# Data Ingestion Tasks

## Task 2: Ingest data from RDS to HBase table

# Install MySQL connector

- Switch to sudo user using “sudo -i”
- Perform the following commands.
  - 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
  - cd mysql-connector-java-8.0.25/
  - sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/

# Create HBase table

- Login to EMR
- Switch to sudo user using “sudo -i” command
- Now login to HBase using “hbase shell” command
- Create table using below syntax:
  - **Syntax:** create ‘taxi_tripdata’, ‘trip_info’
- Check the HBase table:
  - describe ‘taxi_tripdata’

# Run Sqoop job

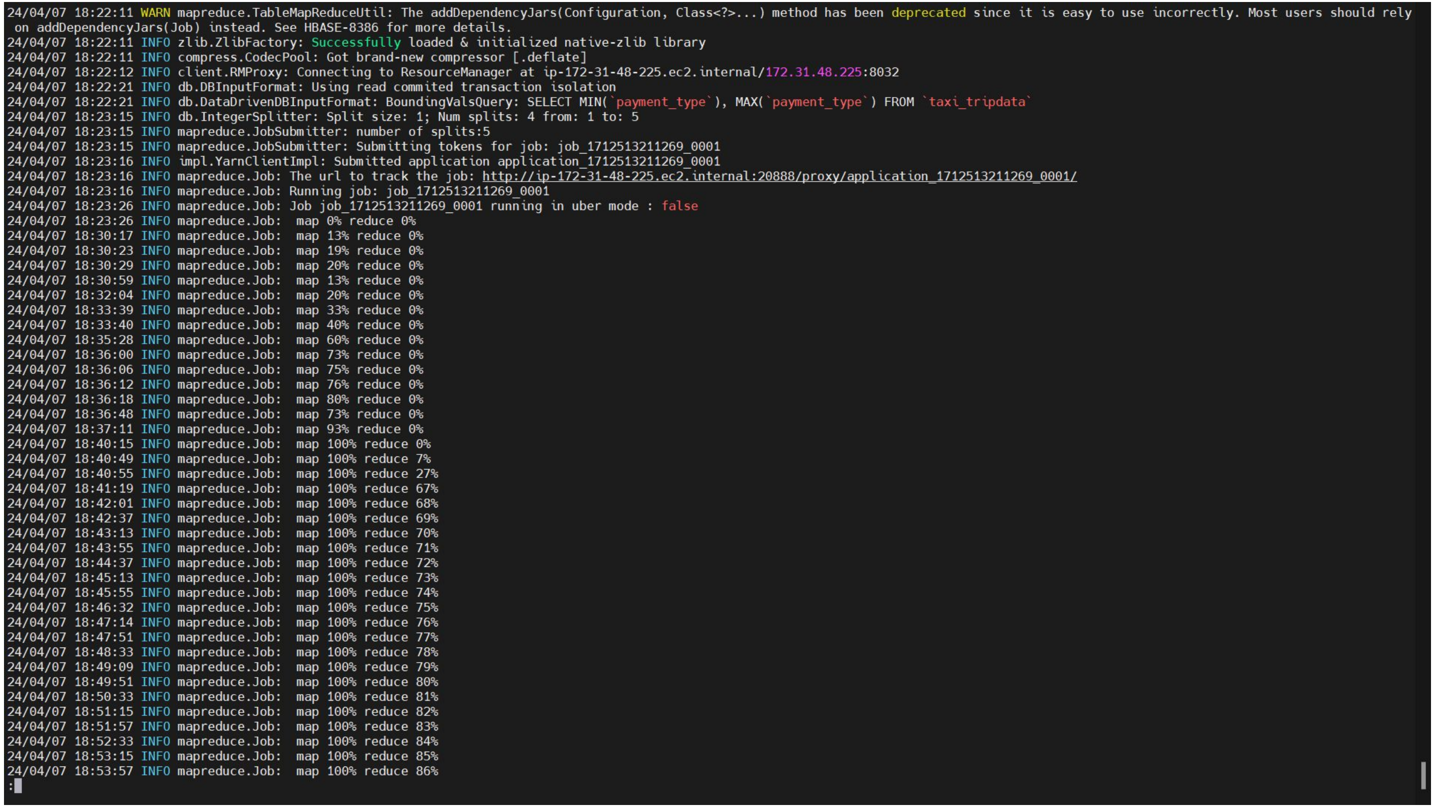
- Run the following sqoop command to migrate data from SQL to NoSQL.

```
sqoop import --connect
jdbc:mysql://rds-for-assignment.cl6qig6uspf6.us-east-1.rds.amazonaws.com/taxi_data \
--username admin --password 12345678 \
--table taxi_tripdata \
--hbase-table taxi_tripdata_hbase --column-family trip_info \
--hbase-row-key tpep_pickup_datetime,tpep_dropoff_datetime \
--hbase-create-table \
--hbase-bulkload \
--split-by payment_type -m 20
```

