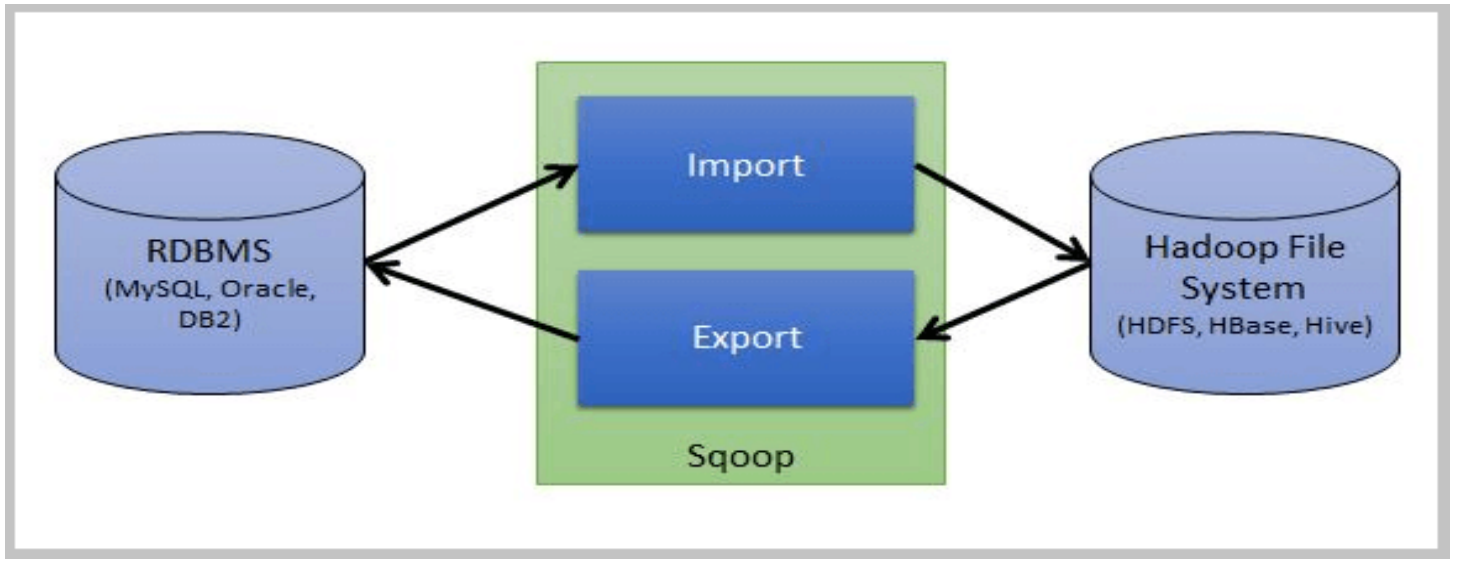
Apache

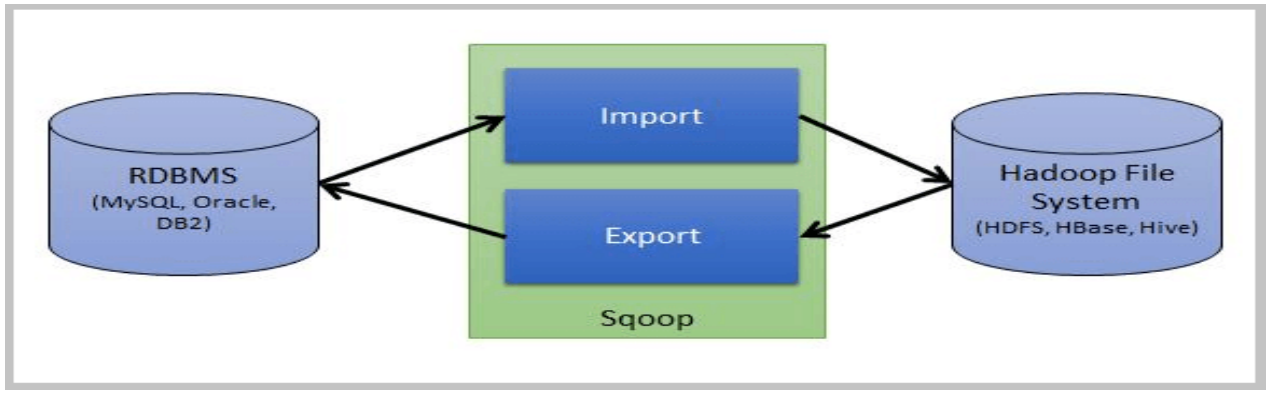
What is Sqoop?

