

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://upgradtest1.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
```

```

(base) [root@ip-10-0-0-135:~]# sqoop import \
> --connect jdbc:mysql://upgradtest1.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
20/11/24 15:33:13 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
20/11/24 15:33:13 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/11/24 15:33:13 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
20/11/24 15:33:13 INFO tool.CodeGenTool: Beginning code generation
20/11/24 15:33:14 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
20/11/24 15:33:14 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
20/11/24 15:33:14 INFO orm.CompilationManager: HADOOP MAPRED HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/96b37900ac5e5d063ae56d219da2bc8c/SRC_ATM_TRANS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
20/11/24 15:33:17 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/96b37900ac5e5d063ae56d219da2bc8c/SRC_ATM_TRANS.jar
20/11/24 15:33:17 WARN manager.MySQLManager: It looks like you are importing from mysql.
20/11/24 15:33:17 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
20/11/24 15:33:17 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
20/11/24 15:33:17 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
20/11/24 15:33:17 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS
20/11/24 15:33:17 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
20/11/24 15:33:18 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
20/11/24 15:33:18 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-0-135.ec2.internal/10.0.0.135:8032
20/11/24 15:33:25 INFO db.DBInputFormat: Using read committed transaction isolation
20/11/24 15:33:27 INFO mapreduce.JobSubmitter: number of splits:1
20/11/24 15:33:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1606230705591_0001
20/11/24 15:33:27 INFO impl.YarnClientImpl: Submitted application application_1606230705591_0001
20/11/24 15:33:27 INFO mapreduce.Job: The url to track the job: http://ip-10-0-0-135.ec2.internal:8088/proxy/application_1606230705591_0001/
20/11/24 15:33:27 INFO mapreduce.Job: Running job: job_1606230705591_0001
20/11/24 15:33:38 INFO mapreduce.Job: Job job_1606230705591_0001 running in uber mode : false
20/11/24 15:33:38 INFO mapreduce.Job: map 0% reduce 0%
20/11/24 15:34:11 INFO mapreduce.Job: map 100% reduce 0%
20/11/24 15:34:12 INFO mapreduce.Job: Job job_1606230705591_0001 completed successfully
20/11/24 15:34:12 INFO mapreduce.Job: Counters: 30
File System Counters

```

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

Hadoop fs -ls

```

(base) [root@ip-10-0-0-135:~]# hadoop fs -ls
Found 8 items
drwxr-xr-x - root supergroup 0 2020-11-18 13:48 .sparkStaging
drwx----- - root supergroup 0 2020-11-24 15:34 .staging
-rw-r--r-- 3 root supergroup 123765458 2020-11-22 09:14 Police_Department_Incident_Reports__2018_to_Present.csv
drwxr-xr-x - root supergroup 0 2020-11-24 15:34 SRC_ATM_TRANS
-rw-r--r-- 3 root supergroup 6971 2020-11-15 12:35 flights_data
-rw-r--r-- 3 root supergroup 44529893 2020-11-15 12:36 online_data
-rw-r--r-- 3 root supergroup 0 2020-11-19 16:11 upgrad.txt
drwxr-xr-x - root supergroup 0 2020-11-19 16:00 upgrad_folder
(base) [root@ip-10-0-0-135:~]#

```

Hadoop fs -ls /user/root/SRC_ATM_TRANS

```
root@ip-10-0-0-135:~
login as: ec2-user
Authenticating with public key "imported-openssh-key"
Last login: Wed Nov 25 15:52:57 2020 from 49.35.120.124
[ec2-user@ip-10-0-0-135 ~]$ sudo -i
(base) [root@ip-10-0-0-135 ~]# hadoop fs -ls
Found 8 items
drwxr-xr-x - root supergroup 0 2020-11-18 13:48 .sparkStaging
drwx----- - root supergroup 0 2020-11-24 15:34 .staging
-rw-r--r-- 3 root supergroup 123765458 2020-11-22 09:14 Police_Department_Inc
ident_Reports_2018_to_Present.csv
drwxr-xr-x - root supergroup 0 2020-11-24 15:34 SRC_ATM_TRANS
-rw-r--r-- 3 root supergroup 6971 2020-11-15 12:35 flights_data
-rw-r--r-- 3 root supergroup 44529893 2020-11-15 12:36 online_data
-rw-r--r-- 3 root supergroup 0 2020-11-19 16:11 upgrad.txt
drwxr-xr-x - root supergroup 0 2020-11-19 16:00 upgrad_folder
(base) [root@ip-10-0-0-135 ~]# hadoop fs -ls /user/root/SRC_ATM_TRANS
Found 2 items
-rw-r--r-- 3 root supergroup 0 2020-11-24 15:34 /user/root/SRC_ATM_TR
ANS/ SUCCESS
-rw-r--r-- 3 root supergroup 522660639 2020-11-24 15:34 /user/root/SRC_ATM_TR
ANS/part-m-00000
(base) [root@ip-10-0-0-135 ~]#
```

Screenshot of the imported data:

```
root@ip-10-0-0-135:~
20/11/24 15:33:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1606230705591_0001
20/11/24 15:33:27 INFO impl.YarnClientImpl: Submitted application application_1606230705591_0001
20/11/24 15:33:27 INFO mapreduce.Job: The url to track the job: http://ip-10-0-0-135.ec2.internal:8088/proxy/application_1
20/11/24 15:33:27 INFO mapreduce.Job: Running job: job_1606230705591_0001
20/11/24 15:33:38 INFO mapreduce.Job: Job job_1606230705591_0001 running in uber mode : false
20/11/24 15:33:38 INFO mapreduce.Job: map 0% reduce 0%
20/11/24 15:34:11 INFO mapreduce.Job: map 100% reduce 0%
20/11/24 15:34:12 INFO mapreduce.Job: Job job_1606230705591_0001 completed successfully
20/11/24 15:34:12 INFO mapreduce.Job: Counters: 30
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=176683
  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=522660639
  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)=30934
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=30934
  Total vcore-milliseconds taken by all map tasks=30934
  Total megabyte-milliseconds taken by all map tasks=31676416
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)=234
  CPU time spent (ms)=32140
  Physical memory (bytes) snapshot=417554432
  Virtual memory (bytes) snapshot=2801614848
  Total committed heap usage (bytes)=388497408
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=522660639
20/11/24 15:34:12 INFO mapreduce.ImportJobBase: Transferred 498.448 MB in 54.1614 seconds (9.203 MB/sec)
20/11/24 15:34:12 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
```

