

# Sqoop Setup Documentation

CSD 526 Big Data Analytics – ISA 3

## Team Members:

1. Shubham Kelkar
2. Pavan Sannaik

**Note:** Apache Sqoop is a retired project and is no longer actively maintained. While it may still function, we do encounter compatibility issues or limitations when using it with newer Hadoop (higher than 2.6.0) or other ecosystem components.

## Aim:

This documentation outlines the step-by-step process for installing and configuring Apache Sqoop (version 1.4.7) to facilitate efficient data transfer between a MySQL database and Hadoop Distributed File System (HDFS).

- It includes downloading and setting up Sqoop, integrating the MySQL JDBC driver, and executing practical examples to import data from MySQL to HDFS and export it back to a new MySQL table.
- The goal is to demonstrate Sqoop's functionality for seamless data migration within a Hadoop ecosystem, despite its status as a retired project, while addressing potential compatibility issues with newer Hadoop versions.

## Sqoop (1.4.7) Installation

1. Download Sqoop Binary Package: Uses wget to download the Sqoop 1.4.7 binary package (built for Hadoop 2.6.0) from the Apache archive.

```
wget https://archive.apache.org/dist/sqoop/1.4.7/sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
```

```
shubham-kelkar@shubham-kelkar:~$ ls
apache-hive-3.1.2-bin.tar.gz  Desktop  Downloads  hive  Music  Public  Templates
derby.log                   Documents  hadoop-3.3.6.tar.gz  metastore_db  Pictures  snap  Videos
shubham-kelkar@shubham-kelkar:~$ wget https://archive.apache.org/dist/sqoop/1.4.7/sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
--2025-03-21 18:27:15-- https://archive.apache.org/dist/sqoop/1.4.7/sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 17953604 (17M) [application/x-gzip]
Saving to: 'sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz'

sqoop-1.4.7.bin__hadoop-2. 100%[=====] 17.12M 755KB/s in 18s

2025-03-21 18:27:34 (976 KB/s) - 'sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz' saved [17953604/17953604]

shubham-kelkar@shubham-kelkar:~$ ls
apache-hive-3.1.2-bin.tar.gz  Downloads  Music  sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
derby.log                   hadoop-3.3.6.tar.gz  Pictures  Templates
Desktop                     hive  Public  Videos
Documents                   metastore_db  snap
shubham-kelkar@shubham-kelkar:~$ file sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz: gzip compressed data, was "sqoop-1.4.7.bin__hadoop-2.6.0.tar", last modified: Mon Dec 18 23:00:00 2017, max compression, from Unix, original size modulo 2^32 26613760
shubham-kelkar@shubham-kelkar:~$
```

2. Extract the Package: Use tar to extract the downloaded.tar.gz file, unpacking its contents.

```
tar -xvzf sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
```

3. Rename the Directory: Renames the extracted folder to “sqoop” for easier reference.

```
mv sqoop-1.4.7.bin__hadoop-2.6.0 sqoop
```

4. Set SQOOP\_HOME Environment Variable: Appends an export statement to ~/.bashrc to define SQOOP\_HOME as the Sqoop installation directory.

```
echo 'export SQOOP_HOME=~/.sqoop' >> ~/.bashrc
```

5. Update PATH Variable: Appends a command to ~/.bashrc that adds Sqoop's bin directory to the system PATH, enabling Sqoop commands to be run from any terminal.

```
echo 'export PATH=$PATH:$SQOOP_HOME/bin' >> ~/.bashrc
```

6. Apply Environment Changes: Sources ~/.bashrc to immediately update the current terminal session with the new environment variable settings.

```
source ~/.bashrc
```

7. Verify Sqoop Installation: Runs the "sqoop version" command to check and confirm the installed Sqoop version.

```
shubham-kelkar@shubham-kelkar: ~  
shubham-kelkar@shubham-kelkar:~$ echo 'export SQOOP_HOME=~/.sqoop' >> ~/.bashrc  
shubham-kelkar@shubham-kelkar:~$ echo 'export PATH=$PATH:$SQOOP_HOME/bin' >> ~/.bashrc  
shubham-kelkar@shubham-kelkar:~$ source ~/.bashrc  
shubham-kelkar@shubham-kelkar:~$ sqoop version  
Warning: /home/shubham-kelkar/sqoop/./hbase does not exist! HBase imports will fail.  
Please set $HBASE_HOME to the root of your HBase installation.  
Warning: /home/shubham-kelkar/sqoop/./hcatalog does not exist! HCatalog jobs will fail.  
Please set $HCAT_HOME to the root of your HCatalog installation.  
Warning: /home/shubham-kelkar/sqoop/./accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
Warning: /home/shubham-kelkar/sqoop/./zookeeper does not exist! Accumulo imports will fail.  
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.  
2025-03-21 18:39:01,124 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7  
Sqoop 1.4.7  
git commit id 2328971411f57f0cb683dfb79d19d4d19d185dd8  
Compiled by maugli on Thu Dec 21 15:59:58 STD 2017  
shubham-kelkar@shubham-kelkar:~$
```

