数据集成

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 02:53:13

数据集成是一套稳定高效、弹性伸缩的数据接入、转换、加工、检核的可视化的数据套件，整个套件包括数据同步、数据加工、数据整合和业务检核四大功能。极大地降低了用户数据上云以及数据开发的门槛。 数据集成主要包括四大功能组件： 数据同步工具不仅能够满足传统数据集成服务在复杂网络环境下进行多种异构数据源的导入导出需求，同时在数据导入导出的过程中同步进行数据清洗、去重、规范化等，提高数据质量，防止脏数据、垃圾数据的传播。 数据加工工具采用可视化拖拽的方式进行数据ETL开发，降低开发门槛，使没有SQL经验的业务人员也能够进行快速的数据逻辑开发。 数据整合工具结合行业经验，沉淀丰富的贴源数据处理算法，用户只需要创建特定的表结构后通过向导式的勾选就可实现数据贴源层加工。 业务检核工具与数据质量模块相结合，对数据进行数据质量，数据波动的统计查询，让用户了解数据质量情况。

产品价值

最近更新时间: 2019-10-28 02:54:16

提供了复杂网络环境下、异构数据源的数据接入和批量同步服务。在向导式，拖拽式的开发过程中通过数据清理、数据转换等，加强数据质量管理，最终实现分布在各个不同源中的数据高质量的汇总。

产品定位

最近更新时间: 2019-10-28 02:55:05

数据集成是大数据云服务核心组件之一，定位于为大数据云项目中离线数据的处理，包括用户线下数据的上云迁移，可视化的ETL加工，以及数据同步中的检核等，是离线数据处理功能组件的一个重要部分。

产品优势

最近更新时间: 2019-11-26 15:30:29

- 多种不同类型数据源传输，有效整合分散的数据资产，解决数据孤岛问题
- 向导式、拖拽式的开发方式实现数据计算逻辑设计，零代码开发，降低使用门槛，提升开发效率
- 对无效数据，异常数据等脏数据进行清洗、规范化等，有效提成数据质量
- 丰富的数据脱敏，加密等转换方式，提升数据安全合规
- 灵活的技术检核与业务检核配置，数据传输过程中进行数据质量全程监控并生成质量报告

应用场景

最近更新时间: 2019-11-26 15:30:29

本地数据迁移上云：

使用数据集成中的数据同步服务，用户可以快速、低成本的创建面向对象存储、标准数据接口服务（JDBC适配的数据库）、NoSQL等多种数据源的数据同步任务，通过调度的周期性任务设置，企业可轻松实现不同数据源的周期性数据接入，大大降低企业本地数据上云门槛。

贴源数据的逻辑加工

使用数据集成中的数据整合功能，用户可以将业务系统每日产生的数据快速的进行逻辑整合，生成拉链表，切片表等。减少了复杂逻辑脚本的开发，降低了数据整合处理门槛。

功能特性

功能特性

最近更新时间: 2019-10-28 03:02:54

数据集成部分包括的功能有数据检核、数据转换、数据同步和数据整合。主要是完成Source到Sink之间同源、异源、文件到库表以及库表到文件等一系列数据操作。功能细节如下：

1. 数据检核部分通常在数据转换、数据同步、数据整合或者数据加工之前进行，分为业务检核和技术检核。
2. 数据同步部分是为了实现从源端到目标端数据的加载、卸载、复制。
3. 数据整合部分使用增量切片算法、全量切片算法、拉链算法、时点快照算法及数据表拆分等方式实现。
4. 数据加工部分主要使用SQL对库表或者非库表数据进行指标、维度、统计等加工计算。

数据同步
数据源和数据目标支持表设置

最近更新时间: 2019-10-28 03:08:54

- 数据同步支持的数据源类型
  - 文件存储(COS)
  - 数据库(Oracle,MySQL)
  - NoSQL(HBase)
  - 大数据类(HIVE)
  - MPP数据库(MPP)

- 数据同步支持的数据目标类型
  - 文件存储(COS)
  - 数据库(Oracle,MySQL)
  - NoSQL(HBase,Redis)
  - 大数据类(HIVE,Elasticsearch)
  - MPP数据库(MPP)

不同数据类型写入方式不同

最近更新时间: 2019-10-28 03:10:11

不同数据源的具有不同的写入方式列表如下

	insert into	insert overwrite	append	其他设置
COS（文本类型）		每次运行是进行文件覆盖	进行数据的追加写入	1.是否写入表头 选择写入的源是否有表头，需要跳过
HIVE	每次运行进行数据追加	档表有分区时，将分区数据进行替换。当表没有分区时，直接将表清空再写入		
MPP	每次运行进行数据的追加			
Oracle	每次运行进行数据追加	每次运行时将表清空再写入		
MySQL	每次运行进行数据追加	每次运行时将表清空再写入		
HBase				1.rowkey设置 在数据管理设置rowkey，这里只进行显示
Redis				1.KeyIndex 表名+选择多列+列间隔符 2.value type 和 mode string-》set hash-》hset、hmset list-》lpush、rpush、mpush set-》sadd 3.写入方式 标准模式 和 value转key模式 两种模式 4.是否设置有效时间

				5.数据有效时间 是否设置有效时间选【是】显示
ES				1.doc id生成方式 拼接列 -- 选择多列和间隔符 特定列 -- 选择一个列 随机UUID

数据加工

最近更新时间: 2019-11-26 15:30:28

数据加工支持的算子简介

- Source算子 作用： 数据加工的数据来源，可以选择多种数据源进行数据 操作方式： 拖拽Source算子到画板中，显示库表选择框，选择需要进行加工的库表点击确定后，Source变为缩略态。双击Source，显示编辑态，在编辑态中可以在过滤语句中添加过滤条件，将希望后续输出的字段‘输出’进行勾选。

- Target算子 作用： 整个数据数据加工的数据目标。 操作方式： 拖拽Target算子到画板中，显示库表选择框，选择需要进行加工的库表点击确定后，target变为缩略态。将上游算子连接到target算子。双击显示编辑态，在编辑态中进行上游算子字段和目标表字段的映射关系设置，并根据不同的目标源进行写入方式设置。

- Map算子 作用： 基于行级的数据项复制、修改、计算。在同行记录中可新增、减少数据项 操作方式： 拖拽Map算子到画板中，将上游算子连线到Map算子，上游算子勾选输出的数据会同步到Map算子中，双击Map算子进入Map编辑状态。可以在每行表达式中可以进行行级数据处理，如：数据类型转换，例如to_date(Port1,’yyyyMMdd’)，数据项计算，例如(Port1+port2)/Port3，新增变量，例如Port2=Port1+1等。将希望后续输出的字段‘输出’进行勾选。

- Filter算子 作用： 按照条件过滤掉不符合条件的行。 操作方式： 拖拽Filter算子到画板中，将上游算子连接到Filter算子，上游算子勾选输出的数据会同步到Filter算子中，双击Filter算子进入编辑状态。在Filter条件中添加过滤条件。将希望后续输出的字段‘输出’进行勾选。

- Sample算子 作用： 按照一定的规律抽取数据，目前只支持按照百分比进行数据抽取 操作方式： 拖拽Sample算子到画板中，将上游算子连接到Sample算子，上游算子勾选输出的数据会同步到Sample算子中，双击算子进入编辑状态。在Sample条件中添加采样条件，按照百分比进行数据抽样。将希望后续输出的字段‘输出’进行勾选。

- Sorter算子 作用： 对数据按照某些字段进行升序/降序的排序。 操作方式： 拖拽Sorter算子到画板中，将上游算子连接到Sorter算子，上游算子勾选输出的数据会同步到Sorter算子中，双击算子进入编辑态。在排序字段中添加需要进行排序的字段，并选择排序类型是升序还是降序。将希望后续输出的字段‘输出’进行勾选。

- Join算子 作用： 对两个数据源进行连接操作。只支持等值连接。Join只支持连接两个数据源，如果有多个数据源进行连接，使用多个Join。 操作方式： 拖拽Join算子到画板中，Join算子可以接收两个输入源，将一个上游算子拖拽到Join作为Join的master，将第二个上游算子拖拽到Join作为Join的detail。

- Union算子 作用： 合并两个数据源到一个结果集。与执行“UNION ALL”SQL语句结果相似，不会删除重复行。Union只支持合并两个数据源，如果有多个数据源进行合并，使用多个Union。 操作方式： 拖拽Union算子到画

板中，Union算子可以接收两个输入源，将一个上游算子拖拽到Union作为Union的第一个输入组，在选另一个上游算子拖拽到Union中作为Union的第二个输入组。第一个输入组的字段信息会显示在Union输出列表中，调整第一输入组，第二输入组和Union输出列表。需要字段类型一致。在Union输出列表中，将希望后续输出的字段‘输出’进行勾选。

- Aggregator算子 作用： 对多组记录进行聚合计算 操作方式： 拖拽Aggregator算子到画板中，将上游算子连线到Aggregator算子，上游算子勾选输出的数据会同步到Aggretator算子中，双击Aggregator算子进入Aggregator算子编辑状态。对于Aggregator算子需要至少有一个分组字段，增加分组字段后，再添加需要进行聚合计算的字段，下拉勾选出对字段进行sum、avg、max、min等聚合运算。 在分组字段和聚合字段上将希望后续输出的字段‘输出’进行勾选。

目前数据加工支持的算子数量。以及每个算子的输入、输出及数据来源。

算子	输入	输出	数据来源
Source算子	无	多	库多表选择
Target算子	1	无	库表选择和上游算子
Map算子	1	多	上游算子
Filter算子	1	多	上游算子
Sample算子	1	多	上游算子
Sorter算子	1	多	上游算子
Join算子	2	多	上游算子
Union算子	2	多	上游算子
Aggregator算子	1	多	上游算子

数据整合

最近更新时间: 2019-10-28 03:06:07

结合多年数据处理行业经验，沉淀固化通用数据整合模型，将贴源数据的处理过程从繁复的代码逻辑中解放，仅需简单配置既可完成复杂贴源数据整合。同时数据整合在应用时不同算法对于源表和目标表有一定的配置要求，一般来说目标表比源表需要新增特定字段，具体新增字段如下。除了新增字段外其他字段需完全保持一致。

数据整合		
全量切片	业务日期字段	ty_data_date
	批次号字段	ty_batch_number
	来源标识	ty_src_flag
	运行job字段	ty_job_name
增量切片	业务日期字段	ty_data_date
	批次号字段	ty_batch_numbe
	删除标识	ty_del_flag
	删除日期	ty_del_date
	删除批次	ty_del_batch
	运行job字段	ty_job_name
拉链表	开始业务日期	ty_start_date
	开始批次号	ty_start_batch
	结束业务日期	ty_end_date
	结束批次号	ty_end_batch
	删除标识	ty_del_flag
	运行job字段	ty_job_name
当前表	业务日期字段	ty_data_date
	批次号字段	ty_batch_number

	运行job字段	ty_job_name
	删除标识	ty_del_flag
	删除日期字段	ty_del_date
	删除批次字段	ty_del_batch
	首次加载日期	ty_rec_init_date
当前全量表	业务日期字段	ty_data_date
	批次号字段	ty_batch_number
	运行job字段	ty_job_name
	删除标识	ty_del_flag
	删除日期字段	ty_del_date
	删除批次字段	ty_del_batch
	首次加载日期	ty_rec_init_date

快速入门

快速入门

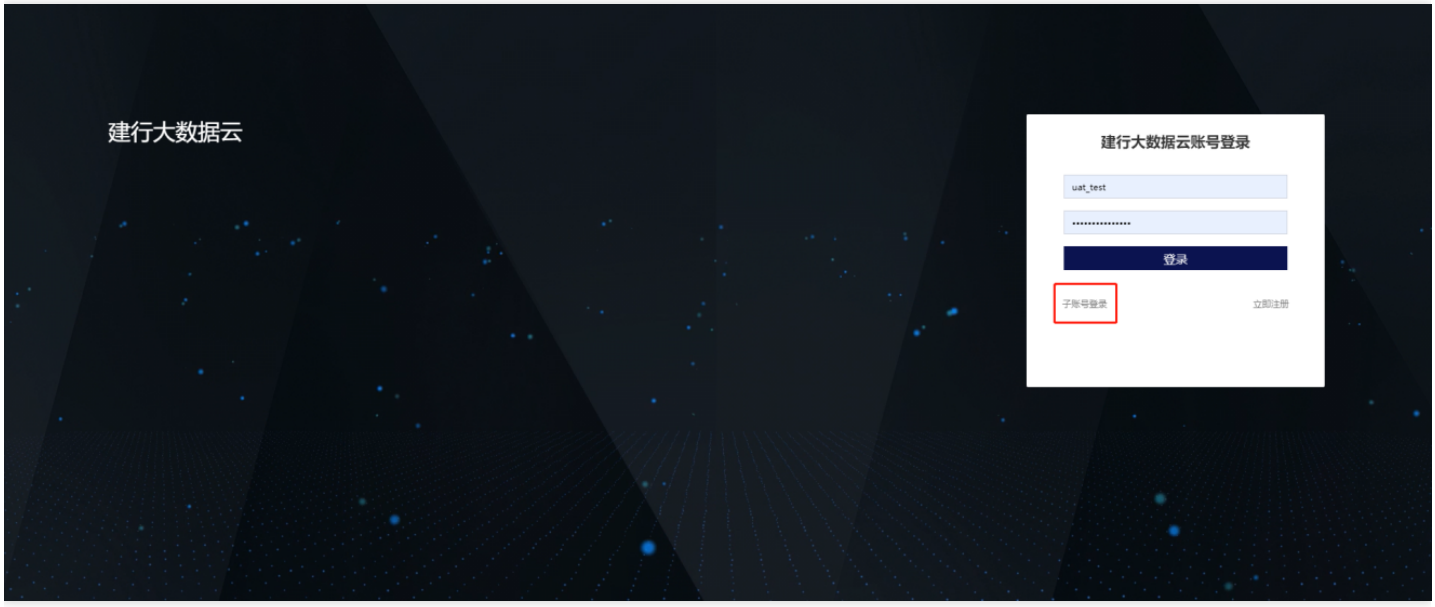
最近更新时间: 2019-10-28 03:14:27

本章节将带领用户创建一个简单的集成任务，并分别介绍进行数据同步、数据加工、数据整合、业务核检的具体步骤。通过这些步骤，用户可以快速了解如何使用大数据开发套件功能完成各类数据集成任务。

