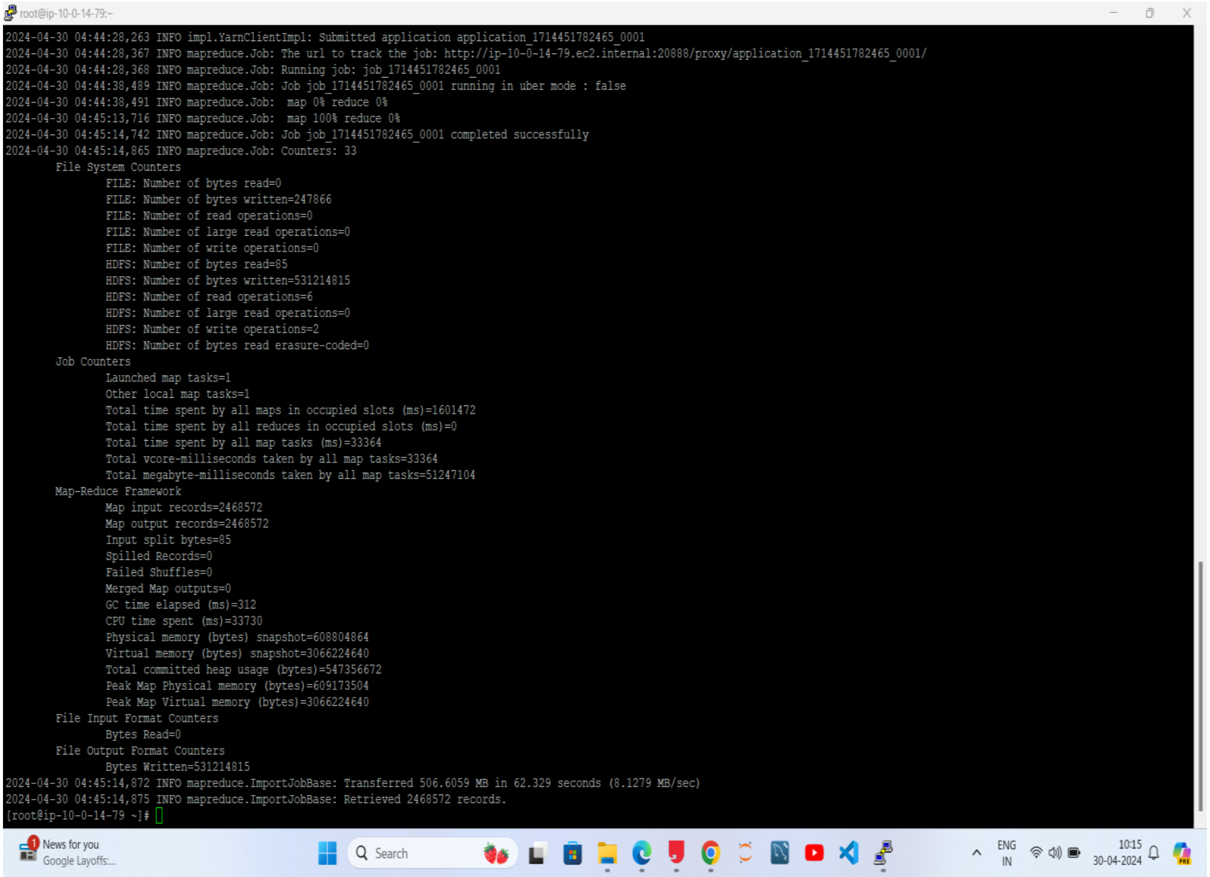
Data Ingestion from the RDS to HDFS using Sqoop

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

```
sqoop import \
--connect jdbc:mysql://upgraddetest.cyaielc9bmnf.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
```

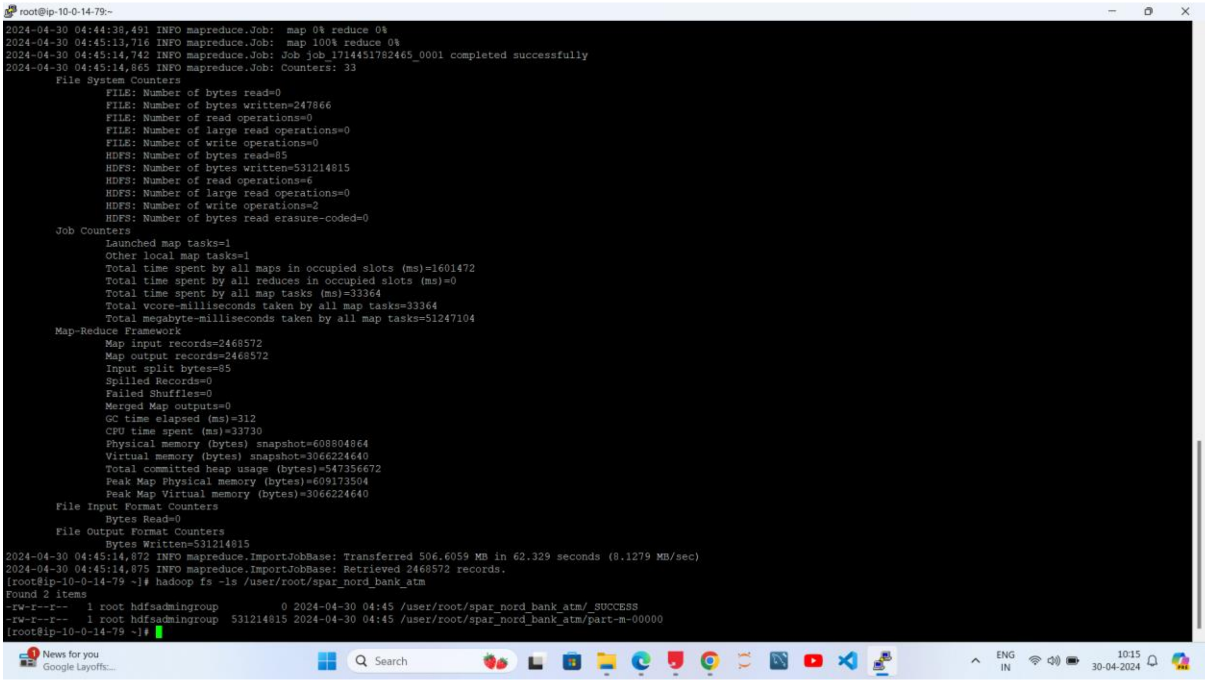


```
2024-04-30 04:44:28,263 INFO impl.YarnClientImpl: Submitted application application_1714451782465_0001
2024-04-30 04:44:28,367 INFO mapreduce.Job: The url to track the job: http://ip-10-0-14-79.ec2.internal:20888/proxy/application_1714451782465_0001/
2024-04-30 04:44:28,368 INFO mapreduce.Job: Running job: job_1714451782465_0001
2024-04-30 04:44:38,489 INFO mapreduce.Job: Job job_1714451782465_0001 running in uber mode : false
2024-04-30 04:44:38,491 INFO mapreduce.Job:  map 0% reduce 0%
2024-04-30 04:45:13,716 INFO mapreduce.Job:  map 100% reduce 0%
2024-04-30 04:45:14,742 INFO mapreduce.Job: Job job_1714451782465_0001 completed successfully
2024-04-30 04:45:14,865 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=247866
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=85
                HDFS: Number of bytes written=531214815
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=1601472
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=33364
                Total vcore-milliseconds taken by all map tasks=33364
                Total megabyte-milliseconds taken by all map tasks=51247104
        Map-Reduce Framework
                Map input records=2468572
                Map output records=2468572
                Input split bytes=85
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=312
                CPU time spent (ms)=33730
                Physical memory (bytes) snapshot=608804864
                Virtual memory (bytes) snapshot=3066224640
                Total committed heap usage (bytes)=547356672
                Peak Map Physical memory (bytes)=609173504
