Data Formula

产 品 操 作 手 册

浪潮工业互联网股份有限公司

目录

1 产品概述

1.1 产品简介

Data Formula 是由 Inspur（浪潮工业互联网股份有限公司）自主研发的一款企业级数据中台软件产品。其基于并行计算技术架构，具备操作简单、部署灵活、快速响应等特点。Data Formula 可广泛应用于各行各业，从百亿级数据量的企业到各垂直中小企业，专注为企业提供数字化运营基础平台，解决各行业的业务数据分析需求。

Data Formula 可以帮助企业搭建功能完整的数据中台 ,提供了从数据汇聚、数据处理、模型管理，任务调度，算法管理，数据服务于一体的完整解决方案。Data Formula 致力于帮助企业快速搭建数据基础平台，及时通过数据发现问题进而改进业务。

1.2 核心价值

- **消除数据孤岛，整合业务数据**

众所周知，我国的企业信息化发展相对落后，企业早期缺乏长远的 IT 建设规划，发展至今，各个业务系统仍相对独立，导致数据孤岛长期存在。随着大数据时代的到来，数据来源变得多样化，数据结构也更加复杂化，这使得企业在数据采集、ETL 等方面会耗费很大精力。业务数据的过度分散，让企业很难从全局角度去分析业务的发展情况。

针对企业数据孤岛问题，Data Formula 提供了完善的解决方案。基于自主研发的数据连接器框架，Data Formula 可以对接企业内各个业务系统，包括 ERP、CRM、财务系统、日志系统等，帮助企业整合所有业务数据，还可以对接数据文件、API、NoSQL 数据库等。此外，Data Formula 基于并行计算架构的存储体系，支持海量业务数据的处理，确保数据

需求响应速度可以达到秒级。

- **数据资产管理，掌控企业数据脉络**

企业数据资产，包括了数据模型，数据目录，数据集，以及针对这些数据资产的转换处理能力，包括数据算法、AI 算法等。

Data Formula 摒除了传统数据管理模式中，复杂，难以理解的数据管理方法，改用业务视角的数据管理模式，从数据模型的增删改查，到数据集的转化处理方法的管理，都让数据管理的复杂度进一步降低。可以使得数据管理人员，甚至业务人员能更好的理解数据。

Data Formula 还重新设计了数据质量模块 ,让业务从数据转换的角度去查看数据质量，更好的体现了数据为业务服务的思想。除此之外 ,数据血缘关系也得到了重新的设计和梳理，更方便的查看数据的处理演变过程，为数据业务化提供了进一步的支持。

- **数据服务 API 化，提升服务效率**

数据中台另一项重大的改进就是数据服务化，通过自定义的数据服务 API，数据中台可以快速的响应前端业务需求变化。

不同于以往的数据仓库产品，Data Formula 无须通过数据库的账号和密码进行数据共享。数仓时代的账号密码方式，有诸多弊端，一个是数据安全性无法保证，另一个数据服务性能无法保证。

在新的 Data Formula 自定义 API 的服务模式下，操作人员可以通过界面配置，直接使用 RestfulAPI 的模式暴露数据，提供给前端应用，可以增加数据权限校验，也可以通过缓存，增加 API 的数据服务性能，更好的为前端业务系统服务。

1.3 产品特性

- 异构数据源整合： 兼容多种数据源，可接入企业内部各类业务系统 API、各种经典

关系行数据库（Oracle， SQL Server， MySQL， DB2 等），各种 NoSQL 数据库（MongoDB 等），各种数据文件（CSV， EXCEL），轻松集成整合所有相关业务数据；

- 数据模型管理：可以根据不同的业务域，组织管理数据模型；支持数据模型的增删改查，能偶满足业务快速变化的需求；
- 数据加工处理：中台产品支持标准的数据处理算子，通过算子堆叠和任务调度，完成数据从一个数据集到另一个数据集的转换过程，替代了传统的人工 ETL 过程，随着模型和数据集的调整，对应的转换过程也可以自动调整；
- 数据算法（AI）管理：为了应对业务的复杂性，Data Formula 数据中台支持复杂的算法加工能力，可以对数据进行标签化，或者通过 AI 机器学习算法，进行复杂数据处理和业务处理。
- 数据质量管理：重新设计了数据质量管理，从最新的数据处理的理念出发，将数据质量的概念和数据处理流程混合，保证用户对数据的质量有一个全面的掌控，并可以对数据质量进行优化；
- 数据 API 服务：数据中台对外输出数据服务可通过 API 完成，Data Formula 提供配置化的 API 能力，可以通过简单的配置将数据集转换为数据 API，供其他业务系统使用，同时 API 还具备权限控制和数据缓存能力，极大的提升了系统响应能力；
- 流式数据处理：针对目前比较流行的 IOT 数据场景，以及其他流式数据处理场景，Data Formula 提供了完整的流数据处理能力，方便将流式数据进行存储推送，并同步进行计算，汇入数据中台。

2 使用说明

2.1 数据总览

2.1.1 资产总览

资产总览是对系统中数据资产情况的整体聚合展示。通过资产总览模块，用户可以更加快速、直观、全面的了解到数据资产的概貌。

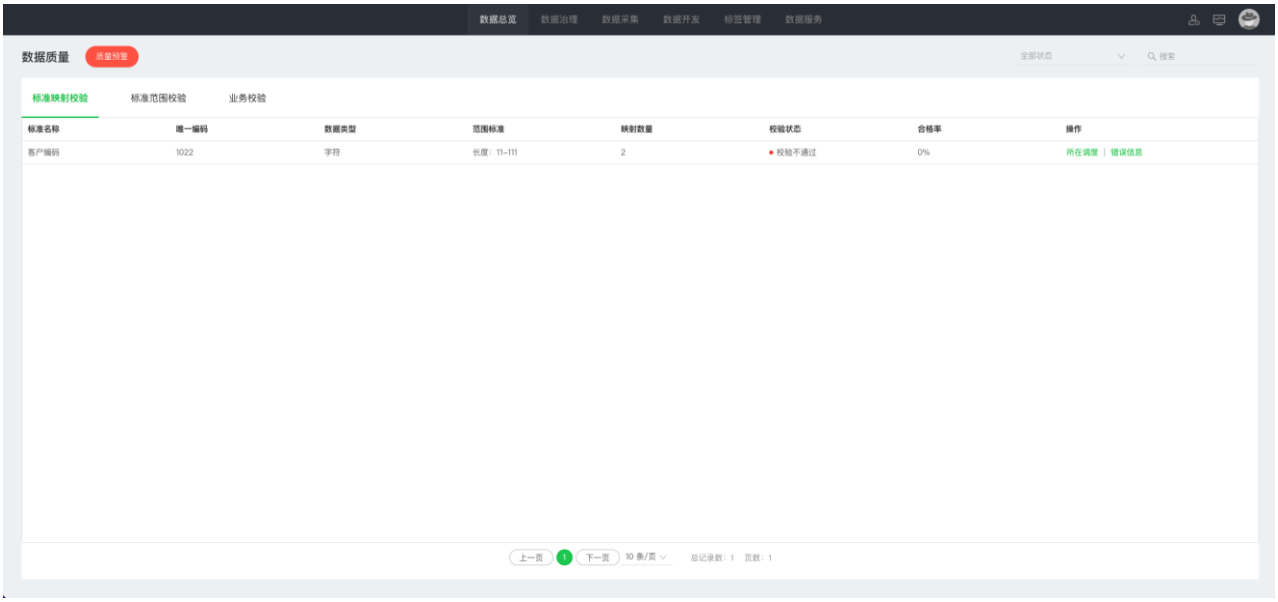
展示的内容包括：数据资产情况、数据质量情况和任务调度执行情况。

2.1.2 数据质量

因为数据的可靠性和实用性，会直接影响到统计分析的结论，所以数据的质量十分重要。而高质量数据的获得，必须通过有效的数据质量控制手段，进行数据的管理和控制，从而消除数据质量问题。

通过此模块能够实现对标准映射、标准范围和业务数据的校验，达到提升数据质量的目的。质量校验以调度任务的方式，统一在【数据总览】-【任务调度】中配置执行周期，调度

执行。此模块可以添加业务校验任务、查看标准映射校验、标准范围校验和业务校验任务的执行结果。



一、 添加业务校验

数据质量校验的类型分为：标准映射校验、标准范围校验、业务校验。

标准映射校验和标准范围校验直接在【数据总览】-【任务调度】中，配置执行周期，调度执行。

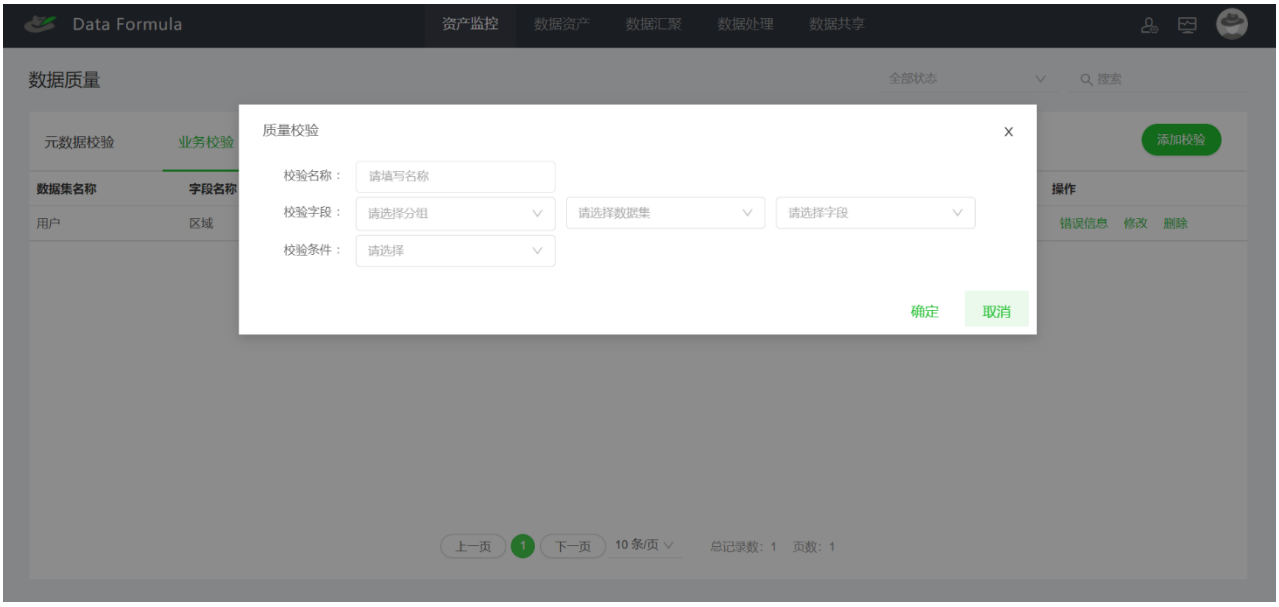
业务校验需要先添加校验，然后在【数据总览】-【任务调度】中，配置执行周期，调度执行。

业务校验添加步骤如下：

在【数据总览】-【数据质量】页面，切换到【业务校验】tab 页，点击“添加校验”，弹出添加校验设置页面，进行业务校验的添加。

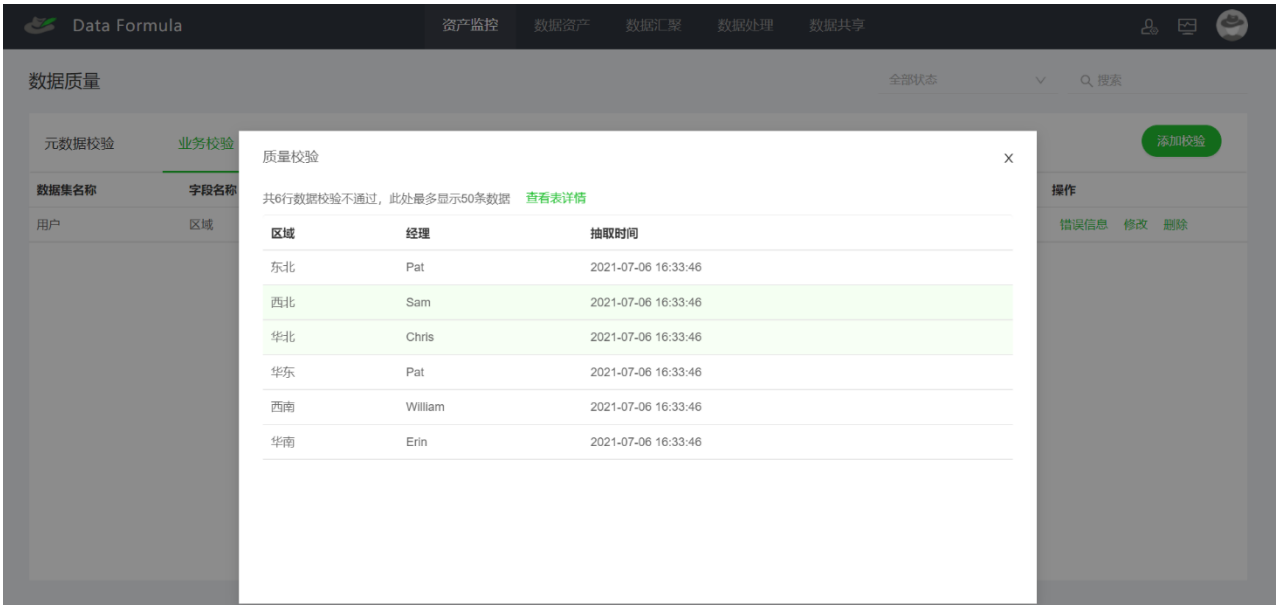
在【添加校验】页面，输入校验名称，选择校验字段和校验条件，点击“确定”，创建完成。

创建完成后，可以在【数据总览】-【任务调度】，点击“添加任务”，选择“质量任务”，找到创建好的任务，调度执行。



二、 查看校验结果

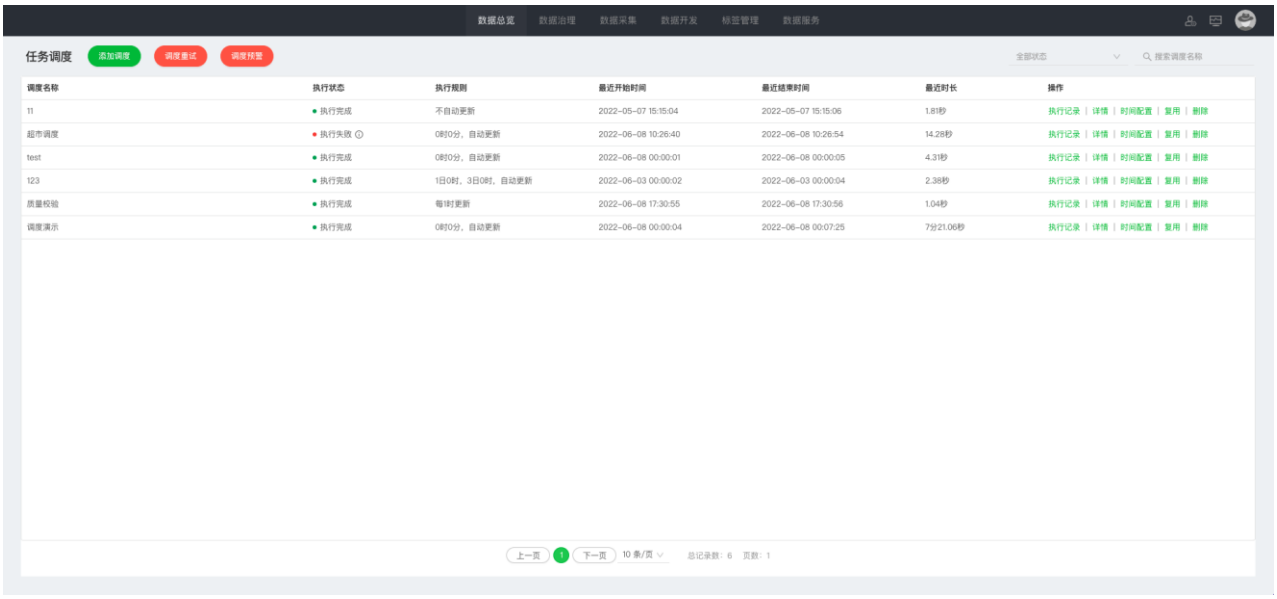
校验任务执行后的“校验状态”包括未校验、校验中、校验通过、校验不通过。如果校验不通过，可以点击“错误信息”，查看校验不通过的数据信息。



2.1.3 任务调度

作为数据处理的核心体系之一，批量式数据处理是企业中最常见的业务场景。Data Formula 中针对这种数据处理方式，提供了统一的任务调度管理功能。

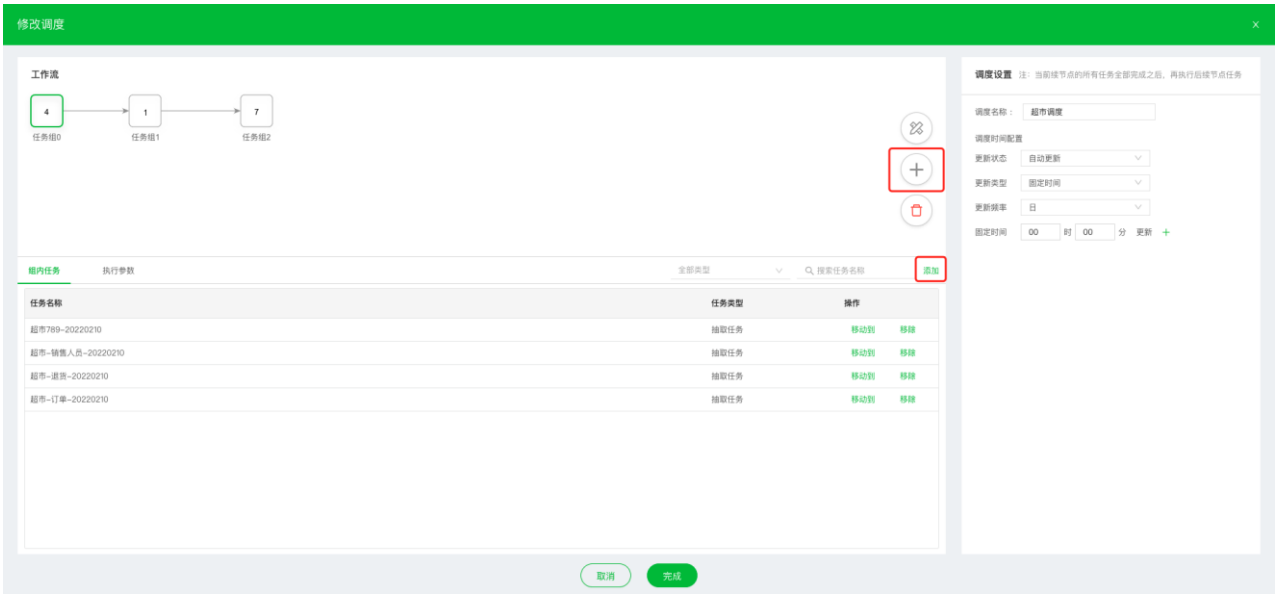
在模块中，系统用户可以看到数据中台系统所有批量数据处理的任务，包括这些任务的状态，执行历史等信息。也可以对这些任务进行操作，包括修改、执行、暂停或者取消等。



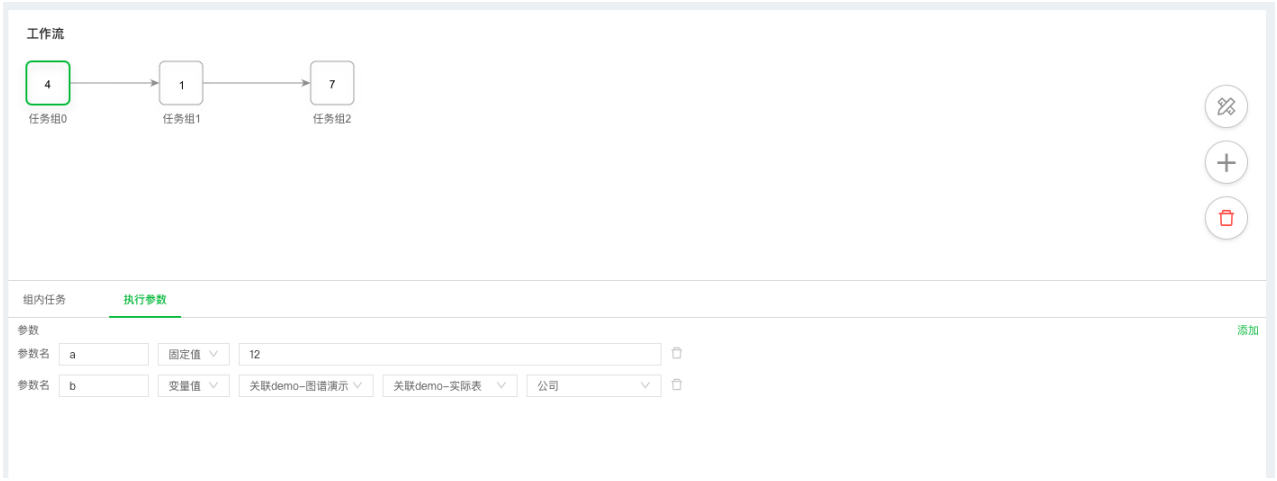
一、添加任务

点击“添加任务”，进入【添加任务】界面。【添加任务】界面分为两个展示区域：工作流设置、调度设置。

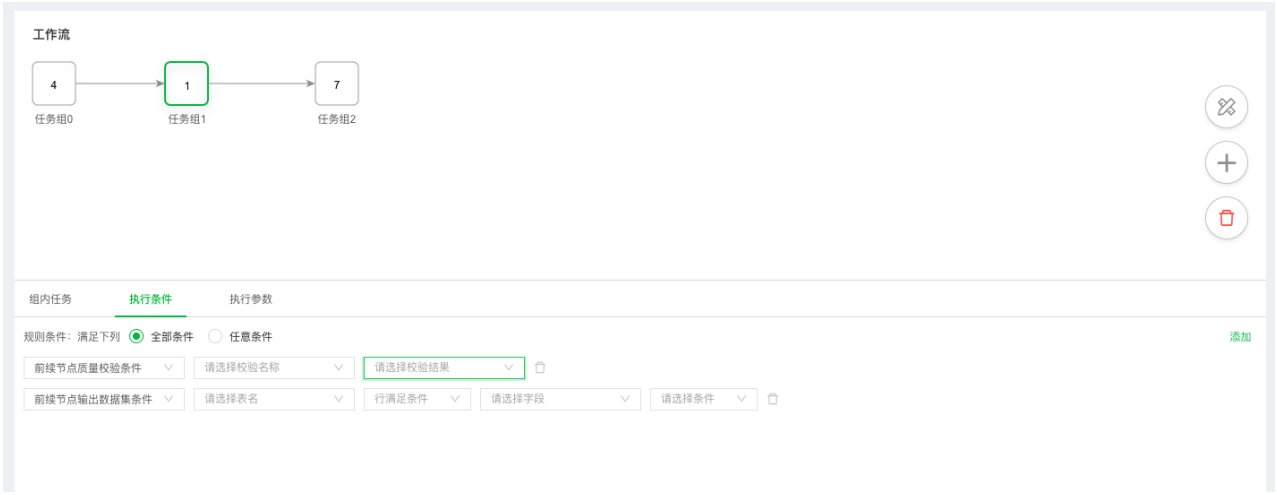
（1）工作流设置：右侧点击“+”号，添加任务组，在任务组内，点击“添加”，添加组内子任务，可以添加多组任务，多组任务会按照各任务之间的依赖关系依次执行。