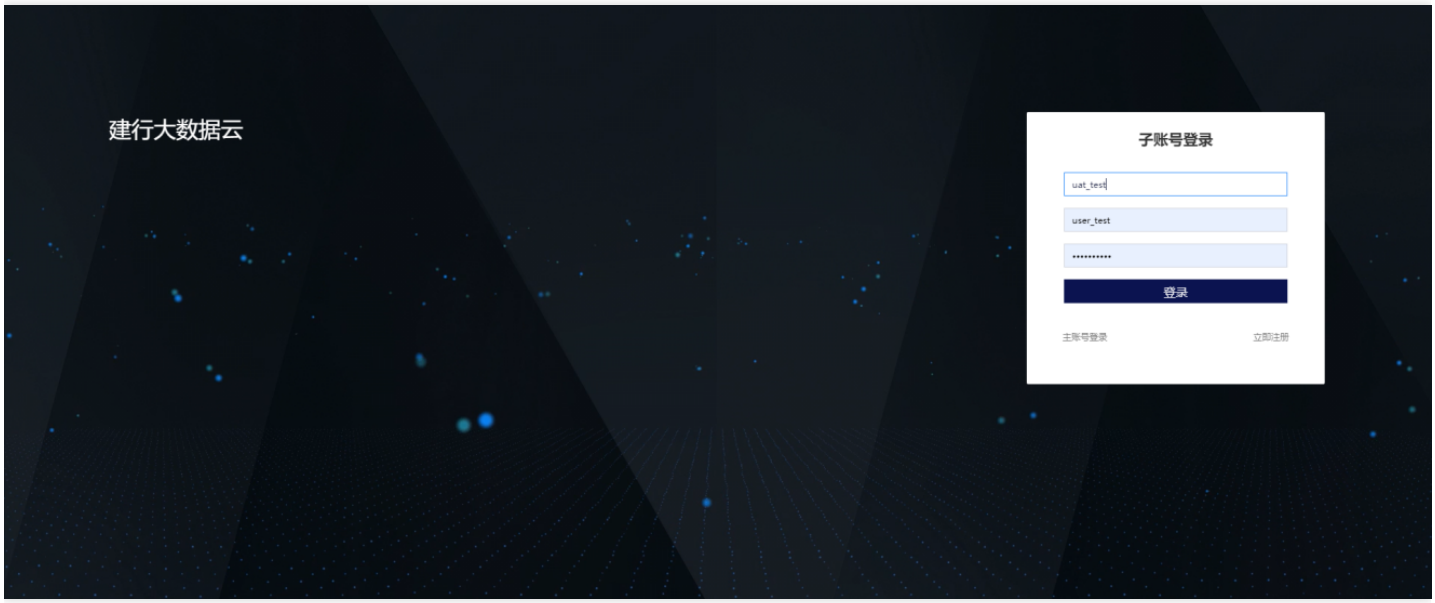
子帐号登录

最近更新时间: 2019-11-12 10:21:54

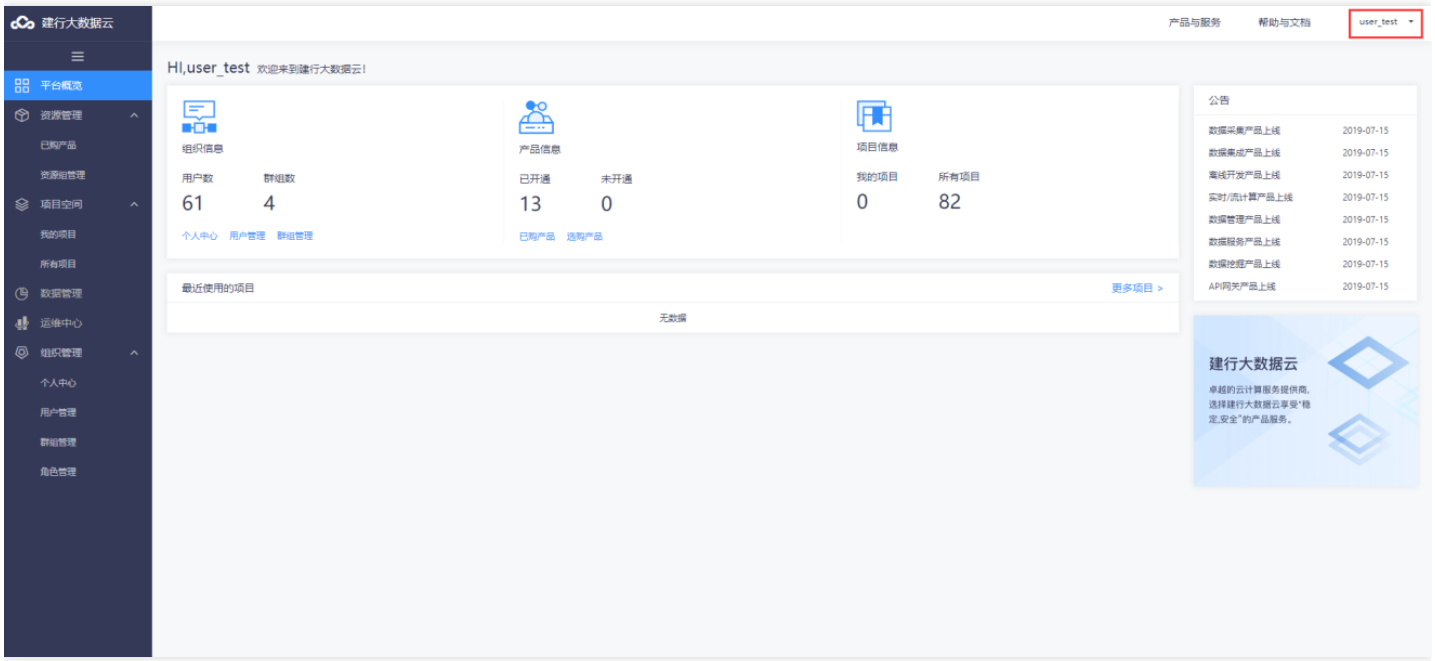
(1)进入租户控制台登录页。点击【子账户登录】



(2)显示子账号登录页面，填写租户名、子账户名、子账户密码，点击登录，即进入。



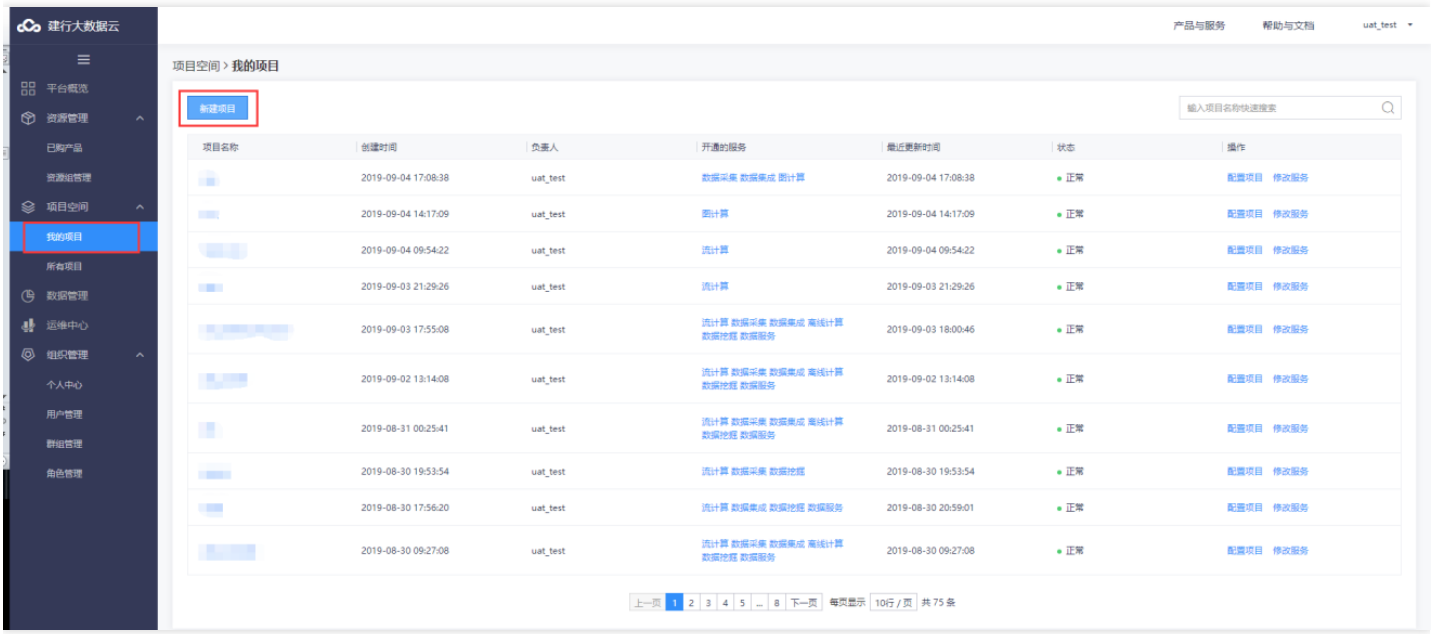
以子账户的身份进入大数据开发平台。



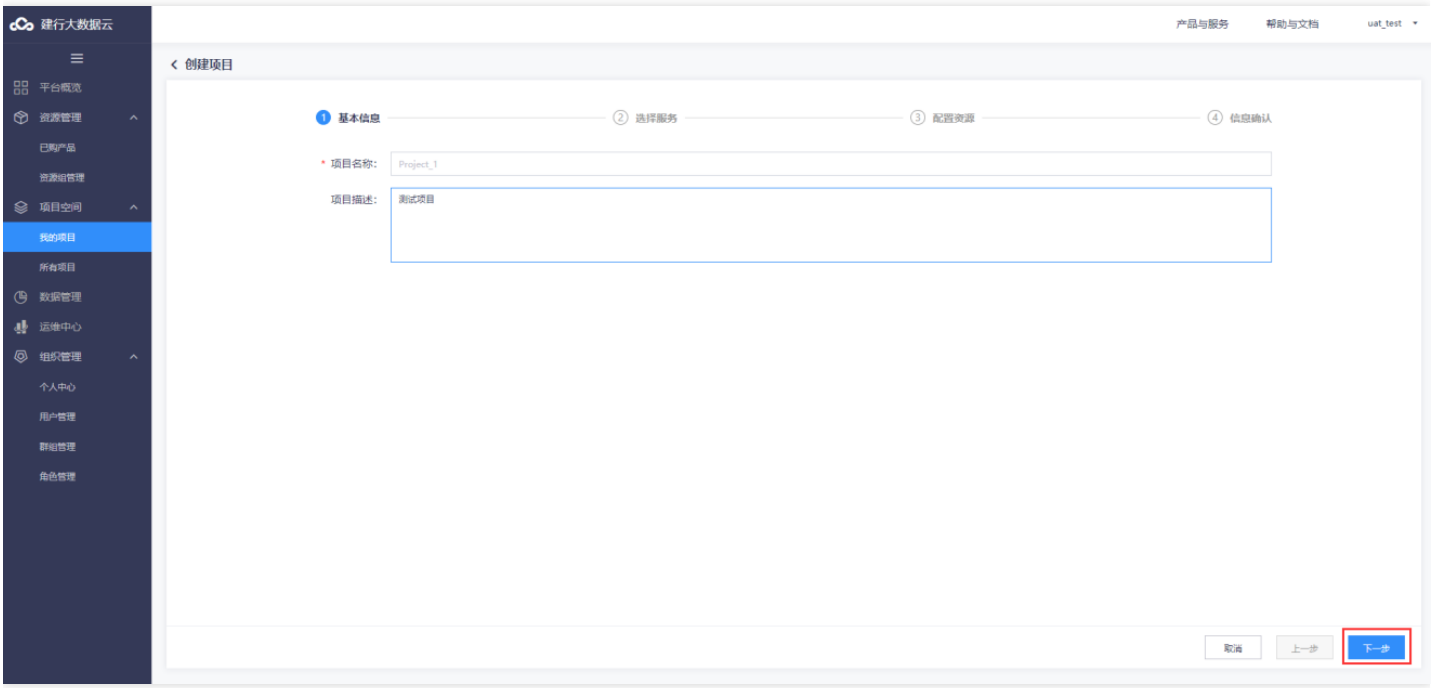
新建项目

最近更新时间: 2019-11-12 10:21:54

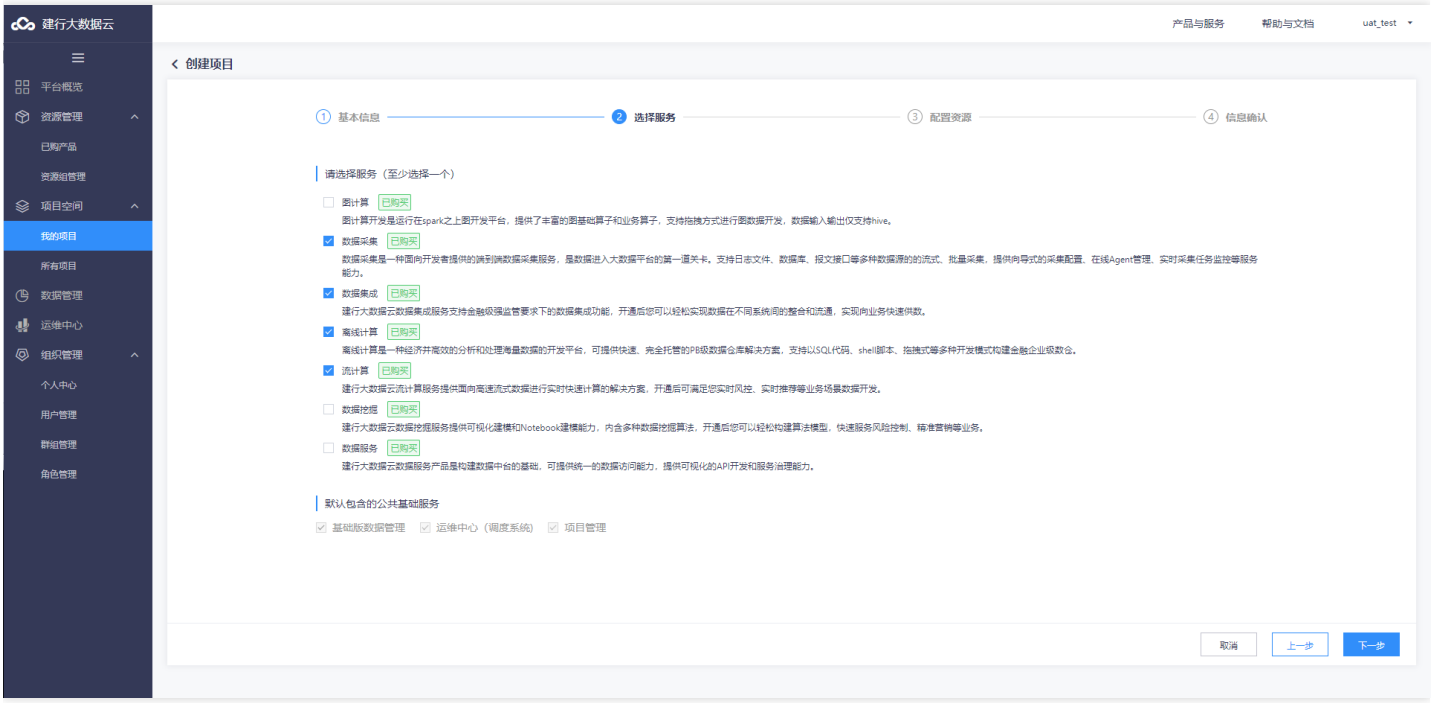
（1）选择项目空间下【我的项目】展示当前用户参与了哪些项目，点击【新建项目】创建一个新的项目。



