

Data Ingestion from the RDS to HDFS using Sqoop

Have mysql connector jar in place as a prerequisite for running sqoop:

```
sudo -i
```

```
wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
```

```
tar -xvf mysql-connector-java-8.0.25.tar.gz
```

```
sudo cp mysql-connector-java-8.0.25/mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/
```

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

```
sqoop import --connect jdbc:mysql://upgraddetest.cyaieic9bmnf.us-east-1.rds.amazonaws.com:3306/testdatabase --username student --password STUDENT123 --table SRC_ATM_TRANS --target-dir /user/hadoop/ETL/ATMDATA -m 1
```

Command to see the list of imported data in HDFS:

```
hadoop fs -ls /user/hadoop/ETL/ATMDATA/
```

count of data records:

```
hadoop fs -cat /user/hadoop/ETL/ATMDATA/part-m-00000 | wc -l
```

top 3 records:

```
hadoop fs -cat /user/hadoop/ETL/ATMDATA/part-m-00000 | head -n 3
```

Screenshot of the imported data:

```
root@ip-172-31-35-173:~  
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)=1121376  
  Total time spent by all reduces in occupied slots (ms)=0  
  Total time spent by all map tasks (ms)=23362  
  Total vcore-milliseconds taken by all map tasks=23362  
  Total megabyte-milliseconds taken by all map tasks=35884032  
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)=190  
  CPU time spent (ms)=27690  
  Physical memory (bytes) snapshot=614600704  
  Virtual memory (bytes) snapshot=3303415808  
  Total committed heap usage (bytes)=534773760  
File Input Format Counters  
  Bytes Read=0  
File Output Format Counters  
  Bytes Written=531214815  
23/09/04 04:51:49 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 44.623 seconds (11.353 MB/sec)  
23/09/04 04:51:49 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.  
[root@ip-172-31-35-173 ~]#
```

```
root@ip-172-31-35-173:~  
[root@ip-172-31-35-173 ~]# hadoop fs -ls /user/hadoop/ETL/ATMDATA/  
Found 2 items  
-rw-r--r-- 1 root hadoop 0 2023-09-04 04:51 /user/hadoop/ETL/ATMDATA/_SUCCESS  
-rw-r--r-- 1 root hadoop 531214815 2023-09-04 04:51 /user/hadoop/ETL/ATMDATA/part-m-00000  
[root@ip-172-31-35-173 ~]# hadoop fs -count /user/hadoop/ETL/ATMDATA/part-m-00000  
0 1 531214815 /user/hadoop/ETL/ATMDATA/part-m-00000  
[root@ip-172-31-35-173 ~]# hadoop fs -cat /user/hadoop/ETL/ATMDATA/part-m-00000 | head -n 3  
2017,January,1,Sunday,0,Active,1,NCR,NÅfÅ|stved,Farinagsvej,8,4700,55.233,11.763,DKK,MasterCard,5643,Withdrawal,,55.230,11.761,26160  
38,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  
cat: Unable to write to output stream.  
[root@ip-172-31-35-173 ~]# hadoop fs -cat /user/hadoop/ETL/ATMDATA/part-m-00000 | wc -l  
2468572  
[root@ip-172-31-35-173 ~]#
```