

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 \
--table SRC_ATM_TRANS \
--username student --password STUDENT123 \
--target-dir /user/root/Atm_data \
-m 1
```

We have imported the data from RDS to HDFS using sqoop import, where the jdbc:mysql://upgradtest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase is the connects string to connect to RDS table, source table name is SRC_ATM_TRANS , target directory on hdfs is SRC_ATM_TRANS and number of mappers used is 1. It will take default field delimiter as ','.

Screenshot of the imported data:

Sqoop command submission:

```
root@ip-10-0-0-96:~#
login as: ec2-user
Authenticating with public key "imported-openssh-key"
Last login: Sat Jan  9 12:38:14 2021 from 49.36.121.235
[ec2-user@ip-10-0-0-96 ~]$ sudo -i
[root@ip-10-0-0-96 ~]# hdfs dfs -ls /user/root/
ls/user/root/: Unknown command
[root@ip-10-0-0-96 ~]# hdfs dfs -ls /user/root/
Found 4 items
drwxr-xr-x  - root supergroup          0 2021-01-04 13:33 /user/root/.sparkStaging
drwx----- - root supergroup          0 2020-12-05 05:19 /user/root/.staging
-rw-r--r--  3 root supergroup        6971 2020-12-03 04:57 /user/root/flights_data
-rw-r--r--  3 root supergroup    44529893 2020-12-03 04:58 /user/root/online_data
[root@ip-10-0-0-96 ~]# 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/Atm_data \
> -m 1
21/01/09 13:06:14 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
21/01/09 13:06:14 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/01/09 13:06:14 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/01/09 13:06:14 INFO tool.CodeGenTool: Beginning code generation
21/01/09 13:06:15 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
21/01/09 13:06:15 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
21/01/09 13:06:15 INFO orm.CompilationManager: HADOOP MAPRED HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/0d8b16cd06bd1b0eb7bec8b718873e7c/SRC_ATM_TRANS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/01/09 13:06:19 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/0d8b16cd06bd1b0eb7bec8b718873e7c/SRC_ATM_TRANS.jar
21/01/09 13:06:19 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/01/09 13:06:19 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/01/09 13:06:19 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/01/09 13:06:19 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/01/09 13:06:19 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS
21/01/09 13:06:19 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/01/09 13:06:20 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
21/01/09 13:06:20 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-0-96.ec2.internal/10.0.0.96:8032
21/01/09 13:06:26 INFO db.DBInputFormat: Using read committed transaction isolation
21/01/09 13:06:26 INFO mapreduce.JobSubmitter: number of splits:1
21/01/09 13:06:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1610195913603_0001
21/01/09 13:06:27 INFO impl.YarnClientImpl: Submitted application application_1610195913603_0001
21/01/09 13:06:27 INFO mapreduce.Job: The url to track the job: http://ip-10-0-0-96.ec2.internal:8088/proxy/application_1610195913603_0001/
21/01/09 13:06:27 INFO mapreduce.Job: Running job: job_1610195913603_0001
21/01/09 13:06:37 INFO mapreduce.Job: Job job_1610195913603_0001 running in uber mode : false
21/01/09 13:06:37 INFO mapreduce.Job:  map 0% reduce 0%
```

Imported Data: 2468572 rows imported

```
21/01/09 13:06:19 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/01/09 13:06:19 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/01/09 13:06:19 INFO mapreduce.ImportJobBase: Beginning import of SRC.ATM_TRANS
21/01/09 13:06:19 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/01/09 13:06:20 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
21/01/09 13:06:20 INFO client.RMPProxy: Connecting to ResourceManager at ip-10-0-0-96.ec2.internal/10.0.0.96:8032
21/01/09 13:06:26 INFO db.DBInputFormat: Using read committed transaction isolation
21/01/09 13:06:26 INFO mapreduce.JobSubmitter: number of splits:1
21/01/09 13:06:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1610195913603_0001
21/01/09 13:06:27 INFO Impl.YarnClientImpl: Submitted application application_1610195913603_0001
21/01/09 13:06:27 INFO mapreduce.Job: The url to track the job: http://ip-10-0-0-96.ec2.internal:8088/proxy/application_1610195913603_0001/
21/01/09 13:06:27 INFO mapreduce.Job: Running job: job_1610195913603_0001
21/01/09 13:06:37 INFO mapreduce.Job: Job job_1610195913603_0001 running in uber mode : false
21/01/09 13:06:37 INFO mapreduce.Job: map 0% reduce 0%
21/01/09 13:07:07 INFO mapreduce.Job: map 100% reduce 0%
21/01/09 13:07:09 INFO mapreduce.Job: Job job_1610195913603_0001 completed successfully
21/01/09 13:07:09 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=176278
    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)=28181
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=28181
    Total vcore-milliseconds taken by all map tasks=28181
    Total megabyte-milliseconds taken by all map tasks=28857344
  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)=207
    CPU time spent (ms)=29080
    Physical memory (bytes) snapshot=445526016
    Virtual memory (bytes) snapshot=2841772032
    Total committed heap usage (bytes)=386924544
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=531214815
21/01/09 13:07:09 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 48.7176 seconds (10.3988 MB/sec)
21/01/09 13:07:09 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-10-0-0-96 ~]#
```

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

hadoop fs -ls /user/root/Atm_data/

```
[root@ip-10-0-0-96 ~]# hadoop fs -ls /user/root/Atm_data/
Found 2 items
-rw-r--r--  3 root supergroup          0 2021-01-09 13:07 /user/root/Atm_data/_SUCCESS
-rw-r--r--  3 root supergroup 531214815 2021-01-09 13:07 /user/root/Atm_data/part-m-000000
[root@ip-10-0-0-96 ~]#
```

Viewing the file content:

hdfs dfs -cat /user/root/Atm_data/part-m-000000 |head -10

```
[root@ig-10-0-0-96 ~]# hdfs dfs -cat /user/root/Atm data/part-m-00000 |head -10
2017,January,1,Sunday,0,Active,1,NCR,NÅfÅ|stved,Farimagsvej,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,RÅfÅ|HusstrÅfÅ|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,BrÅfÅ|nsager,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
```