                Peak Map Virtual memory (bytes)=3066224640
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
2024-04-30 04:45:14,872 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 62.329 seconds (8.1279 MB/sec)
2024-04-30 04:45:14,875 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-10-0-14-79 ~]# 
```

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

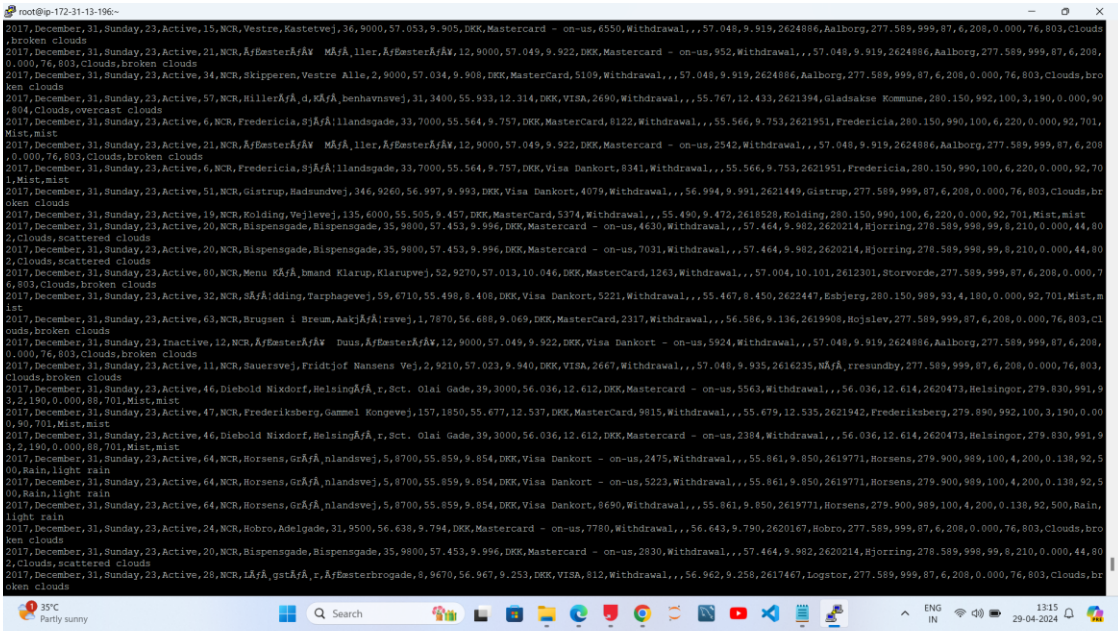


```
2024-04-30 04:44:38,491 INFO mapreduce.Job:  map 0% reduce 0%
2024-04-30 04:45:13,716 INFO mapreduce.Job:  map 100% reduce 0%
2024-04-30 04:45:14,742 INFO mapreduce.Job: Job job_1714451782465_0001 completed successfully
2024-04-30 04:45:14,865 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=247866
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=85
                HDFS: Number of bytes written=531214815
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=1601472
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=33364
                Total vcore-milliseconds taken by all map tasks=33364
                Total megabyte-milliseconds taken by all map tasks=51247104
        Map-Reduce Framework
                Map input records=2468572
                Map output records=2468572
                Input split bytes=85
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=312
                CPU time spent (ms)=33730
                Physical memory (bytes) snapshot=608804864
                Virtual memory (bytes) snapshot=3066224640
                Total committed heap usage (bytes)=547356672
                Peak Map Physical memory (bytes)=609173504
                Peak Map Virtual memory (bytes)=3066224640
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
2024-04-30 04:45:14,872 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 62.329 seconds (8.1279 MB/sec)