- Command explanation

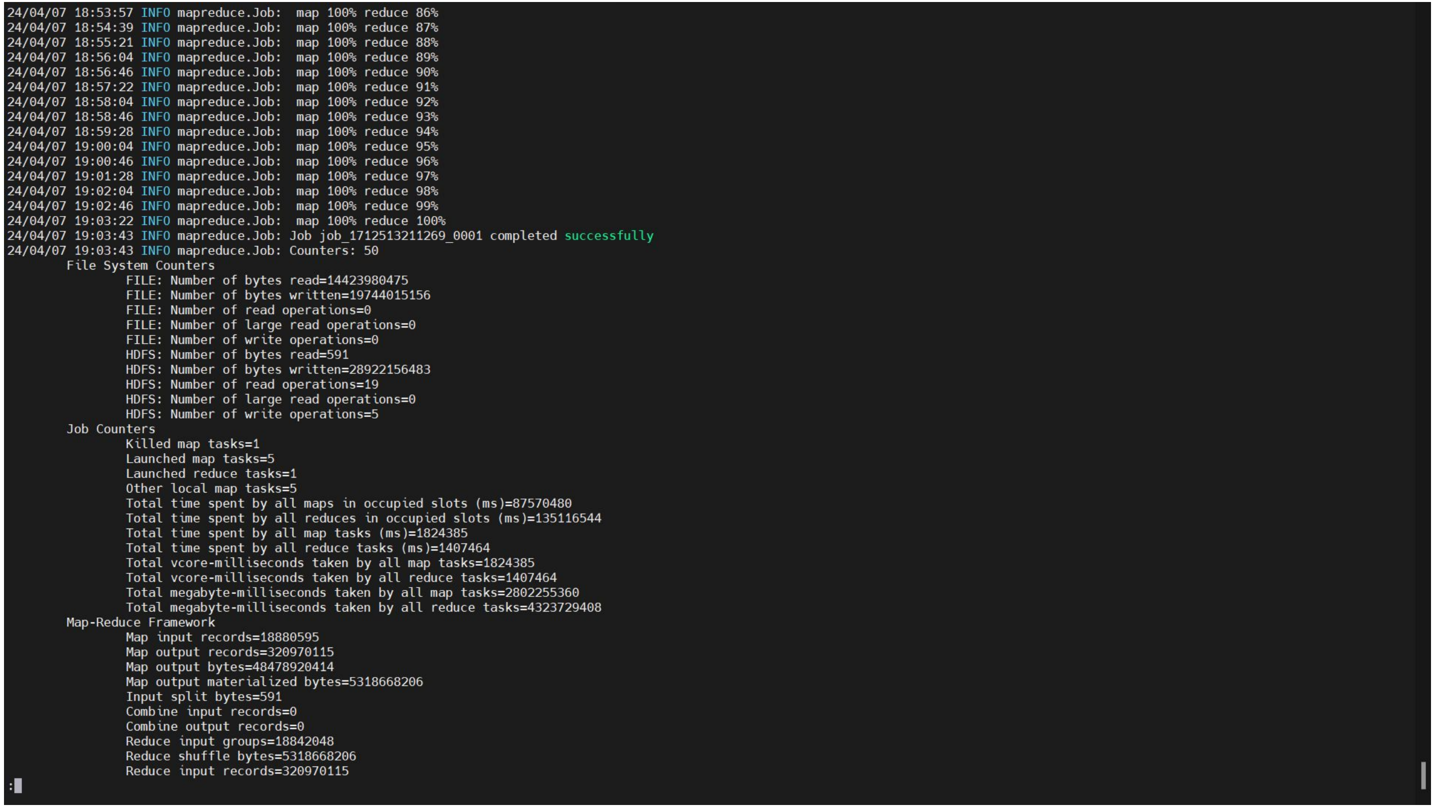
--connect: Defines connection string to the source MySQL database.
--username/--password: Username and password for the MySQL database.
--table: Specifies the table in MySQL to import data from.
--hbase-table: Names the HBase table to store the imported data.
--column-family: Defines the category within HBase to store the data.
--hbase-row-key: Sets the unique identifier for each record in HBase (uses two columns here).
--hbase-create-table: Creates the HBase table if it doesn't exist.
--hbase-bulkload: Uses a faster bulk loading approach for import.
--split-by: Splits the import job based on a specific column.
-m: Specifies the number of mappers (parallel processes) to use for import.

# Running the Sqoop job



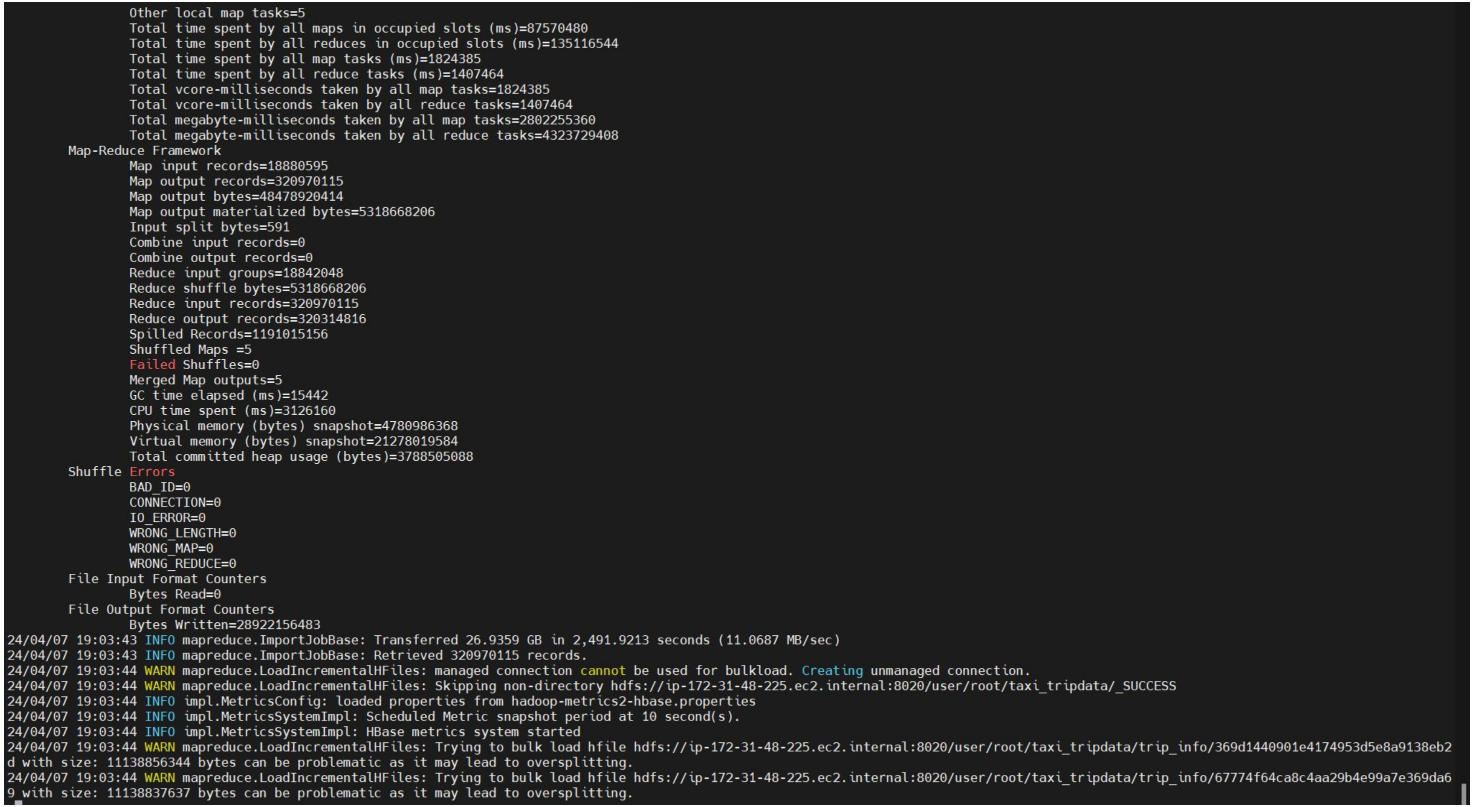
```
24/04/07 18:22:11 WARN mapreduce.TableMapReduceUtil: The addDependencyJars(Configuration, Class<?>...) method has been deprecated since it is easy to use incorrectly. Most users should rely
 on addDependencyJars(Job) instead. See HBASE-8386 for more details.
24/04/07 18:22:11 INFO zlib.ZlibFactory: Successfully loaded & initialized native-zlib library
24/04/07 18:22:11 INFO compress.CodecPool: Got brand-new compressor [.deflate]
24/04/07 18:22:12 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-48-225.ec2.internal/172.31.48.225:8032
24/04/07 18:22:21 INFO db.DBInputFormat: Using read commited transaction isolation
24/04/07 18:22:21 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`payment_type`), MAX(`payment_type`) FROM `taxi_tripdata`
24/04/07 18:23:15 INFO db.IntegerSplitter: Split size: 1; Num splits: 4 from: 1 to: 5
24/04/07 18:23:15 INFO mapreduce.JobSubmitter: number of splits:5
24/04/07 18:23:15 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1712513211269_0001
24/04/07 18:23:16 INFO impl.YarnClientImpl: Submitted application application_1712513211269_0001
24/04/07 18:23:16 INFO mapreduce.Job: The url to track the job: http://ip-172-31-48-225.ec2.internal:20888/proxy/application_1712513211269_0001/
24/04/07 18:23:16 INFO mapreduce.Job: Running job: job_1712513211269_0001
24/04/07 18:23:26 INFO mapreduce.Job: Job job_1712513211269_0001 running in uber mode : false
24/04/07 18:23:26 INFO mapreduce.Job:  map 0% reduce 0%
24/04/07 18:30:17 INFO mapreduce.Job:  map 13% reduce 0%
24/04/07 18:30:23 INFO mapreduce.Job:  map 19% reduce 0%
24/04/07 18:30:29 INFO mapreduce.Job:  map 20% reduce 0%
24/04/07 18:30:59 INFO mapreduce.Job:  map 13% reduce 0%
24/04/07 18:32:04 INFO mapreduce.Job:  map 20% reduce 0%
24/04/07 18:33:39 INFO mapreduce.Job:  map 33% reduce 0%
24/04/07 18:33:40 INFO mapreduce.Job:  map 40% reduce 0%
24/04/07 18:35:28 INFO mapreduce.Job:  map 60% reduce 0%
24/04/07 18:36:00 INFO mapreduce.Job:  map 73% reduce 0%
24/04/07 18:36:06 INFO mapreduce.Job:  map 75% reduce 0%
24/04/07 18:36:12 INFO mapreduce.Job:  map 76% reduce 0%
24/04/07 18:36:18 INFO mapreduce.Job:  map 80% reduce 0%
24/04/07 18:36:48 INFO mapreduce.Job:  map 73% reduce 0%
24/04/07 18:37:11 INFO mapreduce.Job:  map 93% reduce 0%
24/04/07 18:40:15 INFO mapreduce.Job:  map 100% reduce 0%
24/04/07 18:40:49 INFO mapreduce.Job:  map 100% reduce 7%
24/04/07 18:40:55 INFO mapreduce.Job:  map 100% reduce 27%
24/04/07 18:41:19 INFO mapreduce.Job:  map 100% reduce 67%
24/04/07 18:42:01 INFO mapreduce.Job:  map 100% reduce 68%
24/04/07 18:42:37 INFO mapreduce.Job:  map 100% reduce 69%
24/04/07 18:43:13 INFO mapreduce.Job:  map 100% reduce 70%
24/04/07 18:43:55 INFO mapreduce.Job:  map 100% reduce 71%
24/04/07 18:44:37 INFO mapreduce.Job:  map 100% reduce 72%
24/04/07 18:45:13 INFO mapreduce.Job:  map 100% reduce 73%
24/04/07 18:45:55 INFO mapreduce.Job:  map 100% reduce 74%
24/04/07 18:46:32 INFO mapreduce.Job:  map 100% reduce 75%
24/04/07 18:47:14 INFO mapreduce.Job:  map 100% reduce 76%
24/04/07 18:47:51 INFO mapreduce.Job:  map 100% reduce 77%
24/04/07 18:48:33 INFO mapreduce.Job:  map 100% reduce 78%
24/04/07 18:49:09 INFO mapreduce.Job:  map 100% reduce 79%
24/04/07 18:49:51 INFO mapreduce.Job:  map 100% reduce 80%
24/04/07 18:50:33 INFO mapreduce.Job:  map 100% reduce 81%
24/04/07 18:51:15 INFO mapreduce.Job:  map 100% reduce 82%
24/04/07 18:51:57 INFO mapreduce.Job:  map 100% reduce 83%
24/04/07 18:52:33 INFO mapreduce.Job:  map 100% reduce 84%
24/04/07 18:53:15 INFO mapreduce.Job:  map 100% reduce 85%
24/04/07 18:53:57 INFO mapreduce.Job:  map 100% reduce 86%
:
```

