

SPARNOD BANK ETL PROJECT

Sqoop Data Ingestion

Submitted By:

Prerna Gupta (prerna.gupta1792@gmail.com)

Nikhil Mahakal (nikhilmahakal07@gmail.com)

Pradeep Singh Negi (pradeep.negidba@gmail.com)

Data Ingestion from the RDS to HDFS using Sqoop

1. Installed mysql connector on the cluster using below command to run sqoop.
sudo -i
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/
2. Run the below sqoop command to ingest data from SRC_ATM_TRANS table in testdatabase 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/hadoop/ETL/sparnod \  
-m 1
```

```
hadoop@ip-172-31-7-130:~$ 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/hadoop/ETL/sparnod \  
-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.  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/lib/hbase/lib/client-facing-thirdparty/slf4j-reload4j-1.7.33.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.Reload4jLoggerFactory]  
2024-06-02 13:58:35,641 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7  
2024-06-02 13:58:35,679 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
2024-06-02 13:58:35,815 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.  
2024-06-02 13:58:35,816 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 of the driver class is generally unnecessary.  
2024-06-02 13:58:36,496 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1  
2024-06-02 13:58:36,560 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'SRC_ATM_TRANS' AS t LIMIT 1  
2024-06-02 13:58:36,586 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce  
2024-06-02 13:58:41,808 ERROR orm.CompilationManager: Could not rename /tmp/sqoop-hadoop/compile/69b98703835118e509a1bac00d167829/SRC_ATM_TRANS.java to /home/hadoop/./SRC_ATM_TRANS.java. Error: Destination '/home/hadoop/./SRC_ATM_TRANS.java' already exists  
2024-06-02 13:58:41,808 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/69b98703835118e509a1bac00d167829/SRC_ATM_TRANS.jar  
2024-06-02 13:58:41,833 WARN manager.MySQLManager: It looks like you are importing from mysql.  
2024-06-02 13:58:41,833 WARN manager.MySQLManager: This transfer can be faster! Use the --direct  
2024-06-02 13:58:41,834 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.  
2024-06-02 13:58:41,834 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)  
2024-06-02 13:58:42,033 INFO mapreduce.ImportJobBase: Beginning import of SRC_ATM_TRANS  
2024-06-02 13:58:42,712 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps  
2024-06-02 13:58:43,036 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-7-130.ec2.internal/172.31.7.130:10200  
2024-06-02 13:58:43,552 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1717333714660_0001  
2024-06-02 13:58:46,668 INFO db.DBInputFormat: Using read committed transaction isolation  
2024-06-02 13:58:46,741 INFO mapreduce.JobSubmitter: number of splits:1  
2024-06-02 13:58:47,057 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1717333714660_0001  
2024-06-02 13:58:47,058 INFO mapreduce.JobSubmitter: Executing with tokens: []  
2024-06-02 13:58:47,276 INFO conf.Configuration: resource-types.xml not found  
2024-06-02 13:58:47,278 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.  
2024-06-02 13:58:47,664 INFO impl.YarnClientImpl: Submitted application application_1717333714660_0001  
2024-06-02 13:58:47,723 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-130.ec2.internal:20888/proxy/application_1717333714660_0001/  
2024-06-02 13:58:47,724 INFO mapreduce.Job: Running job: job_1717333714660_0001  
2024-06-02 13:58:55,850 INFO mapreduce.Job: Job job_1717333714660_0001 running in uber mode : false  
2024-06-02 13:58:55,853 INFO mapreduce.Job: map 0% reduce 0%
```

```

2024-06-02 13:58:55,853 INFO mapreduce.Job: map 0% reduce 0%
2024-06-02 13:59:33,199 INFO mapreduce.Job: map 100% reduce 0%
2024-06-02 13:59:34,213 INFO mapreduce.Job: Job Job_1717333714660_0001 completed successfully
2024-06-02 13:59:34,290 INFO mapreduce.Job: Counters: 33
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=289437
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=65
  HDFS: Number of bytes written=531214815
  HDFS: Number of read operations=6
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
  HDFS: Number of bytes read erasure-coded=0
Job Counters
  Launched map tasks=1
  Other local map tasks=1
  Total time spent by all maps in occupied slots (ms)=1662000
  Total time spent by all reducers in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=34625
  Total vcore-milliseconds taken by all map tasks=34625
  Total megabyte-milliseconds taken by all map tasks=53184000
Map-Reduce Framework
  Map input records=2468572
  Map output records=2468572
  Input split bytes=85
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=336
  CPU time spent (ms)=37320
  Physical memory (bytes) snapshot=426426368
  Virtual memory (bytes) snapshot=3194728448
  Total committed heap usage (bytes)=306184192
  Peak Map Physical memory (bytes)=426426368
  Peak Map Virtual memory (bytes)=3194728448
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=531214815
2024-06-02 13:59:34,296 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 51.5691 seconds (9.8238 MB/sec)
2024-06-02 13:59:34,302 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[hadoop@ip-172-31-7-130 ~]$

```

- Run below command to check the files on HDFS after completion of import

hadoop fs -ls /user/hadoop/ETL/sparnod