2024-04-30 04:45:14,875 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-10-0-14-79 ~]# hadoop fs -ls /user/root/spar_nord_bank_atm
Found 2 items
-rw-r--r--   1 root hdfsadmingroup          0 2024-04-30 04:45 /user/root/spar_nord_bank_atm/_SUCCESS
-rw-r--r--   1 root hdfsadmingroup  531214815 2024-04-30 04:45 /user/root/spar_nord_bank_atm/part-m-00000
[root@ip-10-0-14-79 ~]#
```

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:

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



```
2017,December,31,Sunday,23,Active,15,NCR,Vestre,Kastetvej,36,9000,57.053,9.905,DKK,Mastercard - on-us,6550,Withdrawal,,,57.048,9.919,2624886,Aalborg,277.589,999,87,6,208,0.000,76,803,Clouds
,broken clouds
2017,December,31,Sunday,23,Active,21,NCR,ÃƒÆsterÃƒÂ¥  MÃƒÂ¸ller,ÃƒÆsterÃƒÂ¥,12,9000,57.049,9.922,DKK,Mastercard - on-us,952,Withdrawal,,,57.048,9.919,2624886,Aalborg,277.589,999,87,6,208,
0.000,76,803,Clouds,broken clouds
2017,December,31,Sunday,23,Active,34,NCR,Skipperen,Vestre Alle,2,9000,57.034,9.908,DKK,MasterCard,5109,Withdrawal,,,57.048,9.919,2624886,Aalborg,277.589,999,87,6,208,0.000,76,803,Clouds,bro
ken clouds
2017,December,31,Sunday,23,Active,57,NCR,HillerÃƒÂ¸d,KÃƒÂ¸benhavnsvej,31,3400,55.933,12.314,DKK,VISA,2690,Withdrawal,,,55.767,12.433,2621394,Gladsakse Kommune,280.150,992,100,3,190,0.000,90
,804,Clouds,overcast clouds
2017,December,31,Sunday,23,Active,6,NCR,Fredericia,SjÃƒÂ¦llandsgade,33,7000,55.564,9.757,DKK,MasterCard,8122,Withdrawal,,,55.566,9.753,2621951,Fredericia,280.150,990,100,6,220,0.000,92,701,
Mist,mist
2017,December,31,Sunday,23,Active,21,NCR,ÃƒÆsterÃƒÂ¥  MÃƒÂ¸ller,ÃƒÆsterÃƒÂ¥,12,9000,57.049,9.922,DKK,Mastercard - on-us,2542,Withdrawal,,,57.048,9.919,2624886,Aalborg,277.589,999,87,6,208
,0.000,76,803,Clouds,broken clouds
2017,December,31,Sunday,23,Active,6,NCR,Fredericia,SjÃƒÂ¦llandsgade,33,7000,55.564,9.757,DKK,Visa Dankort,8341,Withdrawal,,,55.566,9.753,2621951,Fredericia,280.150,990,100,6,220,0.000,92,70
