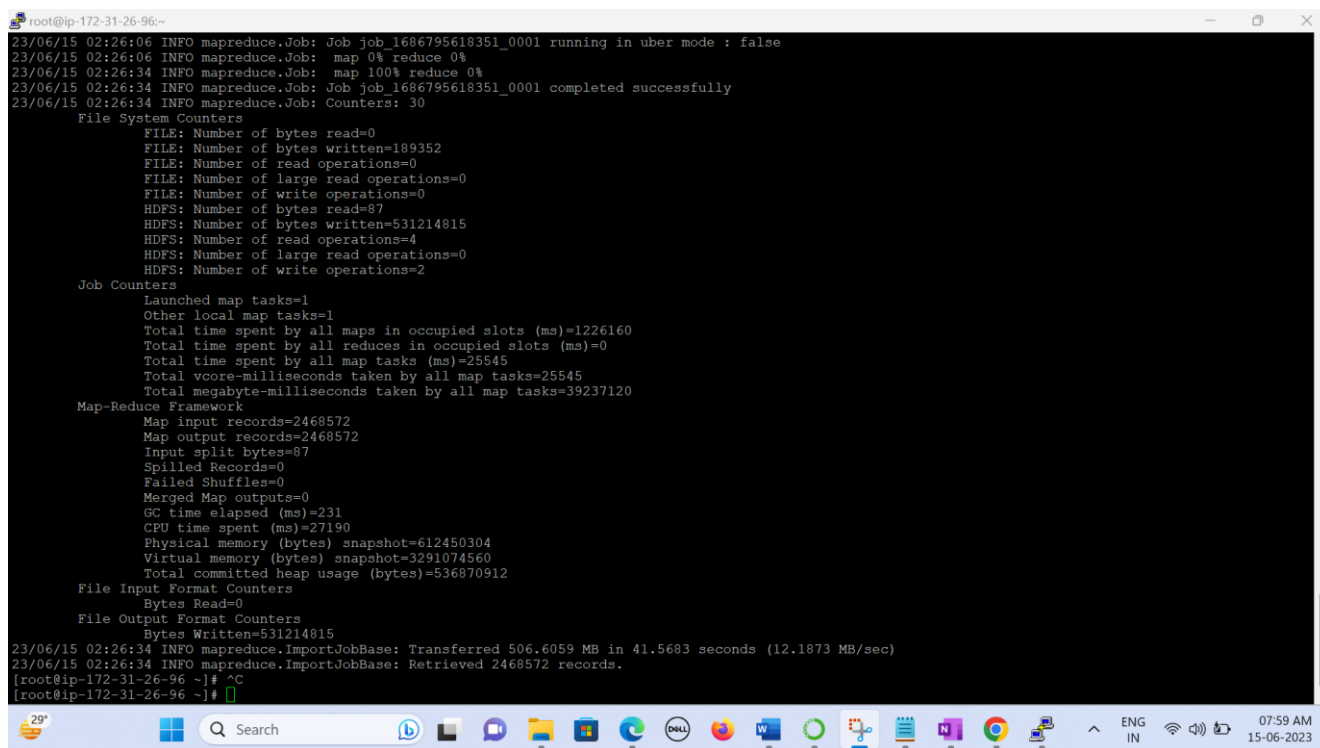# Data Ingestion from the RDS to HDFS using Sqoop

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 used for importing table from RDS to HDFS:

```
[root@ip-172-31-26-96 ~]# sqoop import \
> --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east1.rds.amazonaws.com/testdatabase \
> --table SRC_ATM_TRANS \
> --username student --password STUDENT123 \
> --target-dir /user/root/SRC_ATM_TRANS \
> -m 1
```

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

```
root@ip-172-31-26-96:~
23/06/15 02:26:06 INFO mapreduce.Job: Job job_1686795618351_0001 running in uber mode : false
23/06/15 02:26:06 INFO mapreduce.Job:  map 0% reduce 0%
23/06/15 02:26:34 INFO mapreduce.Job:  map 100% reduce 0%
23/06/15 02:26:34 INFO mapreduce.Job: Job job_1686795618351_0001 completed successfully
23/06/15 02:26:34 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189352
                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)=1226160
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=25545
                Total vcore-milliseconds taken by all map tasks=25545
                Total megabyte-milliseconds taken by all map tasks=39237120
        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)=231
                CPU time spent (ms)=27190
                Physical memory (bytes) snapshot=612450304
                Virtual memory (bytes) snapshot=3291074560
                Total committed heap usage (bytes)=536870912
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
23/06/15 02:26:34 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 41.5683 seconds (12.1873 MB/sec)
23/06/15 02:26:34 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-172-31-26-96 ~]# ^C
[root@ip-172-31-26-96 ~]# []
```

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

<mark>hadoop fs -ls /user/root/SRC_ATM_TRANS</mark>

In the screenshot above we can see two items:
- The first file is the success file, means the MapReduce job was successful.
- The second file 'part-m-00000' is the one that I imported.

Screenshot of the imported data:

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