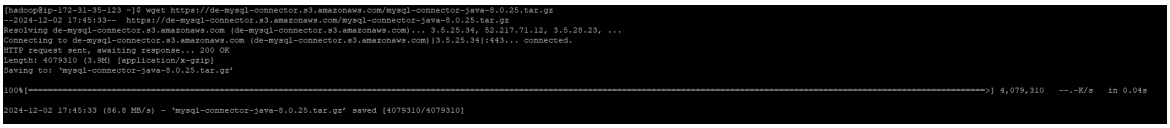
Task 2: Using Sqoop commands to ingest the data from RDS into the HBase Table.

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



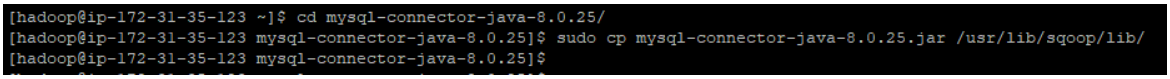
-> Now, we run the following step to extract the MySQL connector tar file

-> 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://database-1.c1e04g20st3d.us-east-1.rds.amazonaws.com:3306/yellow
_taxi   --username admin   --password 1234HelloHiWork   --table trip_records
--target-dir /user/hadoop/nyc_yello_taxi   --hbase-table trip_log_hbase
--column-family cf1   --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 : species 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 : species 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 : species the name of the column family in HBase where the imported data will be stored

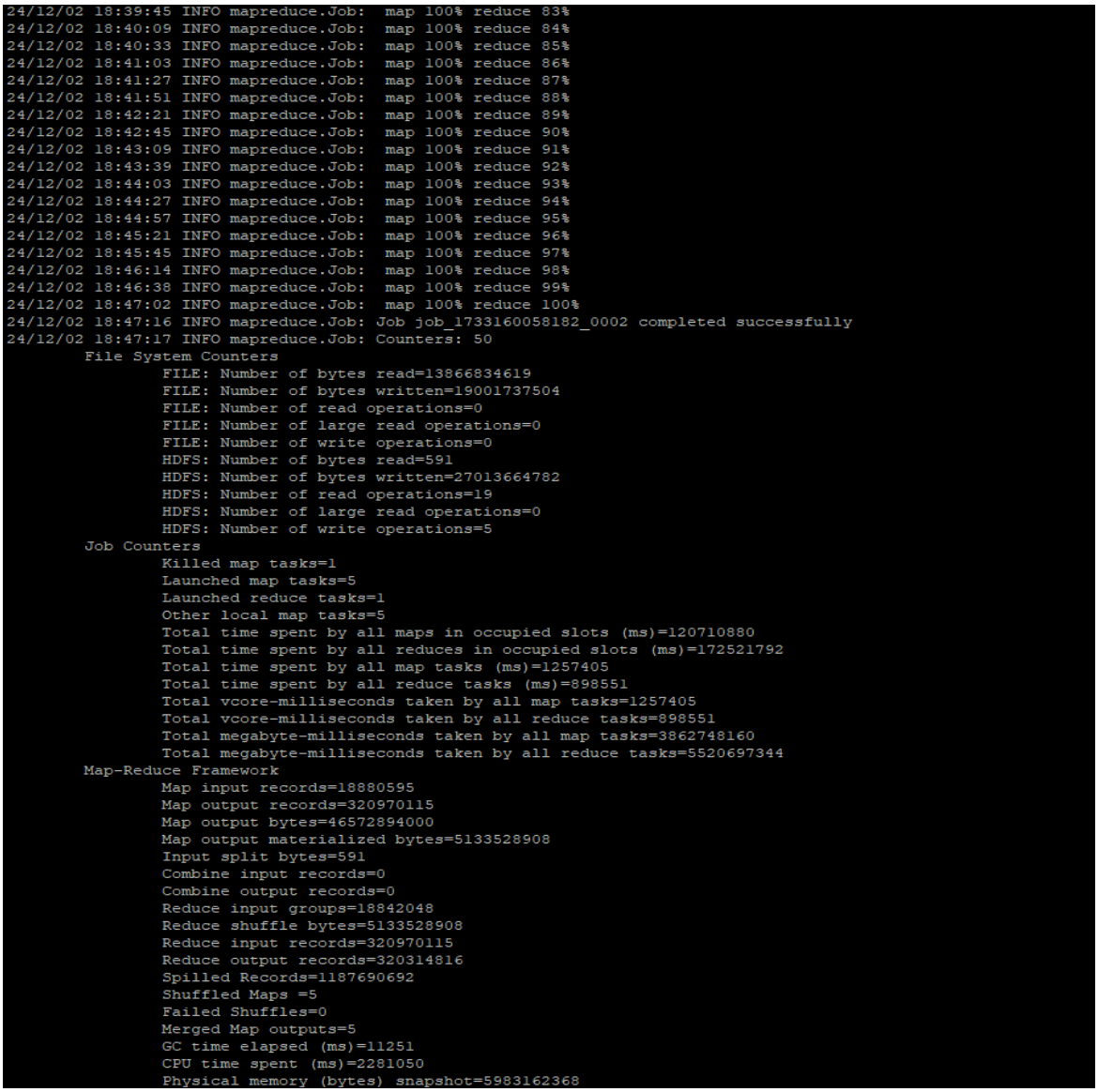
--hbase-table : species the name of the HBase table to import data into

