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 \
--username student \
--password STUDENT123 \
--table SRC_ATM_TRANS \
--target-dir /user/hdfs/SRC_ATM_TRANS \
--m 1
```

```
        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)=239
                CPU time spent (ms)=27880
                Physical memory (bytes) snapshot=628543488
                Virtual memory (bytes) snapshot=3299880960
                Total committed heap usage (bytes)=527433728
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
24/07/28 21:01:58 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 43.4474 seconds (11.6602 MB/sec)
24/07/28 21:01:58 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[hadoop@ip-172-31-50-227 ~]$
```

The above screenshot shows Retrieved 2468572 records

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

```
hadoop fs -ls /user/root/SRC_ATM_TRANS
```

```
[hadoop@ip-172-31-50-227 ~]$ hadoop fs -ls /user/root/SRC_ATM_TRANS
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2024-07-28 21:01 /user/root/SRC_ATM_TRANS/_SUCCESS
-rw-r--r--   1 hadoop hadoop  531214815 2024-07-28 21:01 /user/root/SRC_ATM_TRANS/part-m-00000
[hadoop@ip-172-31-50-227 ~]$
```

In the screenshot above we can see two items:

- The first file is the success file, indicating that the MapReduce job was successful.
- The second file 'part-m-00000' is the one that I imported. Since I used only one mapper in my import command, thus the data is in a single file.

Screenshot of the imported data:

```
hadoop fs -cat/user/root/SRC_ATM_TRANS
```

A portion of data read from part-m-00000 file: