Data Ingestion from the RDS to HDFS using Sqoop

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

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

```
43640888181_0001
22/01/31 15:09:46 INFO impl.YarnClientImpl: Submitted application application_16
43640888181_0001
22/01/31 15:09:46 INFO mapreduce.Job: The url to track the job: http://ip-172-31
-21-158.ec2.internal:20888/proxy/application_1643640888181_0001/
22/01/31 15:09:46 INFO mapreduce.Job: Running job: job_1643640888181_0001
22/01/31 15:09:55 INFO mapreduce.Job: Job job_1643640888181_0001 running in uber
mode : false
22/01/31 15:09:55 INFO mapreduce.Job:  map 0% reduce 0%
22/01/31 15:10:26 INFO mapreduce.Job:  map 100% reduce 0%
22/01/31 15:10:26 INFO mapreduce.Job: Job job_1643640888181_0001 completed succe
ssfully
22/01/31 15:10:26 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189435
                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)=1370304
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=28548
                Total vcore-milliseconds taken by all map tasks=28548
                Total megabyte-milliseconds taken by all map tasks=43849728
        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)=322
                CPU time spent (ms)=28420
                Physical memory (bytes) snapshot=606138368
                Virtual memory (bytes) snapshot=3290247168
                Total committed heap usage (bytes)=538443776
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
22/01/31 15:10:26 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 49.92
26 seconds (10.1478 MB/sec)
22/01/31 15:10:26 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-172-31-21-158 ~]#
```

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

Command : hadoop fs -ls /user/root/SRC_ATM_TRANS

```
[root@ip-172-31-21-158 ~]# hadoop fs -ls /user/root/SRC_ATM_TRANS
Found 2 items
-rw-r--r--   1 root hadoop          0 2022-01-31 15:10 /user/root/SRC_ATM_TRANS/_SUCCESS
-rw-r--r--   1 root hadoop  531214815 2022-01-31 15:10 /user/root/SRC_ATM_TRANS/part-m-00000
[root@ip-172-31-21-158 ~]#
```

As per the above screenshot, I can see that target directory contains two items:
1. The first file is the success file, indicating that the MapReduce job was successful.
2. The second file 'part-m-00000' is the one with all of the data imported. Since in my import command i used only 1 mapper thus the data is in a single file.

Screenshot of the imported data:

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