1. Login to EMR Instance, switch to root user and download mySQL connector :

   //switch to root user
   sudo -i

   // download mySQL connector
   wget hΣps://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz

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

   cd mysql-connector-java-8.0.25/

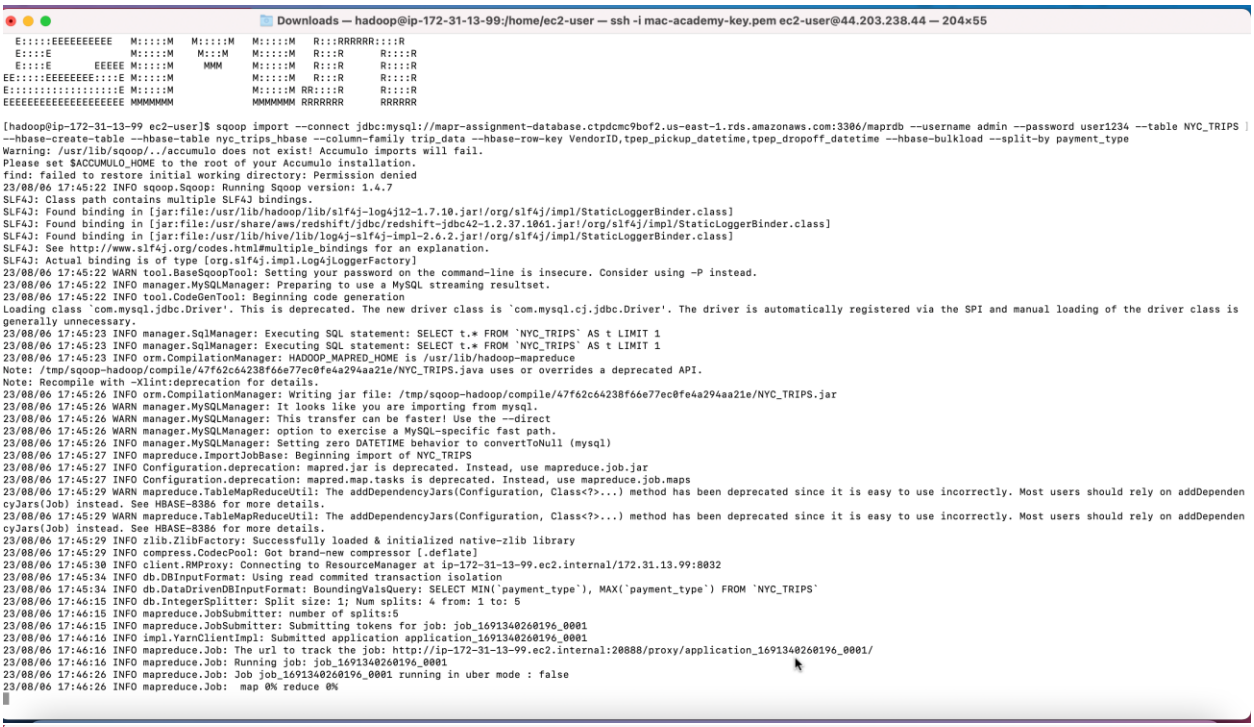
   sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/