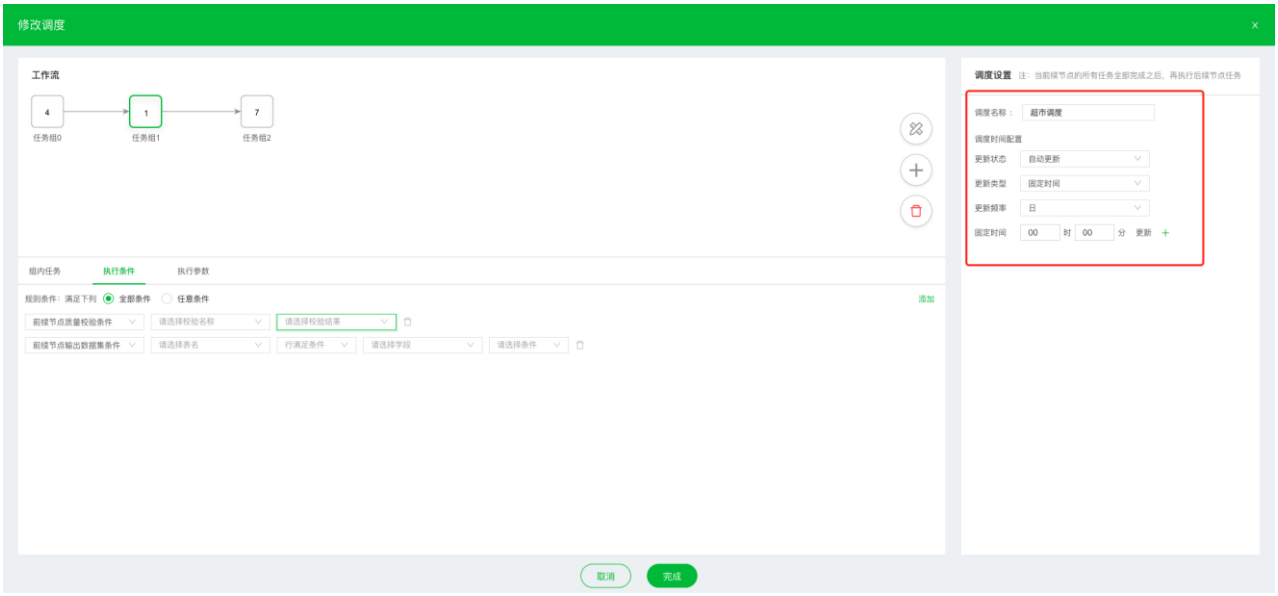
（2）执行参数：当组内开发任务有使用参数时，可在调度中设置本任务执行参数，参数支持设置固定值或变量值，变量值可从某张表的某字段获取。



（3）执行条件：任务组可设置执行条件，根据前续任务组执行完的结果来判断是否执行本任务组，执行条件支持前续节点质量校验条件（通过/不通过）和前续节点输出数据集条件。

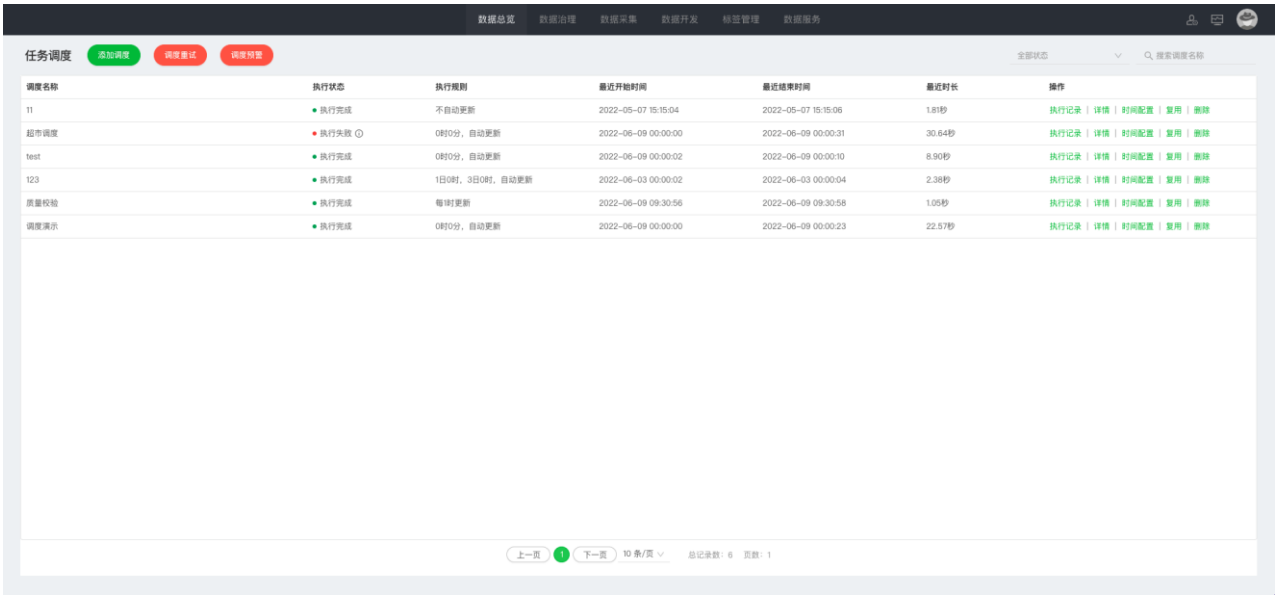


（4）调度设置：填写“调度名称”和“更新时间配置”。【更新时间配置】包括：更新状态和更新时间，更新时间分为固定时间和间隔时间两种方式。【更新状态】包括“自动更新”和“不自动更新”，若选择“自动更新”需要继续填写“更新类型”；若选择“不自动更新”，则任务设置完成。【更新类型】可选择“间隔时间”或“固定时间”。若选择【间隔时间】，则需要输入“间隔时长”；若选择【固定时间】，则需要输入“更新频率”和“固定时间”。设置完成后，点击“完成”，跳转到任务调度页面。

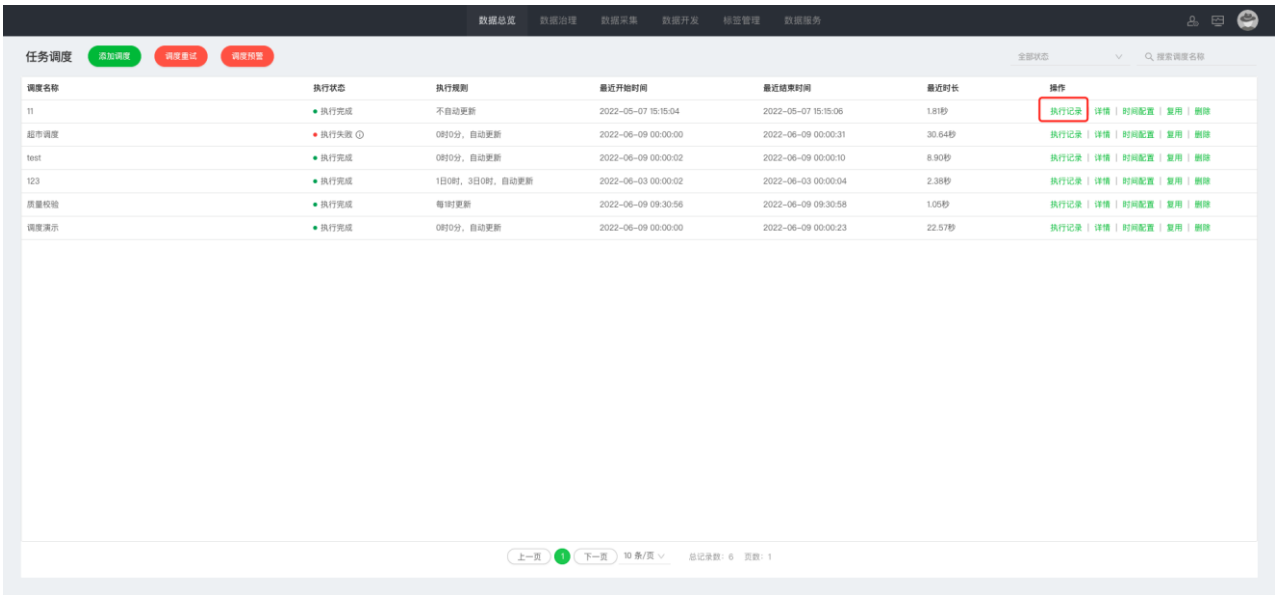


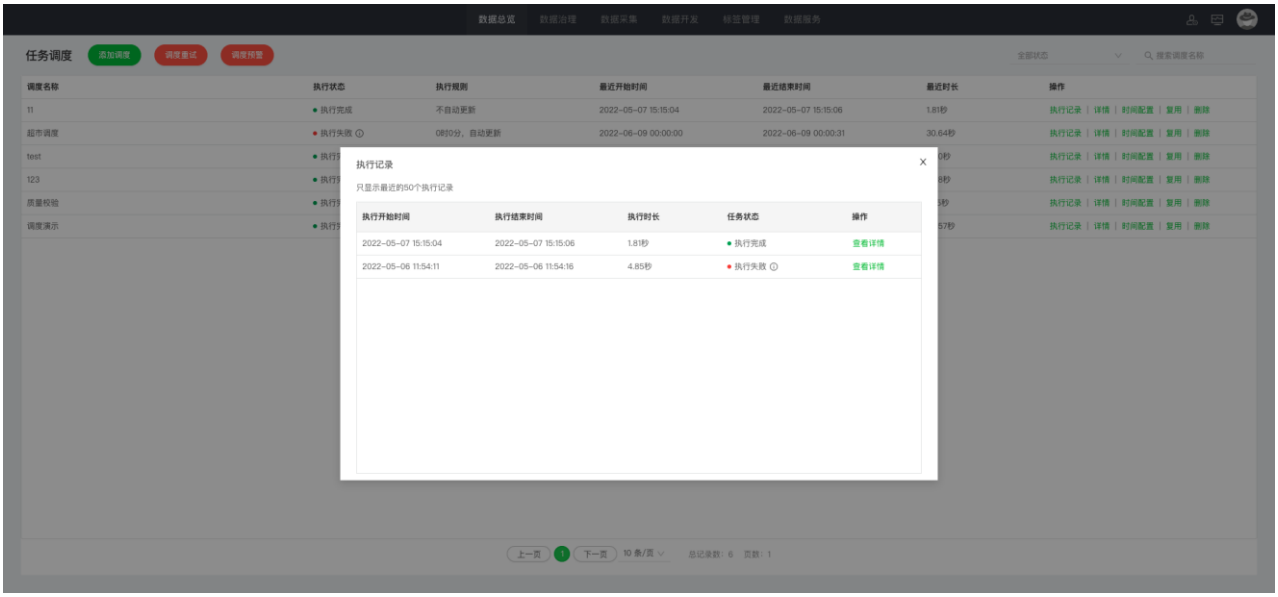
二、查看任务

（1）在任务调度页面，可以查看创建的所有任务。【任务状态】分为“初始化”、“执行中”、“执行完成/执行失败”。新创建的任务为“初始化”状态，系统根据任务设置中的配置自动执行，任务开始执行后，进入“执行中”状态。任务执行完成后，系统根据执行的结果更新任务状态。如果执行成功，则状态变更为“执行完成”；如果执行失败，则状态变更为“执行失败”。

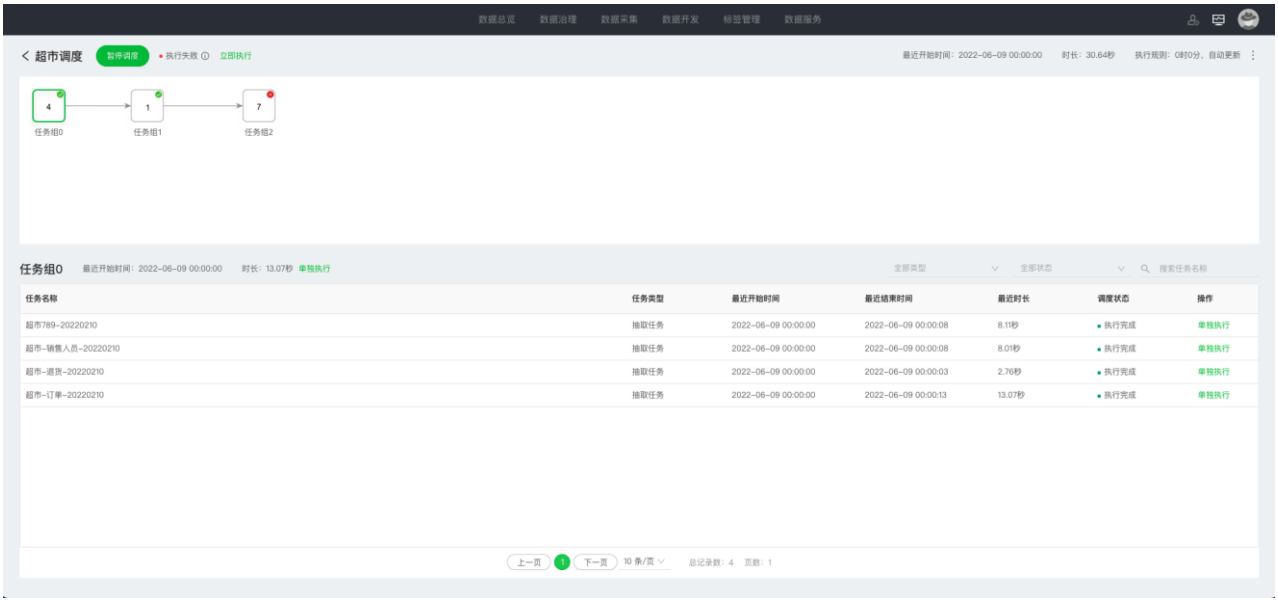


（2）点击“执行记录”可以查看该任务最近 50 次的执行情况。



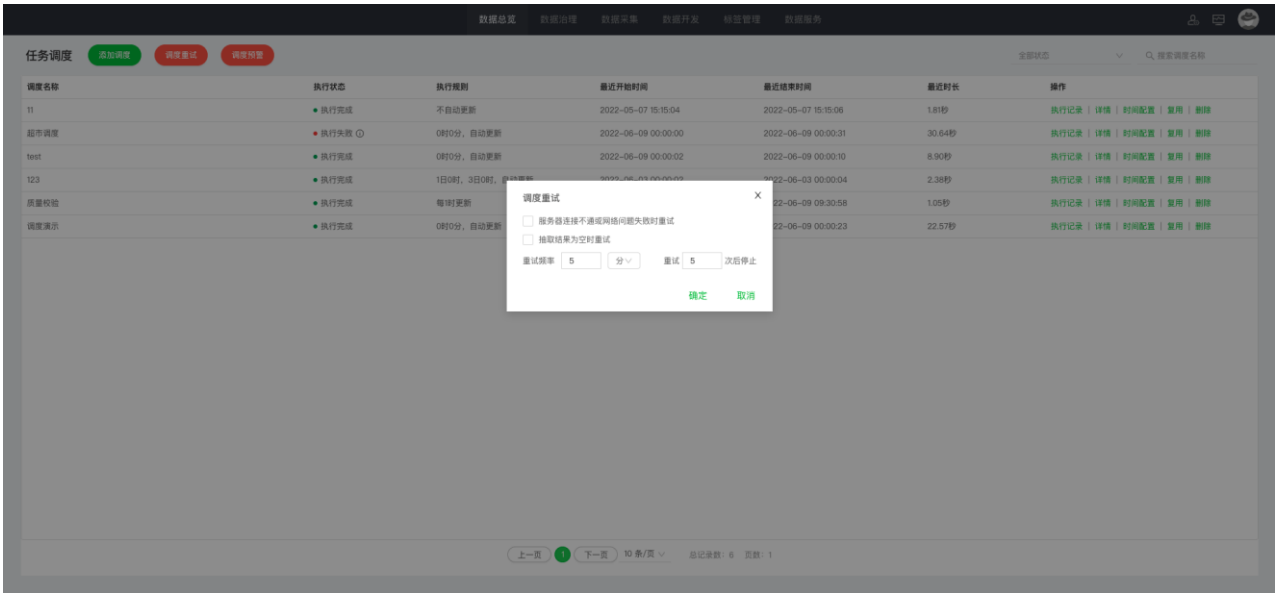


（3）调度详情，点击“详情”可查看当前调度的配置情况，可查看任务组调度流程以及各任务组中的任务，可详细查看组及子任务的执行时间和状态，可暂停/启动调度，可手动执行。



三、任务重试

因服务器连接或网络原因而执行失败的任务，和抽取结果为空的任务，系统可以根据提前设置的重试规则自动重试。若重试后执行成功，任务恢复正常，若重试几次仍未成功，任务将中止，变为任务失败状态。设置的重试规则对所有的任务生效。

