

Data Ingestion from the RDS to HDFS using Sqoop

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

<Sqoop Import Command>

Answer:

creating file in hdfs to import data from MYSQL database

```
[root@ip-172-31-45-59 ~]#hadoop fs -rm -r /user/root/ETL_Proj_Data
```

importing table using sqoop command (using compression)

```
[root@ip-172-31-45-59 ~]#sqoop import \  
--connect jdbc:mysql://upgradetest.cyaieic9bmnf.us-east-1.rds.amazonaws.com/testdatabase \  
--table SRC_ATM_TRANS \  
--username student --password STUDENT123 \  
--target-dir /user/root/ETL_Proj_Data \  
--compression-codec org.apache.hadoop.io.compress.SnappyCodec \  
--null-string '\N' --null-non-string '\N' \  
-m 1
```

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

<Command used>

Answer:

```
hadoop fs -ls /user/root/ETL_Proj_Data
```

Screenshot of the imported data:

<Screenshot of the imported data>

Answer:

```
root@ip-10-0-0-45:~#
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)=29876
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=29876
  Total vcore-milliseconds taken by all map tasks=29876
  Total megabyte-milliseconds taken by all map tasks=30593024
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)=284
  CPU time spent (ms)=29570
  Physical memory (bytes) snapshot=401092608
  Virtual memory (bytes) snapshot=2804875264
  Total committed heap usage (bytes)=389021696
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=94076505
21/06/08 18:04:51 INFO mapreduce.ImportJobBase: Transferred 89.7183 MB in 59.208
3 seconds (11.5153 MB/sec)
21/06/08 18:04:51 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-10-0-0-45 ~]#
[root@ip-10-0-0-45 ~]#
[root@ip-10-0-0-45 ~]# hadoop fs -ls /user/root/ETL_Proj_Data
Found 2 items
-rw-r--r-- 3 root supergroup          0 2021-06-08 18:04 /user/root/ETL_Proj_D
ata/_SUCCESS
-rw-r--r-- 3 root supergroup  94076505 2021-06-08 18:04 /user/root/ETL_Proj_D
ata/part-m-000000.snappy
[root@ip-10-0-0-45 ~]#
[root@ip-10-0-0-45 ~]#
[root@ip-10-0-0-45 ~]#
```

SQOOP COMMAND SS

```
root@ip-10-0-0-45:~#
# login as: ec2-user
# Authenticating with public key "imported-openssh-key"
Last login: Tue Jun  8 16:17:53 2021 from 129.201.163.41
[ec2-user@ip-10-0-0-45 ~]$ sudo -i
[root@ip-10-0-0-45 ~]#
[root@ip-10-0-0-45 ~]# hadoop fs -rm -r /user/root/ETL_Proj_Data
21/06/08 18:01:56 INFO fs.TrashPolicyDefault: Moved: 'hdfs://ip-10-0-0-45.ec2.in
ternal:8020/user/root/ETL_Proj_Data' to trash at: hdfs://ip-10-0-0-45.ec2.intern
al:8020/user/root/.Trash/Current/user/root/ETL_Proj_Data
[root@ip-10-0-0-45 ~]#
[root@ip-10-0-0-45 ~]#
[root@ip-10-0-0-45 ~]# sqoop import \
> --connect jdbc:mysql://upgradetest.cyaieic9bmuf.us-east-1.rds.amazonaws.com/t
estdatabase \
> --table SRC_ATM_TRANS \
> --username student --password STUDENT123 \
> --target-dir /user/root/ETL_Proj_Data \
> --fields-terminated-by ',' --lines-terminated-by '\n' \
> --compression-codec org.apache.hadoop.io.compress.SnappyCodec \
> --null-string '\N' --null-non-string '\N' \
> -m 1
21/06/08 18:03:45 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.15.1
21/06/08 18:03:45 WARN tool.BaseSqoopTool: Setting your password on the command-
line is insecure. Consider using -P instead.
21/06/08 18:03:45 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
21/06/08 18:03:45 INFO tool.CodeGenTool: Beginning code generation
21/06/08 18:03:46 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM 'SRC_ATM_TRANS' AS t LIMIT 1
21/06/08 18:03:46 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM 'SRC_ATM_TRANS' AS t LIMIT 1
21/06/08 18:03:46 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloude
ra/parcels/CDH/11b/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/73bd80e36180a78202f5baa58c0fa276/SRC_ATM_TRANS.jav
a uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/06/08 18:03:51 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root
/compile/73bd80e36180a78202f5baa58c0fa276/SRC_ATM_TRANS.jar
21/06/08 18:03:51 WARN manager.MySQLManager: It looks like you are importing fro
m mysql.
21/06/08 18:03:51 WARN manager.MySQLManager: This transfer can be faster! Use th
e --direct
21/06/08 18:03:51 WARN manager.MySQLManager: option to exercise a MySQL-specific
fast path.
```