```

2024-06-02 13:59:34,302 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[hadoop@ip-172-31-7-130 ~]$ hadoop fs -ls /user/hadoop/ETL/sparnod
Found 2 items
-rw-r--r--  1 hadoop hdfsadmingroup          0 2024-06-02 13:59 /user/hadoop/ETL/sparnod/_SUCCESS
-rw-r--r--  1 hadoop hdfsadmingroup 531214815 2024-06-02 13:59 /user/hadoop/ETL/sparnod/part-m-000000
[hadoop@ip-172-31-7-130 ~]$

```

- Run below command to check the number of records exported from RDS to HDFS file.

hadoop fs -cat /user/hadoop/ETL/sparnod/part-m-000000 | wc -l

```

-rw-r--r--  1 hadoop hdfsadmingroup 531214815 2024-06-02 13:59 /user/hadoop/ETL/sparnod/part-m-000000
[hadoop@ip-172-31-7-130 ~]$ hadoop fs -cat /user/hadoop/ETL/sparnod/part-m-000000 | wc -l
2468572
[hadoop@ip-172-31-7-130 ~]$

```

Above screenshot shows **2468572** rows imported.

5. Screenshot of imported data.

```
[hadoop@ip-172-31-7-130 ~]$ hadoop fs -cat /user/hadoop/ETL/sparnod/part-m-00000 | head -10
2017,January,1,Sunday,0,Active,1,NCR,NÅfÅ\stved,Farimagvej,8,4700,55.233,11.763,DKK,MasterCard,5643,Withdrawal,,,55.230,11.761,2616038,Naestved,281.150,1014,87,7,260,0.215,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,2,NCR,Vejgaard,Hadsundvej,20,9000,57.043,9.950,DKK,MasterCard,1764,Withdrawal,,,57.048,9.935,2616235,NÅfÅ\resundby,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,2,NCR,Vejgaard,Hadsundvej,20,9000,57.043,9.950,DKK,VISA,1891,Withdrawal,,,57.048,9.935,2616235,NÅfÅ\resundby,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Inactive,3,NCR,Ikast,RÅfÅ\vdhusstrÅfÅ\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,Drizzle,light intensity drizzle
2017,January,1,Sunday,0,Active,4,NCR,Svogerslev,BrÅfÅ\nsager,1,4000,55.634,12.018,DKK,MasterCard,5153,Withdrawal,,,55.642,12.080,2614481,Roskilde,280.610,1014,87,7,260,0.000,88,701,Mist,mist
2017,January,1,Sunday,0,Active,5,NCR,Nibe,Torvet,1,9240,56.983,9.639,DKK,MasterCard,3269,Withdrawal,,,56.981,9.639,2616483,Nibe,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,6,NCR,Fredericia,SjÅfÅ\llandsgade,33,7000,55.564,9.757,DKK,MasterCard,887,Withdrawal,,,55.566,9.753,2621951,Fredericia,281.150,1014,93,7,230,0.290,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,7,Diebold Nixdorf,Hjallerup,Hjallerup Centret,18,9320,57.168,10.148,DKK,Mastercard - on-us,4626,Withdrawal,,,57.165,10.146,2620275,Hjallerup,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
2017,January,1,Sunday,0,Active,8,NCR,GlyngÅfÅ\re,FÅfÅ\rgvej,1,7870,56.762,8.867,DKK,MasterCard,470,Withdrawal,,,56.793,8.853,2615964,Nykobing Mors,281.150,1011,100,6,240,0.000,75,300,Drizzle,light intensity drizzle
2017,January,1,Sunday,0,Active,9,Diebold Nixdorf,Hadsund,Storegade,12,9560,56.716,10.114,DKK,VISA,8473,Withdrawal,,,56.715,10.117,2620952,Hadsund,280.640,1020,93,9,250,0.590,92,500,Rain,light rain
cat: Unable to write to output stream.
[hadoop@ip-172-31-7-130 ~]$
```