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/spar_nord_bank_atm \

-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):

```
23/04/09 11:32:01 INFO mapreduce.Job: Job job_1681038000113_0001 running in uber mode : false
23/04/09 11:32:01 INFO mapreduce.Job:  map 0% reduce 0%
23/04/09 11:32:26 INFO mapreduce.Job:  map 100% reduce 0%
23/04/09 11:32:28 INFO mapreduce.Job: Job job_1681038000113_0001 completed successfully
23/04/09 11:32:29 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189996
                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)=1111824
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=23163
                Total vcore-milliseconds taken by all map tasks=23163
                Total megabyte-milliseconds taken by all map tasks=35578368
        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)=198
                CPU time spent (ms)=26210
                Physical memory (bytes) snapshot=613900288
                Virtual memory (bytes) snapshot=3286200320
                Total committed heap usage (bytes)=533725184
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
23/04/09 11:32:29 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 43.263 seconds (11.7099 MB/sec)
23/04/09 11:32:29 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

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 the data I imported. Since I used only one mapper in my import command thus the data is in a single file.

```
[hadoop@ip-172-31-81-61 ~]$ hadoop fs -ls /user/root/spar_nord_bank_atm
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2023-04-09 11:32 /user/root/spar_nord_bank_atm/_SUCCESS
-rw-r--r--   1 hadoop hadoop  531214815 2023-04-09 11:32 /user/root/spar_nord_bank_atm/part-m-00000
```

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 a portion of the imported data: