
```

This command imports data from the MySQL table **'NYC_TRIPS'** into an HBase table named **'NYC_TRIPS_hbase'**. The imported data will be stored in the column family named **'Trip_details'** within HBase. The row key in HBase is composed of three columns from the MySQL table: **'VendorID', 'tpep_pickup_datetime'** and **'tpep_dropoff_datetime'**. The data is loaded into HBase using the bulk load feature for faster loading.

hadoop@ip-172-31-66-76:~

```
[hadoop@ip-172-31-66-76 ~]$ sqoop import \
> --connect "jdbc:mysql://mapreduce-rds.cvg0g4e4op16.us-east-1.rds.amazonaws.com
:3306/nyc_taxi" \
> --username admin --password admin123 \
> --table NYC_TRIPS \
> --columns "VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,Passenger_count
,Trip_distance,RateCodeID,Store_and_fwd_flag,PULocationID,DOLocationID,Payment_t
ype,Fare_amount,Extra,MTA_tax,Tip_amount,Tolls_amount,Improvement_surcharge,Tota
l_amount" \
> --hbase-create-table \
> --hbase-table NYC_TRIPS_hbase \
> --column-family Trip_details \
> --hbase-row-key VendorID,tpep_pickup_datetime,tpep_dropoff_datetime \
> -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
24/02/05 16:34:21 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/StaticLoggerBinder.class]
```

hadoop@ip-172-31-66-76:~

```
24/02/05 16:34:26 INFO Configuration.deprecation: mapred.map.tasks is deprecated
. Instead, use mapreduce.job.maps
24/02/05 16:34:28 INFO mapreduce.HBaseImportJob: Creating missing HBase table NY
C_TRIPS_hbase
24/02/05 16:34:30 WARN mapreduce.TableMapReduceUtil: The addDependencyJars(Confi
guration, Class<?>...) method has been deprecated since it is easy to use incorr
ectly. Most users should rely on addDependencyJars(Job) instead. See HBASE-8386
for more details.
24/02/05 16:34:30 INFO client.RMProxy: Connecting to ResourceManager at ip-172-3
1-66-76.ec2.internal/172.31.66.76:8032
24/02/05 16:34:34 INFO db.DBInputFormat: Using read commited transaction isolati
on
24/02/05 16:34:34 INFO mapreduce.JobSubmitter: number of splits:1
24/02/05 16:34:34 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_17
07150143144_0001
24/02/05 16:34:35 INFO impl.YarnClientImpl: Submitted application application_17
07150143144_0001
24/02/05 16:34:35 INFO mapreduce.Job: The url to track the job: http://ip-172-31
-66-76.ec2.internal:20888/proxy/application_1707150143144_0001/
24/02/05 16:34:35 INFO mapreduce.Job: Running job: job_1707150143144_0001
24/02/05 16:34:46 INFO mapreduce.Job: Job job_1707150143144_0001 running in uber
 mode : false
24/02/05 16:34:46 INFO mapreduce.Job:  map 0% reduce 0%
```

hadoop@ip-172-31-66-76:~

```
24/02/05 16:34:46 INFO mapreduce.Job:  map 0% reduce 0%
24/02/05 17:25:41 INFO mapreduce.Job:  map 100% reduce 0%
24/02/05 17:25:42 INFO mapreduce.Job: Job job_1707150143144_0001 completed succe
ssfully
24/02/05 17:25:42 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=225835
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=1
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=146531424
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=3052738
                Total vcore-milliseconds taken by all map tasks=3052738
                Total megabyte-milliseconds taken by all map tasks=4689005568
        Map-Reduce Framework
                Map input records=18880595
                Map output records=18880595
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=17050
                CPU time spent (ms)=1148070
                Physical memory (bytes) snapshot=823984128
                Virtual memory (bytes) snapshot=3323379712
                Total committed heap usage (bytes)=705167360
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
24/02/05 17:25:42 INFO mapreduce.ImportJobBase: Transferred 0 bytes in 3,072.177
2 seconds (0 bytes/sec)
24/02/05 17:25:42 INFO mapreduce.ImportJobBase: Retrieved 18880595 records.
[hadoop@ip-172-31-66-76 ~]$
```

Task 3:
Bulk import of data from 2 csv files (yellow_tripdata_2017-03.csv & yellow_tripdata_2017-04.csv) in the dataset on your EMR Cluster to your HBase Table using relevant codes.

Upload the **batch_ingest.py** file to EMR. Use command: python batch_ingest.py