# Running the Sqoop job



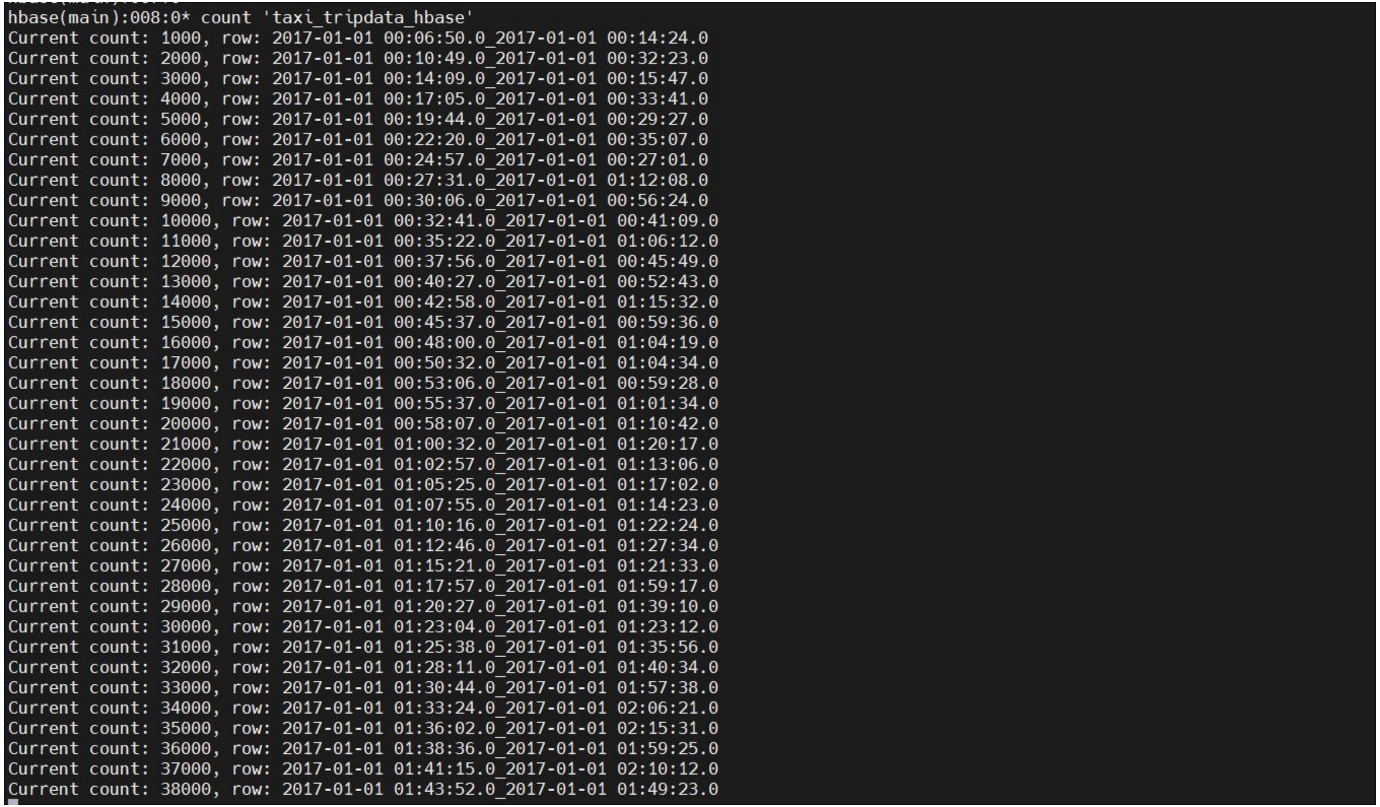
```
24/04/07 18:53:57 INFO mapreduce.Job:  map 100% reduce 86%
24/04/07 18:54:39 INFO mapreduce.Job:  map 100% reduce 87%
24/04/07 18:55:21 INFO mapreduce.Job:  map 100% reduce 88%
24/04/07 18:56:04 INFO mapreduce.Job:  map 100% reduce 89%
24/04/07 18:56:46 INFO mapreduce.Job:  map 100% reduce 90%
24/04/07 18:57:22 INFO mapreduce.Job:  map 100% reduce 91%
24/04/07 18:58:04 INFO mapreduce.Job:  map 100% reduce 92%
24/04/07 18:58:46 INFO mapreduce.Job:  map 100% reduce 93%
24/04/07 18:59:28 INFO mapreduce.Job:  map 100% reduce 94%
24/04/07 19:00:04 INFO mapreduce.Job:  map 100% reduce 95%
24/04/07 19:00:46 INFO mapreduce.Job:  map 100% reduce 96%
24/04/07 19:01:28 INFO mapreduce.Job:  map 100% reduce 97%
24/04/07 19:02:04 INFO mapreduce.Job:  map 100% reduce 98%
24/04/07 19:02:46 INFO mapreduce.Job:  map 100% reduce 99%
24/04/07 19:03:22 INFO mapreduce.Job:  map 100% reduce 100%
24/04/07 19:03:43 INFO mapreduce.Job: Job job_1712513211269_0001 completed successfully
24/04/07 19:03:43 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=14423980475
                FILE: Number of bytes written=19744015156
                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=28922156483
                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)=87570480
                Total time spent by all reduces in occupied slots (ms)=135116544
                Total time spent by all map tasks (ms)=1824385
                Total time spent by all reduce tasks (ms)=1407464
                Total vcore-milliseconds taken by all map tasks=1824385
                Total vcore-milliseconds taken by all reduce tasks=1407464
                Total megabyte-milliseconds taken by all map tasks=2802255360
                Total megabyte-milliseconds taken by all reduce tasks=4323729408
        Map-Reduce Framework
                Map input records=18880595
                Map output records=320970115
                Map output bytes=48478920414
                Map output materialized bytes=5318668206
                Input split bytes=591
                Combine input records=0
                Combine output records=0
                Reduce input groups=18842048
                Reduce shuffle bytes=5318668206
                Reduce input records=320970115
:
```

