

Data Ingestion from the RDS to HDFS using Sqoop

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

```
sqoop import \  
--connect jdbc:mysql://upgraddetest.cyaiehc9bmnf.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
```

```
[hadoop@ip-172-31-3-218:~  
login as: hadoop  
Authenticating with public key "anmol_key_pair"  
  
A newer release of "Amazon Linux" is available.  
Version 2023.4.20240611:  
Version 2023.5.20240624:  
Version 2023.5.20240701:  
Run "/usr/bin/dnf check-release-update" for full release and version update info
```

```

#
~\##### Amazon Linux 2023
~~\_#####\
~~\_###|
~~\_#\ https://aws.amazon.com/linux/amazon-linux-2023
~~V~'--->
~~~~
~~~-./--/
~/m/'-/
Last login: Tue Jul 2 13:05:09 2024
```

```
EEEEEEEEEEEEEEEEEEEE MMMMMMM          MMMMMMMM RRRRRRRRRRRRRR
E::::::::::::::::::E M::::::::M          M::::::::M R::::::::::::R
EE::::::::EEEEEEEE::E M::::::::M          M::::::::M R::::RRRRRR::::R
E::::E      EEEEE M::::::::M          M::::::::M RR::::R          R::::R
E::::E      M::::::::M:M          M::::::::M R:::R          R::::R
E::::EEEEEEEEEE M::::M M::M M::M M::::M R::RRRRRR::::R
E::::::::::::E M::::M M::M:M:M M::::M R::::::::::::RR
E::::EEEEEEEEEE M::::M M::::M M::::M R::RRRRRR::::R
E::::E      M::::M M::M M::::M R:::R          R::::R
E::::E      EEEEE M::::M      MMM M::::M R:::R          R::::R
EE::::::::EEEEEEEE::E M::::M          M::::::::M R:::R          R::::R
E::::::::::::E M::::M          M::::M RR::::R          R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM          MMMMMMMM RRRRRRR          RRRRRR
```

```
[hadoop@ip-172-31-3-218 ~]$ sqoop import --connect jdbc:mysql://upgraddetest.cya
ielc9bmnf.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/../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.
2024-07-02 13:19:07,673 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-07-02 13:19:07,750 WARN tool.BaseSqoopTool: Setting your password on the co
mmmand-line is insecure. Consider using -P instead.
2024-07-02 13:19:07,910 INFO manager.MySQLManager: Preparing to use a MySQL stre
aming resultset.
```

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

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

```
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)=286
CPU time spent (ms)=37060
Physical memory (bytes) snapshot=395886592
Virtual memory (bytes) snapshot=3208441856
Total committed heap usage (bytes)=264241152
Peak Map Physical memory (bytes)=395886592
Peak Map Virtual memory (bytes)=3208441856
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=531214815
2024-07-02 13:20:13,441 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 55.4841 seconds (9.1307 MB/sec)
2024-07-02 13:20:13,444 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[hadoop@ip-172-31-3-218 ~]$
```

Screenshot of the imported data:

A portion of data read from part-m-00000 file

```
root@ip-10-0-7-69:~
24/07/08 12:13:47 INFO mapreduce.Job: The url to track the job: http://ip-10-0-7-69.ec2.internal:20888/proxy/application_1720439247647_0001/
24/07/08 12:13:47 INFO mapreduce.Job: Running job: job_1720439247647_0001
24/07/08 12:13:56 INFO mapreduce.Job: Job job_1720439247647_0001 running in uber mode : false
24/07/08 12:13:56 INFO mapreduce.Job: map 0% reduce 0%
24/07/08 12:14:22 INFO mapreduce.Job: map 100% reduce 0%
24/07/08 12:14:22 INFO mapreduce.Job: Job job_1720439247647_0001 completed successfully
24/07/08 12:14:22 INFO mapreduce.Job: Counters: 30
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=189363
  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)=1145376
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=23862
  Total vcore-milliseconds taken by all map tasks=23862
  Total megabyte-milliseconds taken by all map tasks=36652032
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)=27439
  Physical memory (bytes) snapshot=620478464
  Virtual memory (bytes) snapshot=3301695488
  Total committed heap usage (bytes)=538443776
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=531214815
24/07/08 12:14:22 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 44.2406 seconds (11.4512 MB/sec)
24/07/08 12:14:22 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-10-0-7-69 ~]# hadoop fs -ls /user/root/spar_nord_bank_atm
Found 2 items
-rw-r--r-- 1 root hadoop 0 2024-07-08 12:14 /user/root/spar_nord_bank_atm/ SUCCESS
-rw-r--r-- 1 root hadoop 531214815 2024-07-08 12:14 /user/root/spar_nord_bank_atm/part-m-00000
[root@ip-10-0-7-69 ~]#
```