

课堂实验

一、基本信息

文档编号		文档版本	1.0
实验名称	Sqoop 数据导入		
所属课程	第 5 讲 分布式 ETL 工具 Sqoop	认证等级	数据工程师（初级）
授课形式	上机实验	实验批次	第 5 次 / 共 11 次

二、实验目的

- 掌握 Sqoop 数据导入的基本方法。
- 理解 Sqoop 工作原理。

三、实验准备

- 为了简化实验，采用 Inceptor Metastore 自带的 MySQL（即 TxSQL），作为 Sqoop 数据源。
- 下载并安装 TDH Client，使用 TDH Client 集成的 Sqoop。
- 实验目录与命名规划
 - （1）本地目录
工作目录：/mnt/disk1/{student_name}
 - （2）HDFS 目录
工作目录：/tmp/{student_name}
Sqoop 数据导入目录：/tmp/{student_name}/sqoop_data
 - （3）MySQL 数据库名：{db_student_name}
 - （4）{student_name} 为变量，代表学员姓名全拼
- 文件服务器
 - （1）IP：172.16.140.111
 - （2）目录：/mnt/disk1/de_training

四、实验内容

1、启动 TDH Client

- 任务：执行 TDH Client 的 init.sh 脚本，启动 TDH Client。

- 步骤

Linux:

```
// 执行 TDH Client 的 init.sh 脚本
1. source {TDH_Client_install_dir}/init.sh
```

2、创建 MySQL 用户

- 任务：创建 MySQL 用户，将全部权限赋给集群的三台服务器，并确保赋权生效。

- 步骤

Linux:

```
// 登录集群第一个节点，获取 MySQL 的 root 密码
1. cat /etc/txsql1/conf/db.properties
// 登录 MySQL, -u 用户名, -p 密码, -P 端口号, -h 集群第一个节点的 IP
2. mysql -uroot -p120621016 -P3316 -h172.16.140.85
```

SQL:

```
// 创建 MySQL 用户
1. create user 'sqoopuser'@'172.16.140.85' identified by '111111';
2. create user 'sqoopuser'@'172.16.140.86' identified by '111111';
3. create user 'sqoopuser'@'172.16.140.87' identified by '111111';
// 将全部权限赋给集群的三个节点
4. grant all privileges on *.* to sqoopuser@'172.16.140.85';
5. grant all privileges on *.* to sqoopuser@'172.16.140.86';
6. grant all privileges on *.* to sqoopuser@'172.16.140.87';
// 使赋权生效
7. flush privileges;
```

3、准备实验数据

- 任务：向 MySQL 中插入实验数据。

- 步骤

SQL:

```
// 创建并切换数据库
1. create database {db_student_name};
2. use {db_student_name};
3. show tables;
// 创建表，并插入 10 条数据
4. create table tbl_sqoop(name varchar(11),age int ,sex varchar(11));
5. insert into tbl_sqoop value('zs',11,'girl');
6. insert into tbl_sqoop value('zs',11,'girl');
7. insert into tbl_sqoop value('zs',11,'girl');
8. insert into tbl_sqoop value('zs',11,'girl');
```

```
9. insert into tbl_sqoop value('zs',11,'girl');
10. insert into tbl_sqoop value('zs',11,'girl');
11. insert into tbl_sqoop value('zs',11,'girl');
12. insert into tbl_sqoop value('zs',11,'girl');
13. insert into tbl_sqoop value('zs',11,'girl');
14. insert into tbl_sqoop value('zs',11,'girl');
// 查看表中的数据
15. select * from tbl_sqoop;
// 退出 MySQL
16. exit;
```

4、为 Sqoop 添加 MySQL 驱动

- 任务：为 Sqoop 添加 MySQL 驱动，并将驱动文件的权限修改为 777。
- 步骤

Linux:

```
// 登录文件服务器，将 MySQL 驱动复制到集群第一个节点的 TDH-Client/sqoop/lib/中
1. cd /mnt/disk1/de_traning
2. scp mysql-connector-java-5.1.38-bin.jar
   172.16.140.85:{TDH_Client_install_dir}/sqoop/lib/
// 登录集群第一个节点，将 MySQL 驱动文件的权限修改为 777
3. cd {TDH_Client_install_dir}/sqoop/lib/
4. chmod -R 777 mysql-connector-java-5.1.38-bin.jar
```

5、通过 Sqoop 查看数据库和表

- 任务：通过 Sqoop 列出 MySQL 中的所有数据库，以及指定数据库中的所有表。
- 步骤

Linux:

```
// 列出 MySQL 的所有数据库
1. sqoop list-databases --username sqoopuser --password 111111 --connect
   jdbc:mysql://172.16.140.85:3316/
// 列出数据库下的所有表
2. sqoop list-tables --username sqoopuser --password 111111 --connect
   jdbc:mysql://172.16.140.85:3316/{db_student_name}
```

6、Sqoop 数据导入

- 任务：将 MySQL 中的实验数据导入到 HDFS，并查看导入数据。注意，导入前需确保 HDFS 目录不存在，否则无法导入。
- 步骤

Linux:

```
// 将 tbl_sqoop 表中的数据导入到 HDFS 的/tmp/{student_name}/sqoop_data 目录中
1. sqoop import --username sqoopuser --password 111111 --connect
   jdbc:mysql://172.16.140.85:3316/{db_student_name} --query "select * from tbl_sqoop
   where \$CONDITIONS" --target-dir /tmp/{student_name}/sqoop_data/ --fields-
   terminated-by ',' --split-by id -m 1
// 查看导入 HDFS 的数据文件
2. hadoop fs -ls /tmp/{student_name}/sqoop_data
// 查看导入文件的内容
3. hadoop fs -cat /tmp/{student_name}/sqoop_data/part-m-00000
```