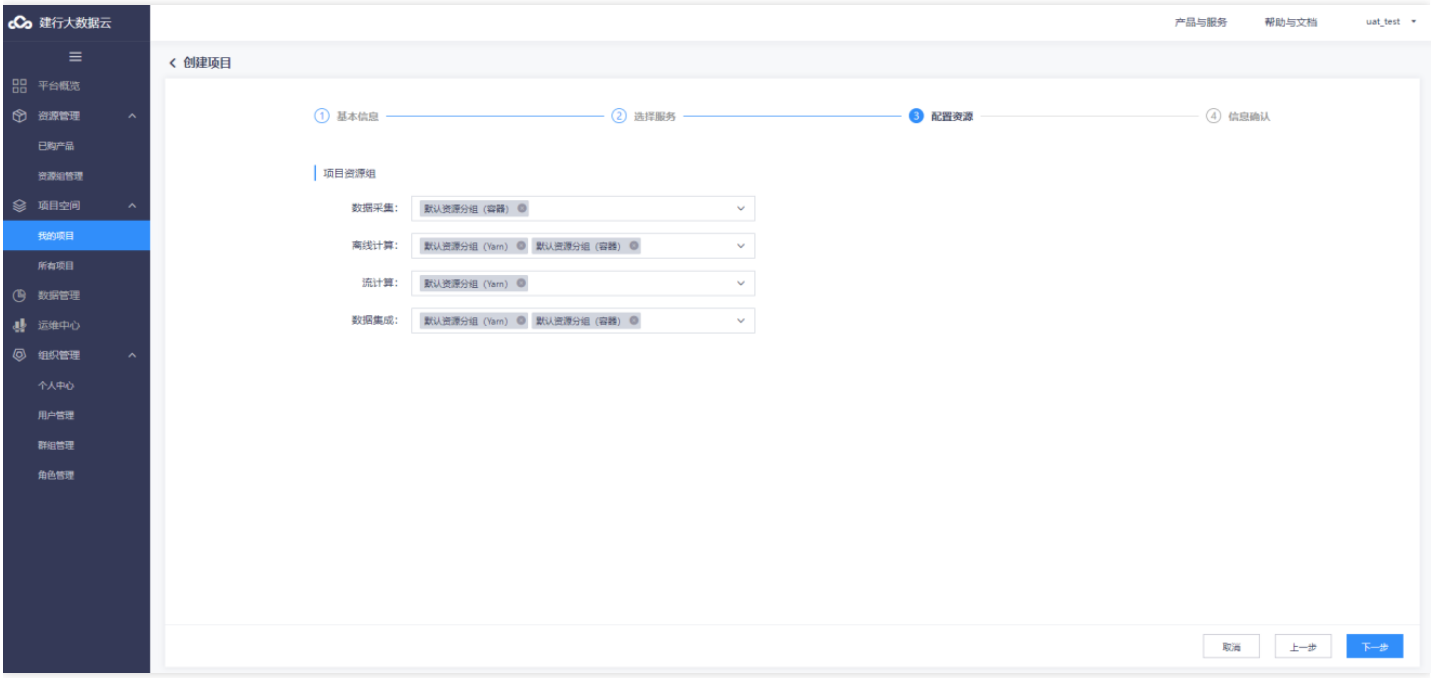
（2）输入项目名称，项目描述，点击【下一步】。



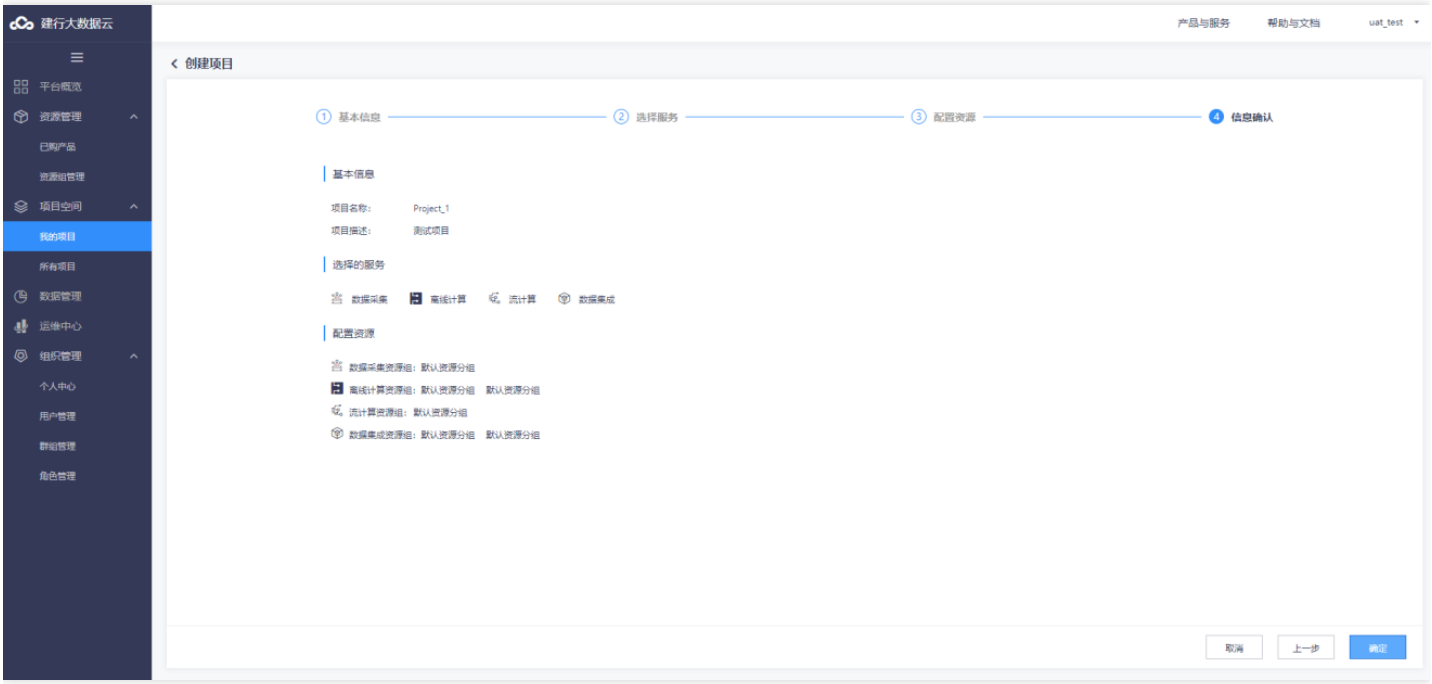
（3）勾选项目需要创建的服务。



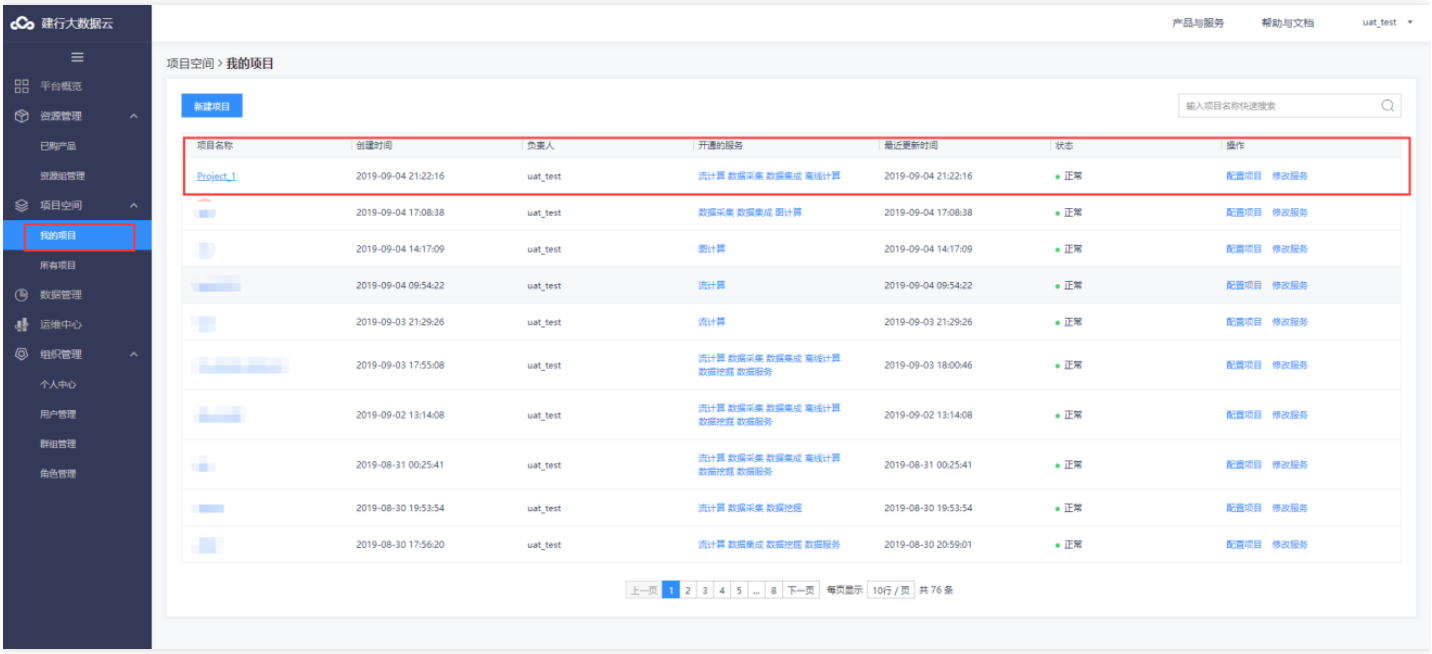
（4）为每个服务进行资源选择，每项服务都有项目资源后，点击【下一步】。



（5）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”。



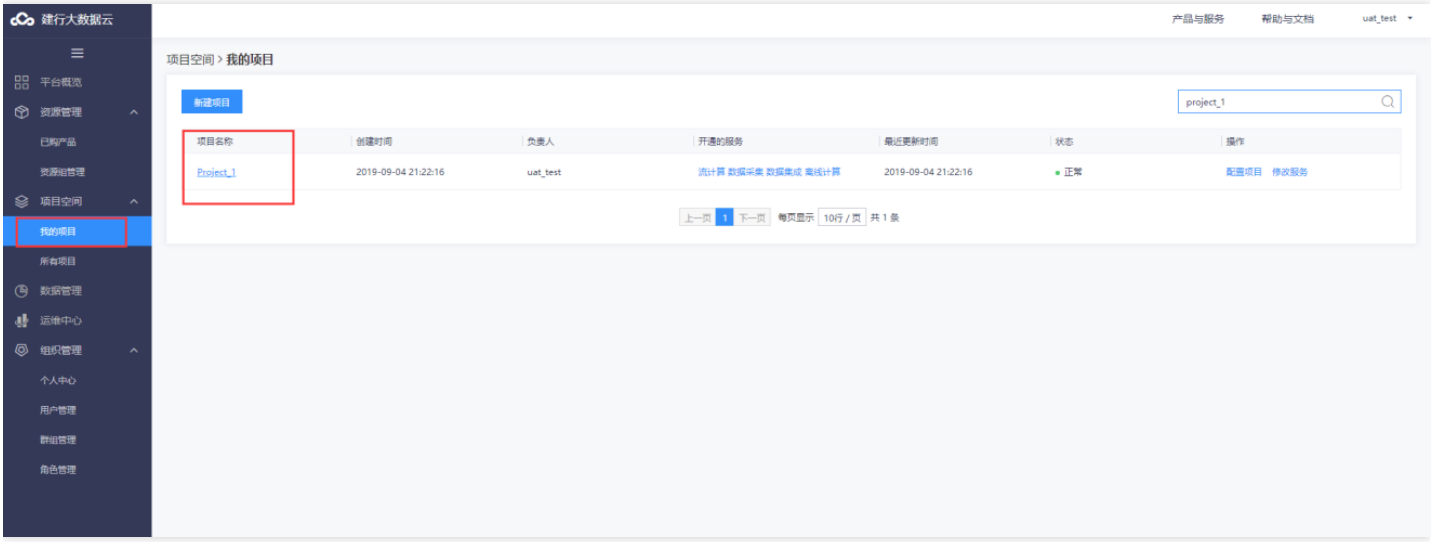
（6）回到列表页显示，刚刚创建的项目空间。



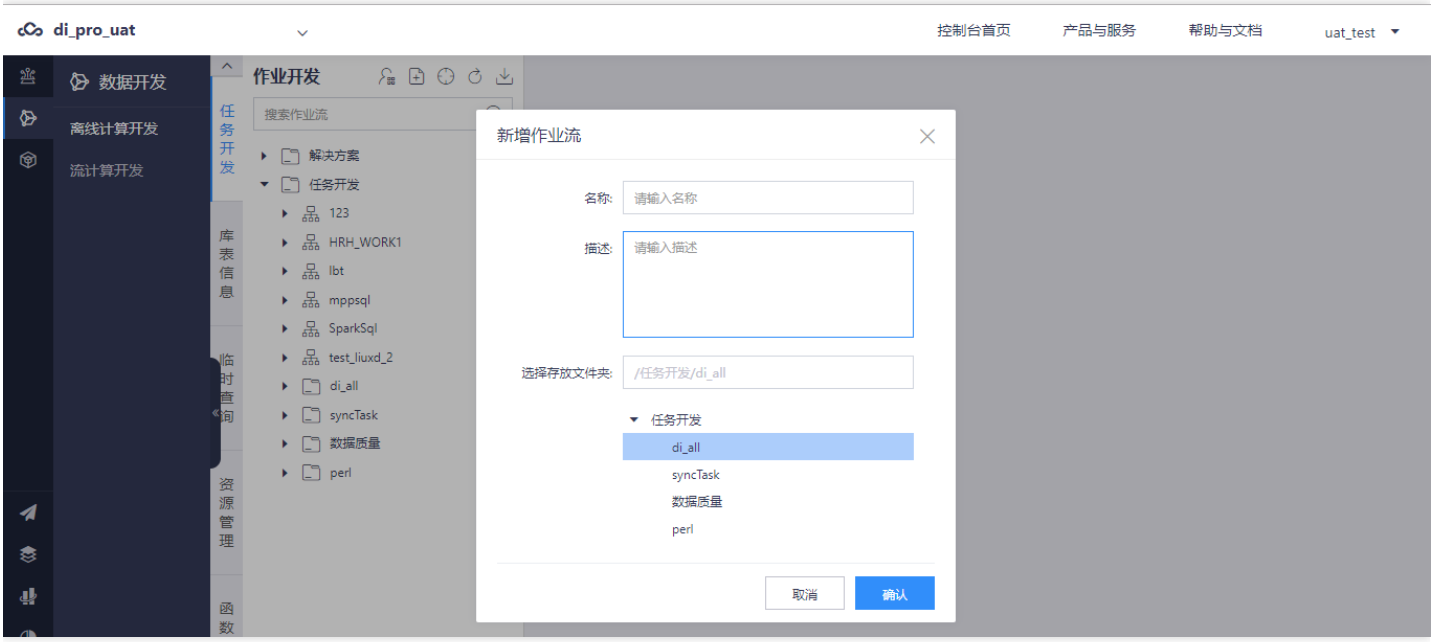
新建数据集成任务

最近更新时间: 2019-11-12 10:21:54

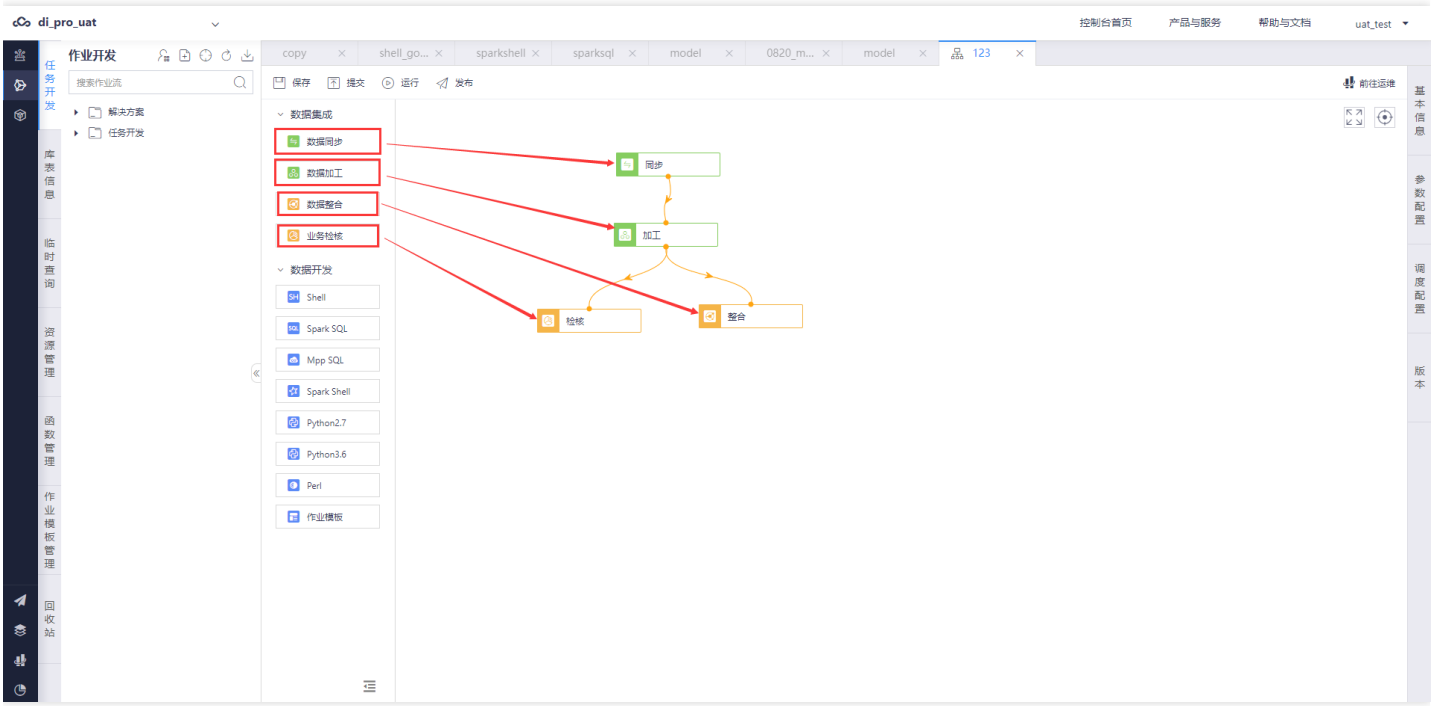
点击【项目空间-我的项目】显示项目列表页面，点击一个有权限的项目。



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。



在新建的作业中拖拽数据集成的某一个模块。生成对应的任务。

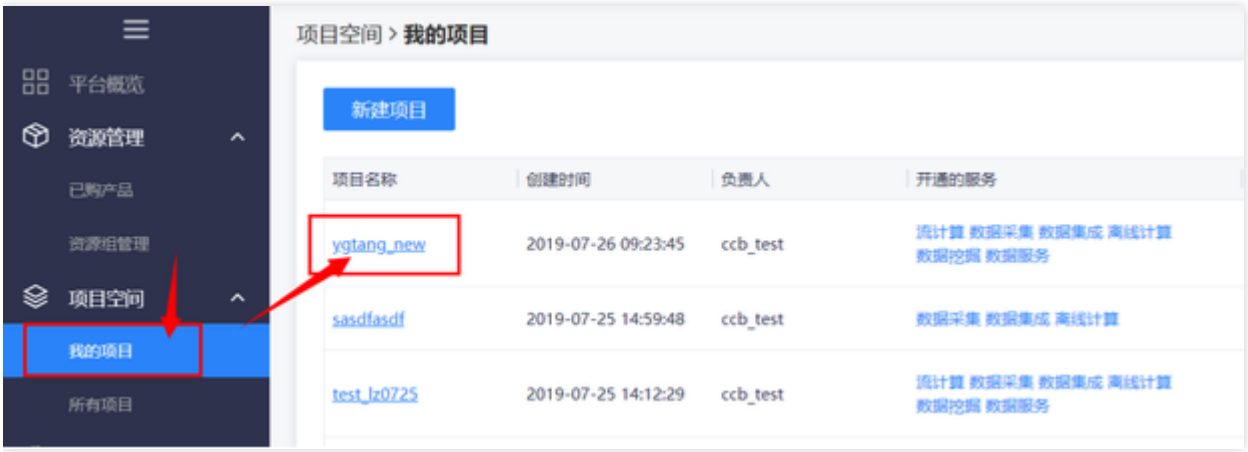


点击具体作业进入编辑界面。具体作业编辑使用参考【第五节–操作指南】

操作指南
数据同步

