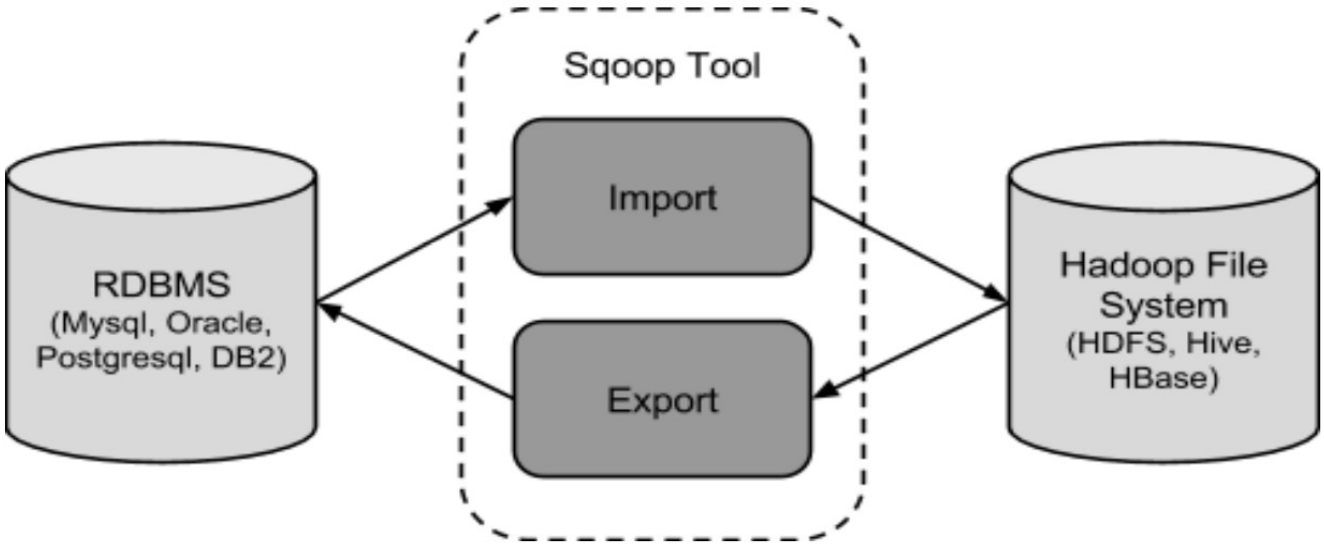
SQOOP

- ***Sqoop:*** *It is used to import and export data to and from between HDFS and RDBMS.*
- ***Pig:*** *It is a procedural language platform used to develop a script for MapReduce operations.*
- ***Hbase:*** *HBase is a distri column-oriented database built on top of the Hadoop file system.*
- ***Hive:*** *It is a platform used to develop SQL type scripts to do MapReduce operations.*
- ***Flume:*** *Used to handle streaming data on the top of Hadoop.*
- ***Oozie:*** *Apache Oozie is a workflow scheduler for Hadoop.*

How Sqoop Works? The following image describes the workflow of Sqoop.



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.*

Importing data from MySQL to HDFS

In order to store data into HDFS, we make use of Apache Hive which provides an SQL-like interface between the user and the Hadoop distributed file system (HDFS) which integrates Hadoop. We perform the following steps:

Step 1: *Login into MySQL*

mysql -u root -p

```
Applications  Places  System                                   Mon Sep 7, 11:21 PM   clouder
cloudera@quickstart:~                                          Click to view your appointments and tasks
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ mysql -u root -pcloudera
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 6
Server version: 5.1.66 Source distribution

Copyright (c) 2000, 2012, 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>
```

Step 2: *Create a database and table and insert data.*

create database library;
create table books(author_name varchar(65), total_no_of_articles int, phone_no int, address varchar(65));
insert into books values("Rohan",10,123456789,"Lucknow");