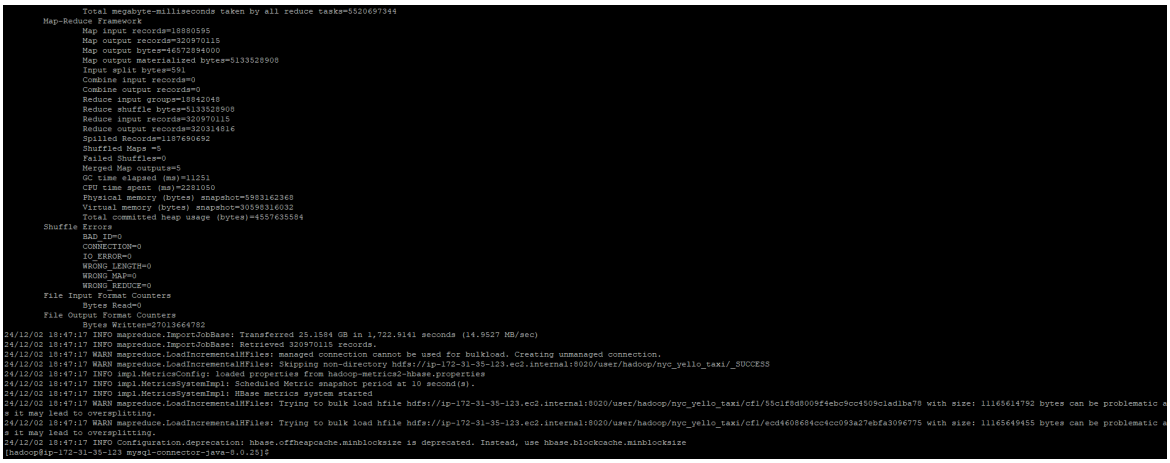
--table : species the name of the MySQL table to import data from

