

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://upgradetest.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
```

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):

```
[hadoop@ip-172-31-81-61 ~]$ 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/spar_nord_bank_atm \  
> -m 1  
Warning: /usr/lib/sqoop/.accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
23/04/09 11:31:40 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift-jdbc/redshift-jdbc42-1.2.37.1061.jar/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.jar/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]  
23/04/09 11:31:40 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
23/04/09 11:31:40 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.  
23/04/09 11:31:40 INFO tool.CodeGenTool: Beginning code generation  
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.  
23/04/09 11:31:41 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1  
23/04/09 11:31:41 INFO manager.MySQLManager: It looks like you are importing from mysql.  
23/04/09 11:31:41 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce  
Note: /tmp/sqoop-hadoop/compile/1df29faalc43b17287fb382c34d95fd/SRC_ATM_TRANS.java uses or overrides a deprecated API.  
Note: Recompile with -Xlint:deprecation for details.  
23/04/09 11:31:44 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/1df29faalc43b17287fb382c34d95fd/SRC_ATM_TRANS.jar  
23/04/09 11:31:44 WARN manager.MySQLManager: This transfer can be faster! Use the --direct  
23/04/09 11:31:44 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.  
23/04/09 11:31:44 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)  
23/04/09 11:31:44 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS  
23/04/09 11:31:45 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps  
23/04/09 11:31:45 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-81-61.ec2.internal/172.31.81.61:8032  
23/04/09 11:31:49 INFO db.DBInputFormat: Using read committed transaction isolation  
23/04/09 11:31:49 INFO mapreduce.JobSubmitter: number of splits:1  
23/04/09 11:31:49 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1681038000113_0001  
23/04/09 11:31:50 INFO impl.YarnClientImpl: Submitted application application_1681038000113_0001  
23/04/09 11:31:50 INFO mapreduce.Job: The url to track the job: http://ip-172-31-81-61.ec2.internal:20888/proxy/application_1681038000113_0001/  
23/04/09 11:31:50 INFO mapreduce.Job: Running job: job_1681038000113_0001  
23/04/09 11:32:01 INFO mapreduce.Job: Job job_1681038000113_0001 running in uber mode : false  
23/04/09 11:32:01 INFO mapreduce.Job: map 0% reduce 0%
```

```

23/04/09 11:32:01 INFO mapreduce.Job: Job job_1681038000113_0001 running in uber mode : false
23/04/09 11:32:01 INFO mapreduce.Job: map 0% reduce 0%
23/04/09 11:32:26 INFO mapreduce.Job: map 100% reduce 0%
23/04/09 11:32:28 INFO mapreduce.Job: Job job_1681038000113_0001 completed successfully
23/04/09 11:32:29 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=189996
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=97
    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)=1111824
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=23163
    Total vcore-milliseconds taken by all map tasks=23163
    Total megabyte-milliseconds taken by all map tasks=35578368
  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)=198
    CPU time spent (ms)=26210
    Physical memory (bytes) snapshot=613900288
    Virtual memory (bytes) snapshot=3286200320
    Total committed heap usage (bytes)=533725184
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=531214815
23/04/09 11:32:29 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 43.263 seconds (11.7099 MB/sec)
23/04/09 11:32:29 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.

```

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 the data I imported. Since I used only one mapper in my import command thus the data is in a single file.

```

[hadoop@ip-172-31-81-61 ~]$ hadoop fs -ls /user/root/spar_nord_bank_atm
Found 2 items
-rw-r--r-- 1 hadoop hadoop 0 2023-04-09 11:32 /user/root/spar_nord_bank_atm/_SUCCESS
-rw-r--r-- 1 hadoop hadoop 531214815 2023-04-09 11:32 /user/root/spar_nord_bank_atm/part-m-00000

```

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:

