第 10 章 Sqoop

10.1 Sqoop 概述

Sqoop 是一款开源的工具，主要用于在 Hadoop(Hive)与传统的数据库(mysql、postgresql...)间进行数据的传递，可以将一个关系型数据库（例如 ： MySQL ,Oracle ,Postgres 等）中的数据导进到 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导进到关系型数据库中。

Sqoop 项目开始于 2009 年，最早是作为 Hadoop 的一个第三方模块存在，后来为了让使用者能够快速部署，也为了让开发人员能够更快速的迭代开发，Sqoop 独立成为一个 Apache 项目。

最新的稳定版本是 1.4.7。Sqoop2 的最新版本是 1.99.7。请注意，1.99.7 与 1.4.7 不兼容，且没有特征不完整，它并不打算用于生产部署。

10.2 Sqoop 下载与安装

10.2.1 Sqoop 安装地址

1）Sqoop 官网地址：

http://sqoop.apache.org/

2）文档查看地址：

http://sqoop.apache.org/docs/1.4.7/index.html

3）下载地址：

https://mirrors.tuna.tsinghua.edu.cn/apache/sqoop/1.4.7/

10.2.2 Sqoop 安装部署

1) 把 sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz 上传到 linux 的/opt/software 目录下
2) 解压 sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz 到/opt/module/目录下

```
[atguigu@hadoop102 software]$ tar -zxvf sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz -C /opt/module/
```

3) 修改 sqoop-1.4.7.bin__hadoop-2.6.0 的名称为 sqoop

```
[atguigu@hadoop102 software]$ mv sqoop-1.4.7.bin__hadoop-2.6.0/ sqoop
```

4) 修改/opt/module/sqoop/conf 目录下的 sqoop-env-template.sh 名称为 sqoop-env.sh

```
[atguigu@hadoop102 software]$ mv sqoop-env-template.sh sqoop-env.sh

[atguigu@hadoop102 software]$ mv sqoop-site-template.xml sqoop-site.xml
```

5）配置 sqoop-env.sh 文件

```
export HADOOP_COMMON_HOME=/opt/module/hadoop-2.7.2
export HADOOP_MAPRED_HOME=/opt/module/hadoop-2.7.2
export HIVE_HOME=/opt/module/hive
export ZOOKEEPER_HOME=/opt/module/zookeeper-3.4.10
export ZOOCFGDIR=/opt/module/zookeeper-3.4.10
```

10.2.3 添加 JDBC 驱动

拷贝/opt/software/mysql-libs/mysql-connector-java-5.1.27 目录下的

mysql-connector-java-5.1.27-bin.jar 到/opt/module/sqoop/lib/

[atguigu@hadoop102 mysql-connector-java-5.1.27]$ cp mysql-connector-java-5.1.27-bin.jar /opt/module/sqoop/lib/

10.2.4 验证 Sqoop

我们可以通过某一个 command 来验证 sqoop 配置是否正确：

```
$ bin/sqoop help
出现一些 Warning 警告（警告信息已省略），并伴随着帮助命令的输出：
Available commands:
  codegen            Generate code to interact with database records
  create-hive-table  Import a table definition into Hive
  eval               Evaluate a SQL statement and display the results
  export             Export an HDFS directory to a database table
  help               List available commands
  import             Import a table from a database to HDFS
  import-all-tables  Import tables from a database to HDFS
  import-mainframe   Import datasets from a mainframe server to HDFS
  job                Work with saved jobs
  list-databases     List available databases on a server
  list-tables        List available tables in a database
  merge              Merge results of incremental imports
  metastore          Run a standalone Sqoop metastore
  version            Display version information
```

10.2.5 测试 Sqoop 是否能够成功连接数据库

```
$ bin/sqoop list-databases --connect jdbc:mysql://hadoop102:3306/ --username root --password 000000
出现如下输出：
information_schema
metastore
mysql
performance_schema
```

