大数据加工

产品文档

文档目录

产品简介

产品概述

最近更新时间: 2019-10-28 05:48:45

大数据加工服务，为用户提供大数据应用开发环境，开发的作业流可直接在云上测试、发布、运行。 平台支持SQL代码、shell脚本、拖拽式等多种开发模式，以提高开发效率。平台提供丰富的算子及其调用入口，进一步减少用户输入，真正做到低门槛、易使用。

相关定义

最近更新时间: 2019-11-26 15:30:16

了解数据开发时会涉及到以下概念： 作业流:是指一个由作业节点组成的图。每个作业节点按照配置完成一定的处理逻辑。作业节点之间通过有向边进行依赖关联，但关联时不能形成环路。一个画布中的全部作业节点及其依赖称为一个作业流。一般来说，在作业流调度模型中，作业流为调度单元，而其中的作业节点为最小粒度的执行单元。 作业:作业流中的一个节点，即由用户定义的完成一定工作的逻辑单元。在任务调度模型中，作业（或任务）是最小执行单元。

插件：一个作业配置模板，它包含了作业类型和该种类型作业的必要参数，通过插件创建作业时，只需要填写作业类型和必要的参数就可以完成作业的创建，可以极大的节省创建作业的时间。

算子：一段可被高度提炼的逻辑，比如一段被高频率使用的SQL，算子必须依赖于插件存在，并最终可被插件解释和执行。

依赖包：被作业依赖的外部资源,比如一个JAR文件。

在线测试：作业流提交到测试环境执行，通过ENV_ID区分，在线测试不强制要求作业流是发布状态，任何状态都可以测试。

作业测试：同在线测试，但仅运行单个节点作业。

立即执行：将作业流提交到生产环境运行，作业流状态必须是已发布状态。

提交调度：将作业流提交到生产环境并按指定频率运行,作业流状态必须是已发布状态。

项目管理员：项目管理员具有项目下的所有权限，可以添加或删除项目成员，项目成员又分为开发人员和运维人员等。

运维人员：主要负责作业流的执行、调度及审批等。

开发人员：负责作业流的开发，资源维护、UDF开发等。

功能和优势

最近更新时间: 2019-11-26 15:30:16

大数据云数据开发提供了一站式的云上数据开发平台，满足用户多种类型的大数据开发需求。并支持任务开发，保存，调试，发布等数据开发全流程，有效提升开发效率。

- 丰富的插件支持，支持Spark SQL，Shell等多种类型插件，一个平台满足用户数据集成，计算分析等需求，同时满足用户自定义插件不同层级的复用。
- 灵活的版本管理，支持不同版本的查看，切换，并支持不同版本的作业上线发布。
- 与调度系统无缝衔接，支持以可视化的方式进行调度编排与配置。
- 用户自有资源的上传，版本管理和引用，并在UDF等层面上支持用户自定义扩展。
- 多人协作，在线开发，共享开发任务，同时支持开发，测试等不同环境充分隔离。

应用场景

最近更新时间: 2019-10-28 05:52:57

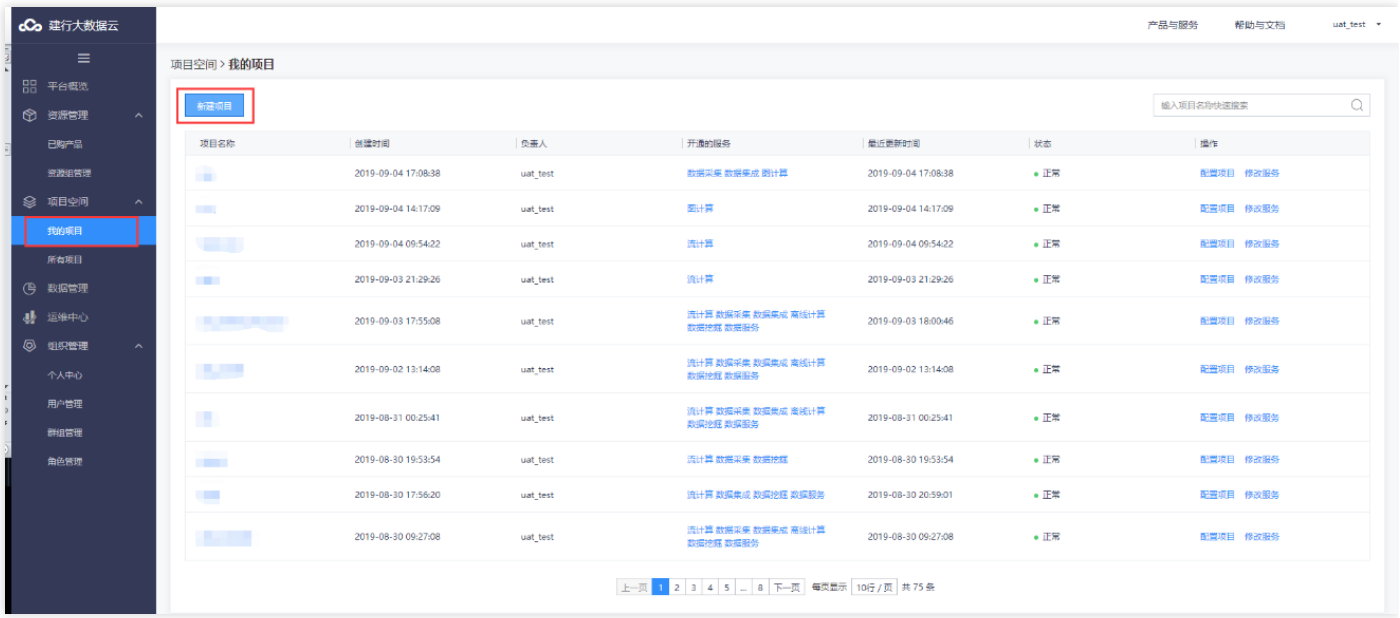
- 离线数据加工 可以搭配数据集成服务，对离线场景的数据进行预处理和加工，为后续的深加工做准备。
- 实时数据加工 搭配流计算服务和可视化BI服务，用户可快速搭建一套实时流数据分析平台。可以解决传统基于批量模型无法实时进行数据分析的问题。