```
[hadoop@ip-172-31-81-61 ~]$ hadoop fs -cat /user/root/spar nord bank atm/part-m-00000 | more
2017,January,1,Sunday,0,Active,1,NCR,NÅfÅ\stved,Farimagsvej,8,4700,55.233,11.763,DKK,MasterCard,5643,Withdrawal,,,55.230,11.761,2616038,Naestved,281.150,1014,87,7,260,0.215,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,2,NCR,Vejgaard,Hadsundvej,20,9000,57.043,9.950,DKK,MasterCard,1764,Withdrawal,,,57.048,9.935,2616235,NÅfÅ_rresundby,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,2,NCR,Vejgaard,Hadsundvej,20,9000,57.043,9.950,DKK,VISA,1891,Withdrawal,,,57.048,9.935,2616235,NÅfÅ_rresundby,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,3,NCR,Ikast,RÅfÅ\husstrÅfÅ\det,12,7430,56.139,9.154,DKK,VISA,4166,Withdrawal,,,56.139,9.158,2619426,Ikast,281.150,1011,100,6,240,0.000,75,300,Drizzle,light intensity drizzle
2017,January,1,Sunday,0,Active,4,NCR,Svogerslev,BrÅfÅ_nsager,1,4000,55.634,12.018,DKK,MasterCard,5153,Withdrawal,,,55.642,12.080,2614481,Roskilde,280.610,1014,87,7,260,0.000,88,701,Mist,mist
2017,January,1,Sunday,0,Active,5,NCR,Nibe,Torvet,1,9240,56.983,9.639,DKK,MasterCard,3269,Withdrawal,,,56.981,9.639,2616483,Nibe,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,6,NCR,Fredericia,SjÅfÅ\llandsgade,33,7000,55.564,9.757,DKK,MasterCard,887,Withdrawal,,,55.566,9.753,2621951,Fredericia,281.150,1014,93,7,230,0.290,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,7,Diebold Nixdorf,Hjallerup,Hjallerup Centret,18,9320,57.168,10.148,DKK,Mastercard - on-us,4626,Withdrawal,,,57.165,10.146,2620275,Hjallerup,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,8,NCR,GlyngÅfÅ_re,FÅfÅ\rgevej,1,7870,56.762,8.867,DKK,MasterCard,470,Withdrawal,,,56.793,8.853,2615964,Nykobing Mors,281.150,1011,100,6,240,0.000,75,300,Drizzle,light intensity drizzle
2017,January,1,Sunday,0,Active,9,Diebold Nixdorf,Hadsund,Storegade,12,9560,56.716,10.114,DKK,VISA,8473,Withdrawal,,,56.715,10.117,2620952,Hadsund,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,10,NCR,NÅfÅ_rresundby,Torvet,6,9400,57.059,9.922,DKK,Dankort,953,Withdrawal,,,57.048,9.919,2624886,Aalborg,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,11,NCR,Sauersvej,Fridtjof Nansens Vej,2,9210,57.023,9.940,DKK,Visa Dankort,9346,Withdrawal,,,57.048,9.935,2616235,NÅfÅ_rresundby,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,2,NCR,Vejgaard,Hadsundvej,20,9000,57.043,9.950,DKK,Mastercard - on-us,3874,Withdrawal,,,57.048,9.935,2616235,NÅfÅ_rresundby,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,12,NCR,ÅfÅesterÅfÅ\ Duus,ÅfÅesterÅfÅ\,12,9000,57.049,9.922,DKK,Mastercard - on-us,1329,Withdrawal,,,57.048,9.919,2624886,Aalborg,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,13,NCR,SÅfÅ\by,Vestergade,3,9300,57.334,10.515,DKK,Mastercard - on-us,5024,Withdrawal,,,57.441,10.537,2621927,Frederikshavn,281.140,1019,94,12,251,1.275,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,14,NCR,HÅfÅ_rning,NÅfÅ_rreal,12,8362,56.086,10.037,DKK,Visa Dankort - on-us,1133,Withdrawal,,,56.157,10.211,2624652,Århus,281.150,1012,87,5,250,0.000,92,300,Drizzle,light intensity drizzle
2017,January,1,Sunday,0,Active,15,NCR,Vestre,Kastetvej,36,9000,57.053,9.905,DKK,MasterCard,594,Withdrawal,,,57.048,9.919,2624886,Aalborg,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,12,NCR,ÅfÅesterÅfÅ\ Duus,ÅfÅesterÅfÅ\,12,9000,57.049,9.922,DKK,Mastercard - on-us,9570,Withdrawal,,,57.048,9.919,2624886,Aalborg,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,16,NCR,Skive,Adelgade,8,7800,56.567,9.027,DKK,Visa Dankort,6068,Withdrawal,,,56.567,9.033,2613731,Skive,281.150,1012,100,7,240,0.000,75,300,Drizzle,light intensity drizzle
2017,January,1,Sunday,0,Active,17,NCR,Randers,ÅfÅestervold,16,8900,56.462,10.038,DKK,VISA,1630,Withdrawal,,,56.715,10.117,2620952,Hadsund,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,17,NCR,Randers,ÅfÅestervold,16,8900,56.462,10.038,DKK,MasterCard,9948,Withdrawal,,,56.715,10.117,2620952,Hadsund,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,6,NCR,Fredericia,SjÅfÅ\llandsgade,33,7000,55.564,9.757,DKK,MasterCard,4849,Withdrawal,,,55.566,9.753,2621951,Fredericia,281.150,1014,93,7,230,0.290,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,10,NCR,NÅfÅ_rresundby,Torvet,6,9400,57.059,9.922,DKK,MasterCard,4400,Withdrawal,,,57.048,9.919,2624886,Aalborg,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
```