

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.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
--table SRC_ATM_TRANS \
--username student --password STUDENT123 \
--target-dir /user/root/SRC_ATM_TRANS \
-m 1
```

```
root@ip-172-30-2-59 ~]# sqoop import --connect jdbc:mysql://upgraddetest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase --table SRC_ATM_TRANS --username student --p
assword STUDENT123 --target-dir /user/root/SRC_ATM_TRANS -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.
23/10/26 15:58:04 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/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.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]
23/10/26 15:58:04 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/10/26 15:58:04 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/10/26 15:58:04 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 SPI and manual
loading of the driver class is generally unnecessary.
23/10/26 15:58:04 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1
23/10/26 15:58:04 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1
23/10/26 15:58:04 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/9932d0788000da75a7cee54341968127/SRC_ATM_TRANS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/10/26 15:58:06 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/9932d0788000da75a7cee54341968127/SRC_ATM_TRANS.jar
23/10/26 15:58:06 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/10/26 15:58:06 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/10/26 15:58:06 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/10/26 15:58:06 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/10/26 15:58:06 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS
23/10/26 15:58:06 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/10/26 15:58:07 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/10/26 15:58:07 INFO client.RMProxy: Connecting to ResourceManager at ip-172-30-2-59.ec2.internal:20888/proxy/application_1698335703231_0001/
23/10/26 15:58:20 INFO db.DBInputFormat: Using read committed transaction isolation
23/10/26 15:58:21 INFO mapreduce.JobSubmitter: number of splits:1
23/10/26 15:58:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1698335703231_0001
23/10/26 15:58:22 INFO impl.YarnClientImpl: Submitted application application_1698335703231_0001
23/10/26 15:58:22 INFO mapreduce.Job: The url to track the job: http://ip-172-30-2-59.ec2.internal:20888/proxy/application_1698335703231_0001/
23/10/26 15:58:22 INFO mapreduce.Job: Running job: job_1698335703231_0001
23/10/26 15:58:29 INFO mapreduce.Job: Job job_1698335703231_0001 running in uber mode : false
23/10/26 15:58:29 INFO mapreduce.Job: map 0% reduce 0%
23/10/26 15:58:57 INFO mapreduce.Job: map 100% reduce 0%
23/10/26 15:58:58 INFO mapreduce.Job: Job job_1698335703231_0001 completed successfully
23/10/26 15:58:58 INFO mapreduce.Job: Counters: 30
File System Counters
FILE: Number of bytes read=0
FILE: Number of bytes written=189370
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=521314816
```

Screenshot: As can be seen, in total **2468572** records are retrieved.

```

root@ip-172-30-2-59:~$
23/10/26 15:58:04 INFO orm.CompilationManager: HADOOP MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/9932d0788000da75a7cee54341968127/SRC_ATM_TRANS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/10/26 15:58:06 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/9932d0788000da75a7cee54341968127/SRC_ATM_TRANS.jar
23/10/26 15:58:06 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/10/26 15:58:06 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/10/26 15:58:06 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/10/26 15:58:06 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/10/26 15:58:06 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS
23/10/26 15:58:06 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/10/26 15:58:07 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/10/26 15:58:07 INFO client.RMProxy: Connecting to ResourceManager at ip-172-30-2-59.ec2.internal/172.30.2.59:8032
23/10/26 15:58:20 INFO db.DBInputFormat: Using read committed transaction isolation
23/10/26 15:58:21 INFO mapreduce.JobSubmitter: number of splits=1
23/10/26 15:58:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1698335703231_0001
23/10/26 15:58:22 INFO impl.YarnClientImpl: Submitted application application_1698335703231_0001
23/10/26 15:58:22 INFO mapreduce.Job: The url to track the job: http://ip-172-30-2-59.ec2.internal:20888/proxy/application_1698335703231_0001/
23/10/26 15:58:22 INFO mapreduce.Job: Running job: job_1698335703231_0001
23/10/26 15:58:29 INFO mapreduce.Job: Job job_1698335703231_0001 running in uber mode : false
23/10/26 15:58:29 INFO mapreduce.Job: map 0% reduce 0%
23/10/26 15:58:57 INFO mapreduce.Job: map 100% reduce 0%
23/10/26 15:58:58 INFO mapreduce.Job: Job job_1698335703231_0001 completed successfully
23/10/26 15:58:58 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=189370
    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)=2474016
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=25771
    Total vcore-milliseconds taken by all map tasks=25771
    Total megabyte-milliseconds taken by all map tasks=79168512
  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)=208
    CPU time spent (ms)=20160
    Physical memory (bytes) snapshot=1027846144
    Virtual memory (bytes) snapshot=4625690624
    Total committed heap usage (bytes)=954204160
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=531214815
23/10/26 15:58:58 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 51.5989 seconds (9.8181 MB/sec)
23/10/26 15:58:58 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-172-30-2-59 ~]#

```

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

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

Screenshot of the imported data:

```

23/10/26 15:58:58 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-172-30-2-59 ~]# hadoop fs -ls /user/root/SRC_ATM_TRANS
Found 2 items
-rw-r--r-- 1 root hadoop 0 2023-10-26 15:58 /user/root/SRC_ATM_TRANS/_SUCCESS
-rw-r--r-- 1 root hadoop 531214815 2023-10-26 15:58 /user/root/SRC_ATM_TRANS/part-m-00000
[root@ip-172-30-2-59 ~]#

```