1,Mist,mist
2017,December,31,Sunday,23,Active,51,NCR,Gistrup,Hadsundvej,346,9260,56.997,9.993,DKK,Visa Dankort,4079,Withdrawal,,,56.994,9.991,2621449,Gistrup,277.589,999,87,6,208,0.000,76,803,Clouds,br
oken clouds
2017,December,31,Sunday,23,Active,19,NCR,Kolding,Vejlevej,135,6000,55.505,9.457,DKK,MasterCard,5374,Withdrawal,,,55.490,9.472,2618528,Kolding,280.150,990,100,6,220,0.000,92,701,Mist,mist
2017,December,31,Sunday,23,Active,20,NCR,Bispensgade,Bispensgade,35,9800,57.453,9.996,DKK,Mastercard - on-us,4630,Withdrawal,,,57.464,9.982,2620214,Hjorring,278.589,998,99,8,210,0.000,44,80
2,Clouds,scattered clouds
2017,December,31,Sunday,23,Active,20,NCR,Bispensgade,Bispensgade,35,9800,57.453,9.996,DKK,Mastercard - on-us,7031,Withdrawal,,,57.464,9.982,2620214,Hjorring,278.589,998,99,8,210,0.000,44,80
2,Clouds,scattered clouds
2017,December,31,Sunday,23,Active,80,NCR,Menu KÃƒÂ¸bmand Klarup,Klarupvej,52,9270,57.013,10.046,DKK,MasterCard,1263,Withdrawal,,,57.004,10.101,2612301,Storvorde,277.589,999,87,6,208,0.000,7
6,803,Clouds,broken clouds
2017,December,31,Sunday,23,Active,32,NCR,SÃƒÂ¦dding,Tarphagevej,59,6710,55.498,8.408,DKK,Visa Dankort,5221,Withdrawal,,,55.467,8.450,2622447,Esbjerg,280.150,989,93,4,180,0.000,92,701,Mist,m
ist
2017,December,31,Sunday,23,Active,63,NCR,Brugsen i Breum,AakjÃƒÂ¦rsvej,1,7870,56.688,9.069,DKK,MasterCard,2317,Withdrawal,,,56.586,9.136,2619908,Hojslev,277.589,999,87,6,208,0.000,76,803,Cl
ouds,broken clouds
2017,December,31,Sunday,23,Inactive,12,NCR,ÃƒÆsterÃƒÂ¥  Duus,ÃƒÆsterÃƒÂ¥,12,9000,57.049,9.922,DKK,Visa Dankort - on-us,5924,Withdrawal,,,57.048,9.919,2624886,Aalborg,277.589,999,87,6,208,
0.000,76,803,Clouds,broken clouds
2017,December,31,Sunday,23,Active,11,NCR,Sauersvej,Fridtjof Nansens Vej,2,9210,57.023,9.940,DKK,VISA,2667,Withdrawal,,,57.048,9.935,2616235,NÃƒÂ¸rresundby,277.589,999,87,6,208,0.000,76,803,
Clouds,broken clouds
