Data Ingestion from the RDS to HDFS using Sqoop

Steps followed:

1. SSH into EMR primary node
2. Install Mysql connector using following commands
    a. wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
    b. tar -xvf mysql-connector-java-8.0.25.tar.gz
    c. cd mysql-connector-java-8.0.25
    d. sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/
    1.
3. Data import from RDS to HDFS using following sqoop import command

sqoop import ****

--connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase ****

--username student ****

--password STUDENT123 ****

--table SRC_ATM_TRANS ****

--target-dir /user/root/SRC_ATM_TRANS ****

-m 1

```
[root@ip-10-0-12-204 ~]# sqoop import \
> --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
> --username student \
> --password STUDENT123 \
> --table SRC_ATM_TRANS \
> --target-dir /user/root/SRC_ATM_TRANS \
> -m 1
Warning: /usr/lib/sqoop/../hbase 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 $HCAT_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.
23/12/03 07:13:33 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: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
23/12/03 07:13:33 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/12/03 07:13:34 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/12/03 07:13:34 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.
23/12/03 07:13:34 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
23/12/03 07:13:34 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
23/12/03 07:13:34 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/ad597a4bd14b593e3e9cff4c7a20c79b/SRC_ATM_TRANS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/12/03 07:13:36 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/ad597a4bd14b593e3e9cff4c7a20c79b/SRC_ATM_TRANS.jar
23/12/03 07:13:36 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/12/03 07:13:36 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/12/03 07:13:36 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/12/03 07:13:36 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/12/03 07:13:36 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS
23/12/03 07:13:37 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/12/03 07:13:38 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/12/03 07:13:38 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-12-204.ec2.internal/10.0.12.204:8032
23/12/03 07:13:46 INFO db.DBInputFormat: Using read commited transaction isolation
23/12/03 07:13:46 INFO mapreduce.JobSubmitter: number of splits:1
23/12/03 07:13:47 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1701586859712_0001
23/12/03 07:13:48 INFO impl.YarnClientImpl: Submitted application application_1701586859712_0001
23/12/03 07:13:48 INFO mapreduce.Job: The url to track the job: http://ip-10-0-12-204.ec2.internal:20888/proxy/application_1701586859712_0001/
23/12/03 07:13:48 INFO mapreduce.Job: Running job: job_1701586859712_0001
23/12/03 07:13:58 INFO mapreduce.Job: Job job_1701586859712_0001 running in uber mode : false
23/12/03 07:13:58 INFO mapreduce.Job:  map 0% reduce 0%
23/12/03 07:14:56 INFO mapreduce.Job:  map 100% reduce 0%
23/12/03 07:14:57 INFO mapreduce.Job: Job job_1701586859712_0001 completed successfully
23/12/03 07:14:57 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189887
                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=531214815
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
```

```
23/12/03 07:13:58 INFO mapreduce.Job:  map 0% reduce 0%
23/12/03 07:14:56 INFO mapreduce.Job:  map 100% reduce 0%
23/12/03 07:14:57 INFO mapreduce.Job: Job job_1701586859712_0001 completed successfully
23/12/03 07:14:57 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189887
                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=531214815
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=2678352
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=55799
                Total vcore-milliseconds taken by all map tasks=55799
                Total megabyte-milliseconds taken by all map tasks=85707264
        Map-Reduce Framework
                Map input records=2468572
                Map output records=2468572
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=222
                CPU time spent (ms)=26670
                Physical memory (bytes) snapshot=618143744
                Virtual memory (bytes) snapshot=3297095680
                Total committed heap usage (bytes)=460849152
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
23/12/03 07:14:57 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 79.6338 seconds (6.3617 MB/sec)
23/12/03 07:14:57 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
```

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

hadoop fs -ls /user/root/SRC_ATM_TRANS

Command used to verify imported rows count

hadoop fs -ls /user/root/SRC_ATM_TRANS/part-m-00000 | wc -l

```
[hadoop@ip-10-0-12-204 ~]$ hadoop fs -ls /user/root/SRC_ATM_TRANS
Found 2 items
-rw-r--r--   1 root hadoop          0 2023-12-03 07:14 /user/root/SRC_ATM_TRANS/_SUCCESS
-rw-r--r--   1 root hadoop  531214815 2023-12-03 07:14 /user/root/SRC_ATM_TRANS/part-m-00000
[hadoop@ip-10-0-12-204 ~]$ hadoop fs -cat /user/root/SRC_ATM_TRANS/part-m-00000 | wc -l
2468572
[hadoop@ip-10-0-12-204 ~]$
```