快速入门
创建项目

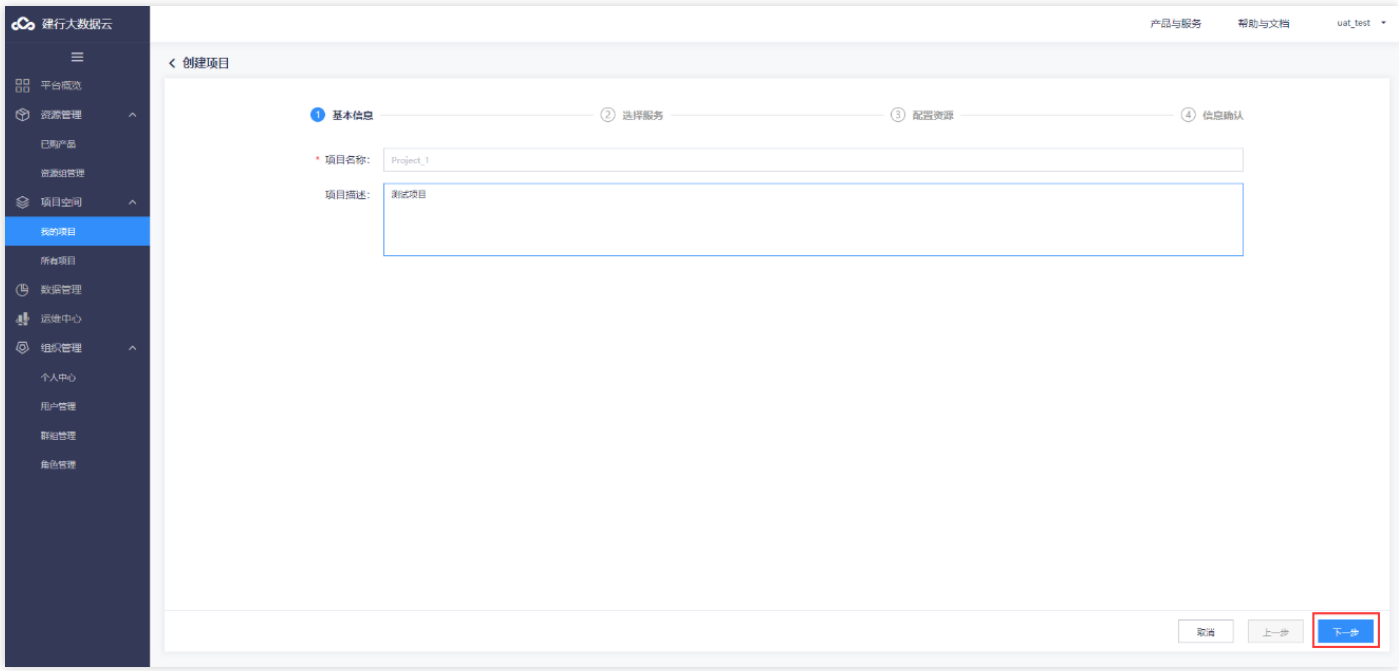
最近更新时间: 2019-11-13 03:21:07

1. 登录大数据云租户控制台。

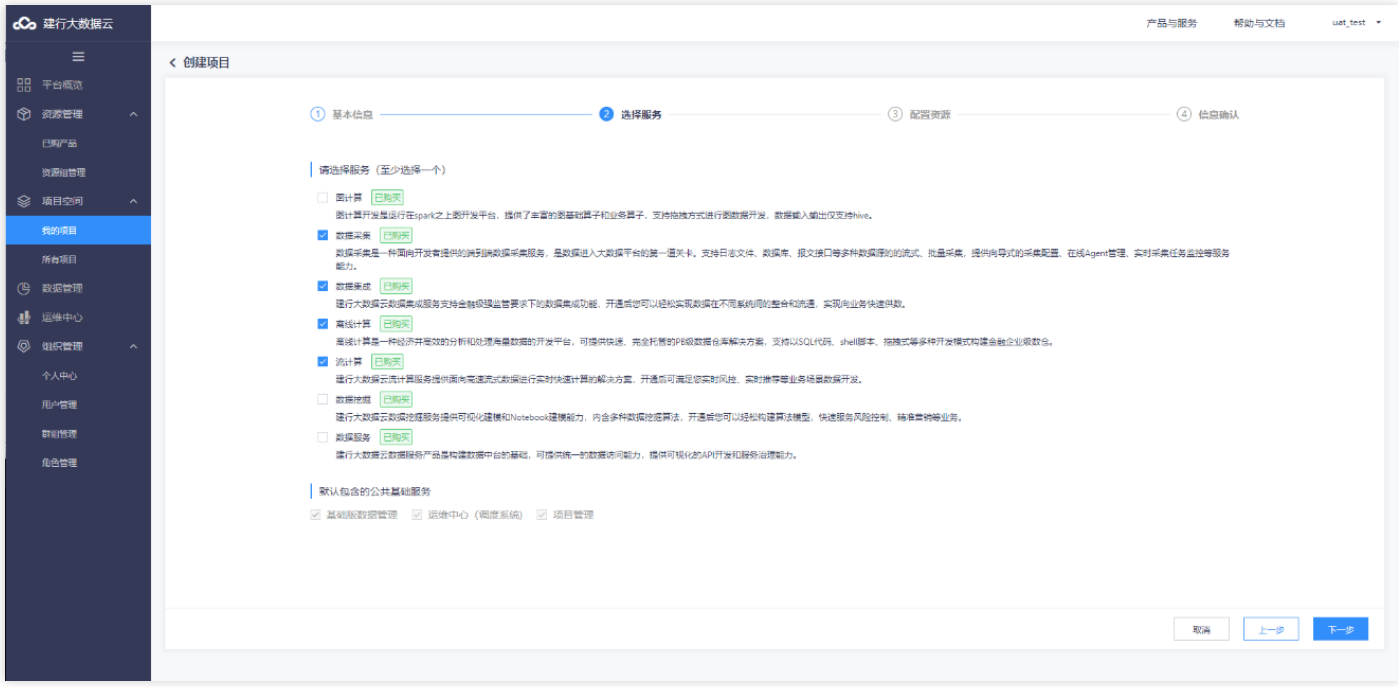
2. 选择项目空间下【新建项目】创建一个新的项目。



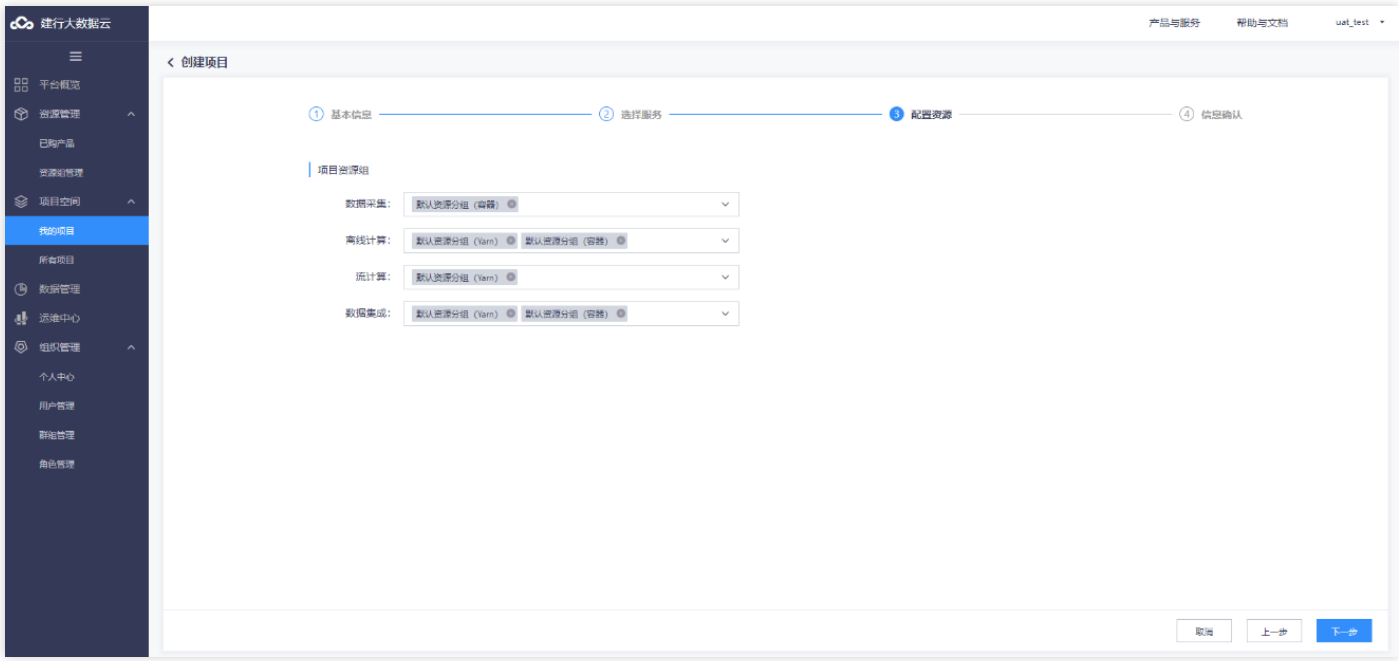
3. 输入项目名称，项目描述，点击【下一步】。



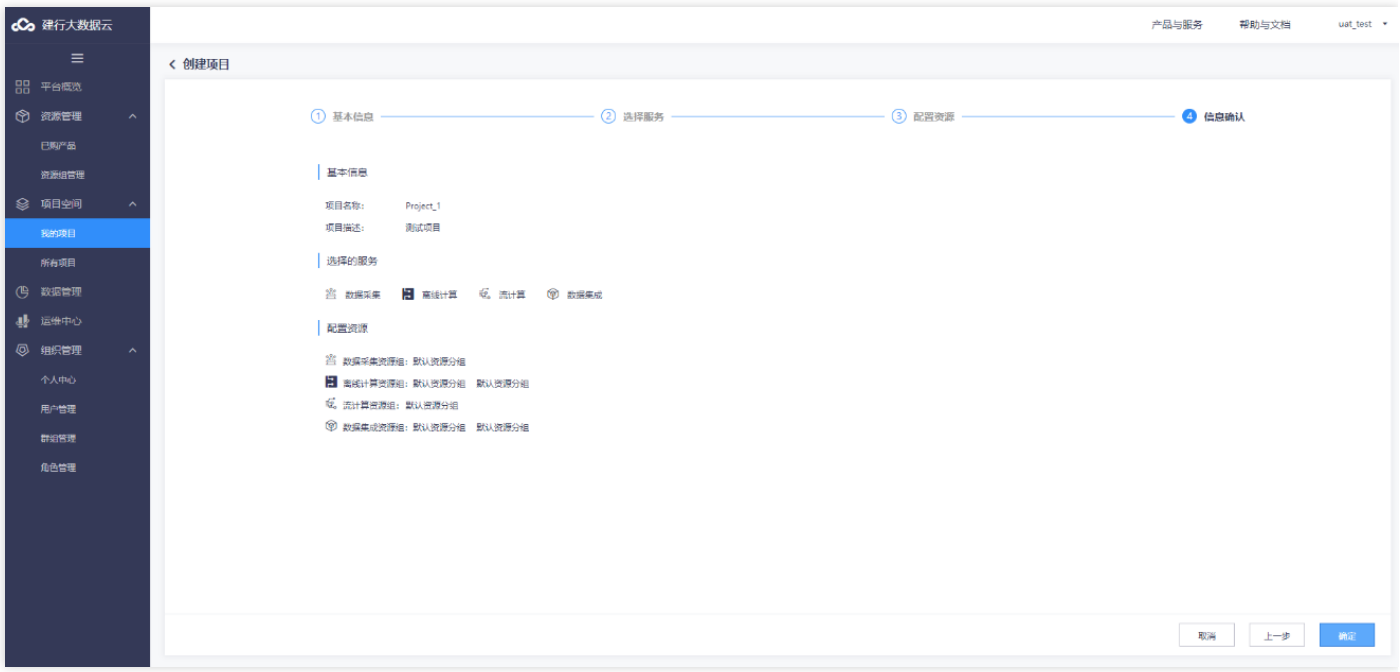
4. 勾选项目需要创建的数据开发及关联服务，点击“下一步”。



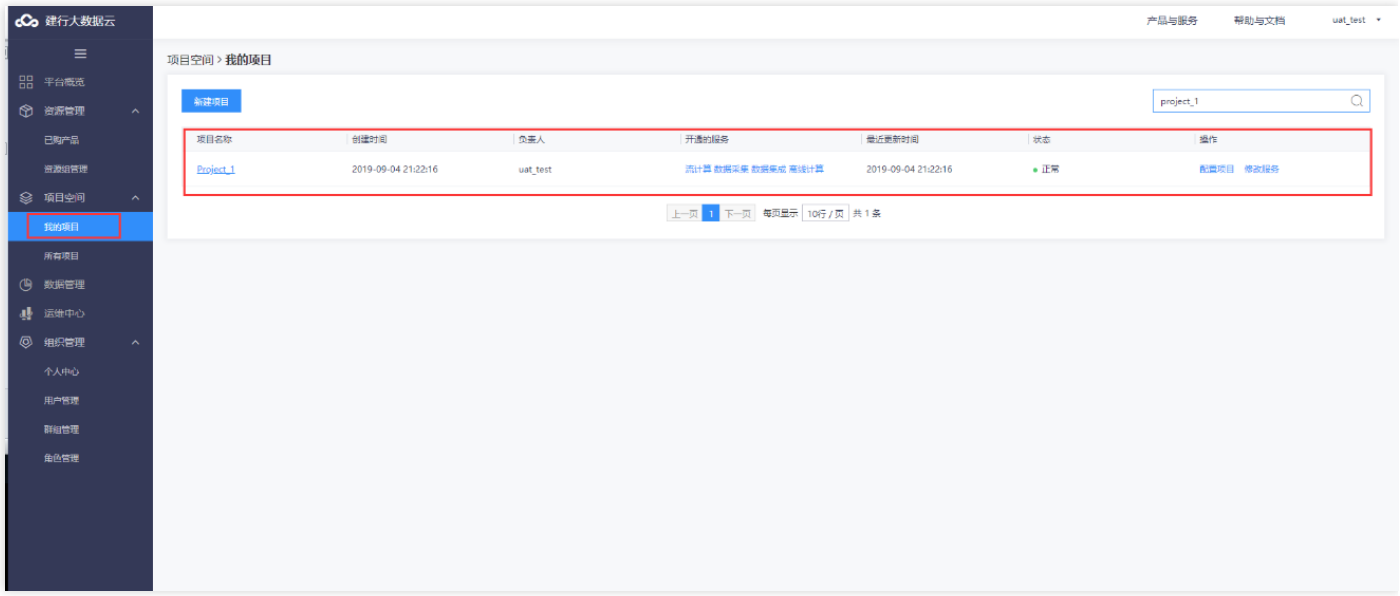
5. 为每个服务进行资源选择，每项服务都有项目资源后，点击【下一步】。



6. 确认项目信息。如需修改，点击“上一步”修改；如无需修改，点击“确定”即可。



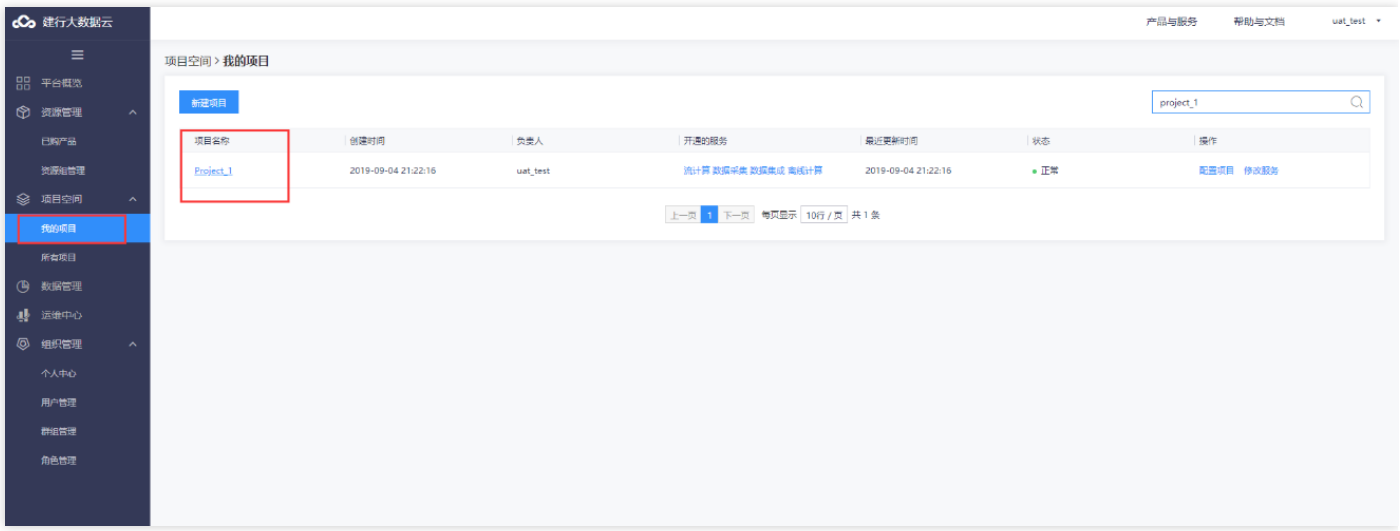
7. 新建项目就会在【我的项目】中展示。



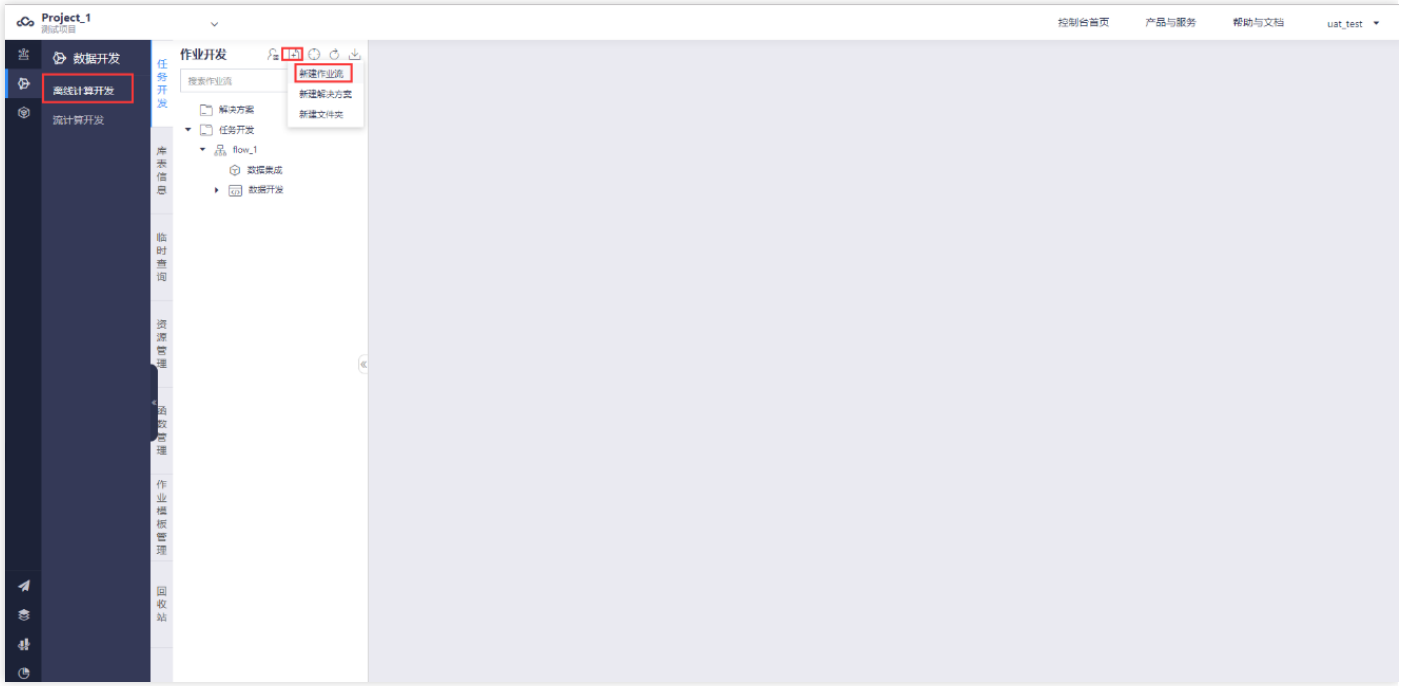
新建数据开发任务

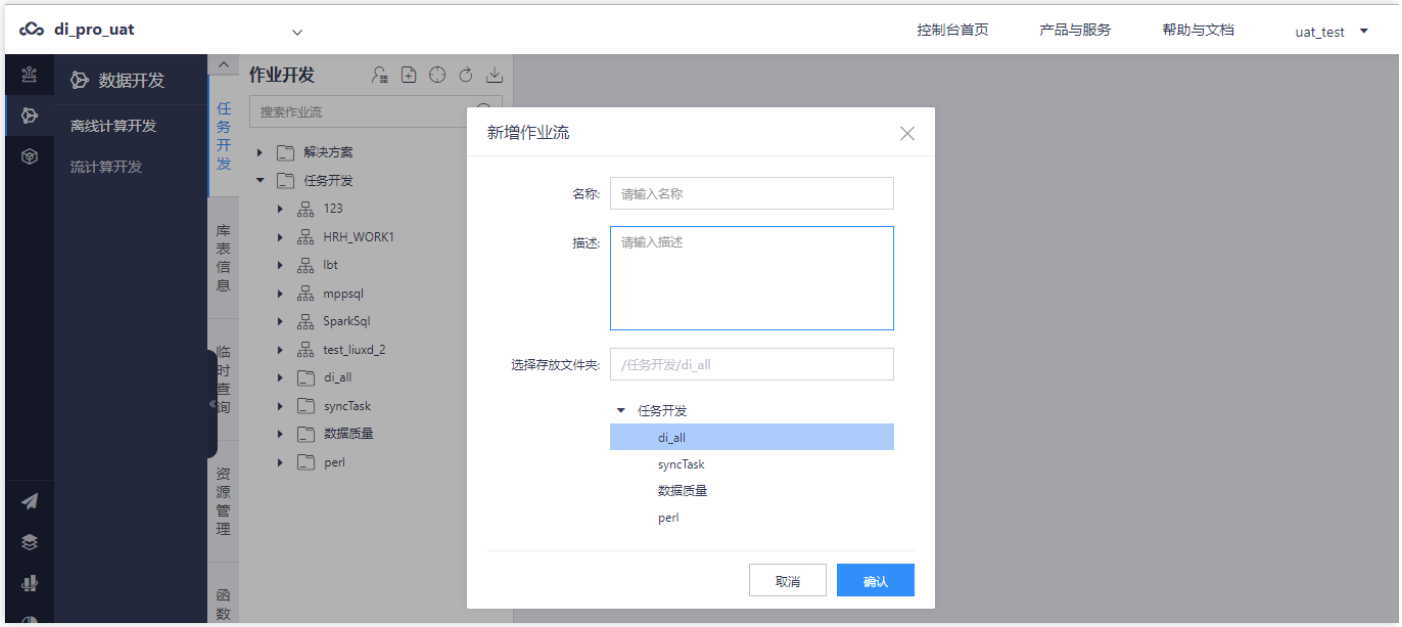
最近更新时间: 2019-11-13 03:42:23

1. 点击【项目空间-我的项目】显示项目列表页面，点击一个有权限的项目。



2. 点击【离线计算开发】，点击新建作业流，进行作业流创建。输入作业流名称，点击确定生成新的作业流。

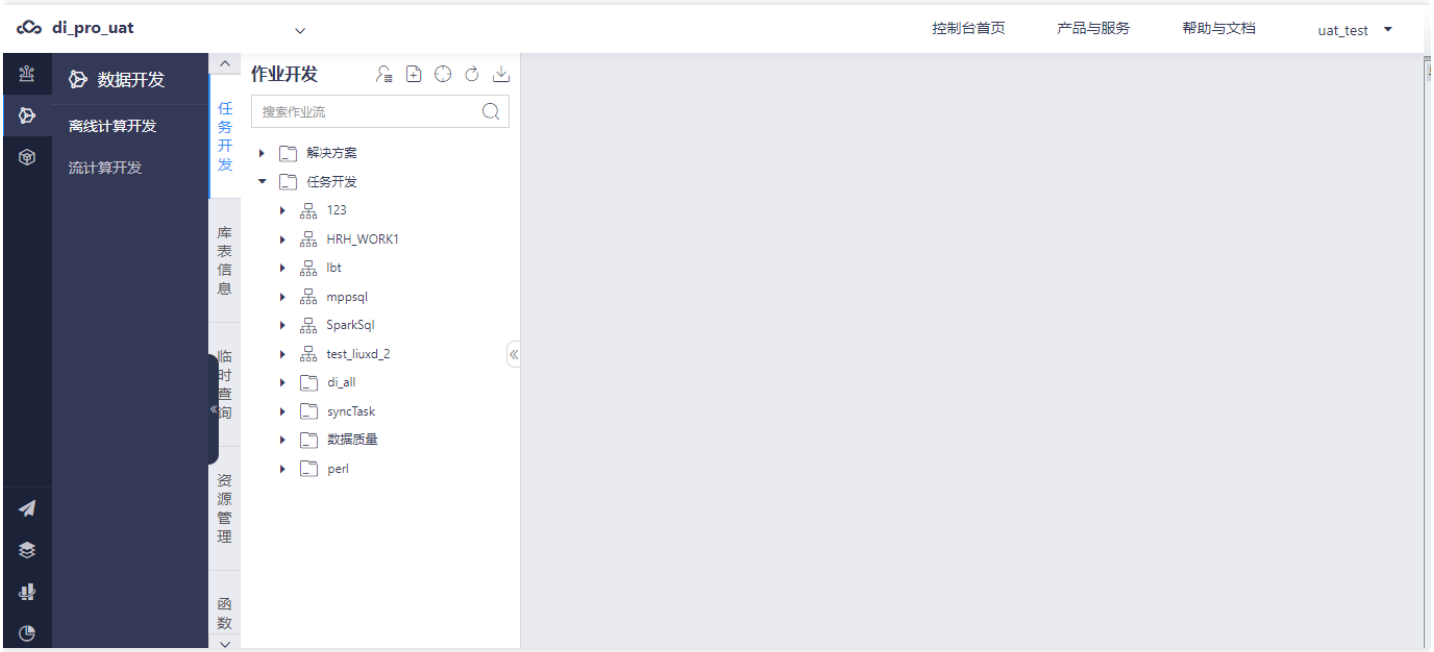


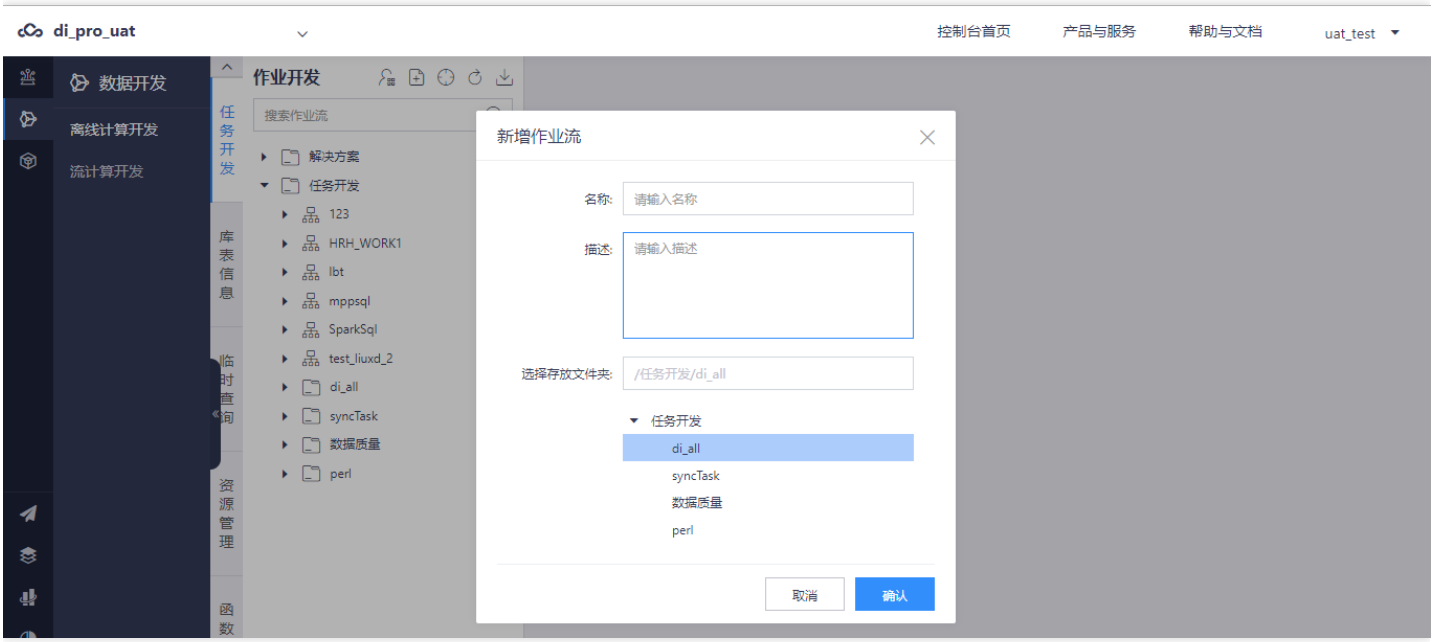

di_pro_uat
控制台首页
产品与服务
帮助与文档
uat_test
数据开发
离线计算开发
流计算开发
作业开发
搜索作业流
任务开发
库表信息
临时查询
资源管理
函数
解决方案
任务开发
123
HRH_WORK1
lbt
mppsql
SparkSql
test_liuxd_2
di_all
syncTask
数据质量
perl
新增作业流
名称:
请输入名称
描述:
请输入描述
选择存放文件夹:
/任务开发/di_all
任务开发
di_all
syncTask
数据质量
perl
取消
确认

操作指南
新建作业流

最近更新时间: 2019-11-13 06:37:00

前置条件：用户有权限在项目下新建作业流且同一个项目下作业流名称不能重复。 操作步骤：用户任务开发页面下选中一个文件夹，在文件夹下新建作业流，填写作业流描述，点击【确认】。

