Data Ingestion from the RDS to HDFS using Sqoop

Steps performed while importing the table from RDS to HDFS:

1. Checking the presence of the output directory and remove it, because this might raise an exception if the directory already exists by chance.
2. Creating a sqoop job to import the data. It contains following parameters:
   a. Connection URL to the remove MySQL server with 'testdatabase' and for 'SRC_ATM_TRANS' table.
   b. Target directory is provided as '**/user/root/ETL_Project/bank_data_import**'.
   c. Parameter are provided for field and line separation, so that after creating the files, the data should be comma separated.
   d. Compression used to compress file as SnappCode compression, to compress the file and avoid the size issues with huge data.
   e. Parameters are provided to identify the null string and null non-string parameters.
   f. The number of Mappers are provided as 1 to avoid the multiple requests for the remove database server. This is mainly because, the job can be done easily by compromising little time (probably a minute or two, not more).
3. Executing the sqoop job is created based on the above parameters.
4. Checking the importing data in HDFS and verify the file formats (Compressed Snappy File.

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

```
hadoop fs -rm -r /user/root/ETL_Project/bank_data_import
```

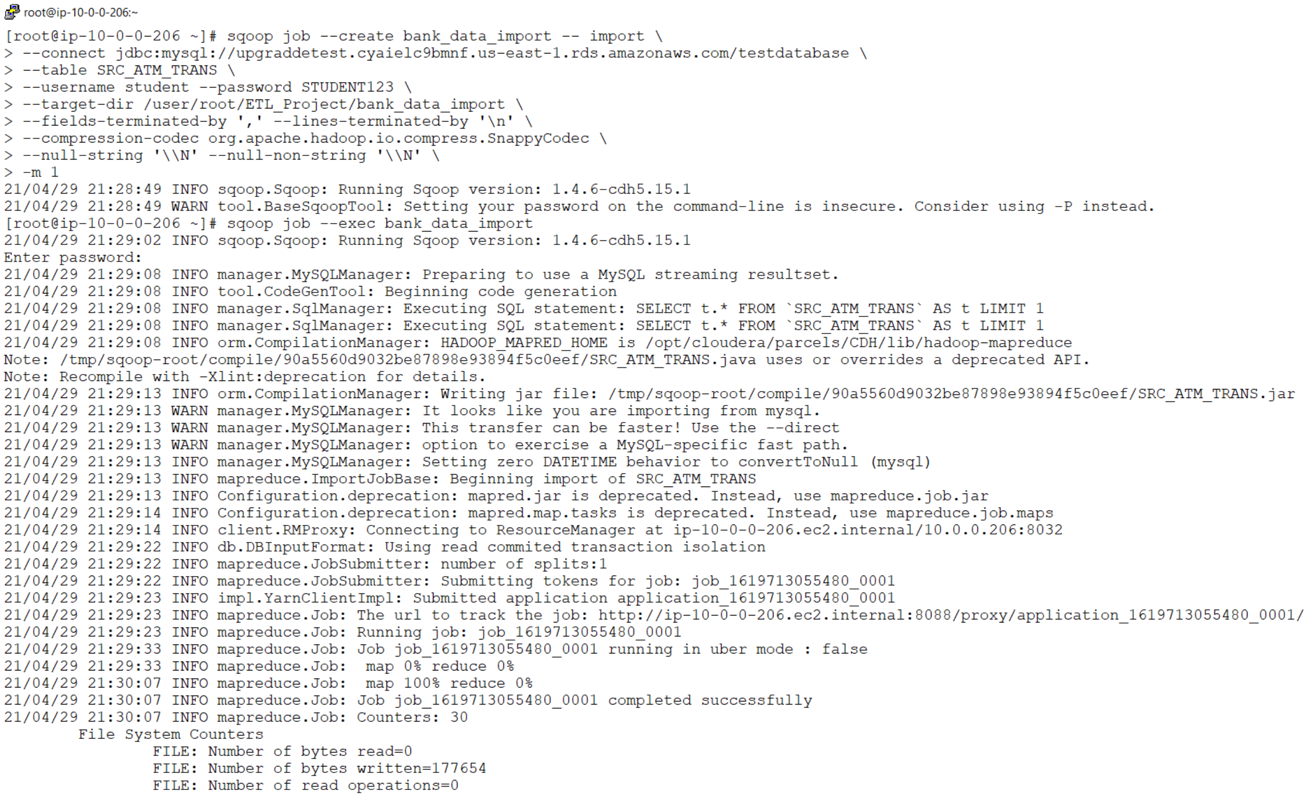
```
sqoop job --create bank_data_import -- import \
--connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-
1.rds.amazonaws.com/testdatabase \
--table SRC_ATM_TRANS \
--username student --password STUDENT123 \
--target-dir /user/root/ETL_Project/bank_data_import \
--fields-terminated-by ',' --lines-terminated-by '\n' \
--compression-codec org.apache.hadoop.io.compress.SnappyCodec \
--null-string '\\N' --null-non-string '\\N' \
-m 1;
```

```
sqoop job --exec bank_data_import
```

