

Task 2

Task 2. Use Sqoop command to ingest the data from RDS into the HBase Table.

- 1) Previous Task we have created the RDS Table and load data from csv files.
- 2) Now we will create HBase table and by using sqoop will import RDS table data to HBase table

Create HBase Table – HBase table named 'trips_data' with a column family named 'cf'

```
hbase(main):007:0* create 'trips_data', 'cf'
0 row(s) in 1.3010 seconds
=> Hbase::Table - trips_data
```

Sqoop import –

We provided RDS server connection details here to get Trips table records. And passed RDS/HBase table Name, Column-family and Row-key to insert the data into HBase table by using 4 mappers.

```
sqoop import --connect jdbc:mysql://case-study-dbb.ck4jzoqb1yn7.us-east-1.rds.amazonaws.com:3306/YellowTaxi \
--username admin \
--password user1234 \
--table taxi \
--hbase-table trips_data \
--column-family cf \
--split-by vendorID \
--num-mappers 4;
```

```
root@ip-172-31-44-105:/home/hadoop
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.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]
23/05/13 08:04:03 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
23/05/13 08:04:03 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/05/13 08:04:04 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/05/13 08:04:04 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.
23/05/13 08:04:04 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'taxi' AS t LIMIT 1
23/05/13 08:04:04 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'taxi' AS t LIMIT 1
23/05/13 08:04:04 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/fc0ec7f0c3e4f5260eeb14af4aa253a5/taxi.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/05/13 08:04:07 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/fc0ec7f0c3e4f5260eeb14af4aa253a5/taxi.jar
23/05/13 08:04:07 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/05/13 08:04:07 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/05/13 08:04:07 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/05/13 08:04:07 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/05/13 08:04:07 INFO mapreduce.ImportJobBase: Beginning import of taxi
23/05/13 08:04:08 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/05/13 08:04:08 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/05/13 08:04:10 INFO mapreduce.HBaseImportJob: Creating missing HBase table taxi_hb
23/05/13 08:04:12 WARN mapreduce.TableMapReduceUtil: The addDependencyJars(Configuration, Class<?>...) method has been deprecated since it is easy to use inc
orrectly. Most users should rely on addDependencyJars(Job) instead. See HBASE-8386 for more details.
23/05/13 08:04:12 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-44-105.ec2.internal/172.31.44.105:8032
23/05/13 08:04:13 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-44-105.ec2.internal/172.31.44.105:10200
23/05/13 08:04:26 INFO db.DBInputFormat: Using read committed transaction isolation
23/05/13 08:04:26 INFO db.DataDrivenDBInputFormat: BoundingValuesQuery: SELECT MIN('vendorID'), MAX('vendorID') FROM 'taxi'
23/05/13 08:05:23 INFO db.IntegerSplitter: Split size: 0; Num splits: 4 from: 1 to: 2
23/05/13 08:05:23 INFO mapreduce.JobSubmitter: number of splits:2
23/05/13 08:05:24 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publi
sher.enabled
23/05/13 08:05:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1683961959498_0001
23/05/13 08:05:24 INFO conf.Configuration: resource-types.xml not found
23/05/13 08:05:24 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/05/13 08:05:24 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
23/05/13 08:05:24 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
23/05/13 08:05:24 INFO impl.YarnClientImpl: Submitted application application_1683961959498_0001
23/05/13 08:05:25 INFO mapreduce.Job: The url to track the job: http://ip-172-31-44-105.ec2.internal:20888/proxy/application_1683961959498_0001/
23/05/13 08:05:25 INFO mapreduce.Job: Running job: job_1683961959498_0001
23/05/13 08:05:34 INFO mapreduce.Job: Job job_1683961959498_0001 running in uber mode : false
23/05/13 08:05:34 INFO mapreduce.Job: map 0% reduce 0%
```

```
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=225212
  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)=264432
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=5509
  Total vcore-milliseconds taken by all map tasks=5509
  Total megabyte-milliseconds taken by all map tasks=8461824
Map-Reduce Framework
  Map input records=198
  Map output records=198
  Input split bytes=87
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=187
  CPU time spent (ms)=5440
  Physical memory (bytes) snapshot=367321088
  Virtual memory (bytes) snapshot=3315101696
  Total committed heap usage (bytes)=328204288
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=0
23/05/07 07:11:34 INFO mapreduce.ImportJobBase: Transferred 0 bytes in 26.9089 seconds (0 bytes
23/05/07 07:11:34 INFO mapreduce.ImportJobBase: Retrieved 198 records.
```

HBase Table Records After Sqoop import –