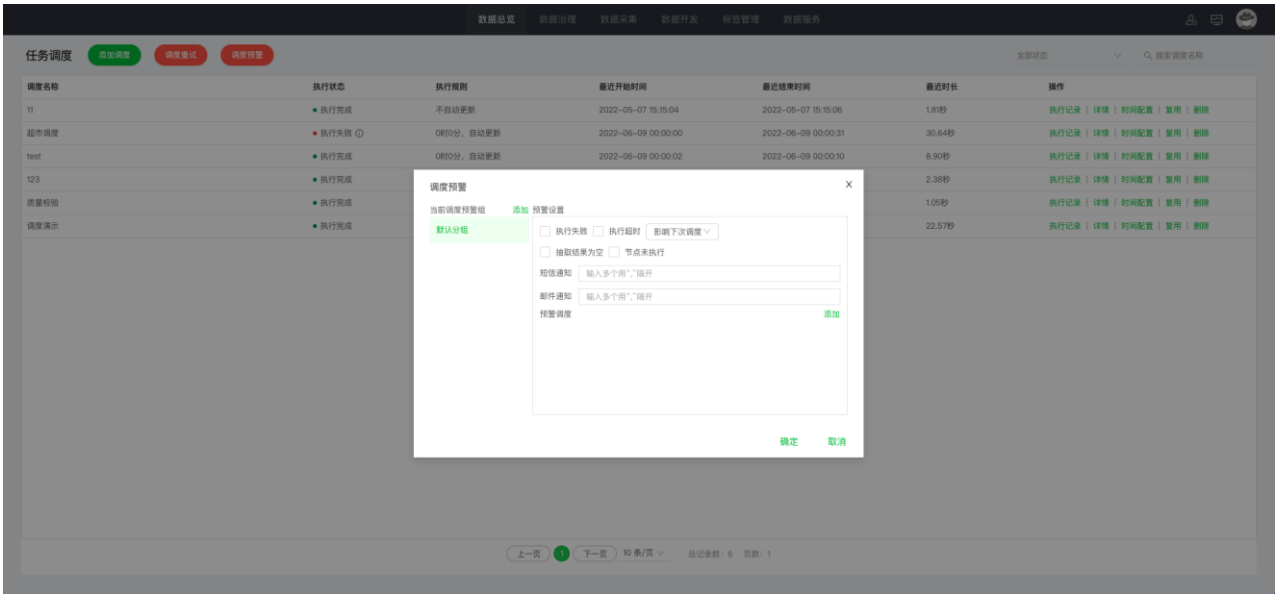


四、调度预警

调度预警是在无人值守情况下，如果任务执行失败、超时、抽取结果为空、节点未执行时，系统将根据手机号或邮箱，以短信或邮件通知的方式，向操作用户推送通知消息。

预警可根据业务情况设置不同的预警组，针对不同的调度进行不同的预警条件。

执行超时支持三种情况：影响下次调度、时长超时、时间点超时。



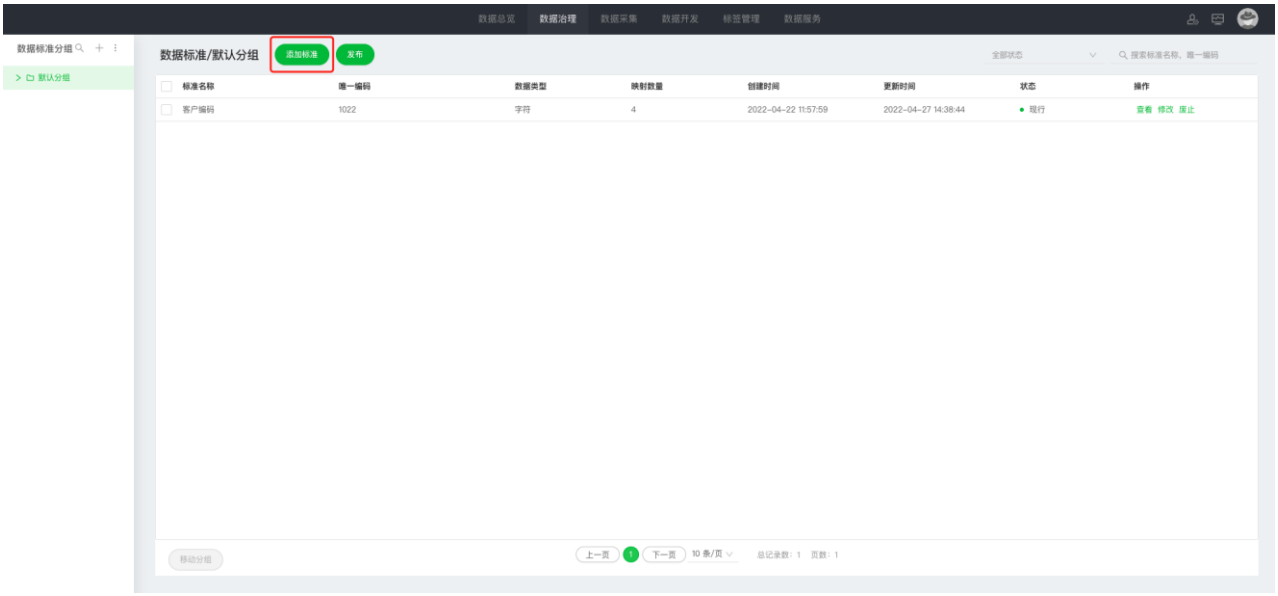
2.2 数据治理

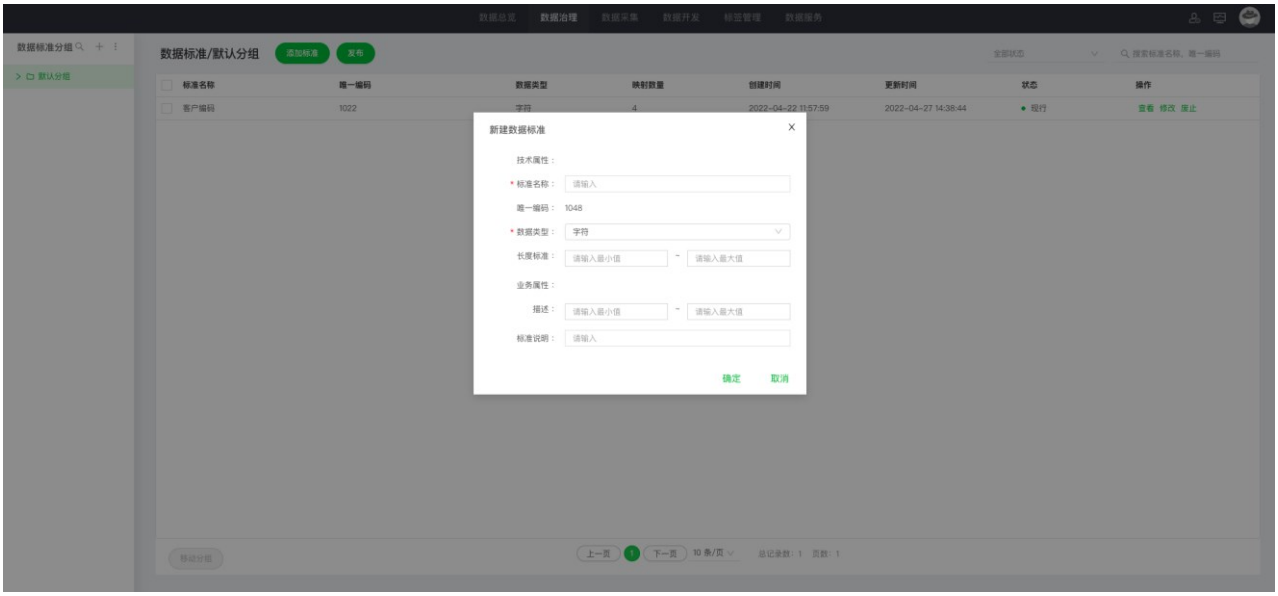
2.2.1 数据标准

数据标准是将字段进行标准货管理，以保证数据的一致性和准确性。

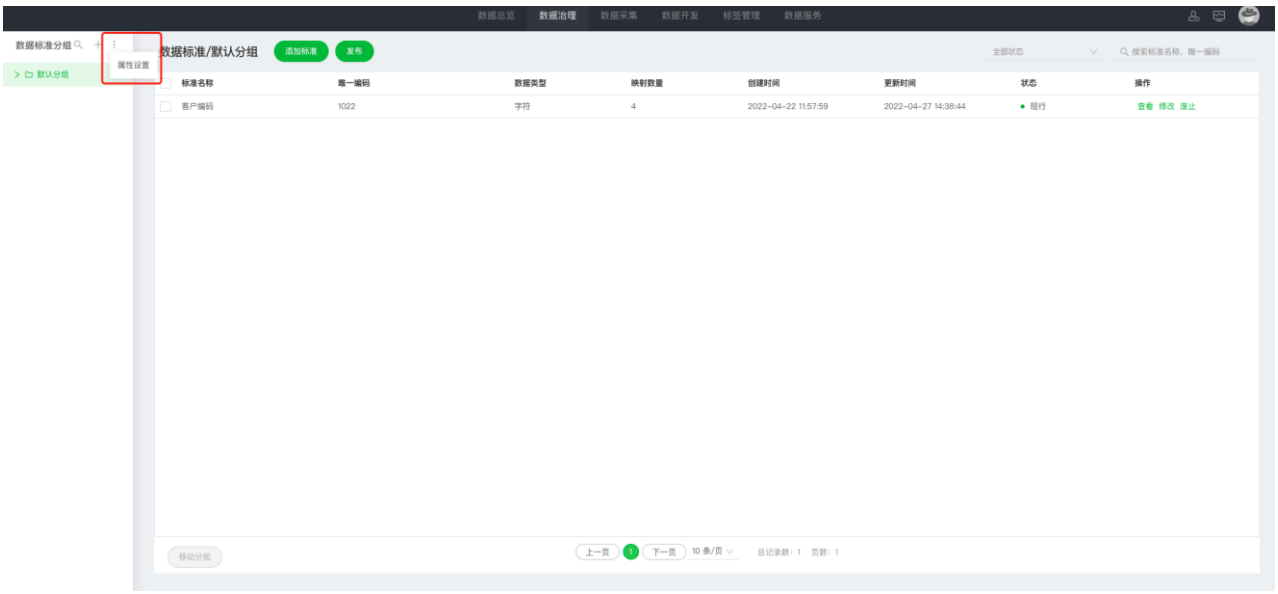
一、 添加标准

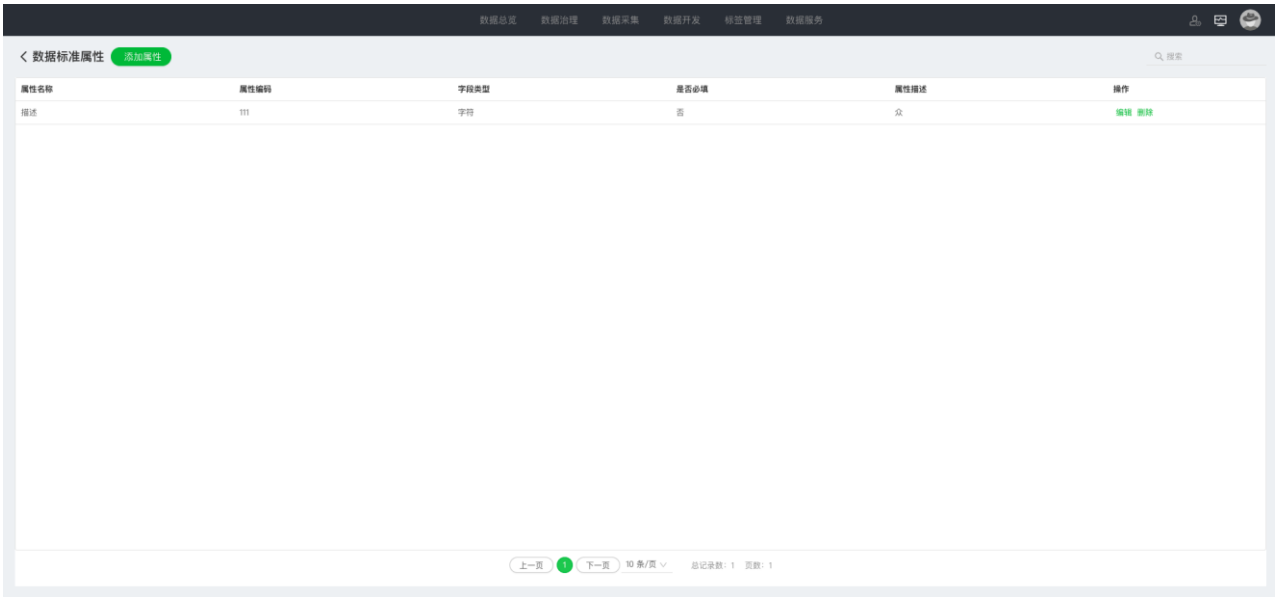
添加标准需填写标准名称、类型、描述等，添加完的标准可查看名称、编码、类型、时间、状态等信息。





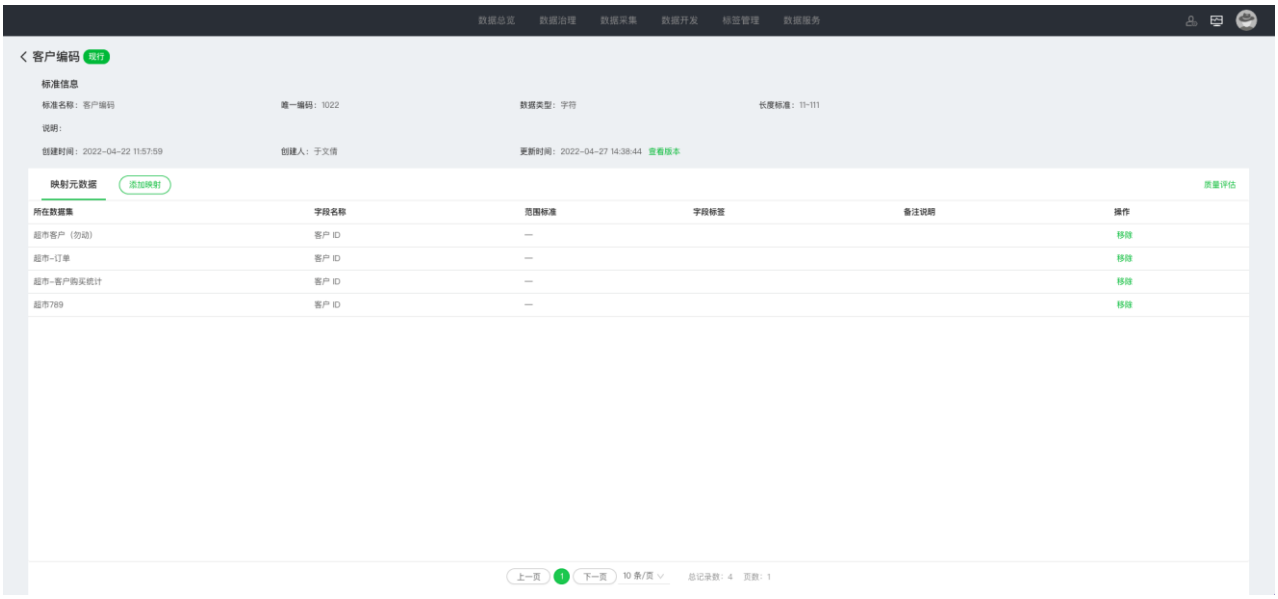
标准的业务属性，可根据企业情况自定义，通过左侧更多菜单里的“属性设置”进入页面进行设置





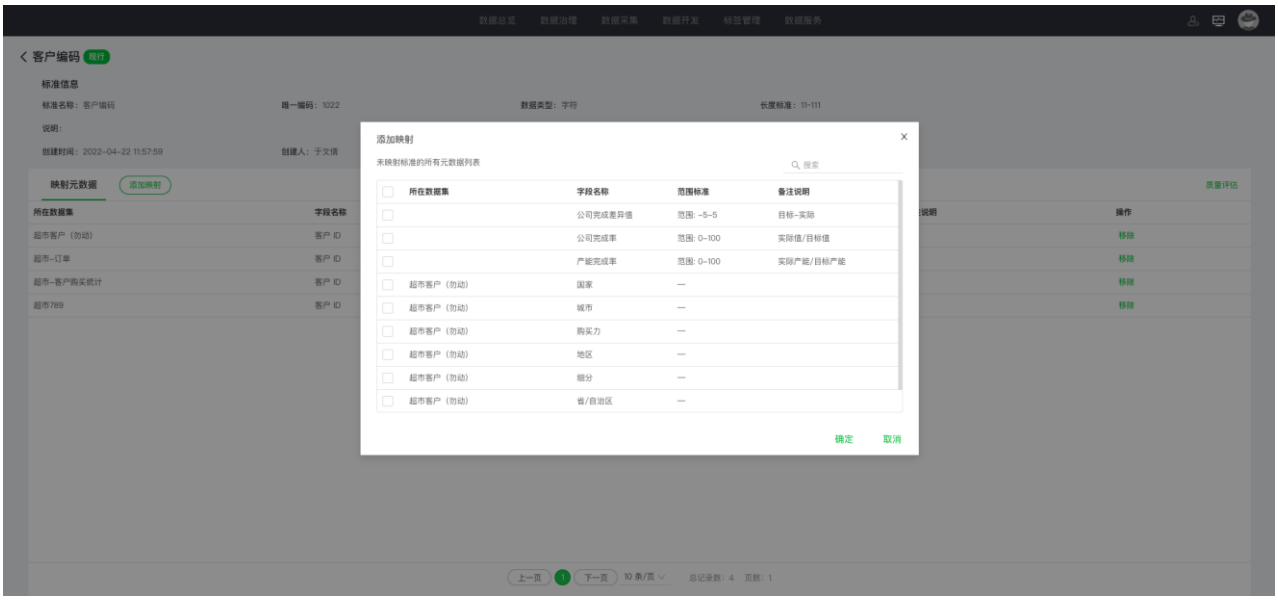
二、 查看标准

点击标准列表右侧的“查看”按钮，进入标准详情页，可查看标准基本信息、版本、映射元数据。



查看版本可查看历史发布的时间和对应的标准信息。

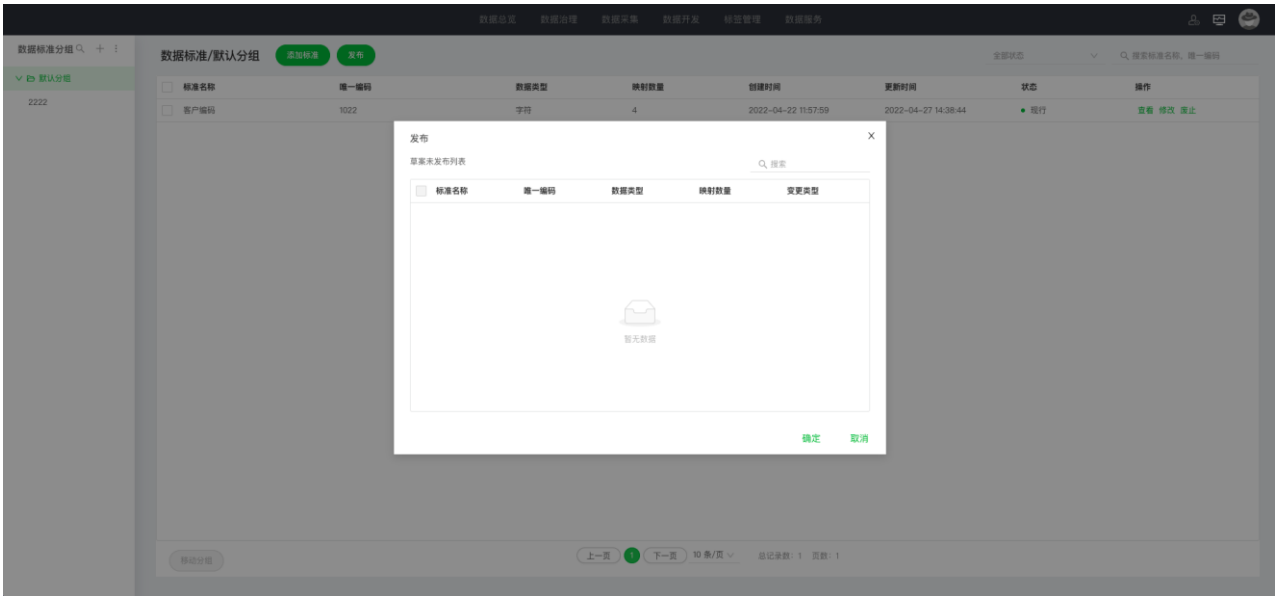
可将未映射到标准的元数据映射到当前标准，映射完的标准将在【数据质量】页面，标准映射校验进行质量校验。



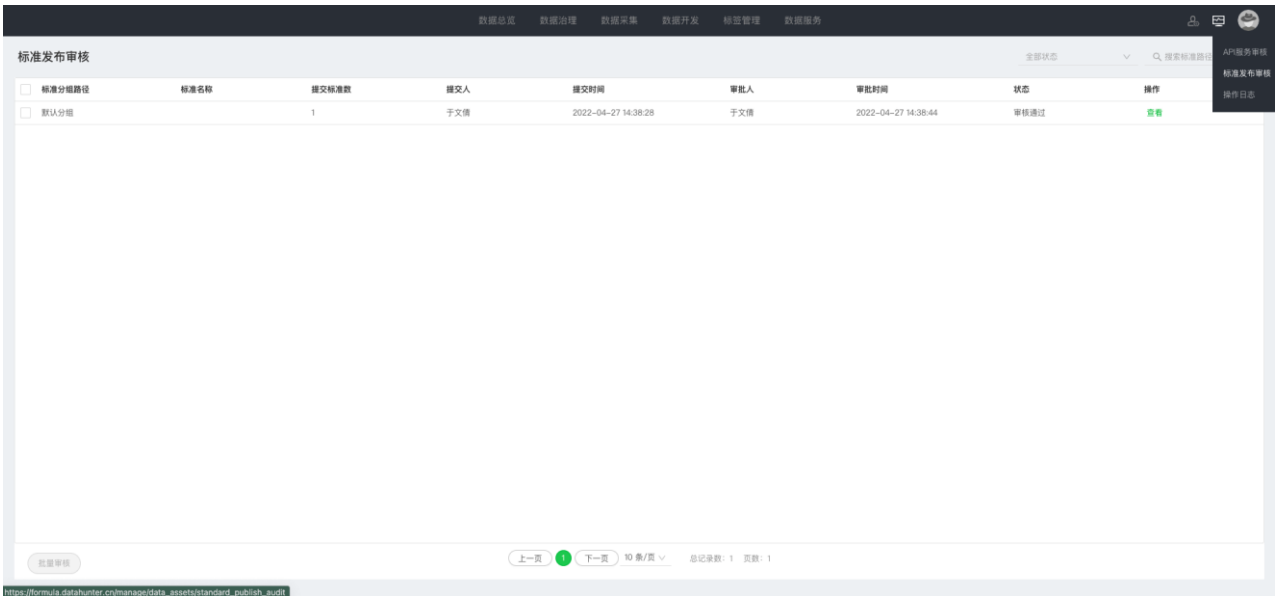
三、 标准发布审核

添加完的标准需经过发布审核才生效。

点击发布按钮，将当前分组下，草案未发布的标准勾选后发布即可。



审核人员在【标准发布审核】页面查看并审核即可生效。



2.2.2 元数据

元数据是描述数据的数据，是数据及信息资源的描述性信息。

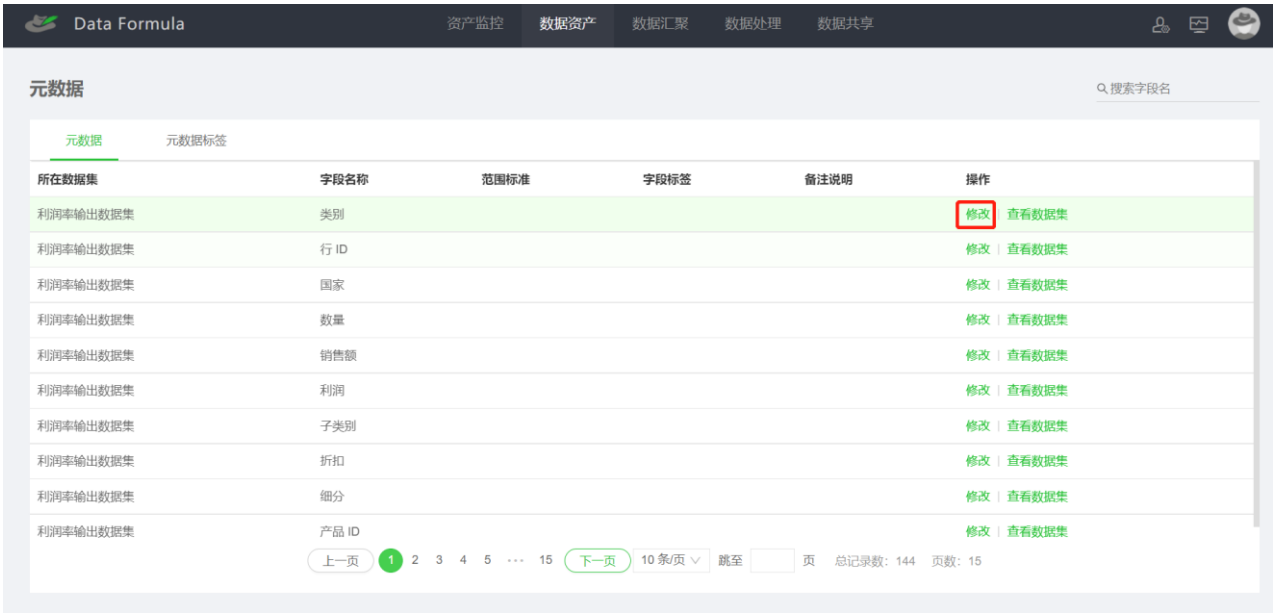
此模块，可以对系统中的所有元数据和元数据标签进行管理。元数据标签可以用来标注图谱中的字段。