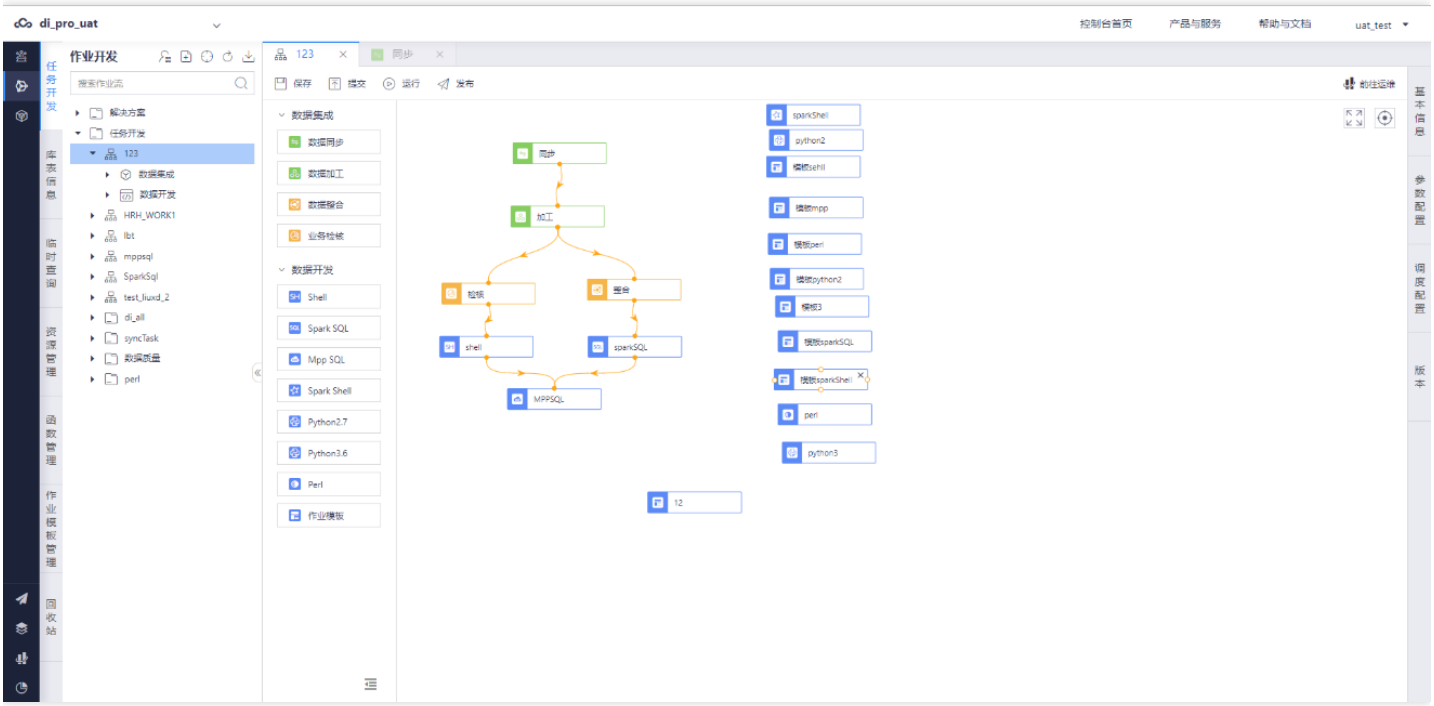


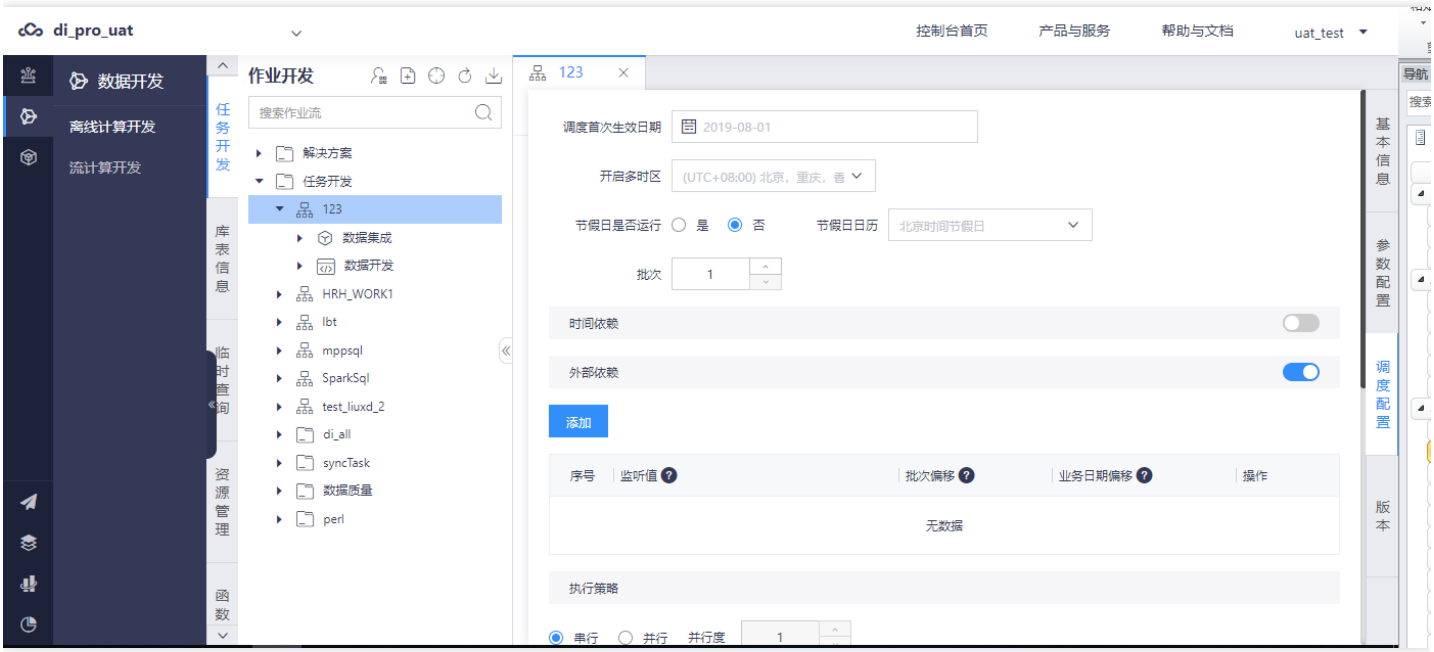


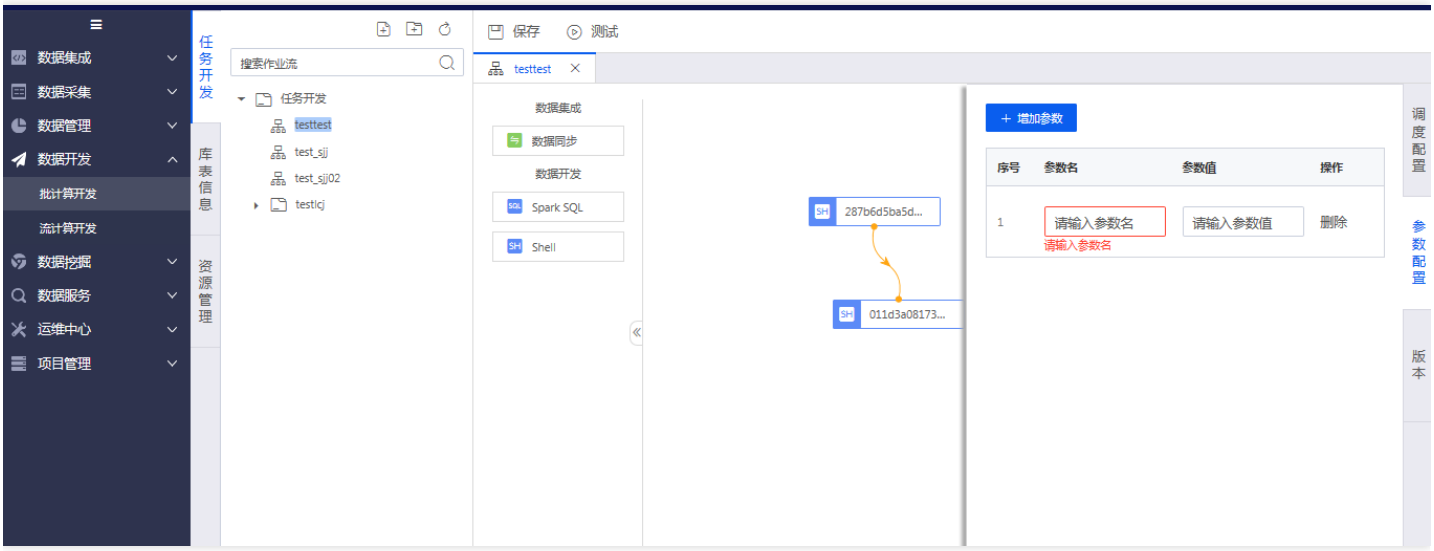
编辑作业流

最近更新时间: 2019-11-13 06:43:01

前置条件：存在的一个开发态的作业流。 操作步骤：作业开发人员拖拽插件，连线形成作业流，并配置作业流的全局参数，依赖配置，调度配置。





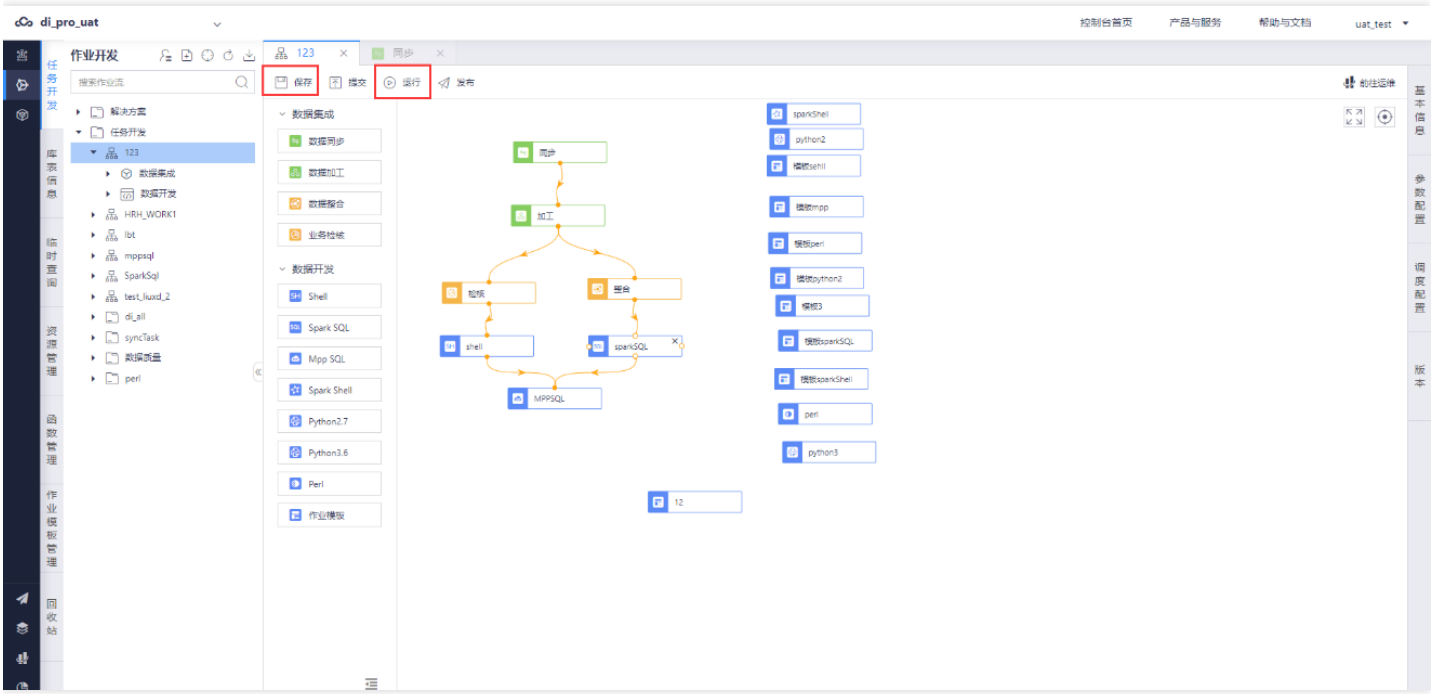

数据集成
数据采集
数据管理
数据开发
批计算开发
流计算开发
数据挖掘
数据服务
运维中心
项目管理
任务开发
库表信息
资源管理
搜索作业流
任务开发
testtest
test_sjj
test_sjj02
testcj
保存
测试
testtest
数据集成
数据同步
数据开发
Spark SQL
Shell
287b6d5ba5d...
011d3a08173...
增加参数
序号
参数名
参数值
操作
1
请输入参数名
请输入参数值
删除
请输入参数名
调度配置
参数配置
版本

测试作业流

最近更新时间: 2019-11-13 07:08:11

前置条件：已经存在的一个开发态的作业流。 操作步骤：

1. 用户在作业流编辑页面，点击【保存】按钮，保存作业流。
2. 用户作业流编辑页面，点击【测试】按钮，将作业流提交给调度在测试环境测试。

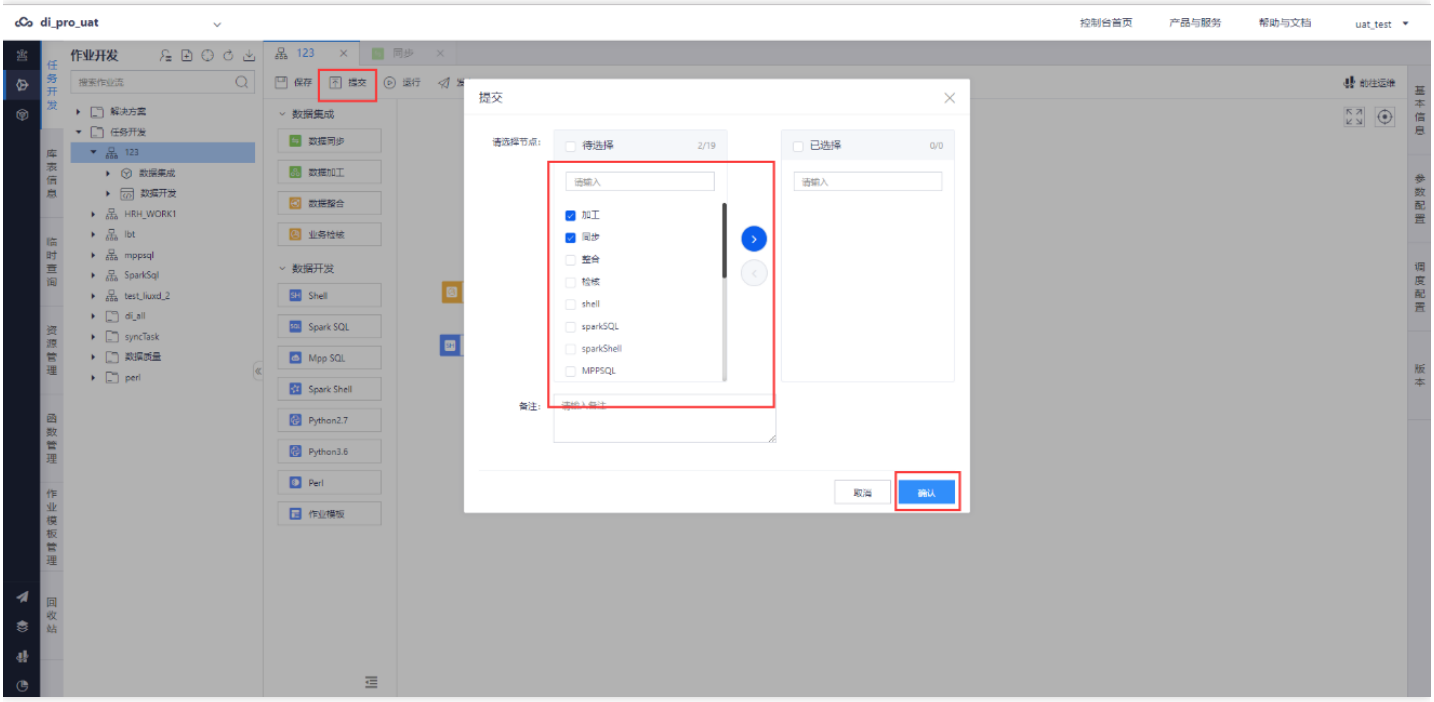


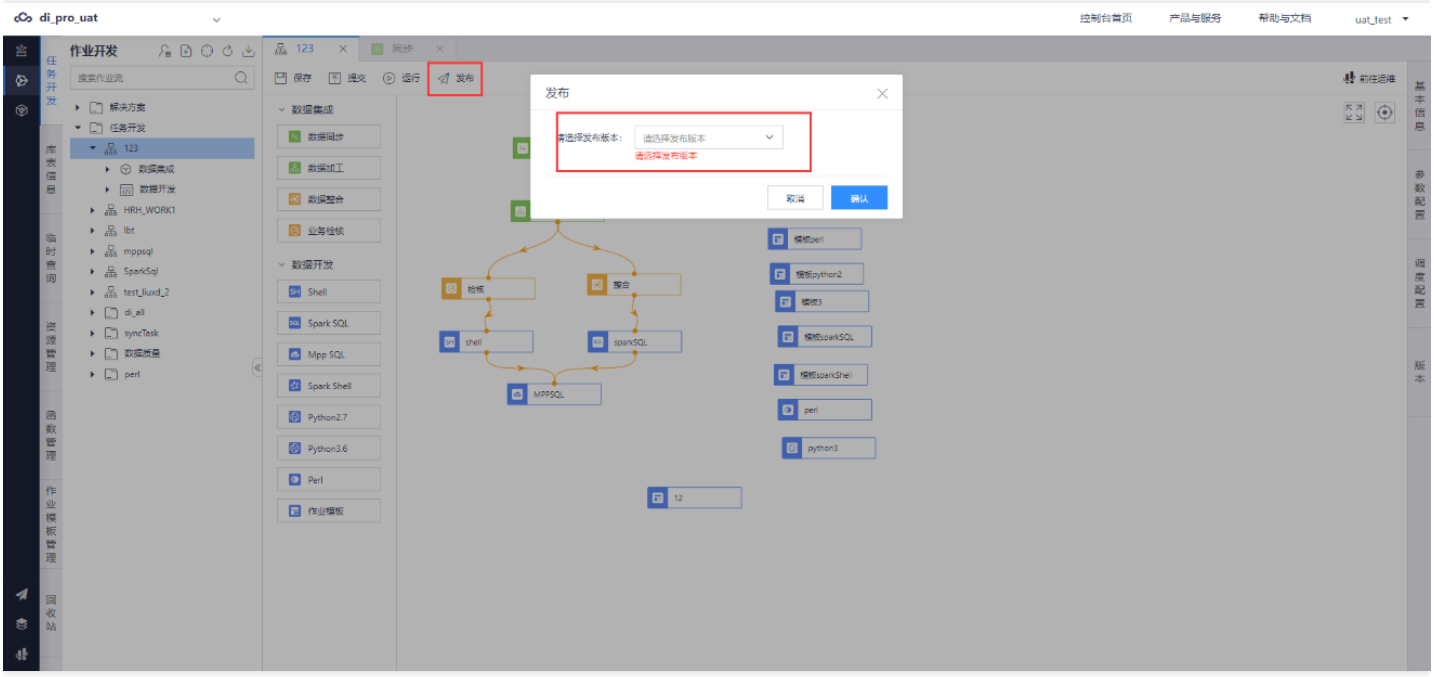
提交作业流

最近更新时间: 2019-11-13 07:08:11

前置条件：已经存在的一个开发态的作业流。 操作步骤：

1. 用户在作业流编辑页面，点击【提交】按钮，提交作业流。
2. 进入到运维中心的【任务管理】界面找到相应的计算作业，进行作业流的发布。

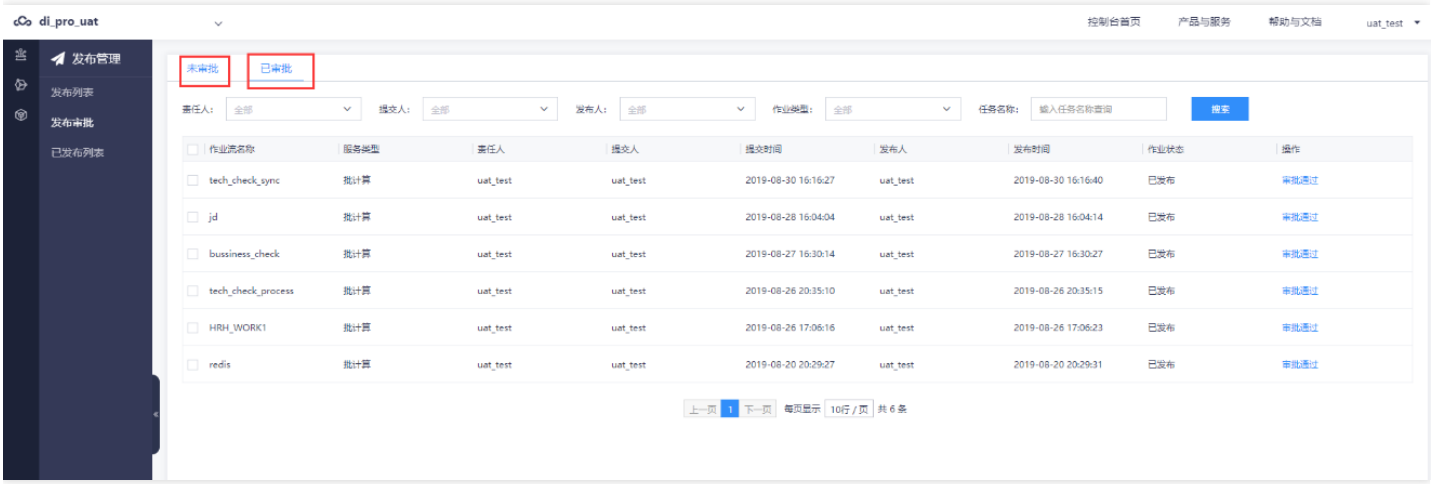

di_pro_uat
作业开发
提交
Shell
Spark SQL
Mpp SQL
Spark Shell
Python2.7
Python3.6
Perl
作业模板

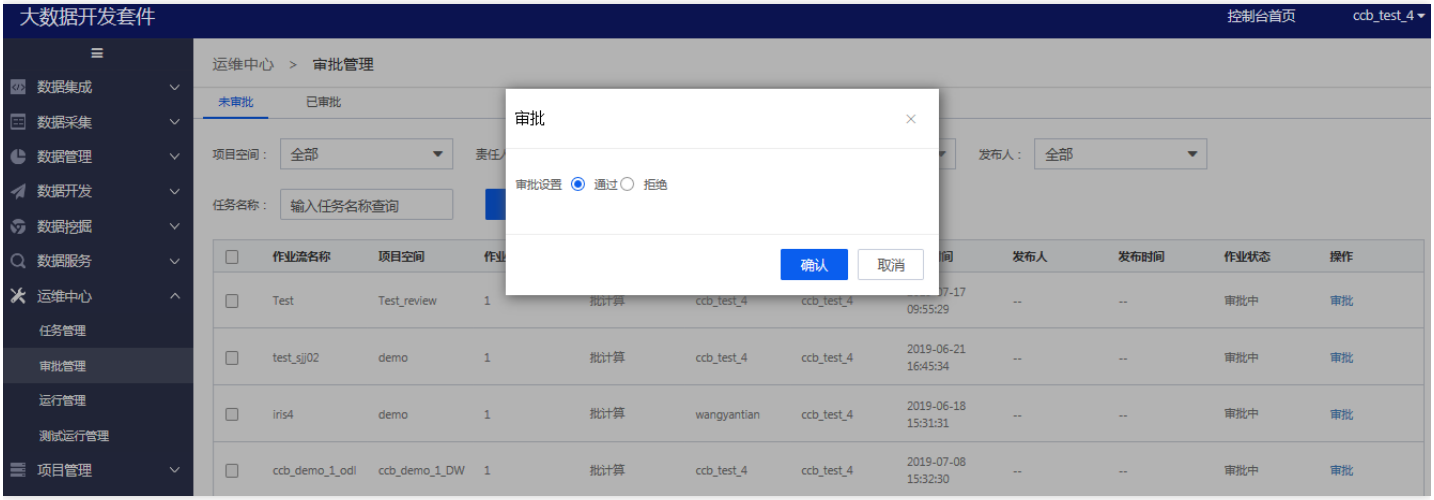

di_pro_uat
作业开发
发布
Shell
Spark SQL
Mpp SQL
Spark Shell
Python2.7
Python3.6
Perl
作业模板

发布作业流

最近更新时间: 2019-11-13 07:08:11

前置条件：作业流已经保存生成版本。 操作步骤：开发人员在发布列表中选中一条已提交的作业流，点击发布申请，项目管理员审批通过后会将作业流信息提交到智能调度。

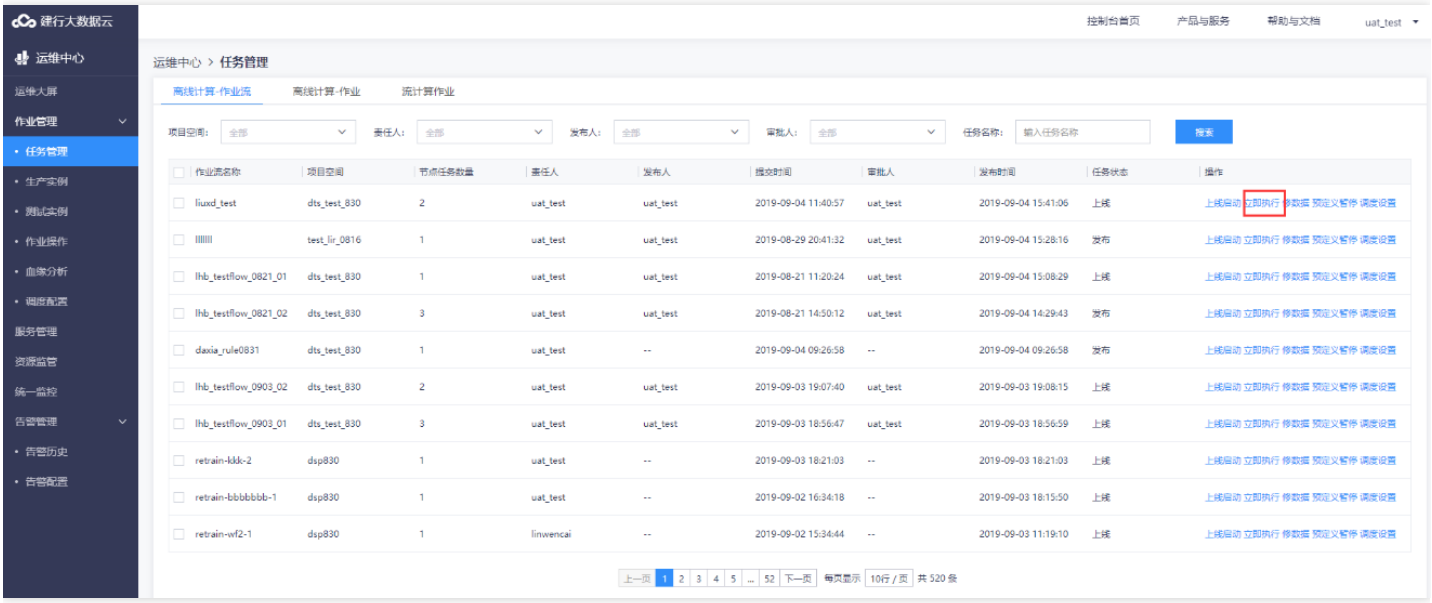




立刻执行作业流

最近更新时间: 2019-11-13 07:07:29

前置条件：作业流处于已发布状态并通过审批。 操作步骤：用户在发布列表中选中一条已发布的作业流，点击立即执行，选中要执行的批次号，数据开发会将作业流的配置信息上传到智能调度。

