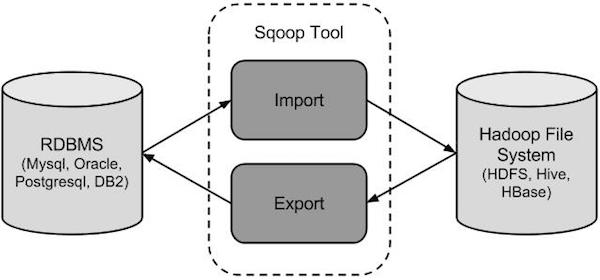
# sqoop数据迁移工具

## 2.1 概述

sqoop是apache旗下一款**“Hadoop和关系数据库服务器之间传送数据”**的工具。

**导入数据**：MySQL，Oracle导入数据到Hadoop的HDFS、HIVE、HBASE等数据存储系统；

**导出数据：**从Hadoop的文件系统中导出数据到关系数据库mysql等



## 2.2 工作机制

将导入或导出命令翻译成mapreduce程序来实现

在翻译出的mapreduce中主要是对inputformat和outputformat进行定制

## 2.3 sqoop实战及原理

### 2.3.1 sqoop安装

安装sqoop的前提是已经具备java和hadoop的环境

#### 1、下载并解压

最新版下载地址http://ftp.wayne.edu/apache/sqoop/1.4.6/

#### 2、修改配置文件

$ cd $SQOOP\_HOME/conf

$ mv sqoop-env-template.sh sqoop-env.sh

打开**sqoop-env.sh**并编辑下面几行：

export HADOOP\_COMMON\_HOME=/home/hadoop/apps/hadoop-2.6.1/

export HADOOP\_MAPRED\_HOME=/home/hadoop/apps/hadoop-2.6.1/

export HIVE\_HOME=/home/hadoop/apps/hive-1.2.1

#### 加入mysql的jdbc驱动包

cp ~/app/hive/lib/mysql-connector-java-5.1.28.jar $SQOOP\_HOME/lib/

#### 4、验证启动

$ cd $SQOOP\_HOME/bin

$ sqoop-version

到这里，整个Sqoop安装工作完成。

验证sqoop到mysql业务库之间的连通性：

bin/sqoop-list-databases --connect jdbc:mysql://localhost:3306 --username root --password root

bin/sqoop-list-tables --connect jdbc:mysql://localhost:3306/userdb --username root --password root

## 2.4 Sqoop的数据导入

“导入工具”导入单个表从RDBMS到HDFS。表中的每一行被视为HDFS的记录。所有记录都存储为文本文件的文本数据（或者Avro、sequence文件等二进制数据）

### 2.4.1 语法

下面的语法用于将数据导入HDFS。

|  |
| --- |
| $ sqoop import (generic-args) (import-args) |

### 2.4.2 示例

#### 导入表表数据到HDFS

下面的命令用于从MySQL数据库服务器中的emp表导入HDFS。

|  |
| --- |
| bin/sqoop import \  --connect jdbc:mysql://hdp-node-01:3306/test \  --username root \  --password root \  --table emp \  --m 1 |

为了验证在HDFS导入的数据，请使用以下命令查看导入的数据

|  |
| --- |
| $ $HADOOP\_HOME/bin/hadoop fs -cat /user/hadoop/emp/part-m-00000 |

#### 导入到HDFS指定目录

在导入表数据到HDFS使用Sqoop导入工具，我们可以指定目标目录。

以下是指定目标目录选项的Sqoop导入命令的语法。

|  |
| --- |
| --target-dir <new or exist directory in HDFS> |

下面的命令是用来导入emp表数据到'/queryresult'目录。

|  |
| --- |
| bin/sqoop import \  --connect jdbc:mysql://hdp-node-01:3306/test \  --username root \  --password root \  --target-dir /queryresult \  --fields-terminated-by ‘\001’ \  --table emp  --split-by id  --m 1 |

注意：如果报错，说emp类找不到，则可以手动从sqoop生成的编译目录(/tmp/sqoop-root/compile)中，找到这个emp.class和emp.jar，拷贝到sqoop的lib目录下：

如果设置了 --m 1，则意味着只会启动一个maptask执行数据导入

如果不设置 --m 1，则默认为启动4个map task执行数据导入，则需要指定一个列来作为划分map task任务的依据

下面的命令是用来验证 /queryresult 目录中 emp表导入的数据形式。

|  |
| --- |
| $HADOOP\_HOME/bin/hadoop fs -cat /queryresult/part-m-\* |

#### 导入关系表到HIVE

|  |
| --- |
| bin/sqoop import --connect jdbc:mysql://hdp-node-01:3306/test --username root --password root --table emp --hive-import --split-by id --m 1 |

#### 导入表数据子集

我们可以导入表的使用Sqoop导入工具，"where"子句的一个子集。它执行在各自的数据库服务器相应的SQL查询，并将结果存储在HDFS的目标目录。

where子句的语法如下。

|  |
| --- |
| --where <condition> |

下面的命令用来导入emp表数据的子集。子集查询检索员工ID和地址，居住城市为：Secunderabad

|  |
| --- |
| bin/sqoop import \  --connect jdbc:mysql://hdp-node-01:3306/test \  --username root \  --password root \  --where "city ='sec-bad'" \  --target-dir /wherequery \  --table emp\_add \  --m 1 |

按需导入

|  |
| --- |
| bin/sqoop import \  --connect jdbc:mysql://hdp-node-01:3306/test \  --username root \  --password root \  --target-dir /wherequery2 \  --query 'select id,name,deg from emp WHERE id>1207 **and $CONDITIONS'** \  --split-by id \  --fields-terminated-by '\t' \  --m 2 |

下面的命令用来验证数据从emp\_add表导入/wherequery目录

|  |
| --- |
| $HADOOP\_HOME/bin/hadoop fs -cat /wherequery/part-m-\* |

#### 增量导入

增量导入是仅导入新添加的表中的行的技术。

sqoop支持两种增量MySql导入到hive的模式，

一种是append，即通过指定一个递增的列，比如：

--incremental append --check-column num\_id --last-value 0

另种是可以根据时间戳，比如：

--incremental lastmodified --check-column created --last-value '2012-02-01 11:0:00'

就是只导入created 比'2012-02-01 11:0:00'更大的数据。

**1/ append模式**

它需要添加‘incremental’, ‘check-column’, 和 ‘last-value’选项来执行增量导入。

下面的语法用于Sqoop导入命令增量选项。

|  |
| --- |
| --incremental <mode>  --check-column <column name>  --last value <last check column value> |

假设新添加的数据转换成emp表如下：

1206, satish p, grp des, 20000, GR

下面的命令用于在EMP表执行增量导入。

|  |
| --- |
| bin/sqoop import \  --connect jdbc:mysql://hdp-node-01:3306/test \  --username root \  --password root \  --table emp --m 1 \  --incremental append \  --check-column id \  --last-value 1208 |

以下命令用于从emp表导入HDFS emp/ 目录的数据验证。

下面的命令是从表emp 用来查看修改或新添加的行

|  |
| --- |
| $HADOOP\_HOME/bin/hadoop fs -cat /emp/part-m-\*1  1206, satish p, grp des, 20000, GR |

## 2.5 Sqoop的数据导出

**1/ 将数据从HDFS把文件导出到RDBMS数据库**

导出前，目标表必须存在于目标数据库中。

* 默认操作是从将文件中的数据使用INSERT语句插入到表中
* 更新模式下，是生成UPDATE语句更新表数据

### 语法

以下是export命令语法。

|  |
| --- |
| $ sqoop export (generic-args) (export-args) |

1、首先需要手动创建mysql中的目标表

|  |
| --- |
| $ mysql  mysql> USE db;  mysql> CREATE TABLE employee (  id INT NOT NULL PRIMARY KEY,  name VARCHAR(20),  deg VARCHAR(20),  salary INT,  dept VARCHAR(10)); |

2、然后执行导出命令

|  |
| --- |
| bin/sqoop export \  --connect jdbc:mysql://hdp-node-01:3306/test \  --username root \  --password root \  --table employee \  --export-dir /user/hadoop/emp/ |