10.3 导入数据

在 Sqoop 中，“导入”概念指：从非大数据集群（RDBMS）向大数据集群（HDFS，HIVE，

HBASE）中传输数据，叫做：导入，即使用 import 关键字。

10.3.1 RDBMS 到 HDFS

1) 确定 Mysql 服务开启正常

2) 在 Mysql 中新建一张表并插入一些数据

```
$ mysql -uroot -p123456
mysql> create database company;
mysql> create table company.staff(id int(4) primary key not null auto_increment, name varchar(255), sex varchar(255));
mysql> insert into company.staff(name, sex) values('Thomas', 'Male');
mysql> insert into company.staff(name, sex) values('Catalina', 'FeMale');
```

3) 导入数据

（1）全部导入

```
$ bin/sqoop import \
--connect jdbc:mysql://linux01:3306/company \
--username root \
--password 123456 \
--table staff \
--target-dir /user/company \
--delete-target-dir \
--num-mappers 1 \
--fields-terminated-by "\t"
```

（2）查询导入

```
$ bin/sqoop import \
--connect jdbc:mysql://linux01:3306/company \
--username root \
--password 123456 \
--target-dir /user/company \
--delete-target-dir \
--num-mappers 1 \
--fields-terminated-by "\t" \
--query 'select name,sex from staff where id <=1 and $CONDITIONS;'
```

尖叫提示：must contain '$CONDITIONS' in WHERE clause.

尖叫提示：如果 query 后使用的是双引号，则$CONDITIONS 前必须加转移符，防止 shell 识别为自己的变量。

（3）导入指定列

```
$ bin/sqoop import \
--connect jdbc:mysql://linux01:3306/company \
--username root \
--password 123456 \
--target-dir /user/company \
--delete-target-dir \
--num-mappers 1 \
--fields-terminated-by "\t" \
```

```
--columns id,sex \
--table staff
```

尖叫提示：columns 中如果涉及到多列，用逗号分隔，分隔时不要添加空格

（4）使用 sqoop 关键字筛选查询导入数据

```
$ bin/sqoop import \
--connect jdbc:mysql://linux01:3306/company \
--username root \
--password 123456 \
--target-dir /user/company \
--delete-target-dir \
--num-mappers 1 \
--fields-terminated-by "\t" \
--table staff \
--where "id=1"
```

尖叫提示：在 Sqoop 中可以使用 sqoop import -D property.name=property.value 这样的方式加入执行任务的参数，多个参数用空格隔开。

10.3.2 RDBMS 到 Hive

```
$ bin/sqoop import \
--connect jdbc:mysql://linux01:3306/company \
--username root \
--password 123456 \
--table staff \
--num-mappers 1 \
--hive-import \
--fields-terminated-by "\t" \
--hive-overwrite \
--hive-table staff_hive
```

提示：该过程分为两步，第一步将数据导入到 HDFS，第二步将导入到 HDFS 的数据迁移到 Hive 仓库，第一步默认的临时目录是/user/admin/表名

10.4 导出数据

在 Sqoop 中，“导出”概念指：从大数据集群（HDFS，HIVE，HBASE）向非大数据集群（RDBMS）中传输数据，叫做：导出，即使用 export 关键字。

4.2.1、HIVE/HDFS 到 RDBMS

```
$ bin/sqoop export \
--connect jdbc:mysql://linux01:3306/company \
--username root \
--password 000000 \
--table staff \
--num-mappers 1 \
```

```
--export-dir /user/hive/warehouse/staff_hive \
--input-fields-terminated-by "\t"
```

注意：Mysql 中如果表不存在，不会自动创建

思考：数据是覆盖还是追加

10.5 脚本打包

使用 opt 格式的文件打包 sqoop 命令，然后执行

1) 创建一个.opt 文件

```
$ mkdir opt
$ touch opt/job_HDFS2RDBMS.opt
```

2) 编写 sqoop 脚本

```
$ vi opt/job_HDFS2RDBMS.opt

export
--connect
jdbc:mysql://hadoop102:3306/company
--username
root
--password
000000
--table
staff
--num-mappers
1
--export-dir
/user/hive/warehouse/staff_hive
--input-fields-terminated-by
"\t"
```

3) 执行该脚本

```
$ bin/sqoop --options-file opt/job_HDFS2RDBMS.opt
```