Sqoop is a tool designed to import data from a relational database management system (RDBMS) such as MySQL or Oracle into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS.



Sqoop Import:-

The import tool imports individual tables from RDBMS to HDFS. Each row in a table is treated as a record in HDFS. All records are stored as text data in text files or as binary data in Avro and Sequence files.



Sqoop Export:-

The export tool exports a set of files from HDFS back to an RDBMS. The files given as input to Sqoop contain records, which are called as rows in table. Those are read and parsed into a set of records and delimited with user-specified delimiter.

How to install Sqoop?

- “Sqoop Installation” explains all the steps to install Sqoop on Ubuntu and we know **Sqoop** is **Hadoop**’s sub-project.
- **Prerequisites**

1. Verifying Java installation.
2. Verifying Hadoop installation.

Step by step installation:-

You can download the latest version of Sqoop from the following link For this tutorial, we are using version 1.4.5, that is, **sqoop-1.4.5.bin__hadoop-2.0.4-alpha.tar.gz**.

Step1:- Extract the Sqoop tar ball.

```
samiksha@samiksha-HP-Pavilion-Notebook:~$ tar -xvf sqoop-1.4.6.bin__hadoop-2.0.4-alpha.tar.gz
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/bin/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/conf/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/api/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/api/com/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/api/com/cloudera/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/api/com/cloudera/sqoop/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/api/com/cloudera/sqoop/lib/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/api/com/cloudera/sqoop/lib/class-use/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/api/org/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/api/org/apache/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/api/org/apache/sqoop/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/api/org/apache/sqoop/lib/
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/docs/api/org/apache/sqoop/lib/class-use/
```

```
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/src/scripts/run-perftest.sh
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/src/scripts/write-version-info.sh
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/testdata/hcatalog/conf/hive-log4j.properties
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/testdata/hcatalog/conf/hive-site.xml
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/testdata/hcatalog/conf/log4j.properties
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/testdata/hive/bin/hive
sqoop-1.4.6.bin__hadoop-2.0.4-alpha/testdata/hive/bin/hive.cmd
```

- Step2:- Configuring bashrc

```
samiksha@samiksha-HP-Pavilion-Notebook:~$ nano .bashrc
samiksha@samiksha-HP-Pavilion-Notebook:~$ source .bashrc
```

- Step3:- Configuring Sqoop:- To configure Sqoop with Hadoop, you need to edit the **sqoop-env.sh** file, which is placed in the **$SQOOP_HOME/conf** directory. First of all, Redirect to Sqoop config directory and copy the template file using the following command –

```
samiksha@samiksha-HP-Pavilion-Notebook:~$ cd $SQOOP_HOME/conf
samiksha@samiksha-HP-Pavilion-Notebook:~/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/conf$ mv sqoop-env-template.sh sqoop-env.sh
```

- Step4:- Download and Configure mysql-connector-java

We can download **mysql-connector-java-5.1.30.tar.gz** file from the following link.

The following commands are used to extract mysql-connector-java tarball and move **mysql-connector-java-5.1.30-bin.jar** to /sqoop/lib directory.

```
samiksha@samiksha-HP-Pavilion-Notebook:~$ tar -zxf mysql-connector-java-8.0.12.tar.gz
samiksha@samiksha-HP-Pavilion-Notebook:~$ cd mysql-connector-java-8.0.12
samiksha@samiksha-HP-Pavilion-Notebook:~/mysql-connector-java-8.0.12$ mv mysql-connector-java-8.0.12.jar /home/samiksha/sqoop-1.4.6.bin__hadoo
p-2.0.4-alpha/lib
```

- Step5:- Verifying Sqoop

```
samiksha@samiksha-HP-Pavilion-Notebook:~/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/bin$ cd
samiksha@samiksha-HP-Pavilion-Notebook:~$ cd $SQOOP_HOME/bin
```

```
samiksha@samiksha-HP-Pavilion-Notebook:~/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/bin$ sqoop-version
Warning: /home/samiksha/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/samiksha/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/samiksha/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/samiksha/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
18/08/23 11:53:13 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
Sqoop 1.4.6
git commit id c0c5a81723759fa575844a0a1eae8f510fa32c25
Compiled by root on Mon Apr 27 14:38:36 CST 2015
```

Sqoop installation is complete.

