

Data Ingestion from the RDS to HDFS using Sqoop

Before initiating the Sqoop import command, it's necessary to first download and install the MySQL connector. Once our EMR cluster is set up and we've accessed the EMR console, we need to execute the following steps:

1. **sudo -i**
 - Switch to the root user to ensure we have the necessary permissions for the following operations.
2. **wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz**
 - Download the MySQL connector from the provided link. This connector is required for Sqoop to interact with MySQL databases.
3. **tar -xvf mysql-connector-java-8.0.25.tar.gz**
 - Extract the contents of the downloaded tar.gz file.
4. **cd mysql-connector-java-8.0.25/**
 - Navigate into the directory that we created from extracting the tar.gz file. This directory contains the MySQL connector JAR file.
5. **sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/**
 - Copy the MySQL connector JAR file to the Sqoop library directory. This step is essential for Sqoop to use the MySQL connector during the import process.

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

```
sqoop import \  
--connect jdbc:mysql://upgraddetest.cyaie1c9bmnf.us-east-1.rds.amazonaws.com/testdatabase \  
--table SRC_ATM_TRANS \  
--username student -P \  
--target-dir /user/root/ETL/Spar_Nord_Bank_ATM \  
-m 1
```

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

→ Lists the contents of the directory

/user/root/ETL/Spar_Nord_Bank_ATM in the Hadoop Distributed File System (HDFS)

```
[hadoop@ip-172-31-19-74 ~]$ hadoop fs -ls
```

```
/user/root/ETL/Spar_Nord_Bank_ATM
```

```
Found 2 items
```

```
-rw-r--r--    1 root hadoop          0 2023-11-15 15:36
```

```
/user/root/ETL/Spar_Nord_Bank_ATM/_SUCCESS
```

```
-rw-r--r--    1 root hadoop 531214815 2023-11-15 15:36
```

```
/user/root/ETL/Spar_Nord_Bank_ATM/part-m-00000
```

→ displays the first 10 lines of the file **part-m-00000** located in the **/user/root/ETL/Spar_Nord_Bank_ATM** directory on HDFS.

```
[hadoop@ip-172-31-19-74 ~]$ hadoop fs -cat
```

```
/user/root/ETL/Spar_Nord_Bank_ATM/part-m-00000 | head -n 10
```

```
2017,January,1,Sunday,0,Active,1,NCR,NÃfÃ|stved,Farimagsvej,8,4700,55.233,11.763,DKK,MasterCard,5643,Withdrawal,,,55.230,11.761,2616038,Naes tved,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,NÃfÃ,r resundby,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,NÃfÃ,rresund by,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
```

```
2017,January,1,Sunday,0,Inactive,3,NCR,Ikast,RÃfÃ¥dhusstrÃfÃ|det,12,74 30,56.139,9.154,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,BrÃfÃ,nsager,1,4000,55 .634,12.018,DKK,MasterCard,5153,Withdrawal,,,55.642,12.080,2614481,Ros kilde,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,D KK,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,SjÃfÃ|llandsgade,33,70 00,55.564,9.757,DKK,MasterCard,887,Withdrawal,,,55.566,9.753,2621951,F redericia,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,GlyngÃfÃ,re,FÃfÃ|rgevej,1,7870,56 .762,8.867,DKK,MasterCard,470,Withdrawal,,,56.793,8.853,2615964,Nykobi
```

ng 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,Withdraw
al,,,56.715,10.117,2620952,Hadsund,280.640,1020,93,9,250,0.590,92,500,
Rain,light rain
cat: Unable to write to output stream.

Screenshot of the imported data:

➔ The screenshot below shows the output from the Sqoop import command.

```
23/11/15 16:07:51 INFO db.InputFormat: Using read committed transaction isolation
23/11/15 16:07:51 INFO mapreduce.JobSubmitter: number of splits=1
23/11/15 16:07:51 INFO impl.YarnClientImpl: Submitted application job: job_1700057222234_0001
23/11/15 16:07:52 INFO mapreduce.Job: The url to track the job: http://ip-172-31-45-58.ec2.internal:20888/proxy/application_1700057222234_0001/
23/11/15 16:07:52 INFO mapreduce.Job: Running job: job_1700057222234_0001
23/11/15 16:08:01 INFO mapreduce.Job: Job job_1700057222234_0001 running in uber mode : false
23/11/15 16:08:01 INFO mapreduce.Job: map 0% reduce 0%
23/11/15 16:08:28 INFO mapreduce.Job: map 100% reduce 0%
23/11/15 16:08:28 INFO mapreduce.Job: Job job_1700057222234_0001 completed successfully
23/11/15 16:08:28 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=189085
    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)=1159392
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=24154
    Total vcore-milliseconds taken by all map tasks=24154
    Total megabyte-milliseconds taken by all map tasks=37100544
  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)=214
    CPU time spent (ms)=24590
    Physical memory (bytes) snapshot=617836544
    Virtual memory (bytes) snapshot=3289292800
    Total committed heap usage (bytes)=533725184
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=531214815
23/11/15 16:08:28 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 44.1824 seconds (11.4662 MB/sec)
23/11/15 16:08:28 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
ip-172-31-45-58 mysql-connector-java-8.0.25#
```

➔ The screenshot below shows the commands to list the data as required in previous step.

```
[hadoop@ip-172-31-19-74 ~]$ clear
[hadoop@ip-172-31-19-74 ~]$ hadoop fs -ls /user/root/ETL/Spar_Nord_Bank_ATM
Found 2 items
-rw-r--r-- 1 root hadoop 0 2023-11-15 15:36 /user/root/ETL/Spar_Nord_Bank_ATM/_SUCCESS
-rw-r--r-- 1 root hadoop 531214815 2023-11-15 15:36 /user/root/ETL/Spar_Nord_Bank_ATM/part-m-00000
[hadoop@ip-172-31-19-74 ~]$ hadoop fs -cat /user/root/ETL/Spar_Nord_Bank_ATM/part-m-00000 | head -n 10
2017,January,1,Sunday,0,Active,1,NCR,NÅFÅ[stved,Farimagvej,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,NÅFÅHusstrÅFÅdet,12,7430,56.139,9.154,DKK,VISA,4166,Withdrawal,,,56.139,9.150,2619426,Ikast,201.150,1011,100,6,240,0.000,75,300,Drizzle,light intensity driz
zle
2017,January,1,Sunday,0,Active,4,NCR,Svogerslev,BrÅFÅ,nsager,1,4000,55.634,12.018,DKK,MasterCard,5153,Withdrawal,,,55.642,12.000,2614481,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,Fredricia,SjÅFÅIllundsgade,33,7000,55.564,9.757,DKK,MasterCard,687,Withdrawal,,,55.566,9.753,2621951,Fredricia,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,280.640,1020,93,9,250,0.590,92,5
00,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 intens
ity 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
cat: Unable to write to output stream.
[hadoop@ip-172-31-19-74 ~]$
```