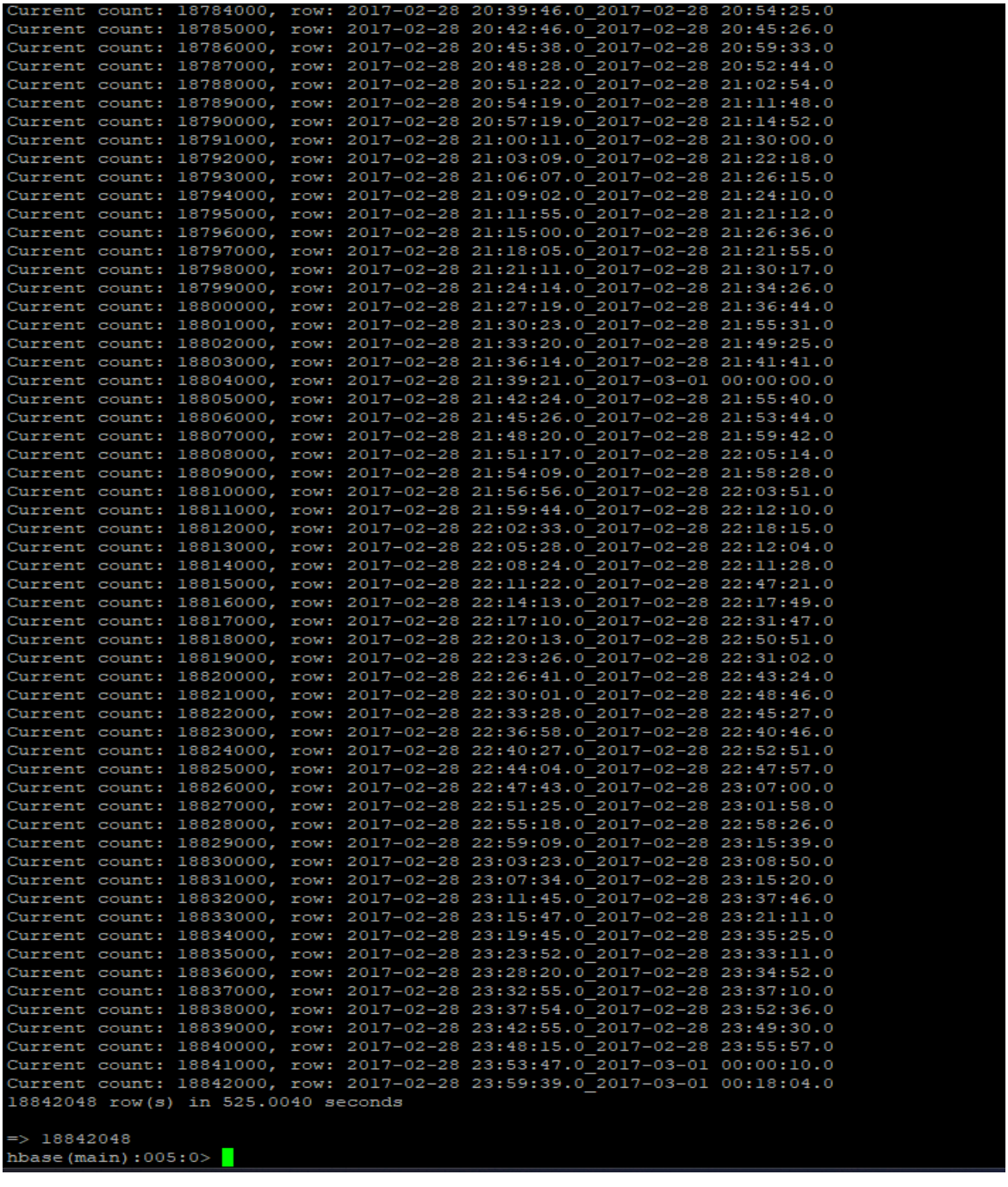
--password : species the password to use when connecting to the MySQL database

--username : species the username to use when connecting to the MySQL database

–target-dir : specifies the name of the HDFS directory where Sqoop will store the imported data from the MySQL table