Installation of Mysql in Ubuntu:-

- To install it, simply update the package index on your server and install the default package with apt-get.

```
samiksha@samiksha-HP-Pavilion-Notebook:~$ sudo apt-get install mysql-server
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  mysql-client-5.7 mysql-server-5.7 mysql-server-core-5.7
```

- Configuring MySQL

```
samiksha@samiksha-HP-Pavilion-Notebook:~$ sudo mysql_secure_installation

Securing the MySQL server deployment.

Connecting to MySQL using a blank password.
```

There will be some query:-

1. VALIDATE PASSWORD PULGIN :- Press(y)
This allows the user to set only those passwords which are secure enough.

1. Then enter 0 for LOW
2. New Password:-
3. R-enter-Password:-
4. Remove anonymous users – press(no)
5. Disallow root login remotely? :- press(no)
6. Remove the test database:- press(yes)
7. Reload the privileges tables:- press(yes)

```
 - Removing privileges on test database...
Success.

Reloading the privilege tables will ensure that all changes
made so far will take effect immediately.

Reload privilege tables now? (Press y|Y for Yes, any other key for No) : y
Success.

All done!
```

- Testing MySQL

```
samiksha@samiksha-HP-Pavilion-Notebook:~$ systemctl status mysql.service
● mysql.service - MySQL Community Server
   Loaded: loaded (/lib/systemd/system/mysql.service; enabled; vendor preset: enabled)
   Active: active (running) since Thu 2018-08-23 13:58:39 IST; 3min 26s ago
  Process: 32734 ExecStart=/usr/sbin/mysqld --daemonize --pid-file=/run/mysqld/mysqld.pid (code=exited, status=0/SUCCESS)
  Process: 32725 ExecStartPre=/usr/share/mysql/mysql-systemd-start pre (code=exited, status=0/SUCCESS)
 Main PID: 32736 (mysqld)
    Tasks: 29 (limit: 4915)
   CGroup: /system.slice/mysql.service
           └─32736 /usr/sbin/mysqld --daemonize --pid-file=/run/mysqld/mysqld.pid

Aug 23 13:58:38 samiksha-HP-Pavilion-Notebook systemd[1]: Starting MySQL Community Server...
Aug 23 13:58:39 samiksha-HP-Pavilion-Notebook systemd[1]: Started MySQL Community Server.
```

- Start Mysql:- -u (user), -p (password)

```
samiksha@samiksha-HP-Pavilion-Notebook:~$ sudo mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 13
Server version: 5.7.23-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2018, Oracle and/or its affiliates. All rights reserved.

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>
```

Example :- For importing Mysql data to HDFS

- First create a database to mysql
- In mysql write Command:- show databases;

```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
4 rows in set (0.00 sec)

mysql>
```

- Command:- create database sample;
- Command:- show databases;

```
mysql> create database sample;
Query OK, 1 row affected (0.00 sec)

mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mysql              |
| performance_schema |
| sample             |
| sys                |
+--------------------+
5 rows in set (0.00 sec)
```

- Command:- use sample;

```
mysql> use sample;
Database changed
```

- Now, create table books

```
mysql> CREATE TABLE books( title VARCHAR(50) NOT NULL, price INT NOT NULL, language VARCHAR(50) DEFAULT "ENGLISH");
Query OK, 0 rows affected (0.65 sec)
```

- Command:- DESCRIBE books;

```
mysql> DESCRIBE books;
+----------+-------------+------+-----+---------+-------+
| Field    | Type        | Null | Key | Default | Extra |
+----------+-------------+------+-----+---------+-------+
| title    | varchar(50) | NO   |     | NULL    |       |
| price    | int(11)     | NO   |     | NULL    |       |
| language | varchar(50) | YES  |     | ENGLISH |       |
+----------+-------------+------+-----+---------+-------+
3 rows in set (0.01 sec)
```

- Now, insert the data into books

```
mysql> insert into books values("Five Littile Pigs",100.00,"English")
    -> ;
Query OK, 1 row affected (0.24 sec)

mysql> insert into books values("Rules",110.00,"English")
    -> ;
Query OK, 1 row affected (0.12 sec)

mysql> insert into books values("Gitanjali",110.00,"Hindi");
Query OK, 1 row affected (0.12 sec)

mysql> insert into books values("Kabir",220.00,"Hindi");
Query OK, 1 row affected (0.05 sec)

mysql> select * from books;
+-------------------+-------+----------+
| title             | price | language |
+-------------------+-------+----------+
| book title        |    50 | hindi    |
| Five Littile Pigs |   100 | English  |
| Rules             |   110 | English  |
| Gitanjali         |   110 | Hindi    |
| Kabir             |   220 | Hindi    |
+-------------------+-------+----------+
5 rows in set (0.00 sec)
```

- Create user on localhost

```
mysql> CREATE USER  'samiksha'@'localhost' IDENTIFIED BY 'Abc@0307'
    -> ;
Query OK, 0 rows affected (0.00 sec)
```

