

大数据处理套件

快速入门

产品文档



腾讯云

【版权声明】

©2013-2018 腾讯云版权所有

本文档著作权归腾讯云单独所有，未经腾讯云事先书面许可，任何主体不得以任何形式复制、修改、抄袭、传播全部或部分本文档内容。

【商标声明】

及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。

【服务声明】

本文档意在向客户介绍腾讯云全部或部分产品、服务的当时的整体概况，部分产品、服务的内容可能有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或模式的承诺或保证。

文档目录

快速入门

快速入门 workflow 概述

创建工作流

创建任务

编辑任务

workflow 任务继承关系建立

任务运行与运行过程查看

删除任务

删除 workflow

快速入门

快速入门 workflow 概述

最近更新时间：2017-10-27 15:03:30

简介

workflow 是腾讯自研的任务调度系统，是 TBDS（大数据处理套件）最重要的功能之一，具有毫秒级任务下发，高可靠的特性，同时支持插件式扩展任务类型。

快速入门流程图

快速入门主要通过 workflow 的基础操作，使您快速了解大数据处理套件这个产品，流程图如下：



申请 TBDS 体验账号试用

申请体验账号请联系 [腾讯云大数据TBDS](#) 进行产品试用申请。

注意：

大数据处理套件为私有化部署产品，在腾讯云官网上目前没有直接入口。登录控制台需使用 Chrome 或 Firefox 浏览器。

创建 workflow

最近更新时间：2018-12-19 11:07:53

workflow 是腾讯自研的任务调度系统。作为 TBDS 最重要的功能之一，workflow 对于数据处理的基本流程包含数据接入，计算和数据导出三部分，workflow 原理如下图：