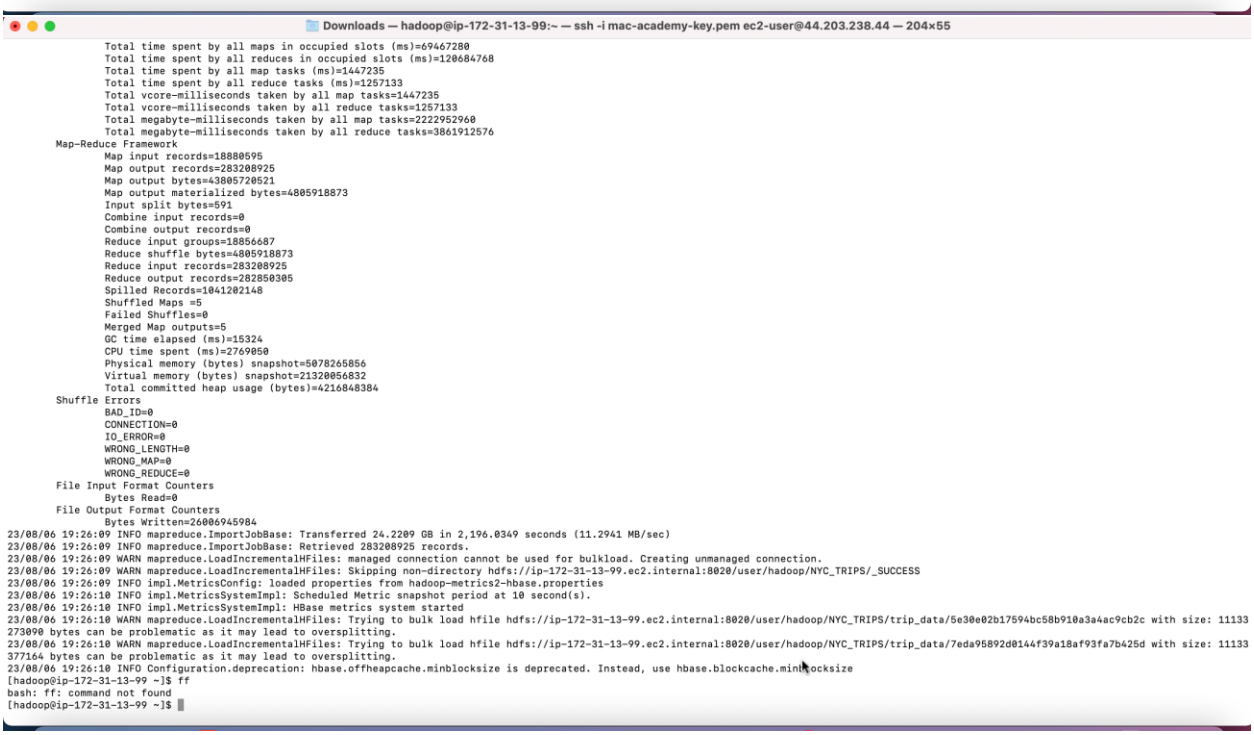
2. Use Sqoop command to Ingest data from MySQL table to Hbase Table:

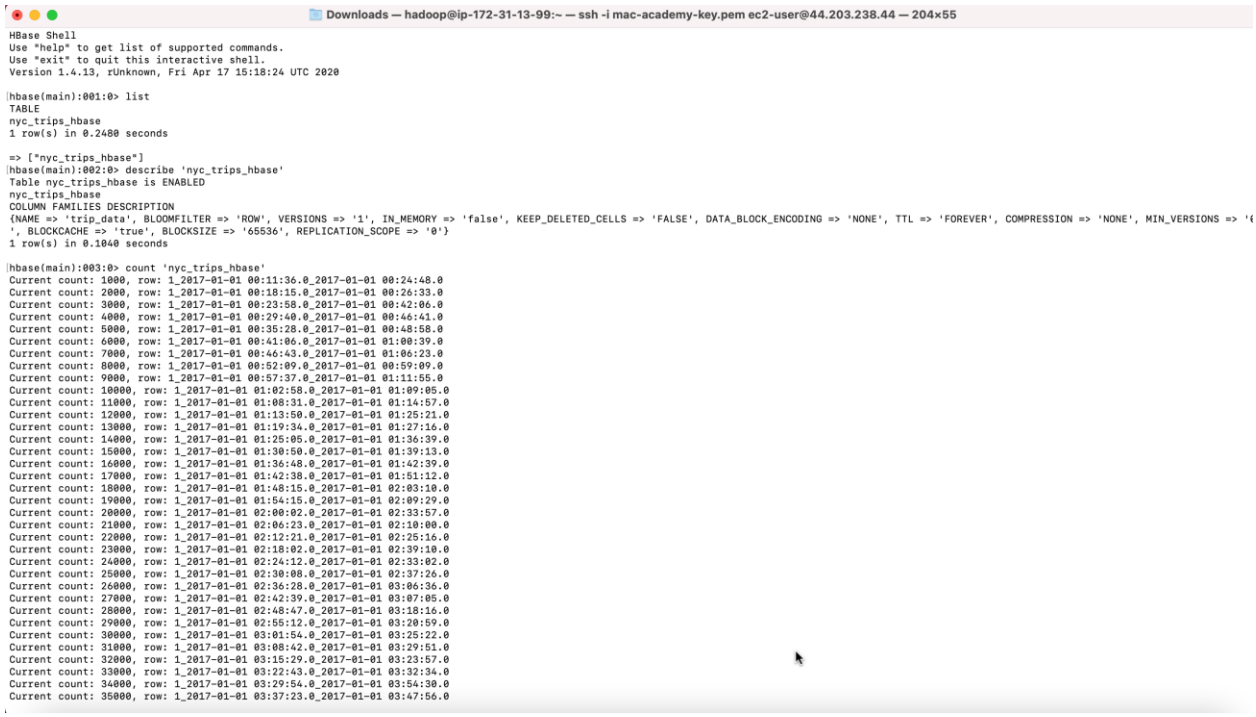
   sqoop import --connect jdbc:mysql://mapr-assignment-database.ctpdcmc9bof2.us-east-1.rds.amazonaws.com:3306/maprdb --username admin -P --table NYC_TRIPS --hbase-create-table --hbase-table nyc_trips_hbase --column-family trip_data --hbase-row-key VendorID,tpep_pickup_datetime,tpep_dropoff_datetime --hbase-bulkload --split-by payment_type