四、 元数据

对元数据的操作，包括：修改、查看数据集。

（1）元数据修改

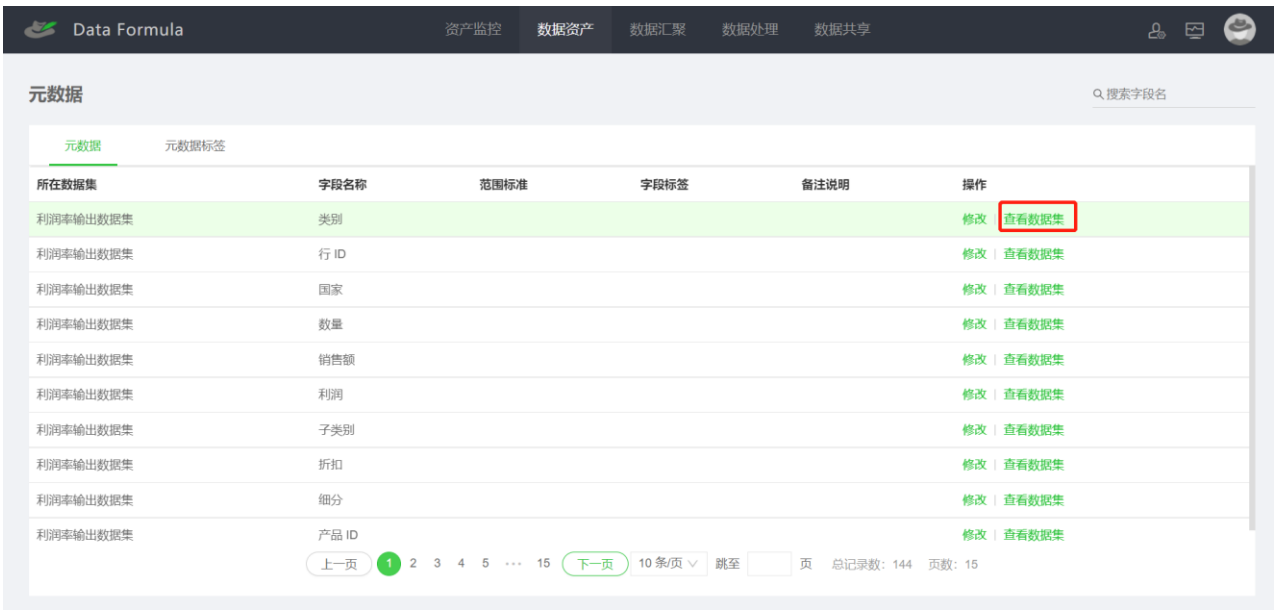
点击【修改】，弹出【修改字段】页面。



在【修改字段】页面，可以修改字段名称、字段类型、字段长度、字段标签、备注说明。点击【字段标签】后面的“+”号，可以为字段添加标签。

（2）查看数据集

点击“查看数据集”，跳转到【数据集】内容展示页面。



五、 元数据标签

对元数据标签的操作包括：添加、修改、删除。

（1） 添加标签

在【元数据标签】页面，点击“添加标签”，弹出【添加标签】页面。

在【添加标签】页面，可以输入标签名称，选择标签颜色，然后点击“确定”，创建该标签。

（2）修改标签

在【元数据标签】页面中单个标签的后面，点击“修改”，弹出【编辑标签】页面。

在【编辑标签】页面，可以修改标签名称，重新选择标签颜色，然后点击“确定”，对该标签进行修改。

（3）删除标签

在【元数据标签】页面中单个标签的后面，点击“删除”，可以删除该标签。

2.2.3 数据图谱

数据图谱代表着数据的血缘关系，通过数据集与数据集之间的转换关系进行描述，每一个数据变迁过程，都可以追溯到具体的字段以及变化内容。Data Formula 系统中的数据指标字段可以采用数据图谱的方式来展示，这样能够更好的体现出指标字段的由来。

模块支持对数据图谱进行创建、查看、编辑、删除操作。

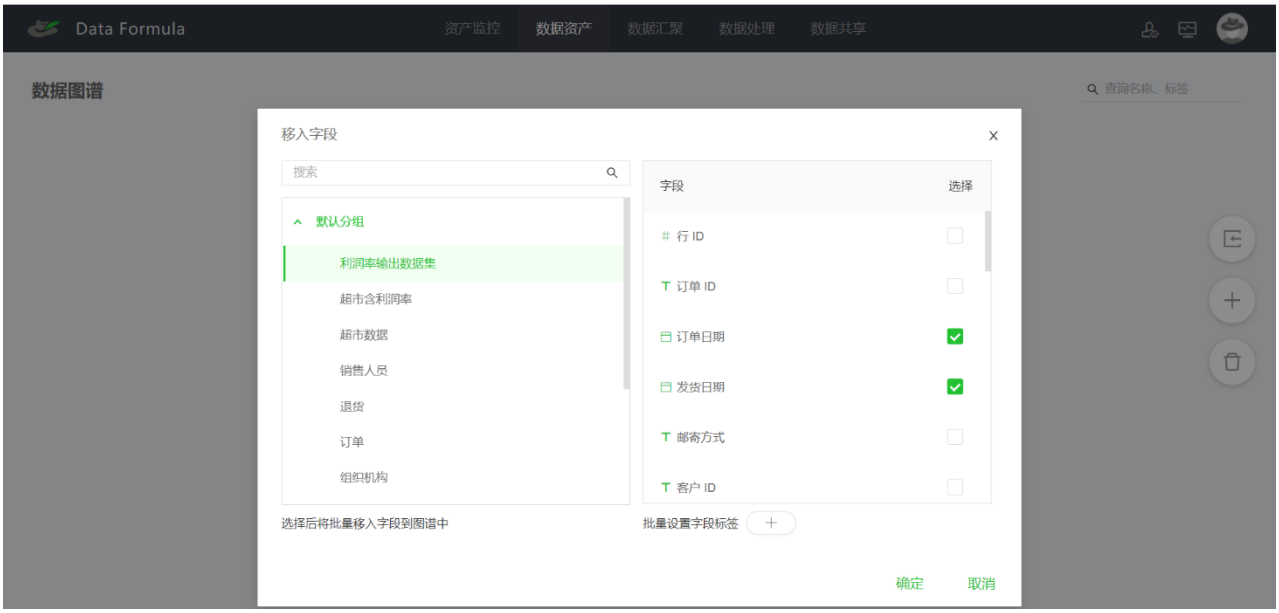
一、创建图谱字段

数据图谱的创建，支持【导入现有字段】和【添加新字段】两种方式。

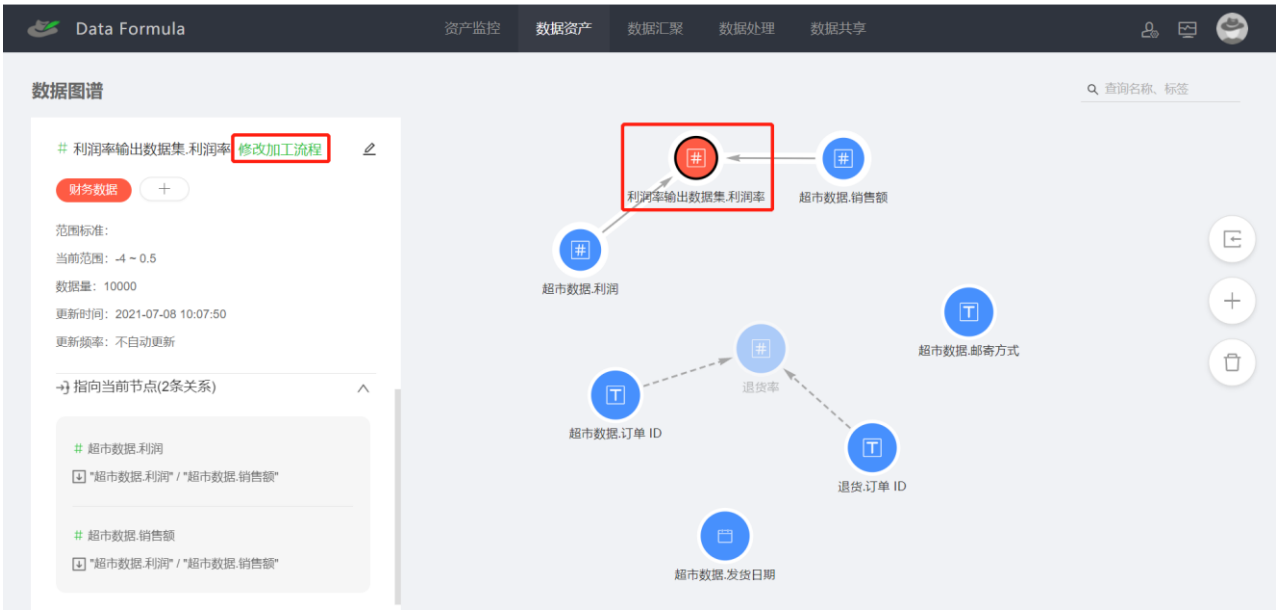
（1）导入现有字段

点击“导入”符号，弹出【移入字段】页面。

在【移入字段】页面，在页面左侧选择数据集所在分组，然后选中数据表。页面右侧显示出所选数据集中的所有字段，此处支持一次选中多个字段，然后点击“确认”，完成字段导入。在导入字段时，可以点击“+”号给字段设置标签。



导入字段成功后，选中图谱中该字段对应的节点，弹出【字段详细信息】弹窗，弹窗中出现“修改加工流程”按钮，点击“修改加工流程”，跳转到【数据加工】页面，可以在数据加工流程中对该字段进行修改加工。



（2）添加新字段

点击“+”符号，弹出【添加新字段】页面。

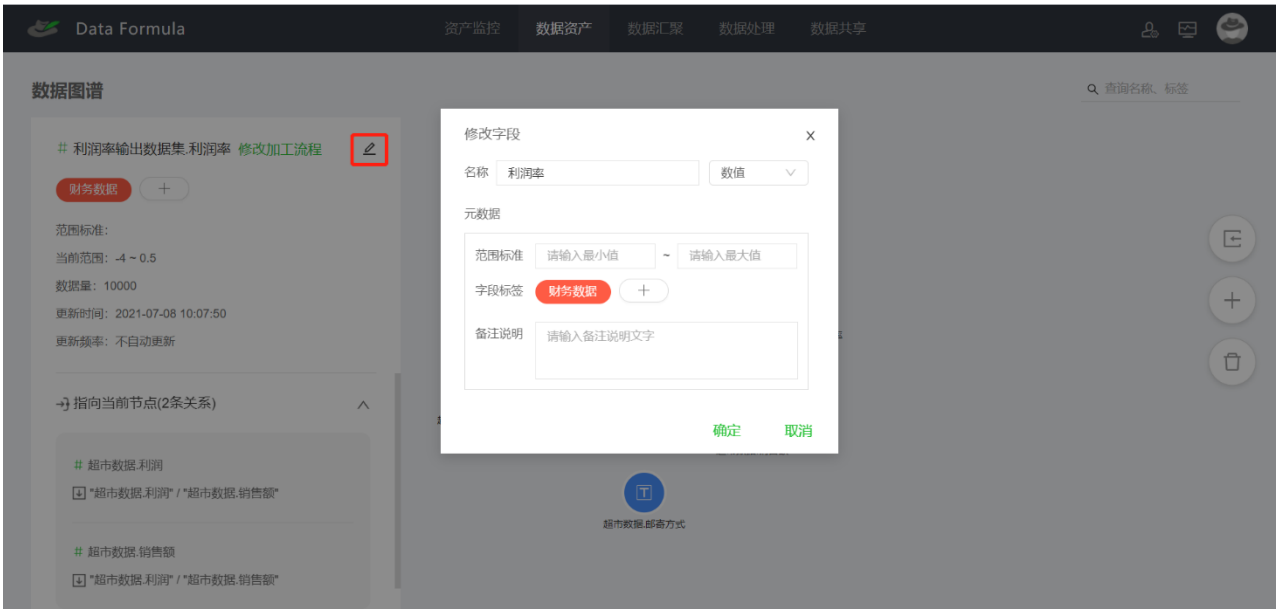
在【添加新字段】页面，需要输入字段信息和元数据信息。字段信息包括：字段名称、字段类型。

元数据信息包括：范围标准、字段标签、备注说明、计算关系。【字段标签】是为该字段打标签，点击“+”号，可以选择提前创建好的元数据标签进行打标签操作。【计算关系】是输入新字段的计算公式。点击右下角“问号”图标，弹出【计算管理】的使用帮助。当需要在【计算关系】中引用字段时，可直接输入字段名称，系统会进行实时搜索，并显示出匹配的字段，在列表中选择对应字段即可。【计算关系】的填写格式示例："超市数据.利润" / "超市数据.销售额"（运算符号的前后需要留出空格）。

新字段添加成功后，选中图谱中该字段对应的节点，弹出【字段详细信息】弹窗，弹窗中出现“加工生成”按钮，点击“加工生成”，跳转到【离线开发】页面，可以在数据开发流程中加工生成该字段。

二、编辑图谱字段

选中图谱中的节点后弹出【字段详细信息】弹窗，点击“编辑”图标，跳转到【修改字段】页面，可以进行图谱字段的修改。



三、 删除图谱字段

选中图谱中的节点后，点击右侧的“删除”图标，可以删除选中的图谱字段。

2.2.4 数据模型

数据模型是数据表之间的关系模型。模块从业务数据应用场景出发，采用业务域的方式，对数据模型进行管理，用户可以直观的通过业务模型视角，观察数据集之间的关系，也可以对数据模型进行增删改查操作。

一、 添加模型组

（1）新增模型组

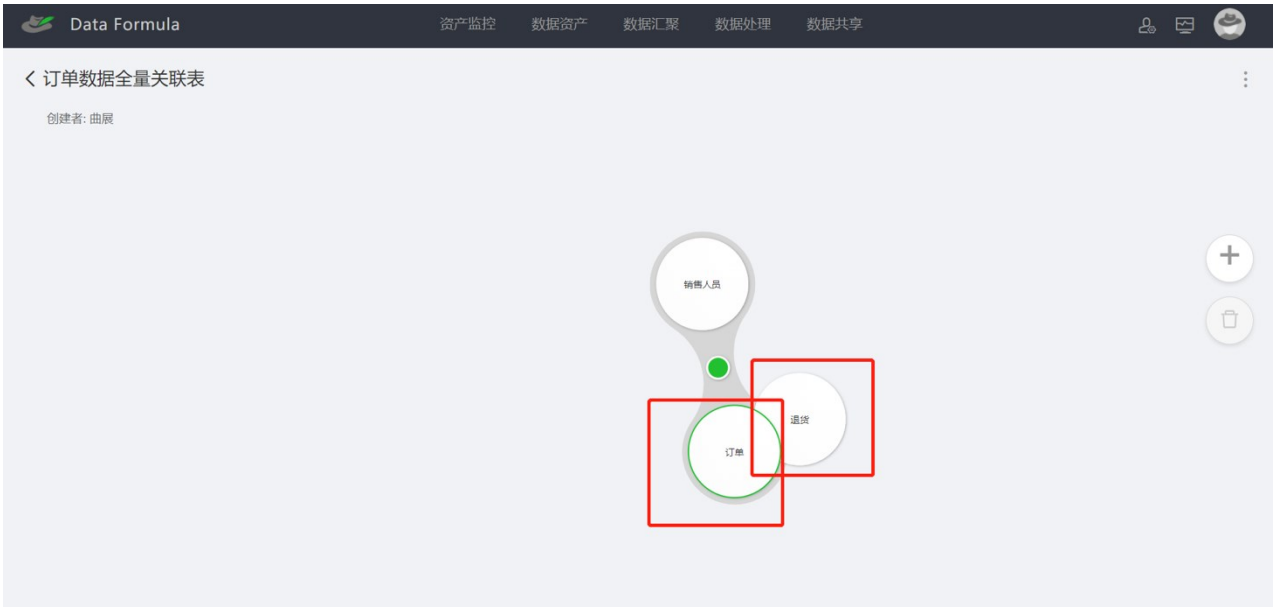
点击右侧的“+”号，弹出【添加模型组】弹窗。

在【增加模型组】弹窗页面，填写“模型组名称”，点击“确认”，生成模型组。

（2）配置模型组

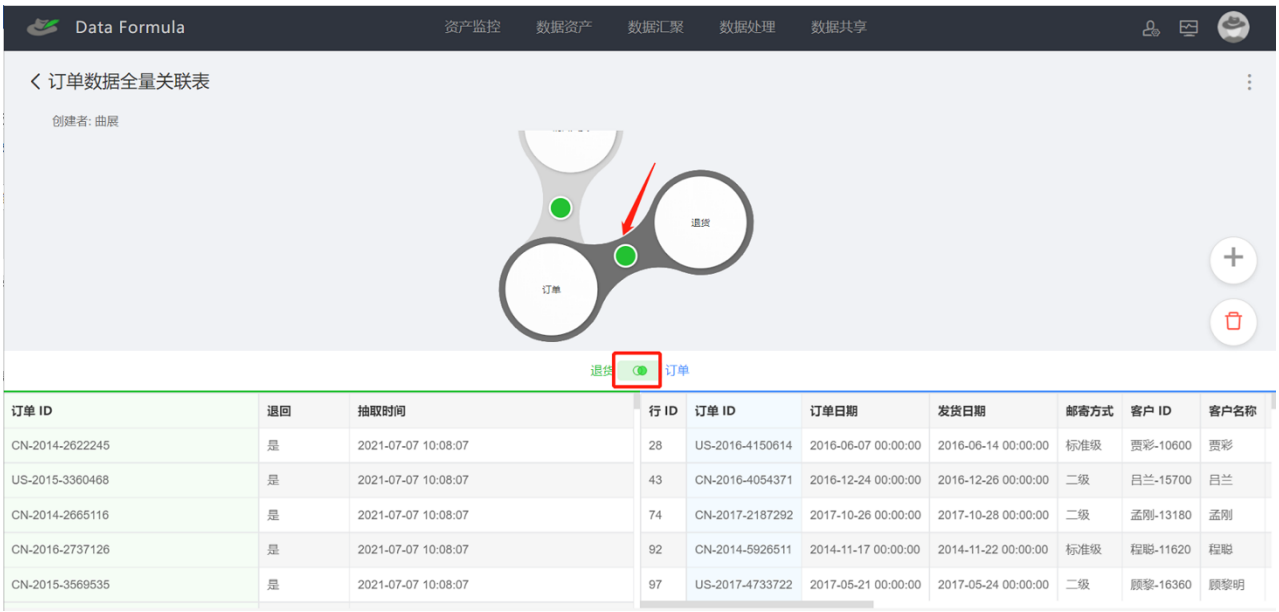
在【数据模型】页面，点击模型组，跳转到【模型组详情】页面。

在【模型组详情】页面，点击右侧“+”号，可以添加数据集，然后可以通过鼠标拖拽数据集，建立数据集之间的关联关系。若将两个数据集拖拽到一起时，页面弹出【表关联配置】弹窗，需要选择关联字段来建立关联关系。



