Big Data Analytics and Reasoning - Practice 05

Giuseppe Mazzotta

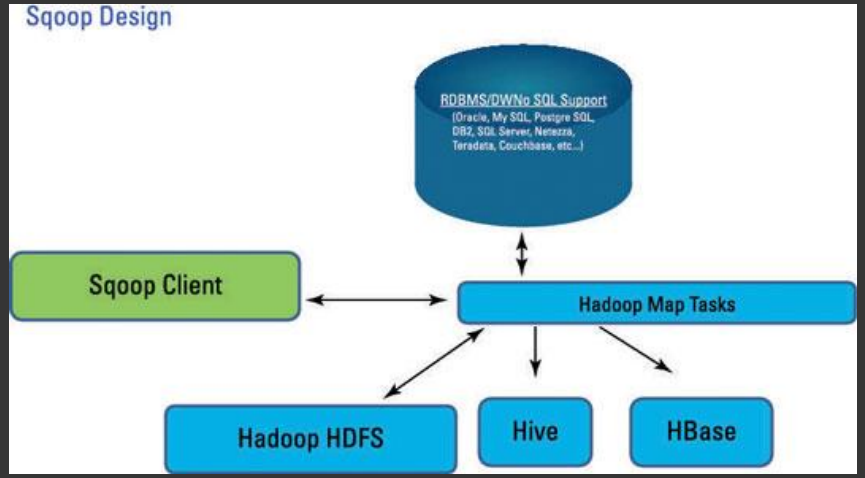
Sqoop

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

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

Tip

Number of mappers could affect data source

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

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