   explanation:
   Above sqoop command will ingest data from MySQL table NYC_TRIPS to a newly created Hbase Table nyc_trips_hbase, with column family trip_data and row key VendorID,tpep_pickup_datetime,tpep_dropoff_datetime

   `--connect`: to connect to given MySQL url
   `--username`: to provide username for connecting MySQL Database
   `-P`: used for password
   `--table`: MySQL table to ingest data from
   `--hbase-create-table`: to create a new HBase table if it does not exist.
   `--hbase-table`: Hbase Table to ingest data to
   `--column-family`: to provide column family name for Hbase table
   `--hbase-row-key`: specifies one or more columns from the MySQL table that will be used as the row key in HBase.
   `--hbase-bulkload`: uses HBase bulk load feature for faster data loading.
   `--split-by`: specifies a column from the MySQL table that will be used to split data into multiple HBase regions.


Downloads — hadoop@ip-172-31-13-99:/home/ec2-user — ssh -i mac-academy-key.pem ec2-user@44.203.238.44 — 204×55
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
EEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMM RRRRRRR RRRRRR
[hadoop@ip-172-31-13-99 ec2-user]$ sqoop import --connect jdbc:mysql://mapr-assignment-database.ctpdcmc9bof2.us-east-1.rds.amazonaws.com:3306/maprdb --username admin --password user1234 --table NYC_TRIPS
--hbase-create-table --hbase-table nyc_trips_hbase --column-family trip_data --hbase-row-key VendorID,tpep_pickup_datetime,tpep_dropoff_datetime --hbase-bulkload --split-by payment_type
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
find: failed to restore initial working directory: Permission denied
23/08/06 17:45:22 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]
23/08/06 17:45:22 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/08/06 17:45:22 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/08/06 17:45:22 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 driver is automatically registered via the SPI and manual loading of the driver class is
generally unnecessary.
23/08/06 17:45:23 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `NYC_TRIPS` AS t LIMIT 1
23/08/06 17:45:23 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `NYC_TRIPS` AS t LIMIT 1
23/08/06 17:45:23 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/47f62c64238f66e77ec0fe4a294aa21e/NYC_TRIPS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/08/06 17:45:26 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/47f62c64238f66e77ec0fe4a294aa21e/NYC_TRIPS.jar
23/08/06 17:45:26 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/08/06 17:45:26 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/08/06 17:45:26 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/08/06 17:45:26 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/08/06 17:45:27 INFO mapreduce.ImportJobBase: Beginning import of NYC_TRIPS
23/08/06 17:45:27 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/08/06 17:45:27 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/08/06 17:45:29 WARN mapreduce.TableMapReduceUtil: The addDependencyJars(Configuration, Class<?>...) method has been deprecated since it is easy to use incorrectly. Most users should rely on addDependen
cyJars(Job) instead. See HBASE-8386 for more details.
23/08/06 17:45:29 WARN mapreduce.TableMapReduceUtil: The addDependencyJars(Configuration, Class<?>...) method has been deprecated since it is easy to use incorrectly. Most users should rely on addDependen
cyJars(Job) instead. See HBASE-8386 for more details.
23/08/06 17:45:29 INFO zlib.ZlibFactory: Successfully loaded & initialized native-zlib library
23/08/06 17:45:29 INFO compress.CodecPool: Got brand-new compressor [.deflate]
23/08/06 17:45:30 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-13-99.ec2.internal/172.31.13.99:8032
23/08/06 17:45:34 INFO db.DBInputFormat: Using read commited transaction isolation
23/08/06 17:45:34 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`payment_type`), MAX(`payment_type`) FROM `NYC_TRIPS`
23/08/06 17:46:15 INFO db.IntegerSplitter: Split size: 1; Num splits: 4 from: 1 to: 5
23/08/06 17:46:15 INFO mapreduce.JobSubmitter: number of splits:5
23/08/06 17:46:15 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1691340260196_0001
23/08/06 17:46:16 INFO impl.YarnClientImpl: Submitted application application_1691340260196_0001
23/08/06 17:46:16 INFO mapreduce.Job: The url to track the job: http://ip-172-31-13-99.ec2.internal:20888/proxy/application_1691340260196_0001/
23/08/06 17:46:16 INFO mapreduce.Job: Running job: job_1691340260196_0001
23/08/06 17:46:26 INFO mapreduce.Job: Job job_1691340260196_0001 running in uber mode : false
23/08/06 17:46:26 INFO mapreduce.Job: map 0% reduce 0%