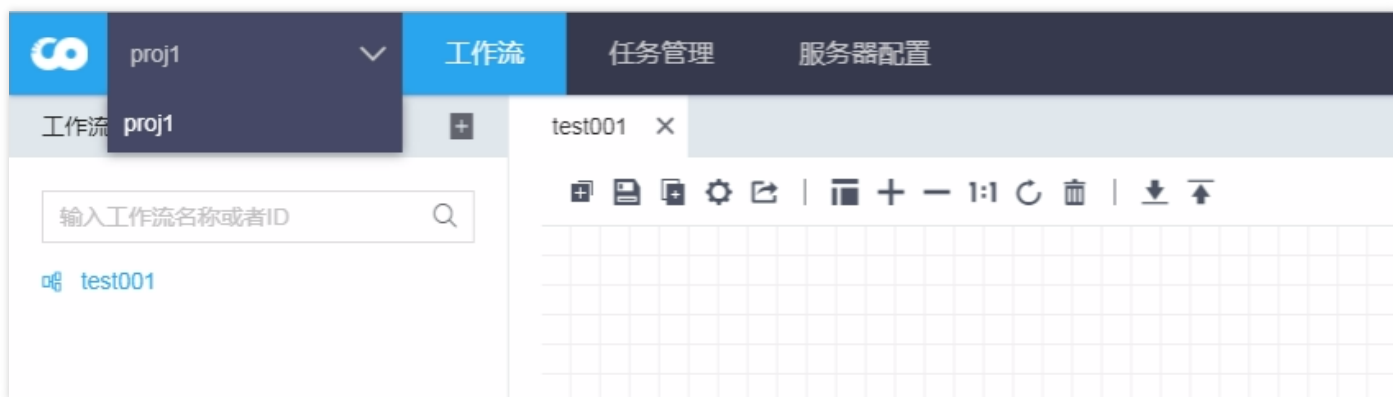


操作步骤

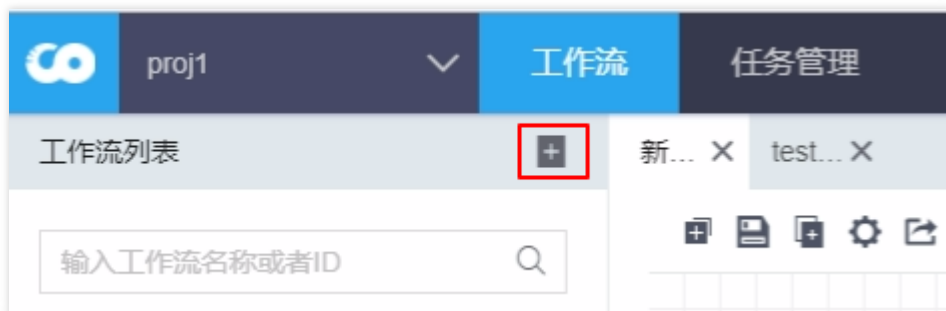
1. 登录 [腾讯大数据处理套件](#)，进入业务主界面后，单击【工作流】，即可进行工作流和任务的创建。



2. 在如下下拉框会显示当前登录账号加入的项目，选择一个项目，在项目中创建工作流。

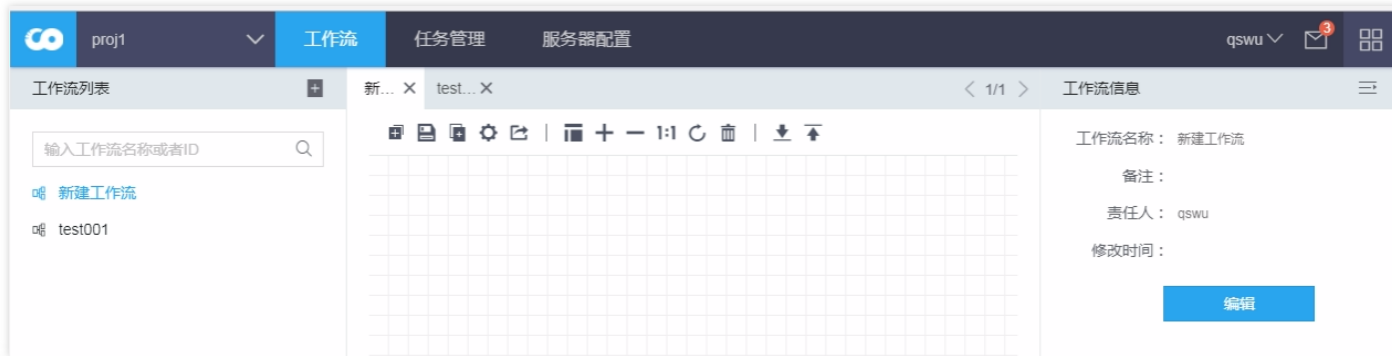


3. 在工作流的页面中，有创建工作流的入口，单击【+】新建工作流。



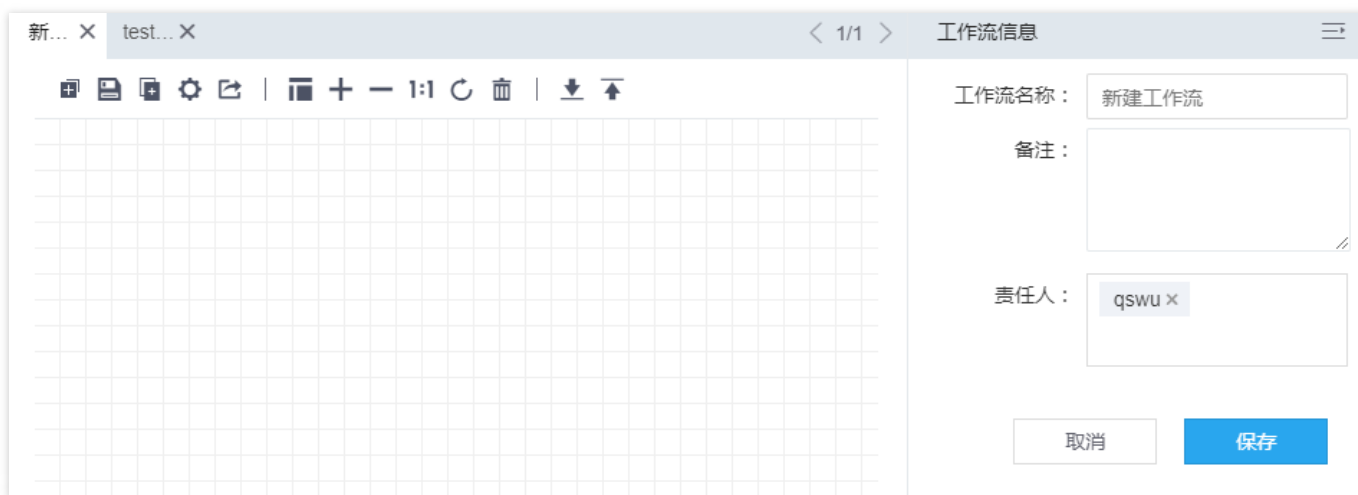
依次填写信息后，单击【确定】。

4. 创建完成后返回工作界面。



单击右侧【编辑】，可配置工作流，编辑完成后单击【保存】即可。

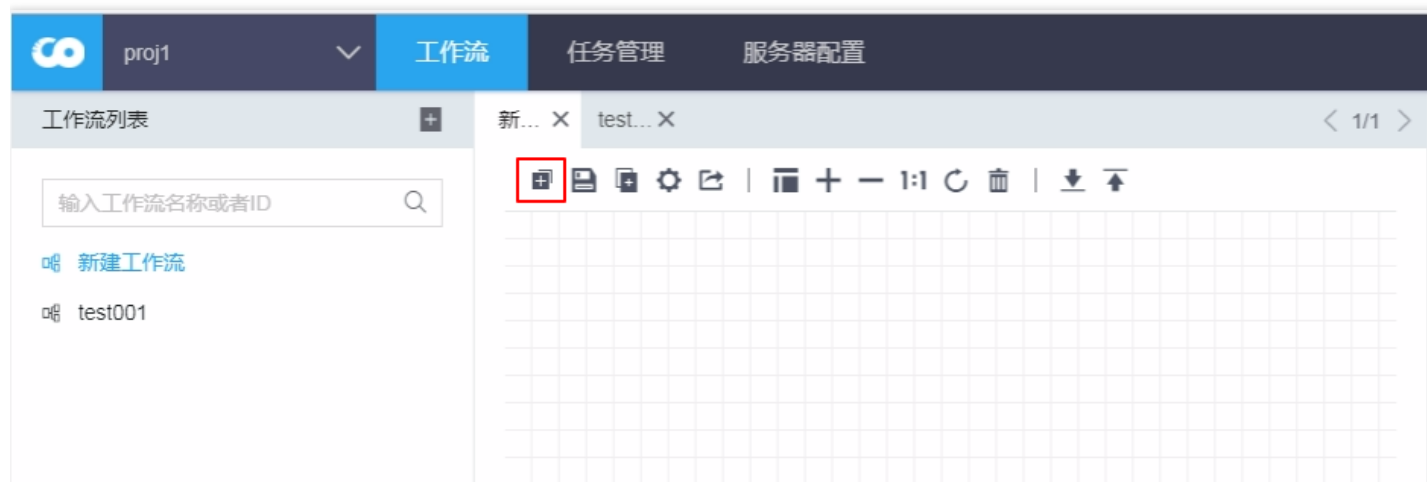
- **工作流名称**（可修改）：工作流的名称，不唯一。
- **备注**（可修改）：工作流的描述信息。
- **责任人**（必填，多选）：责任人只能是“所属项目”中的项目成员，可以填多个。
- **修改时间**（只读）：展示最近修改时间。



