

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>

```
[hadoop@ip-172-31-35-123 ~]$ wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
--2024-12-02 17:45:33-- 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)... 3.5.25.34, 52.217.71.12, 3.5.28.23, ...
Connecting to de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com) (3.5.25.34):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'

mysql-connector-java-8.0.25.tar.gz 4079310 100% 4.079MB/s 0s
2024-12-02 17:45:33 (86.6 MB/s) - 'mysql-connector-java-8.0.25.tar.gz' saved [4079310/4079310]
```

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

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

```
[hadoop@ip-172-31-35-123 ~]$ 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/obian.in/
```

-> Now, we 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/

```
[hadoop@ip-172-31-35-123 ~]$ cd mysql-connector-java-8.0.25/
[hadoop@ip-172-31-35-123 mysql-connector-java-8.0.25]$ sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/
[hadoop@ip-172-31-35-123 mysql-connector-java-8.0.25]$
```

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 cfl --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

```
24/12/02 18:39:45 INFO mapreduce.Job: map 100% reduce 83%
24/12/02 18:40:09 INFO mapreduce.Job: map 100% reduce 84%
24/12/02 18:40:33 INFO mapreduce.Job: map 100% reduce 85%
24/12/02 18:41:03 INFO mapreduce.Job: map 100% reduce 86%
24/12/02 18:41:27 INFO mapreduce.Job: map 100% reduce 87%
24/12/02 18:41:51 INFO mapreduce.Job: map 100% reduce 88%
24/12/02 18:42:21 INFO mapreduce.Job: map 100% reduce 89%
24/12/02 18:42:45 INFO mapreduce.Job: map 100% reduce 90%
24/12/02 18:43:09 INFO mapreduce.Job: map 100% reduce 91%
24/12/02 18:43:39 INFO mapreduce.Job: map 100% reduce 92%
24/12/02 18:44:03 INFO mapreduce.Job: map 100% reduce 93%
24/12/02 18:44:27 INFO mapreduce.Job: map 100% reduce 94%
24/12/02 18:44:57 INFO mapreduce.Job: map 100% reduce 95%
24/12/02 18:45:21 INFO mapreduce.Job: map 100% reduce 96%
24/12/02 18:45:45 INFO mapreduce.Job: map 100% reduce 97%
24/12/02 18:46:14 INFO mapreduce.Job: map 100% reduce 98%
24/12/02 18:46:38 INFO mapreduce.Job: map 100% reduce 99%
24/12/02 18:47:02 INFO mapreduce.Job: map 100% reduce 100%
24/12/02 18:47:16 INFO mapreduce.Job: Job job_1733160058182_0002 completed successfully
24/12/02 18:47:17 INFO mapreduce.Job: Counters: 50
  File System Counters
    FILE: Number of bytes read=13866834619
    FILE: Number of bytes written=19001737504
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=591
    HDFS: Number of bytes written=27013664782
    HDFS: Number of read operations=19
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=5
  Job Counters
    Killed map tasks=1
    Launched map tasks=5
    Launched reduce tasks=1
    Other local map tasks=5
    Total time spent by all maps in occupied slots (ms)=120710880
    Total time spent by all reduces in occupied slots (ms)=172521792
    Total time spent by all map tasks (ms)=1257405
    Total time spent by all reduce tasks (ms)=898551
    Total vcore-milliseconds taken by all map tasks=1257405
    Total vcore-milliseconds taken by all reduce tasks=898551
    Total megabyte-milliseconds taken by all map tasks=3862748160
    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
```

```
Total mapreduce-millisecs taken by all reduce tasks=5520497344
Map-Reduce Framework
  Map input records=18880595
  Map output records=22879313
  Map output bytes=4657284000
  Map output materialized bytes=513328508
  Input splits bytes=551
  Combine input records=0
  Combine output records=0
  Reduce input groups=18892048
  Reduce shuffle bytes=513328508
  Reduce input records=3207015
  Reduce output records=320314816
  Spillable Records=118780692
  Shuffled Maps=45
  Failed Shuffles=0
  Merged Map outputs=5
  GC time elapsed (ms)=11281
  CPU time spent (ms)=2261080
  Physical memory (bytes) snapshot=5983162368
  Virtual memory (bytes) snapshot=30588316032
  Total committed heap usage (bytes)=4517625584
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 25.1884 GB in 1,722.9141 seconds (14.9527 MB/sec)
24/12/02 18:47:17 INFO mapreduce.ImportJobBase: Retrieved 22879313 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 INFO mapreduce.LoadIncrementalHFiles: Skipping non-directory hdfs://ip-172-31-35-123.ec2.internal:8020/user/hadoop/nyv_yello_taxi/_SUCCESS
24/12/02 18:47:17 INFO impl.MetricConfigImpl: Loaded properties from hadoop-metric2-hbase.properties
24/12/02 18:47:17 INFO impl.MetricConfigImpl: Scheduled Metric snapshot period at 10 second(s).
24/12/02 18:47:17 INFO impl.MetricConfigImpl: HBase metric 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/nyv_yello_taxi/cfi/35c1f8d800f4eb9cc4509c0ad1d8a78 with size: 11165614792 bytes can be problematic as 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/nyv_yello_taxi/cfi/ec4d46086840cc0c093a27ebfa3096775 with size: 11165649455 bytes can be problematic as 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 mygl-cometecor-java-8-0.25]6
```

```
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> █