Data Ingestion from the RDS to HDFS using Sqoop

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

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:

```
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/SRC_ATM_TRANS \
-m 1
```

In the screenshot below, I can see that as a result of Sqoop Import Job, 2468572 records have been retrieved (same as the checkpoint mentioned in the Validation document):

```
21/10/16   18:16:28 INFO mapreduce.Job:  map 0% reduce 0%
21/10/16   18:17:04 INFO mapreduce.Job:  map 100% reduce 0%
21/10/16   18:17:13 INFO mapreduce.Job: Job job_1628531322983_0001 completed succe
ssfully
21/10/16   18:17:14 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=176686
                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)=41548
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=41548
                Total vcore-milliseconds taken by all map tasks=41548
                Total megabyte-milliseconds taken by all map tasks=42545152
        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)=249
                CPU time spent (ms)=28410
                Physical memory (bytes) snapshot=419827712
                Virtual memory (bytes) snapshot=2802429952
                Total committed heap usage (bytes)=384827392
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
21/10/16   18:17:14 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 64.38
95 seconds (7.8678 MB/sec)
21/10/16   18:17:14 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-10-0-0-166 ~]#
```

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

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

In the screenshot below, I can see that the target directory contains 2 items:

- The first file is the success file, indicating that the MapReduce job was successful.
- The second file 'part-m-00000' is the one with all of the data I imported. Since I used only one mapper in my import command thus the data is in a single file.

```
[root@ip-10-0-0-166 ~]# hadoop fs -ls /user/root/SRC_ATM_TRANS
Found 2 items
-rw-r--r--   3 root supergroup          0 2021/10/16 18:17 /user/root/SRC_ATM_TR
ANS/_SUCCESS
-rw-r--r--   3 root supergroup  531214815 2021/10/16 18:16 /user/root/SRC_ATM_TR
ANS/part-m-00000
[root@ip-10-0-0-166 ~]#
```

When I open the 'part-m-00000' file using the following command, I can see all of the data that has been imported:

```
hadoop fs -cat /user/root/SRC_ATM_TRANS/part-m-00000
```

Screenshot of a portion of the imported data:

```
2017,January,4,Wednesday,10,Active,104,NCR,Intern  ÃƒÆœsterÃƒÂ¥,ÃƒÆœsterÃƒÂ¥,12,
4886,Aalborg,275.079,1008,69,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Inactive,3,NCR,Ikast,RÃƒÂ¥dhusstrÃƒÂ¦det,12,7430,56.
,1001,69,7,350,0.000,0,800,Clear,Sky is Clear
2017,January,4,Wednesday,10,Active,33,NCR,Vadum,Ellehammersvej,43,9430,57.118,9.
008,69,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Inactive,54,NCR,Durup,Torvet,4,7870,56.745,8.949,DKK
,1001,69,7,350,0.000,32,802,Clouds,scattered clouds
2017,January,4,Wednesday,10,Active,38,NCR,Hasseris,Hasserisvej,113,9000,57.044,9
5.079,1008,69,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Active,69,NCR,Taars,Bredgade,91,9830,57.385,10.116,D
69,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Active,31,NCR,Slagelse,Mariendals Alle,29,4200,55.39
6.380,994,74,10,350,0.000,12,801,Clouds,few clouds
2017,January,4,Wednesday,10,Active,62,Diebold Nixdorf,Terndrup,Bymidten,2,9575,5
und,275.079,1008,69,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Active,41,Diebold Nixdorf,Skagen,Sct. Laurentiivej,3
613939,Skagen,272.280,997,86,9,20,0.000,90,600,Snow,light snow
2017,January,4,Wednesday,10,Active,13,NCR,SÃƒÂ¦by,Vestergade,3,9300,57.334,10.51
avn,276.229,1009,100,13,15,0.000,76,803,Clouds,broken clouds
2017,January,4,Wednesday,10,Active,15,NCR,Vestre,Kastetvej,36,9000,57.053,9.905,
9,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Active,48,Diebold Nixdorf,BrÃƒÂ¸nderslev,Algade,4,97
,Vadum,275.079,1008,69,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Active,104,NCR,Intern  ÃƒÆœsterÃƒÂ¥,ÃƒÆœsterÃƒÂ¥,12,
86,Aalborg,275.079,1008,69,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Active,4,NCR,Svogerslev,BrÃƒÂ¸nsager,1,4000,55.634,1
,994,74,10,350,0.235,36,500,Rain,light rain
2017,January,4,Wednesday,10,Active,40,Diebold Nixdorf,Frederikshavn,Danmarksgade
7,2621927,Frederikshavn,276.229,1009,100,13,15,0.000,76,803,Clouds,broken clouds
2017,January,4,Wednesday,10,Active,69,NCR,Taars,Bredgade,91,9830,57.385,10.116,D
08,69,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Active,39,NCR,Svenstrup,GodthÃƒÂ¥bsvej,14,9230,56.97
rup,275.079,1008,69,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Active,49,NCR,Bindslev,NÃƒÂ¸rrebro,18,9881,57.541,10
9,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Active,62,Diebold Nixdorf,Terndrup,Bymidten,2,9575,5
und,275.079,1008,69,11,3,0.000,8,800,Clear,sky is clear
2017,January,4,Wednesday,10,Active,28,NCR,LÃƒÂ¸gstÃƒÂ¸r,ÃƒÆœsterbrogade,8,9670,5
y,56.962,9.258,2617467,Logstor,275.079,1008,69,11,3,0.000,8,800,Clear,sky is cle
```