

ETL Project : ATM Transactions

Sqoop Data Ingestion

➤ **Step-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
cd mysql-connector-java-8.0.25/
sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/
```

➤ Step 2 - Sqoop command to import data from RDS to HDFS

```
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/SRC_ATM_TRANS/ \
-m 1
```

```
hadoop@ip-172-31-68-8:/mysql-connector-java-8.0.25$ 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/SRC_ATM_TRANS/ \
-m 1
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set ACCUMULO_HOME to the root of your Accumulo installation.
24/06/15 02:00:25 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jarfile:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jarfile:/usr/share/awx/redshift-jdbc42-1.2.37.1061.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]
24/06/15 02:00:25 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/06/15 02:00:25 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/06/15 02:00:25 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 registered via the SPI and manual loading is not needed.
24/06/15 02:00:26 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1
24/06/15 02:00:26 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/ooaf898bd89f2c0a10a7742bod7595/SRC_ATM_TRANS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/06/15 02:00:28 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/ooaf898bd89f2c0a10a7742bod7595/SRC_ATM_TRANS.jar
24/06/15 02:00:28 WARN manager.MySQLManager: It looks like you are importing from mysql.
24/06/15 02:00:28 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
24/06/15 02:00:28 INFO manager.MySQLManager: option to exercise a MySQL-specific fast path.
24/06/15 02:00:28 INFO manager.MySQLManager: Setting auto DEDUP behavior to convertToNull (mysql)
24/06/15 02:00:28 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS
24/06/15 02:00:29 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
24/06/15 02:00:29 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/06/15 02:00:29 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-68-8.ec2.internal/172.31.68.8:8032
24/06/15 02:00:43 INFO mapreduce.JobSubmitter: Using read committed transaction isolation
24/06/15 02:00:43 INFO mapreduce.JobSubmitter: number of splits=1
24/06/15 02:00:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1718416171485_0001
24/06/15 02:00:44 INFO impl.YarnClientImpl: Submitted application application_1718416171485_0001
24/06/15 02:00:44 INFO mapreduce.Job: The url to track the job: http://ip-172-31-68-8.ec2.internal:20888/proxy/application_1718416171485_0001/
24/06/15 02:00:44 INFO mapreduce.Job: Running job: job_1718416171485_0001
24/06/15 02:00:53 INFO mapreduce.Job: Job job_1718416171485_0001 running in uber mode : false
24/06/15 02:01:21 INFO mapreduce.Job: map 0% reduce 0%
24/06/15 02:01:21 INFO mapreduce.Job: map 100% reduce 0%
24/06/15 02:01:21 INFO mapreduce.Job: Job job_1718416171485_0001 completed successfully
24/06/15 02:01:21 INFO mapreduce.Job: Counters: 30
File System Counters
FILE: Number of bytes read=0
FILE: Number of bytes written=189867
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=37
HDFS: Number of bytes written=531214815
HDFS: Number of read operations=4
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
Job Counters
Launched map tasks=1
Other local map tasks=1
Total time spent by all maps in occupied slots (ms)=1240512
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=25844
Total vcore-milliseconds taken by all map tasks=25844
Total megabyte-milliseconds taken by all map tasks=39696384
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)=208
CPU time spent (ms)=27160
Physical memory (bytes) snapshot=602841088
Virtual memory (bytes) snapshot=3285446656
Total committed heap usage (bytes)=533968064
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=531214815
24/06/15 02:01:21 INFO mapreduce.ImportJobBase: Transferred 506.4059 MB in 51.8102 seconds (9.7781 MB/sec)
24/06/15 02:01:21 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
(hadoop@ip-172-31-68-8 mysql-connector-java-8.0.25)
```

ETL Project : ATM Transactions - Sqoop Data Ingestion

➤ Step 3 – Validation of imported data in HDFS

a) `hadoop fs -ls /user/root/SRC_ATM_TRANS/`

```
[hadoop@ip-172-31-68-8 ~]$ hadoop fs -ls /user/root/SRC_ATM_TRANS/
Found 2 items
-rw-r--r-- 1 hadoop hadoop          0 2024-06-15 02:01 /user/root/SRC_ATM_TRANS/_SUCCESS
-rw-r--r-- 1 hadoop hadoop 531214815 2024-06-15 02:01 /user/root/SRC_ATM_TRANS/part-m-00000
```

b) `hadoop fs -cat /user/root/SRC_ATM_TRANS/part-m-00000 | wc -l`

```
[hadoop@ip-172-31-68-8 ~]$ hadoop fs -cat /user/root/SRC_ATM_TRANS/part-m-00000 | wc -l
2468572
[hadoop@ip-172-31-68-8 ~]$
```

c) `hadoop fs -cat /user/root/SRC_ATM_TRANS/part-m-00000 | head -n 5`

```
[hadoop@ip-172-31-68-8 ~]$ hadoop fs -cat /user/root/SRC_ATM_TRANS/part-m-00000 | head -n 5
2017,January,1,Sunday,0,Active,1,MCR,NÅfÅ|stved,Farimagvej,5,4700,55.239,11.763,DKK,MasterCard,5643,Withdrawal,,,55.230,11.761,2616038,Masterwed,281.150,1014,87,7,260,0.215,92,500
,Rain,light rain
2017,January,1,Sunday,0,Inactive,2,MCR,Vejgaard,Hadsundvej,20,9000,57.043,9.950,DKK,MasterCard,1764,Withdrawal,,,57.048,9.935,2616235,NÅfÅ|,rresundby,280.640,1020,93,9,250,0.590,92
,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,2,MCR,Vejgaard,Hadsundvej,20,9000,57.043,9.950,DKK,VISA,1891,Withdrawal,,,57.048,9.935,2616235,NÅfÅ|,rresundby,280.640,1020,93,9,250,0.590,92,500,R
ain,light rain
2017,January,1,Sunday,0,Inactive,3,MCR,Ikast,NÅfÅ|VdhustrÅfÅ|det,12,7430,56.139,9.154,DKK,VISA,4166,Withdrawal,,,56.139,9.158,2619426,Ikast,281.150,1011,100,6,240,0.000,75,300,Dri
zzle,light intensity drizzle
2017,January,1,Sunday,0,Active,4,MCR,Svogerslev,BrÅfÅ|nsager,1,4000,55.634,12.018,DKK,MasterCard,5153,Withdrawal,,,55.642,12.080,2614481,Roskilde,280.610,1014,87,7,260,0.000,88,70
1,Mist,mist
```

ETL Project : ATM Transactions - Sqoop Data Ingestion