- Grant all the privileges

```
mysql> GRANT ALL PRIVILEGES ON *.* TO 'samiksha'@'localhost' WITH GRANT OPTION;
Query OK, 0 rows affected (0.00 sec)

mysql> EXIT
```

- Command:- ssh localhost
- Start Hadoop by Command :- ./start-all.sh

```
samiksha@samiksha-HP-Pavilion-Notebook:~$ cd /home/samiksha/hadoop-2.7.3/sbin
samiksha@samiksha-HP-Pavilion-Notebook:~/hadoop-2.7.3/sbin$ ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
samiksha@localhost's password:
localhost: starting namenode, logging to /home/samiksha/hadoop-2.7.3/logs/hadoop-samiksha-nameno
samiksha@localhost's password:
localhost: starting datanode, logging to /home/samiksha/hadoop-2.7.3/logs/hadoop-samiksha-datano
Starting secondary namenodes [0.0.0.0]
samiksha@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /home/samiksha/hadoop-2.7.3/logs/hadoop-samiksha
```

Import data to HDFS

As we created database **sample** and in that database we created table **books** now, we will import this table to hdfs

```
samiksha@samiksha-HP-Pavilion-Notebook:~/hadoop-2.7.3/sbin$  sqoop import --connect "jdbc:mysql://localhost/sample" --username samiksha --pass
word "Abc@0307" --table books -m 1 --target-dir /user/samiksha/sqoop
Warning: /home/samiksha/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/samiksha/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/samiksha/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/samiksha/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
18/08/29 18:24:53 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
18/08/29 18:24:53 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
18/08/29 18:24:53 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
18/08/29 18:24:53 INFO tool.CodeGenTool: Beginning code generation
Wed Aug 29 18:24:53 IST 2018 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 existi
ng applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting use
SSL=false, or set useSSL=true and provide truststore for server certificate verification.
18/08/29 18:24:53 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `books` AS t LIMIT 1
18/08/29 18:24:53 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `books` AS t LIMIT 1
18/08/29 18:24:53 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/samiksha/hadoop-2.7.3
Note: /tmp/sqoop-samiksha/compile/ae558539e2825e01af61813936a39990/books.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
18/08/29 18:24:56 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-samiksha/compile/ae558539e2825e01af61813936a39990/books.jar
18/08/29 18:24:56 WARN manager.MySQLManager: It looks like you are importing from mysql.
18/08/29 18:24:56 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
18/08/29 18:24:56 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
```

```
                Total committed heap usage (bytes)=148375504
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=20
18/08/29 18:26:55 INFO mapreduce.ImportJobBase: Transferred 20 bytes in 24.9339 seconds (0.8021 bytes/sec)
18/08/29 18:26:55 INFO mapreduce.ImportJobBase: Retrieved 1 records.
samiksha@samiksha-HP-Pavilion-Notebook:~/hadoop-2.7.3/sbin$
```

You can see the output

Browse Directory

/ Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	samiksha	supergroup	0 B	8/15/2018, 12:56:08 AM	0	0 B	Pig_Dat
drwxr-xr-x	samiksha	supergroup	0 B	8/15/2018, 4:13:59 AM	0	0 B	Pig_Data
drwxr-xr-x	cloudera	supergroup	0 B	7/25/2018, 5:20:33 AM	0	0 B	home
drwxr-xr-x	samiksha	supergroup	0 B	8/29/2018, 6:26:52 PM	0	0 B	sqoop
drwx------	samiksha	supergroup	0 B	8/15/2018, 1:06:54 AM	0	0 B	tmp
drwxr-xr-x	samiksha	supergroup	0 B	7/25/2018, 4:27:13 AM	0	0 B	user

Browse Directory

/sqoop Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	samiksha	supergroup	0 B	8/29/2018, 6:26:52 PM	1	128 MB	_SUCCESS
-rw-r--r--	samiksha	supergroup	20 B	8/29/2018, 6:26:52 PM	1	128 MB	part-m-00000

Hadoop, 2016.

Here, you can download and check the output

- Or, you can check like this also

```
samiksha@samiksha-HP-Pavilion-Notebook:~/hadoop-2.7.3/sbin$ hadoop fs -cat /sqoop/*
Five Littile Pigs,100,English
Rules,110,English
Gitanjali,110,Hindi
Kabir,220,Hindi
samiksha@samiksha-HP-Pavilion-Notebook:~/hadoop-2.7.3/sbin$
```

So, we successfully import the data into HDFS.

THANK YOU