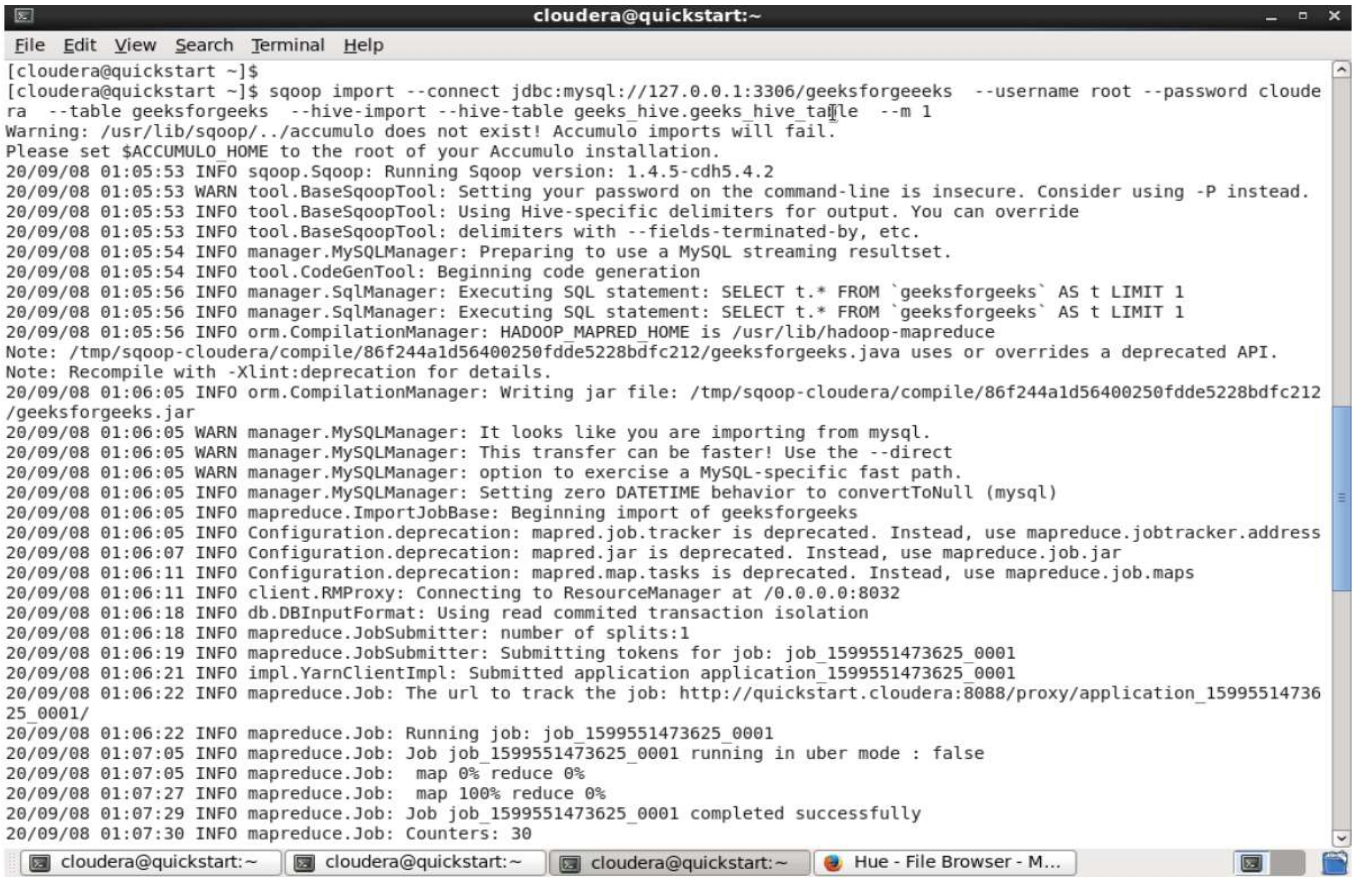
```
+-------------+----------------------+-----------+----------+
| author_name | total_no_of_articles | phone_no  | address  |
+-------------+----------------------+-----------+----------+
| Rohan       |                   10 | 123456789 | Lucknow  |
| McCallan    |                  250 | 234567890 | New York |
| Palak       |                   50 | 345678901 | Delhi    |
| Arjun       |                  120 | 456789012 | Mumbai   |
| Robert      |                 1000 | 567890123 | Texas    |
+-------------+----------------------+-----------+----------+
5 rows in set (0.00 sec)

mysql>
```

Step 3: *Create a database and table in the hive where data should be imported.*

create table books_hive_table(name string, total_articles int, phone_no int, address string) row format delimited fields terminated by ';';

Step 4: *Run below the import command on Hadoop. Step 4: Run below the import command on Hadoop.*

*sqoop import --connect *
*jdbc:mysql://127.0.0.1:3306/database_name_in_mysql *
*--username root --password cloudera *
*--table table_name_in_mysql *
*--hive-import --hive-table database_name_in_hive.table_name_in_hive *
--m 1



```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://127.0.0.1:3306/geeksforgeeeks  --username root --password cloude
ra  --table geeksforgeeks  --hive-import --hive-table geeks_hive.geeks_hive_table  --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.
20/09/08 01:05:53 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.2
20/09/08 01:05:53 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/09/08 01:05:53 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
20/09/08 01:05:53 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
20/09/08 01:05:54 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
20/09/08 01:05:54 INFO tool.CodeGenTool: Beginning code generation
20/09/08 01:05:56 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `geeksforgeeks` AS t LIMIT 1
20/09/08 01:05:56 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `geeksforgeeks` AS t LIMIT 1
20/09/08 01:05:56 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/86f244a1d56400250fdde5228bdfc212/geeksforgeeks.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
20/09/08 01:06:05 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/86f244a1d56400250fdde5228bdfc212
/geeksforgeeks.jar
20/09/08 01:06:05 WARN manager.MySQLManager: It looks like you are importing from mysql.
20/09/08 01:06:05 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
20/09/08 01:06:05 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
20/09/08 01:06:05 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
20/09/08 01:06:05 INFO mapreduce.ImportJobBase: Beginning import of geeksforgeeks
20/09/08 01:06:05 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
20/09/08 01:06:07 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
20/09/08 01:06:11 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
20/09/08 01:06:11 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/09/08 01:06:18 INFO db.DBInputFormat: Using read commited transaction isolation
20/09/08 01:06:18 INFO mapreduce.JobSubmitter: number of splits:1
20/09/08 01:06:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1599551473625_0001
20/09/08 01:06:21 INFO impl.YarnClientImpl: Submitted application application_1599551473625_0001
20/09/08 01:06:22 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_15995514736
25_0001/
20/09/08 01:06:22 INFO mapreduce.Job: Running job: job_1599551473625_0001
20/09/08 01:07:05 INFO mapreduce.Job: Job job_1599551473625_0001 running in uber mode : false
20/09/08 01:07:05 INFO mapreduce.Job:  map 0% reduce 0%
20/09/08 01:07:27 INFO mapreduce.Job:  map 100% reduce 0%
20/09/08 01:07:29 INFO mapreduce.Job: Job job_1599551473625_0001 completed successfully
20/09/08 01:07:30 INFO mapreduce.Job: Counters: 30
cloudera@quickstart:~   cloudera@quickstart:~   cloudera@quickstart:~   Hue - File Browser - M...
```

Step 5: *Check-in hive if data is imported successfully or not*

```
hive> select * from geeks_hive_table;
OK
Time taken: 0.831 seconds
hive> select * from geeks_hive_table;
OK
Rohan   10      123456789       Lucknow
McCallan        250     234567890       New York
Palak   50      345678901       Delhi
Arjun   120     456789012       Mumbai
Robert  1000    567890123       Texas
Time taken: 0.191 seconds, Fetched: 5 row(s)
hive>
```

cloudera@quickstart:~ | cloudera@quickstart:~ | cloudera@quickstart:~ | Hue - File Browser - M...

Exporting data from HDFS to MySQL

To export data into MySQL from HDFS, perform the following steps:

Step 1: *Create a database and table in the hive.*

create table hive_table_export(name string,company string, phone int, age int) row format delimited fields terminated by ';';

Applications Places System

Browse and run installed applications

cloudera@quickstart:~

File Edit View Search Terminal Help

```
hive> create database hive_export;
OK
Time taken: 0.083 seconds
hive> use hive_export;
OK
Time taken: 0.034 seconds
hive> create table hive_table_export(name string,company string, phone int, age int);
OK
Time taken: 0.164 seconds
hive>
```

Step 2: *Insert data into the hive table.*

insert into hive_table_export values("Ritik","Amazon",234567891,35);

Applications Places System

Browse and run installed applications

File Edit View Search Terminal Help

```
hive> select * from hive_table_export;
OK
Rohan   GeeksforGeeks   123456789       21
Ritik   Amazon  234567891       35
Pratham Barclays        345678912       27
Time taken: 0.17 seconds, Fetched: 3 row(s)
hive>
```

Step 3: *Create a database and table in MySQL in which data should be exported.*

```
Applications  Places  System
Browse and run installed applications        cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
mysql> create database mysql_export;
Query OK, 1 row affected (0.00 sec)

mysql> use mysql_export;
Database changed
mysql> create table mysql_table_export(name varchar(65),company varchar(65),phone int, age int);
Query OK, 0 rows affected (0.02 sec)

mysql> ▌
```

Step 4: *Run the following command on Hadoop.*

*sqoop export --connect *
*jdbc:mysql://127.0.0.1:3306/database_name_in_mysql *
*--table table_name_in_mysql *
*--username root --password cloudera *
*--export-dir /user/hive/warehouse/hive_database_name.db/table_name_in_hive *
*--m 1 *
-- driver com.mysql.jdbc.Driver
--input-fields-terminated-by ','

```
[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://127.0.0.1:3306/mysql_export  --table mysql_table_export  --usern
ame root --password cloudera  --export-dir /user/hive/warehouse/hive_export.db/hive_table_export  --m 1  --driver com.mysql.j
dbc.Driver --input-fields-terminated-by ','
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/09/08 02:10:05 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.2
20/09/08 02:10:05 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/09/08 02:10:05 WARN sqoop.ConnFactory: Parameter --driver is set to an explicit driver however appropriate connection mana
ger is not being set (via --connection-manager). Sqoop is going to fall back to org.apache.sqoop.manager.GenericJdbcManager.
Please specify explicitly which connection manager should be used next time.
20/09/08 02:10:06 INFO manager.SqlManager: Using default fetchSize of 1000
20/09/08 02:10:06 INFO tool.CodeGenTool: Beginning code generation
20/09/08 02:10:08 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM mysql_table_export AS t WHERE 1=0
20/09/08 02:10:08 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM mysql_table_export AS t WHERE 1=0
20/09/08 02:10:08 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/3337bf5a79cf6ef945aa0f7d87de28a4/mysql_table_export.java uses or overrides a deprecated API
.
Note: Recompile with -Xlint:deprecation for details.
20/09/08 02:10:17 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/3337bf5a79cf6ef945aa0f7d87de28a4
/mysql_table_export.jar
20/09/08 02:10:17 INFO mapreduce.ExportJobBase: Beginning export of mysql_table_export
20/09/08 02:10:17 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
20/09/08 02:10:18 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
20/09/08 02:10:23 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM mysql_table_export AS t WHERE 1=0
20/09/08 02:10:23 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapre
duce.reduce.speculative
20/09/08 02:10:23 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduc
e.map.speculative
20/09/08 02:10:23 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
20/09/08 02:10:23 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/09/08 02:10:28 INFO input.FileInputFormat: Total input paths to process : 3
20/09/08 02:10:28 INFO input.FileInputFormat: Total input paths to process : 3
20/09/08 02:10:28 INFO mapreduce.JobSubmitter: number of splits:1
20/09/08 02:10:28 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduc
e.map.speculative
20/09/08 02:10:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1599551473625_0010
20/09/08 02:10:31 INFO impl.YarnClientImpl: Submitted application application_1599551473625_0010
cloudera@quickstart:~  cloudera@quickstart:~  cloudera@quickstart:~  Hue - File Browser - M...
```

Step 5: *Check-in MySQL if data is exported successfully or not.*

```
mysql> select * from mysql_table_export;
+---------+---------------+-----------+------+
| name    | company       | phone     | age  |
+---------+---------------+-----------+------+
| Rohan   | GeeksforGeeks | 123456789 |   21 |
| Ritik   | Amazon        | 234567891 |   35 |
| Pratham | Barclays      | 345678912 |   27 |
+---------+---------------+-----------+------+
3 rows in set (0.00 sec)

mysql> ▌
```