Downloads — hadoop@ip-172-31-13-99:~ — ssh -i mac-academy-key.pem ec2-user@44.203.238.44 — 204×55
Total time spent by all maps in occupied slots (ms)=69467280
Total time spent by all reduces in occupied slots (ms)=120684768
Total time spent by all map tasks (ms)=1447235
Total time spent by all reduce tasks (ms)=1257133
Total vcore-milliseconds taken by all map tasks=1447235
Total vcore-milliseconds taken by all reduce tasks=1257133
Total megabyte-milliseconds taken by all map tasks=2222952960
Total megabyte-milliseconds taken by all reduce tasks=3861912576
Map-Reduce Framework
Map input records=18880595
Map output records=283208925
Map output bytes=43805720521
Map output materialized bytes=4805918873
Input split bytes=591
Combine input records=0
Combine output records=0
Reduce input groups=18856687
Reduce shuffle bytes=4805918873
Reduce input records=283208925
Reduce output records=282850305
Spilled Records=1041202148
Shuffled Maps =5
Failed Shuffles=0
Merged Map outputs=5
GC time elapsed (ms)=15324
CPU time spent (ms)=2769050
Physical memory (bytes) snapshot=5078265856
Virtual memory (bytes) snapshot=21320056832
Total committed heap usage (bytes)=4216848384
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=26006945984
23/08/06 19:26:09 INFO mapreduce.ImportJobBase: Transferred 24.2289 GB in 2,196.0349 seconds (11.2941 MB/sec)
23/08/06 19:26:09 INFO mapreduce.ImportJobBase: Retrieved 283208925 records.
23/08/06 19:26:09 WARN mapreduce.LoadIncrementalHFiles: managed connection cannot be used for bulkload. Creating unmanaged connection.
23/08/06 19:26:09 WARN mapreduce.LoadIncrementalHFiles: Skipping non-directory hdfs://ip-172-31-13-99.ec2.internal:8020/user/hadoop/NYC_TRIPS/_SUCCESS
23/08/06 19:26:09 INFO impl.MetricsConfig: loaded properties from hadoop-metrics2-hbase.properties
23/08/06 19:26:10 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
23/08/06 19:26:10 INFO impl.MetricsSystemImpl: HBase metrics system started
23/08/06 19:26:10 WARN mapreduce.LoadIncrementalHFiles: Trying to bulk load hfile hdfs://ip-172-31-13-99.ec2.internal:8020/user/hadoop/NYC_TRIPS/trip_data/5e30e02b17594bc58b910a3a4ac9cb2c with size: 11133
273090 bytes can be problematic as it may lead to oversplitting.
23/08/06 19:26:10 WARN mapreduce.LoadIncrementalHFiles: Trying to bulk load hfile hdfs://ip-172-31-13-99.ec2.internal:8020/user/hadoop/NYC_TRIPS/trip_data/7eda95892d0144f39a18af93fa7b425d with size: 11133
377164 bytes can be problematic as it may lead to oversplitting.
23/08/06 19:26:10 INFO Configuration.deprecation: hbase.offheapcache.minblocksize is deprecated. Instead, use hbase.blockcache.minblocksize
[hadoop@ip-172-31-13-99 ~]$ ff
bash: ff: command not found
[hadoop@ip-172-31-13-99 ~]$