## Sqoop Configuration for MySQL

1. MySQL JDBC Driver: Download the MySQL Connector/J JAR file and place it in the \$SQOOP\_HOME/lib/ directory

```
wget https://repo1.maven.org/maven2/mysql/mysql-connector-java/5.1.49/mysql-connector-java-5.1.49.jar -P ~/.sqoop/lib/
```

2. Verify Driver Installation: The command lists the files in the lib directory and filters for "mysql" to confirm the driver is there. output is mysql-connector-java-5.1.49.jar

```
ls ~/.sqoop/lib/ | grep mysql
```

3. Download commons-lang: Download the commons-lang library and saves it to the lib directory under the user's sqoop folder. Check if the commons-lang library (specifically the commons-lang-2.6.jar file) is present in the ~/sqoop/lib/ and then save the changes.

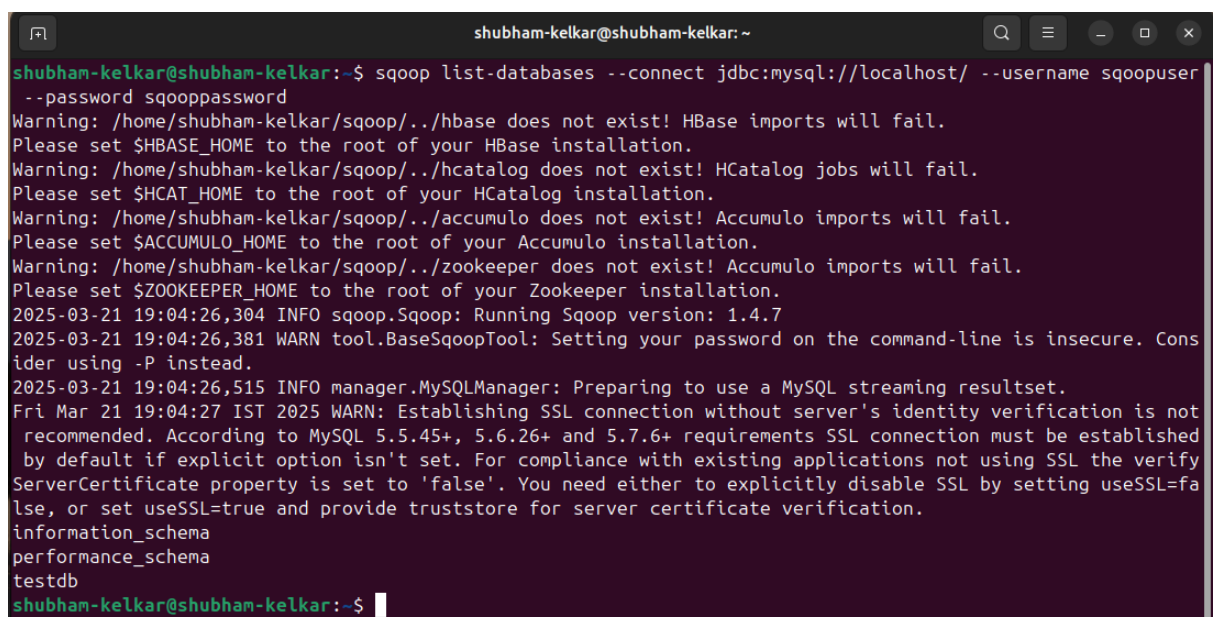
```
wget https://repo1.maven.org/maven2/commons-lang/commons-lang/2.6/commons-lang-2.6.jar -P ~/sqoop/lib/
```

```
ls ~/sqoop/lib/ | grep commons-lang
```

```
commons-lang-2.6.jar  
commons-lang3-3.4.jar
```