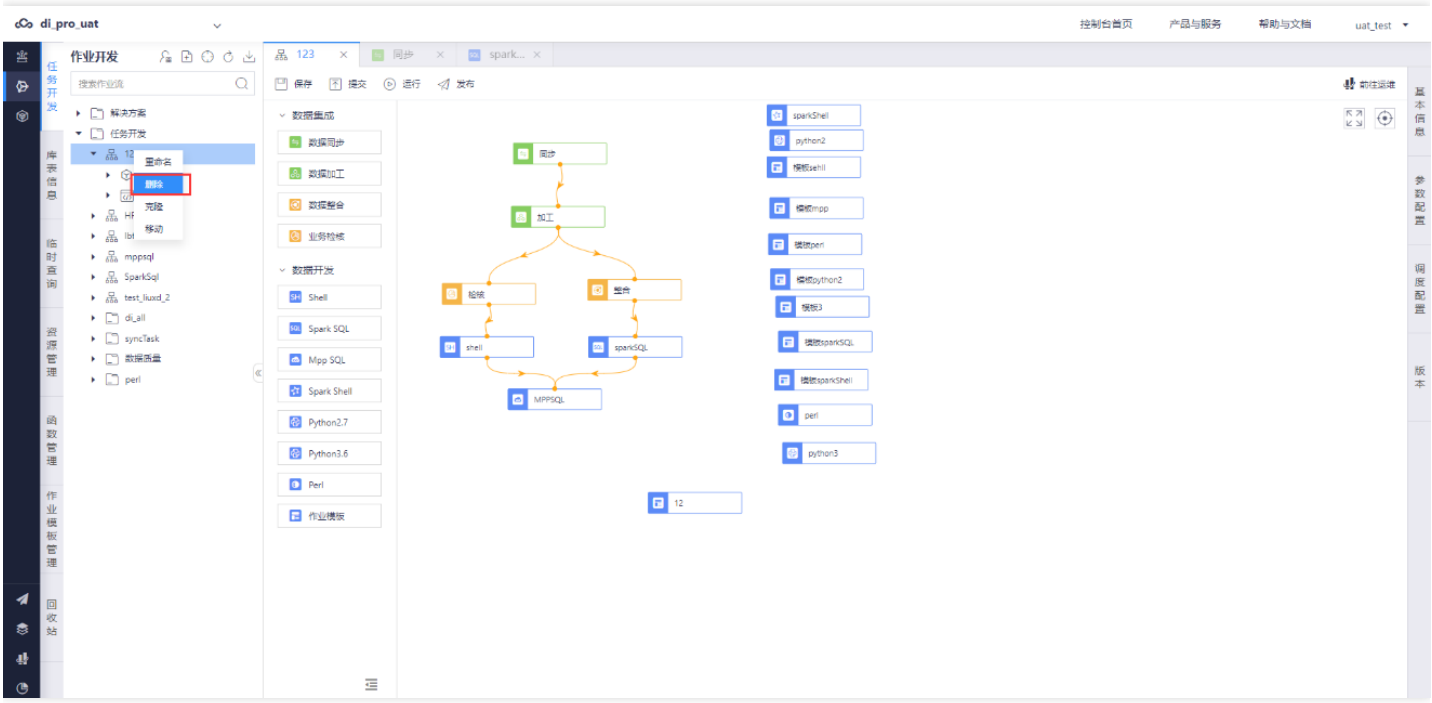


删除作业流

最近更新时间: 2019-11-13 07:19:57

前置条件：已经存在的一个开发态的作业流，且该作业流处于未发布或已下线状态。 操作步骤：

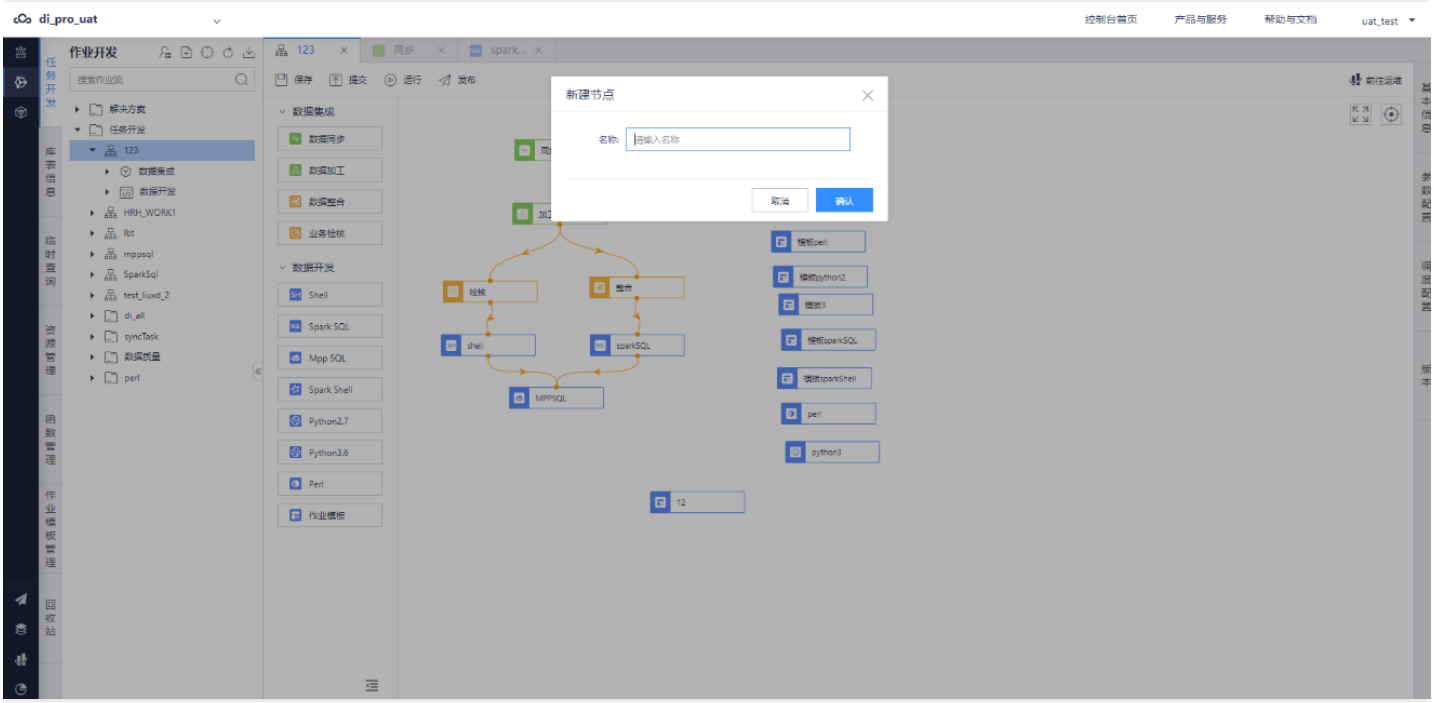
1. 用户在作业流编辑页面选中一条作业流。
2. 右键删除操作，如果该作业流有实例在运行，则删除失败，如果无实例运行，删除成功。



新建作业

最近更新时间: 2019-11-13 07:11:02

前置条件：已经存在的一个开发态的作业流。 操作步骤：用户在作业流编辑页面选中一条作业流。拖拉拽一个插件新建一个作业, 如。



编辑作业

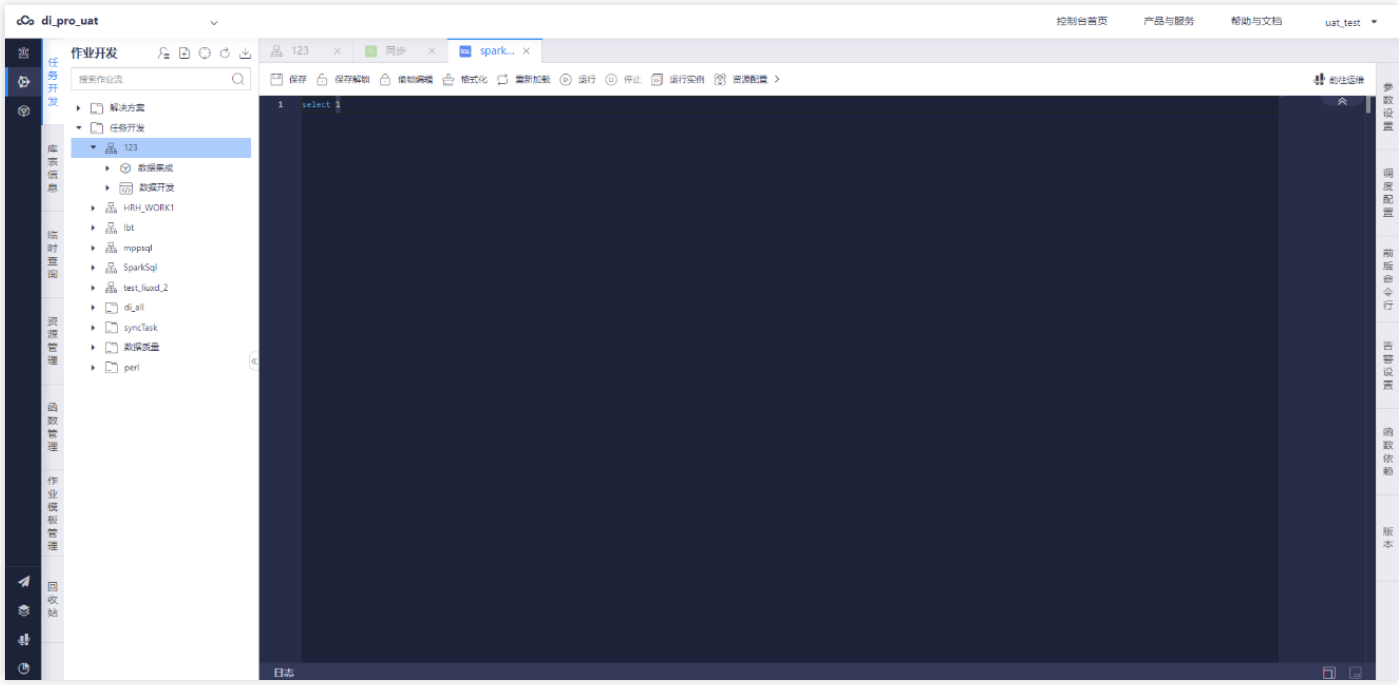
最近更新时间: 2019-11-13 07:19:57

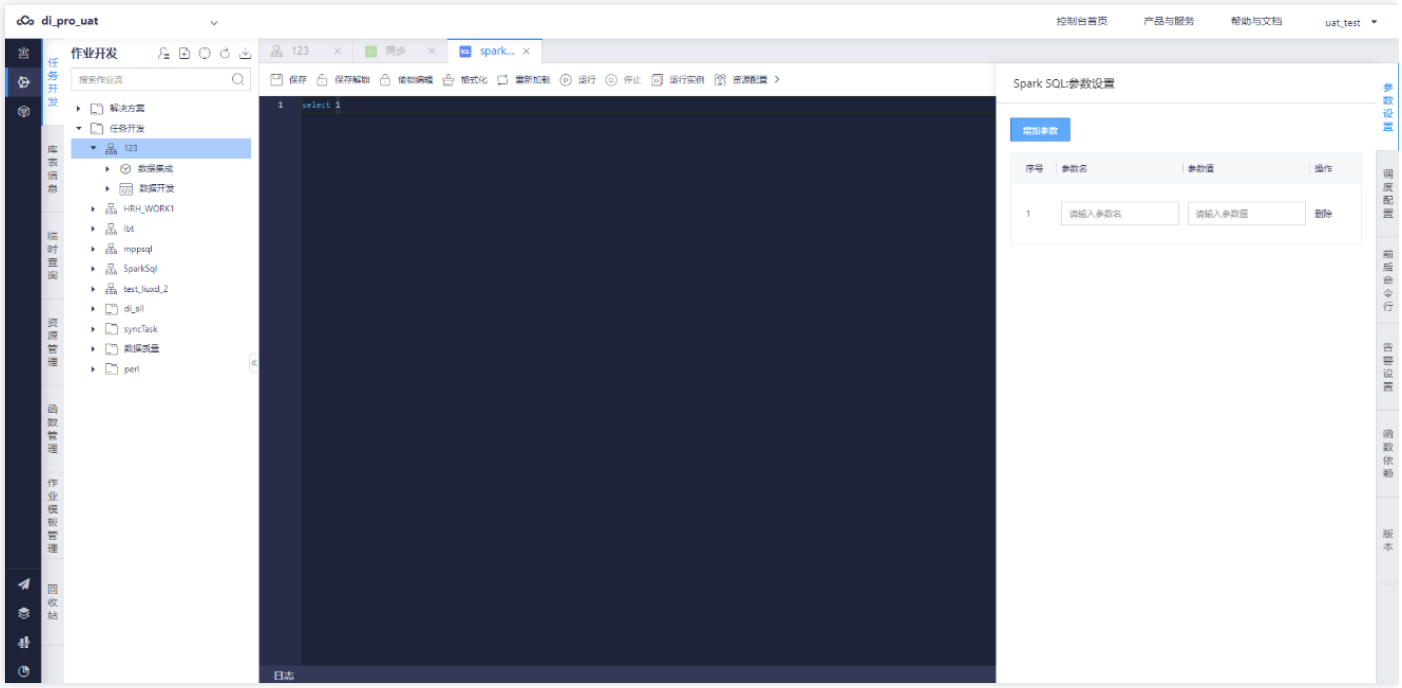
前置条件：已经存在的一个开发态的作业流并且里面包含作业。 操作步骤：

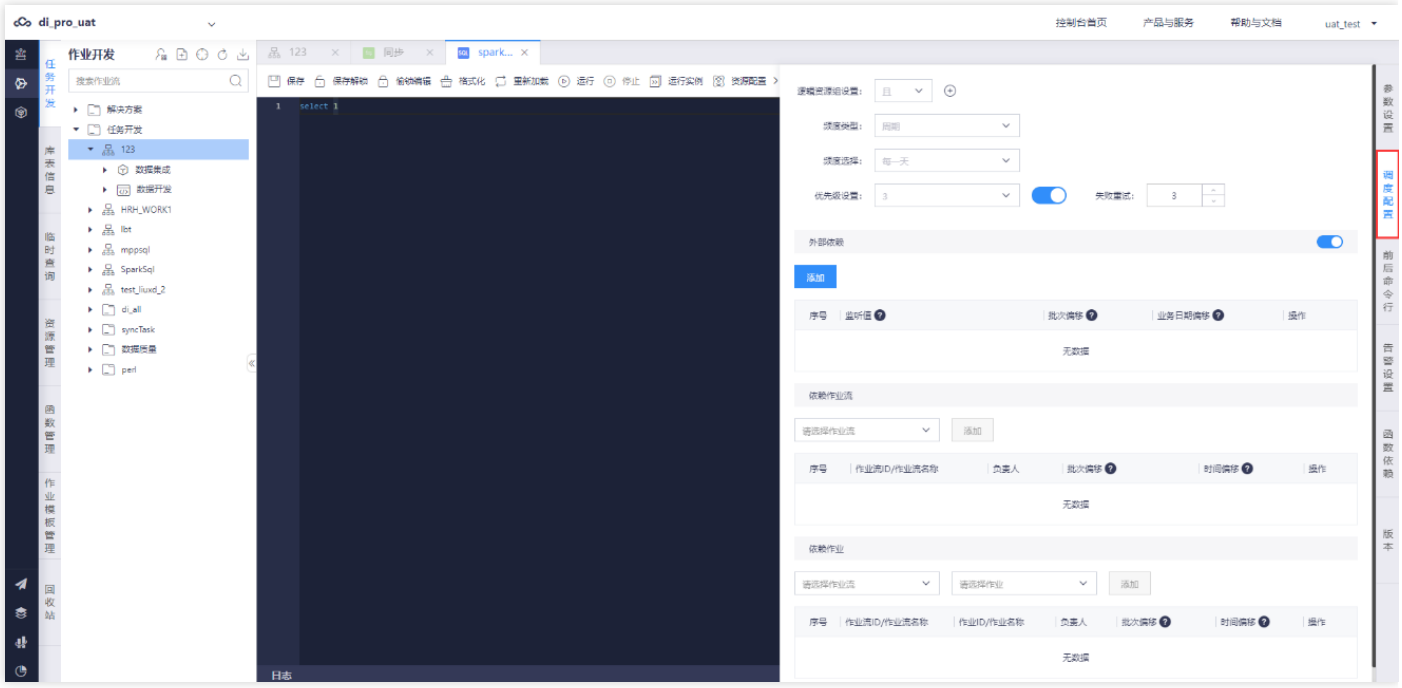
1. 用户在作业流编辑页面选中一条作业流。

2. 在作业流里选中一个作业节点，双击编辑业务。

3. 进行作业参数属性配置，作业依赖配置。


di_pro_uat
控制台首页
产品与服务
帮助与文档
uat_test
作业开发
select 1
日志

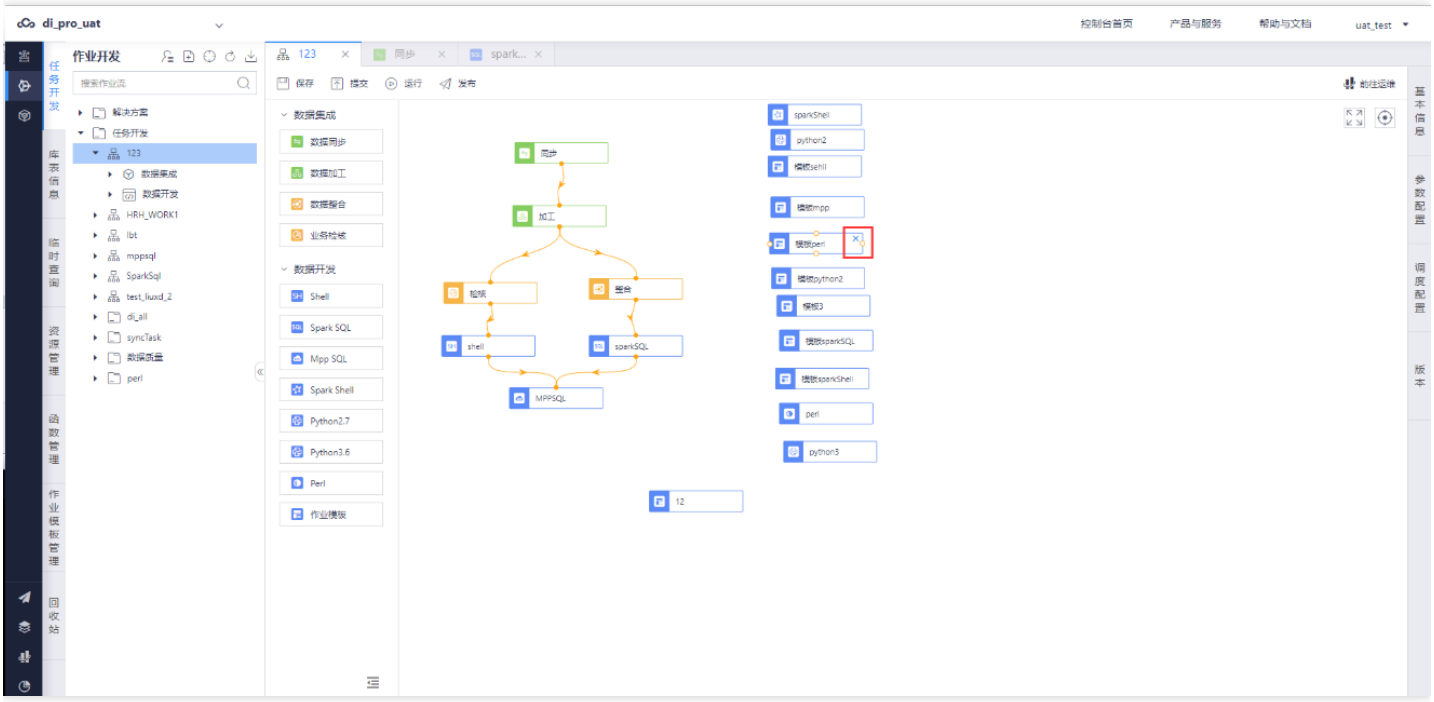

di_pro_uat
控制台首页
产品与服务
帮助与文档
uat_test
作业开发
select 1
Spark SQL参数设置
日志


di_pro_uat
控制台首页
产品与服务
帮助与文档
uat_test
作业开发
123
同步
spark...
保存
保存解析
检验编辑
格式化
重新加载
运行
停止
运行实例
资源配置
select 1
日志
参数设置
调度配置
前后命令行
告警设置
函数依赖
版本
外部依赖
添加
序号
监听值
批次偏移
业务日期偏移
操作
无数据
依赖作业流
依赖作业
作业流ID/作业流名称
负责人
时间偏移
作业ID/作业名称

删除作业

最近更新时间: 2019-11-13 07:19:57

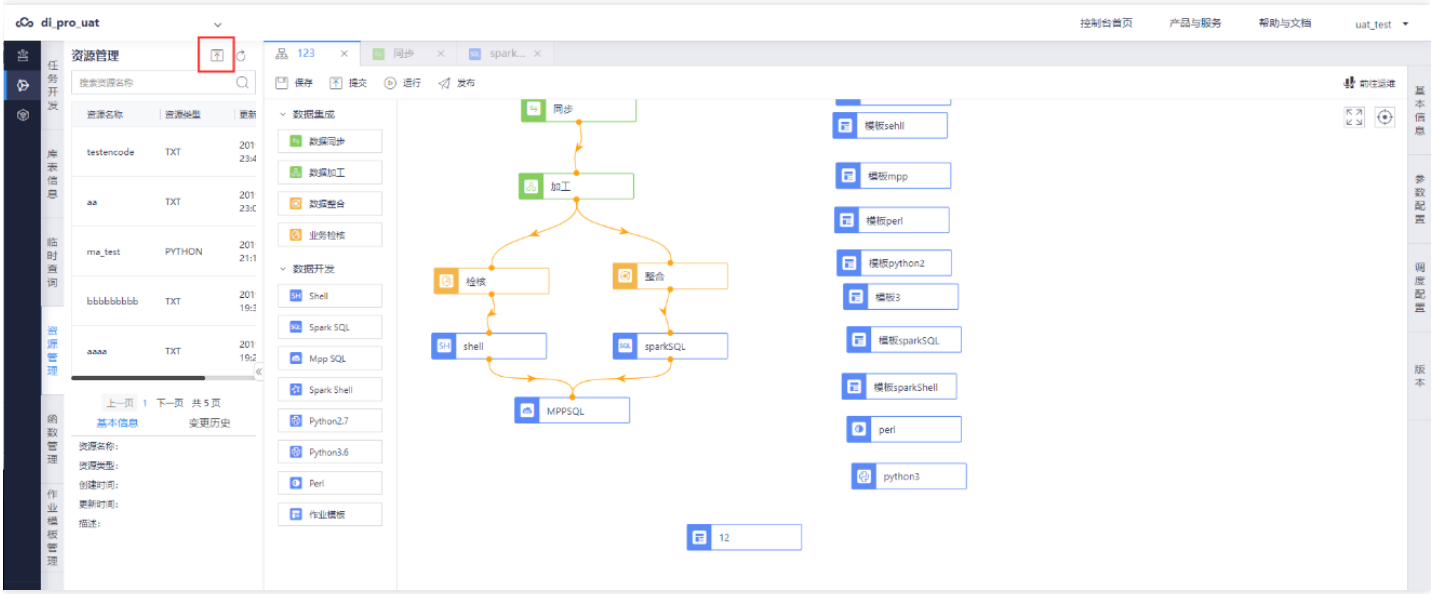
前置条件：已经存在的一个开发态作业流，里面包含一个或多个作业节点，并且作业流处于非运行态。 操作步骤：用户选中一条作业流进入作业流编辑页面，选中作业流的某个作业，右键或者直接点击作业右上角的×进行删除。

