

Data Ingestion from the RDS to HDFS using Sqoop

```
wget http://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
```

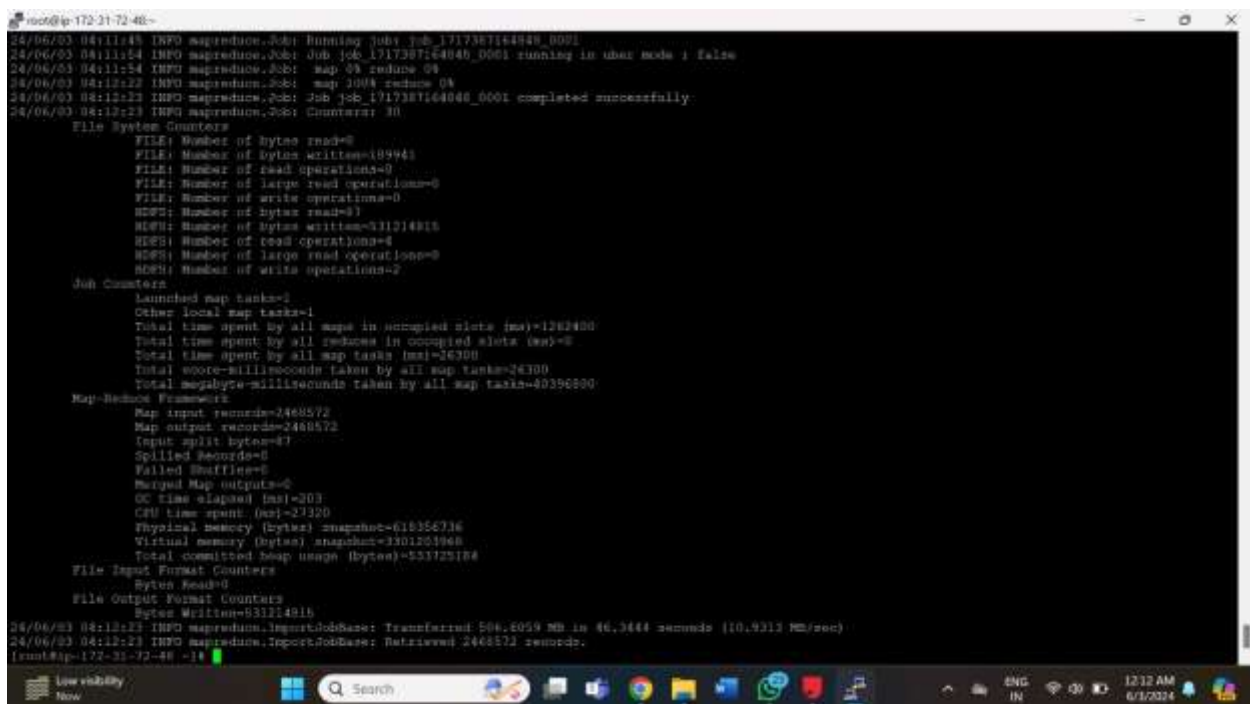
```
tar -xvf mysql-connector-java-8.0.25.tar.gz
```

```
cd mysql-connector-java-8.0.25 /
```

```
sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib
```

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

```
sqoop import \  
> --connect jdbc:mysql://upgradetest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \  
> --table SRC_ATM_TRANS \  
> --username student --password STUDENT123 \  
> --target-dir /user/root/etl_spar_nord_atm \  
> --m 1
```



```
root@ip-172-31-72-46:~# sqoop import \  
24/06/23 04:11:45 INFO mapreduce.Job: Running job: job_1717387164046_0001  
24/06/23 04:11:54 INFO mapreduce.Job: Job job_1717387164046_0001 running in uber mode : false  
24/06/23 04:11:54 INFO mapreduce.Job: map 0% reduce 0%  
24/06/23 04:12:22 INFO mapreduce.Job: map 100% reduce 0%  
24/06/23 04:12:23 INFO mapreduce.Job: Job job_1717387164046_0001 completed successfully  
24/06/23 04:12:23 INFO mapreduce.Job: Counters: 30  
File System Counters  
FILE: Number of bytes read=0  
FILE: Number of bytes written=189944  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=0  
HDFS: Number of bytes written=531214816  
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)=1262400  
Total time spent by all reducers in occupied slots (ms)=0  
Total time spent by all map tasks (ms)=26300  
Total write-milliseconds taken by all map tasks=24300  
Total megabyte-milliseconds taken by all map tasks=40396800  
Map-Reduce Framework  
Map input records=2468572  
Map output records=2468572  
Input split bytes=47  
Spilled Records=0  
Failed Shuffles=0  
Merged Map outputs=0  
GC time elapsed (ms)=203  
CPU time spent (ms)=23320  
Physical memory (bytes) snapshot=616356736  
Virtual memory (bytes) snapshot=330128960  
Total committed heap usage (bytes)=531725184  
File Input Format Counters  
Bytes Read=0  
File Output Format Counters  
Bytes Written=531214816  
24/06/23 04:12:23 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 46.3444 seconds (10.9213 MB/sec)  
24/06/23 04:12:23 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.  
[root@ip-172-31-72-46 ~]#
```

We can see in this screenshot that 506.6059 MB of data was imported. Sqoop command retrieved 2468572 records in total.