最近更新时间: 2019-11-12 10:21:54

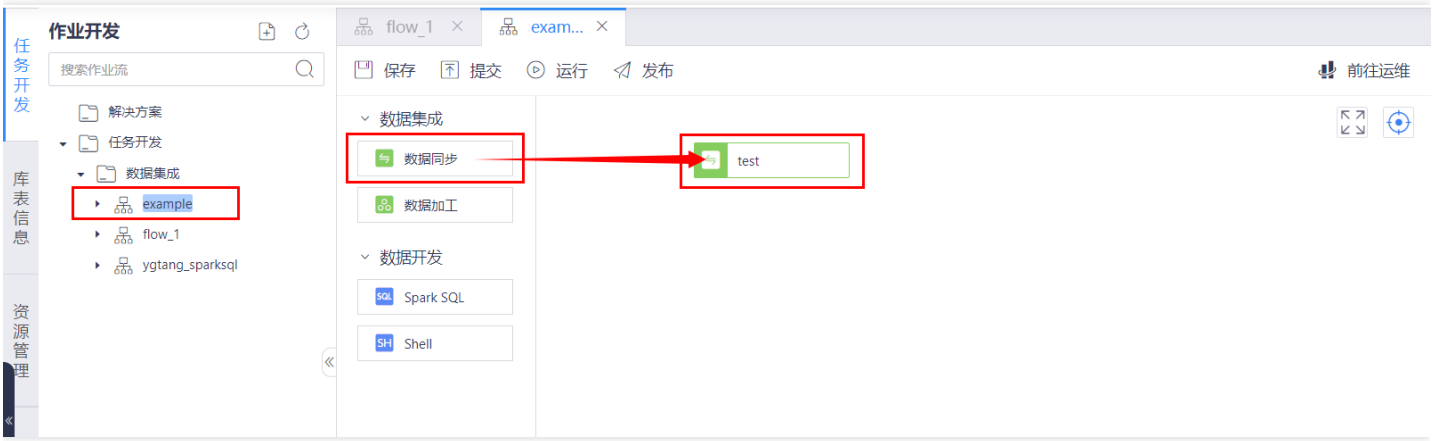
数据同步工具不仅能够满足传统数据集成服务在复杂网络环境下进行多种异构数据源的导入导出需求，同时在数据导入导出的过程中的进行数据清洗、去重、规范化等提高数据质量。防止脏数据、垃圾数据的传播。 进入【项目空间】->【我的项目】，点击项目名称进入大数据开发套件

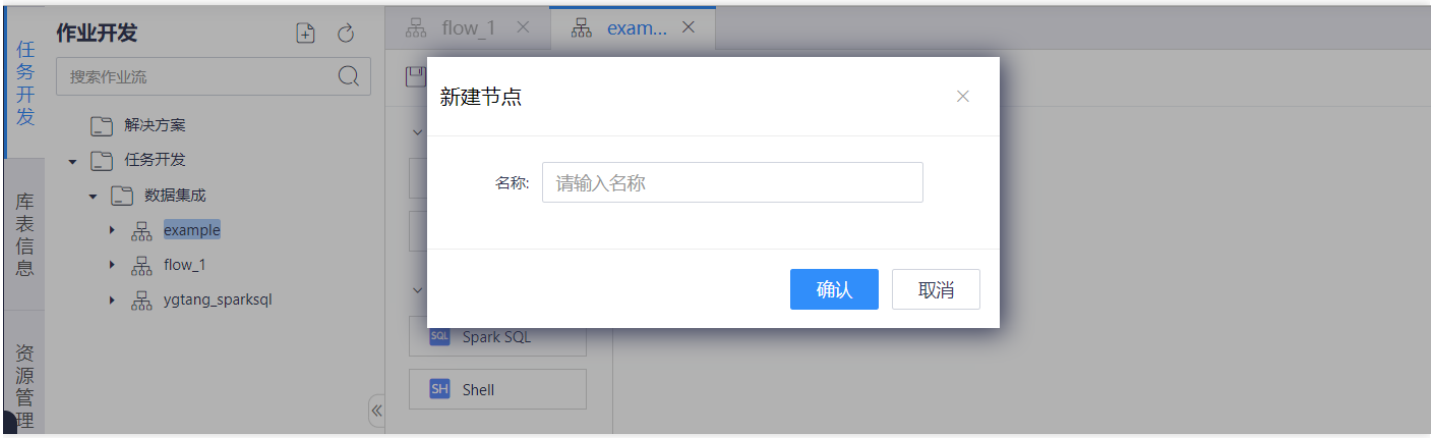


点击进入【数据开发】->【离线作业开发】。

选择【任务开发】，在左侧目录点击创建的作业流，新建一个作业流

双击作业流，进入作业流开发面板，拖拽数据同步插件，输入节点名称。





双击打开新建的同步任务，打开同步任务页面后整个同步任务分成三步。

- 第一步选择数据源表：

选择数据源的过程中可以在【数据过滤】中添加过滤语句，进行数据的增量同步。具体支持变量请参考。调度系统中变量设置章节。

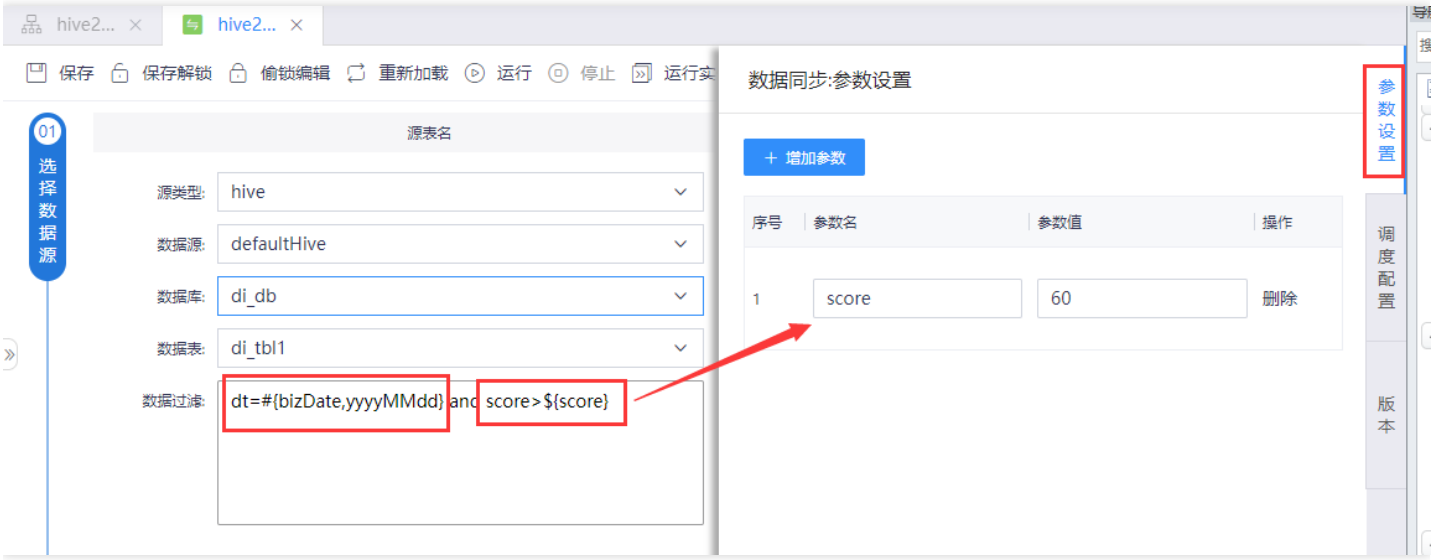
- 第二步选择数据目标表

- 第三步设置数据源表和数据目标表的映射管理。 在映射过程中左边字段信息来自源表，右边字段信息来自目标表。 用户可以在源表字段上进行字段的行级信息转换：进行字段格式转换、对字段应用系统函数、常量设置等。也可以新增字段进行字段转换。 在目标表字段中可以设置默认值，如有上游有数据传输下来使用上游字段，如果上游数据为空，使用默认值设置。

