
Big Data Analytics and Reasoning - Practice 05

Giuseppe Mazzotta

Sqoop

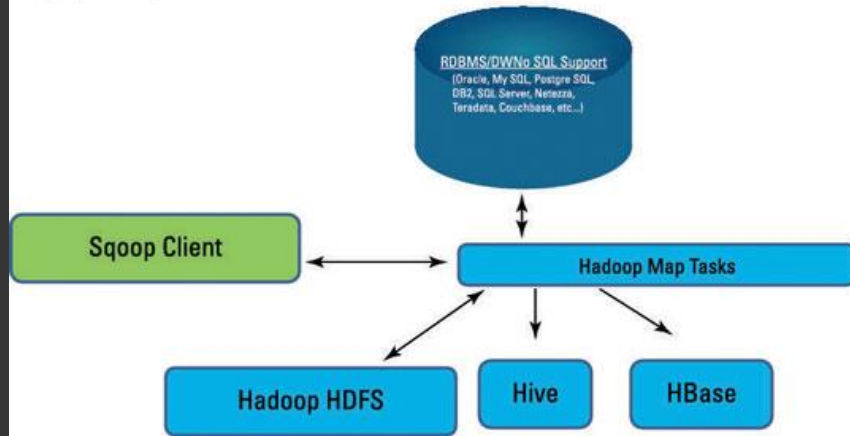
Sqoop (SQL to Hadoop) is tool designed for efficiently transferring large amount of data between external sources and hadoop

ETL tool (Extract Transform Load)

Use mapreduce engine for transferring data in parallel



Sqoop Design



Tip

Number of mappers
could affect data source

General Overview

Sqoop can import and export data between RDBMS and Hadoop.

Main sqoop commands:

- Import
- Export

Sqoop fetch metadata from the RDBMS about the input/output table

By using metadata sqoop will generate and compile a java class used for importing/exporting data

Sqoop installation

Sqoop is a client tool that will act as an hdfs user

Download the binary archive of a sqoop distribution (1.4.7) from the official website

Unfold the downloaded archive

Export the SQOOP_HOME

Add SQOOP_HOME/bin to the PATH environment variable

A screenshot of a web browser showing the 'Index of /dist/sqoop/' page from archive.apache.org. The browser's address bar shows the URL 'archive.apache.org/dist/sqoop/'. The page title is 'Index of /dist/sqoop/'. Below the title is a table with four columns: 'Name', 'Last modified', 'Size', and 'Description'. The table lists various Sqoop distribution versions as directories, including '1.4.0-incubating/', '1.4.1-incubating/', '1.4.2/', '1.4.3/', '1.4.4/', '1.4.5/', '1.4.6/', '1.4.7/', '1.99.1/', '1.99.2/', '1.99.3/', '1.99.4/', '1.99.5/', '1.99.6/', '1.99.7/', and a file named 'KEYS'. The 'Last modified' column shows dates and times for each entry. The 'Size' column shows dashes for directories and '69K' for the 'KEYS' file. The 'Description' column is empty for all entries.

Name	Last modified	Size	Description
 Parent Directory		-	
 1.4.0-incubating/	2012-06-13 23:13	-	
 1.4.1-incubating/	2013-07-30 22:11	-	
 1.4.2/	2013-07-30 22:11	-	
 1.4.3/	2013-07-30 22:11	-	
 1.4.4/	2013-07-30 22:11	-	
 1.4.5/	2014-08-11 20:18	-	
 1.4.6/	2017-10-04 11:10	-	
 1.4.7/	2020-07-06 15:20	-	
 1.99.1/	2012-12-24 15:16	-	
 1.99.2/	2013-04-23 21:48	-	
 1.99.3/	2013-11-02 15:33	-	
 1.99.4/	2014-11-24 22:18	-	
 1.99.5/	2015-02-24 20:14	-	
 1.99.6/	2015-05-05 22:34	-	
 1.99.7/	2020-07-06 15:20	-	
 KEYS	2020-07-06 15:19	69K	



1. Required Libraries

Sqoop internally use different libraries. Some of these have to be added manually

→ **commons-lang-2.6**

Used internally by sqoop

→ **mysql-connector**

Used to communicate with mysql server (this could be different depending on the RDBMS).

→ **hive-common**

Used to import data into hive

Let us proceed by examples!



Tip

Remember. Sqoop is a client tool, it doesn't need to be installed on the cluster machines

Sqoop acts as hadoop client -> it has to be able to run hadoop command