

Data Ingestion from the RDS to HDFS using Sqoop

We Need MySQL connector jar file in our Sqoop lib directory before running Sqoop jobs

- Switch to root user
sudo -i
- Download the tar file on to your cluster
wget <https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz>
- Extract the file
tar -xvf mysql-connector-java-8.0.25.tar.gz
- Copy the file and move it to /usr/lib/sqoop/lib/
sudo cp mysql-connector-java-8.0.25/mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/

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

```
sqoop import \  
--connect jdbc:mysql://upgradtest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase\  
--table SRC_ATM_TRANS \  
--username student --password STUDENT123 \  
--target-dir /user/root/spar_bank_data/ \  
--m 1
```

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

- To check the files present in destination directory.
hdfs dfs -ls /user/root/spar_bank_data
- To check the first three records
hdfs dfs -cat /user/root/spar_bank_data/part-m-00000 | head -n 3
- To check the total count of the records
hdfs dfs -cat /user/root/spar_bank_data/part-m-00000 | wc -l

Screenshot of the imported data:

Successful loading of data with 2 files present in the destination. Success file and the part file.

```

Map output records=2468572
Input split bytes=85
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=277
CPU time spent (ms)=26610
Physical memory (bytes) snapshot=620314624
Virtual memory (bytes) snapshot=3072421888
Total committed heap usage (bytes)=542638080
Peak Map Physical memory (bytes)=625438720
Peak Map Virtual memory (bytes)=3090731008

File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=531214815
2024-03-28 06:57:45,207 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 42.5637 seconds (11.9023 MB/sec)
2024-03-28 06:57:45,209 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-172-31-48-227 ~]#
[root@ip-172-31-48-227 ~]#

FILE: Number of bytes written=247065
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)=1228320
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=25590
  Total vcore-milliseconds taken by all map tasks=25590
  Total megabyte-milliseconds taken by all map tasks=39306240

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)=277
  CPU time spent (ms)=26610
  Physical memory (bytes) snapshot=620314624
  Virtual memory (bytes) snapshot=3072421888
  Total committed heap usage (bytes)=542638080
  Peak Map Physical memory (bytes)=625438720
  Peak Map Virtual memory (bytes)=3090731008

File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=531214815
2024-03-28 06:57:45,207 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 42.5637 seconds (11.9023 MB/sec)
2024-03-28 06:57:45,209 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-172-31-48-227 ~]#
[root@ip-172-31-48-227 ~]#
[root@ip-172-31-48-227 ~]# hdfs dfs -ls /user/root/spar_bank_data
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/ava/emr/emfs/lib/slf4j-log4j12-1.7.12.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]
Found 2 items
-rw-r--r-- 1 root hdfsadmin group 0 2024-03-28 06:57 /user/root/spar_bank_data/_SUCCESS
-rw-r--r-- 1 root hdfsadmin group 531214815 2024-03-28 06:57 /user/root/spar_bank_data/part-m-000000
[root@ip-172-31-48-227 ~]# hdfs dfs

```

Viewing the first 3 records in the file

```

SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
cat: '/user/root/spar_bank_data': Is a directory
[root@ip-172-31-48-227 ~]# hdfs dfs -ls /user/root/spar_bank_data/
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/ava/emr/emfs/lib/slf4j-log4j12-1.7.12.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]
Found 2 items
-rw-r--r-- 1 root hdfsadmin group 0 2024-03-28 06:57 /user/root/spar_bank_data/_SUCCESS
-rw-r--r-- 1 root hdfsadmin group 531214815 2024-03-28 06:57 /user/root/spar_bank_data/part-m-000000
[root@ip-172-31-48-227 ~]# hdfs dfs -cat /user/root/spar_bank_data/part-m-000000 | head -n 3
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/ava/emr/emfs/lib/slf4j-log4j12-1.7.12.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]
2017, January, 1, Sunday, 0, Inactive, 1, NCR, NAF, Istved, Farinagsvej, 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, NAF, 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, NAF, rresundby, 280.640, 1020, 93, 9, 250, 0.590, 92, 500, Rain, light rain
cat: Unable to write to output stream.
[root@ip-172-31-48-227 ~]#

```

Checking the total record count in the part file.

```

[root@ip-172-31-48-227 ~]#
[root@ip-172-31-48-227 ~]# hdfs dfs -cat /user/root/spar_bank_data/part-m-000000 | wc -l
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/ava/emr/emfs/lib/slf4j-log4j12-1.7.12.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]
2468572
[root@ip-172-31-48-227 ~]#

```