Data Ingestion from the RDS to HDFS using Sqoop

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

Setting up MYSQL connector

```
[root@ip-172-31-12-193 ~]# wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
--2024-04-24 08:47:19--  https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
Resolving de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)... 3.5.17.120, 52.216.57.1, 54.231.132.217, ...
Connecting to de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)|3.5.17.120|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4079310 (3.9M) [application/x-gzip]
Saving to: 'mysql-connector-java-8.0.25.tar.gz.1'

100%[================================================================================================================================================>] 4,079,310   --.-K/s   in 0.05s

2024-04-24 08:47:19 (83.3 MB/s) - 'mysql-connector-java-8.0.25.tar.gz.1' saved [4079310/4079310]

[root@ip-172-31-12-193 ~]# tar -xvf mysql-connector-java-8.0.25.tar.gz
mysql-connector-java-8.0.25/
mysql-connector-java-8.0.25/src/
mysql-connector-java-8.0.25/src/build/
mysql-connector-java-8.0.25/src/build/java/
mysql-connector-java-8.0.25/src/build/java/documentation/
mysql-connector-java-8.0.25/src/build/java/instrumentation/
mysql-connector-java-8.0.25/src/build/misc/
mysql-connector-java-8.0.25/src/build/misc/debian.in/
mysql-connector-java-8.0.25/src/build/misc/debian.in/source/
mysql-connector-java-8.0.25/src/demo/
mysql-connector-java-8.0.25/src/demo/java/
mysql-connector-java-8.0.25/src/demo/java/demo/
mysql-connector-java-8.0.25/src/demo/java/demo/x/
mysql-connector-java-8.0.25/src/demo/java/demo/x/devapi/
mysql-connector-java-8.0.25/src/generated/
mysql-connector-java-8.0.25/src/generated/java/
mysql-connector-java-8.0.25/src/generated/java/com/
mysql-connector-java-8.0.25/src/generated/java/com/mysql/
mysql-connector-java-8.0.25/src/generated/java/com/mysql/cj/
```

```
[root@ip-172-31-12-193 mysql-connector-java-8.0.25]# mysql_secure_installation

NOTE: RUNNING ALL PARTS OF THIS SCRIPT IS RECOMMENDED FOR ALL MariaDB
      SERVERS IN PRODUCTION USE!  PLEASE READ EACH STEP CAREFULLY!

In order to log into MariaDB to secure it, we'll need the current
password for the root user.  If you've just installed MariaDB, and
you haven't set the root password yet, the password will be blank,
so you should just press enter here.

Enter current password for root (enter for none):
ERROR 1045 (28000): Access denied for user 'root'@'localhost' (using password: NO)
Enter current password for root (enter for none):
OK, successfully used password, moving on...

Setting the root password ensures that nobody can log into the MariaDB
root user without the proper authorisation.

You already have a root password set, so you can safely answer 'n'.

Change the root password? [Y/n] ^C
Aborting!

Cleaning up...
[root@ip-172-31-12-193 mysql-connector-java-8.0.25]# clear
[root@ip-172-31-12-193 mysql-connector-java-8.0.25]# mysql_secure_installation

NOTE: RUNNING ALL PARTS OF THIS SCRIPT IS RECOMMENDED FOR ALL MariaDB
      SERVERS IN PRODUCTION USE!  PLEASE READ EACH STEP CAREFULLY!

In order to log into MariaDB to secure it, we'll need the current
password for the root user.  If you've just installed MariaDB, and
you haven't set the root password yet, the password will be blank,
so you should just press enter here.

Enter current password for root (enter for none):
OK, successfully used password, moving on...

Setting the root password ensures that nobody can log into the MariaDB
root user without the proper authorisation.

You already have a root password set, so you can safely answer 'n'.

Change the root password? [Y/n] y
New password:
Re-enter new password:
Password updated successfully!
Reloading privilege tables..
 ... Success!

By default, a MariaDB installation has an anonymous user, allowing anyone
```

Sqoop Import command used for importing table from RDS to HDFS

```
sqoop import \
  --connect
jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/tes
tdatabase \
  --table SRC_ATM_TRANS \
  --username student \
  --password STUDENT123 \
  --target-dir /user/root/sn_bank_atm \
  -m 1
```

```
[hadoop@ip-172-31-12-193 ~]$ 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/sn_bank_atm \
> -m 1
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
24/04/24 09:02:54 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/St
```

In the screenshot below we can see 2468572 rows have been retrieved

```
24/04/24 09:03:12 INFO mapreduce.Job: Job job_1713946851004_0003 running in uber mode : false
24/04/24 09:03:12 INFO mapreduce.Job:  map 0% reduce 0%
24/04/24 09:03:44 INFO mapreduce.Job:  map 100% reduce 0%
24/04/24 09:03:44 INFO mapreduce.Job: Job job_1713946851004_0003 completed successfully
24/04/24 09:03:45 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=189981
                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)=1416960
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=29520
                Total vcore-milliseconds taken by all map tasks=29520
                Total megabyte-milliseconds taken by all map tasks=45342720
        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)=270
                CPU time spent (ms)=33950
                Physical memory (bytes) snapshot=608428032
                Virtual memory (bytes) snapshot=3295764480
                Total committed heap usage (bytes)=538968064
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
24/04/24 09:03:45 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 45.7984 seconds (11.0616 MB/sec)
24/04/24 09:03:45 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
```

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

hadoop fs -ls /user/root/sn_bank_atm

In the screenshot below 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' since only one mapper was used in the import command, as a result the data is in a single file.

```
[hadoop@ip-172-31-12-193 ~]$ hadoop fs -ls /user/root/
Found 1 items
drwxr-xr-x   - hadoop hadoop          0 2024-04-24 09:06 /user/root/sn_bank_atm
[hadoop@ip-172-31-12-193 ~]$ hadoop fs -ls /user/root/sn_bank_atm
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2024-04-24 09:06 /user/root/sn_bank_atm/_SUCCESS
-rw-r--r--   1 hadoop hadoop  531214815 2024-04-24 09:06 /user/root/sn_bank_atm/part-m-00000
[hadoop@ip-172-31-12-193 ~]$
```

Screenshot of the imported data:

```
hadoop@ip-172-31-12-193:~
[hadoop@ip-172-31-12-193 ~]$ hadoop fs -cat /user/root/sn_bank_atm/part-m-00000 | head -n 20
2017,January,1,Sunday,0,Active,1,NCR,NÃƒÂ¦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,NÃƒÂ¸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,NÃƒÂ¸rresundby,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,3,NCR,Ikast,RÃƒÂ¥dhusstrÃƒÂ¦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,BrÃƒÂ¸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,SjÃƒÂ¦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,GlyngÃƒÂ¸re,FÃƒÂ¦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,NÃƒÂ¸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,NÃƒÂ¸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,NÃƒÂ¸rresundby,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,12,NCR,ÃƒÂ˜sterÃƒÂ¥  Duus,ÃƒÂ˜sterÃƒÂ¥,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,SÃƒÂ¦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,HÃƒÂ¸rning,NÃƒÂ¸rrealle,12,8362,56.086,10.037,DKK,Visa Dankort - on-us,1133,Withdrawal,,,56.157,10.211,2624652,Arhus,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,ÃƒÂ˜sterÃƒÂ¥  Duus,ÃƒÂ˜sterÃƒÂ¥,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
```