Data Ingestion from the RDS to HDFS using Sqoop

Connecting to msql-connector -

- 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
- cd mysql-connector-java-8.0.25
- sudo cp 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.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase --username student --password STUDENT123 --table SRC_ATM_TRANS --target-dir /user/root/nord_bank_atm --m 1

```
[hadoop@ip-172-31-54-94 mysql-connector-java-8.0.25]$ sqoop import --connect jdb
c:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase --u
sername student --password STUDENT123 --table SRC_ATM_TRANS --target-dir /user/r
oot/nord_bank_atm --m 1
                Total megabyte-milliseconds taken by all map tasks=40648704
        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)=182
                CPU time spent (ms)=25430
                Physical memory (bytes) snapshot=620044288
                Virtual memory (bytes) snapshot=3303927808
                Total committed heap usage (bytes)=518520832
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=531214815
23/06/18 13:57:47 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 43.16
16 seconds (11.7374 MB/sec)
23/06/18 13:57:47 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
```

Explanation-

- Using the above sqoop command we can confirm that a total of 2468572 records were created.
- For connection we are using the jdbc connection, username, password and the table provided to us.
- For this task we are using only 1 mapper to move data to user/root/nord_bank_atm

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

- hadoop fs -ls /user/root/nord_bank_atm
- hadoop fs -cp /user/root/nord_bank_atm/part-m-00000 /user/livy/

```
[hadoop@ip-172-31-54-94 mysql-connector-java-8.0.25]$ hadoop fs -ls /user/root/n
ord_bank_atm
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-06-18 13:57 /user/root/nord
_bank_atm/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup  531214815 2023-06-18 13:57 /user/root/nord
_bank_atm/part-m-00000
```

Explanation –

- We can confirm from the first command that a part file was created
- Using the 2nd command we are moving the data to /livy , so spark can access it

Screenshot of the imported data: