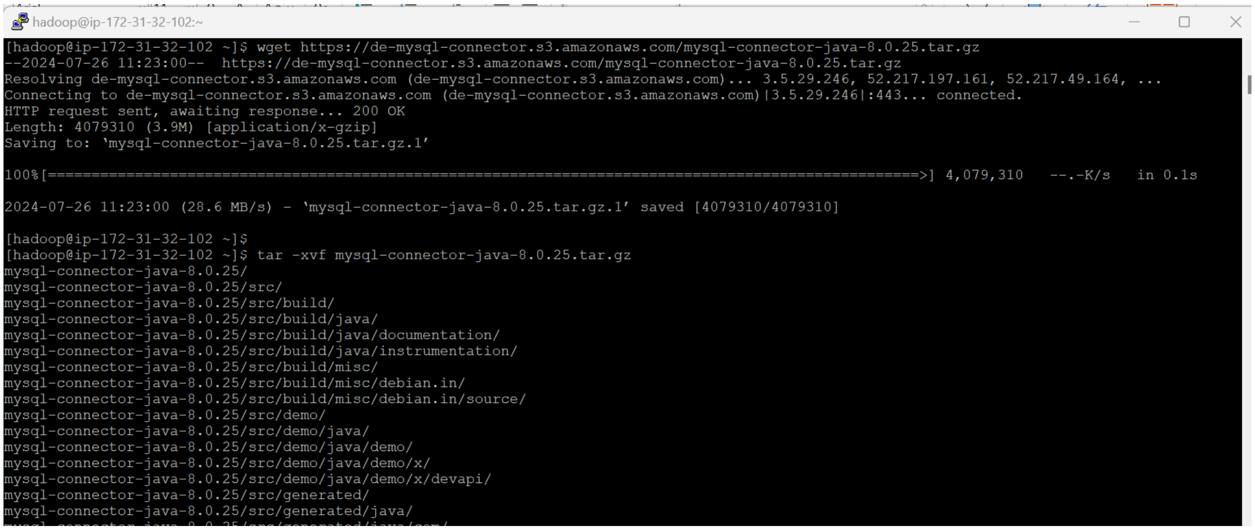
Load data from AWS RDS to Hadoop

Ingest data from AWS RDS to Hadoop using Sqoop

1. Installing MySQL 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



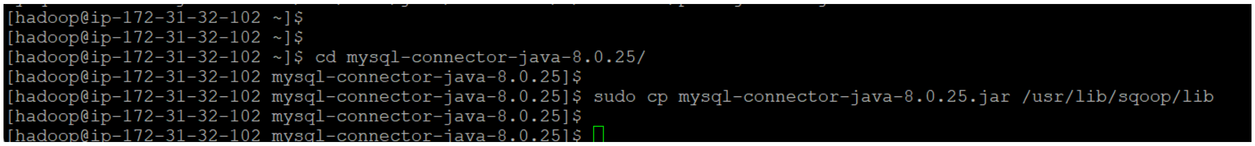
```
[hadoop@ip-172-31-32-102 ~]$ wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
--2024-07-26 11:23:00--  https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz
Resolving de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)... 3.5.29.246, 52.217.197.161, 52.217.49.164, ...
Connecting to de-mysql-connector.s3.amazonaws.com (de-mysql-connector.s3.amazonaws.com)|3.5.29.246|: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.1'

100%[=====================================================================================================>] 4,079,310   --.-K/s   in 0.1s

2024-07-26 11:23:00 (28.6 MB/s) - 'mysql-connector-java-8.0.25.tar.gz.1' saved [4079310/4079310]

[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ tar -xvf mysql-connector-java-8.0.25.tar.gz
mysql-connector-java-8.0.25/
mysql-connector-java-8.0.25/src/
mysql-connector-java-8.0.25/src/build/
mysql-connector-java-8.0.25/src/build/java/
mysql-connector-java-8.0.25/src/build/java/documentation/
mysql-connector-java-8.0.25/src/build/java/instrumentation/
mysql-connector-java-8.0.25/src/build/misc/
mysql-connector-java-8.0.25/src/build/misc/debian.in/
mysql-connector-java-8.0.25/src/build/misc/debian.in/source/
mysql-connector-java-8.0.25/src/demo/
mysql-connector-java-8.0.25/src/demo/java/
mysql-connector-java-8.0.25/src/demo/java/demo/
mysql-connector-java-8.0.25/src/demo/java/demo/x/
mysql-connector-java-8.0.25/src/demo/java/demo/x/devapi/
mysql-connector-java-8.0.25/src/generated/
mysql-connector-java-8.0.25/src/generated/java/
mysql-connector-java-8.0.25/src/generated/java/com/
```

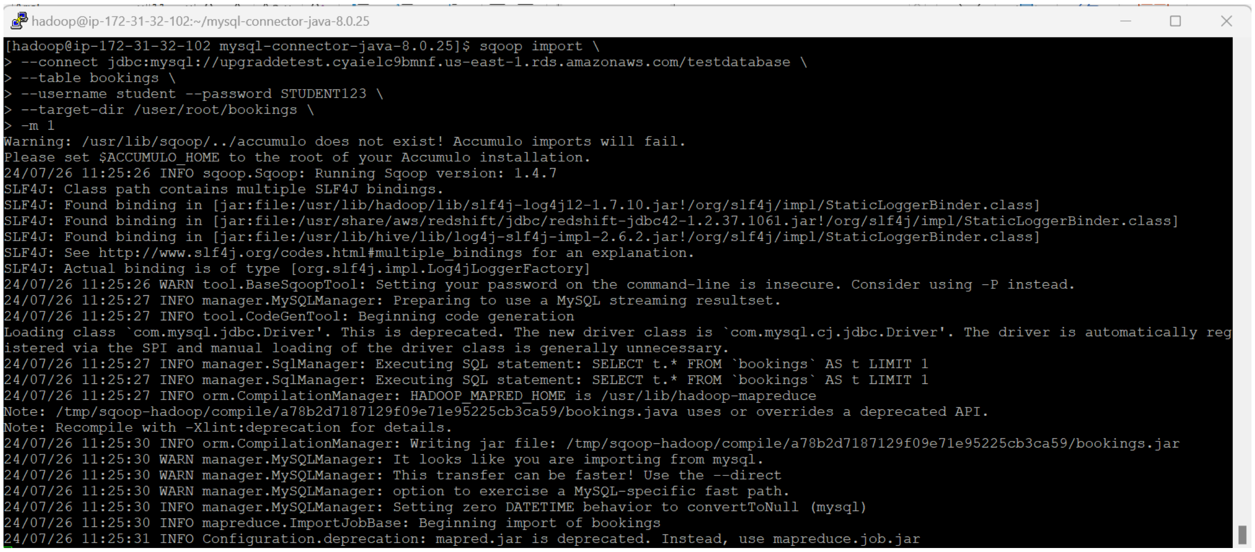
cd mysql-connector-java-8.0.25/
sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib



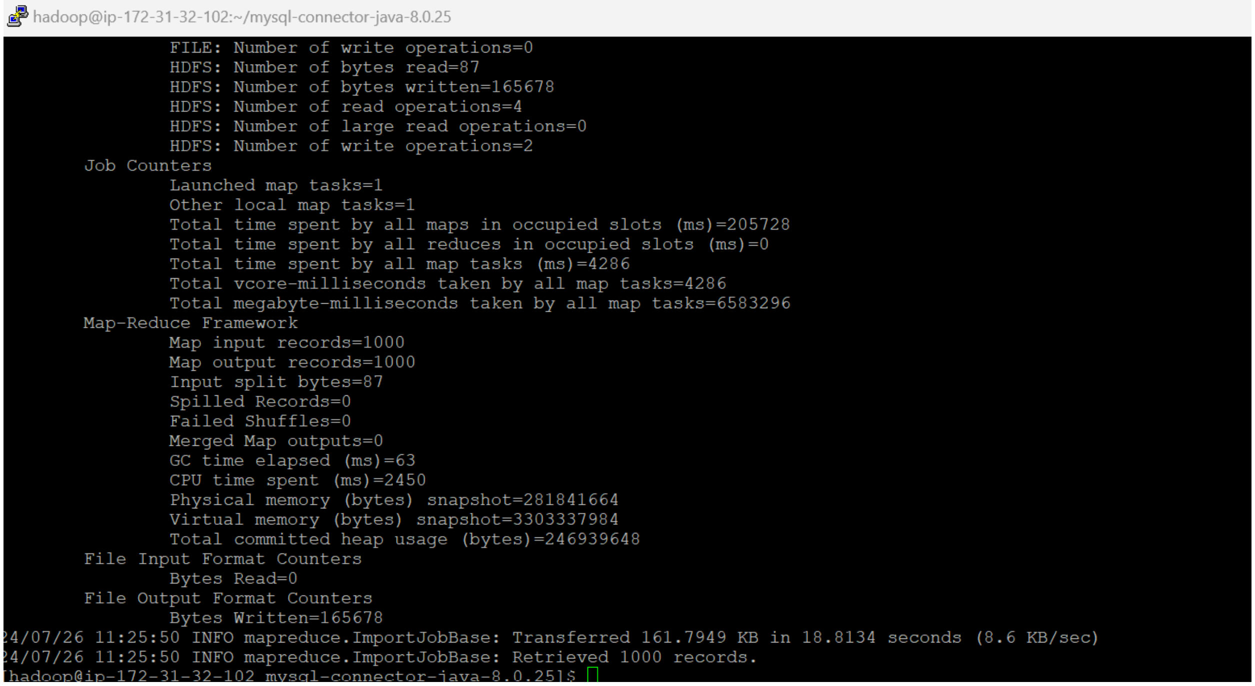
```
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ cd mysql-connector-java-8.0.25/
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$ sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
```

2. Ingesting batch data (bookings data stored in the RDS) to Hadoop using Sqoop

```
sqoop import \
--connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
--table bookings \
--username student --password STUDENT123 \
--target-dir /user/root/bookings \
-m 1
```

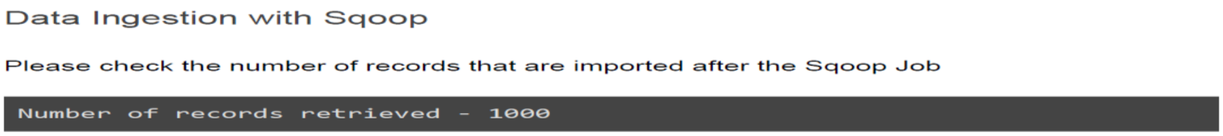


```
hadoop@ip-172-31-32-102:~/mysql-connector-java-8.0.25
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$ sqoop import \
> --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
> --table bookings \
> --username student --password STUDENT123 \
> --target-dir /user/root/bookings \
> -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.
24/07/26 11:25:26 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]
24/07/26 11:25:26 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/07/26 11:25:27 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/07/26 11:25:27 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically reg
istered via the SPI and manual loading of the driver class is generally unnecessary.
24/07/26 11:25:27 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `bookings` AS t LIMIT 1
24/07/26 11:25:27 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `bookings` AS t LIMIT 1
24/07/26 11:25:27 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/a78b2d7187129f09e71e95225cb3ca59/bookings.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/07/26 11:25:30 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/a78b2d7187129f09e71e95225cb3ca59/bookings.jar
24/07/26 11:25:30 WARN manager.MySQLManager: It looks like you are importing from mysql.
24/07/26 11:25:30 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
24/07/26 11:25:30 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
24/07/26 11:25:30 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
24/07/26 11:25:30 INFO mapreduce.ImportJobBase: Beginning import of bookings
24/07/26 11:25:31 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
```



```
hadoop@ip-172-31-32-102:~/mysql-connector-java-8.0.25
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=165678
                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)=205728
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4286
                Total vcore-milliseconds taken by all map tasks=4286
                Total megabyte-milliseconds taken by all map tasks=6583296
        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)=63
                CPU time spent (ms)=2450
                Physical memory (bytes) snapshot=281841664
                Virtual memory (bytes) snapshot=3303337984
                Total committed heap usage (bytes)=246939648
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=165678
24/07/26 11:25:50 INFO mapreduce.ImportJobBase: Transferred 161.7949 KB in 18.8134 seconds (8.6 KB/sec)
24/07/26 11:25:50 INFO mapreduce.ImportJobBase: Retrieved 1000 records.
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
```

*** No of records ingested is matching is matching with the validation documents*



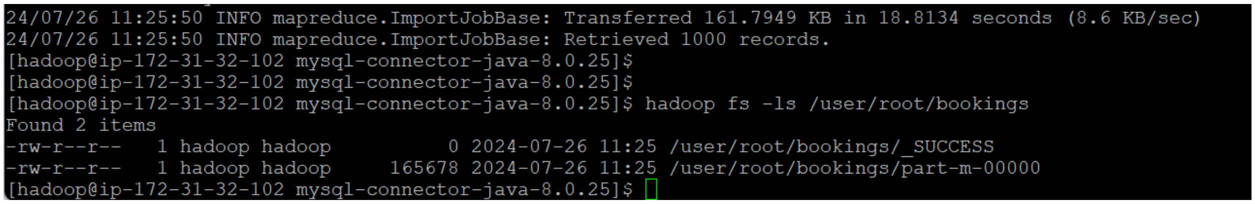
Data Ingestion with Sqoop

Please check the number of records that are imported after the Sqoop Job

```
Number of records retrieved - 1000
```

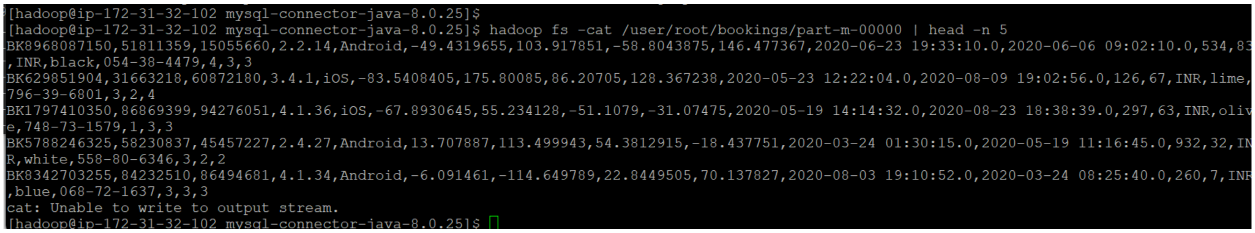
3. Validating the directory and files created in HDFS

hadoop fs -ls /user/root/bookings



```
24/07/26 11:25:50 INFO mapreduce.ImportJobBase: Transferred 161.7949 KB in 18.8134 seconds (8.6 KB/sec)
24/07/26 11:25:50 INFO mapreduce.ImportJobBase: Retrieved 1000 records.
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$ hadoop fs -ls /user/root/bookings
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2024-07-26 11:25 /user/root/bookings/_SUCCESS
-rw-r--r--   1 hadoop hadoop     165678 2024-07-26 11:25 /user/root/bookings/part-m-00000
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
```

hadoop fs -cat /user/root/bookings/part-m-00000 | head -n 5



```
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$ hadoop fs -cat /user/root/bookings/part-m-00000 | head -n 5
BK8968087150,51811359,15055660,2.2.14,Android,-49.4319655,103.917851,-58.8043875,146.477367,2020-06-23 19:33:10.0,2020-06-06 09:02:10.0,534,83
,INR,black,054-38-4479,4,3,3
BK629851904,31663218,60872180,3.4.1,iOS,-83.5408405,175.80085,86.20705,128.367238,2020-05-23 12:22:04.0,2020-08-09 19:02:56.0,126,67,INR,lime,
796-39-6801,3,2,4
BK1797410350,86869399,94276051,4.1.36,iOS,-67.8930645,55.234128,-51.1079,-31.07475,2020-05-19 14:14:32.0,2020-08-23 18:38:39.0,297,63,INR,oliv
e,748-73-1579,1,3,3
BK5788246325,58230837,45457227,2.4.27,Android,13.707887,113.499943,54.3812915,-18.437751,2020-03-24 01:30:15.0,2020-05-19 11:16:45.0,932,32,IN
R,white,558-80-6346,3,2,2
BK8342703255,84232510,86494681,4.1.34,Android,-6.091461,-114.649789,22.8449505,70.137827,2020-08-03 19:10:52.0,2020-03-24 08:25:40.0,260,7,INR
,blue,068-72-1637,3,3,3
cat: Unable to write to output stream.
[hadoop@ip-172-31-32-102 mysql-connector-java-8.0.25]$
```