Data Ingestion from the RDS to HDFS using Sqoop

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

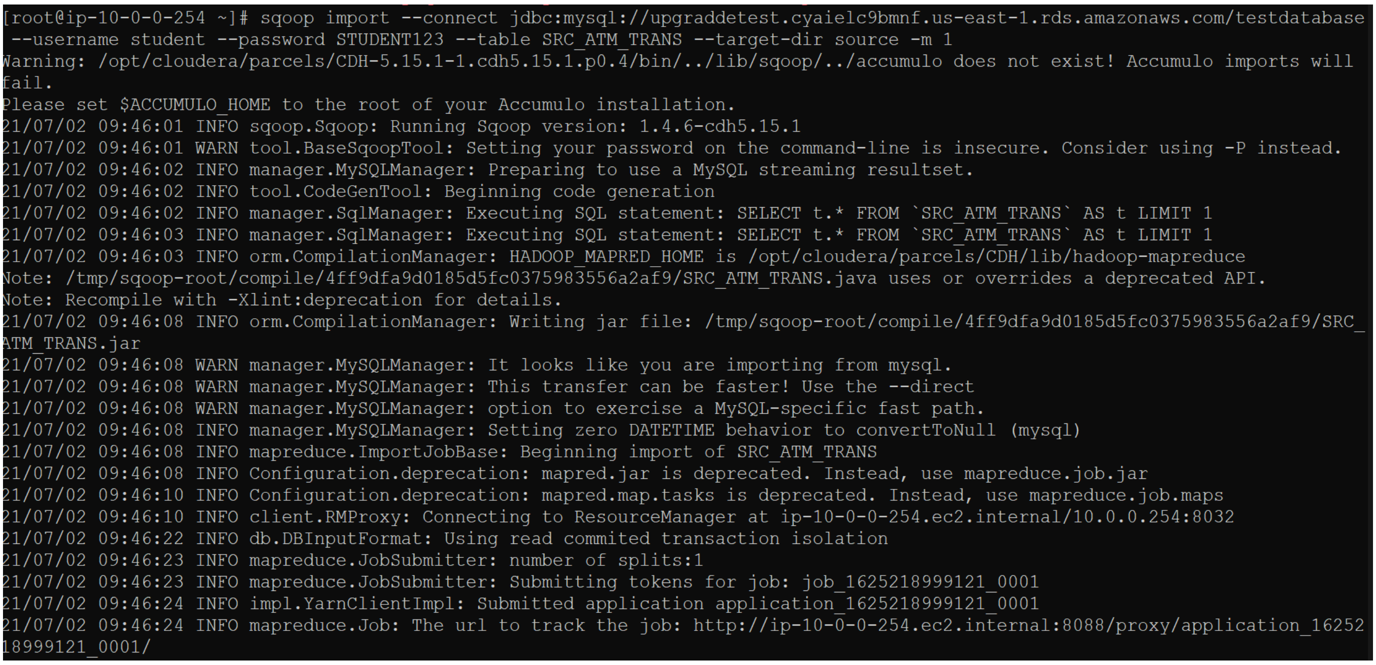
sqoop import --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase --username student --password STUDENT123 --table SRC_ATM_TRANS --target-dir source -m 1

Comment: Used the authentication details provide in the problem statement to import data from the RDS to HDFS on ec2 instance. “source” is the target directory I’ve choosen.

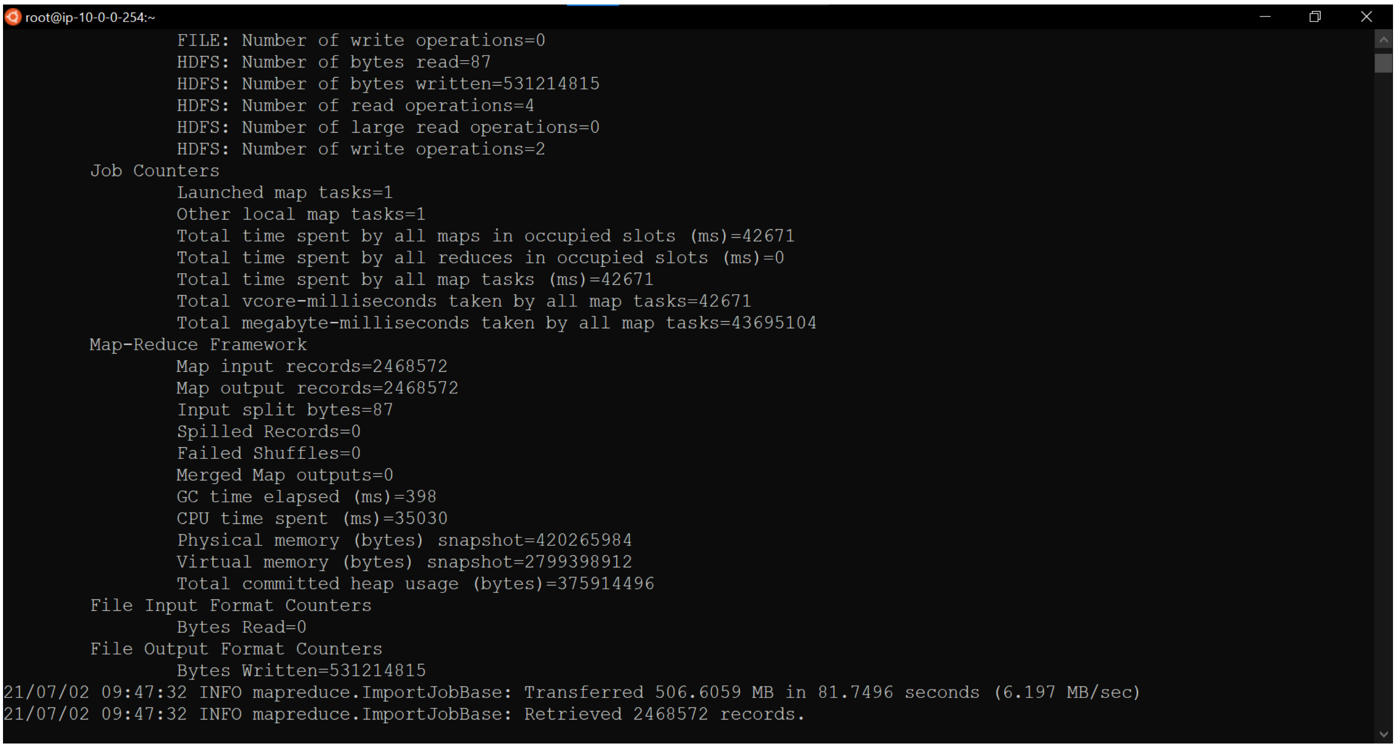
Command used to see the list of imported data in HDFS:

hadoop fs -ls source/

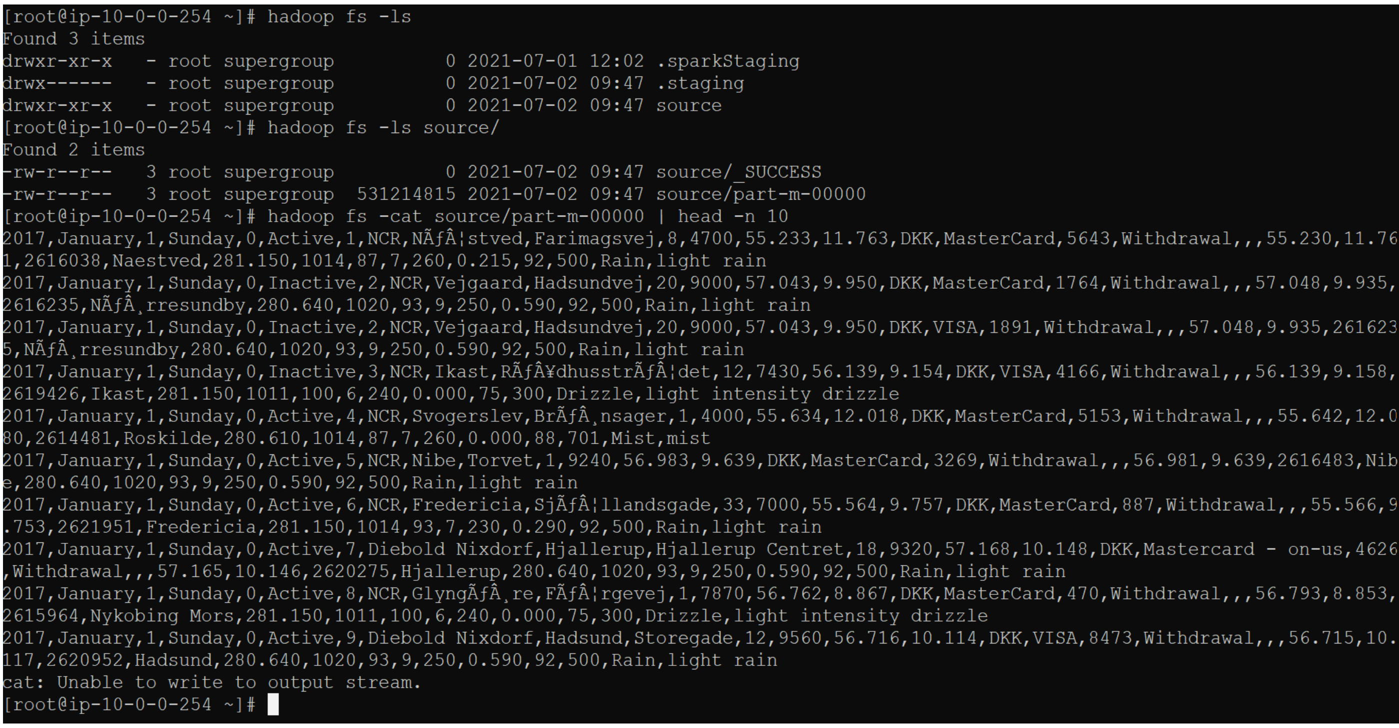
Screenshots :



```
[root@ip-10-0-0-254 ~]# sqoop import --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase
 --username student --password STUDENT123 --table SRC_ATM_TRANS --target-dir source -m 1
Warning: /opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will
fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
21/07/02 09:46:01 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
21/07/02 09:46:01 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/07/02 09:46:02 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/07/02 09:46:02 INFO tool.CodeGenTool: Beginning code generation
21/07/02 09:46:02 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
21/07/02 09:46:03 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
21/07/02 09:46:03 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/4ff9dfa9d0185d5fc0375983556a2af9/SRC_ATM_TRANS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/07/02 09:46:08 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/4ff9dfa9d0185d5fc0375983556a2af9/SRC_
ATM_TRANS.jar
21/07/02 09:46:08 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/07/02 09:46:08 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/07/02 09:46:08 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/07/02 09:46:08 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/07/02 09:46:08 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS
21/07/02 09:46:08 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/07/02 09:46:10 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
21/07/02 09:46:10 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-0-254.ec2.internal/10.0.0.254:8032
21/07/02 09:46:22 INFO db.DBInputFormat: Using read commited transaction isolation
21/07/02 09:46:23 INFO mapreduce.JobSubmitter: number of splits:1
21/07/02 09:46:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1625218999121_0001
21/07/02 09:46:24 INFO impl.YarnClientImpl: Submitted application application_1625218999121_0001
21/07/02 09:46:24 INFO mapreduce.Job: The url to track the job: http://ip-10-0-0-254.ec2.internal:8088/proxy/application_16252
18999121_0001/
```



```
                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)=42671
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=42671
                Total vcore-milliseconds taken by all map tasks=42671
                Total megabyte-milliseconds taken by all map tasks=43695104
        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)=398
                CPU time spent (ms)=35030
                Physical memory (bytes) snapshot=420265984
                Virtual memory (bytes) snapshot=2799398912
                Total committed heap usage (bytes)=375914496
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
21/07/02 09:47:32 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 81.7496 seconds (6.197 MB/sec)
21/07/02 09:47:32 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
```



```
[root@ip-10-0-0-254 ~]# hadoop fs -ls
Found 3 items
drwxr-xr-x   - root supergroup          0 2021-07-01 12:02 .sparkStaging
drwx------   - root supergroup          0 2021-07-02 09:47 .staging
drwxr-xr-x   - root supergroup          0 2021-07-02 09:47 source
[root@ip-10-0-0-254 ~]# hadoop fs -ls source/
Found 2 items
-rw-r--r--   3 root supergroup          0 2021-07-02 09:47 source/_SUCCESS
-rw-r--r--   3 root supergroup  531214815 2021-07-02 09:47 source/part-m-00000
[root@ip-10-0-0-254 ~]# hadoop fs -cat source/part-m-00000 | head -n 10
2017,January,1,Sunday,0,Active,1,NCR,NÃƒÂ¦stved,Farimagsvej,8,4700,55.233,11.763,DKK,MasterCard,5643,Withdrawal,,,55.230,11.76
1,2616038,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,NÃƒÂ¸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,,,57.048,9.935,261623
5,NÃƒÂ¸rresundby,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,3,NCR,Ikast,RÃƒÂ¥dhusstrÃƒÂ¦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,BrÃƒÂ¸nsager,1,4000,55.634,12.018,DKK,MasterCard,5153,Withdrawal,,,55.642,12.0
80,2614481,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,Nib
e,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,6,NCR,Fredericia,SjÃƒÂ¦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,GlyngÃƒÂ¸re,FÃƒÂ¦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.
[root@ip-10-0-0-254 ~]# 
```