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

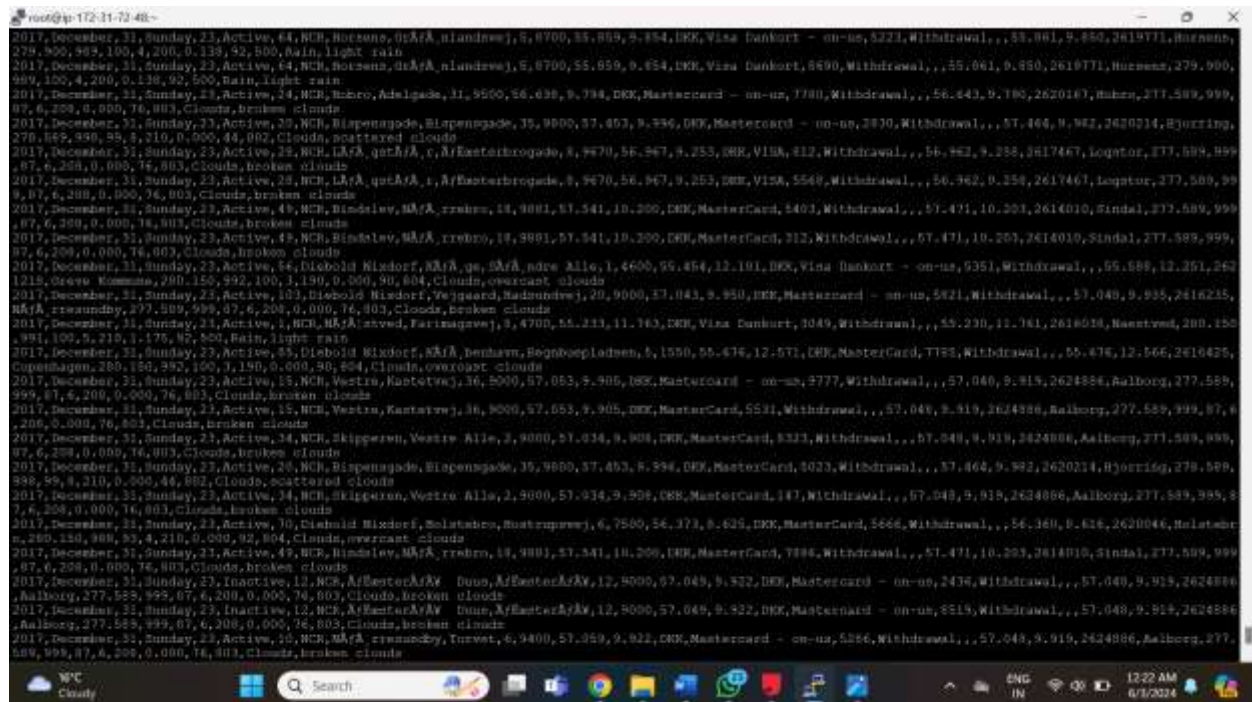
```
hdfs dfs -ls /user/root/etl_spar_nord_atm
```

```
[root@ip-172-31-72-48 ~]# hdfs dfs -ls /user/root/etl_spar_nord_atm
Found 2 items
-rw-r--r-- 1 root hadoop 0 2024-06-03 04:12 /user/root/etl_spar_nord_atm/ SUCCESS
-rw-r--r-- 1 root hadoop 531214815 2024-06-03 04:12 /user/root/etl_spar_nord_atm/part-m-00000
```

Since, I used only one mapper, data is imported in only one file which is “**part-m-0000**”.

Screenshot of the imported data:

1. `hdfs dfs -cat /user/root/etl_spar_nord_atm/part-m-0000`



Screenshot of the portion of the data read from “**part-m-0000**”.

2. `hdfs dfs -cat /user/root/etl_spar_nord_atm/part-m-00000 | wc -l`

```
[root@ip-172-31-72-48 ~]# hdfs dfs -cat /user/root/etl_spar_nord_atm/part-m-00000 | wc -l
2468572
```