Total megabyte-milliseconds taken by all reduce tasks=5520697344
Map-Reduce Framework
Map input records=18880595
Map output records=320970115
Map output bytes=46572894000
Map output materialized bytes=5133528908
Input split bytes=591
Combine input records=0
Combine output records=0
Reduce input groups=18842048
Reduce shuffle bytes=5133528908
Reduce input records=320970115
Reduce output records=320314816
Spilled Records=1187690692
Shuffled Maps =5
Failed Shuffles=0
Merged Map outputs=5
GC time elapsed (ms)=11251
CPU time spent (ms)=2281050
Physical memory (bytes) snapshot=5983162368
Virtual memory (bytes) snapshot=30598316032
Total committed heap usage (bytes)=4557635584
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=27013664782
24/12/02 18:47:17 INFO mapreduce.ImportJobBase: Transferred 25.1584 GB in 1,722.9141 seconds (14.9527 MB/sec)
24/12/02 18:47:17 INFO mapreduce.ImportJobBase: Retrieved 320970115 records.
24/12/02 18:47:17 WARN mapreduce.LoadIncrementalHFiles: managed connection cannot be used for bulkload. Creating unmanaged connection.
24/12/02 18:47:17 WARN mapreduce.LoadIncrementalHFiles: Skipping non-directory hdfs://ip-172-31-35-123.ec2.internal:8020/user/hadoop/nyc_yello_taxi/_SUCCESS
24/12/02 18:47:17 INFO impl.MetricsConfig: loaded properties from hadoop-metrics2-hbase.properties
24/12/02 18:47:17 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
24/12/02 18:47:17 INFO impl.MetricsSystemImpl: HBase metrics system started
24/12/02 18:47:17 WARN mapreduce.LoadIncrementalHFiles: Trying to bulk load hfile hdfs://ip-172-31-35-123.ec2.internal:8020/user/hadoop/nyc_yello_taxi/cf1/55c1f0d8009f4ebc9cc4509c1ad1ba78 with size: 11165614792 bytes can be problematic a
s it may lead to oversplitting.
24/12/02 18:47:17 WARN mapreduce.LoadIncrementalHFiles: Trying to bulk load hfile hdfs://ip-172-31-35-123.ec2.internal:8020/user/hadoop/nyc_yello_taxi/cf1/ecd4608684cc4cc093a27ebfa3096775 with size: 11165649455 bytes can be problematic a
s it may lead to oversplitting.
24/12/02 18:47:17 INFO Configuration.deprecation: hbase.offheapcache.minblocksize is deprecated. Instead, use hbase.blockcache.minblocksize
[hadoop@ip-172-31-35-123 mysql-connector-java-8.0.25]$