2017,December,31,Sunday,23,Active,46,Diebold Nixdorf,HelsingÃƒÂ¸r,Sct. Olai Gade,39,3000,56.036,12.612,DKK,Mastercard - on-us,5563,Withdrawal,,,56.036,12.614,2620473,Helsingor,279.830,991,9
3,2,190,0.000,88,701,Mist,mist
2017,December,31,Sunday,23,Active,47,NCR,Frederiksberg,Gammel Kongevej,157,1850,55.677,12.537,DKK,MasterCard,9815,Withdrawal,,,55.679,12.535,2621942,Frederiksberg,279.890,992,100,3,190,0.00
0,90,701,Mist,mist
2017,December,31,Sunday,23,Active,46,Diebold Nixdorf,HelsingÃƒÂ¸r,Sct. Olai Gade,39,3000,56.036,12.612,DKK,Mastercard - on-us,2384,Withdrawal,,,56.036,12.614,2620473,Helsingor,279.830,991,9
3,2,190,0.000,88,701,Mist,mist
2017,December,31,Sunday,23,Active,64,NCR,Horsens,GrÃƒÂ¸nlandsvej,5,8700,55.859,9.854,DKK,Visa Dankort - on-us,2475,Withdrawal,,,55.861,9.850,2619771,Horsens,279.900,989,100,4,200,0.138,92,5
00,Rain,light rain
2017,December,31,Sunday,23,Active,64,NCR,Horsens,GrÃƒÂ¸nlandsvej,5,8700,55.859,9.854,DKK,Visa Dankort - on-us,5223,Withdrawal,,,55.861,9.850,2619771,Horsens,279.900,989,100,4,200,0.138,92,5
00,Rain,light rain
2017,December,31,Sunday,23,Active,64,NCR,Horsens,GrÃƒÂ¸nlandsvej,5,8700,55.859,9.854,DKK,Visa Dankort,8690,Withdrawal,,,55.861,9.850,2619771,Horsens,279.900,989,100,4,200,0.138,92,500,Rain,
light rain
2017,December,31,Sunday,23,Active,24,NCR,Hobro,Adelgade,31,9500,56.638,9.794,DKK,Mastercard - on-us,7780,Withdrawal,,,56.643,9.790,2620167,Hobro,277.589,999,87,6,208,0.000,76,803,Clouds,bro
ken clouds
2017,December,31,Sunday,23,Active,20,NCR,Bispensgade,Bispensgade,35,9800,57.453,9.996,DKK,Mastercard - on-us,2830,Withdrawal,,,57.464,9.982,2620214,Hjorring,278.589,998,99,8,210,0.000,44,80
2,Clouds,scattered clouds
2017,December,31,Sunday,23,Active,28,NCR,LÃƒÂ¸gstÃƒÂ¸r,ÃƒÆsterbrogade,8,9670,56.967,9.253,DKK,VISA,812,Withdrawal,,,56.962,9.258,2617467,Logstor,277.589,999,87,6,208,0.000,76,803,Clouds,br
oken clouds
```