源和目标之间的连线设置表示数据的流向关系。

在数据同步开发过程中可以进行参数设置如下。 其中#{}为系统参数，具体提供系统参数可参考【调度系统–功能特性–变量设置】章节。系统参数不需要用户进行赋值，只需要进行格式设置既可。 ${}为用户自定义变量，用户自定

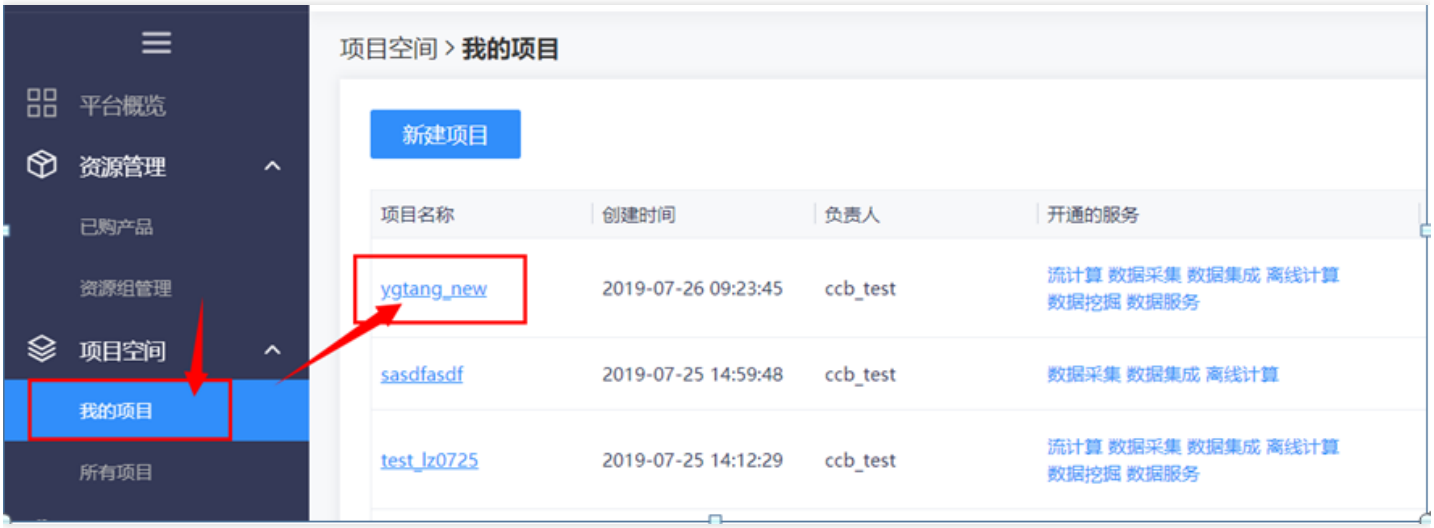
义变量需要用户在作业【参数设置】中进行参数赋值。



数据加工

最近更新时间: 2019-11-12 10:21:54

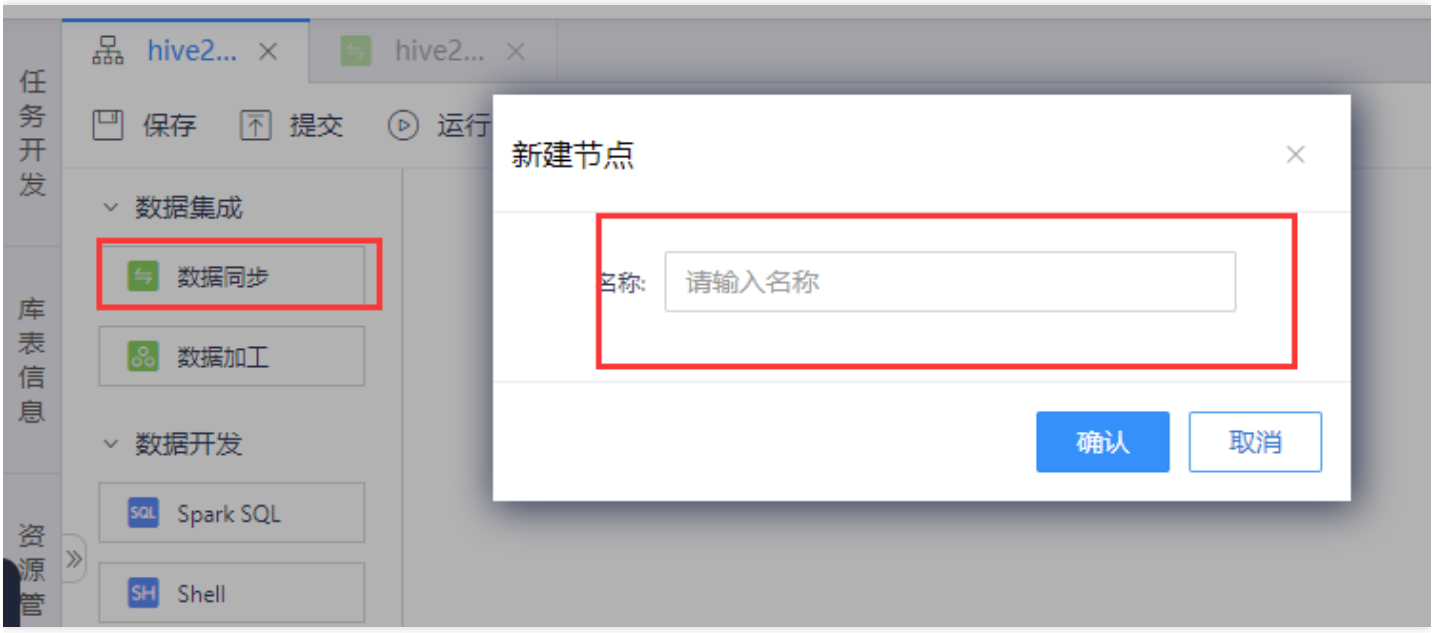
数据加工工具采用可视化拖拽的方式进行数据开发，降低开发门槛，使没有SQL经验的业务人员也能够进行快速的数据逻辑开发。 进入【项目空间】->【我的项目】，点击项目名称进入大数据开发套件

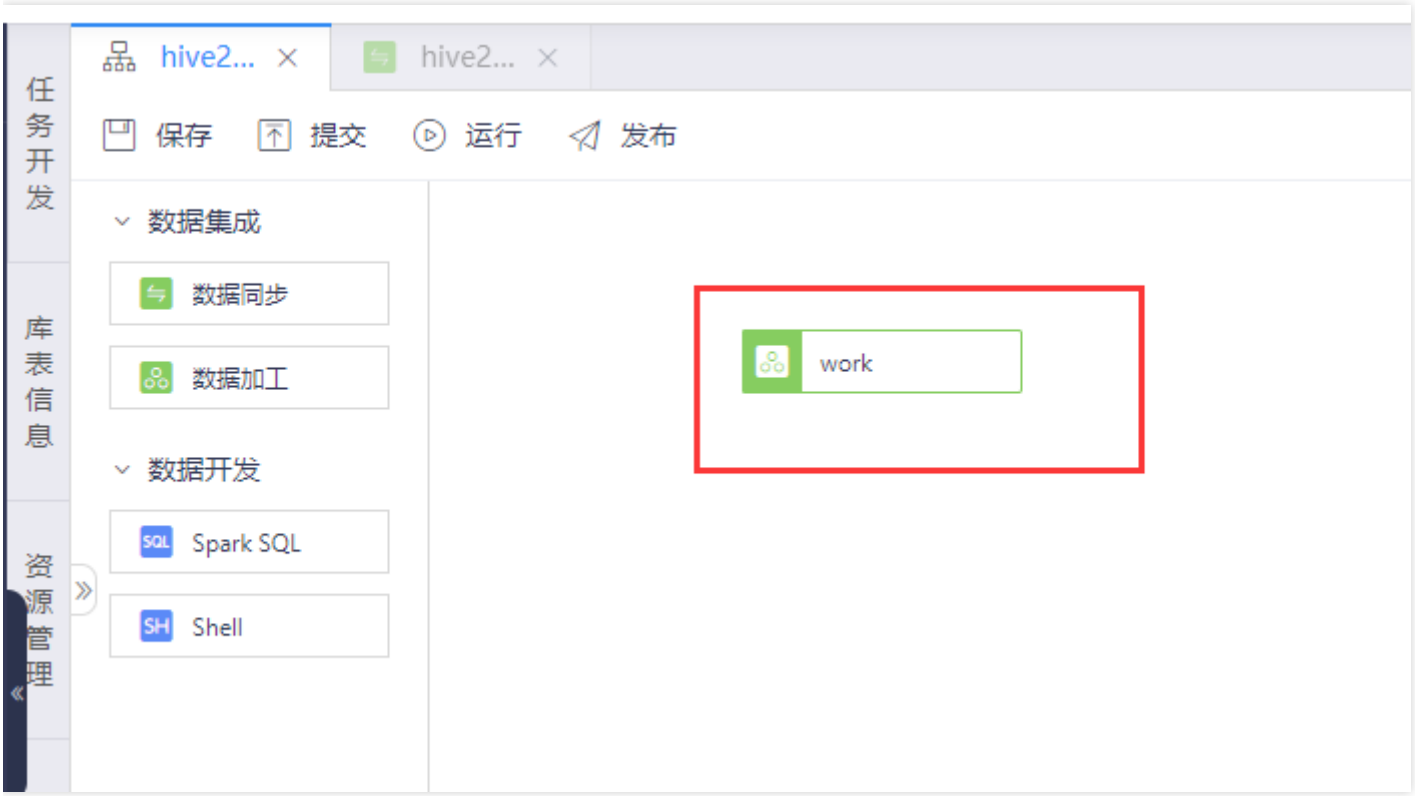


点击进入【数据开发】->【离线作业开发】。

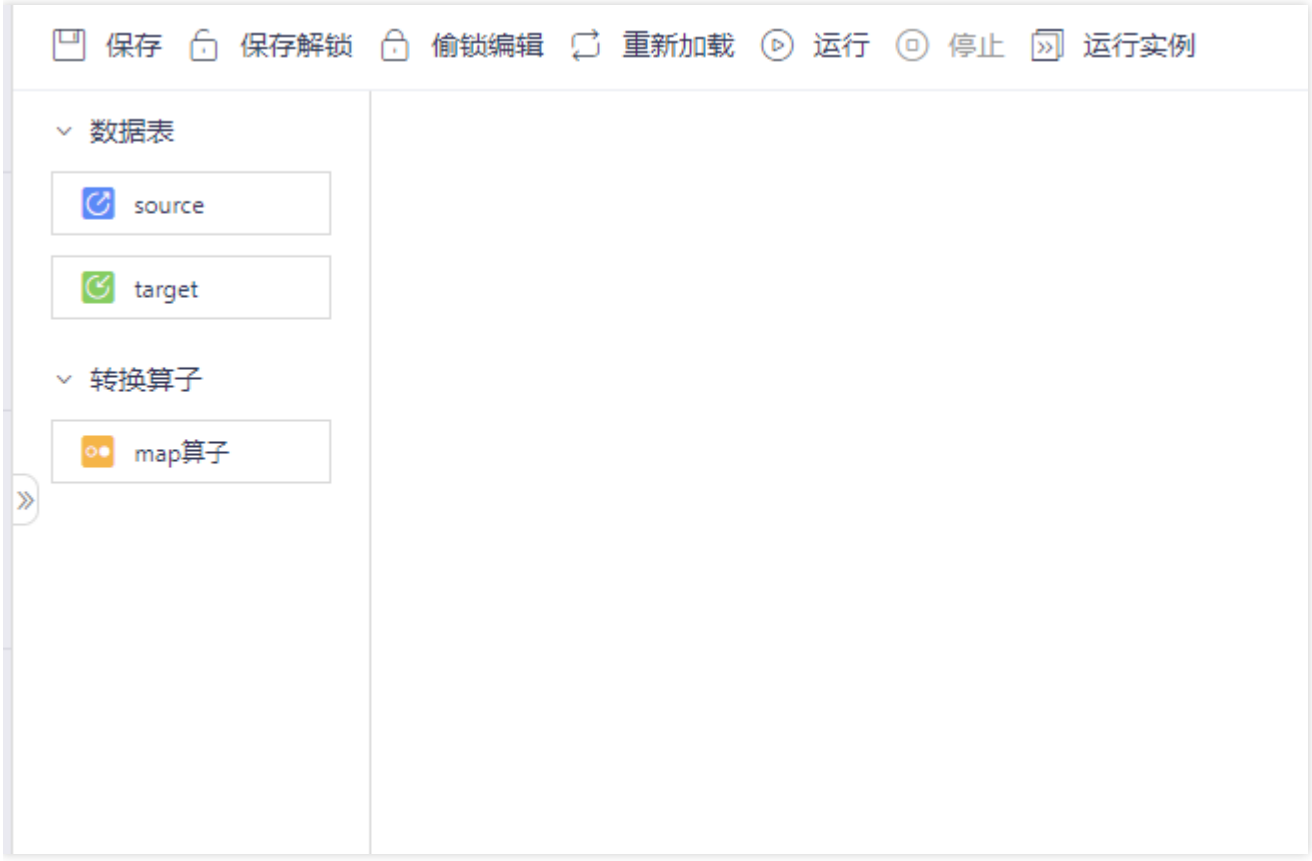
选择【任务开发】，在左侧目录点击创建的作业流，新建一个作业流

双击作业流，进入作业流开发面板，拖拽数据加工插件，输入节点名称。生成一个数据加工作业节点。





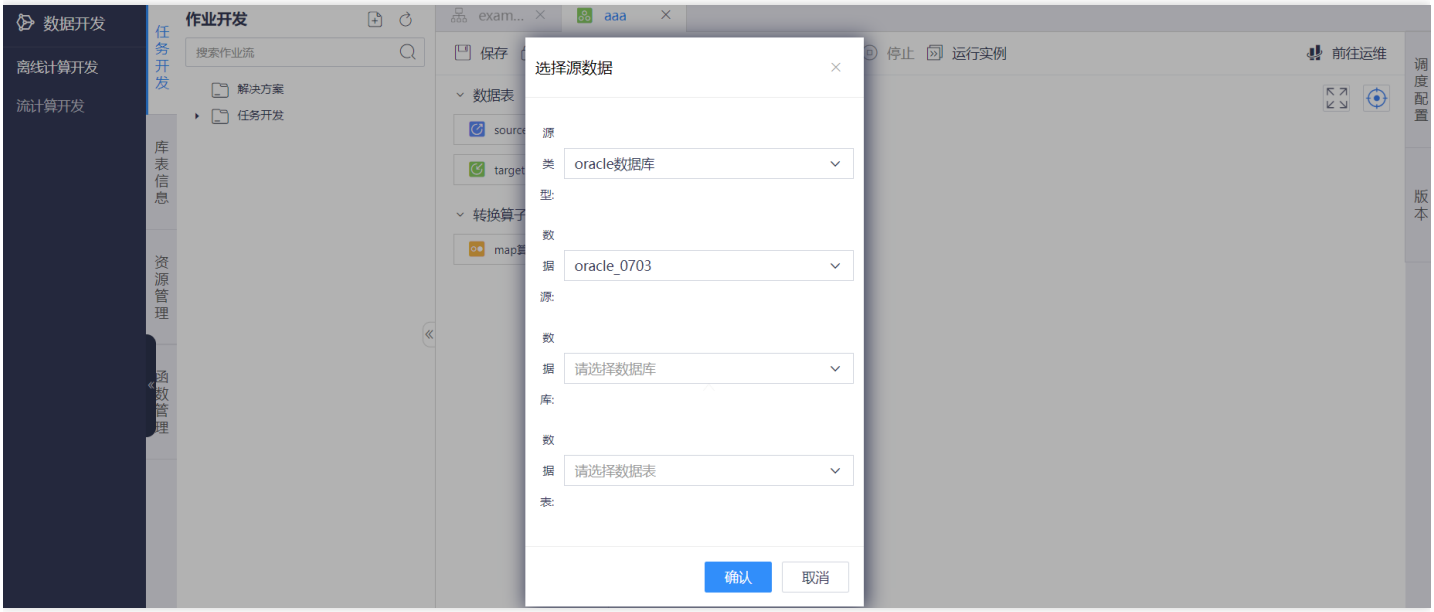
双击打开新建的数据加工任务，进入数据加工的开发界面。数据加工是拖拽式的开发过程，左侧显示了用户可拖拽的开发算子。



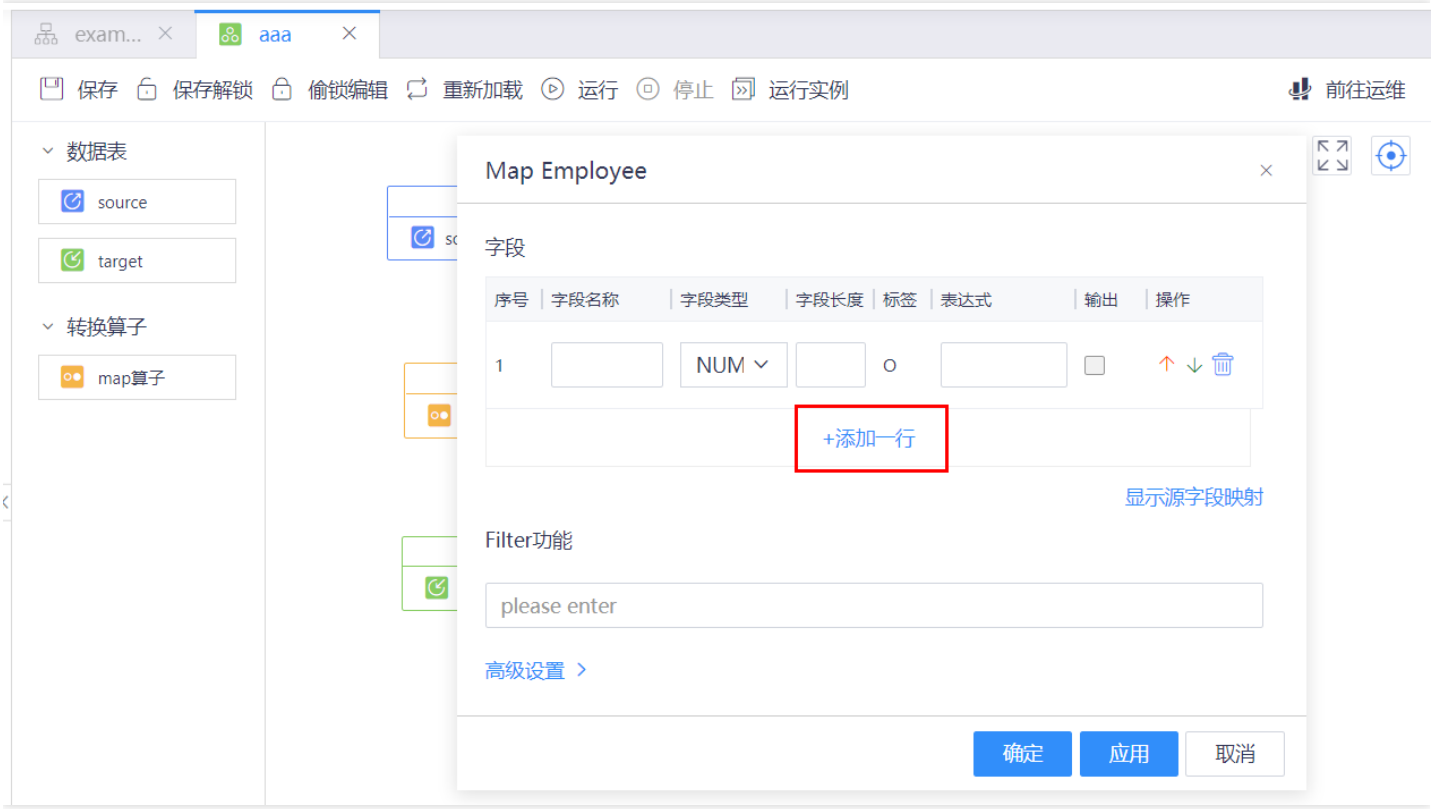
双击进入

加工任务，拖动添加源表和目标表

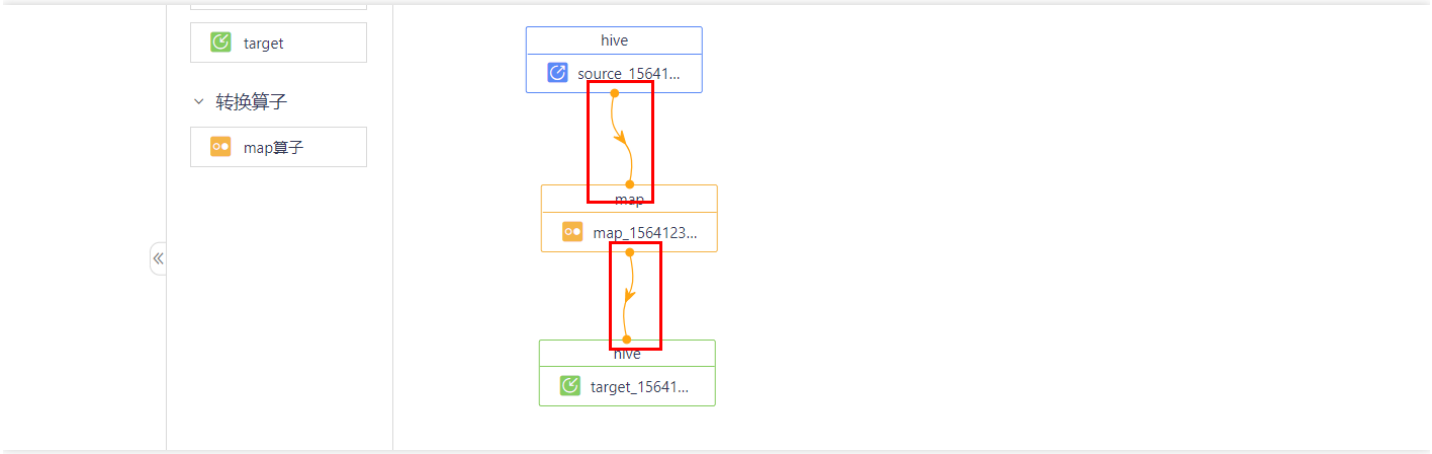
依次选择源类型–数据源–数据库–数据表



拖动添加转换算子，双击图标进行添加字段和填写功能备注



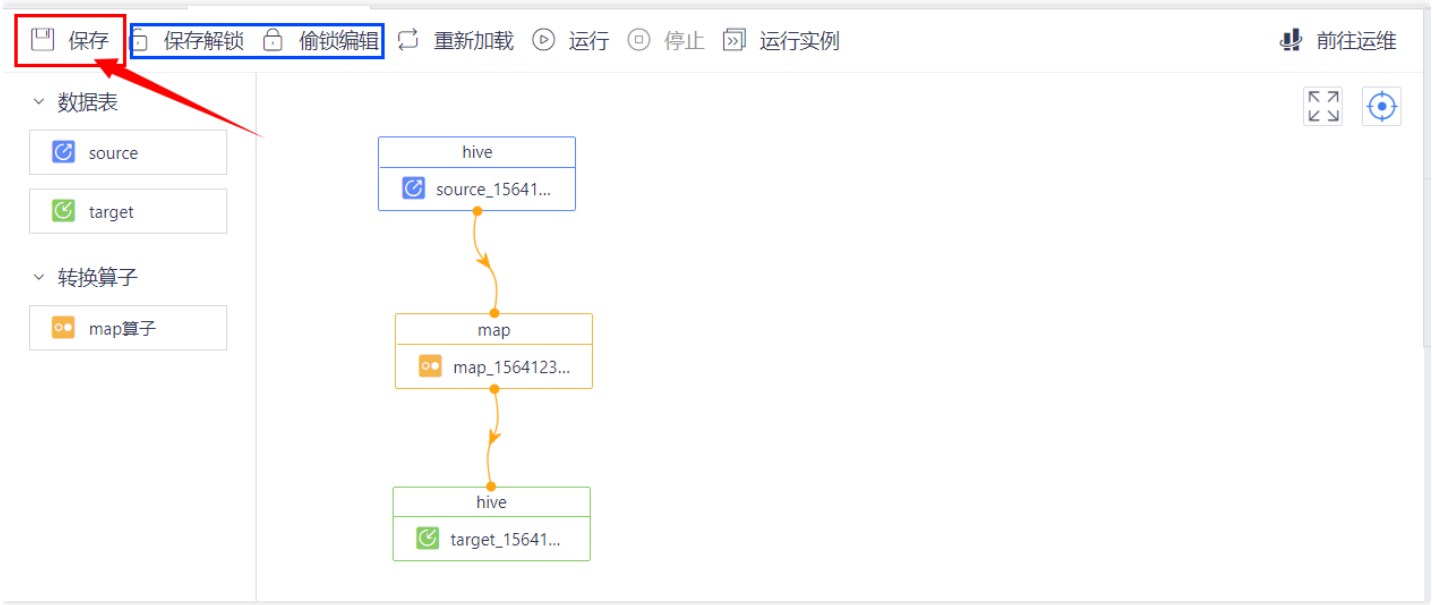
拖动连线确定关系



点击上方【运行】按钮进行测试，点击【停止】停止运行，点击【运行实例】进行查看

完成后点击【保存】保存当前编辑，如果选择了【偷锁编辑】，那么在同一时间其他用户不能进行修改，点击【保

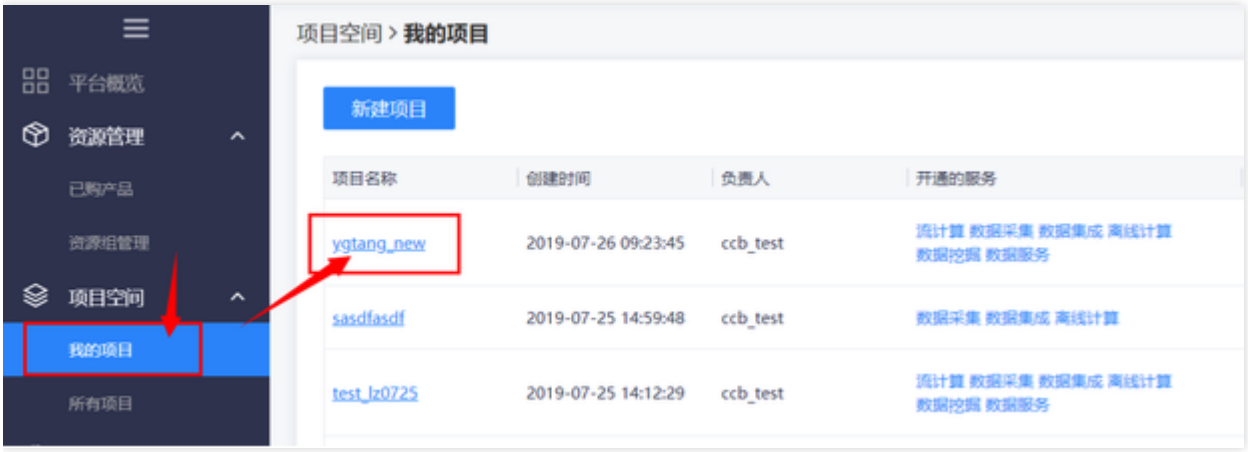
存解锁】可以解除锁定。


保存
保存解锁
偷锁编辑
重新加载
运行
停止
运行实例
前往运维
数据表
source
target
转换算子
map算子
hive
source_15641...
map
map_1564123...
hive
target_15641...

数据整合

最近更新时间: 2019-11-12 10:00:51

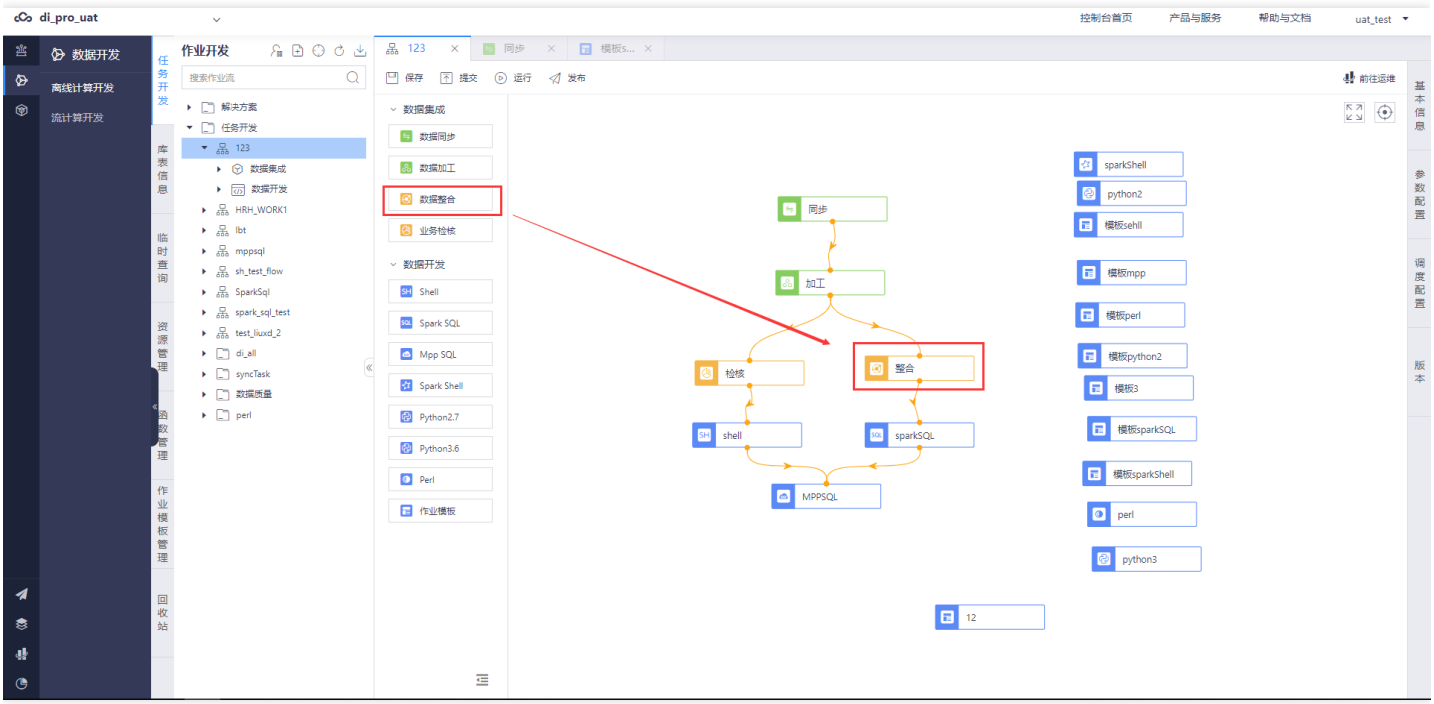
结合多年数据处理行业经验，沉淀固化通用数据整合模型，将贴源数据的处理过程从繁复的代码逻辑中解放，仅需简单配置既可完成复杂贴源数据整合。 进入【项目空间】->【我的项目】，点击项目名称进入大数据开发套件



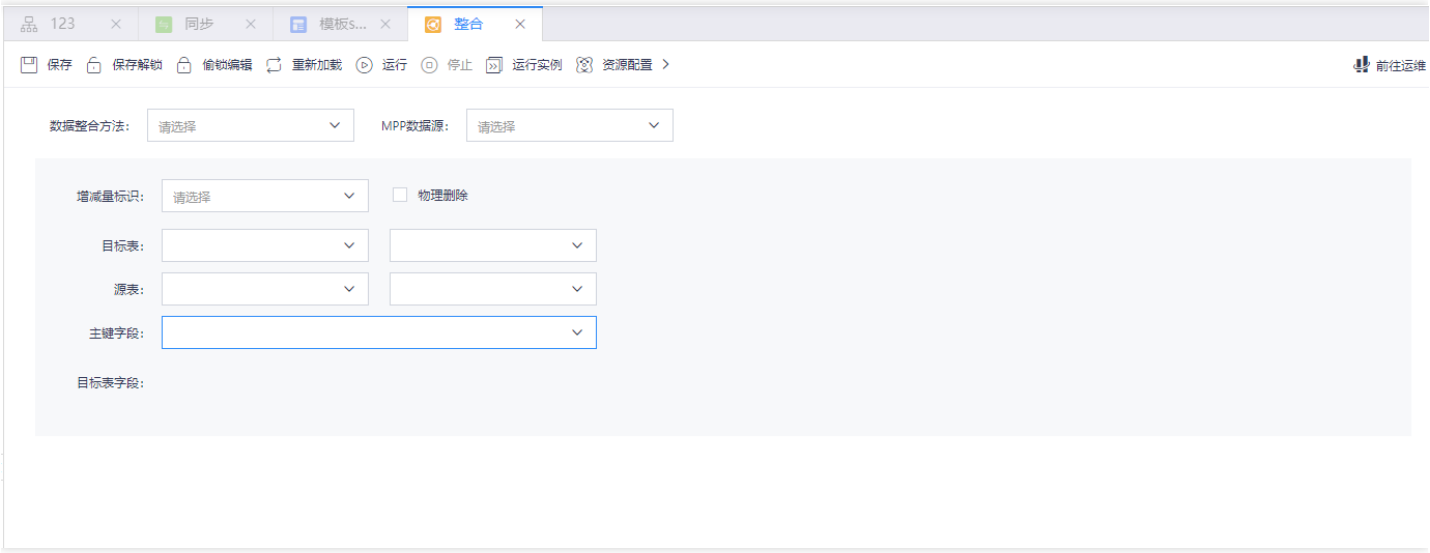
点击进入【数据开发】->【离线作业开发】。

选择【任务开发】，在左侧目录点击创建的作业流，新建一个作业流

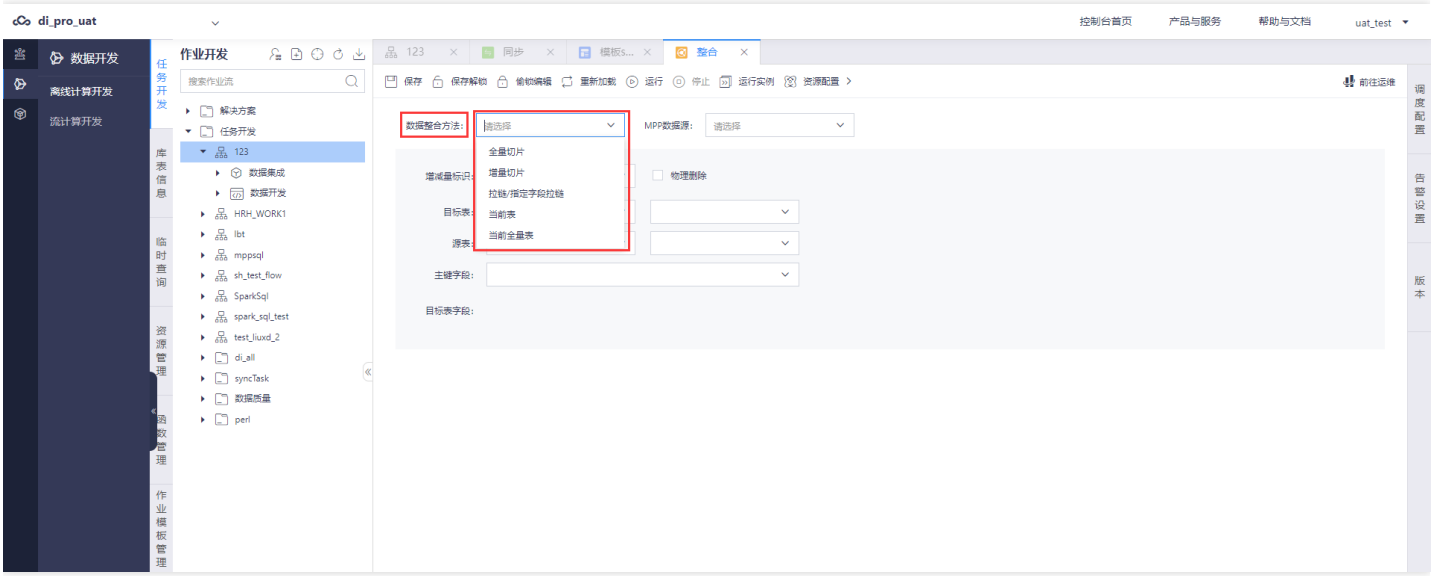
双击作业流，进入作业流开发面板，拖拽【数据整合】插件，输入节点名称，生成一个数据整合作业。



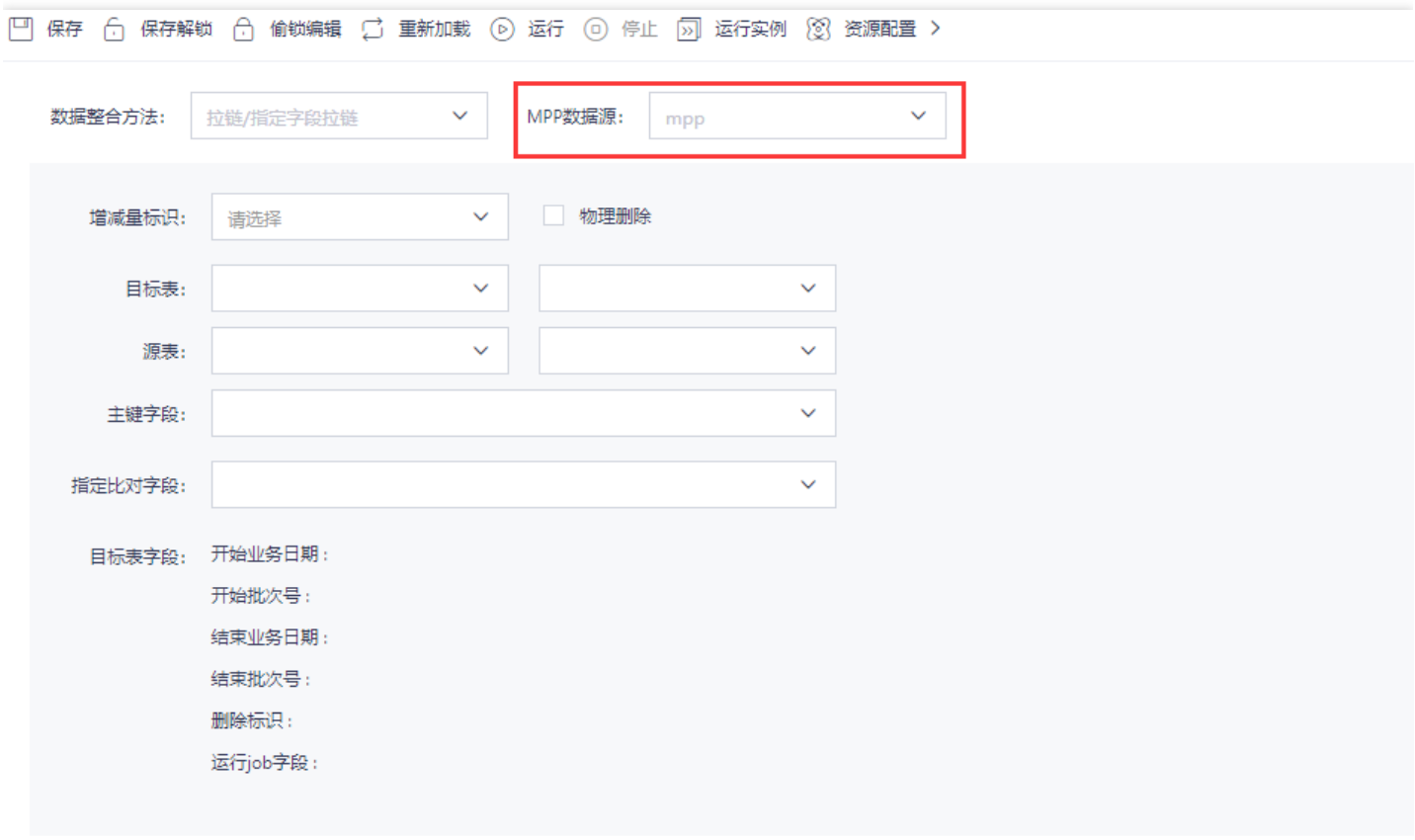
双击打开新建的数据整合作业，进入数据整合开发面板。



在数据整合面板中可以选择需要应用的数据整合算法，目前数据整合算法包括：全量切片、增量切片、拉链/指定字段拉链、当前表、当前全量表。用户可根据需求选择不同类型整合算法。

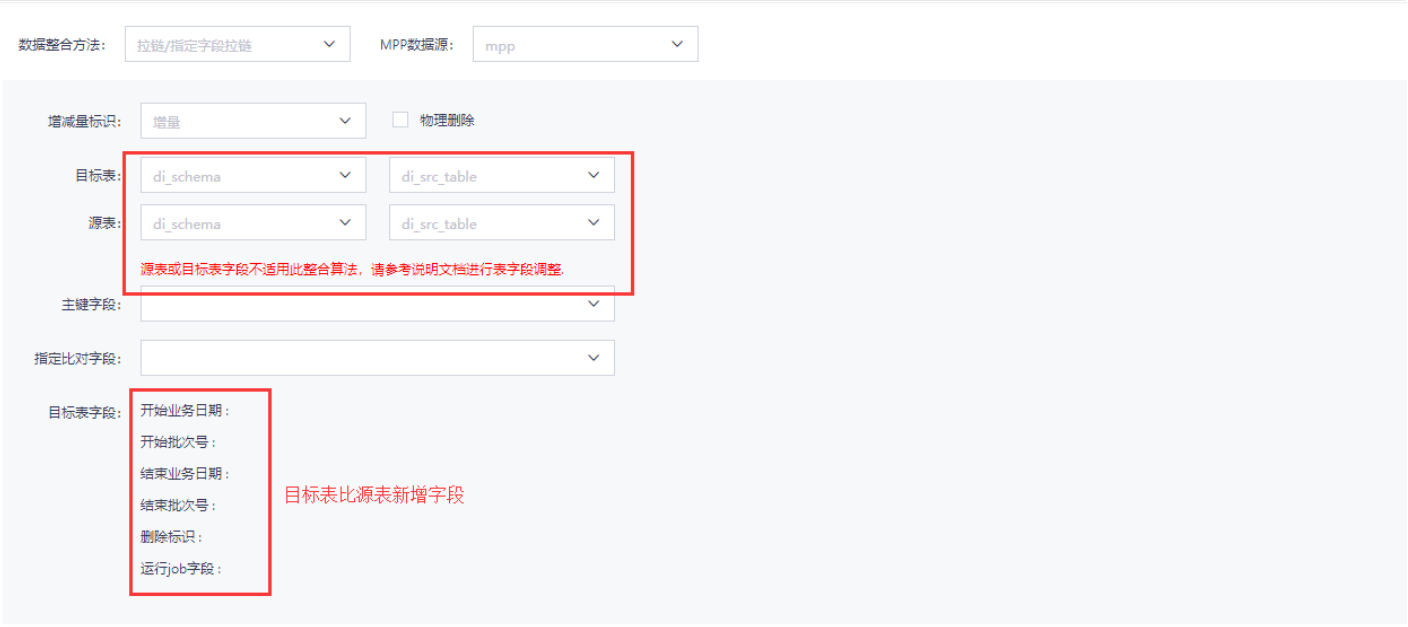


确定好使用的拉链表后，选择一个应用数据源，目前数据整合算法仅支持MPP数据源。



确定好数据源后，选择应用算法的源表和目标表，每种算法在应用时目标表会比源表多一定的特定字段，除了新增

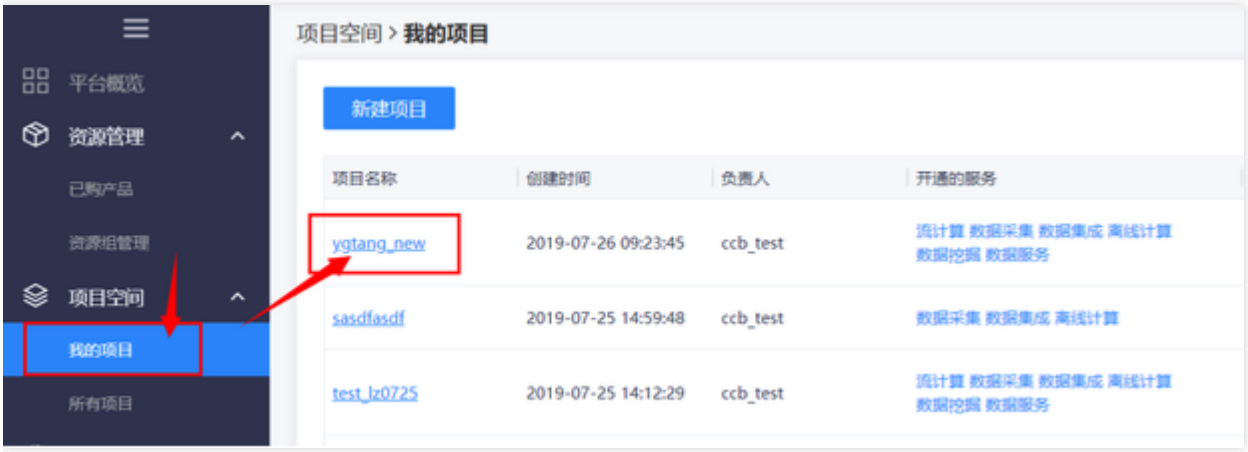
的特定字段外，其他字段需要完全保持一致。 如果选取字段不一致界面将提示进行表字段调整。



业务检核

最近更新时间: 2019-11-12 10:00:51

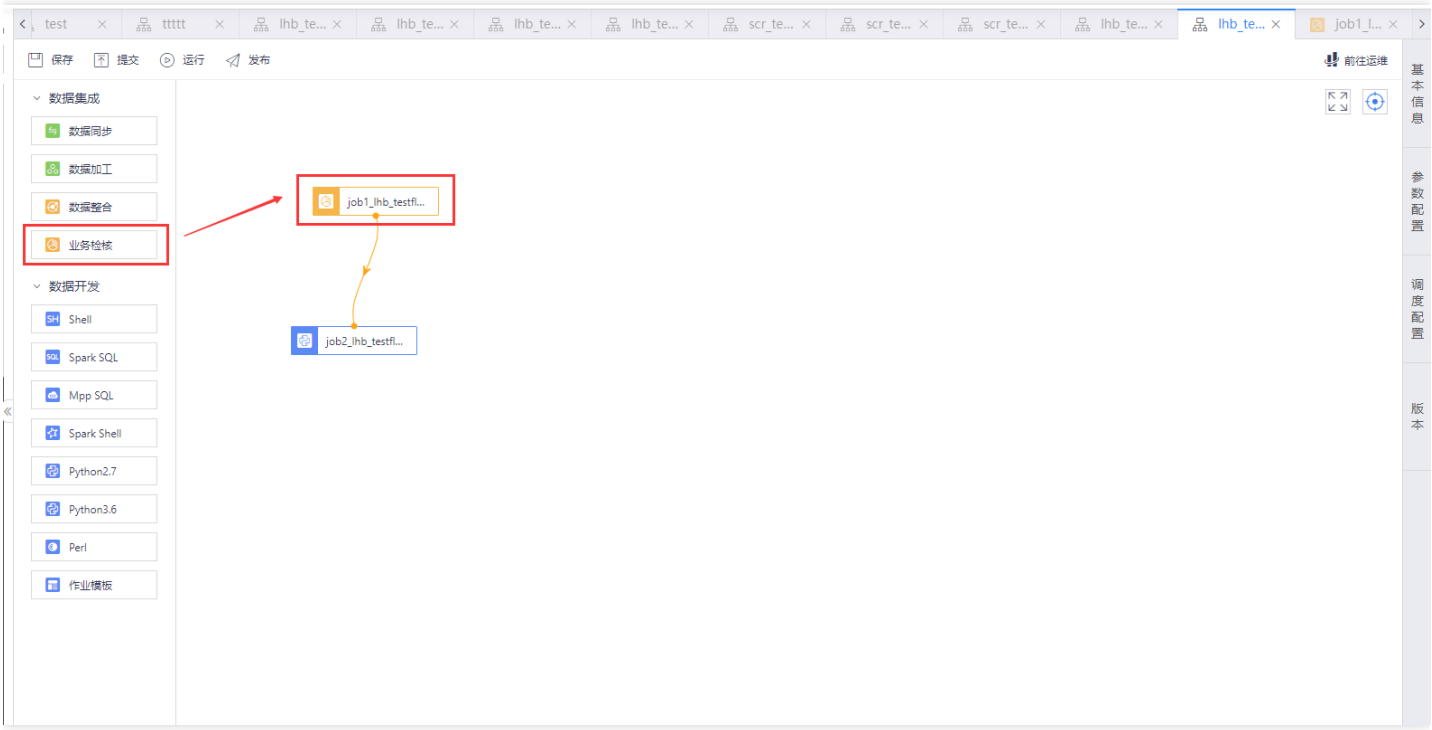
和数据质量中的业务规则无缝衔接，对数据进行全方位的规则检核。 进入【项目空间】->【我的项目】，点击项目名称进入大数据开发套件



点击进入【数据开发】->【离线作业开发】。

选择【任务开发】，在左侧目录点击创建的作业流，新建一个作业流

双击作业流，进入作业流开发面板，拖拽数据同步插件，输入节点名称。

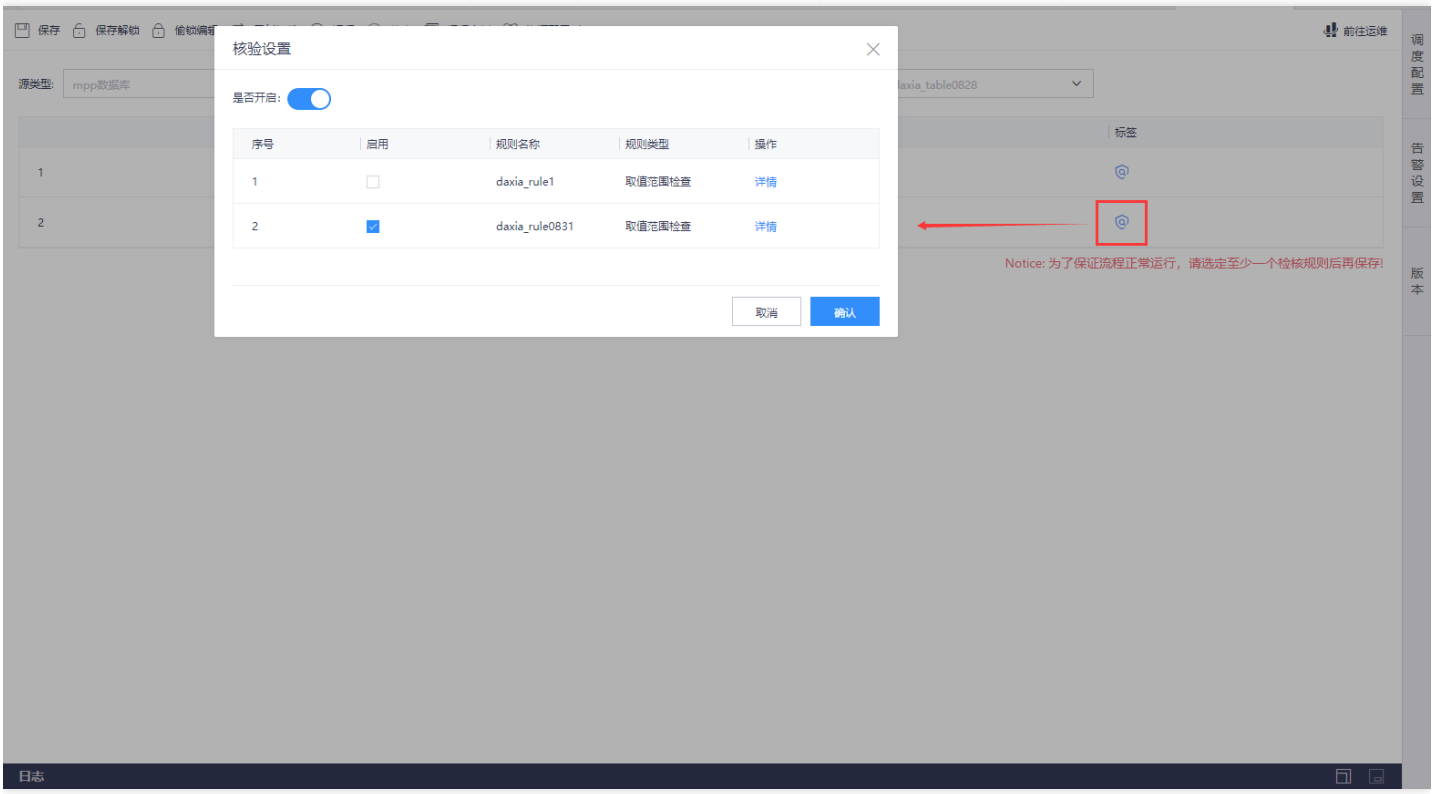


双击打开新建的业务检核作业，显示业务检核操作界面。

选择一种数据源后，确定表，表上面的字段信息就会展开。如果某个字段上配置了业务检核则会在标签字段上显示检核图标。

点击检核图标，弹出字段上的业务检核设置，可以看到字段上配置了哪些检核规则。也可以进行勾选确定是否在作

业运行是应用某个具体的检核规则。应用业务检核后，对检核结果查看可以去【数据管理–数据质量】中查询结果。



最佳实践

最近更新时间: 2019-10-28 03:13:45

1. 数据同步切分键设置：当同步的数据源数据过多时，可以使用切分设置，选取某个特定字段作为切分键，并指定一定的切分数量。目前切分键字段仅支持数字类型，切分数量按照表数据量大小来指定。
2. 每日增量同步数据：在进行数据源同步时需要对源表进行每日增量数据同，此时可以在【数据过滤】中添加系统变量${BizDate},在调度周期性运行时，调度系统的业务时间会对变量赋值，从而实现了数据过滤的效果。

常见问题

最近更新时间: 2021-09-16 10:08:31

创建的集成类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。 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。 出现这种问题需要从两个方面进行检查：

- 数据源确认是否包含测试和生产 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。

- 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。

词汇表

最近更新时间: 2019-11-26 15:30:16

- 数据集成 数据集成提供了一整套包括数据同步，数据加工，数据整合以及业务检核的数据加工处理工具集合。满足多种业务场景，快速上手。
- 数据同步 稳定高效的数据同步工具。能够在复杂的网络情况下进行异构数据源之间数据的同步迁移。
- 数据加工 可视化拖拽式的数据加工工具，满足不同数据源在数据加工过程中的整合，转换，聚合等。降低数据加工门槛，快速获得数据加工处理能力。
- 数据整合 结合多年数据处理行业经验，沉淀固化通用数据整合模型，将贴源数据的处理过程从繁复的代码逻辑中解放，仅需简单配置既可完成复杂贴源数据整合。
- 业务检核 与数据质量中的业务规则无缝衔接，对数据进行全方位的规则检核。
- 数据源 数据集成所处理的数据来源，支持多种不同类型的数据来源，且支持不同数据源之间的转换。
- 脏数据 脏数据是指数据格式本身不符合规范，或者不满足用户定义的格式的数据。脏数据会影响干扰后续的数据处理，造成数据偏差，数据错误等。
- 插件 插件指用户在开发界面上可操作的最小单元。一个插件相当于一个作业类型，当用户拖拽一个插件后生成一个具体的作业。
- 算子 算子指在以拖拽形式开发的插件内部用户可进行操作的最小单元。单个算子无法进行运行，需组合成一个处理逻辑后作为一个作业整体运行。