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

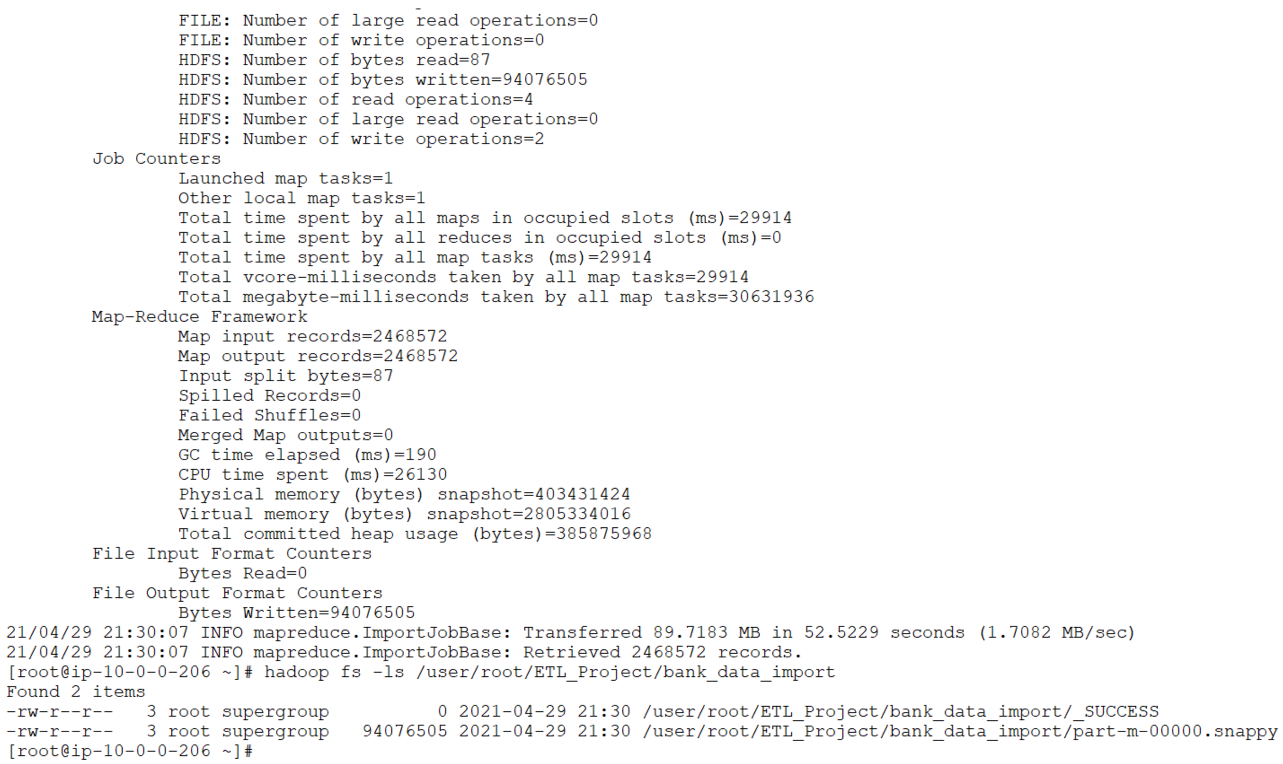
```
hadoop fs -ls /user/root/ETL_Project/bank_data_import
```

