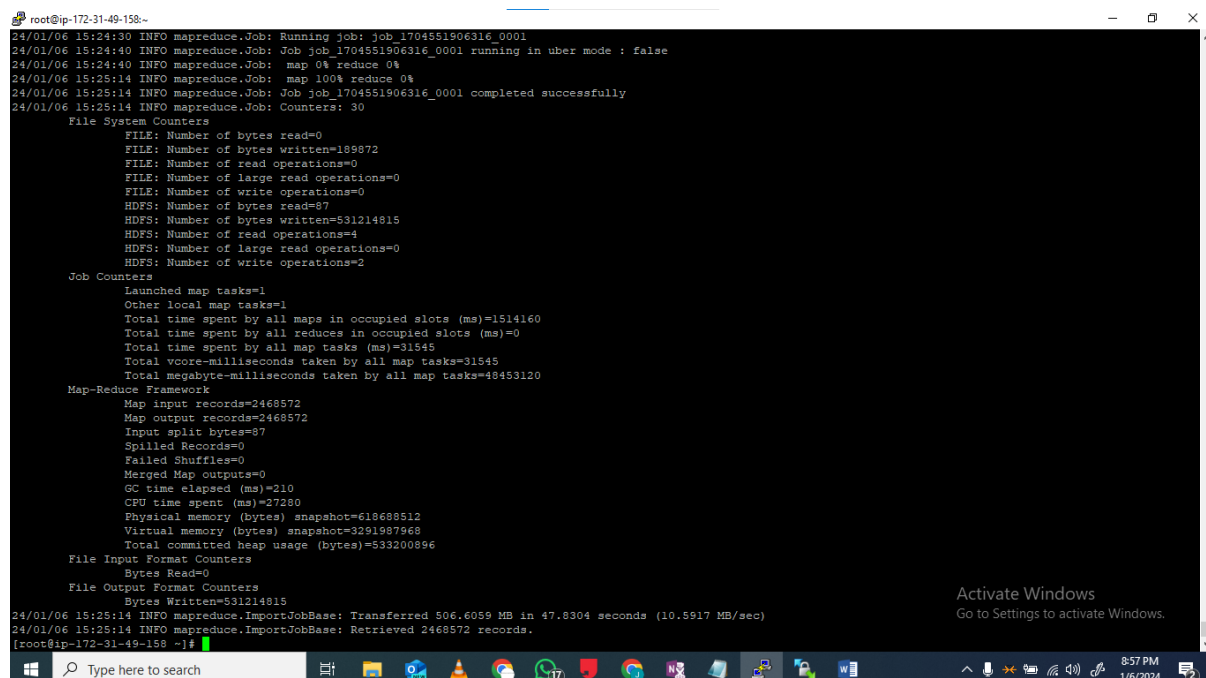


Data Ingestion from the RDS to HDFS using Sqoop

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

```
sqoop import --connect jdbc:mysql://upgraddetest.cyaieic9bmnf.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
```

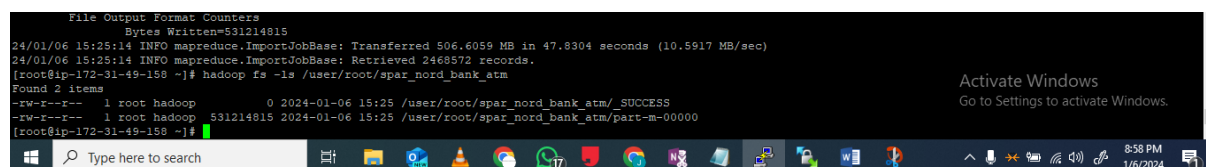


```
root@ip-172-31-49-158:~  
24/01/06 15:24:30 INFO mapreduce.Job: Running job: job_1704551906316_0001  
24/01/06 15:24:40 INFO mapreduce.Job: Job job_1704551906316_0001 running in uber mode : false  
24/01/06 15:24:40 INFO mapreduce.Job: map 0% reduce 0%  
24/01/06 15:25:14 INFO mapreduce.Job: map 100% reduce 0%  
24/01/06 15:25:14 INFO mapreduce.Job: Job job_1704551906316_0001 completed successfully  
24/01/06 15:25:14 INFO mapreduce.Job: Counters: 30  
File System Counters  
  FILE: Number of bytes read=0  
  FILE: Number of bytes written=189872  
  FILE: Number of read operations=0  
  FILE: Number of large read operations=0  
  FILE: Number of write operations=0  
  HDFS: Number of bytes read=37  
  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)=1514160  
  Total time spent by all reduces in occupied slots (ms)=0  
  Total time spent by all map tasks (ms)=31545  
  Total vcore-milliseconds taken by all map tasks=31545  
  Total megabyte-milliseconds taken by all map tasks=48453120  
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)=27280  
  Physical memory (bytes) snapshot=618688512  
  Virtual memory (bytes) snapshot=3291997968  
  Total committed heap usage (bytes)=533200896  
File Input Format Counters  
  Bytes Read=0  
File Output Format Counters  
  Bytes Written=531214815  
24/01/06 15:25:14 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 47.8304 seconds (10.5917 MB/sec)  
24/01/06 15:25:14 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.  
[root@ip-172-31-49-158 ~]#
```

In the screenshot above we can see 2468572 rows have been retrieved.

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

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



```
File Output Format Counters  
  Bytes Written=531214815  
24/01/06 15:25:14 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 47.8304 seconds (10.5917 MB/sec)  
24/01/06 15:25:14 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.  
[root@ip-172-31-49-158 ~]# hadoop fs -ls /user/root/spar_nord_bank_atm  
Found 2 items  
-rw-r--r-- 1 root hadoop 0 2024-01-06 15:25 /user/root/spar_nord_bank_atm/_SUCCESS  
-rw-r--r-- 1 root hadoop 531214815 2024-01-06 15:25 /user/root/spar_nord_bank_atm/part-m-00000  
[root@ip-172-31-49-158 ~]#
```

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/spar_nord_bank_atm/part-m-00000