Downloads — hadoop@ip-172-31-13-99:~ — ssh -i mac-academy-key.pem ec2-user@44.203.238.44 — 204×55
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
nyc_trips_hbase
1 row(s) in 0.2480 seconds
=> ["nyc_trips_hbase"]
hbase(main):002:0> describe 'nyc_trips_hbase'
Table nyc_trips_hbase is ENABLED
nyc_trips_hbase
COLUMN FAMILIES DESCRIPTION
{NAME => 'trip_data', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '
', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.1040 seconds
hbase(main):003:0> count 'nyc_trips_hbase'
Current count: 1000, row: 1_2017-01-01 00:11:36.0_2017-01-01 00:24:48.0
Current count: 2000, row: 1_2017-01-01 00:18:15.0_2017-01-01 00:26:33.0
Current count: 3000, row: 1_2017-01-01 00:23:58.0_2017-01-01 00:42:06.0
Current count: 4000, row: 1_2017-01-01 00:29:40.0_2017-01-01 00:46:41.0
Current count: 5000, row: 1_2017-01-01 00:35:28.0_2017-01-01 00:48:58.0
Current count: 6000, row: 1_2017-01-01 00:41:06.0_2017-01-01 01:00:39.0
Current count: 7000, row: 1_2017-01-01 00:46:43.0_2017-01-01 01:06:23.0
Current count: 8000, row: 1_2017-01-01 00:52:09.0_2017-01-01 00:59:09.0
Current count: 9000, row: 1_2017-01-01 00:57:37.0_2017-01-01 01:11:55.0
Current count: 10000, row: 1_2017-01-01 01:02:58.0_2017-01-01 01:09:05.0
Current count: 11000, row: 1_2017-01-01 01:08:31.0_2017-01-01 01:14:57.0
Current count: 12000, row: 1_2017-01-01 01:13:50.0_2017-01-01 01:25:21.0
Current count: 13000, row: 1_2017-01-01 01:19:34.0_2017-01-01 01:27:16.0
Current count: 14000, row: 1_2017-01-01 01:25:05.0_2017-01-01 01:36:39.0
Current count: 15000, row: 1_2017-01-01 01:30:50.0_2017-01-01 01:39:13.0
Current count: 16000, row: 1_2017-01-01 01:36:48.0_2017-01-01 01:42:39.0
Current count: 17000, row: 1_2017-01-01 01:42:38.0_2017-01-01 01:51:12.0
Current count: 18000, row: 1_2017-01-01 01:48:15.0_2017-01-01 02:03:10.0
Current count: 19000, row: 1_2017-01-01 01:54:15.0_2017-01-01 02:09:29.0
Current count: 20000, row: 1_2017-01-01 02:00:02.0_2017-01-01 02:33:57.0
Current count: 21000, row: 1_2017-01-01 02:06:23.0_2017-01-01 02:10:00.0
Current count: 22000, row: 1_2017-01-01 02:12:21.0_2017-01-01 02:25:16.0
Current count: 23000, row: 1_2017-01-01 02:18:02.0_2017-01-01 02:39:10.0
Current count: 24000, row: 1_2017-01-01 02:24:12.0_2017-01-01 02:33:02.0
Current count: 25000, row: 1_2017-01-01 02:30:08.0_2017-01-01 02:37:26.0
Current count: 26000, row: 1_2017-01-01 02:36:28.0_2017-01-01 03:06:36.0
Current count: 27000, row: 1_2017-01-01 02:42:39.0_2017-01-01 03:07:05.0
Current count: 28000, row: 1_2017-01-01 02:48:47.0_2017-01-01 03:18:16.0
Current count: 29000, row: 1_2017-01-01 02:55:12.0_2017-01-01 03:20:59.0
Current count: 30000, row: 1_2017-01-01 03:01:54.0_2017-01-01 03:25:22.0
Current count: 31000, row: 1_2017-01-01 03:08:42.0_2017-01-01 03:29:51.0
Current count: 32000, row: 1_2017-01-01 03:15:29.0_2017-01-01 03:23:57.0
Current count: 33000, row: 1_2017-01-01 03:22:43.0_2017-01-01 03:32:34.0
Current count: 34000, row: 1_2017-01-01 03:29:54.0_2017-01-01 03:54:30.0
Current count: 35000, row: 1_2017-01-01 03:37:23.0_2017-01-01 03:47:56.0