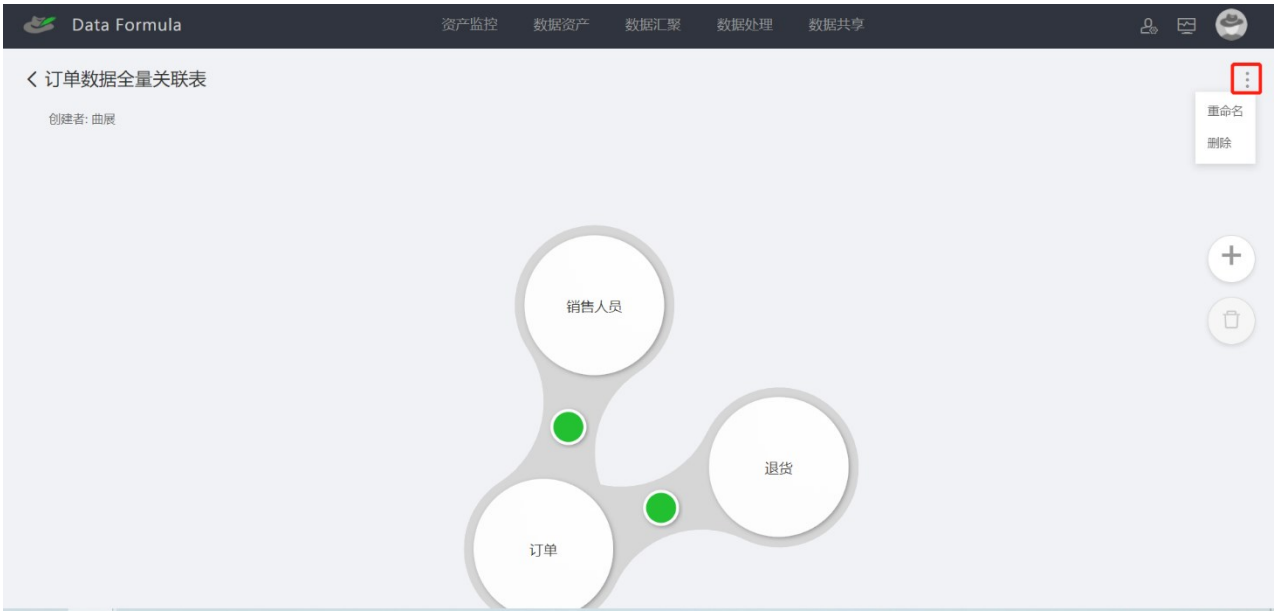
创建好的关联如下图所示：

点击关联图表，页面会弹出两个关联表的关联关系和所有字段信息。可以点击“关联关系”小图标，修改关联关系和关联类型。关联类型包括：全关联、内关联、左关联、右关联。



二、 删除模型组

进入单个模型组后，在右侧点击“更多”，然后点击“删除”，可以删除模型组。



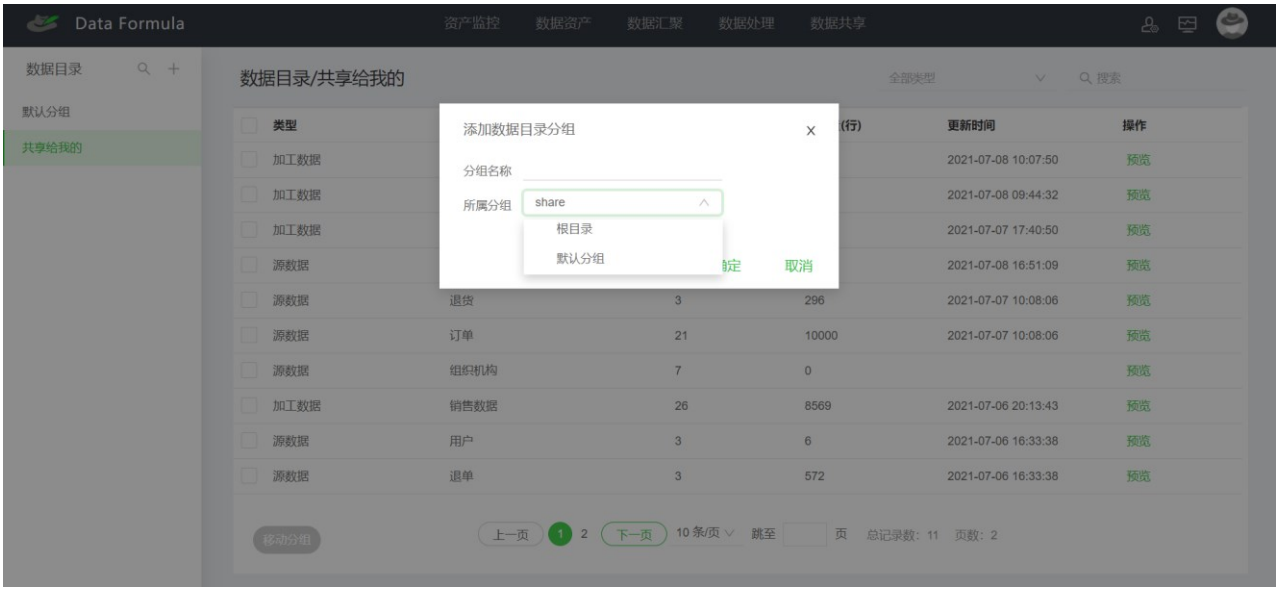
2.2.5 资产目录

在此模块中，可以通过数据目录定位到数据集。针对每一个数据集（归属于模型或离散的），用户都可以查看浏览该数据集的基本信息，血缘关系，以及对应的数据质量。用户也可以针对该数据集进行进一步的后续操作，进行数据转换加工，纳入某一特定业务模型或将

此数据集共享至数据服务。

一、 目录分组

点击左侧树形菜单，可以选择不同的分组。点击左侧的“+”按钮，可以创建新的分组。



选中一数据集后，可以点击左下方的【移动分组】，来调整数据集的分组。

二、数据集操作

在单个分组中，可以对数据集进行搜索、预览、删除操作。在单个数据集上点击右侧的“预览”，可以查看该数据集的详情。

2.3 数据采集

2.3.1 数据抽取

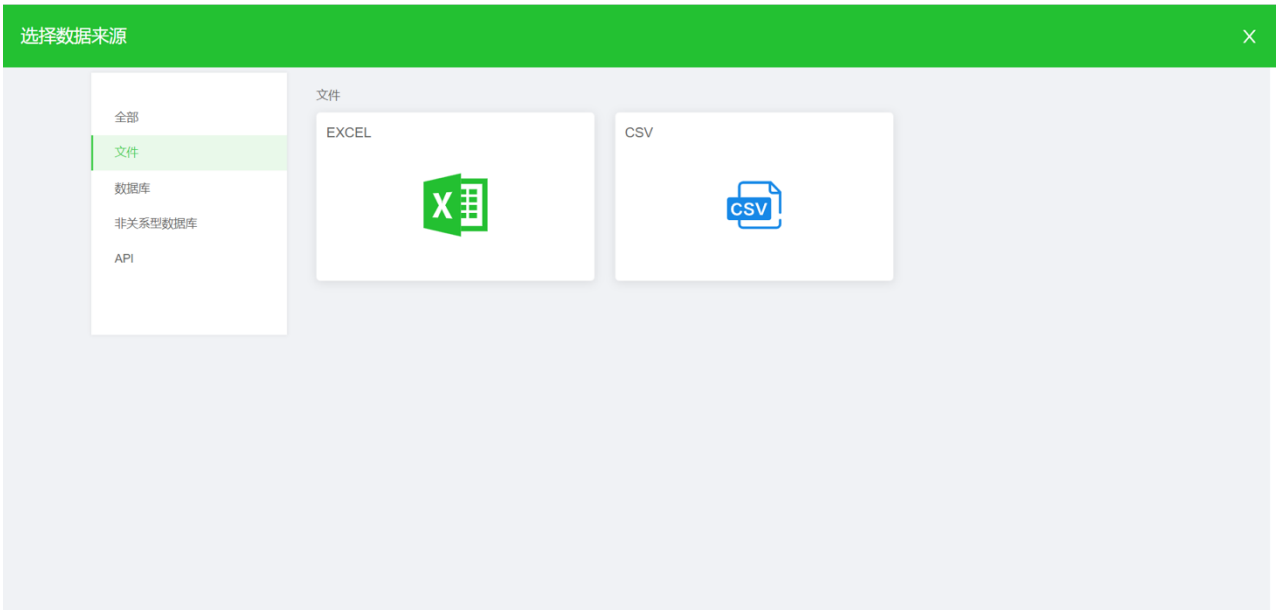
模块支持对多种类型的数据源进行数据抽取。数据源的类型包括 :文件、关系型数据库、非关系型数据库和 API。

一、文件类型数据源的数据抽取

系统可以支持的文件类型的数据源包括 Excel、CSV。

（1）选择数据源类型

点击“添加抽取任务”，跳转到【选择数据来源】页面，点击左侧菜单中的【文件】，右侧显示出可以支持的文件数据源类型。



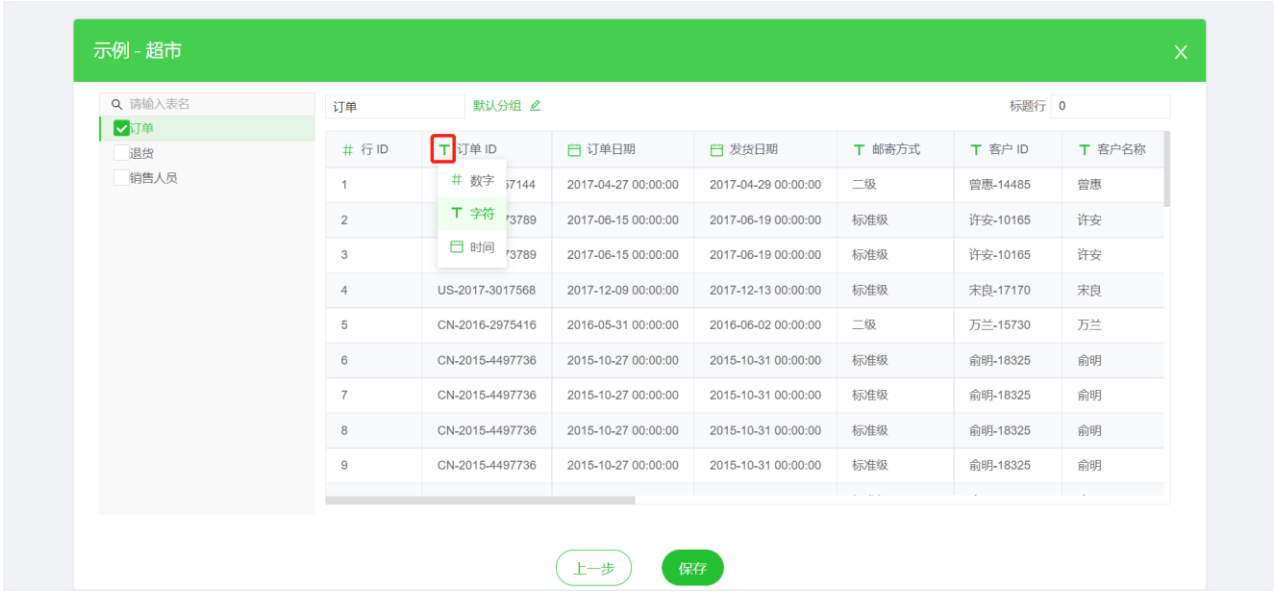
（2）上传文件

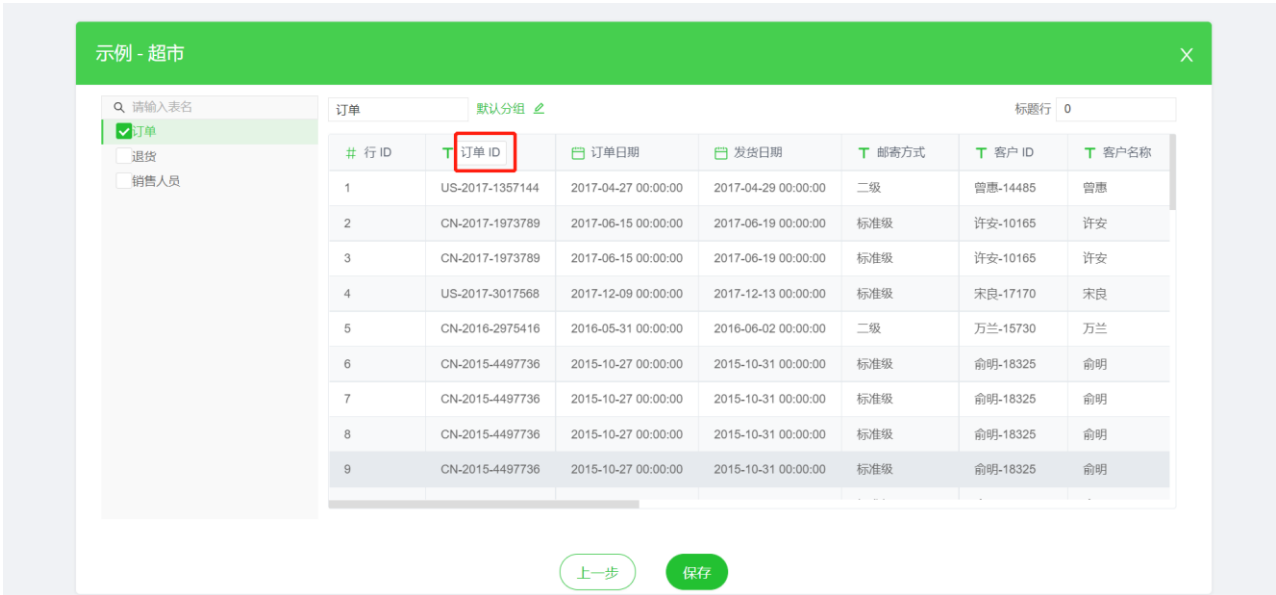
在文件数据源类型中，再点击选中的文件数据源类型，弹出文件上传页面。

（3）数据预览

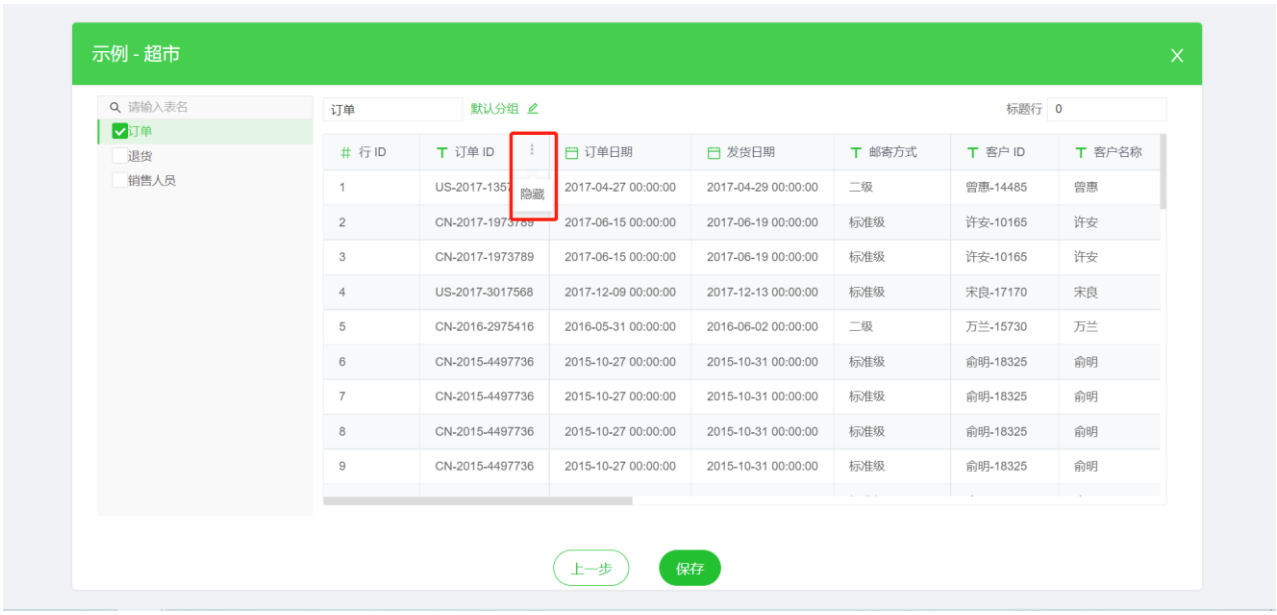
文件上传成功后，系统自动跳转到【数据预览】页面，在左侧数据表列表中选择希望抽

取的数据表，针对表字段，可以进行【更改字段类型】、【修改字段名称】、【隐藏/显示字段】的操作。



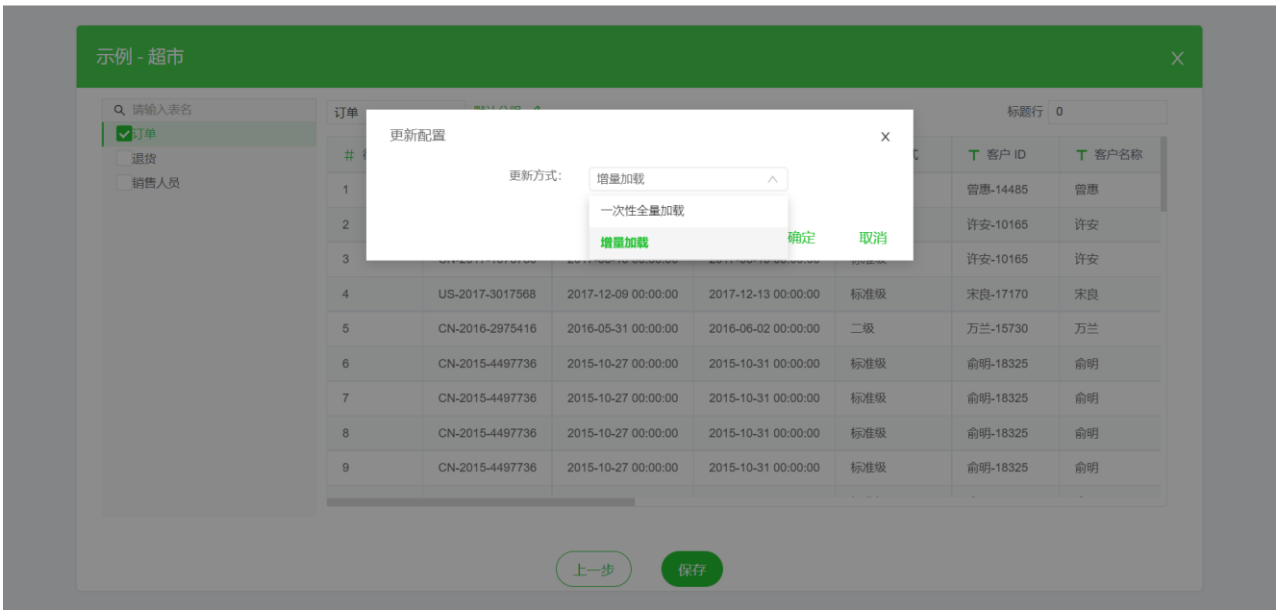


不希望抽取的字段可以进行【隐藏】。



（４）提交保存

点击“保存”，系统弹出更新设置弹窗，用户可以选择数据更新方式，更新方式包括：一次性全量加载、增量加载。



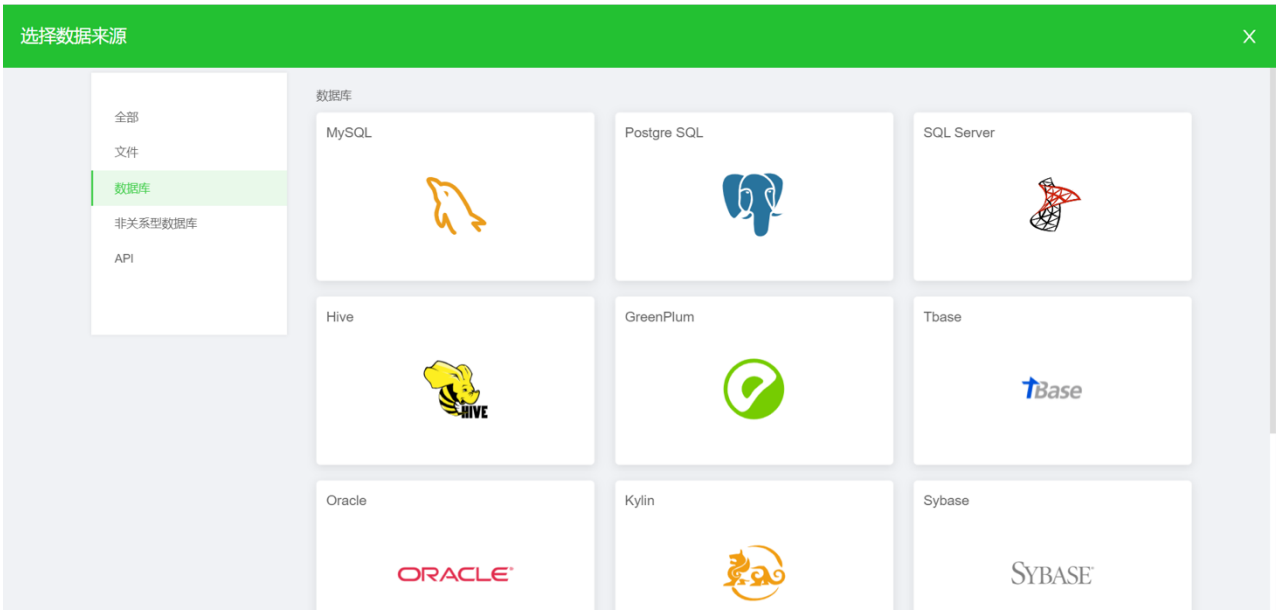
二、数据库类型数据源的数据抽取

Data Formula 系统可以支持的关系型数据库的类型包括：MySQL、Postgre SQL、SQL Server、Hive、GreenPlum、Tbase、Oracle、Kylin、Sybase、Vertica、DaMeng。

系统可以支持的非关系型数据库的类型包括：MongoDB、Elasticsearch。

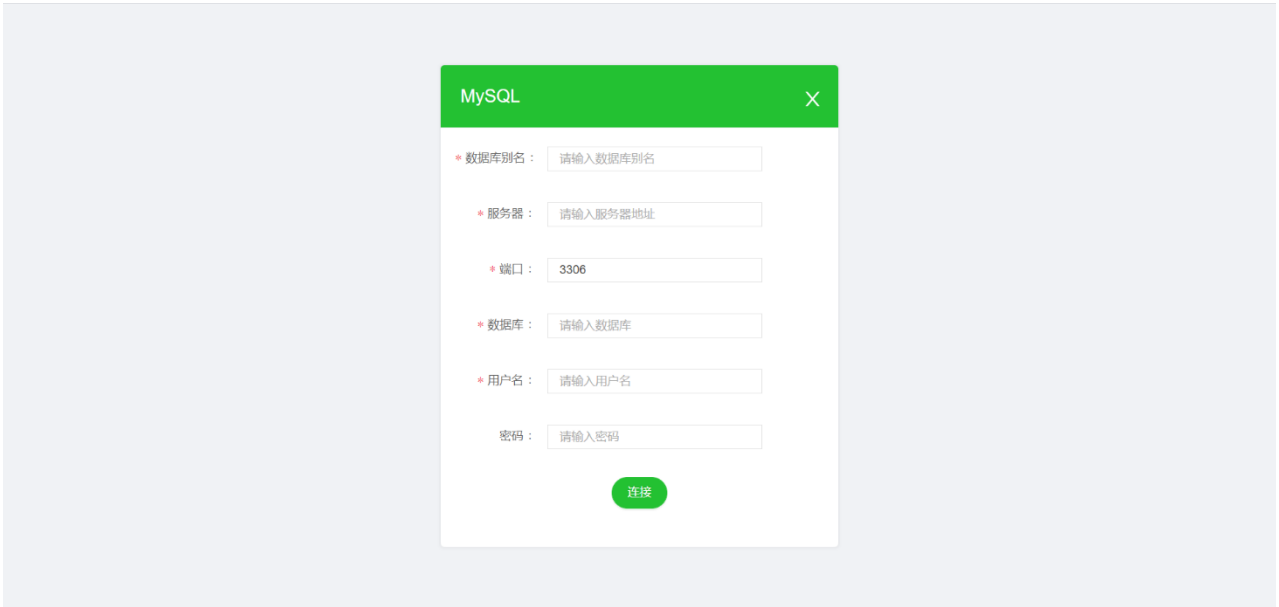
（1）选择数据源类型

点击“添加抽取任务”，跳转到【选择数据来源】页面，点击左侧菜单中的【数据库】/【非关系型数据库】，右侧显示出可以支持的数据库数据源类型。



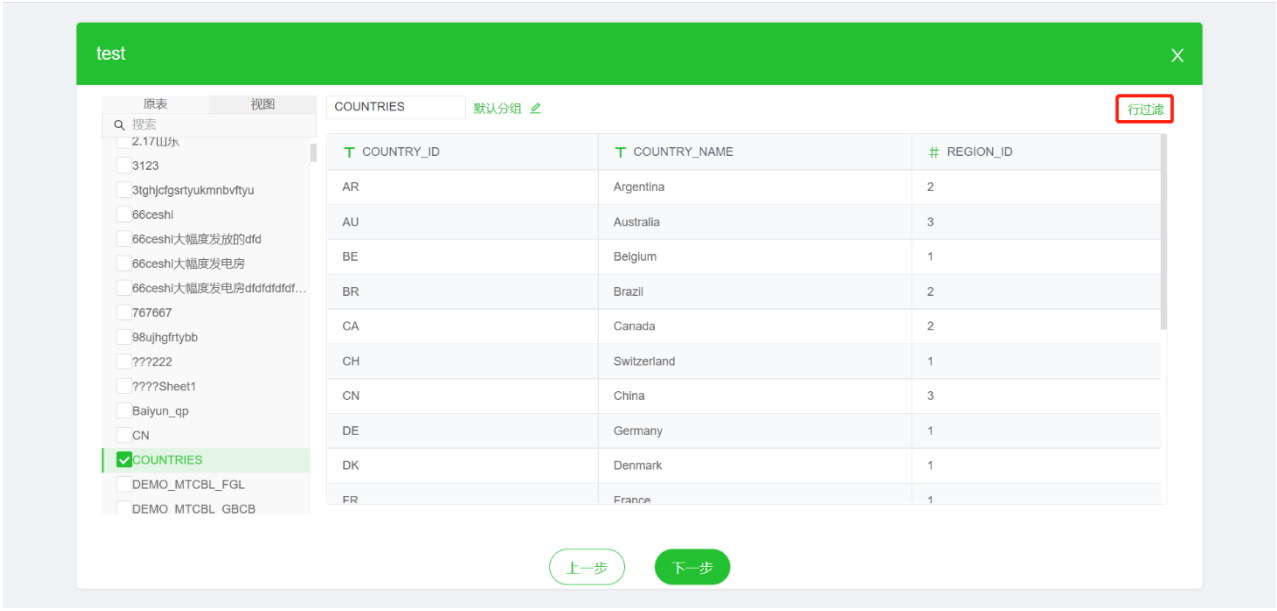
（2）连接数据库

以 MySQL 数据库为例，点击【MySQL 数据库图标】，弹出数据库链接页面。按照页面提示，数据库别名、服务器、端口、数据库名称、用户名、密码，点击“链接”，完成数据的链接操作。

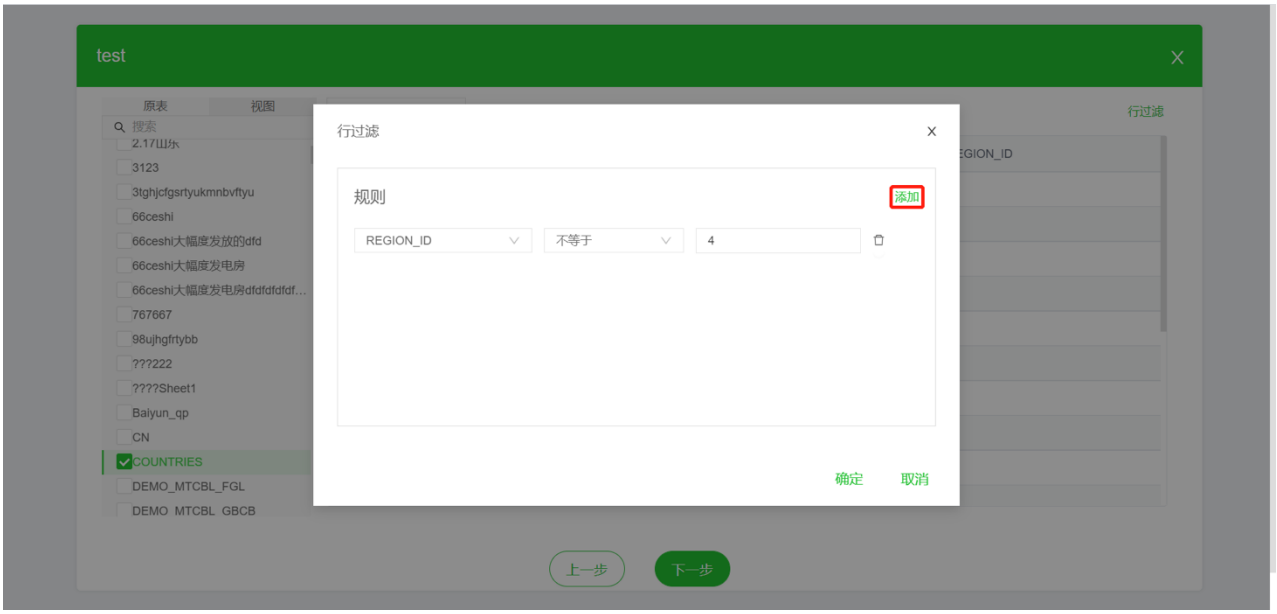


（ 3 ）数据预览

数据库连接成功后，系统自动跳转到【数据预览】页面，在左侧数据表中选择希望抽取的数据表。在页面右侧，点击“行过滤” ，弹出【行过滤】弹窗。



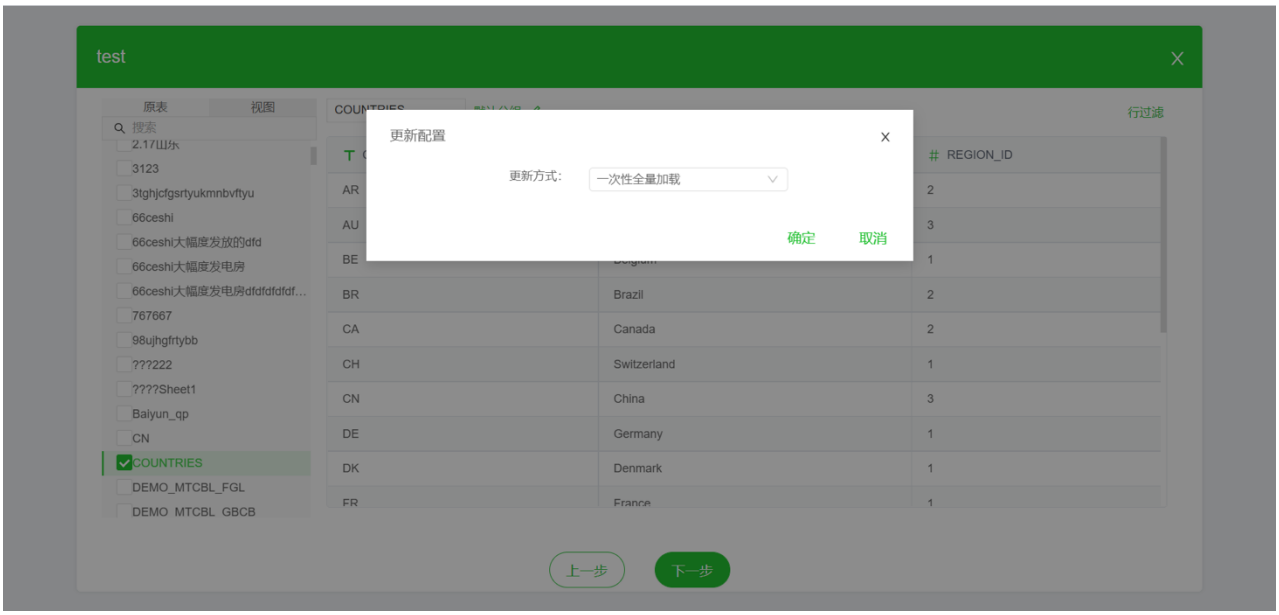
在【行过滤】弹窗，通过添加过滤规则，点击“确定” ，返回到【数据预览】页面，此时已经过滤掉了不符合规则的行信息。



（ 4 ）提交保存

在【数据预览】页面点击“下一步” ，系统弹出更新设置弹窗，用户可以选择数据更新

方式，点击“确定”，操作完成。

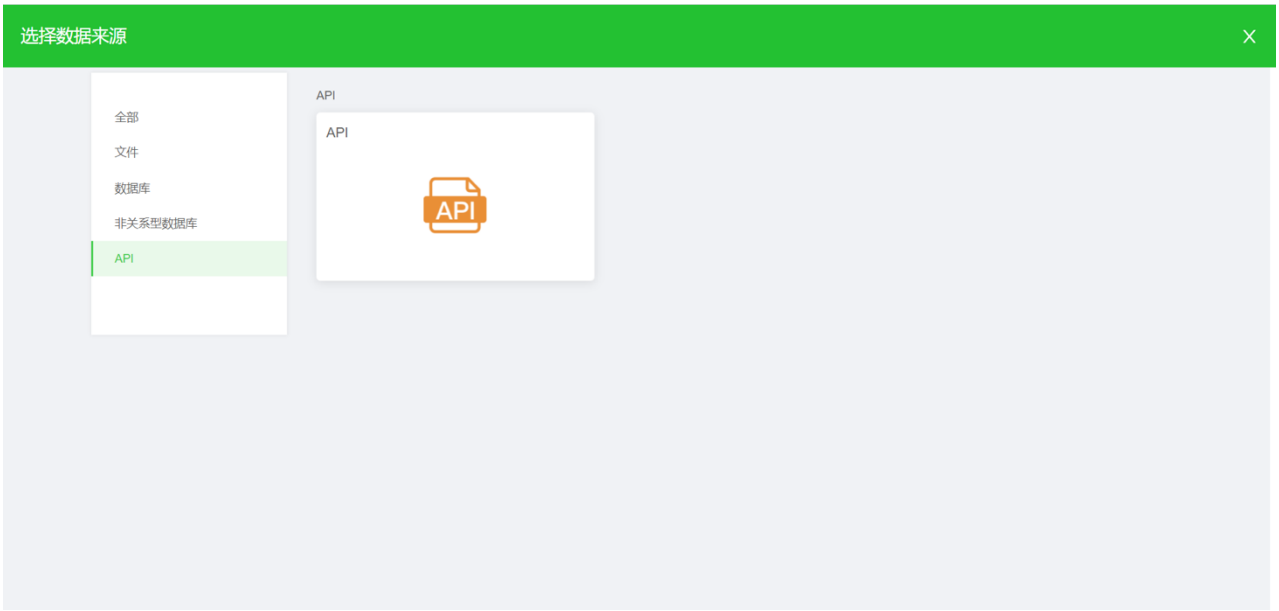


二、 API 数据源抽取

API 类型的数据源抽取，需要先建立 API 连接。

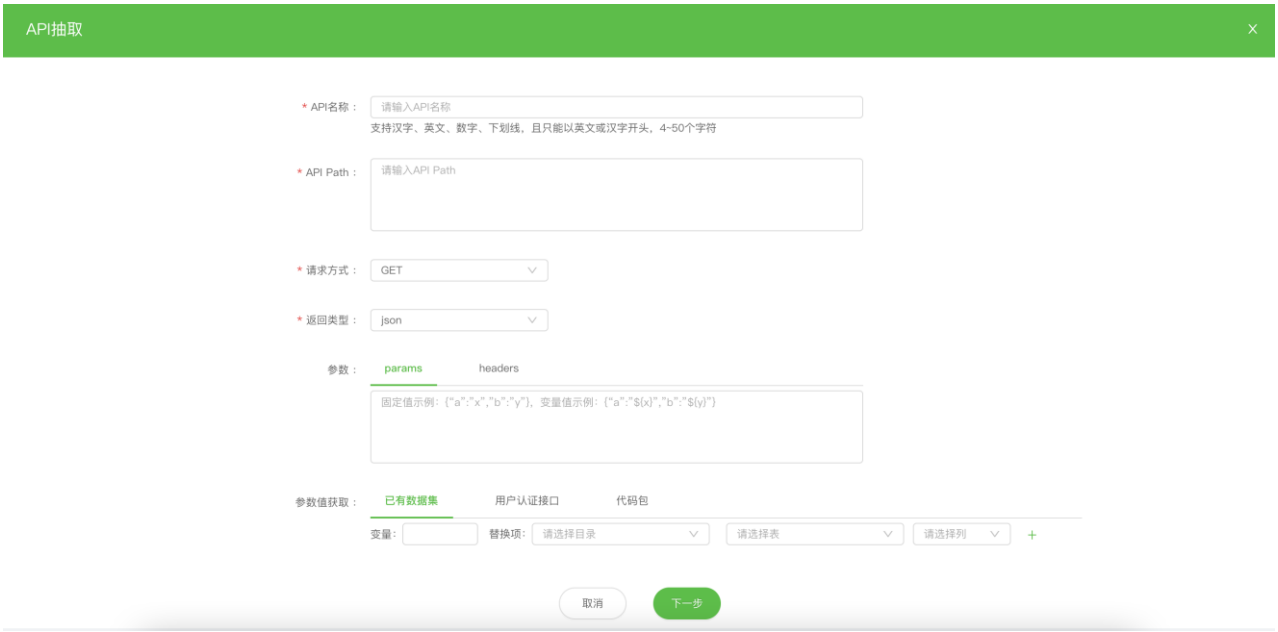
（1）选择链接类型

点击“添加抽取任务”，跳转到【选择数据来源】页面，点击左侧菜单中的“API”，右侧显示出【API 图标】，点击“API”图标，跳转到【创建 API 连接】页面。



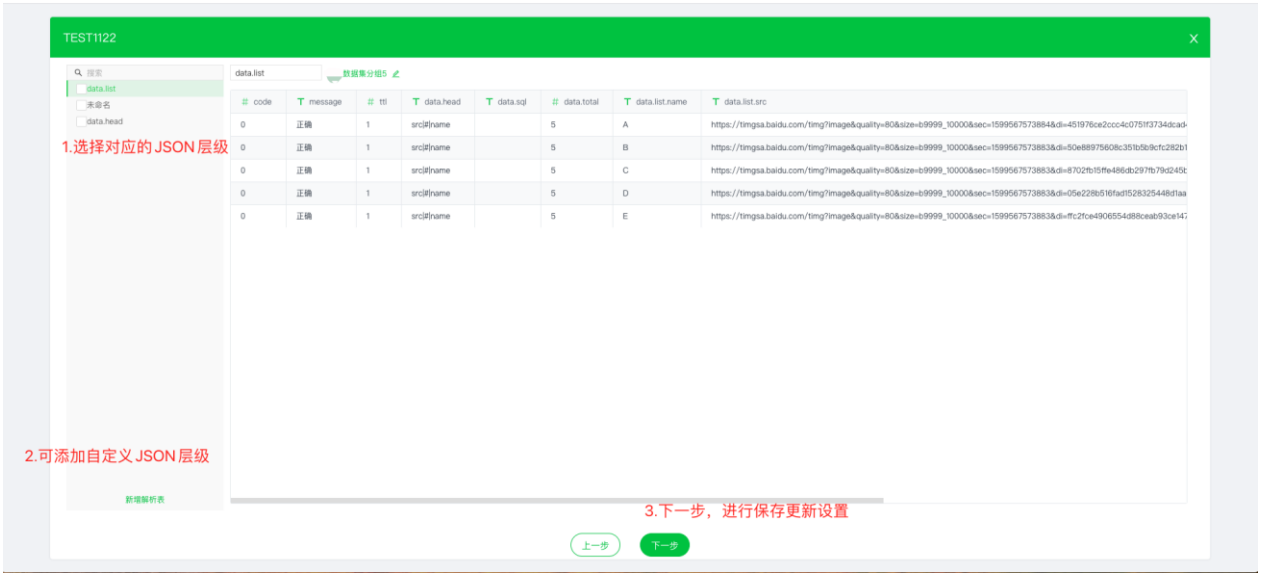
（2）创建 API 连接

在【创建 API 连接】页面，输入 API 名称、API path、请求方式、返回类型、参数、参数值获取后，点击“下一步”



（ 3 ）数据预览

API 连接创建成功后，系统自动跳转到 API 接口【数据预览】页面。在页面左侧，可以选择对应的 JSON 层级，也可以添加自定义 JSON 层级，然后点击“下一步”，操作完成。



2.3.2 数据订阅

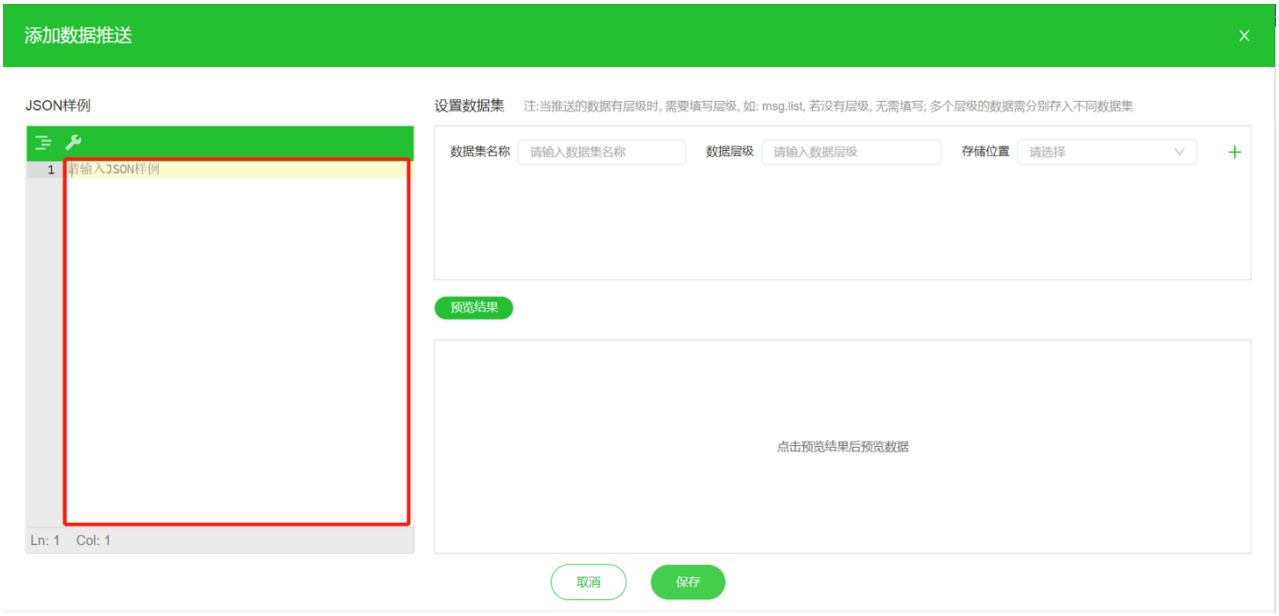
通过功能，可以在 Data Formula 系统创建 JSON 数据推送接口，供外部系统调用。外部系统通过调用推送接口能够向 Data Formula 系统推送 JSON 格式的数据。

一、添加任务

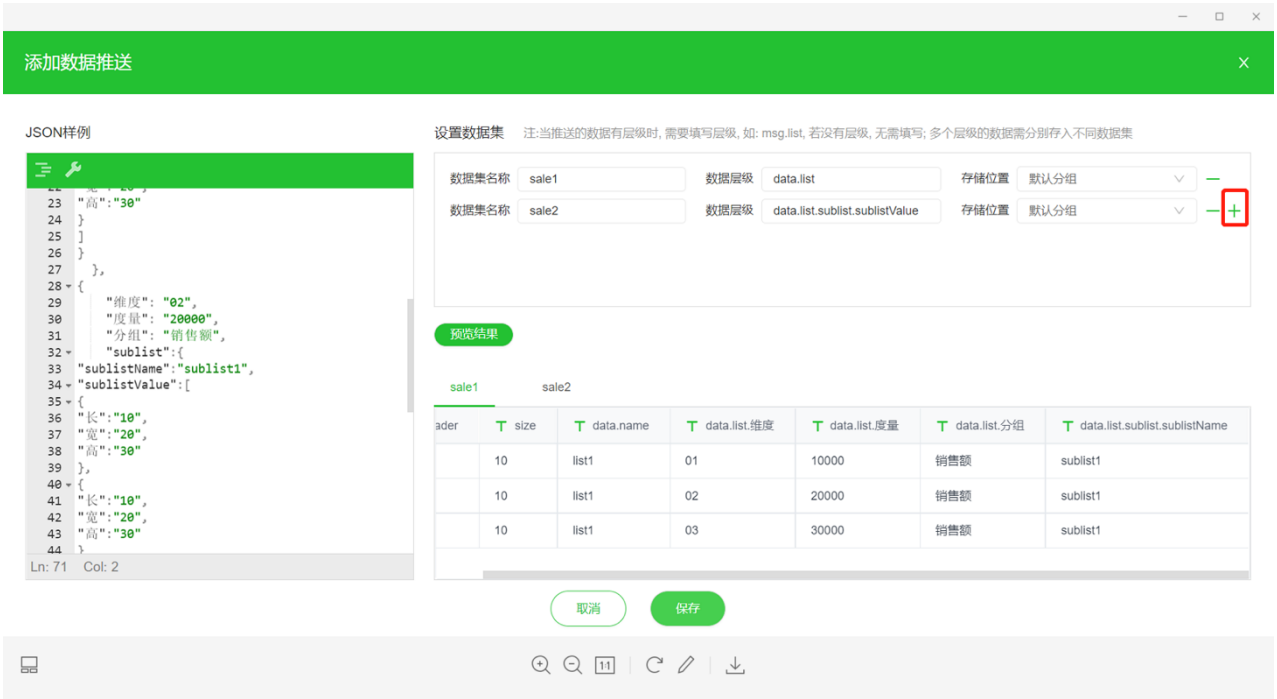
在【数据订阅】页面，点击“添加订阅任务”，跳转到【添加数据订阅】页面。

在【添加数据订阅】页面的【JSON 样例】区域，输入 JSON 样例，然后在【设置数据集】区域输入【数据集名称】、【数据层级】、【存储位置】，点击“预览结果”，可以预

览导入的 JSON 数据。当推送的数据有层级时，需要填写层级；若没有层级，则不用填写。多个层级的数据需分别存入不同数据集。



点击【设置数据集】区域右侧的“+”号，可以【新增数据集】。新增加的数据集在【预览结果】区域以 tab 的方式展现。预览无误后，点击“保存”。



二、推送任务列表操作

推送任务接口建好后，在推送任务列表中生成一条推送任务接口记录。外部系统将通过

该接口按照创建接口时的 JSON 格式向 Data Formula 系统推送数据。推送过来的数据可以在【数据治理】-【资产目录】中查看，并作为数据集在 Data Formula 系统中供用户进行操作。

在推送任务列表中不仅可以对推送任务进行修改和删除操作，还可以查看任务详情和复制接口地址。

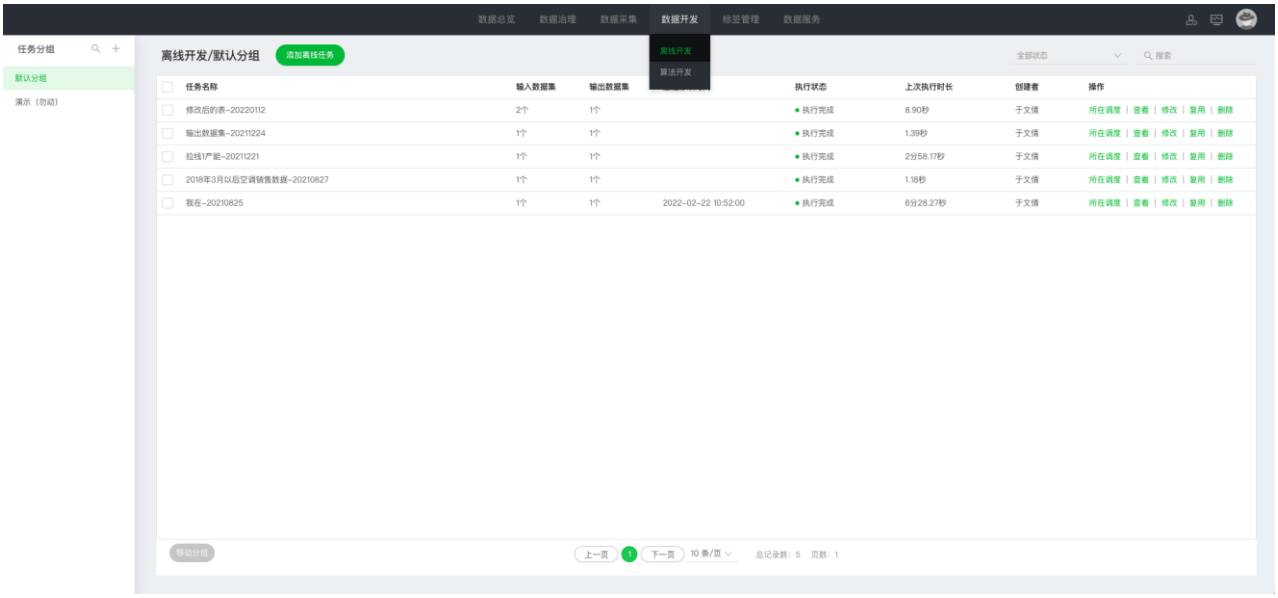
2.3.3 数据源

【数据采集】-【数据源】模块能够聚合展示已执行过抽取任务的数据源信息，并可以对数据源进行修改、删除、查看详情操作。

2.4 数据开发

2.4.1 离线开发

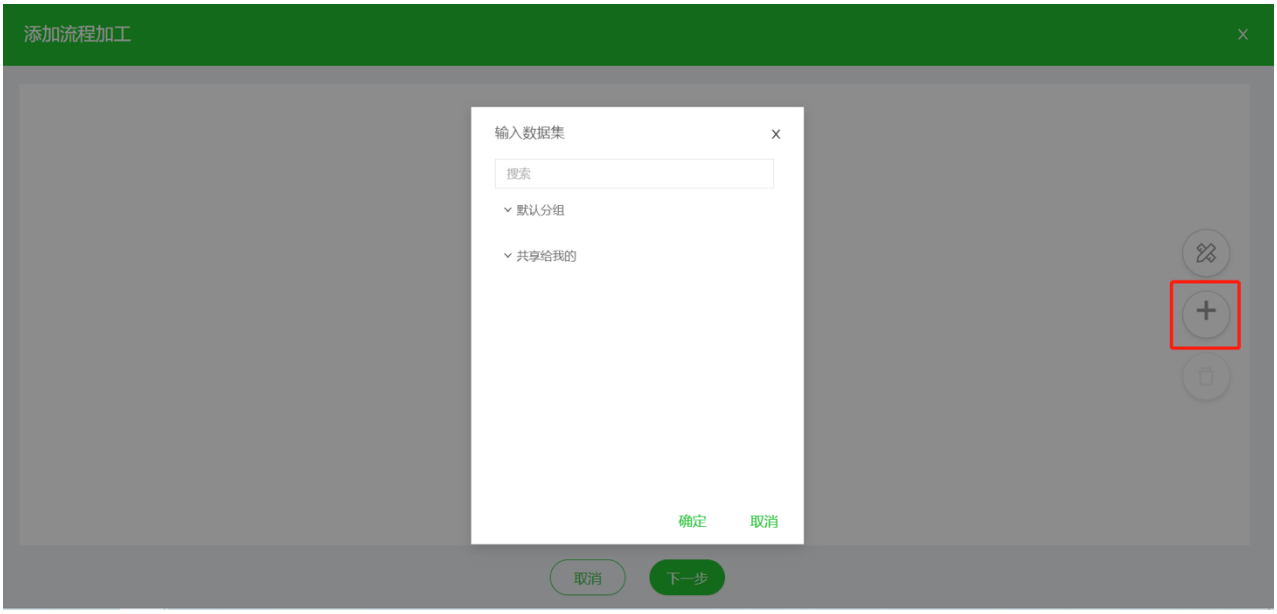
【数据开发】-【离线开发】模块支持以工作流节点拖拽的方式，配置数据加工流程，然后在【数据总览】-【任务调度】模块按照工作流任务统一调度执行。



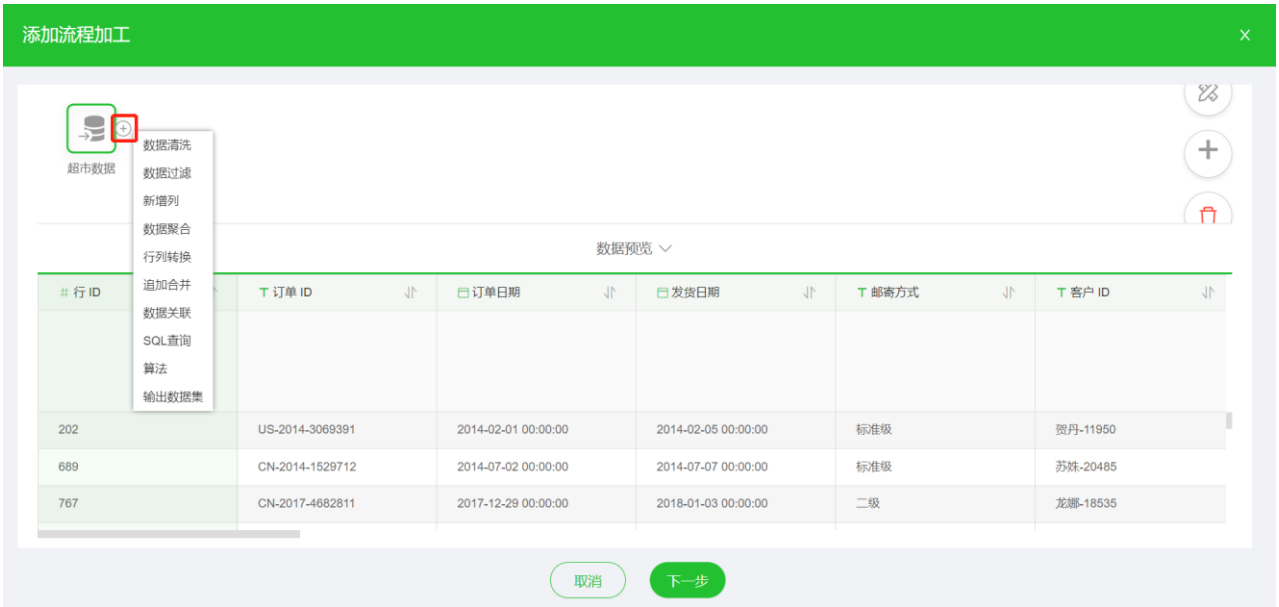
一、添加加工任务

点击“添加加工任务”，跳转到【加工工作流配置】页面。

在【加工工作流配置】页面，点击右侧的“+”号，添加一个或多个数据集。在弹出的【数据集配置】页面，选择需要加工的数据集，并点“确定”，完成数据集的添加。



数据集添加成功后，点击数据集节点右边的小“+”号，弹出全部加工节点，可以选择对应的加工节点创建加工流程。加工节点包括：数据清洗、数据过滤、新增列、数据聚合、行列转换、追加合并、数据关联、SQL 关联、算法、输出数据集。

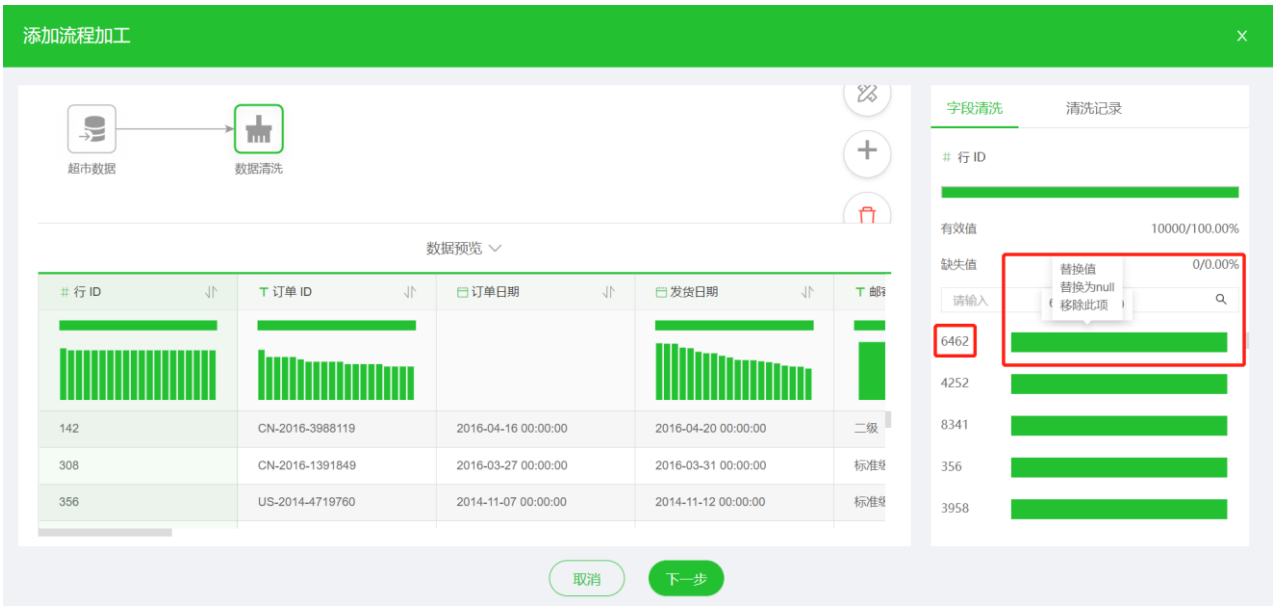


（1）数据清洗

点击【数据集】节点右边的小“+”号，在弹出的加工节点选择【数据清洗】，生成一

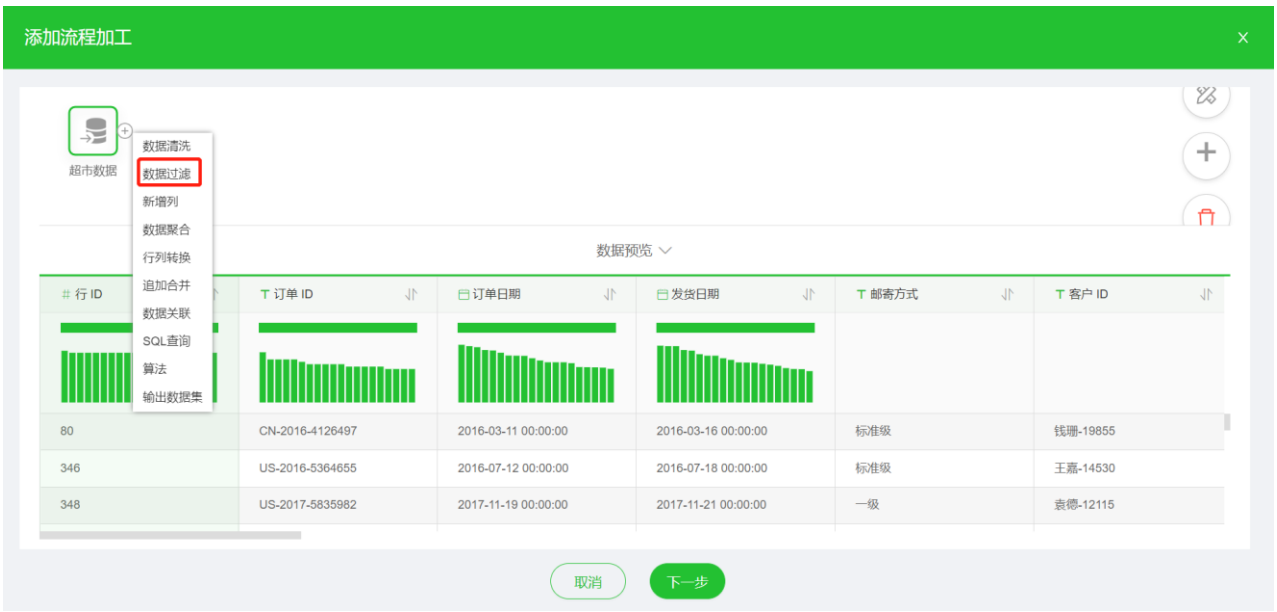
个【数据清洗】的节点。

点击“数据清洗节点”，页面右侧默认显示第一列字段信息，右键点击“单个字段值”弹出【替换值】、【替换为 null】、【移除此项】的清洗操作选项，可以选择其中一个选项进行相应操作。

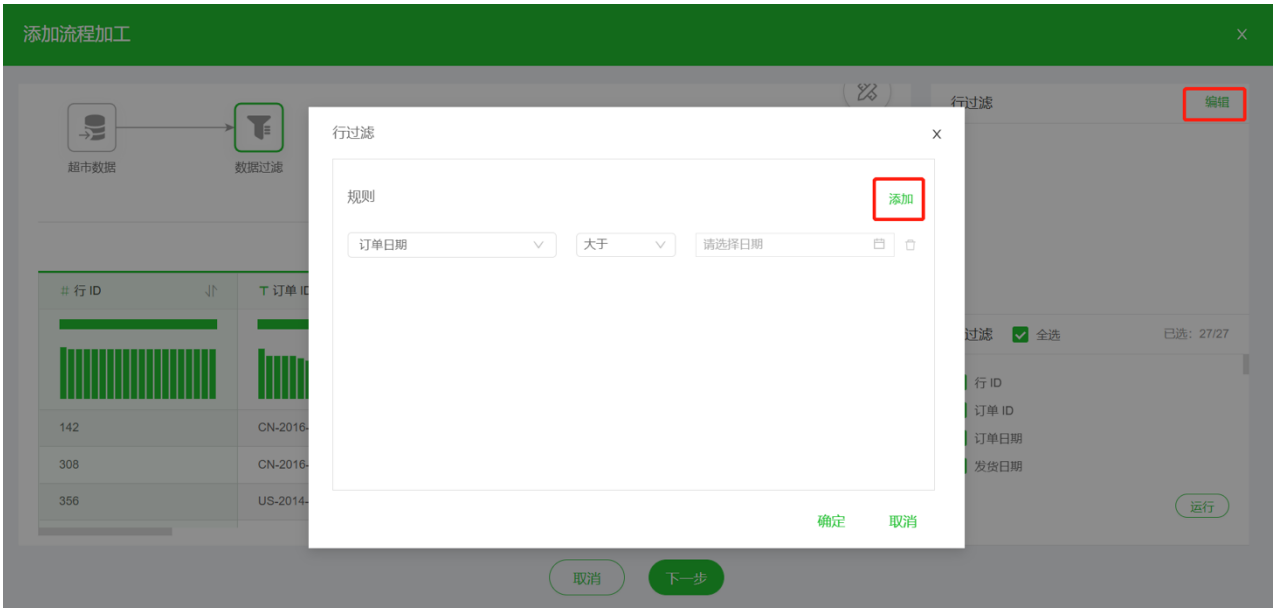


（2） 数据过滤

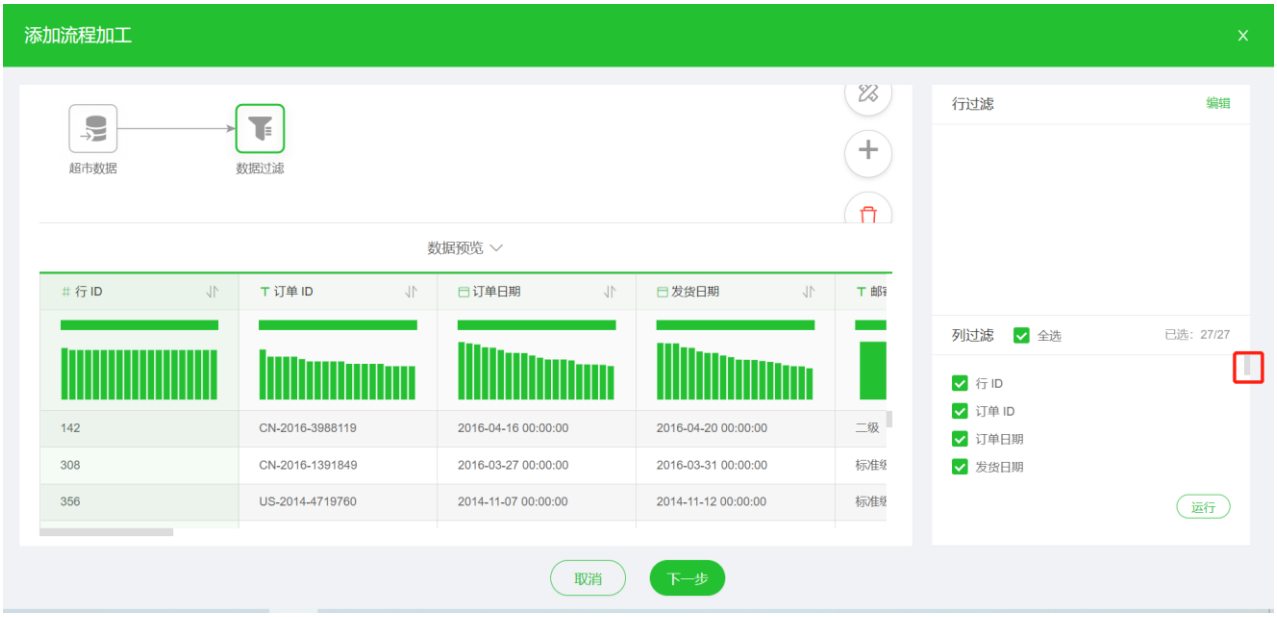
点击【数据集】节点右边的小“+”号，在弹出的加工节点选择【数据过滤】，生成一个【数据过滤】的节点。



