Data Ingestion from the RDS to HDFS using Sqoop

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

Before running the Sqoop import command, I made sure the target directory is not already created. Otherwise, the Sqoop import command would throw an error:

```
hadoop fs -rm -r /user/root/SRC_ATM_TRANS
```

Sqoop Import Command:

```
sqoop import \
--connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-
1.rds.amazonaws.com/testdatabase \
--table SRC_ATM_TRANS \
--username student --password STUDENT123 \
--target-dir /user/root/SRC_ATM_TRANS \
-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):

```
hadoop@ip-172-31-81-150:~
2024-05-02 06:23:10,208 INFO mapreduce.Job: Job job_1714629725995_0001 running i
n uber mode : false
2024-05-02 06:23:10,210 INFO mapreduce.Job:  map 0% reduce 0%
2024-05-02 06:23:41,556 INFO mapreduce.Job:  map 100% reduce 0%
2024-05-02 06:23:42,569 INFO mapreduce.Job: Job job_1714629725995_0001 completed
 successfully
2024-05-02 06:23:42,687 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=297853
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=85
                HDFS: Number of bytes written=531214815
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=1403808
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=29246
                Total vcore-milliseconds taken by all map tasks=29246
                Total megabyte-milliseconds taken by all map tasks=44921856
        Map-Reduce Framework
                Map input records=2468572
                Map output records=2468572
                Input split bytes=85
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=264
                CPU time spent (ms)=33120
                Physical memory (bytes) snapshot=630841344
                Virtual memory (bytes) snapshot=3085635584
                Total committed heap usage (bytes)=600309760
                Peak Map Physical memory (bytes)=630841344
                Peak Map Virtual memory (bytes)=3085635584
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
2024-05-02 06:23:42,693 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in
 54.1102 seconds (9.3625 MB/sec)
2024-05-02 06:23:42,696 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[hadoop@ip-172-31-81-150 ~]$
```

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.

When I open the 'part-m-00000' file using the following command, I can see all of the data that has been imported:

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

Screenshot of the imported data: