

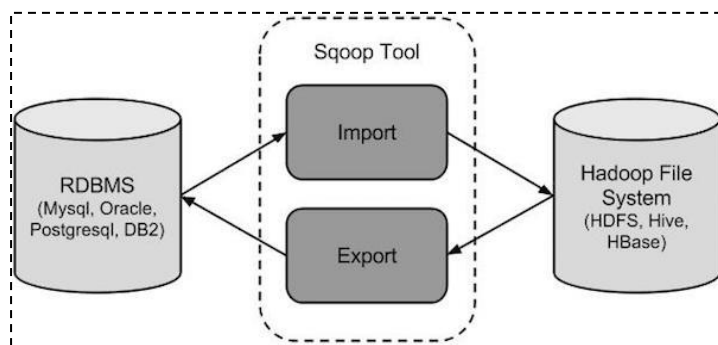
sqoop 数据迁移

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。

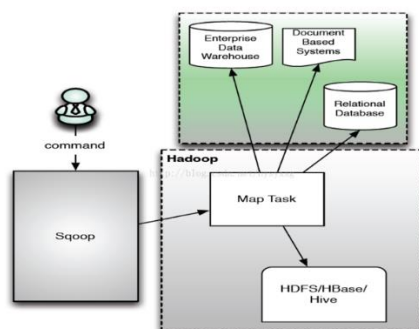
导入数据: MySQL, Oracle 导入数据到 Hadoop 的 HDFS、HIVE、HBASE 等数据存储系统;

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

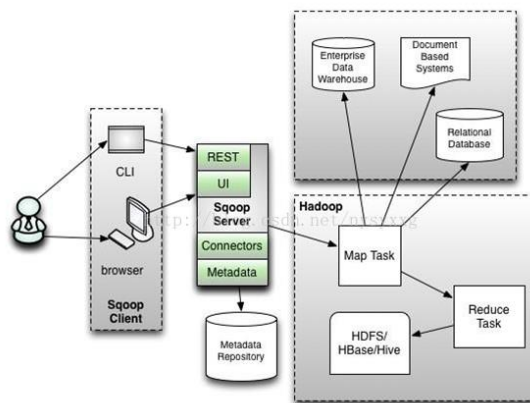


sqoop1 与 sqoop2 架构对比

sqoop1 架构



sqoop2 架构



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

sqoop 实战及原理

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

下载地址

<http://archive.cloudera.com/cdh5/cdh/5/>

sqoop1 版本详细下载地址

<http://archive.cloudera.com/cdh5/cdh/5/sqoop-1.4.6-cdh5.14.0.tar.gz>

sqoop2 版本详细下载地址

<http://archive.cloudera.com/cdh5/cdh/5/sqoop2-1.99.5-cdh5.14.0.tar.gz>

我们这里使用 sqoop1 的版本，下载之后上传到/export/softwares 目录下，然后进行解压

```
cd /export/softwares
tar -zxvf sqoop-1.4.6-cdh5.14.0.tar.gz -C ../servers/
```

修改配置文件

```
cd /export/servers/sqoop-1.4.6-cdh5.14.0/conf/
cp sqoop-env-template.sh sqoop-env.sh
vim sqoop-env.sh
```

```
export HADOOP_COMMON_HOME=/export/servers/hadoop-2.6.0-cdh5.14.0
export HADOOP_MAPRED_HOME=/export/servers/hadoop-2.6.0-cdh5.14.0
export HIVE_HOME=/export/servers/hive-1.1.0-cdh5.14.0
```


加入额外的依赖包


sqoop 的使用需要添加两个额外的依赖包，一个是 mysql 的驱动包，一个是 java-json 的依赖包，不然就会报错

mysql-connector-java-5.1.40.jar

java-json.jar

名称

 java-json.jar

 mysql-connector-java-5.1.40.jar

将这个两个 jar 包添加到 sqoop 的 lib 目录下

验证启动

```
cd /export/servers/sqoop-1.4.6-cdh5.14.0
bin/sqoop-version
```

```
[root@node03 sqoop-1.4.6-cdh5.14.0]# bin/sqoop-version
Warning: /export/servers/sqoop-1.4.6-cdh5.14.0/bin/../../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /export/servers/sqoop-1.4.6-cdh5.14.0/bin/../../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /export/servers/sqoop-1.4.6-cdh5.14.0/bin/../../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /export/servers/sqoop-1.4.6-cdh5.14.0/bin/../../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
19/07/23 13:14:31 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.14.0
sqoop 1.4.6-cdh5.14.0
git commit id
compiled by jenkins on Sat Jan 6 13:24:40 PST 2018
```

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）

列举出所有的数据库

命令行查看帮助

```
bin/sqoop list-databases --help
```

列出 win10 主机所有的数据库:

开启本地数据库远程连接权限:

```
GRANT ALL PRIVILEGES ON *.* TO 'root'@'%' IDENTIFIED BY '123' WITH  
GRANT OPTION;  
flush privileges;
```



```
bin/sqoop list-databases --connect jdbc:mysql://192.168.52.120:3306/ --  
username root --password 123456
```

```
bin/sqoop list-databases --connect jdbc:mysql://10.6.67.200:3306/ --username  
root --password 123
```

查看某一个数据库下面的所有数据表

```
bin/sqoop list-tables --connect jdbc:mysql://10.6.67.200:3306/test00 --  
username root --password 123
```

导入数据库表数据到 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://10.6.67.200:3306/test00 --password  
123 --username root --table emp --m 1
```

如果成功执行，那么会得到下面的输出。

```

19/07/23 13:38:06 INFO orm.compilationManager: writing jar file: /tmp/sqoop-root/compile/d052b32aef49efa9c5ec2c605b1
90b57/emp.jar
19/07/23 13:38:06 WARN manager.MySQLManager: It looks like you are importing from mysql.
19/07/23 13:38:06 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
19/07/23 13:38:06 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
19/07/23 13:38:06 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
19/07/23 13:38:06 INFO mapreduce.ImportJobBase: Beginning import of emp
19/07/23 13:38:07 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
19/07/23 13:38:10 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
19/07/23 13:38:10 INFO client.RMProxy: Connecting to ResourceManager at node01/192.168.52.100:8032
19/07/23 13:38:22 INFO db.DBInputFormat: Using read committed transaction isolation
19/07/23 13:38:22 INFO mapreduce.JobSubmitter: number of splits:1
19/07/23 13:38:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1563824945492_0001
19/07/23 13:38:25 INFO impl.YarnClientImpl: Submitted application application_1563824945492_0001
19/07/23 13:38:25 INFO mapreduce.Job: The url to track the job: http://node01:8088/proxy/application_1563824945492_0001/
19/07/23 13:38:25 INFO mapreduce.Job: Running job: job_1563824945492_0001
19/07/23 13:38:59 INFO mapreduce.Job: Job job_1563824945492_0001 running in uber mode : true
19/07/23 13:38:59 INFO mapreduce.Job: map 0% reduce 0%
19/07/23 13:39:03 INFO mapreduce.Job: map 100% reduce 0%
19/07/23 13:39:04 INFO mapreduce.Job: Job job_1563824945492_0001 completed successfully
19/07/23 13:39:05 INFO mapreduce.Job: Counters: 32

```

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

```
hdfs dfs -ls /user/root/emp
```

```

[root@node03 sqoop-1.4.6-cdh5.14.0]# hdfs dfs -ls /user/root/emp
Found 2 items
-rw-r--r--  3 root supergroup          0 2019-07-23 13:39 /user/root/emp/_SUCCESS
-rw-r--r--  3 root supergroup        809 2019-07-23 13:39 /user/root/emp/part-m-00000
You have new mail in /var/spool/mail/root

```

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。

使用参数 `--target-dir` 来指定导出目的地，

使用参数 `--delete-target-dir` 来判断导出目录是否存在，如果存在就删掉

```
bin/sqoop import --connect jdbc:mysql://10.6.67.200/test00 --username root
--password 123 --delete-target-dir --table emp --target-dir /sqoop/emp --m 1
```

```

19/07/23 13:42:39 INFO tool.ImportTool: Destination directory /sqoop/emp is not present, hence not deleting.
19/07/23 13:42:39 WARN manager.MySQLManager: It looks like you are importing from mysql.
19/07/23 13:42:39 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
19/07/23 13:42:39 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
19/07/23 13:42:39 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
19/07/23 13:42:39 INFO mapreduce.ImportJobBase: Beginning import of emp
19/07/23 13:42:40 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
19/07/23 13:42:40 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
19/07/23 13:42:40 INFO client.RMProxy: Connecting to ResourceManager at node01/192.168.52.100:8032
19/07/23 13:42:48 INFO db.DBInputFormat: Using read committed transaction isolation
19/07/23 13:42:48 INFO mapreduce.JobSubmitter: number of splits:1
19/07/23 13:42:48 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1563824945492_0002
19/07/23 13:42:49 INFO impl.YarnClientImpl: Submitted application application_1563824945492_0002
19/07/23 13:42:49 INFO mapreduce.Job: The url to track the job: http://node01:8088/proxy/application_1563824945492_0002/
19/07/23 13:42:49 INFO mapreduce.Job: Running job: job_1563824945492_0002
19/07/23 13:43:09 INFO mapreduce.Job: Job job_1563824945492_0002 running in uber mode : true
19/07/23 13:43:09 INFO mapreduce.Job: map 0% reduce 0%
19/07/23 13:43:12 INFO mapreduce.Job: map 100% reduce 0%
19/07/23 13:43:12 INFO mapreduce.Job: Job job_1563824945492_0002 completed successfully
19/07/23 13:43:12 INFO mapreduce.Job: Counters: 32

```

查看导出的数据

```
hdfs dfs -text /sqoop/emp/part-m-00000
```

它会用逗号（,）分隔 emp_add 表的数据和字段。

导入到 hdfs 指定目录并指定字段之间的分隔符

```
bin/sqoop import --connect jdbc:mysql://10.6.67.200:3306/test00 --username
root --password 123 --delete-target-dir --table emp --target-dir /sqoop/emp3
```

```
--m 1 --fields-terminated-by '\t'
```

```
19/07/23 13:49:08 INFO db.DBInputFormat: Using read committed transaction isolation
19/07/23 13:49:08 INFO mapreduce.JobSubmitter: number of splits:1
19/07/23 13:49:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1563824945492_0003
19/07/23 13:49:09 INFO impl.YarnClientImpl: Submitted application application_1563824945492_0003
19/07/23 13:49:09 INFO mapreduce.Job: The url to track the job: http://node01:8088/proxy/application_1563824945492_0003/
19/07/23 13:49:09 INFO mapreduce.Job: Running job: job_1563824945492_0003
19/07/23 13:49:27 INFO mapreduce.Job: Job job_1563824945492_0003 running in uber mode : true
19/07/23 13:49:27 INFO mapreduce.Job: map 0% reduce 0%
19/07/23 13:49:31 INFO mapreduce.Job: map 100% reduce 0%
19/07/23 13:49:31 INFO mapreduce.Job: Job job_1563824945492_0003 completed successfully
```

查看文件内容

hdfs dfs -text /sqoop/emp3/part-m-00000

```
[root@node03 sqoop-1.4.6-cdh5.14.0]# hdfs dfs -text /sqoop/emp2/part-m-00000
7369 SMITH CLERK 7902 1980-12-17 800.0 null 20
7499 ALLEN SALESMAN 7698 1981-02-20 1600.0 300.0 30
7521 WARD SALESMAN 7698 1981-02-22 1250.0 500.0 30
7566 JONES MANAGER 7839 1981-04-02 2975.0 null 20
7654 MARTIN SALESMAN 7698 1981-09-28 1250.0 1400.0 30
7698 BLAKE MANAGER 7839 1981-05-01 2850.0 null 30
7782 CLARK MANAGER 7839 1981-06-09 2450.0 null 10
7788 SCOTT ANALYST 7566 1987-07-03 3000.0 null 20
7839 KING PRESIDENT null 1981-11-17 5000.0 null 10
7844 TURNER SALESMAN 7698 1981-09-08 1500.0 0.0 30
7876 ADAMS CLERK 7788 1987-07-13 1100.0 null 20
7900 JAMES CLERK 7698 1981-12-03 950.0 null 30
7902 FORD ANALYST 7566 1981-12-03 3000.0 null 20
7934 MILLER CLERK 7782 1981-01-23 1300.0 null 10
[root@node03 sqoop-1.4.6-cdh5.14.0]#
```

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

hdfs 导出到 mysql

数据是在 HDFS 当中的如下目录/sqoop/emp，数据内容如下

第一步：创建 mysql 表

```
CREATE TABLE emp_out(
  EMPNO int PRIMARY KEY, #员工编号
  ENAME VARCHAR(10), #员工姓名
  JOB VARCHAR(9), #员工工作
  MGR int, #员工直属领导编号
  HIREDATE DATE, #入职时间
  SAL double, #工资
  COMM double, #奖金
  DEPTNO int #对应 dept 表的外键
```

$$);$$

第二步：执行导出命令

通过 export 来实现数据的导出，将 hdfs 的数据导出到 mysql 当中去

```
bin/sqoop export \  
--connect jdbc:mysql://10.6.67.200:3306/test00 \  
--username root --password 123 \  
--table emp_out \  
--export-dir /sqoop/emp \  
--input-fields-terminated-by ","
```

```

19/07/23 13:54:04 INFO input.FileInputFormat: Total input paths to process : 1
19/07/23 13:54:04 INFO input.FileInputFormat: Total input paths to process : 1
19/07/23 13:54:04 INFO mapreduce.JobSubmitter: number of splits:4
19/07/23 13:54:04 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated
19/07/23 13:54:04 INFO mapreduce.map.speculative
19/07/23 13:54:05 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1563824945492_0004
19/07/23 13:54:05 INFO impl.YarnClientImpl: Submitted application application_1563824945492_0004
19/07/23 13:54:05 INFO mapreduce.Job: The url to track the job: http://node01:8088/proxy/application_1563824945492_0004/
19/07/23 13:54:05 INFO mapreduce.Job: Running job: job_1563824945492_0004
19/07/23 13:54:36 INFO mapreduce.Job: Job job_1563824945492_0004 running in uber mode : true
19/07/23 13:54:36 INFO mapreduce.Job: map 0% reduce 0%
19/07/23 13:54:39 INFO mapreduce.Job: map 50% reduce 0%
19/07/23 13:54:40 INFO mapreduce.Job: map 100% reduce 0%
19/07/23 13:54:41 INFO mapreduce.Job: Job job_1563824945492_0004 completed successfully

```

第三步：验证 mysql 表数据

[illegible]