# Running the Sqoop job



```
                Other local map tasks=5
                Total time spent by all maps in occupied slots (ms)=87570480
                Total time spent by all reduces in occupied slots (ms)=135116544
                Total time spent by all map tasks (ms)=1824385
                Total time spent by all reduce tasks (ms)=1407464
                Total vcore-milliseconds taken by all map tasks=1824385
                Total vcore-milliseconds taken by all reduce tasks=1407464
                Total megabyte-milliseconds taken by all map tasks=2802255360
                Total megabyte-milliseconds taken by all reduce tasks=4323729408
        Map-Reduce Framework
                Map input records=18880595
                Map output records=320970115
                Map output bytes=48478920414
                Map output materialized bytes=5318668206
                Input split bytes=591
                Combine input records=0
                Combine output records=0
                Reduce input groups=18842048
                Reduce shuffle bytes=5318668206
                Reduce input records=320970115
                Reduce output records=320314816
                Spilled Records=1191015156
                Shuffled Maps =5
                Failed Shuffles=0
                Merged Map outputs=5
                GC time elapsed (ms)=15442
                CPU time spent (ms)=3126160
                Physical memory (bytes) snapshot=4780986368
                Virtual memory (bytes) snapshot=21278019584
                Total committed heap usage (bytes)=3788505088
        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=28922156483
24/04/07 19:03:43 INFO mapreduce.ImportJobBase: Transferred 26.9359 GB in 2,491.9213 seconds (11.0687 MB/sec)
24/04/07 19:03:43 INFO mapreduce.ImportJobBase: Retrieved 320970115 records.
24/04/07 19:03:44 WARN mapreduce.LoadIncrementalHFiles: managed connection cannot be used for bulkload. Creating unmanaged connection.
24/04/07 19:03:44 WARN mapreduce.LoadIncrementalHFiles: Skipping non-directory hdfs://ip-172-31-48-225.ec2.internal:8020/user/root/taxi_tripdata/_SUCCESS
24/04/07 19:03:44 INFO impl.MetricsConfig: loaded properties from hadoop-metrics2-hbase.properties
24/04/07 19:03:44 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
24/04/07 19:03:44 INFO impl.MetricsSystemImpl: HBase metrics system started
24/04/07 19:03:44 WARN mapreduce.LoadIncrementalHFiles: Trying to bulk load hfile hdfs://ip-172-31-48-225.ec2.internal:8020/user/root/taxi_tripdata/trip_info/369d1440901e4174953d5e8a9138eb2
d with size: 11138856344 bytes can be problematic as it may lead to oversplitting.
24/04/07 19:03:44 WARN mapreduce.LoadIncrementalHFiles: Trying to bulk load hfile hdfs://ip-172-31-48-225.ec2.internal:8020/user/root/taxi_tripdata/trip_info/67774f64ca8c4aa29b4e99a7e369da6
9 with size: 11138837637 bytes can be problematic as it may lead to oversplitting.
```

# Data migrated to HBase table



```
hbase(main):008:0* count 'taxi_tripdata_hbase'
Current count: 1000, row: 2017-01-01 00:06:50.0_2017-01-01 00:14:24.0
Current count: 2000, row: 2017-01-01 00:10:49.0_2017-01-01 00:32:23.0
Current count: 3000, row: 2017-01-01 00:14:09.0_2017-01-01 00:15:47.0
Current count: 4000, row: 2017-01-01 00:17:05.0_2017-01-01 00:33:41.0
Current count: 5000, row: 2017-01-01 00:19:44.0_2017-01-01 00:29:27.0
Current count: 6000, row: 2017-01-01 00:22:20.0_2017-01-01 00:35:07.0
Current count: 7000, row: 2017-01-01 00:24:57.0_2017-01-01 00:27:01.0
Current count: 8000, row: 2017-01-01 00:27:31.0_2017-01-01 01:12:08.0
Current count: 9000, row: 2017-01-01 00:30:06.0_2017-01-01 00:56:24.0
Current count: 10000, row: 2017-01-01 00:32:41.0_2017-01-01 00:41:09.0
Current count: 11000, row: 2017-01-01 00:35:22.0_2017-01-01 01:06:12.0
Current count: 12000, row: 2017-01-01 00:37:56.0_2017-01-01 00:45:49.0
Current count: 13000, row: 2017-01-01 00:40:27.0_2017-01-01 00:52:43.0
Current count: 14000, row: 2017-01-01 00:42:58.0_2017-01-01 01:15:32.0
Current count: 15000, row: 2017-01-01 00:45:37.0_2017-01-01 00:59:36.0
Current count: 16000, row: 2017-01-01 00:48:00.0_2017-01-01 01:04:19.0
Current count: 17000, row: 2017-01-01 00:50:32.0_2017-01-01 01:04:34.0
Current count: 18000, row: 2017-01-01 00:53:06.0_2017-01-01 00:59:28.0
Current count: 19000, row: 2017-01-01 00:55:37.0_2017-01-01 01:01:34.0
Current count: 20000, row: 2017-01-01 00:58:07.0_2017-01-01 01:10:42.0
Current count: 21000, row: 2017-01-01 01:00:32.0_2017-01-01 01:20:17.0
Current count: 22000, row: 2017-01-01 01:02:57.0_2017-01-01 01:13:06.0
Current count: 23000, row: 2017-01-01 01:05:25.0_2017-01-01 01:17:02.0
Current count: 24000, row: 2017-01-01 01:07:55.0_2017-01-01 01:14:23.0
Current count: 25000, row: 2017-01-01 01:10:16.0_2017-01-01 01:22:24.0
Current count: 26000, row: 2017-01-01 01:12:46.0_2017-01-01 01:27:34.0
Current count: 27000, row: 2017-01-01 01:15:21.0_2017-01-01 01:21:33.0
Current count: 28000, row: 2017-01-01 01:17:57.0_2017-01-01 01:59:17.0
Current count: 29000, row: 2017-01-01 01:20:27.0_2017-01-01 01:39:10.0
Current count: 30000, row: 2017-01-01 01:23:04.0_2017-01-01 01:23:12.0
Current count: 31000, row: 2017-01-01 01:25:38.0_2017-01-01 01:35:56.0
Current count: 32000, row: 2017-01-01 01:28:11.0_2017-01-01 01:40:34.0
Current count: 33000, row: 2017-01-01 01:30:44.0_2017-01-01 01:57:38.0
Current count: 34000, row: 2017-01-01 01:33:24.0_2017-01-01 02:06:21.0
Current count: 35000, row: 2017-01-01 01:36:02.0_2017-01-01 02:15:31.0
Current count: 36000, row: 2017-01-01 01:38:36.0_2017-01-01 01:59:25.0
Current count: 37000, row: 2017-01-01 01:41:15.0_2017-01-01 02:10:12.0
Current count: 38000, row: 2017-01-01 01:43:52.0_2017-01-01 01:49:23.0
```