4. List MySQL Databases: Executes a Sqoop command to list all databases on your MySQL server, verifying connectivity and proper configuration.

```
sqoop list-databases --connect jdbc:mysql://localhost/ --username root --password root
```



```
shubham-kelkar@shubham-kelkar: ~  
shubham-kelkar@shubham-kelkar:~$ sqoop list-databases --connect jdbc:mysql://localhost/ --username sqoopuser  
--password sqooppassword  
Warning: /home/shubham-kelkar/sqoop/../hbase does not exist! HBase imports will fail.  
Please set $HBASE_HOME to the root of your HBase installation.  
Warning: /home/shubham-kelkar/sqoop/../hcatalog does not exist! HCatalog jobs will fail.  
Please set $HCAT_HOME to the root of your HCatalog installation.  
Warning: /home/shubham-kelkar/sqoop/../accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
Warning: /home/shubham-kelkar/sqoop/../zookeeper does not exist! Accumulo imports will fail.  
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.  
2025-03-21 19:04:26,304 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7  
2025-03-21 19:04:26,381 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Cons  
ider using -P instead.  
2025-03-21 19:04:26,515 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.  
Fri Mar 21 19:04:27 IST 2025 WARN: Establishing SSL connection without server's identity verification is not  
recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established  
by default if explicit option isn't set. For compliance with existing applications not using SSL the verify  
ServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting useSSL=fa  
lse, or set useSSL=true and provide truststore for server certificate verification.  
information_schema  
performance_schema  
testdb  
shubham-kelkar@shubham-kelkar:~$
```

5. Configure Hadoop Classpath: Appends a script to ~/.bashrc to automatically set the HADOOP\_CLASSPATH with necessary paths for Sqoop and the MySQL connector, ensuring proper class loading during execution. Apply the changes with source ~/.bashrc

```
echo '  
# Automatically set HADOOP_CLASSPATH for Sqoop  
export SQOOP_COMPILE_DIR=$(ls -td /tmp/sqoop-$(whoami)/compile/* 2>/dev/null | head -1)  
export HADOOP_CLASSPATH=$SQOOP_COMPILE_DIR:$(find $SQOOP_COMPILE_DIR -name "*.jar" | tr "\n" ":")/home/$(whoami)/sqoop/lib/mysql-  
connector-java-5.1.49.jar  
' >> ~/.bashrc
```

```
source ~/.bashrc
```

## Sqoop Implementation

This example demonstrates using Sqoop to import data from a MySQL table into HDFS, and then export it back to a new table in MySQL.

1. Create a new database (testdb) and create an employees table with appropriate columns in MySQL.

```
mysql> SHOW TABLES;  
+-----+  
| Tables_in_testdb |  
+-----+  
| employees        |  
+-----+  
1 row in set (0.12 sec)  
  
mysql> SELECT * FROM employees;  
+-----+  
| id | name   | age | department |  
+-----+  
| 1  | Alice  | 30  | HR         |  
| 2  | Bob    | 28  | Engineering|  
| 3  | Charlie| 35  | Finance    |  
+-----+  
3 rows in set (0.00 sec)  
  
mysql>
```

2. Use Sqoop's import command to transfer data from the MySQL employees table into HDFS, storing the output in /user/hadoop/employees\_data

```
--connect jdbc:mysql://localhost:3306/testdb \  
--username root \  
--password root \  
--table employees \  
--columns "id,name,age,department" \  
--m 1 \  
--target-dir /user/hadoop/employees_dataC
```

```
shubham-kelkar@shubham-kelkar:~$ sqoop import \  
--connect jdbc:mysql://localhost:3306/testdb?useSSL=false \  
--username root \  
--password root \  
--table employees \  
--m 1 \  
--target-dir /user/hadoop/employees_data  
Warning: /home/shubham-kelkar/sqoop/../hbase does not exist! HBase imports will fail.  
Please set $HBASE_HOME to the root of your HBase installation.  
Warning: /home/shubham-kelkar/sqoop/../hcatalog does not exist! HCatalog jobs will fail.  
Please set $HCAT_HOME to the root of your HCatalog installation.  
Warning: /home/shubham-kelkar/sqoop/../accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
Warning: /home/shubham-kelkar/sqoop/../zookeeper does not exist! Accumulo imports will fail.  
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.  
2025-03-21 21:21:33,510 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7  
2025-03-21 21:21:33,634 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
2025-03-21 21:21:33,838 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.  
2025-03-21 21:21:33,843 INFO tool.CodeGenTool: Beginning code generation  
2025-03-21 21:21:34,303 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employees` AS t LIMIT 1  
2025-03-21 21:21:34,361 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employees` AS t LIMIT 1  
2025-03-21 21:21:34,379 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop  
Note: /tmp/sqoop-shubham-kelkar/compile/030bf6e907ba9cec0721607e463b4c5e/employees.java uses or overrides a deprecated API.  
Note: Recompile with -Xlint:deprecation for details.  
2025-03-21 21:21:36,009 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-shubham-kelkar/compile/030bf6e907ba9cec0721607e463b4c5e/employees.jar  
2025-03-21 21:21:36,044 WARN manager.MySQLManager: It looks like you are importing from mysql.  
2025-03-21 21:21:36,045 WARN manager.MySQLManager: This transfer can be faster! Use the --direct  
2025-03-21 21:21:36,045 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.  
2025-03-21 21:21:36,045 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)  
2025-03-21 21:21:36,078 INFO mapreduce.ImportJobBase: Beginning import of employees  
2025-03-21 21:21:36,083 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.job.tracker address
```

```
Bytes Read=0  
File Output Format Counters  
Bytes Written=56  
2025-03-21 21:21:40,243 INFO mapred.LocalJobRunner: Finishing task: attempt_local1791197269_0001_m_000000_0  
2025-03-21 21:21:40,245 INFO mapred.LocalJobRunner: map task executor complete.  
2025-03-21 21:21:40,856 INFO mapreduce.Job: map 100% reduce 0%  
2025-03-21 21:21:40,856 INFO mapreduce.Job: Job job_local1791197269_0001 completed successfully  
2025-03-21 21:21:40,895 INFO mapreduce.Job: Counters: 21  
File System Counters  
FILE: Number of bytes read=6966  
FILE: Number of bytes written=676367  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=0  
HDFS: Number of bytes written=56  
HDFS: Number of read operations=6  
HDFS: Number of large read operations=0  
HDFS: Number of write operations=3  
HDFS: Number of bytes read erasure-coded=0  
Map-Reduce Framework  
Map input records=3  
Map output records=3  
Input split bytes=87  
Spilled Records=0  
Failed Shuffles=0  
Merged Map outputs=0  
GC time elapsed (ms)=0  
Total committed heap usage (bytes)=420478976  
File Input Format Counters  
Bytes Read=0  
File Output Format Counters  
Bytes Written=56  
2025-03-21 21:21:40,898 INFO mapreduce.ImportJobBase: Transferred 56 bytes in 3.5725 seconds (15.6754 bytes/sec)  
2025-03-21 21:21:40,900 INFO mapreduce.ImportJobBase: Retrieved 3 records.
```

