# 数仓工具

## 学习目标

能够使用Hue操作HDFS

能够使用Hue操作Hive

理解为什么选择Sqoop

理解Sqoop1和Sqoop2的区别

理解Sqoop抽取数据的两种方式

能够使用Sqoop导入完整数据到HDFS

能够使用Sqoop导入完整数据到Hive

能够使用Sqoop导入条件数据到HDFS

能够使用Sqoop导入条件数据到Hive

能够使用Sqoop导出数据到Mysql

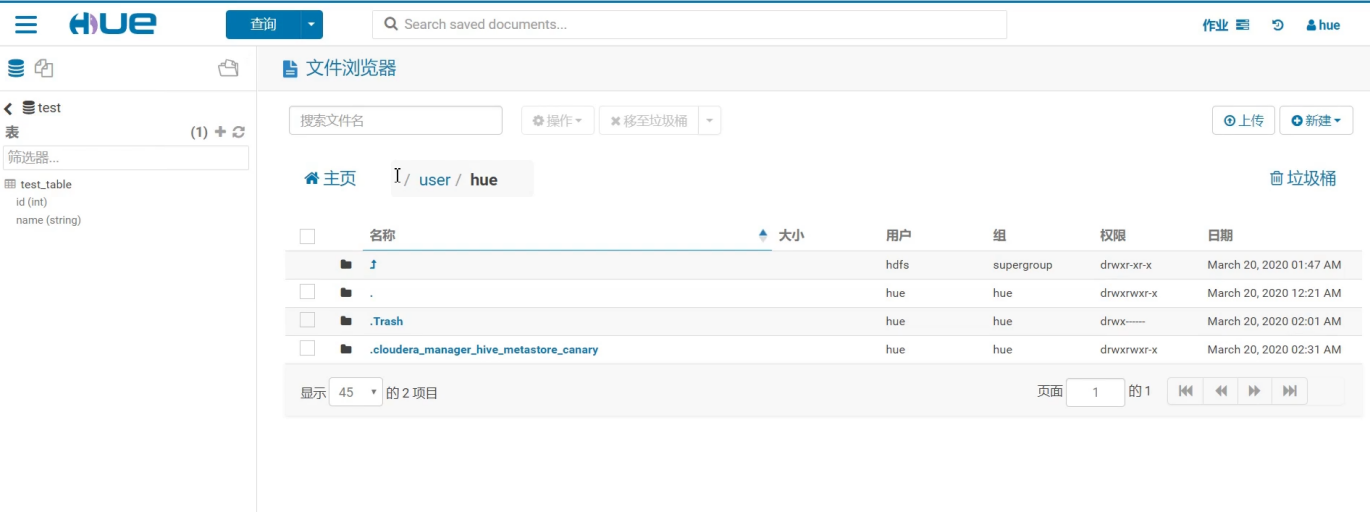
理解为什么选择Oozie

能够使用Hue操作Oozie

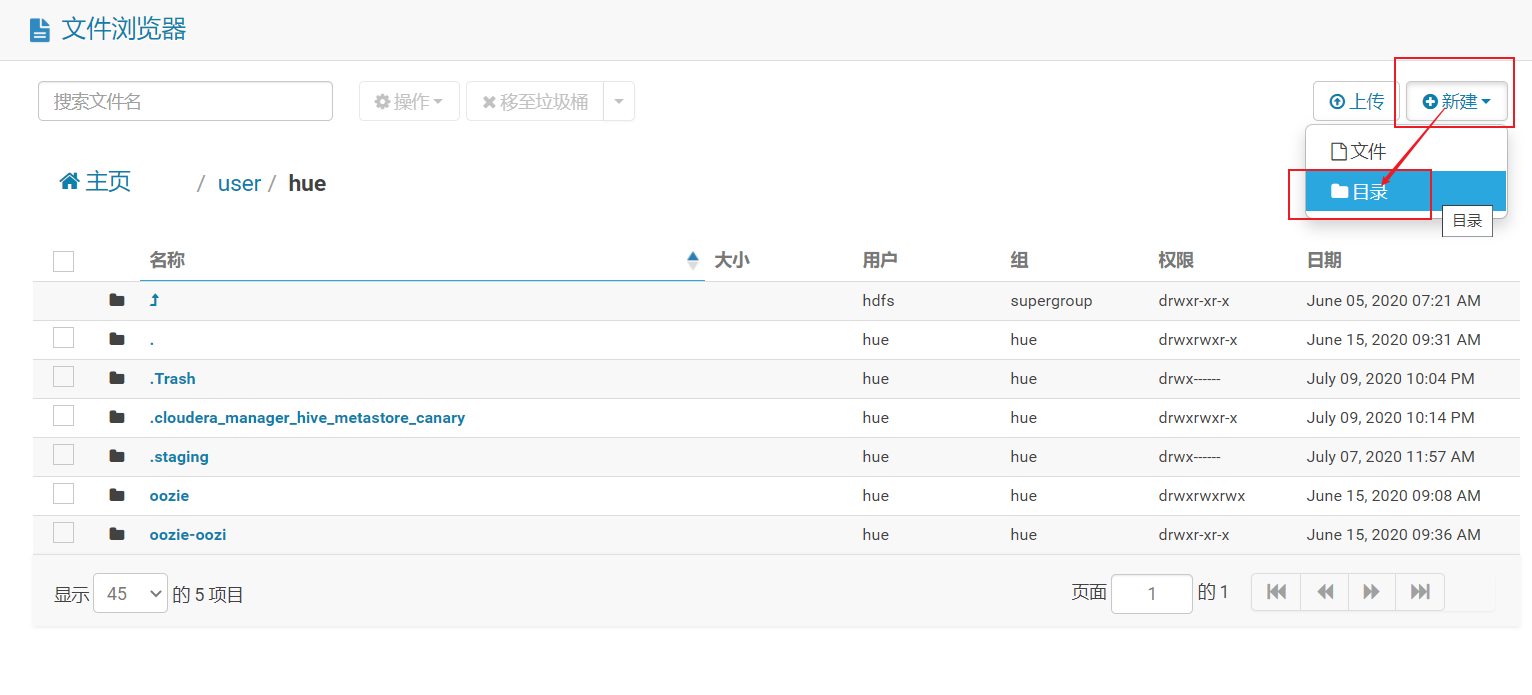
## Hue操作HDFS

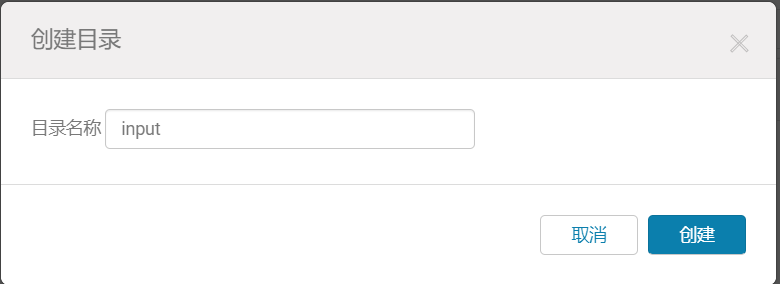
### 进入HDFS管理界面

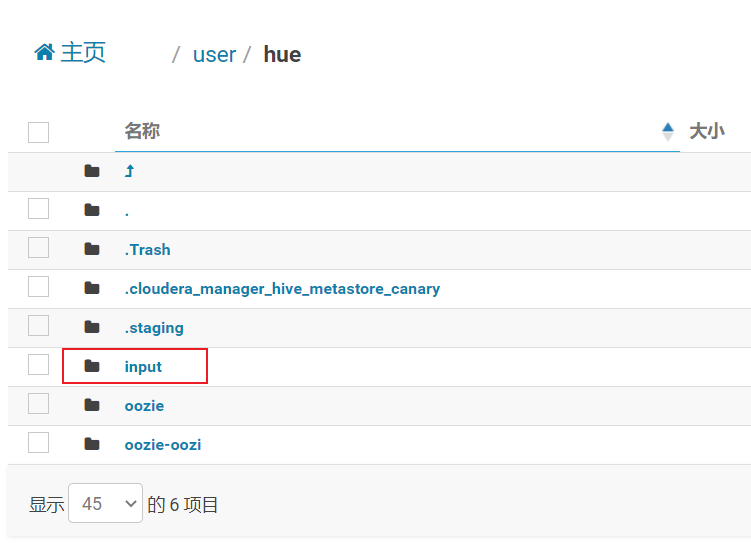




### HDFS新建文件夹

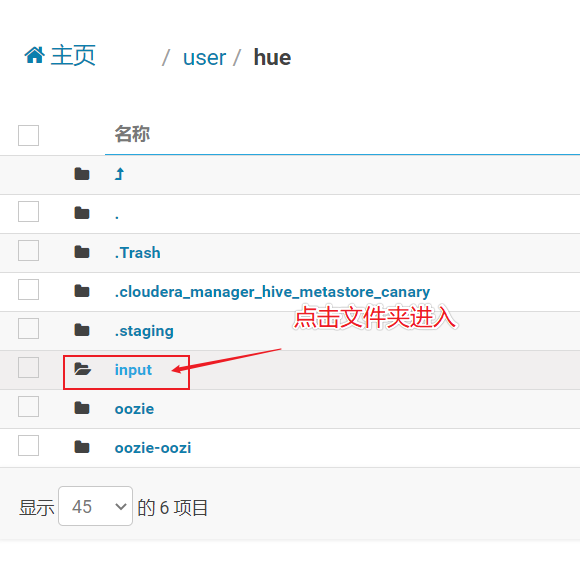




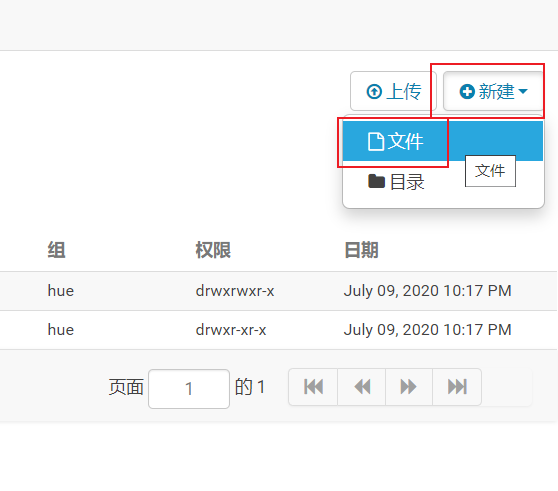


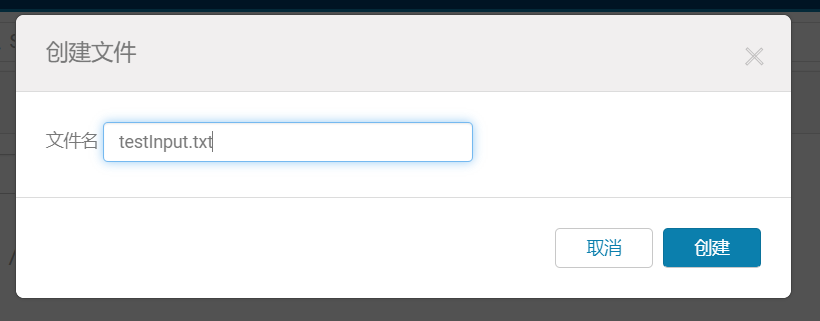
### 新建文件

#### 进入文件夹



#### 新建文件





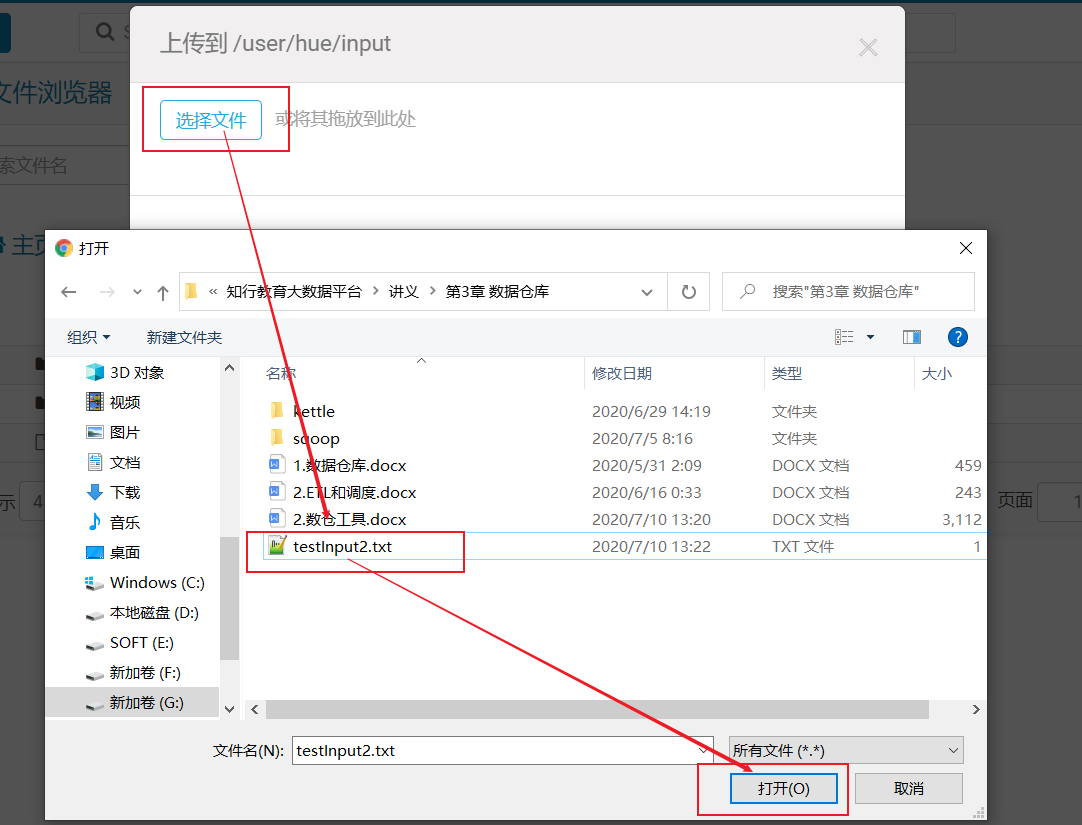
#### 创建成功



### 上传文件

#### 选择文件

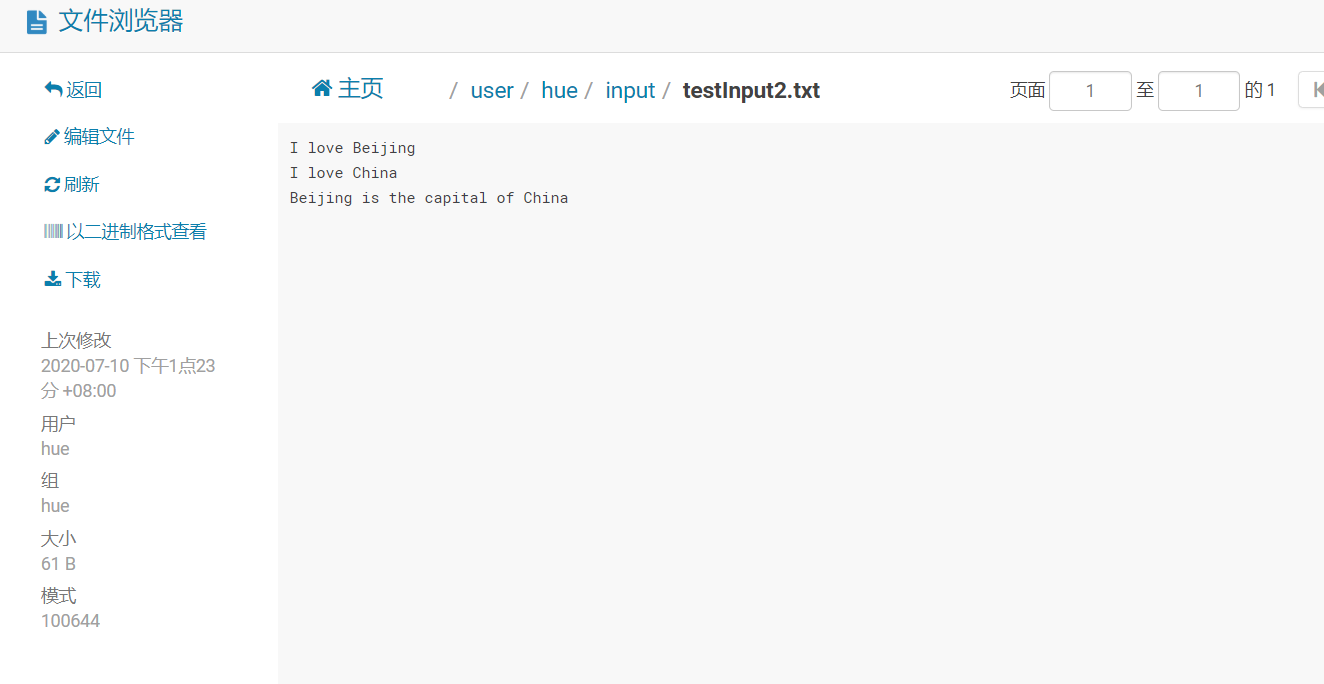




#### 上传成功



### 查看HDFS文件内容



### 编辑HDFS文件



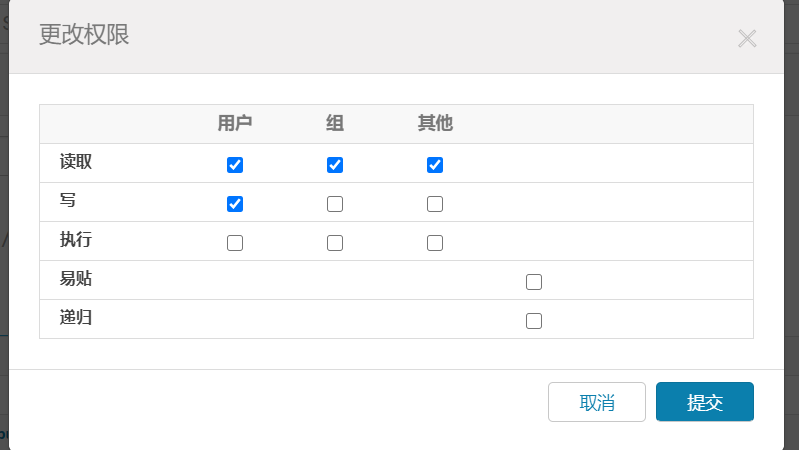


### 删除文件



### 更改文件权限





## Hue操作Hive

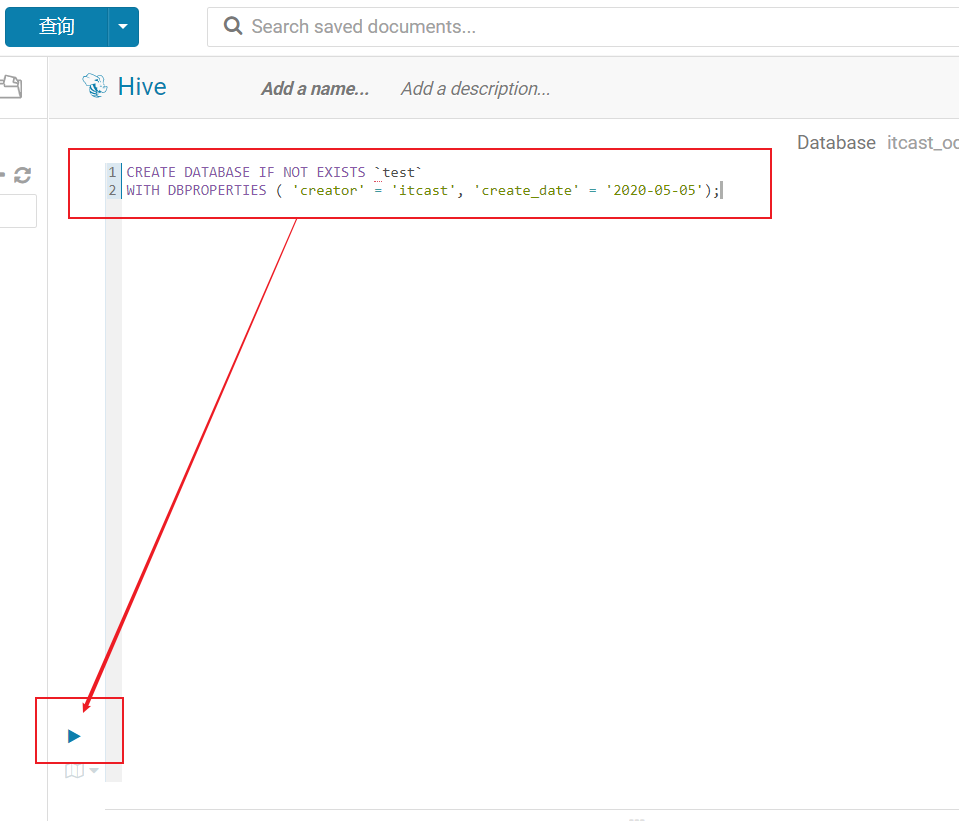
### 进入Hive面板



### 测试

#### 创建数据库

|  |
| --- |
| CREATE DATABASE IF NOT EXISTS `test`; |



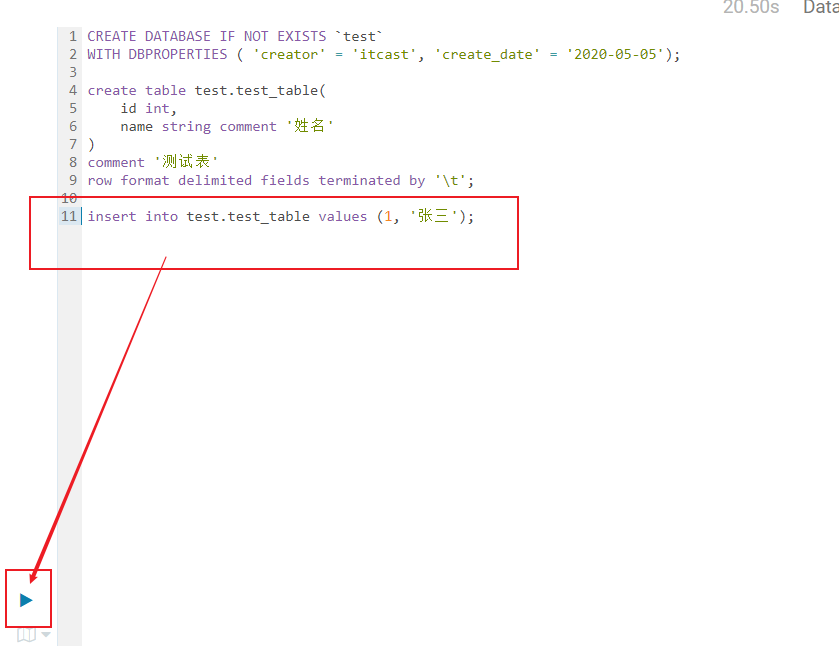
#### 创建表

|  |
| --- |
| create table test.test\_table(  id int,  name string comment '姓名' ) comment '测试表' row format delimited fields terminated by '\t'; |



#### 插入数据

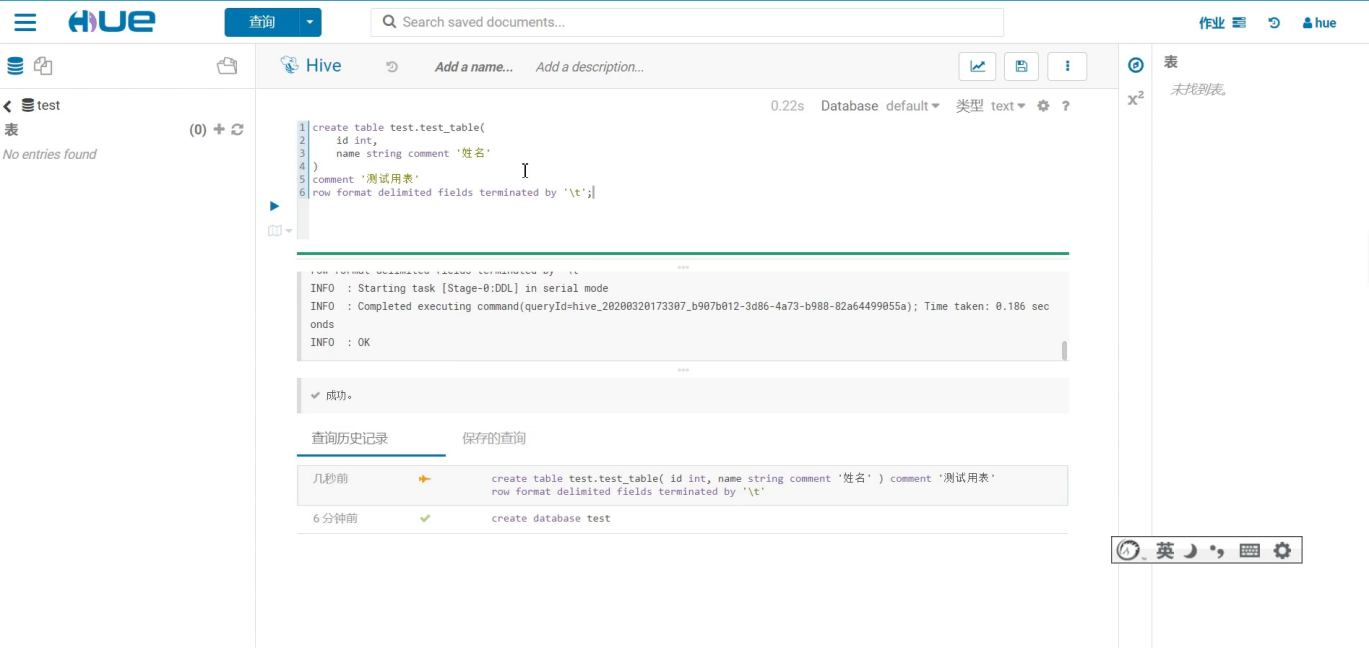
|  |
| --- |
| insert into test.test\_table values (1, '张三'); |



#### 查询数据

|  |
| --- |
| select *\** from test.test\_table; |





#### 调整区域大小



## Sqoop

### Sqoop介绍

Sqoop是Apache下的顶级项目，用来将Hadoop和关系型数据库中的数据相互转移，可以将一个关系型数据库（例如：MySQL，Oracle，PostgreSQL等）中的数据导入到Hadoop的HDFS中，也可以将HDFS的数据导入到关系型数据库中。目前在各个公司应用广泛，且发展前景比较乐观。其特点在于：

1. 专门为Hadoop而生，随Hadoop版本更新支持程度好，且原本即是从CDH版本孵化出来的开源项目，支持CDH的各个版本号。
2. 它支持多种关系型数据库，比如mysql、oracle、postgresql等。
3. 可以高效、可控的利用资源。
4. 可以自动的完成数据映射和转换。
5. 大部分企业还在使用sqoop1版本，sqoop1能满足公司的基本需求。
6. 自带的辅助工具比较丰富，如sqoop-import、sqoop-list-databases、sqoop-list-tables等。

### 为什么选择Sqoop

我们常用的ETL工具有Sqoop、Kettle、Nifi。

知行教育大数据平台，ETL的数据量较大，但是数据来源的类型简单（mysql）：

1. Kettle虽然功能较完善，但当处理大数据量的时候瓶颈问题比较突出，不适合此项目；
2. NiFi的功能强大，且支持大数据量操作，但NiFi集群是独立于Hadoop集群的，需要独立的服务器来支撑，强大也就意味着有上手门槛，学习难度大，用人成本高；
3. Sqoop专为关系型数据库和Hadoop之间的ETL而生，支持海量数据，符合项目的需求，且操作简单门槛低，因此选择Sqoop作为ETL工具。

### Sqoop操作

Sqoop详细资料见 《Home\讲义\第2章 数据仓库与工具\sqoop\sqoop教程\_V1.0.docx》。

## 调度

大数据常用的调度工具有：Oozie和Azkaban。

### **Oozie和Azkaban对比**

1.Oozie是基于Hadoop系统进行操作，而Azkaban是基于命令行进行操作。使用hadoop提供的第三方包JobClient比直接在底层跑shell命令开发成本小，可能遇到的坑也少（一个是基于Hadoop平台，一个是基于Linux系统）。

2.Oozie的操作是放在Hadoop中，而Azkaban的运行是服务器运行shell命令。为保证服务器的稳定，使用Oozie靠谱点。

3.Ooize提供查询任务执行状态，Azkaban查询的是进程执行的结果，如果某进程执行的shell命令出错，其进程仍展示位成功，混淆了任务输出。

4.Oozie将任务执行的状态持久化到数据库中，Azkaban将任务的状态存储在服务器内存中，如果掉电，则Azkaban会丢失任务信息。

5.Ooize中定义的action类型更为丰富，而Azkaban中的依赖较为简单，当面对复杂的逻辑时Oozie执行的比较顺畅。

6.两者均可以通过WEB界面进行配置操作，Azkaban和Oozie均拥有自带WEB，而Oozie同时还被HUE集成，通过HUE集成管理，可以更加方便的在Web页面上完成workflow的启动、停止、恢复。

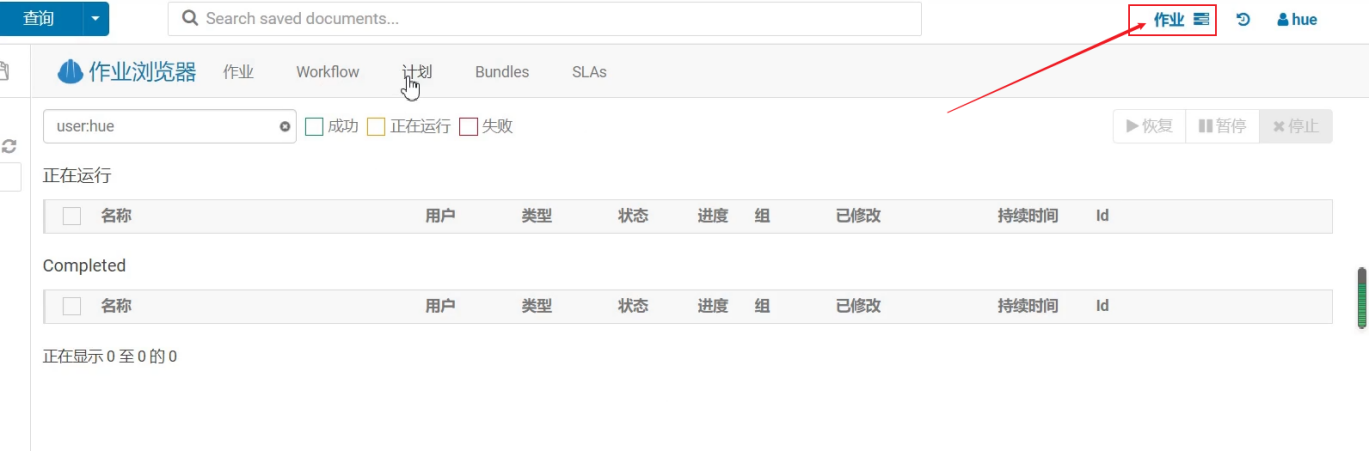
7.CDH原生支持Oozie，而Azkaban还需要额外的编译及安装配置。

### 项目选型

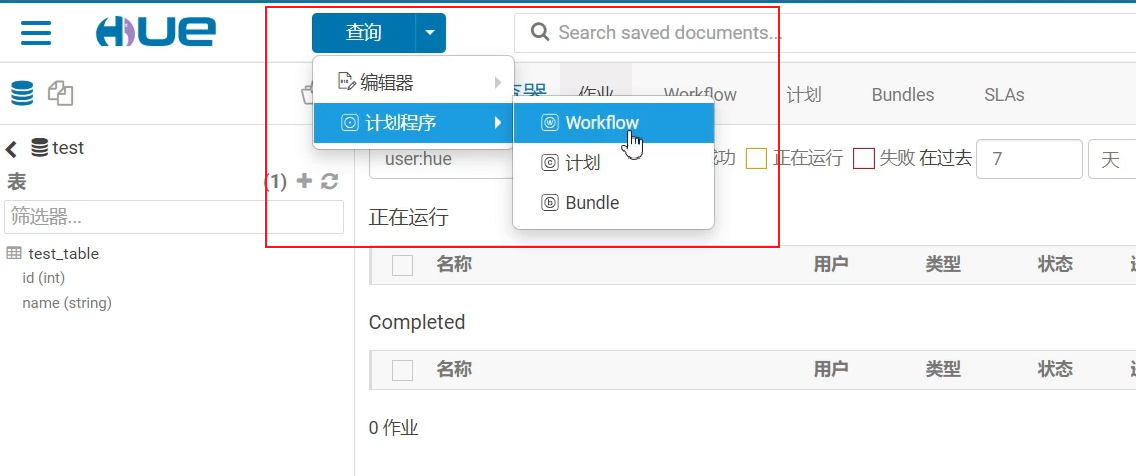
基于以上对比，本课程采用被ClouderaManager和Hue所支持的Oozie来进行调度。

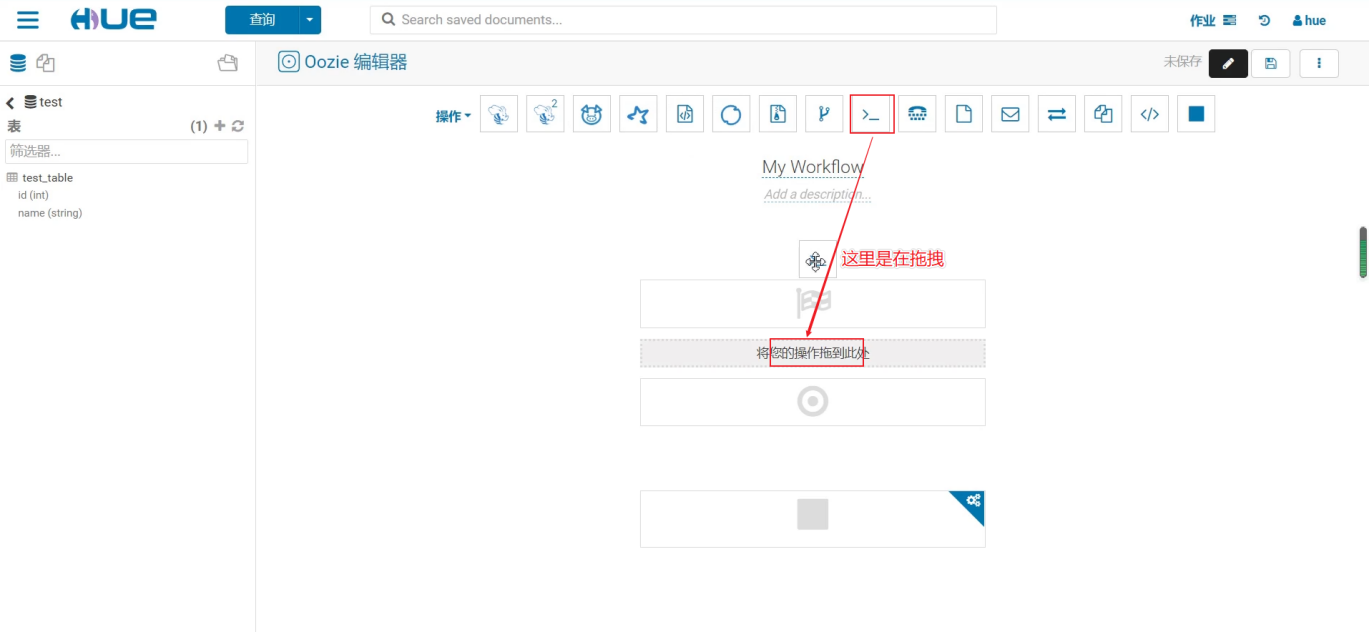
### Oozie使用案例

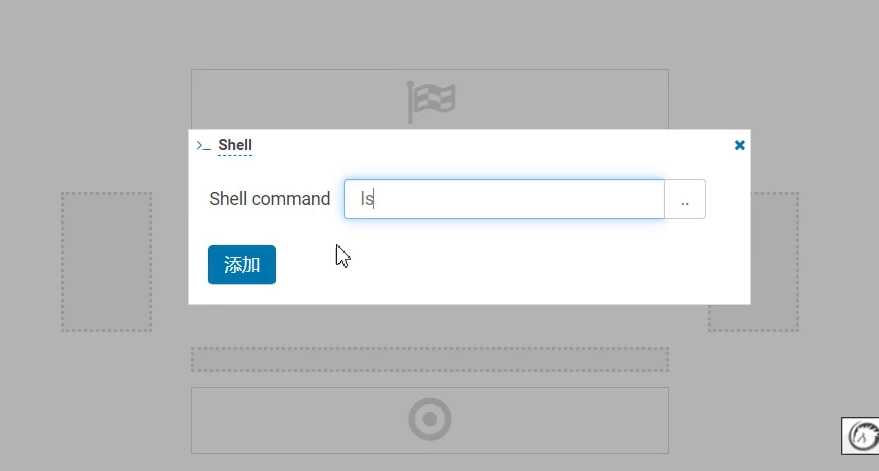
#### 查看已有的oozie程序



#### 新增oozie程序

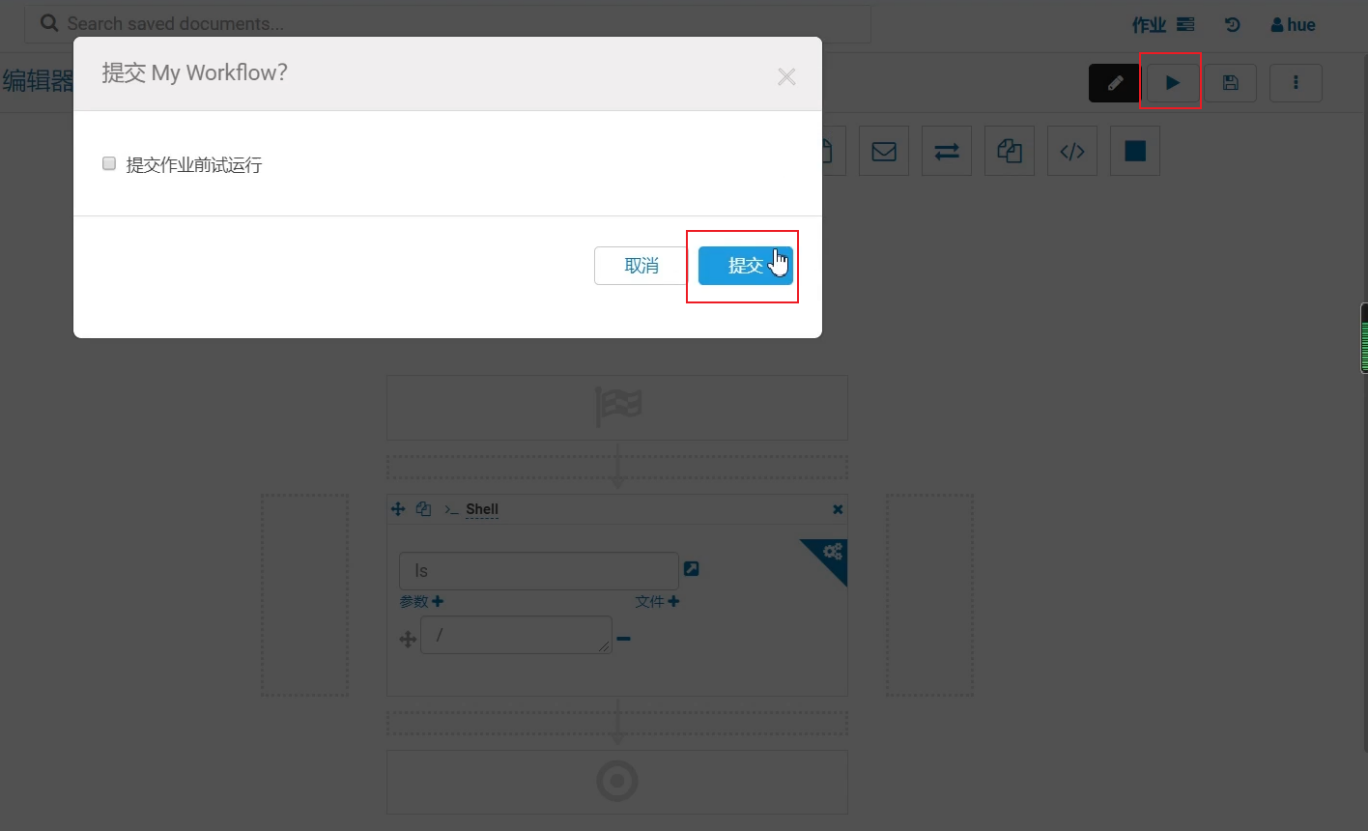


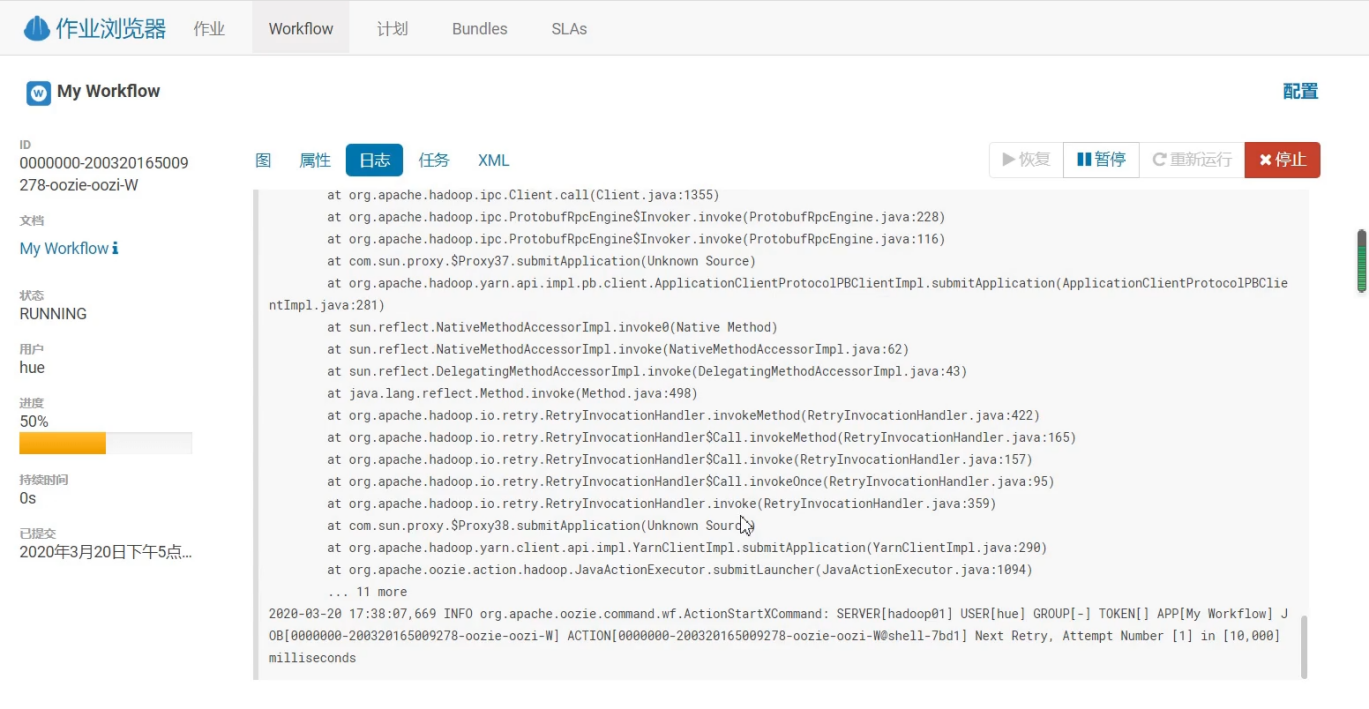






#### 提交任务





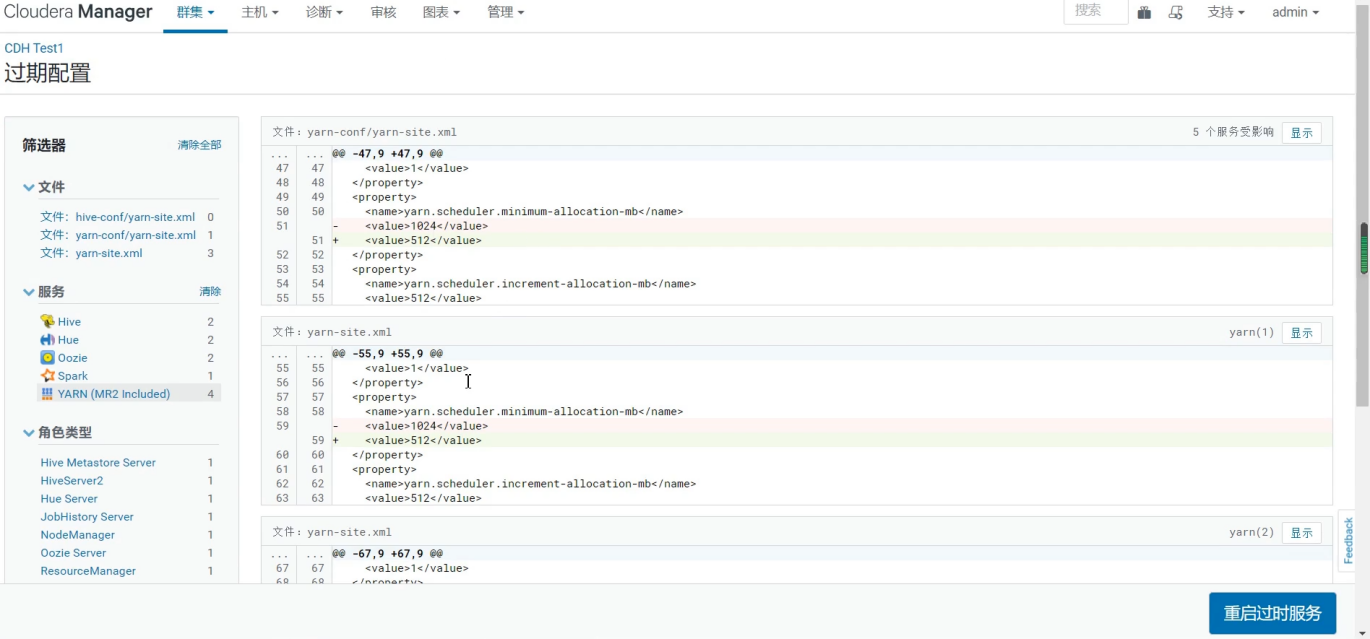
#### 修复异常

org.apache.oozie.action.ActionExecutorException: JA009: Invalid resource request! Cannot allocate containers as requested resource is greater than maximum allowed allocation. Requested resource type=[memory-mb], Requested resource=<memory:2048, vCores:1>, maximum allowed allocation=<memory:1024, vCores:4>, please note that maximum allowed allocation is calculated by scheduler based on maximum resource of registered NodeManagers, which might be less than configured maximum allocation=<memory:1024, vCores:4>

这是内存分配过小导致的错误，需要设置内存并重启Yarn

|  |
| --- |
| <property>  <name>yarn.scheduler.minimum-allocation-mb</name>  <value>526</value>  </property>  <property>  <name>yarn.scheduler.maximum-allocation-mb</name>  <value>4086</value>  </property>  <property>  <name>yarn.scheduler.minimum-allocation-vcores</name>  <value>1</value>  </property>  <property>  <name>yarn.scheduler.maximum-allocation-vcores</name>  <value>2</value>  </property>  <property>  <name>yarn.nodemanager.resource.memory-mb</name>  <value>2048</value>  </property>  <property>  <name>yarn.nodemanager.vmem-pmem-ratio</name>  <value>4.1</value>  </property> |

#### 修改后重启服务



1. 执行成功

