Load data from AWS RDS to Hadoop

<Command to import data from AWS RDS to Hadoop>

1. First we need to setup MySQL Connector on AWS EMR

    a. Run the following command to install the MySQL connector jar file:

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

    b. Run the following step to extract the MySQL connector tar file

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

    c. go to the MySQL Connector directory and then copy it to the Sqoop library to complete the installation.

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

2. Now we run the Sqoop import command to import data from AWS RDS to Hadoop

```
sqoop import \
--connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
--table bookings \
--username student --password STUDENT123 \
--null-string '\\N' --null-non-string '\\N' \
--target-dir bookings_data \
-m 1
```

<Command to view the imported data>

```
hadoop fs -ls bookings_data
hadoop fs -cat bookings_data/part-m-00000 | wc -l
```

<Screenshot of the data>

```
[hadoop@ip-172-31-8-156 ~]$ sqoop import \
> --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
> --table bookings \
> --username student --password STUDENT123 \
> --null-string '\\N' --null-non-string '\\N' \
> --target-dir bookings_data \
> -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/07/22 10:10:57 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
22/07/22 10:10:57 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/07/22 10:10:57 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/07/22 10:10:57 INFO tool.CodeGenTool: Beginning code generation
```

```
        Map-Reduce Framework
                Map input records=1000
                Map output records=1000
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=76
                CPU time spent (ms)=2650
                Physical memory (bytes) snapshot=273190912
                Virtual memory (bytes) snapshot=3287744512
                Total committed heap usage (bytes)=242221056
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=165678
22/07/22 10:11:22 INFO mapreduce.ImportJobBase: Transferred 161.7949 KB in 19.8253 seconds (8.161 KB/sec)
22/07/22 10:11:22 INFO mapreduce.ImportJobBase: Retrieved 1000 records.
[hadoop@ip-172-31-8-156 ~]$
```

```
[hadoop@ip-172-31-8-156 ~]$ hadoop fs -ls
Found 4 items
drwxr-xr-x   - hadoop hadoop          0 2022-07-22 10:11 bookings_data
drwxr-xr-x   - hadoop hadoop          0 2022-07-22 07:53 clickstream_checkpoint
drwxr-xr-x   - hadoop hadoop          0 2022-07-22 07:53 clickstream_data
drwxr-xr-x   - hadoop hadoop          0 2022-07-22 09:02 clickstream_data_flatten
[hadoop@ip-172-31-8-156 ~]$ hadoop fs -ls bookings_data
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2022-07-22 10:11 bookings_data/_SUCCESS
-rw-r--r--   1 hadoop hadoop     165678 2022-07-22 10:11 bookings_data/part-m-00000
[hadoop@ip-172-31-8-156 ~]$ hadoop fs -cat bookings_data/part-m-00000 | wc -l
1000
[hadoop@ip-172-31-8-156 ~]$
```