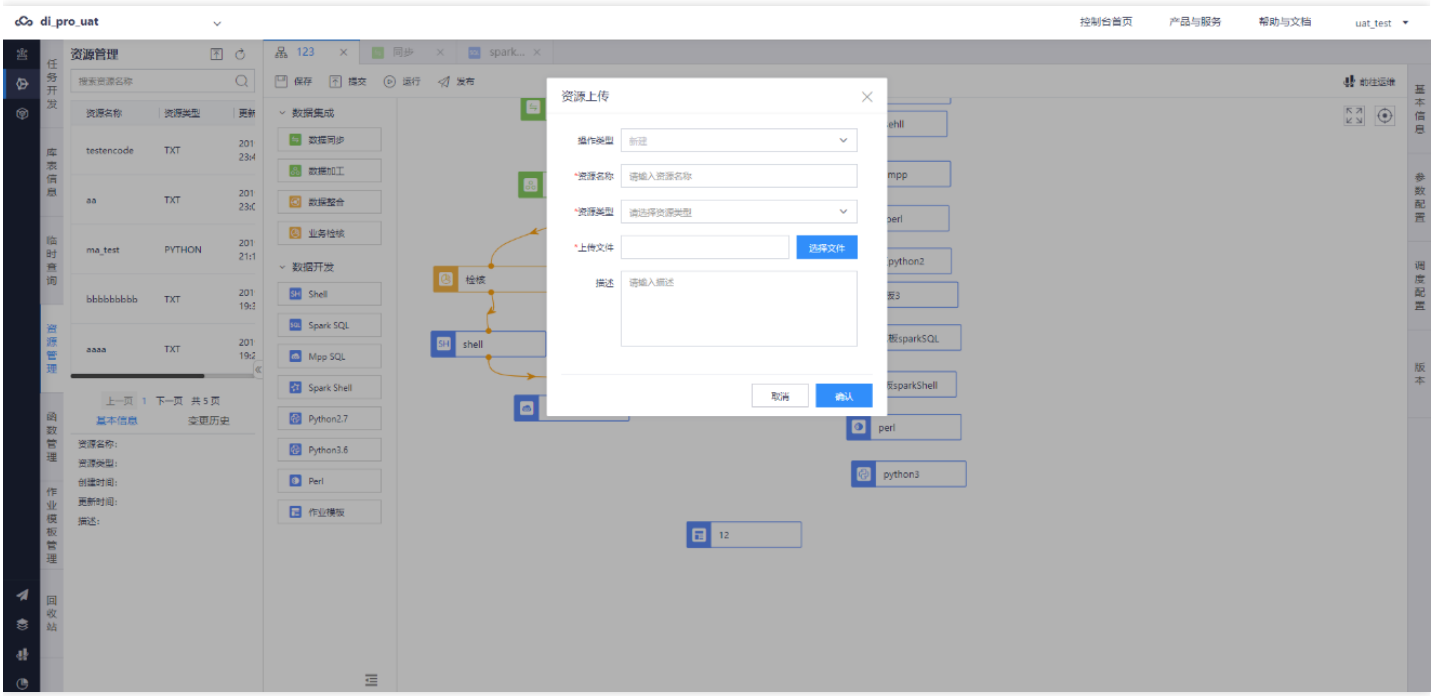


上传资源

最近更新时间: 2019-11-13 07:29:33

前置条件：不存在同名资源，资源名称符合规范，资源大小符合规范 操作步骤：开发人员点击创建资源按扭，填写资源的必要信息，选择要上传的资源，点击确定。





引用资源

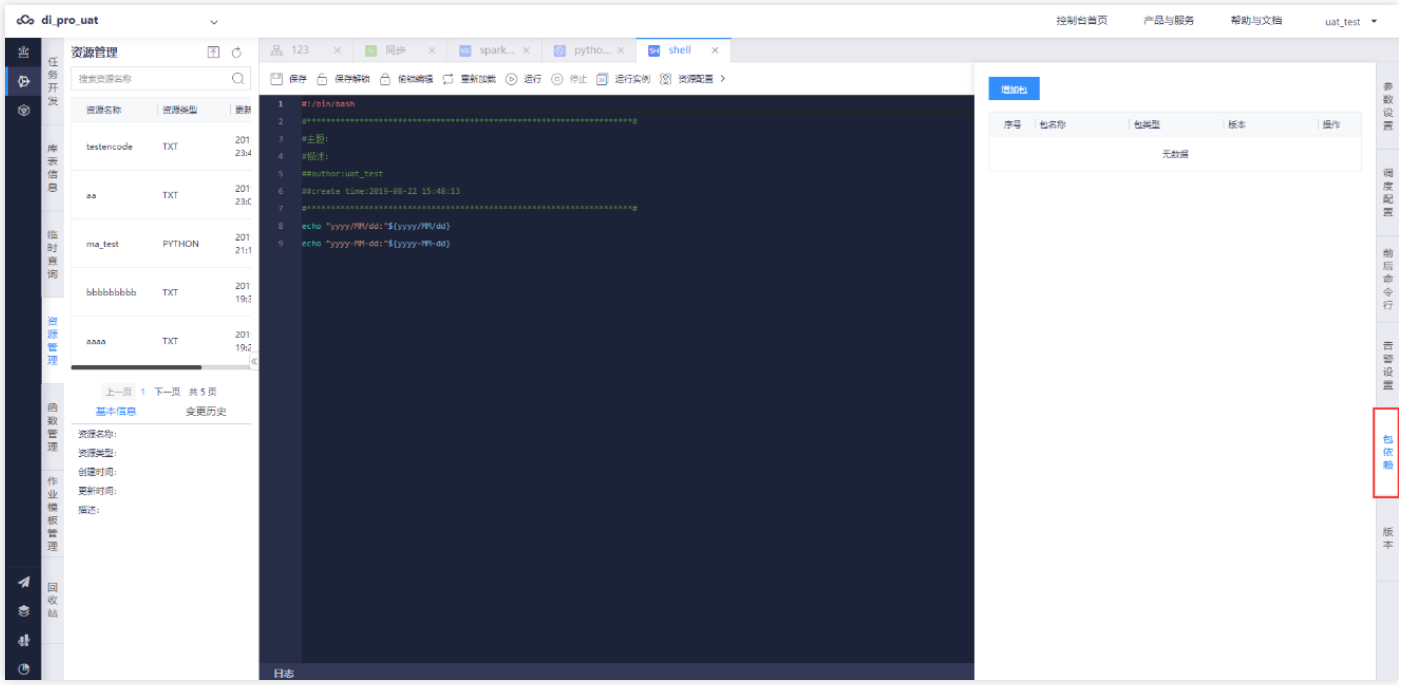
最近更新时间: 2019-11-13 07:29:33

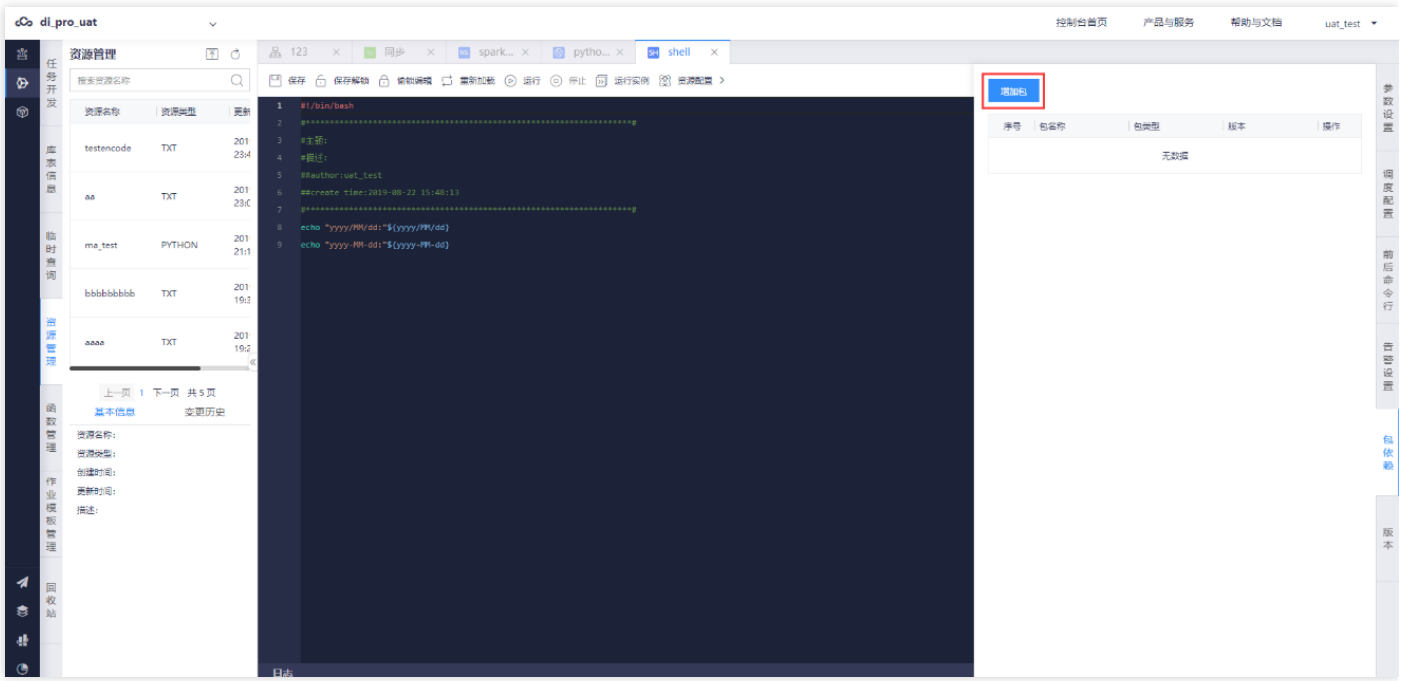
前置条件：当前租户下有可用资源。 操作步骤：

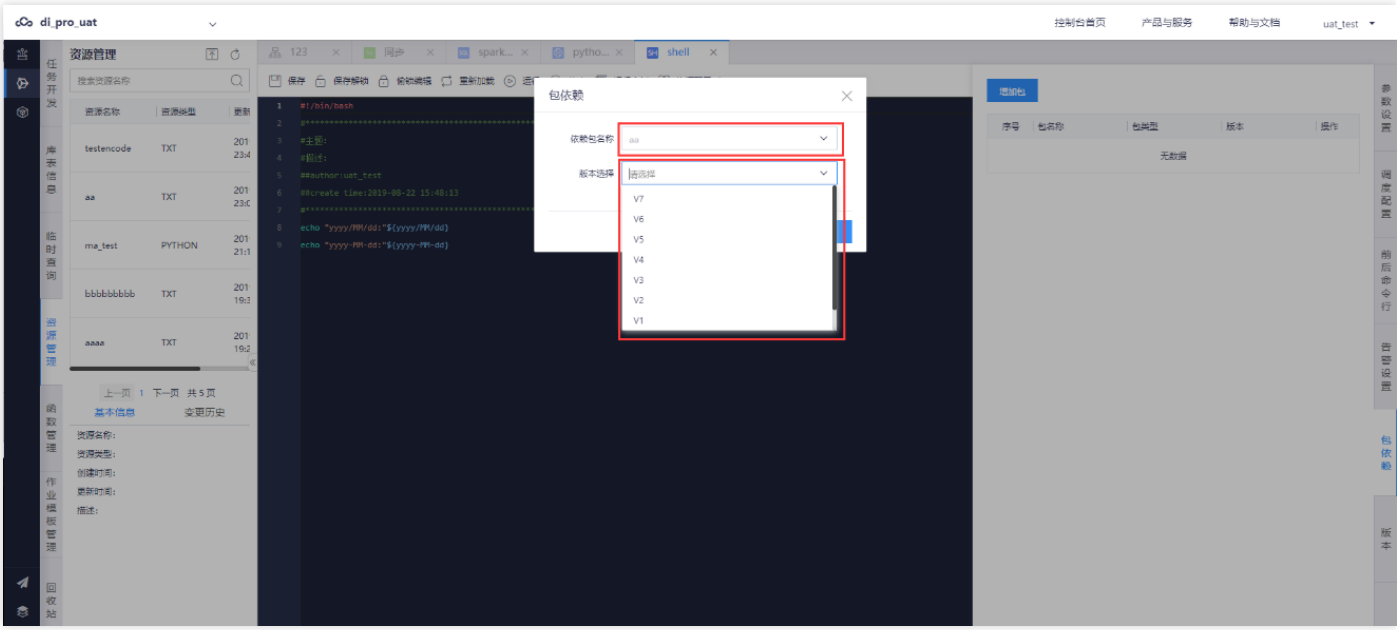
1. 用户进行批计算数据开发，进入包依赖模块。

2. 点击【添加包】。

3. 选择依赖包名称和版本。


di_pro_uat
控制台首页
产品与服务
帮助与文档
uat_test
资源管理
testencode
TXT
aa
TXT
ma_test
PYTHON
bbbbbbbbb
TXT
aaaa
TXT
基本信息
变更历史
123
同步
spark...
pytho...
shell
保存
保存解锁
编辑编辑
重新加载
运行
停止
运行实例
资源配置
#!/bin/bash
#主题:
#描述:
##author:uat_test
##create time:2019-08-22 15:48:13
echo "yyyy/MM/dd:"${yyyy/MM/dd}
echo "yyyy-MM-dd:"${yyyy-MM-dd}
增加包
序号
包名称
包类型
版本
操作
无数据
日志

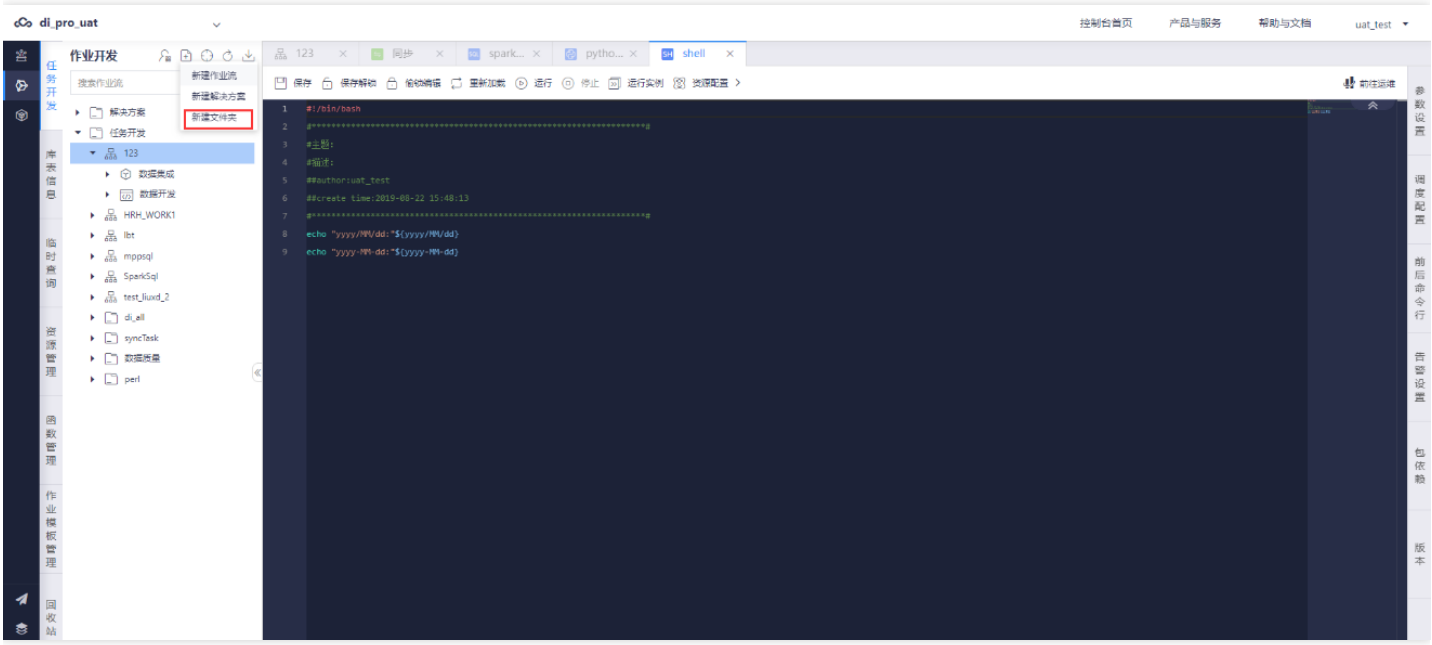

di_pro_uat
控制台首页
产品与服务
帮助与文档
uat_test
资源管理
testencode
TXT
aa
TXT
ma_test
PYTHON
bbbbbbbbb
TXT
aaaa
TXT
基本信息
变更历史
123
同步
spark...
pytho...
shell
保存
保存解锁
编辑编辑
重新加载
运行
停止
运行实例
资源配置
#!/bin/bash
#主题:
#描述:
##author:uat_test
##create time:2019-08-22 15:48:13
echo "yyyy/MM/dd:"${yyyy/MM/dd}
echo "yyyy-MM-dd:"${yyyy-MM-dd}
增加包
序号
包名称
包类型
版本
操作
无数据
日志

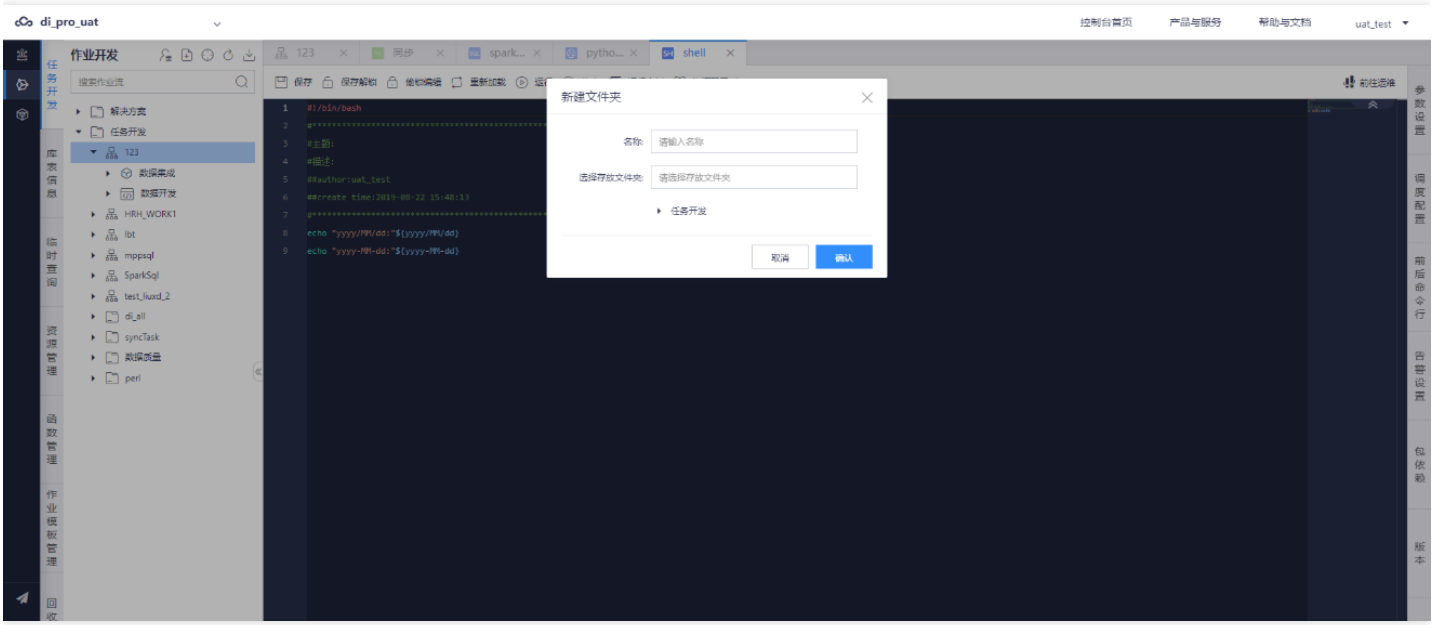

di_pro_uat
控制台首页
产品与服务
帮助与文档
uat_test
资源管理
包依赖
依赖包名称
版本选择
V7
V6
V5
V4
V3
V2
V1
增加包

