Data Ingestion from the RDS to HDFS using Sqoop

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

Sql jar file import:

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/

```
[root@ip-10-0-0-103 ~]# wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
--2021-11-20 17:15:01--  https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
Resolving de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)... 52.217.39.76
Connecting to de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)|52.217.39.76|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4079310 (3.9M) [application/x-gzip]
Saving to: 'mysql-connector-java-8.0.25.tar.gz'

100%[=========================================================================================================================================>]

2021-11-20 17:15:01 (120 MB/s) - 'mysql-connector-java-8.0.25.tar.gz' saved [4079310/4079310]

[root@ip-10-0-0-103 ~]# tar -xvf mysql-connector-java-8.0.25.tar.gz
```

Import command:
sqoop import --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase --table SRC_ATM_TRANS --username student --password STUDENT123 --null-string '\\N' --fields-terminated-by , --escaped-by \\ --enclosed-by '\"' --null-non-string '\\N' --target-dir /user/livy/de1 -m 1

Explanation :

Along with default options, we have used the Null handling parameter to treat the null as blank instead of string 'NULL' .
We have used the escaped and enclosed parameter to handle the comma values in the Message_text column.
Used one mapper to import the data and imported the data to /user/livy/de1 folder

```
[root@ip-10-0-0-105 mysql-connector-java-8.0.25]# sqoop import --connect jdbc:my
sql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase --table
 SRC_ATM_TRANS --username student --password STUDENT123 --null-string '\\N' --fi
elds-terminated-by , --escaped-by \\ --enclosed-by '\"' --null-non-string '\\N'
--target-dir /user/livy/de1 -m 1
```

```
21/11/23 17:20:25 INFO mapreduce.ImportJobBase: Transferred 666.7002 MB in 100.0
76 seconds (6.6619 MB/sec)
21/11/23 17:20:25 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-10-0-0-187 ~]# hadoop fs -ls /user/livy/
Found 1 items
drwxr-xr-x   - root livy          0 2021-11-23 17:20 /user/livy/de1
[root@ip-10-0-0-187 ~]# hadoop fs -ls /user/livy/de1
Found 2 items
-rw-r--r--   1 root livy          0 2021-11-23 17:20 /user/livy/de1/_SUCCESS
-rw-r--r--   1 root livy  699085798 2021-11-23 17:20 /user/livy/de1/part-m-00000
[root@ip-10-0-0-187 ~]#
```

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

hadoop fs -ls /user/livy/de1

hadoop fs -cp /user/livy/de1/part-m-00000 /user/livy/

Explanation:

Imported the file to /user/livy/de1 folder than moved in /user/livy/ folder for Spark access.

Screenshot of the imported data: