Data Ingestion from the RDS to HDFS using Sqoop

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

```
sqoop import \
> --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-
east-1.rds.amazonaws.com/testdatabase \
> --table SRC_ATM_TRANS \
> --username student --password STUDENT123 \
> -m 1
```

In the screenshot below, I can see that as a result of Sqoop Import Job, 2468572 records have been retrieved (same as the checkpoint mentioned in the Validation document):

```
22/10/11 07:29:06 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
22/10/11 07:29:06 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
22/10/11 07:29:06 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/c644a33c0b4ee25e4e80a9719a584779/SRC_ATM_TRANS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/10/11 07:29:08 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/c644a33c0b4ee25e4e80a9719a584779/SRC_ATM_TRANS.jar
22/10/11 07:29:08 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/10/11 07:29:08 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/10/11 07:29:08 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/10/11 07:29:08 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/10/11 07:29:08 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS
22/10/11 07:29:09 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/10/11 07:29:09 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/10/11 07:29:10 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-70-144.ec2.internal/172.31.70.144:8032
22/10/11 07:29:14 INFO db.DBInputFormat: Using read commited transaction isolation
22/10/11 07:29:14 INFO mapreduce.JobSubmitter: number of splits:1
22/10/11 07:29:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1665463394387_0001
22/10/11 07:29:15 INFO impl.YarnClientImpl: Submitted application application_1665463394387_0001
22/10/11 07:29:15 INFO mapreduce.Job: The url to track the job: http://ip-172-31-70-144.ec2.internal:20888/proxy/application_1665463394387_0001/
22/10/11 07:29:15 INFO mapreduce.Job: Running job: job_1665463394387_0001
22/10/11 07:29:26 INFO mapreduce.Job: Job job_1665463394387_0001 running in uber mode : false
22/10/11 07:29:26 INFO mapreduce.Job:  map 0% reduce 0%
22/10/11 07:29:53 INFO mapreduce.Job:  map 100% reduce 0%
22/10/11 07:29:54 INFO mapreduce.Job: Job job_1665463394387_0001 completed successfully
22/10/11 07:29:54 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189412
                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)=1160016
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=24167
                Total vcore-milliseconds taken by all map tasks=24167
                Total megabyte-milliseconds taken by all map tasks=37120512
        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)=205
                CPU time spent (ms)=26460
                Physical memory (bytes) snapshot=616771584
                Virtual memory (bytes) snapshot=3288940544
                Total committed heap usage (bytes)=535298048
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
22/10/11 07:29:54 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 44.883 seconds (11.2872 MB/sec)
22/10/11 07:29:54 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-172-31-70-144 ~]#
```

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

Hadoop fs -ls /user/root/SRC_ATM_TRANS
In the screenshot below, I can see that the target directory contains 2 items:

- - The first file is the success file, indicating that the MapReduce job was successful.

- - The second file ‘part-m-00000’ is the one with all of the data I imported. Since I used
  only one mapper in my import command thus the data is in a single file.

```
                Bytes Written=531214815
22/10/11 07:29:54 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 44.883 seconds (11.2872 MB/sec)
22/10/11 07:29:54 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-172-31-70-144 ~]# hadoop fs -ls /user/root/SRC_ATM_TRANS
Found 2 items
-rw-r--r--   1 root hadoop          0 2022-10-11 07:29 /user/root/SRC_ATM_TRANS/_SUCCESS
-rw-r--r--   1 root hadoop  531214815 2022-10-11 07:29 /user/root/SRC_ATM_TRANS/part-m-00000
```

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

Screenshot of the imported data: