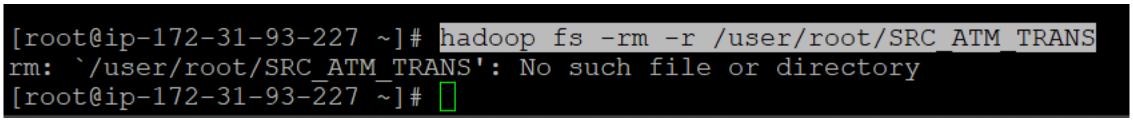
Data Ingestion from the RDS to HDFS using Sqoop

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

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

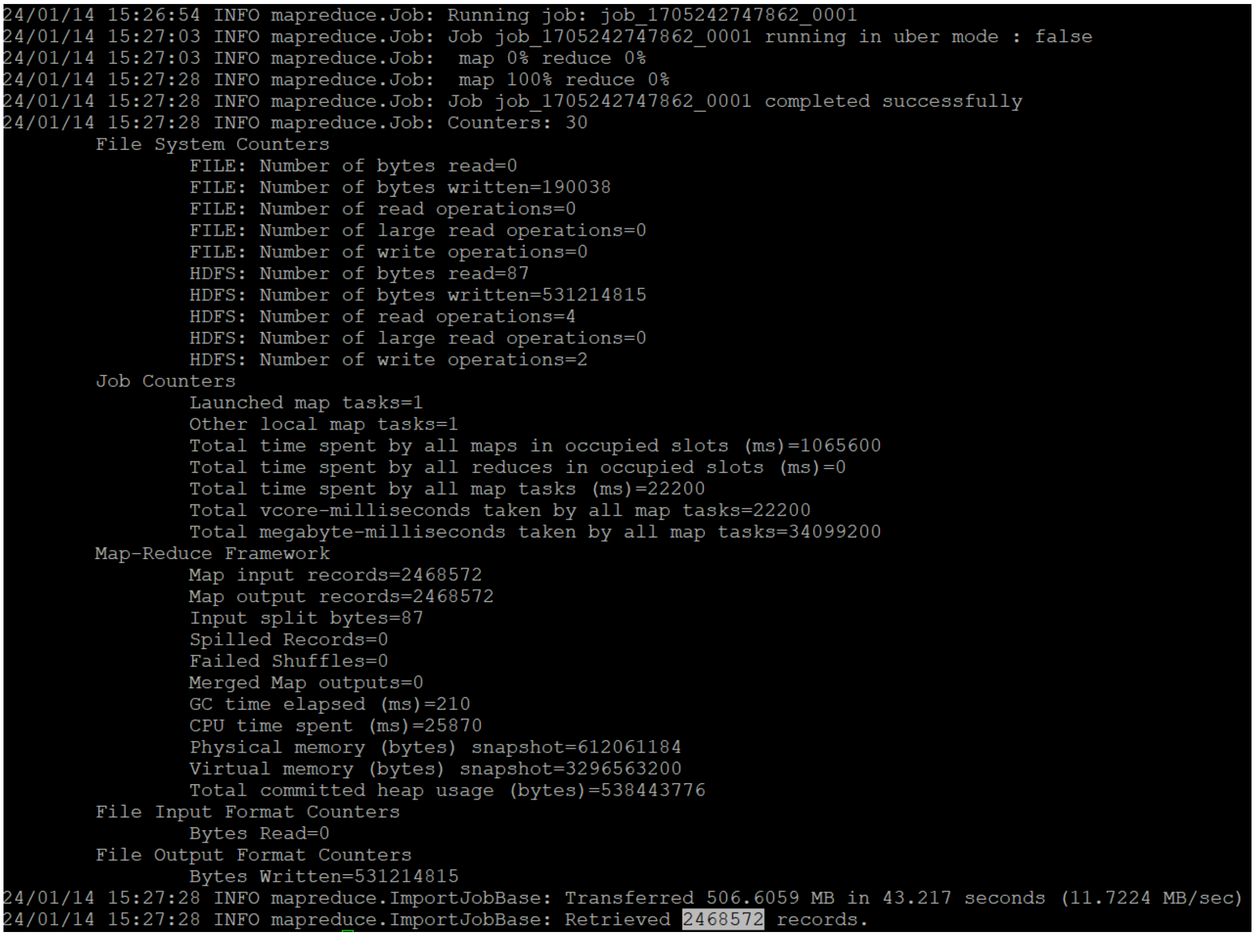
hadoop fs -rm -r /user/root/SRC_ATM_TRANS



```
[root@ip-172-31-93-227 ~]# hadoop fs -rm -r /user/root/SRC_ATM_TRANS
rm: `/user/root/SRC_ATM_TRANS': No such file or directory
[root@ip-172-31-93-227 ~]# 
```

Sqoop import command is used to import the data from mysql server to HDFS folder.

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

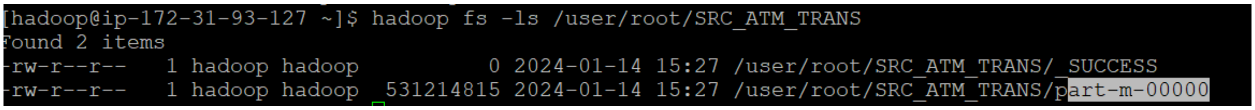


```
24/01/14 15:26:54 INFO mapreduce.Job: Running job: job_1705242747862_0001
24/01/14 15:27:03 INFO mapreduce.Job: Job job_1705242747862_0001 running in uber mode : false
24/01/14 15:27:03 INFO mapreduce.Job:  map 0% reduce 0%
24/01/14 15:27:28 INFO mapreduce.Job:  map 100% reduce 0%
24/01/14 15:27:28 INFO mapreduce.Job: Job job_1705242747862_0001 completed successfully
24/01/14 15:27:28 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=190038
                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)=1065600
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=22200
                Total vcore-milliseconds taken by all map tasks=22200
                Total megabyte-milliseconds taken by all map tasks=34099200
        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)=210
                CPU time spent (ms)=25870
                Physical memory (bytes) snapshot=612061184
                Virtual memory (bytes) snapshot=3296563200
                Total committed heap usage (bytes)=538443776
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
24/01/14 15:27:28 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 43.217 seconds (11.7224 MB/sec)
24/01/14 15:27:28 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

Screenshot of the imported data:



```
[hadoop@ip-172-31-93-127 ~]$ hadoop fs -ls /user/root/SRC_ATM_TRANS
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2024-01-14 15:27 /user/root/SRC_ATM_TRANS/_SUCCESS
-rw-r--r--   1 hadoop hadoop  531214815 2024-01-14 15:27 /user/root/SRC_ATM_TRANS/part-m-00000
```