

Data Ingestion from the RDS to HDFS using Sqoop

Sqoop Import command used for importing table from RDS to HDFS:

I used the sqoop command to import the ATM transactions data from RDS to Hadoop HDFS directory.

```
sqoop import --connect jdbc:mysql://upgradtest.cyaie1c9bmnf.us-east-1.rds.amazonaws.com/testdatabase --username student --password STUDENT123 --table SRC_ATM_TRANS --target-dir /user/hadoop/SRC_ATM_TRANS --m 1
```

```
Thanks for using MariaDB!
D:\adoop>ip-172-31-91-112 mysql-connector-java-8.0.25\% sqoop import --connect jdbc:mysql://upgradtest.cyaie1c9bmnf.us-east-1.rds.amazonaws.com/testdatabase --username student --password STUDENT123 --table SRC_ATM_TRANS --target-dir /user/hadoop/SRC_ATM_TRANS --m 1
Warning: /usr/lib/sqoop/.hibase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/.hcatalog does not exist! HCatalog jobs will fail.
Please set $HCATALOG_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/.accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/06/19 18:14:01 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
22/06/19 18:14:01 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/06/19 18:14:02 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/06/19 18:14:02 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.
22/06/19 18:14:02 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1
22/06/19 18:14:02 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1
22/06/19 18:14:03 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/debb95c6f8e45b461c3cfab8dc54106c/SRC_ATM_TRANS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/06/19 18:14:03 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/debb95c6f8e45b461c3cfab8dc54106c/SRC_ATM_TRANS.jar
22/06/19 18:14:05 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/06/19 18:14:05 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/06/19 18:14:05 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/06/19 18:14:05 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/06/19 18:14:05 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS

Total time spent by all maps in occupied slots (ms)=1209128
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=26440
Total vcore-milliseconds taken by all map tasks=26440
Total megabyte-milliseconds taken by all map tasks=40611840
Map-Reduce Framework
  Map input records=2468372
  Map output records=2468372
  Input split bytes=87
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=213
  CPU time spent (ms)=10790
  Physical memory (bytes) snapshot=631484416
  Virtual memory (bytes) snapshot=3309596072
  Total committed heap usage (bytes)=520617984
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=33214815
22/06/19 18:15:06 INFO mapreduce.ImportJobBase: Transferred 506.6859 MB in 60.17 seconds (8.4196 MB/sec)
22/06/19 18:15:06 INFO mapreduce.ImportJobBase: Retrieved 2468372 records.
```

Command used to see the list of imported data in HDFS:

Used the command line to see the imported data in the following hdfs directory:

```
hadoop fs -cat /user/hadoop/SRC_ATM_TRANS/part-m-00000
```

part-m-00000 contains the csv of the imported file

Screenshot of the imported data:

[illegible]