Screenshot of the imported data:



root@ip-10-0-0-206:~

```
[root@ip-10-0-0-206 ~]# sqoop job --create bank_data_import -- import \
> --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
> --table SRC_ATM_TRANS \
> --username student --password STUDENT123 \
> --target-dir /user/root/ETL_Project/bank_data_import \
> --fields-terminated-by ',' --lines-terminated-by '\n' \
> --compression-codec org.apache.hadoop.io.compress.SnappyCodec \
> --null-string '\\N' --null-non-string '\\N' \
> -m 1
21/04/29 21:28:49 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
21/04/29 21:28:49 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
[root@ip-10-0-0-206 ~]# sqoop job --exec bank_data_import
21/04/29 21:29:02 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
Enter password:
21/04/29 21:29:08 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/04/29 21:29:08 INFO tool.CodeGenTool: Beginning code generation
21/04/29 21:29:08 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
21/04/29 21:29:08 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `SRC_ATM_TRANS` AS t LIMIT 1
21/04/29 21:29:08 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/90a5560d9032be87898e93894f5c0eef/SRC_ATM_TRANS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/04/29 21:29:13 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/90a5560d9032be87898e93894f5c0eef/SRC_ATM_TRANS.jar
21/04/29 21:29:13 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/04/29 21:29:13 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/04/29 21:29:13 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/04/29 21:29:13 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/04/29 21:29:13 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS
21/04/29 21:29:13 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/04/29 21:29:14 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
21/04/29 21:29:14 INFO client.RMProxy: Connecting to ResourceManager at ip-10-0-0-206.ec2.internal/10.0.0.206:8032
21/04/29 21:29:22 INFO db.DBInputFormat: Using read commited transaction isolation
21/04/29 21:29:22 INFO mapreduce.JobSubmitter: number of splits:1
21/04/29 21:29:22 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1619713055480_0001
21/04/29 21:29:23 INFO impl.YarnClientImpl: Submitted application application_1619713055480_0001
21/04/29 21:29:23 INFO mapreduce.Job: The url to track the job: http://ip-10-0-0-206.ec2.internal:8088/proxy/application_1619713055480_0001/
21/04/29 21:29:23 INFO mapreduce.Job: Running job: job_1619713055480_0001
21/04/29 21:29:33 INFO mapreduce.Job: Job job_1619713055480_0001 running in uber mode : false
21/04/29 21:29:33 INFO mapreduce.Job:  map 0% reduce 0%
21/04/29 21:30:07 INFO mapreduce.Job:  map 100% reduce 0%
21/04/29 21:30:07 INFO mapreduce.Job: Job job_1619713055480_0001 completed successfully
21/04/29 21:30:07 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=177654
                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=94076505
                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)=29914
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=29914
                Total vcore-milliseconds taken by all map tasks=29914
                Total megabyte-milliseconds taken by all map tasks=30631936
        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)=26130
                Physical memory (bytes) snapshot=403431424
                Virtual memory (bytes) snapshot=2805334016
                Total committed heap usage (bytes)=385875968
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=94076505
21/04/29 21:30:07 INFO mapreduce.ImportJobBase: Transferred 89.7183 MB in 52.5229 seconds (1.7082 MB/sec)
21/04/29 21:30:07 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-10-0-0-206 ~]# hadoop fs -ls /user/root/ETL_Project/bank_data_import
Found 2 items
-rw-r--r--   3 root supergroup          0 2021-04-29 21:30 /user/root/ETL_Project/bank_data_import/_SUCCESS
-rw-r--r--   3 root supergroup   94076505 2021-04-29 21:30 /user/root/ETL_Project/bank_data_import/part-m-00000.snappy
[root@ip-10-0-0-206 ~]#
```