Current count: 18784000, row: 2017-02-28 20:39:46.0_2017-02-28 20:54:25.0
Current count: 18785000, row: 2017-02-28 20:42:46.0_2017-02-28 20:45:26.0
Current count: 18786000, row: 2017-02-28 20:45:38.0_2017-02-28 20:59:33.0
Current count: 18787000, row: 2017-02-28 20:48:28.0_2017-02-28 20:52:44.0
Current count: 18788000, row: 2017-02-28 20:51:22.0_2017-02-28 21:02:54.0
Current count: 18789000, row: 2017-02-28 20:54:19.0_2017-02-28 21:11:48.0
Current count: 18790000, row: 2017-02-28 20:57:19.0_2017-02-28 21:14:52.0
Current count: 18791000, row: 2017-02-28 21:00:11.0_2017-02-28 21:30:00.0
Current count: 18792000, row: 2017-02-28 21:03:09.0_2017-02-28 21:22:18.0
Current count: 18793000, row: 2017-02-28 21:06:07.0_2017-02-28 21:26:15.0
Current count: 18794000, row: 2017-02-28 21:09:02.0_2017-02-28 21:24:10.0
Current count: 18795000, row: 2017-02-28 21:11:55.0_2017-02-28 21:21:12.0
Current count: 18796000, row: 2017-02-28 21:15:00.0_2017-02-28 21:26:36.0
Current count: 18797000, row: 2017-02-28 21:18:05.0_2017-02-28 21:21:55.0
Current count: 18798000, row: 2017-02-28 21:21:11.0_2017-02-28 21:30:17.0
Current count: 18799000, row: 2017-02-28 21:24:14.0_2017-02-28 21:34:26.0
Current count: 18800000, row: 2017-02-28 21:27:19.0_2017-02-28 21:36:44.0
Current count: 18801000, row: 2017-02-28 21:30:23.0_2017-02-28 21:55:31.0
Current count: 18802000, row: 2017-02-28 21:33:20.0_2017-02-28 21:49:25.0
Current count: 18803000, row: 2017-02-28 21:36:14.0_2017-02-28 21:41:41.0
Current count: 18804000, row: 2017-02-28 21:39:21.0_2017-03-01 00:00:00.0
Current count: 18805000, row: 2017-02-28 21:42:24.0_2017-02-28 21:55:40.0
Current count: 18806000, row: 2017-02-28 21:45:26.0_2017-02-28 21:53:44.0
Current count: 18807000, row: 2017-02-28 21:48:20.0_2017-02-28 21:59:42.0
Current count: 18808000, row: 2017-02-28 21:51:17.0_2017-02-28 22:05:14.0
Current count: 18809000, row: 2017-02-28 21:54:09.0_2017-02-28 21:58:28.0
Current count: 18810000, row: 2017-02-28 21:56:56.0_2017-02-28 22:03:51.0
Current count: 18811000, row: 2017-02-28 21:59:44.0_2017-02-28 22:12:10.0
Current count: 18812000, row: 2017-02-28 22:02:33.0_2017-02-28 22:18:15.0
Current count: 18813000, row: 2017-02-28 22:05:28.0_2017-02-28 22:12:04.0
Current count: 18814000, row: 2017-02-28 22:08:24.0_2017-02-28 22:11:28.0
Current count: 18815000, row: 2017-02-28 22:11:22.0_2017-02-28 22:47:21.0
Current count: 18816000, row: 2017-02-28 22:14:13.0_2017-02-28 22:17:49.0
Current count: 18817000, row: 2017-02-28 22:17:10.0_2017-02-28 22:31:47.0
Current count: 18818000, row: 2017-02-28 22:20:13.0_2017-02-28 22:50:51.0
Current count: 18819000, row: 2017-02-28 22:23:26.0_2017-02-28 22:31:02.0
Current count: 18820000, row: 2017-02-28 22:26:41.0_2017-02-28 22:43:24.0
Current count: 18821000, row: 2017-02-28 22:30:01.0_2017-02-28 22:48:46.0
Current count: 18822000, row: 2017-02-28 22:33:28.0_2017-02-28 22:45:27.0
Current count: 18823000, row: 2017-02-28 22:36:58.0_2017-02-28 22:40:46.0
Current count: 18824000, row: 2017-02-28 22:40:27.0_2017-02-28 22:52:51.0
Current count: 18825000, row: 2017-02-28 22:44:04.0_2017-02-28 22:47:57.0
Current count: 18826000, row: 2017-02-28 22:47:43.0_2017-02-28 23:07:00.0
Current count: 18827000, row: 2017-02-28 22:51:25.0_2017-02-28 23:01:58.0
Current count: 18828000, row: 2017-02-28 22:55:18.0_2017-02-28 22:58:26.0
Current count: 18829000, row: 2017-02-28 22:59:09.0_2017-02-28 23:15:39.0
Current count: 18830000, row: 2017-02-28 23:03:23.0_2017-02-28 23:08:50.0
Current count: 18831000, row: 2017-02-28 23:07:34.0_2017-02-28 23:15:20.0
Current count: 18832000, row: 2017-02-28 23:11:45.0_2017-02-28 23:37:46.0
Current count: 18833000, row: 2017-02-28 23:15:47.0_2017-02-28 23:21:11.0
Current count: 18834000, row: 2017-02-28 23:19:45.0_2017-02-28 23:35:25.0
Current count: 18835000, row: 2017-02-28 23:23:52.0_2017-02-28 23:33:11.0
Current count: 18836000, row: 2017-02-28 23:28:20.0_2017-02-28 23:34:52.0
Current count: 18837000, row: 2017-02-28 23:32:55.0_2017-02-28 23:37:10.0
Current count: 18838000, row: 2017-02-28 23:37:54.0_2017-02-28 23:52:36.0
Current count: 18839000, row: 2017-02-28 23:42:55.0_2017-02-28 23:49:30.0
Current count: 18840000, row: 2017-02-28 23:48:15.0_2017-02-28 23:55:57.0
Current count: 18841000, row: 2017-02-28 23:53:47.0_2017-03-01 00:00:10.0
Current count: 18842000, row: 2017-02-28 23:59:39.0_2017-03-01 00:18:04.0
18842048 row(s) in 525.0040 seconds
=> 18842048
hbase(main):005:0>