3. List the HDFS directory and view the data file to confirm the data has been imported correctly

```
shubham-kelkar@shubham-kelkar:~$ hdfs dfs -cat /user/hadoop/employees_data/part-m-000000
1,Alice,30,HR
2,Bob,28,Engineering
3,Charlie,35,Finance
shubham-kelkar@shubham-kelkar:~$
```

4. Use Sqoop's export command to move the data from HDFS to the new MySQL table, specifying field delimiters and ensuring proper column mapping.

```
--connect jdbc:mysql://localhost:3306/testdb?useSSL=false \
--username root \
--password root \
--table employee_exported \
--export-dir /user/hadoop/employees_data \
--input-fields-terminated-by ',' \
--m 1 \
--columns "id,name,age,department"
```

```
shubham-kelkar@shubham-kelkar:~$ sqoop export \
--connect jdbc:mysql://localhost:3306/testdb?useSSL=false \
--username root \
--password root \
--table employee_exported \
--export-dir /user/hadoop/employees_data \
--input-fields-terminated-by ',' \
--m 1 \
--class-name employee_exported \
--jar-file /tmp/sqoop-shubham-kelkar/compile/7f8c6bc626229a934be8ac714c7c761e/employee_exported.jar
Warning: /home/shubham-kelkar/sqoop/./hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/shubham-kelkar/sqoop/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/shubham-kelkar/sqoop/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/shubham-kelkar/sqoop/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2025-03-21 23:15:00,062 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2025-03-21 23:15:00,209 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using
-P instead.
2025-03-21 23:15:00,480 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2025-03-21 23:15:00,489 INFO tool.CodeGenTool: Using existing jar: /tmp/sqoop-shubham-kelkar/compile/7f8c6bc626229a934be8ac714c7c761e/employee_exported.jar
2025-03-21 23:15:00,504 INFO mapreduce.ExportJobBase: Beginning export of employee_exported
2025-03-21 23:15:00,505 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtra
cker.address
2025-03-21 23:15:00,917 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2025-03-21 23:15:03,058 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instea
d, use mapreduce.reduce.speculative
2025-03-21 23:15:03,063 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead,
use mapreduce.map.speculative
2025-03-21 23:15:03,065 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2025-03-21 23:15:03,283 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-03-21 23:15:03,494 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-03-21 23:15:03,494 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-03-21 23:15:03,902 INFO input.FileInputFormat: Total input files to process : 1
```

5. Query the employee\_exported table in MySQL to verify that the data has been successfully exported.

```
shubham-kelkar@shubham-kelkar:~$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 80
Server version: 8.0.41-0ubuntu0.24.04.1 (Ubuntu)

Copyright (c) 2000, 2025, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> use testdb;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> SELECT * FROM employee_exported;
+-----+-----+-----+-----+
| id | name  | age | department |
+-----+-----+-----+-----+
| 1  | Alice | 30  | HR         |
| 2  | Bob   | 28  | Engineering|
| 3  | Charlie| 35  | Finance    |
+-----+-----+-----+-----+
3 rows in set (0.00 sec)
```