一、Sqoop介绍

Sqoop(发音：skup)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。

二、安装

```
tar -zxf sqoop-1.4.6.bin__hadoop-2.0.4-alpha.tar.gz  -C /opt/instals
mv sqoop-1.4.6.bin__hadoop-2.0.4-alpha sqoop1.4.6

cd /opt/installs/sqoop1.4.6/conf
mv sqoop-env-template.sh sqoop-env.sh
vi sqoop-env.sh
#增加配置，注意修改路径
export HADOOP_COMMON_HOME=/opt/installs/hadoop2.9.2
export HADOOP_MAPRED_HOME=/opt/installs/hadoop2.9.2
export ZOOCFGDIR=/opt/installs/zookeeper3.4.6
export HIVE_HOME=/opt/installs/hive1.2.1

将mysql的驱动jar复制到sqoop的lib目录下（底层需要用JDBC操作MySQL数据库）
mysql-connector-java-5.1.38.jar
```

三、sqoop-import

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。

```
-- 测试数据库的表是否可以连接，显示库中的所有表
bin/sqoop list-tables --connect jdbc:mysql://hadoop10:3306/test1 --username root --password root
```

1). RDBMS -> HDFS

```
bin/sqoop import \
--driver com.mysql.jdbc.Driver \
--connect jdbc:mysql://hadoop10:3306/test1?characterEncoding=UTF-8 \
--username root \
--password root \
--table t_user \
--num-mappers 4 \
--fields-terminated-by '\t' \
--target-dir /mysql/t_user \
--delete-target-dir
```

> -m或--num-mappers
>
> 作用：启动N个map来并行导入数据，默认4个
>
> --fields-terminated-by
>
> 作用：设定每个字段是以什么符号作为结束，默认为逗号
>
> --target-dir
>
> 作用：指定HDFS路径
>
> --delete-target-dir
>
> 作用：如果目标目录已存在，则先删除

```
bin/sqoop import \
--driver com.mysql.jdbc.Driver \
--connect jdbc:mysql://hadoop10:3306/test1?characterEncoding=UTF-8 \
--username root \
--password root \
--table t_user \
--columns "id,name" \
--where "id > 2" \
--target-dir /mysql/t_user \
--delete-target-dir \
--num-mappers 1 \
--fields-terminated-by '\t'
```

> --columns <col1, col2, col3>
>
> 作用：指定要导入的字段
>
> --where
>
> 作用：从关系数据库导入数据时的查询条件

```
bin/sqoop import \
--driver com.mysql.jdbc.Driver \
--connect jdbc:mysql://hadoop10:3306/test1?characterEncoding=UTF-8 \
--username root \
--password root \
--num-mappers 1 \
--fields-terminated-by '\t' \
--query 'select id, name from t_user where $CONDITIONS LIMIT 3' \
--target-dir /mysql/t_user \
--delete-target-dir
```

> --query或--e
>
> 作用：将查询结果的数据导入，使用时必须伴随参--target-dir，--hive-table，如果查询中有where条件，则条件后必须加上$CONDITIONS关键字

2). RDBMS -> Hive

```
bin/sqoop import \
--connect jdbc:mysql://hadoop10:3306/test1 \
--username root \
--password root \
--table t_user \
--num-mappers 1 \
--hive-import \
--fields-terminated-by "\t" \
--hive-overwrite \
--hive-database baizhi \
--hive-table hive_user
```

> --table
>
> 作用：指定关系数据库的表名
>
> --hive-import
>
> 作用：将数据从关系数据库中导入到hive表中
>
> --hive-overwrite
>
> 作用：覆盖掉在hive表中已经存在的数据
>
> --hive-table
>
> 后面接要创建的hive表,默认使用MySQL的表名

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/用户名/表名

四、sqoop-export

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。

1). HDFS|hive -> RDBMS

① 准备数据，上传hdfs的sqoop目录下

```
vi a.txt
# 在文件中添加如下内容
1 zhangsan true 20 2020-01-11
2 lisi false 25 2020-01-10
3 wangwu true 36 2020-01-17
4 zhaoliu false 50 1990-02-08
5 win7 true 20 1991-02-08


#在hdfs上创建sqoop目录(目录名称随意，不过需要和后边对应),将文件上传到sqoop目录下
hdfs dfs -mkidr /sqoop
hdfs dfs -put a.txt /sqoop
```

② 在mysql中创建表

> 注意：导出并不会自动创建对应的表，需要提前自己创建

```
create table t_user2(
    id int primary key auto_increment,
    name VARCHAR(32),
    sex boolean,
    age int,
    birthDay date
) CHARACTER SET=utf8;
```

③ 将hdfs上的数据导入mysql表中

```
bin/sqoop export \
--connect jdbc:mysql://hadoop10:3306/test1 \
--username root \
--password root \
--table t_user2 \
--update-key id \
--update-mode allowinsert \
--export-dir /sqoop \
--input-fields-terminated-by ' '
```

> Mysql中如果表不存在，不会自动创建
>
> --export-dir 导出的数据 --input-fields-terminated-by 字段分割符号
>
> 导出模式可选值可以是updateonly或者allowinsert，
>
> updateonly仅仅会更新已经存在的记录。

> allowinsert既可以更新，又可以添加新数据