

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://upgradtest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \  
--username student \  
--password STUDENT123 \  
--table SRC_ATM_TRANS \  
--target-dir /user/root/SRC_ATM_TRANS \  
--null-string '\N' \  
--null-non-string '\N' \  
--num-mappers 1 \  
--fetch-size 10000
```

```
hadoop@ip-172-31-54-106:~/mysql-connector-java-8.0.25$ sqoop import \  
> --connect jdbc:mysql://upgradtest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \  
> --username student \  
> --password STUDENT123 \  
> --table SRC_ATM_TRANS \  
> --target-dir /user/root/SRC_ATM_TRANS \  
> --null-string '\N' \  
> --null-non-string '\N' \  
> --num-mappers 1 \  
> --fetch-size 1000  
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 $HCATALOG_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.  
2024-09-01 08:30:46,961 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7  
2024-09-01 08:30:47,001 WARN tool.BaseSqoopTool: Setting your password on the co  
mmand-line is insecure. Consider using -P instead.  
2024-09-01 08:30:47,116 INFO manager.MySQLManager: Argument '--fetch-size 1000'  
will probably get ignored by MySQL JDBC driver.  
2024-09-01 08:30:47,116 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 SP  
T and manual loading of the driver class is generally unnecessary.  
2024-09-01 08:30:47,479 INFO manager.SqlManager: Executing SQL statement: SELECT  
t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1  
2024-09-01 08:30:47,527 INFO manager.SqlManager: Executing SQL statement: SELECT  
t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1  
2024-09-01 08:30:47,558 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/  
lib/hadoop-mapreduce  
2024-09-01 08:30:49,538 ERROR orm.CompilationManager: Could not rename /tmp/sqoo  
p-hadoop/compile/5916ad6d30d8a0ad99f1e3e7037ec49/SRC_ATM_TRANS.java to /home/ha  
doop/mysql-connector-java-8.0.25/./SRC_ATM_TRANS.java. Error: Destination '/home  
/hadoop/mysql-connector-java-8.0.25/./SRC_ATM_TRANS.java' already exists  
2024-09-01 08:30:49,539 INFO orm.CompilationManager: Writing jar file: /tmp/sqoo  
p-hadoop/compile/5916ad6d30d8a0ad99f1e3e7037ec49/SRC_ATM_TRANS.jar  
2024-09-01 08:30:49,565 WARN manager.MySQLManager: It looks like you are importi  
ng from mysql.  
2024-09-01 08:30:49,565 WARN manager.MySQLManager: This transfer can be faster!  
Use the --direct  
2024-09-01 08:30:49,565 WARN manager.MySQLManager: option to exercise a MySQL-sp  
ecific fast path.  
2024-09-01 08:30:49,566 INFO manager.MySQLManager: Setting zero DATETIME behavio  
r to convertToNull (mysql)  
2024-09-01 08:30:49,574 INFO mapreduce.ImportJobBase: Beginning import of SRC AT  
M_TRANS  
2024-09-01 08:30:49,862 INFO Configuration.deprecation: mapred.jar is deprecate  
d. Instead, use mapreduce.job.jar  
2024-09-01 08:30:50,582 INFO Configuration.deprecation: mapred.map.tasks is depr  
ecated. Instead, use mapreduce.job.maps  
2024-09-01 08:30:50,962 INFO client.DefaultNoHARMFaloverProxyProvider: Connecti  
ng to ResourceManager at ip-172-31-54-106.ec2.internal/172.31.54.106:8032  
2024-09-01 08:30:51,343 INFO client.AHSProxy: Connecting to Application History
```

```
hadoop@ip-172-31-54-106:~/mysql-connector-java-8.0.25
2024-09-01 08:30:57,370 INFO mapreduce.JobSubmitter: number of splits:1
2024-09-01 08:30:57,564 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1725173192783_0001
2024-09-01 08:30:57,564 INFO mapreduce.JobSubmitter: Securing with tokens: []
2024-09-01 08:30:57,827 INFO conf.Configuration: resource-types.xml not found
2024-09-01 08:30:57,828 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-09-01 08:30:58,354 INFO impl.YarnClientImpl: Submitted application application_1725173192783_0001
2024-09-01 08:30:58,441 INFO mapreduce.Job: The url to track the job: http://ip-172-31-54-106.ec2.internal:20888/proxy/application_1725173192783_0001/
2024-09-01 08:30:58,442 INFO mapreduce.Job: Running job: job_1725173192783_0001
2024-09-01 08:31:00,642 INFO mapreduce.Job: Job job_1725173192783_0001 running in uber mode : false
2024-09-01 08:31:00,643 INFO mapreduce.Job: map 0% reduce 0%
2024-09-01 08:31:36,840 INFO mapreduce.Job: map 100% reduce 0%
2024-09-01 08:31:40,863 INFO mapreduce.Job: Job job_1725173192783_0001 completed successfully
2024-09-01 08:31:40,962 INFO mapreduce.Job: Counters: 33
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=298007
    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)=1241520
    Total time spent by all reducers in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=25865
    Total vcore-milliseconds taken by all map tasks=25865
    Total megabyte-milliseconds taken by all map tasks=39728640
  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)=269
    CPU time spent (ms)=26360
    Physical memory (bytes) snapshot=632877056
    Virtual memory (bytes) snapshot=3096727552
    Total committed heap usage (bytes)=566755328
    Peak Map Physical memory (bytes)=633942016
    Peak Map Virtual memory (bytes)=3105837056
  File Input Format Counters
    Bytes Read 0
  File Output Format Counters
    Bytes Written=531214815
2024-09-01 08:31:40,967 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 50.3732 seconds (10.0571 MB/sec)
2024-09-01 08:31:40,969 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[hadoop@ip-172-31-54-106 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-54-106 mysql-connector-java-8.0.25]$
```

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

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

```
hadoop fs -ls /user/root/SRC_ATM_TRANS
hdfs dfs -mv /user/root/SRC_ATM_TRANS/part-m-00000
/user/root/SRC_ATM_TRANS/src_atm_trans.csv
```

```
hadoop@ip-172-31-63-108:~
[hadoop@ip-172-31-63-108 ~]$ hadoop fs -ls /user/root/SRC_ATM_TRANS
Found 2 items
-rw-r--r-- 1 hadoop hdfsadmin group 0 2024-09-04 05:31 /user/root/SRC_ATM_TRANS/_SUCCESS
-rw-r--r-- 1 hadoop hdfsadmin group 531214815 2024-09-04 05:31 /user/root/SRC_ATM_TRANS/src_atm_trans.csv
[hadoop@ip-172-31-63-108 ~]$
```

In the screenshot above 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

Command used to see the first 10 rows of the imported data

hadoop fs -cat /user/root/SRC_ATM_TRANS/src_atm_trans.csv | head -1

```
[hadoop@ip-172-31-63-108 ~]$ hadoop fs -cat /user/root/SRC_ATM_TRANS/src_atm_trans.csv | 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,RÅfÅvðhusstrÅfÅ: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,BrÅfÅnsager,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
cat: Unable to write to output stream.
[hadoop@ip-172-31-63-108 ~]$
```