新建文件夹

最近更新时间: 2019-11-13 07:53:53

前置条件：已经存在项目可用的项目空间。 操作步骤：进入任务开发界面，点击【新建文件夹】按扭，弹出新建文件夹页面，填写文件夹名称后保存，创建完成。

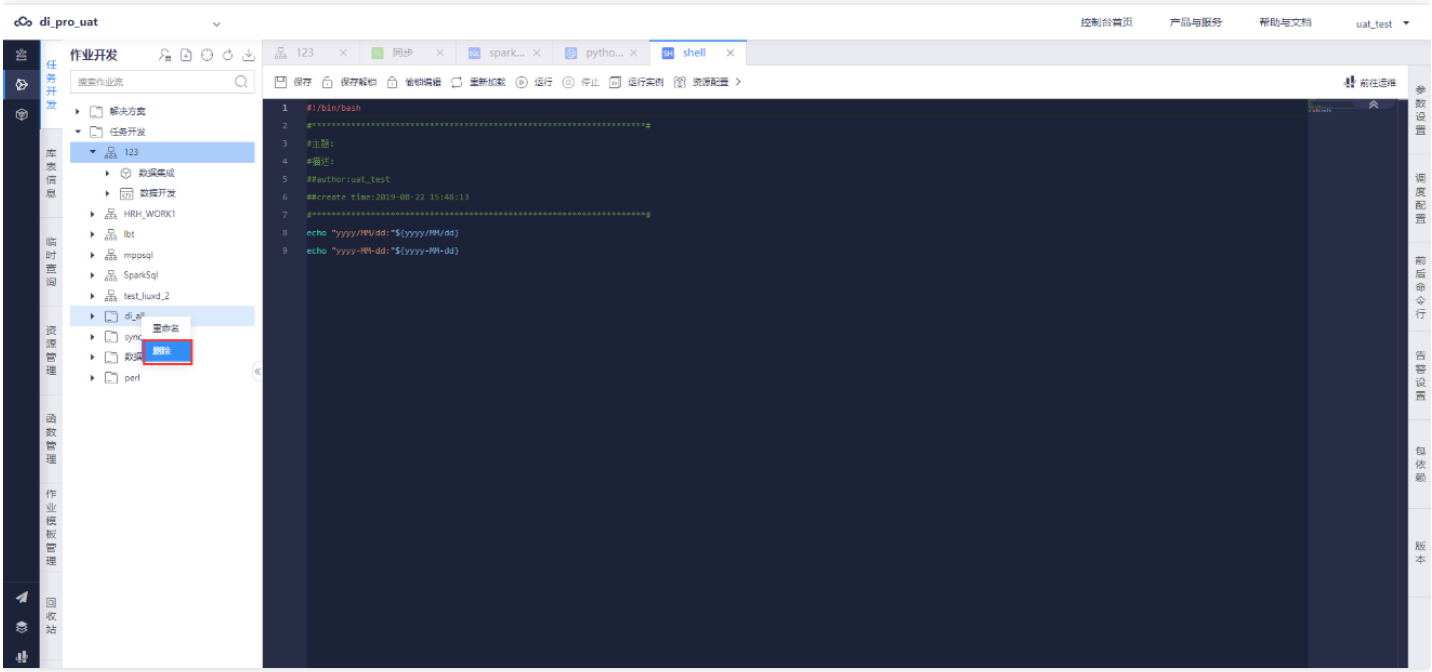




删除文件夹

最近更新时间: 2019-11-13 07:53:53

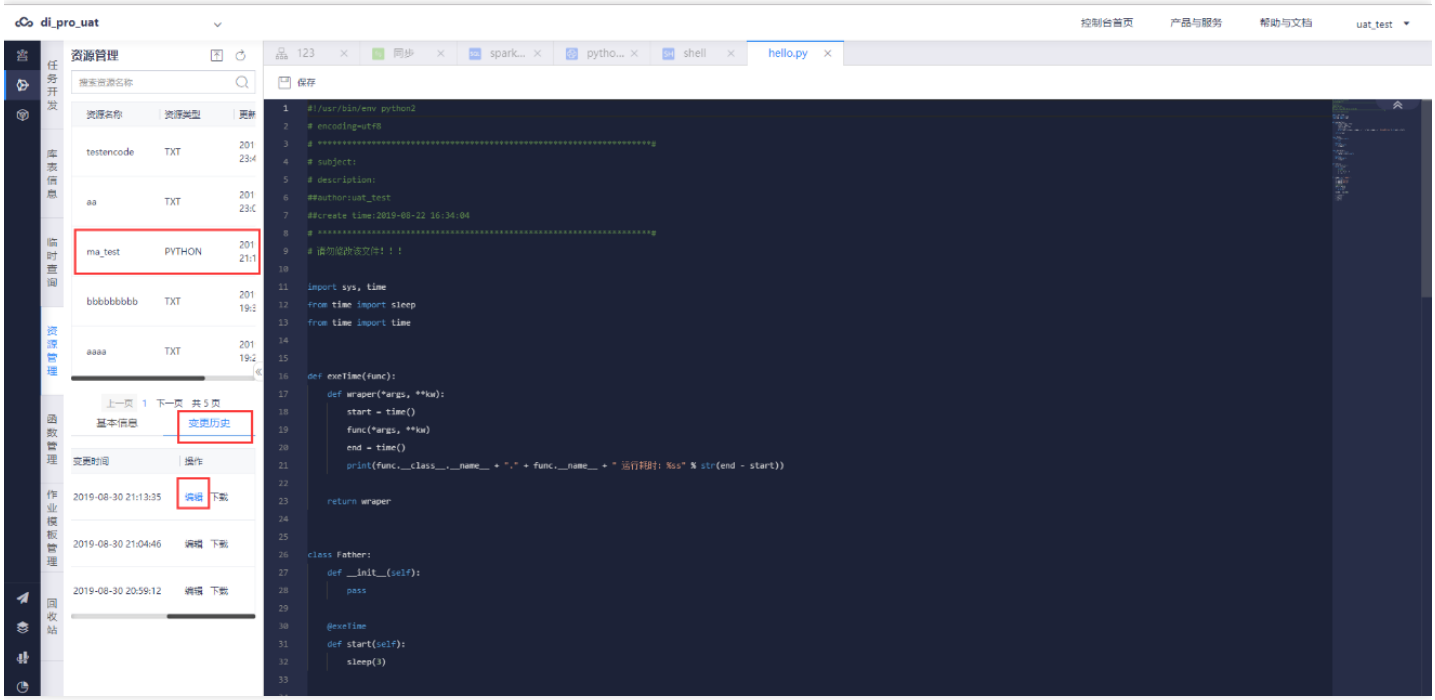
前置条件：已经存在的文件夹，并且文件夹下的作业流都是未发布状态。 操作步骤：选中文件夹，右键删除操作，提示是否要删除文件夹和下面的作业流。



在线开发资源

最近更新时间: 2019-11-13 07:53:53

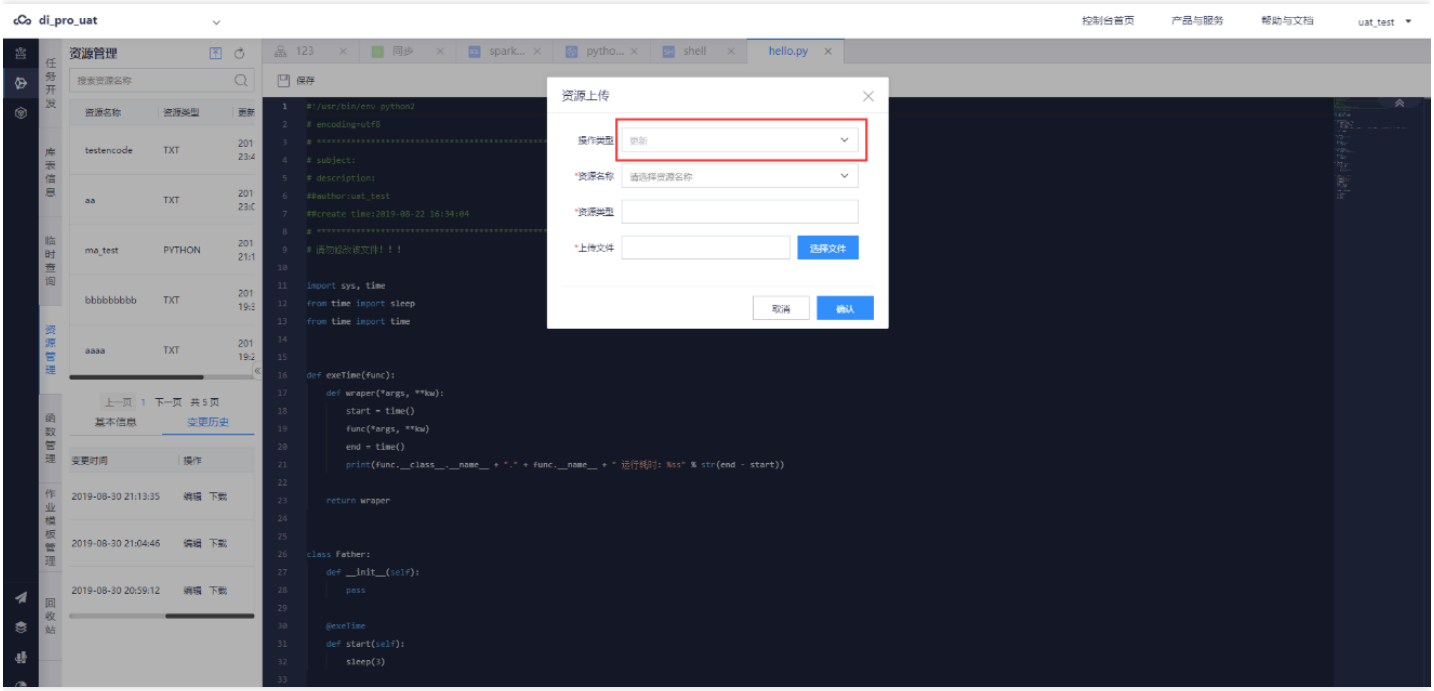
前置条件：在线开发的资源只能是文本类型，比如SQL片段、Shell脚本、JavaScript片段等，不能是需编译型源码，比JAVA源码 操作步骤：用户进入资源管理页面，打开在线开发页面，编辑文件，点击保存成资源，提交资源生成资源版本



资源升级

最近更新时间: 2019-11-13 07:53:53

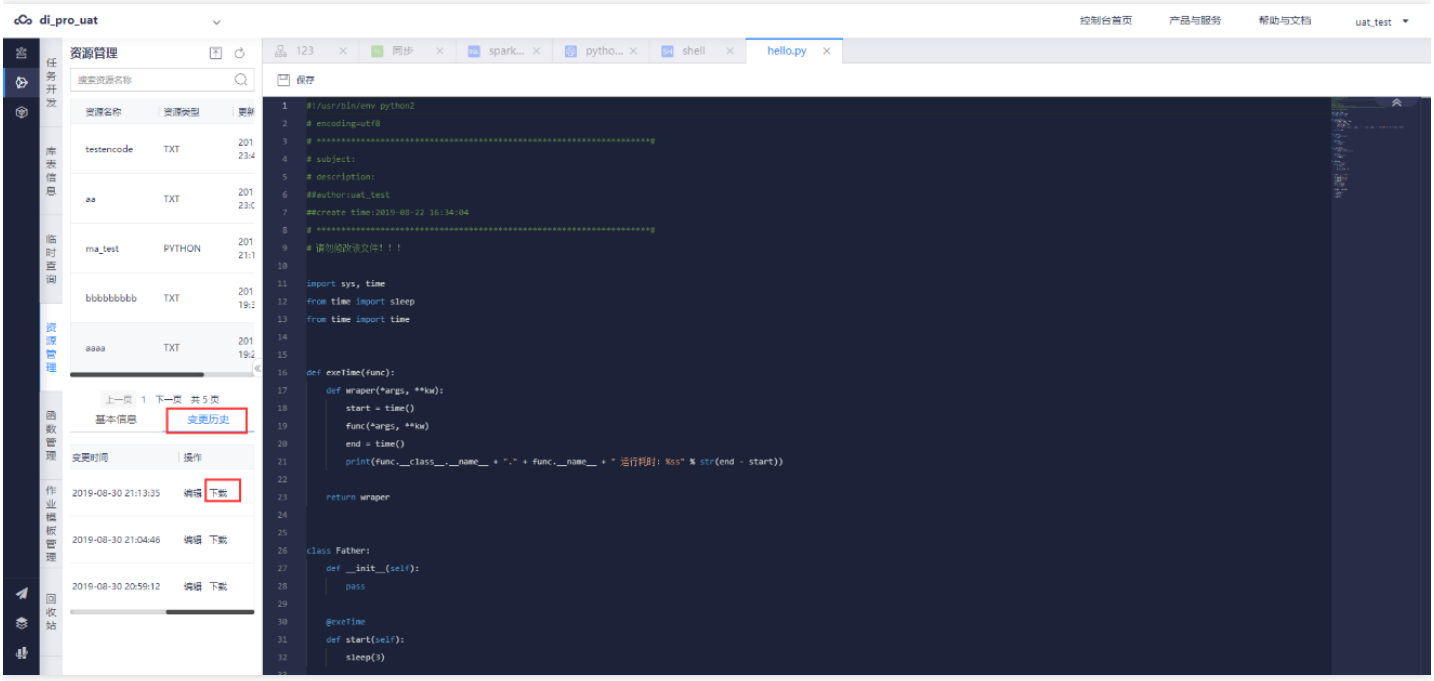
前置条件：资源升级 操作步骤：用户进入资源管理页面，选择资源更新，由用户选择更新资源的方式



资源下载

最近更新时间: 2019-11-13 07:53:53

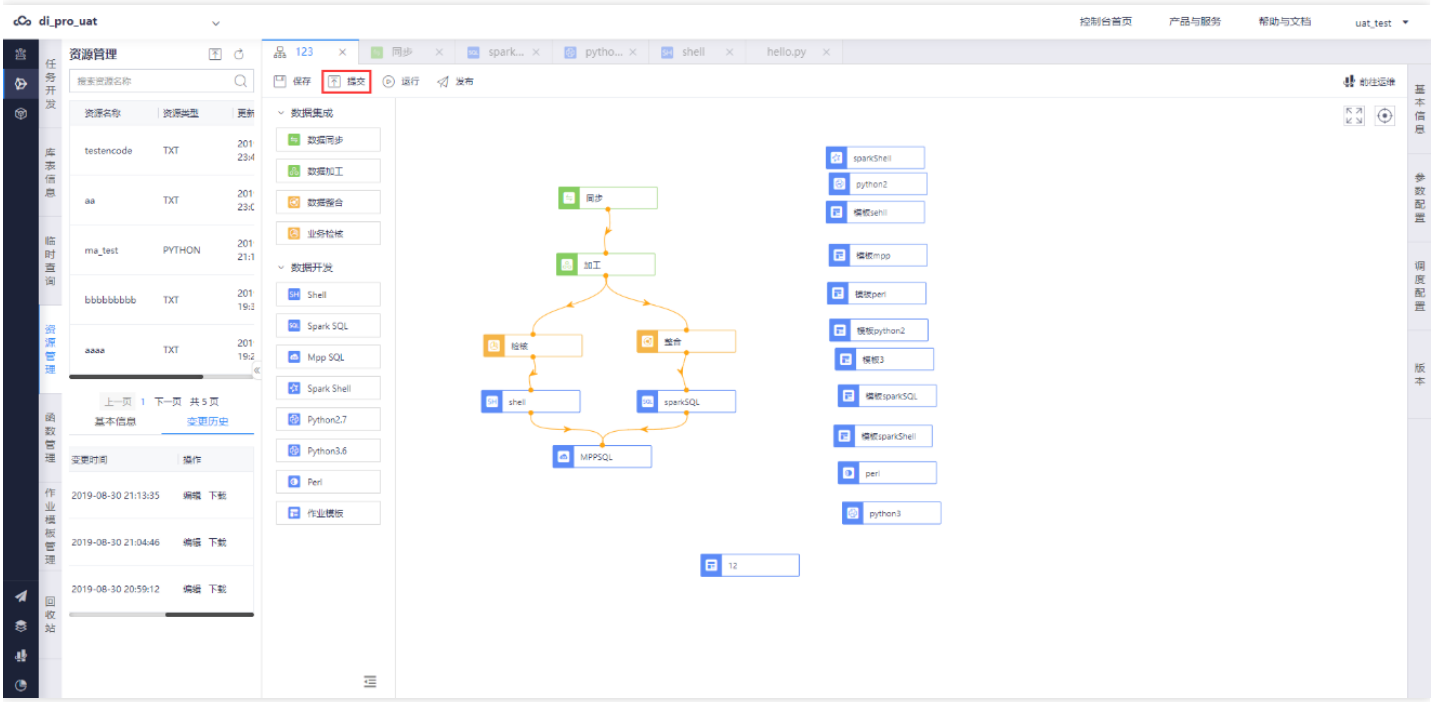
前置条件：已经存在的资源。 操作步骤：用户进入资源管理页面，选择某个资源，某个资源的版本，点击下载。



更新作业流

最近更新时间: 2019-11-13 07:53:38

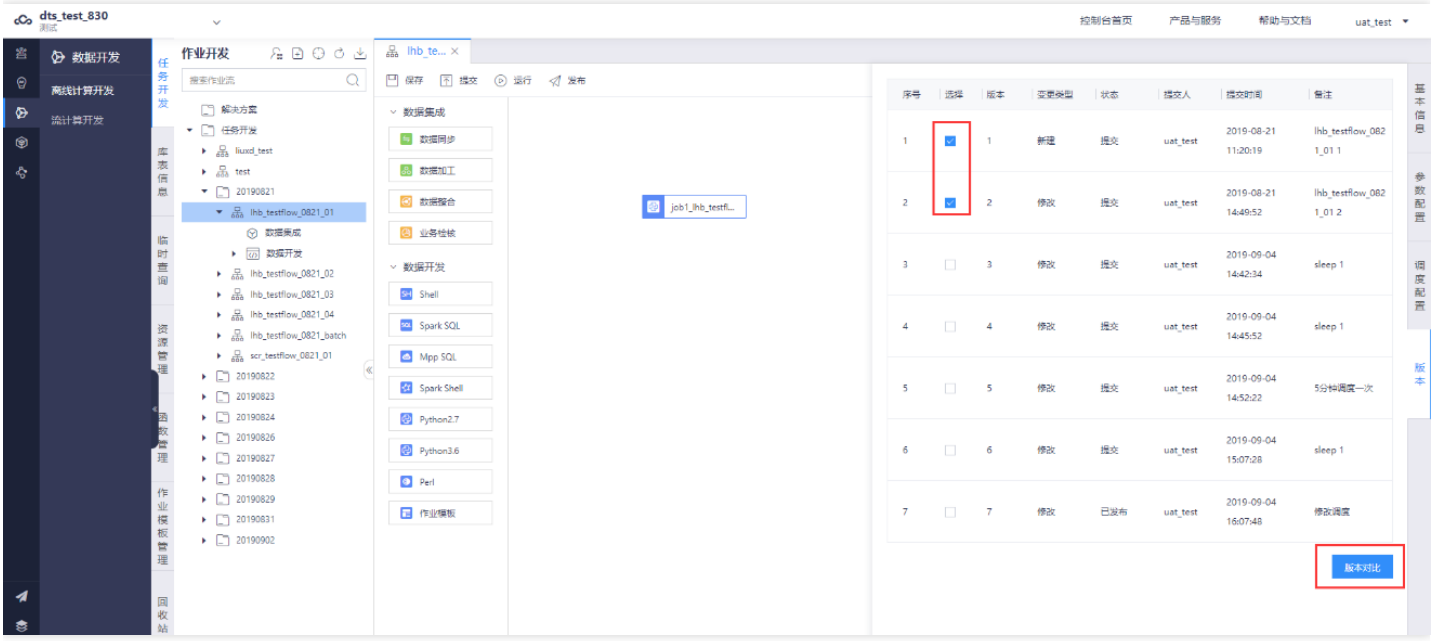
前置条件：作业节点新编辑保存过。 操作步骤：用户双击作业流，进入编辑页面，点击提交按钮，选择有内容变化的节点进行提交。提交成功后，选中的节点版本会增加1，作业流的版本也会递增1

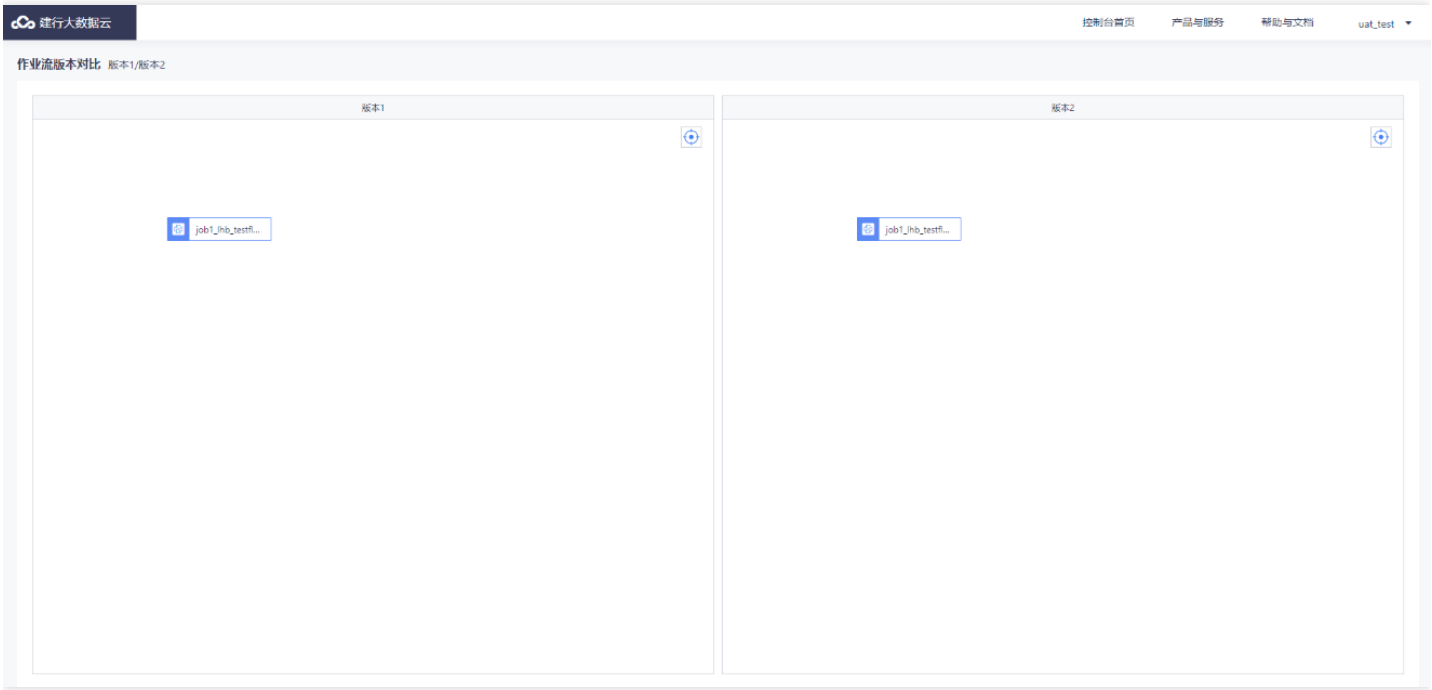


作业节点版本比对

最近更新时间: 2019-11-13 07:53:37

前置条件：作业节点新编辑保存过。 操作步骤：选中一个作业流的作业，选择版本管理，对比两个版本的变化

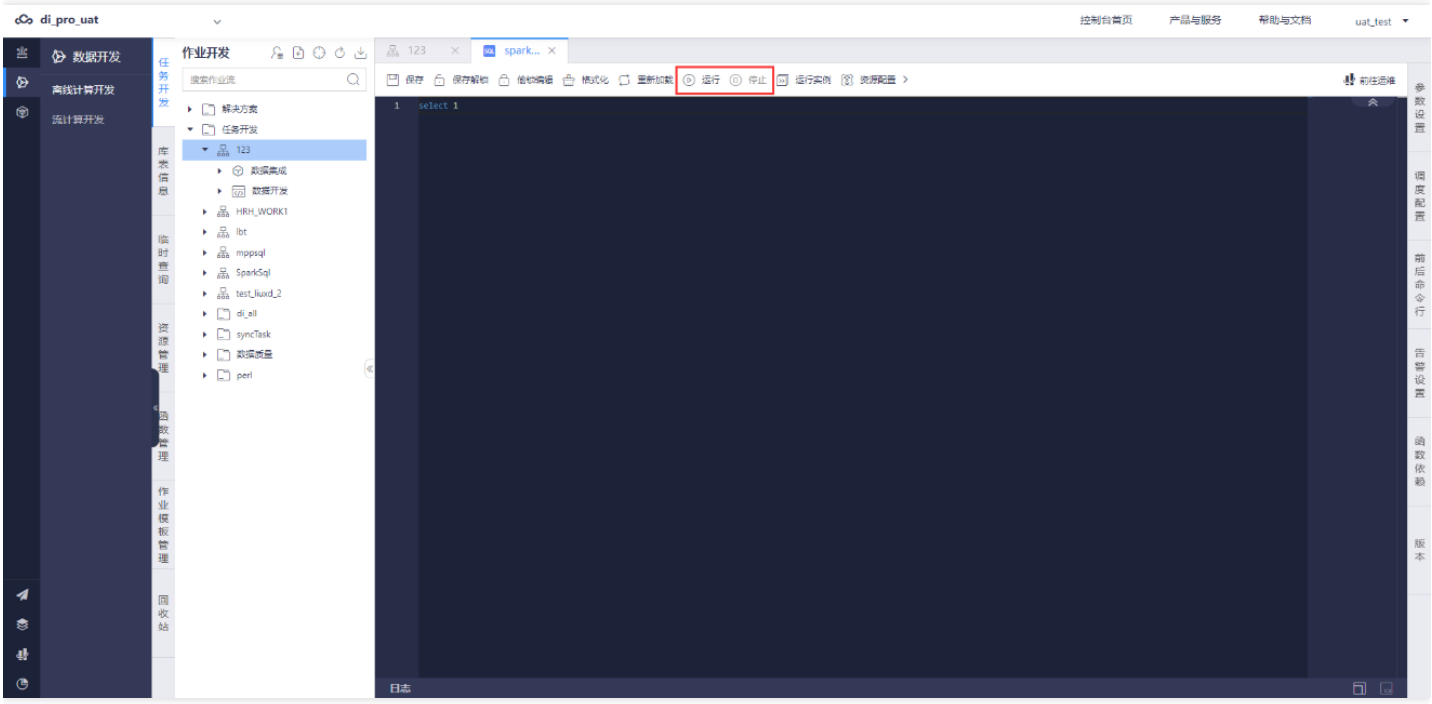




单节点作业测试

最近更新时间: 2019-11-13 07:53:37

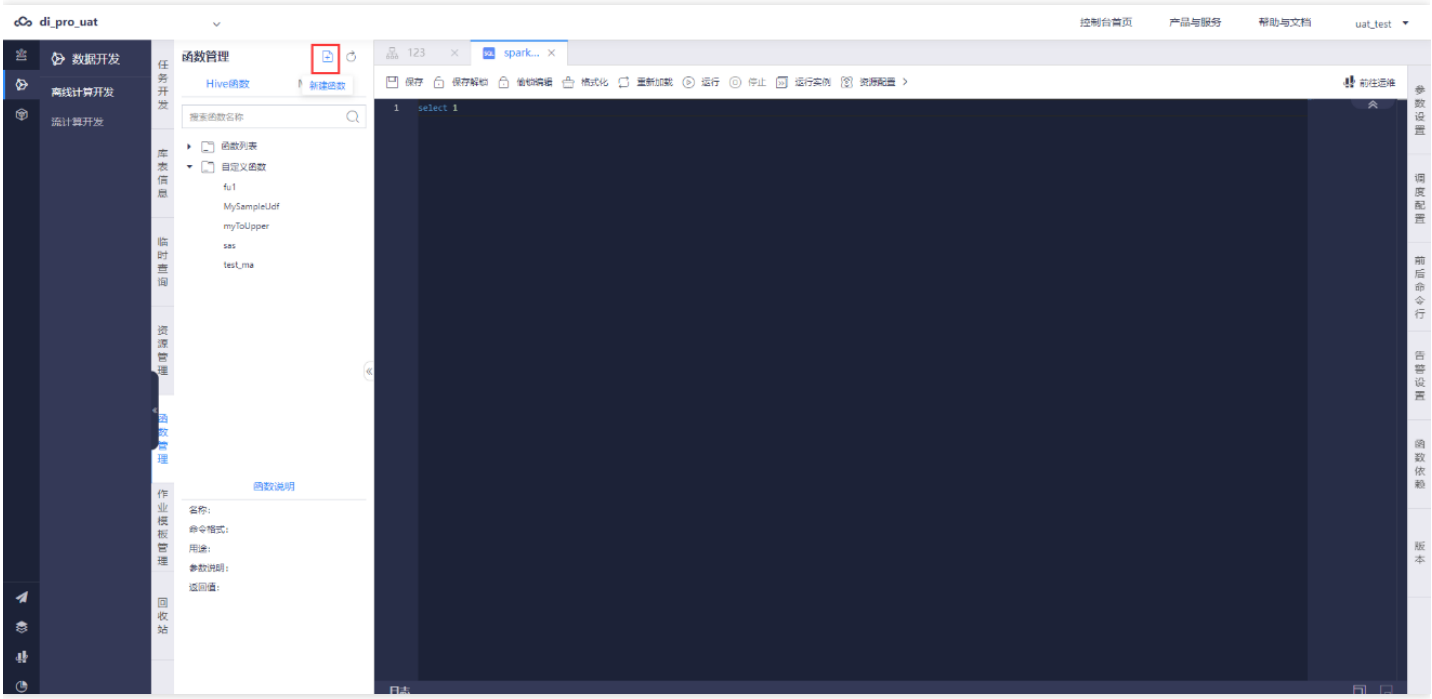
前置条件：无 操作步骤：在开发作业流过程中，选中一个要执行的作业节点，点击在线测试按钮，作业会提交到智能调度测试。

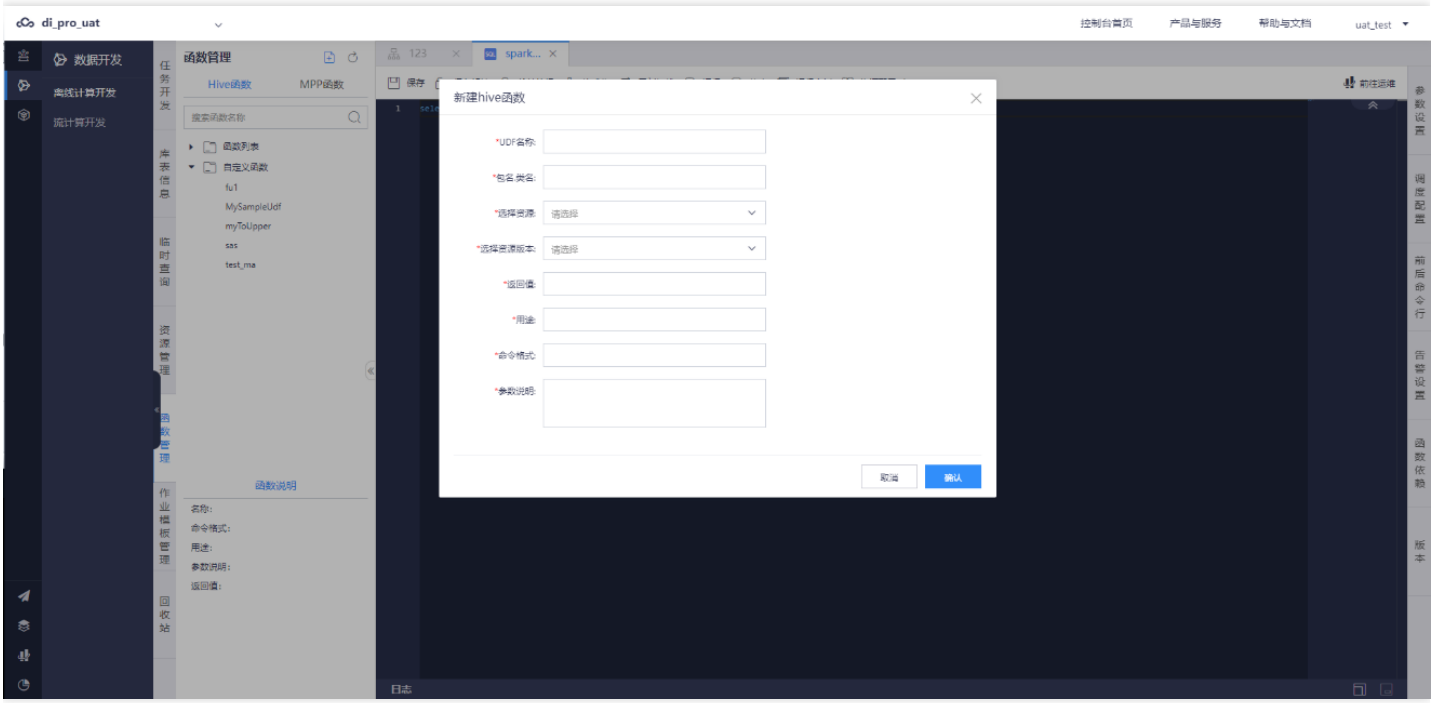


新增函数

最近更新时间: 2019-11-13 07:53:37

前置条件：用户需要使用自定义函数 操作步骤：用户新建函数，定义函数名，类型，用户，参数说明，选择依赖的资源

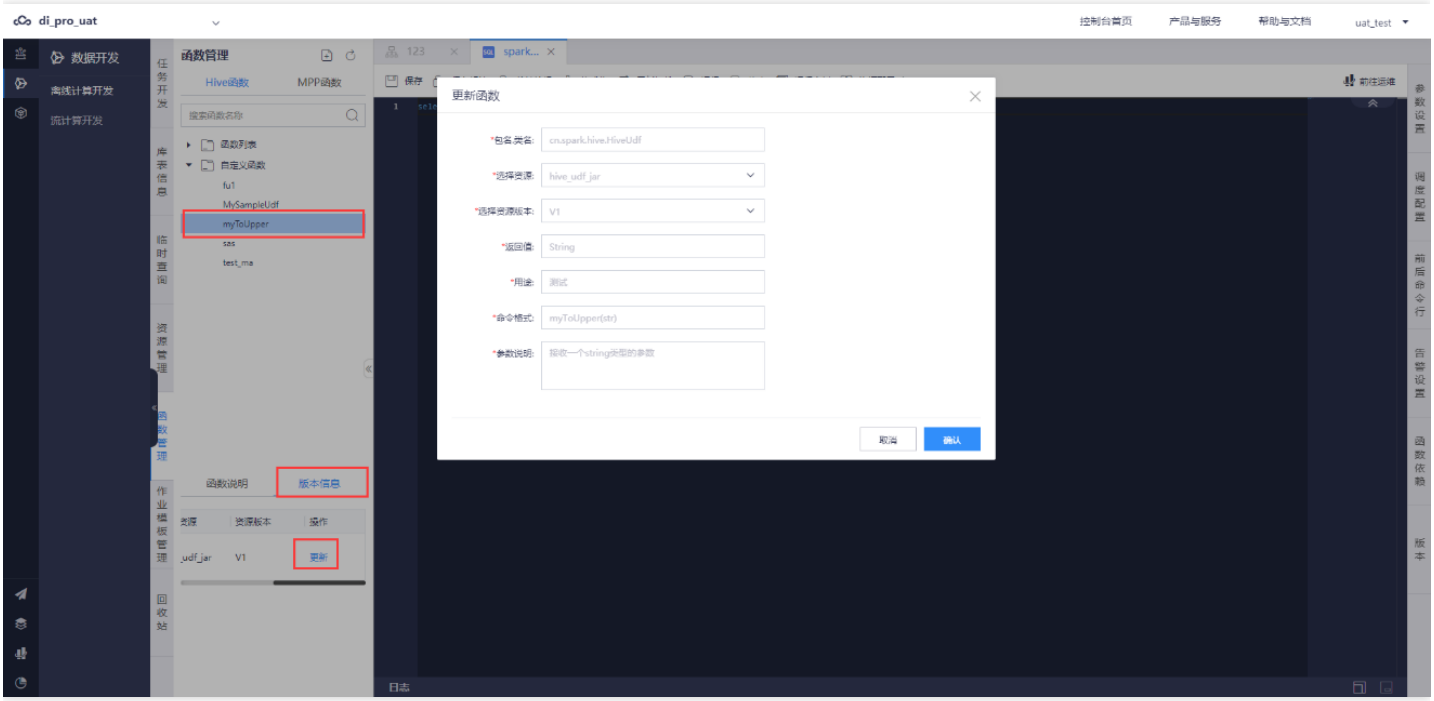




修改函数

最近更新时间: 2019-11-13 07:53:37

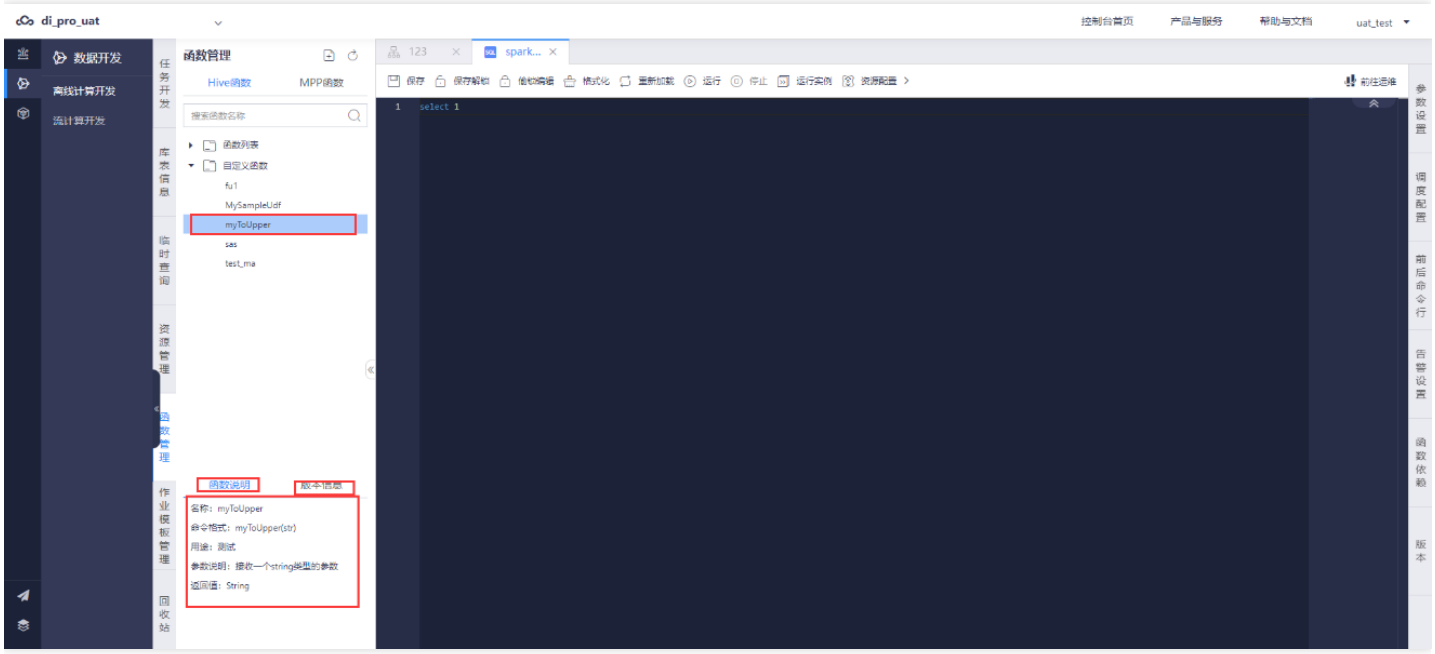
前置条件：已经存在的自定义函数 操作步骤：用户点击编辑函数，重新编辑函数的依赖资源，用途，参数格式



查看函数

最近更新时间: 2019-11-13 07:53:37

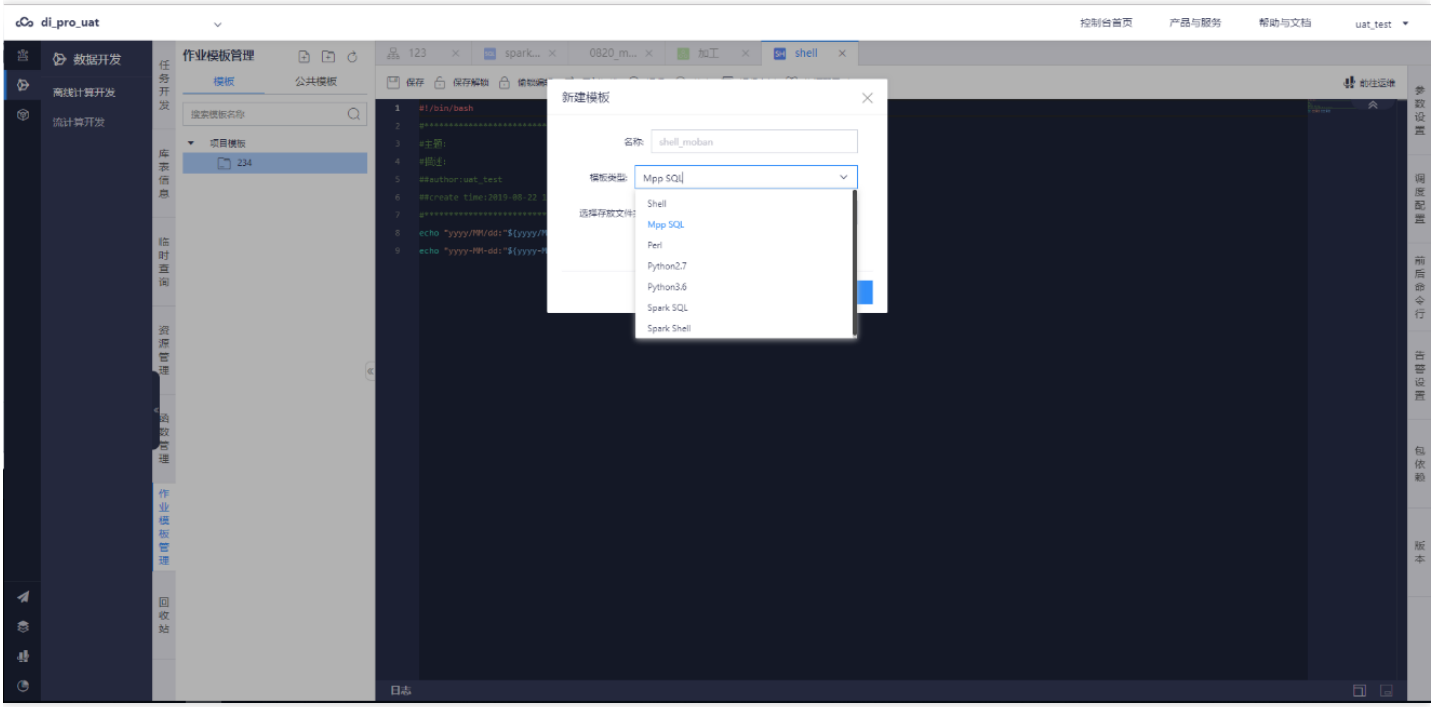
前置条件：已经存在项目可用的项目空间。 操作步骤：用户点击查看函数，返回函数的详细信息



新增作业模板

最近更新时间: 2019-11-13 07:53:37

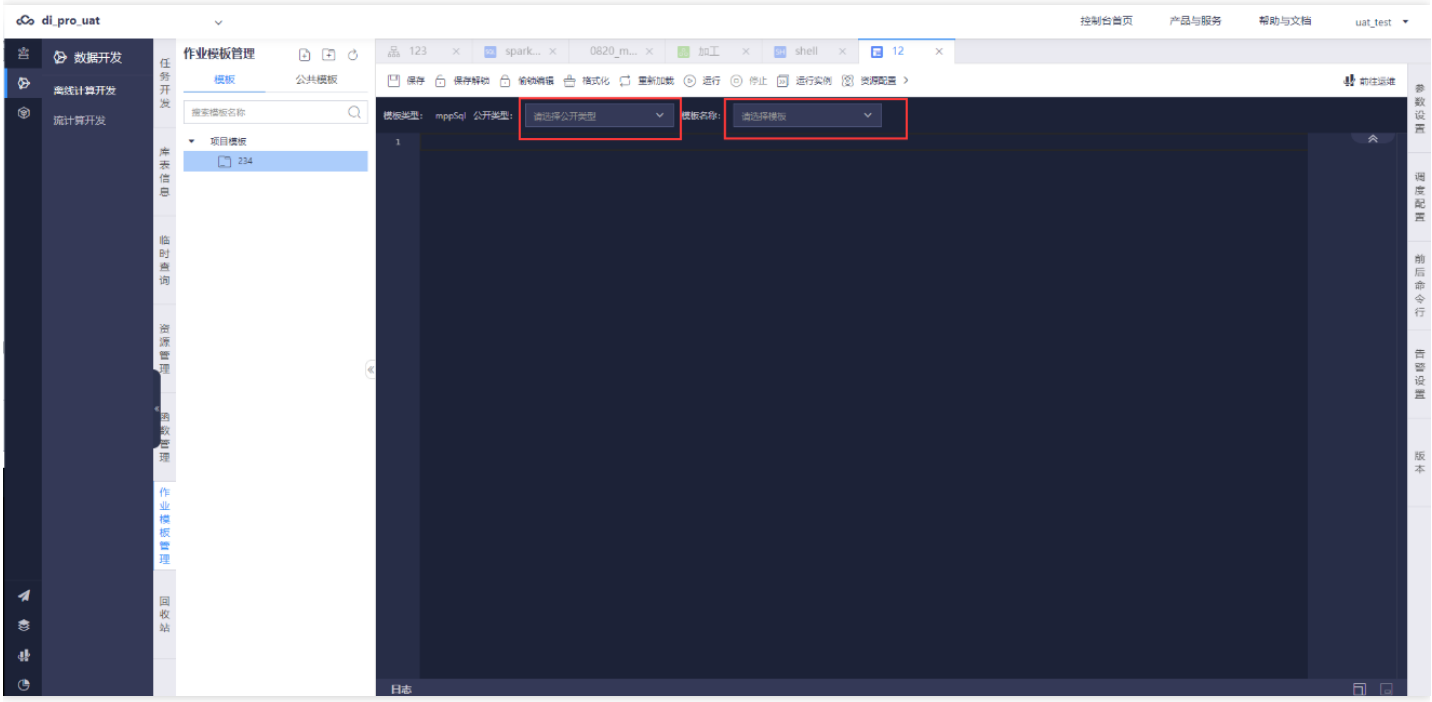
前置条件：平台需要新增作业模板。 操作步骤：用户点击作业模板管理模块，点击创建作业模板，选择作业模板类型



应用作业模板

最近更新时间: 2019-11-13 07:53:37

前置条件：已经存在一个作业模板。 操作步骤：在开发界面中选择一个作业模板，双击点开作业模板，在作业模板内选择需要应用的模板名称。



新建解决方案

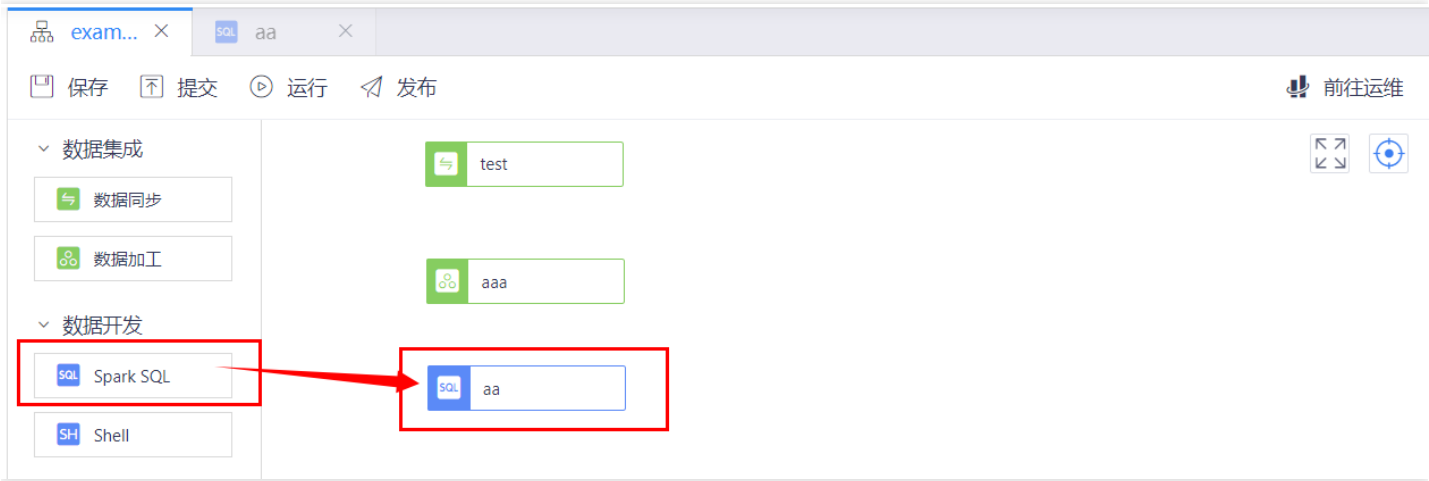
最近更新时间: 2019-11-13 07:53:37

点击【新建解决方案】，输入名称，选择创建的工作流完成解决方案的创建，新建的解决方案会存放在名为【解决方案】的文件夹下。

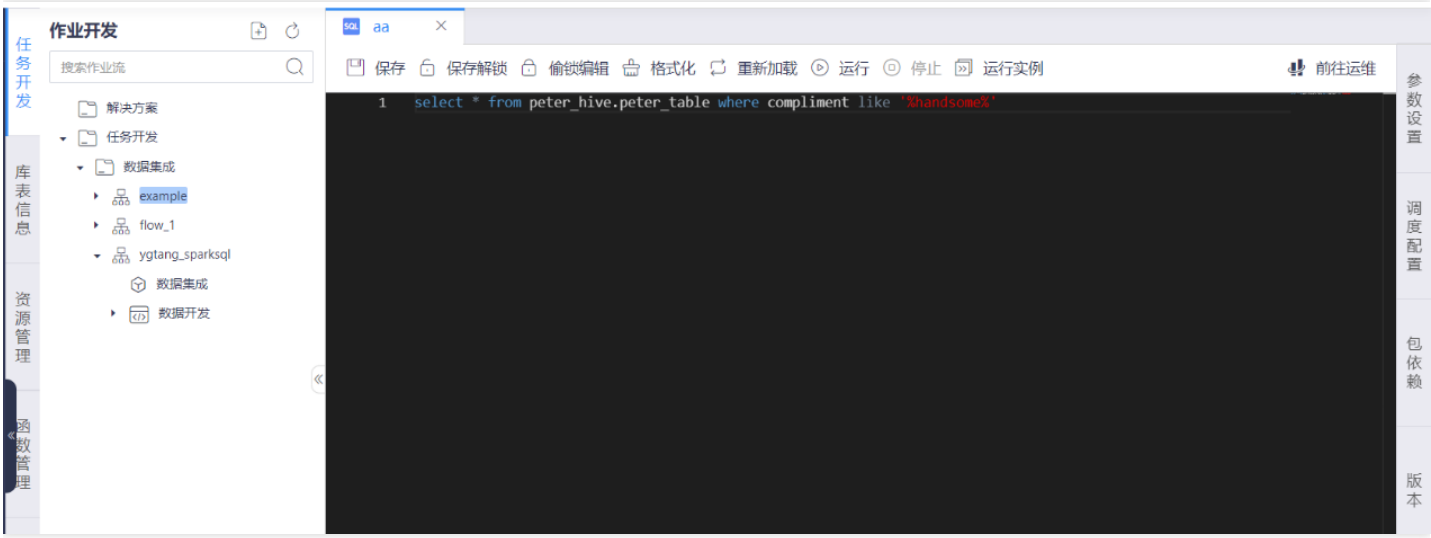
离线计算新增作业
数据开发（以Spark SQL为例）

最近更新时间: 2019-11-13 07:53:37

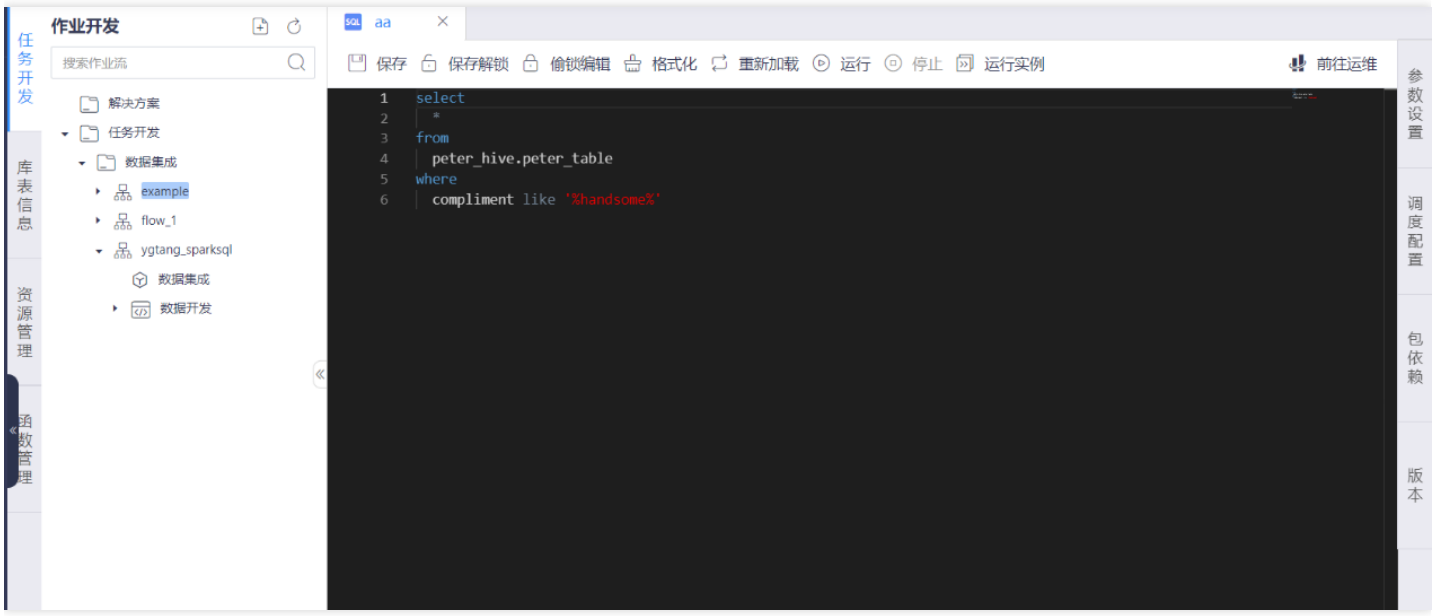
将Spark SQL拖拽至右侧，输入节点名称添加节点



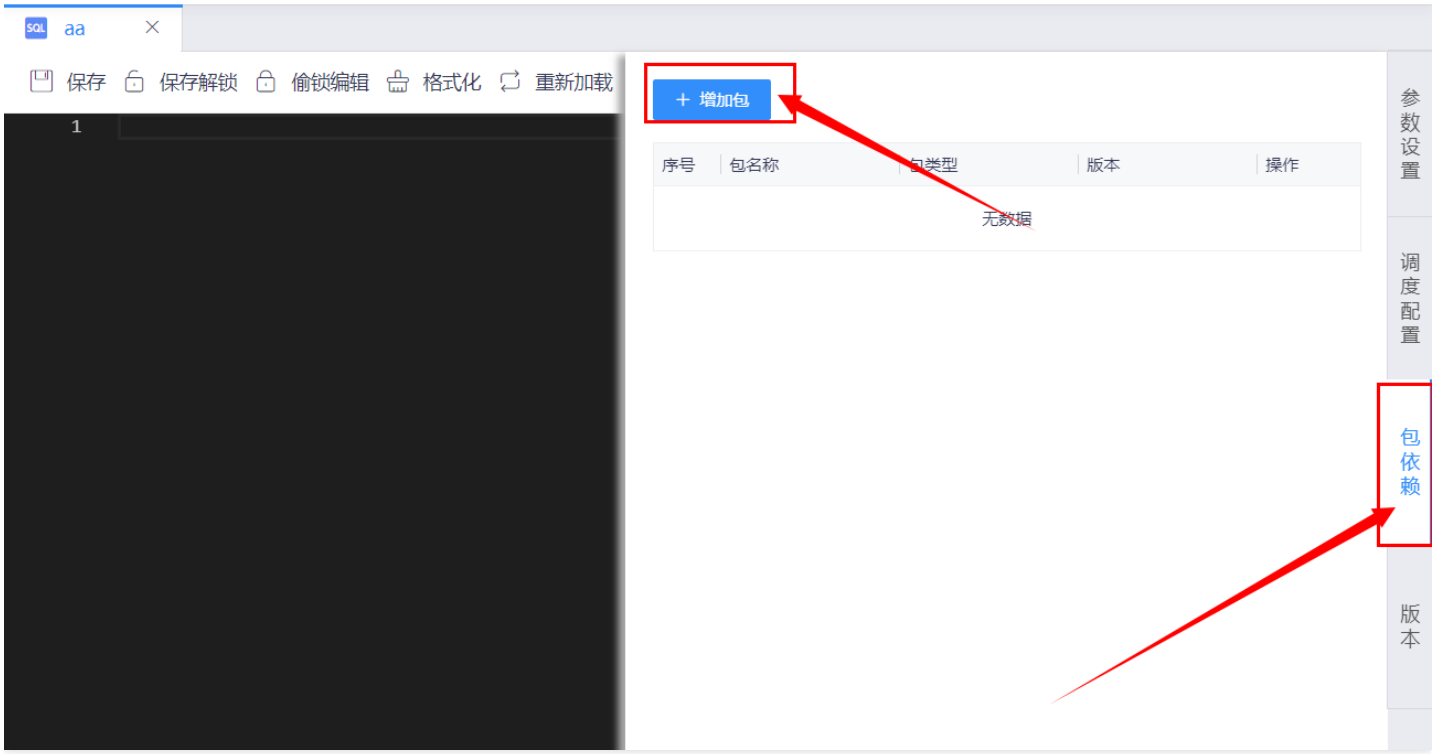
双击打开，在深色区域输入指令

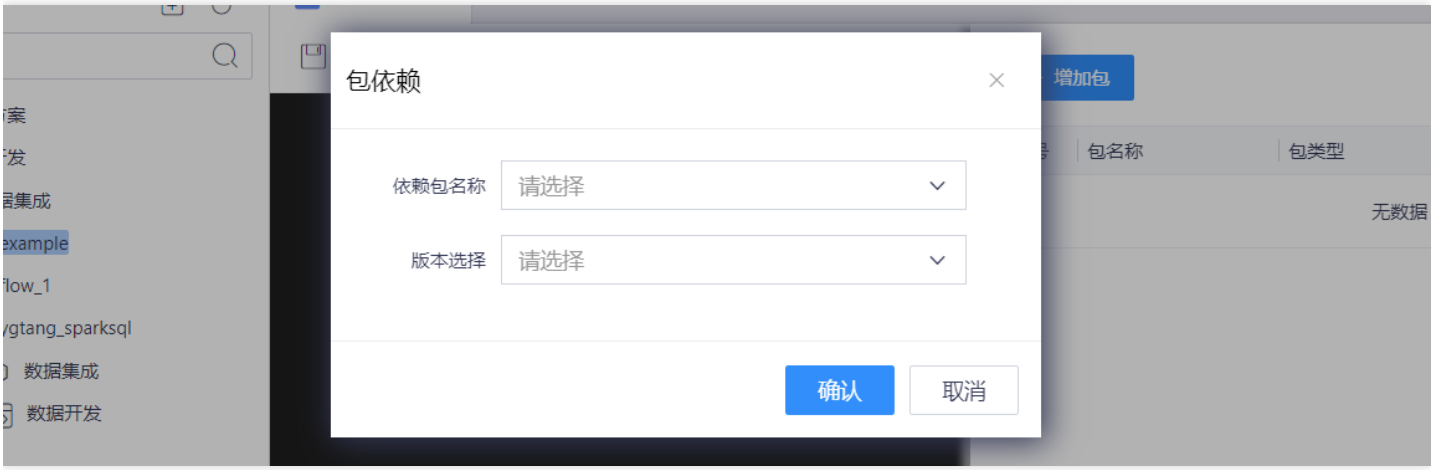


点击【格式化】进行自动的换行和缩进



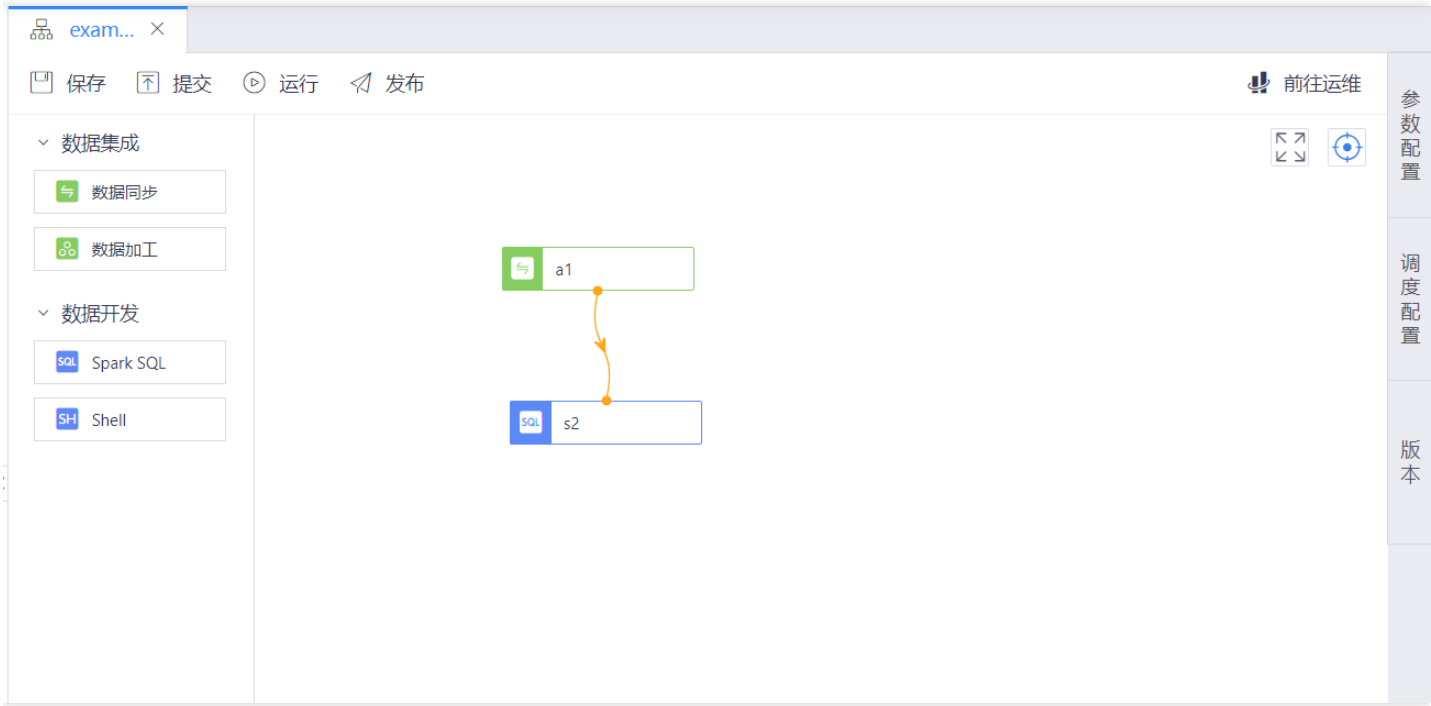
点击右侧包依赖添加依赖包





点击【运行】进行测试，完成后点击【保存】。偷锁相关功能与数据加工处相同。

回到作业流页面，对各节点进行连线建立依赖关系，完成后点击【保存】。



故障处理
作业在测试环境能够正常运行发布到生产环境后运行失败
场景描述

最近更新时间: 2019-10-28 06:35:35

大数据环境分成测试和生产两套环境，两套环境相互之间隔离，用户在数据开发过程中面向的测试环境，当在测试环境运行成功后，将作业发布到生产环境任务才能在生产环境正式运行。有时，可能会出现作业在开发环境正常运行，发布到线上运行失败的问题。

检查处理方式

最近更新时间: 2019-11-26 15:30:16

出现这种问题需要从两个方面进行检查：

- 运行资源是否包含测试资源和生产资源 在开通产品的资源的时候，分成测试资源和生产资源，在测试环境能够正常运行，发布到生产运行失败，需要确认下开通的资源组是不是仅配置了测试环境资源组，没有生产环境资源。

- 数据源是否包含生产测试两个数据源 用户在创建数据管理数据源的时候可以创建到测试和生产两个数据源，在测试环境能够正常运行，发布到生产后运行失败，需要确认下是不是只添加了测试环境的链接信息，没有相应的生产环境链接信息。

创建的作业模板别人无法引用

场景描述

最近更新时间: 2019-10-28 06:39:08

在作业模板中创建模板后，为什么别人无法看到我创建的模板，进行引用

检查处理方式

最近更新时间: 2019-10-28 06:39:51

作业模板管理分成模板和公共模板两部分。其中用户新建模板只会出现在模板tab中，可见范围仅为当前用户。只用将模板公开后，模板变为公共模板，同项目下的其他人才可见，并进行公共模板引用创建作业。

对模板进行修改后，我引用模本创建的作业没有相应的更新

场景描述

最近更新时间: 2019-10-28 06:40:16

使用模板创建一个作业后，对模板进行更新，为什么使用模板创建的作业没有进行同步更新

检查处理方式

最近更新时间: 2019-10-28 06:40:39

作业模板是一个脚本逻辑的固化，可以快速创建生成一个作业。使用模板创建作业的过程相当于将作业模板进行复制过程，生成一个全新作业，生成新作业后和原有模板完全没有关系。对模板的更新不会同步更新到曾经创建的作业上。

有些作业不能够进行编辑操作

场景描述

最近更新时间: 2019-10-28 06:41:03

为什么有的作业可以正常编辑，有的作业不能够进行编辑？保存和保存解锁功能有什么区别

检查处理方式

最近更新时间: 2019-11-26 15:30:16

数据开发是一个多人在线协同开发的IDEA工具，在多人开发的过程中存在一个编辑冲突的问题。数据开发在作业流层面上允许多个用户任意创建作业，最终作业流是一个并集。在作业层面上引入锁的机制，同一时间只允许一个用户进行作业编辑，相当于对作业进行了锁定，其他用户不能对此作业进行编辑操作，如果需要强制修改，则可使用偷锁编辑操作。

保存：仅仅对代码进行了保存，锁定状态仍旧为当前用户，其他用户仍旧不可编辑。

保存解锁：不仅对代码进行了保存，还将锁释放，其他用户此时可以进行编辑。

最佳实践
作业模板管理

最近更新时间: 2019-10-28 06:42:26

当一段脚本比较具有通用型时，可以将这段脚本创建成为模板，用户在创建作业的时候能够引用模板，快速的进行任务创建。

使用解决方案

最近更新时间: 2019-10-28 06:43:03

当一些作业流具有相关性，都是处理同一主题时，可以将这组相关的作业流放在同一个解决方案中，便于用户进行管理。

运行时资源使用量设置

最近更新时间: 2019-10-28 06:43:59

运行不同任务时需要的资源使用量有区别，资源使用量越大运行越快，针对大数据类任务建议资源使用量配置为3（CU）.

常见问题

Q：什么是数据开发？数据开发包含哪些功能？

最近更新时间: 2019-10-28 06:45:00

A：数据开发是大数据云提供的一套离线数据脚本处理服务，全称为离线数据开发。提供了Shell，Spark SQL，MPP SQL,Spark SQL,Python 2.7,Python 3.6,Perl以及作业模板等插件功能。帮助用户在线进行脚本开发、测试、提交、发布上线等一整套流程。

Q：目前数据开发脚本插件支持哪些数据源?

最近更新时间: 2019-10-28 06:45:49

A：特定类型的脚本插件支持特定类型的数据源。目前不同脚本的数据源支持情况如下： Spark SQL：default HIVE数据源 Spark Shell:MPP 数据源 Shell：不支持选择数据源 Python2.7\Python3.6：MPP数据源 Perl：MPP 数据源 MPP SQL：MPP数据源

Q：数据开发支持的插件哪些是大数据类，哪些不是大数据类?

最近更新时间: 2019-10-28 06:46:50

A：在6类插件中，大数据类插件包括：Spark SQL，Spark Shell。非大数据类插件包括：Shell，Python2.7\Python 3.6，Perl，MPP SQL。运行大数据类作业主要使用CU资源，运行非大数据作业主要使用DCU资源(当数据源链接选择MPP 数据源时，需要选择MCU资源)

Q：数据开发对于作业的版本是如何管理?

最近更新时间: 2019-10-28 06:47:45

A：数据开发过程中提交/发布操作等都是以作业流的粒度进行操作，用户提交作业流时选择作业流内需要提交的作业，被提交的作业就会生成一个新的版本。可以选择两个版本进行版本间差异对比。也可以进行历史版本的回退操作。