创建任务

最近更新时间：2017-12-07 17:02:31

创建工作流之后，得到一个空白的工作流画布，可以在该画布上部署不同类型的任务。拖动工作流列表右侧的【+】图标，摆放到任意位置，就会弹出所有的工作流任务。



在搜索框里搜索常用的任务，选择一个任务后，单击【确定】，任务即创建成功，如下图：

任务类型选择



搜索标签或任务



自定义任务类型

任务类型选择：选择计算需要的任务

选择	任务	描述	标签
<input checked="" type="radio"/>	FTP导入HDFS	FTP导入HDFS	FTP,HDFS
<input type="radio"/>	HDFS导出HBASE	HDFS写入HBASE	HDFS,HBASE
<input type="radio"/>	HDFS导出HIVE	HDFS导出HIVE	HDFS,HIVE
<input type="radio"/>	HDFS导出MYSQL	HDFS导出MYSQL	HDFS,MYSQL
<input type="radio"/>	HDFS导出Oracle	HDFS导出ORACLE	HDFS,ORACLE
<input type="radio"/>	HDFS导出PG	HDFS导出POSTGRES	HDFS,POSTGRES

取消

确定

任务创建成功后，工作界面会弹出新建任务的小窗口，如下图：



编辑任务

最近更新时间：2018-06-12 10:11:47

在当前任务窗口右键单击【编辑】。



填写配置信息，任务编辑完成后将在下次运行时生效。

配置说明：

- (1) 源服务器和目标服务器：用户做数据分析处理时需将外部数据从源服务器导入到平台，最终将处理结果导出到目标服务器。配置详情请参见 [服务器配置](#)。
- (2) 源目录：FTP 上面源文件所在的目录，例：`/stage/outface/sng/test_table/${YYYYMMDD}/`。
- (3) 文件名正则表达式：支持时间格式`${YYYYMMDD}`，`*` 为所有文件，基于 Java 正则规则。
- (4) 是否遍历子目录:选择是，则递归遍历子目录。
- (5) 最大遍历层数:默认递归遍历 5 层。
- (6) 目标目录:HDFS 上存放文件的目录，例：`/stage/outface/sng/test_table/${YYYYMMDD}/`。
- (7) 重试次数:重试次数。（重试次数为 0，任务不下发）
- (8) 步长:间隔多少周期执行一次。（若是天任务，步数为 2，每两天执行一次）
- (9) 代理 IP：任务执行所在机器 IP。

新建 workflow > 新建任务 | 保存 返回



新建任务

FTP导入HDFS

基本信息 调度设置 参数配置

* 任务名称：

* 任务类型：FTP导入HDFS

* 负责人：

任务将用第一个责任人相关权限来执行任务，画布责任人也具有任务修改相关权限。

任务说明：

告警：

基本信息 调度设置 参数配置

* 周期类型：

* 起始数据时间：

周期类型和起始数据时间一旦选定保存后将不允许再修改。





任务获取数据的时间起点，任务将于下一个周期（执行时间）调度运行。例如：数据时间为3点的小时任务，将会在下一个周期4点运行。

* 自身依赖： 是 否 并行

调度时间：

任务到了执行时间后，会按照这里的配置延迟一段时间才会开始执行。

基本信息 调度设置 **参数配置**

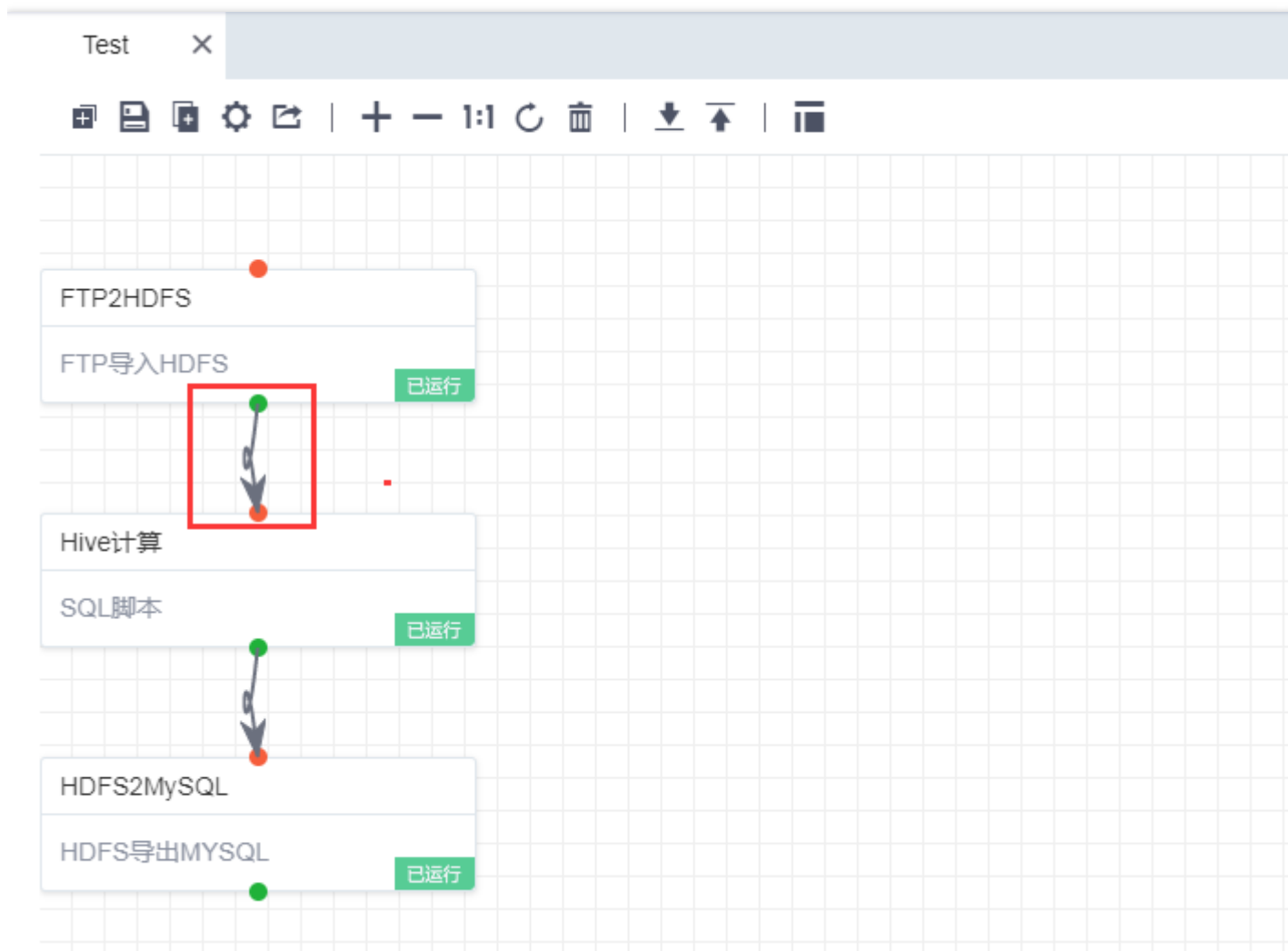
参数	值
* 终止时间：	2019-09-22 
是否可重试：	<input checked="" type="radio"/> 是
* 源服务器：	petertestftp  编辑服务配置
* 目标服务器：	petertesthdfs  编辑服务配置
* 源目录：	<input type="text"/>
* 文件名正则表达式：	<input type="text"/>
* 是否遍历子目录：	否 

FTP上面源文件所在的目录，例：/stage/outface/sng/test_table/\${YYYYMMDD}y

workflow 任务继承关系建立

最近更新时间：2017-12-07 17:02:48

各个 workflow 任务之间可通过拖动任务块下方箭头，建立任务之间的继承关系。建立任务间的继承关系，就是建立任务间的依赖关系。例如任务 A 和任务 B 有依赖关系，那么同周期 A 任务实例成功运行后，同周期 B 任务实例才会开始运行，否则为等待下发状态。



如果要删除某个继承关系，只需将鼠标移动到该继承关系相应的箭头，右键单击【删除】即可。

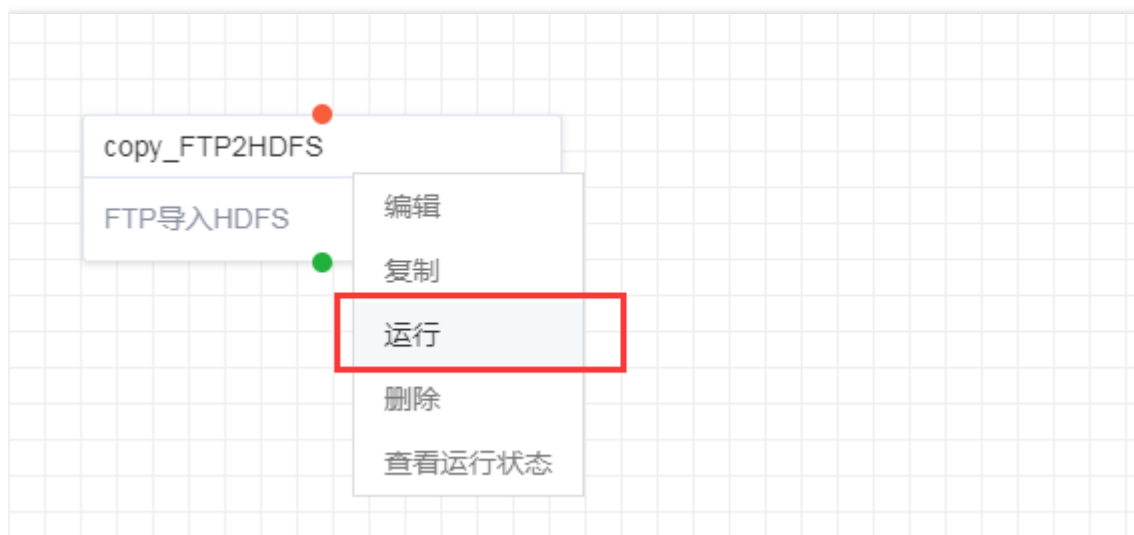


任务运行与运行过程查看

最近更新时间：2017-12-07 17:02:56

运行任务

workflows 任务编辑完成后，就可以开启任务的运行。在任务上右击选择【运行】即可触发任务开始运行。



单击【运行】后，会提示用户任务需要审批才可运行；审批权限为项目管理员所有，可选择审批通过后自动运行，单击【确定】。



查看运行状态

右击 workflow 任务选择【查看运行状态】，可以打开查看该任务各个周期任务运行的详细状态。

任务运行状态

任务名称： FTP2HDFS 状态：所有 查询

时间： 2017-07-03 至 2017-07-10 更多实例 >

终止
重跑

	数据时间	运行耗时	状态	操作	日志
<input type="checkbox"/>	2017-07-03 00:00:00	00:00:17.00 0	✔ 成功	编辑任务	查看
<input type="checkbox"/>	2017-07-04 00:00:00	00:00:17.00 0	✔ 成功	编辑任务	查看
<input type="checkbox"/>	2017-07-05 00:00:00	00:00:17.00 0	✔ 成功	编辑任务	查看
<input type="checkbox"/>	2017-07-06 00:00:00	00:00:17.00 0	✔ 成功	编辑任务	查看
<input type="checkbox"/>	2017-07-07 00:00:00	00:00:17.00 0	✔ 成功	编辑任务	查看
<input type="checkbox"/>	2017-07-08 00:00:00	00:00:17.00 0	✔ 成功	编辑任务	查看

查看运行明细

单击日志下的【查看】可以查看详细运行内容，也可以单击【更多实例】>【任务名称】下的内容，到详细的父子任务查看页面。

[基本信息](#) [父子任务](#) [日志](#)

任务名称： FTP2HDFS

工作流： [Test](#)

责任人： 

周期类型： 天

起始数据时间： 2017-7-1 00:00:00

调度时间： 0

任务类型： FTP导入HDFS

所属项目： 大数据平台

最近修改时间： 2017-7-10 00:33:34

运行信息

当前状态：  成功

数据时间： 2017-07-03 00:00:00

删除任务

最近更新时间：2017-12-07 17:03:04

对新建未运行的任务、已经停止的任务可以进行删除，右键单击【删除】即可，如下图：



注意：如果任务正处于运行状态，则不能删除。

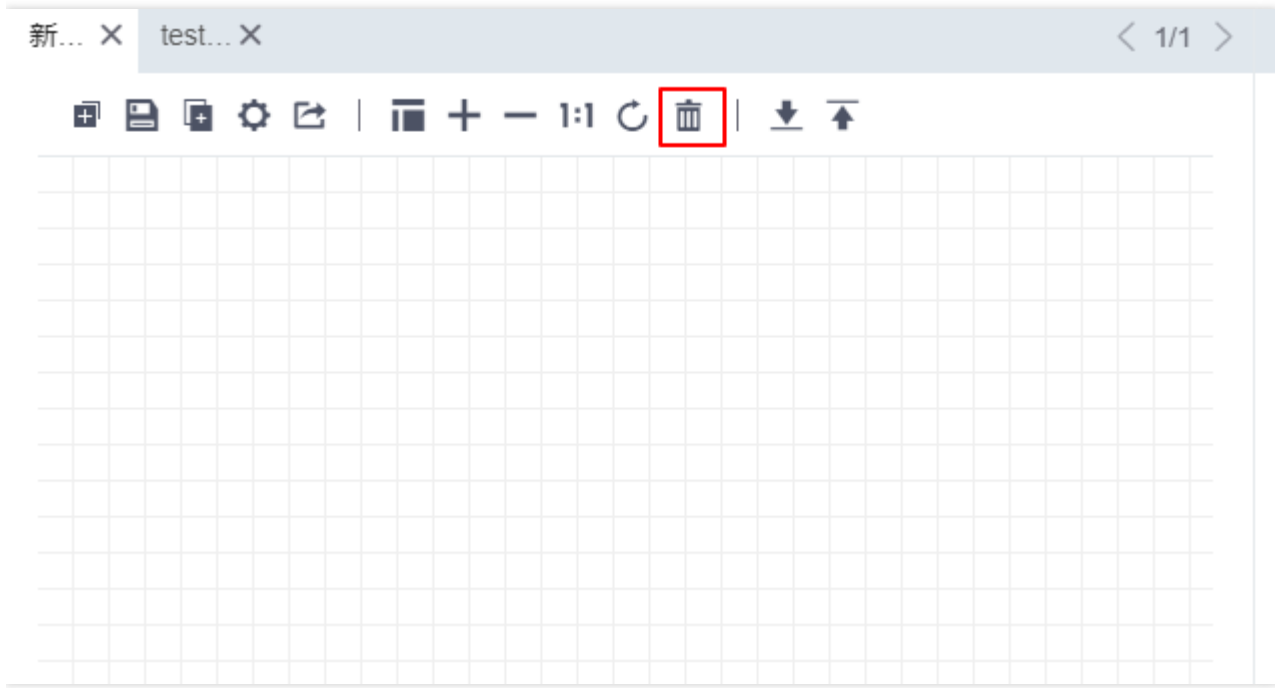
单击【删除】后弹出窗口，单击【确定】完成删除，如下图：



删除 workflow

最近更新时间：2017-12-07 17:03:12

首先在画布的工具栏中，单击【删除】图标。



单击【删除】图标后，弹出确认删除窗口，单击【确定】即可删除整个 workflow。