行过滤：点中【数据过滤】节点，右侧出现【行过滤】和【列过滤】的操作区域。点击右上角的“编辑”按钮，弹出【行过滤】弹窗。系统可以根据添加的过滤条件，过滤掉不符合条件的数据行。



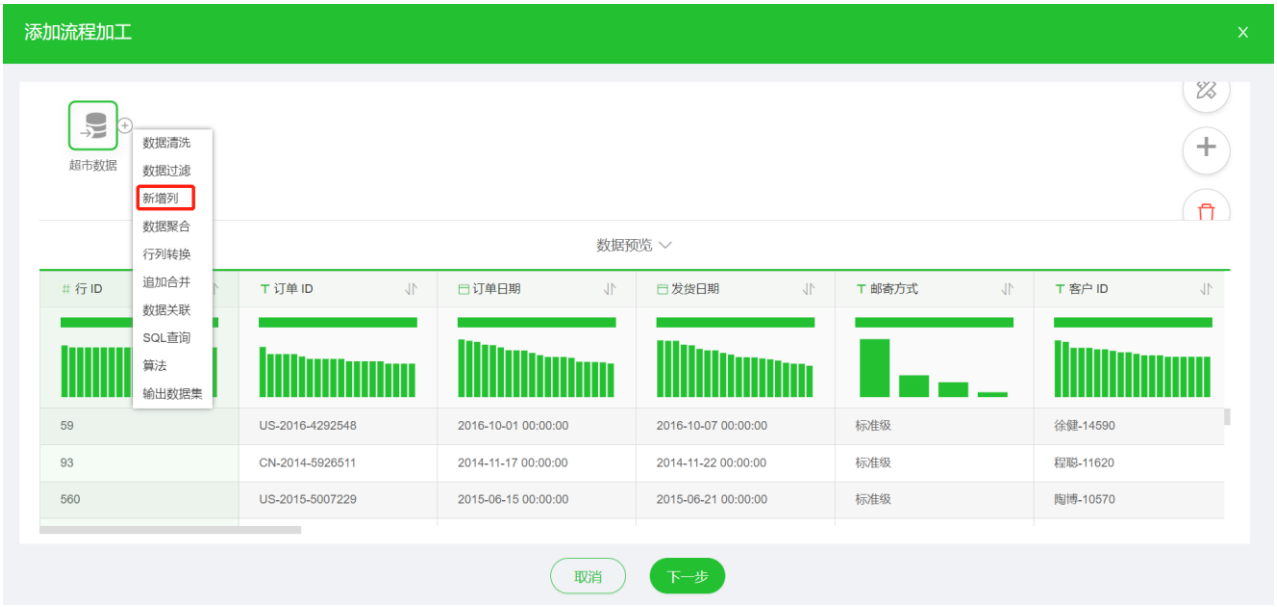
列过滤：在列过滤操作区域，可以通过勾选的方式，对不希望显示的列字段进行取消勾选操作，从而达到过滤掉列的目的。



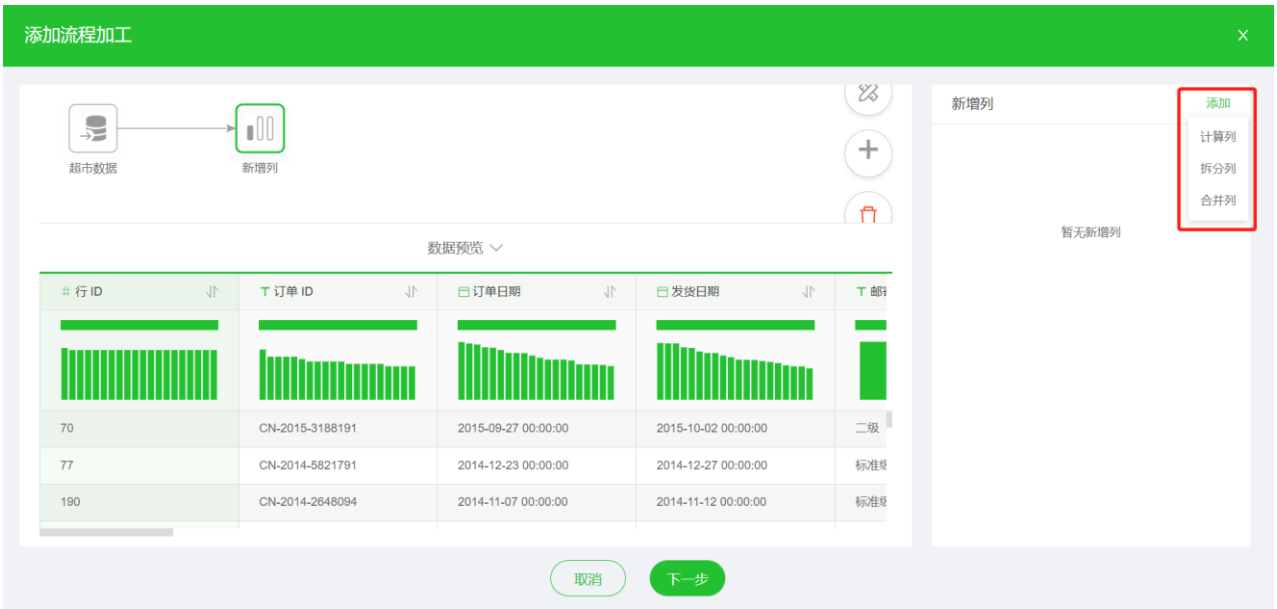
（3） 新增列

点击【数据集】节点右边的小“+”号，在弹出的加工节点选择【新增列】，生成一

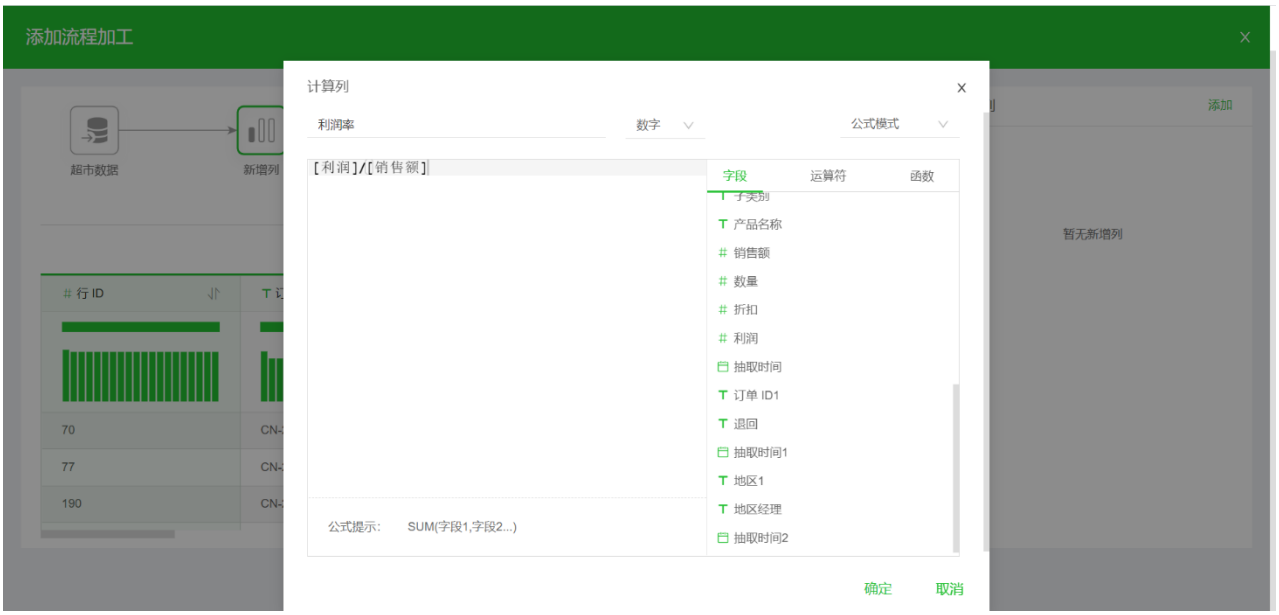
个【新增列】的节点。



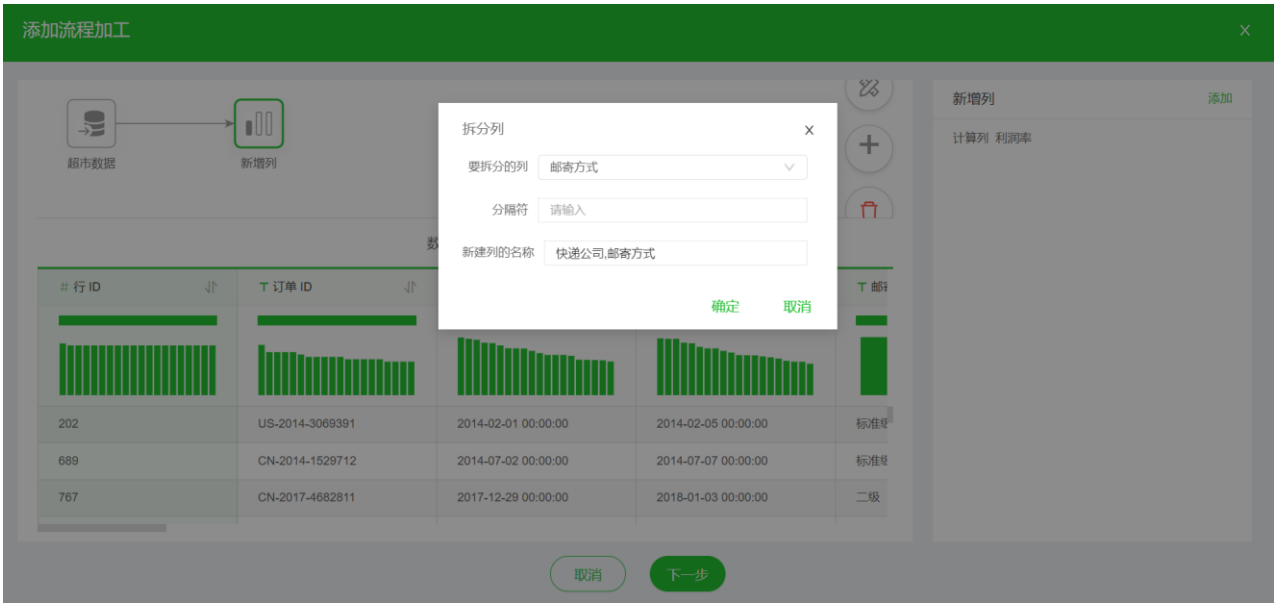
点中【新增列】节点，右侧出现【新增列】的操作区域。点击右上角的“添加”按钮，弹出【计算列】、【拆分列】、【合并列】的操作选项。



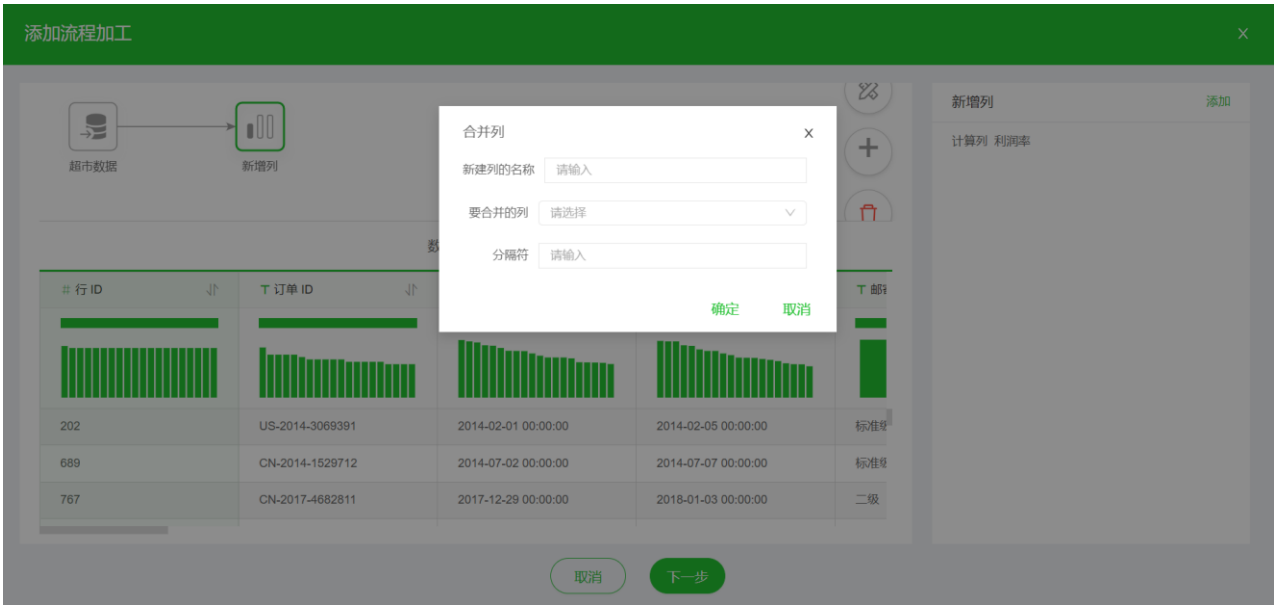
计算列：点击“计算列”，弹出【计算列】页面。在【计算列】页面中，输入要增加的字段名称、字段类型，然后在右侧选择计算字段、运算符/函数，点击“确认”，系统检测计算公式，若公式无误将保存创建的列。



拆分列：点击“拆分列”，弹出【拆分列】页面，输入要拆分的列、分隔符、新建列的名称（列与列之间用逗号分隔），点击“确定”。



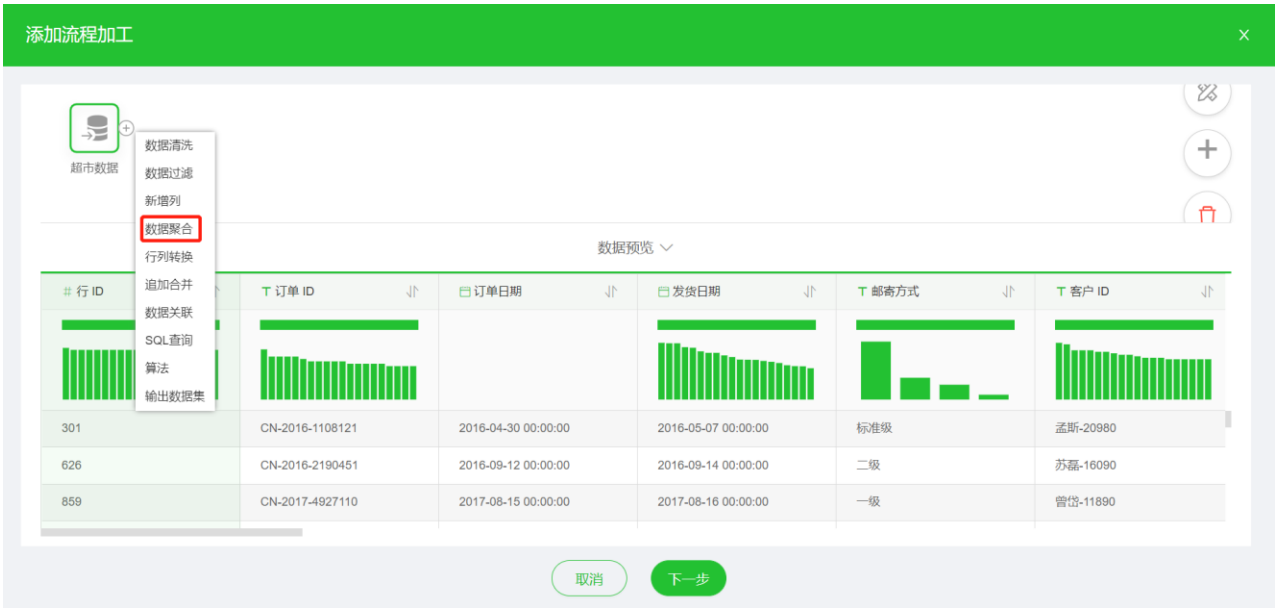
合并列：点击“合并列”，弹出【合并列】页面，输入新建列的名称、要合并的列、分隔符，点击“确定”。



（4）数据聚合

数据聚合其实是对数据集中的行数据进行数值合并。

点击【数据集】节点右边的小“+”号，在弹出的加工节点选择【数据聚合】，生成一个【数据聚合】的节点。



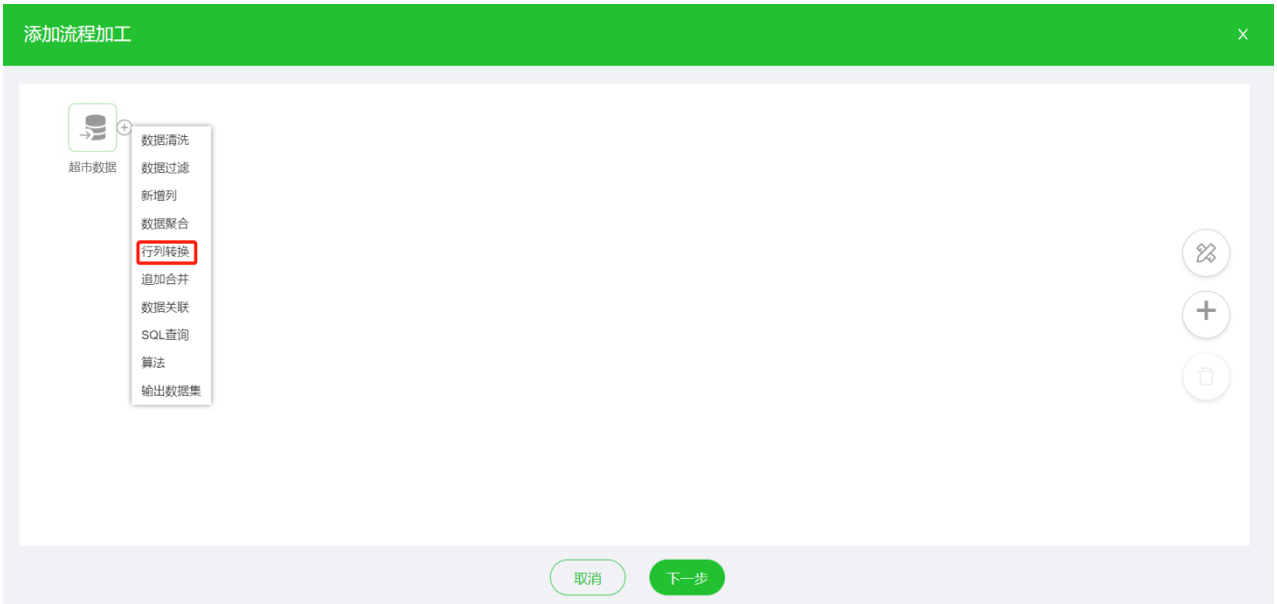
点中【数据聚合】节点，右侧出现【数据聚合】的操作区域。其中包括：选择聚合维度和选择聚合数值。【选择聚合维度】是将该字段做为计算维度进行聚合。【选择聚合数值】是参与聚合计算的字段。选择合适的【聚合维度】和【聚合数值】后，点击“运行”

进行计算。

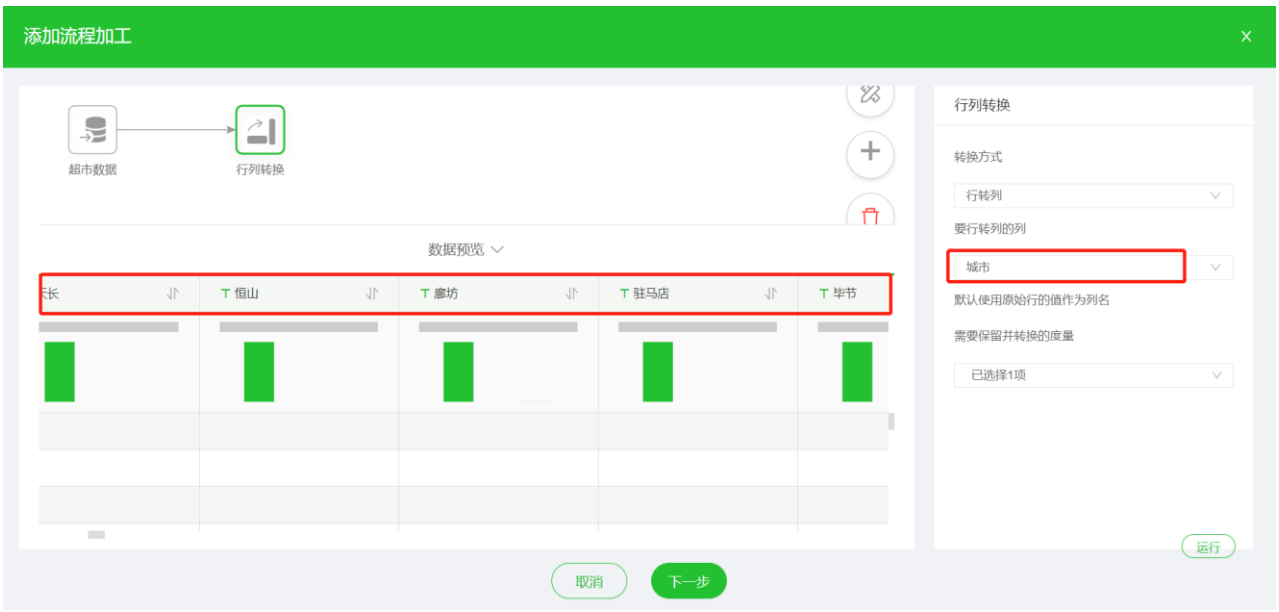
（5） 行列转换

行列转换是将数据集的行转换成列，列转换成行。

点击【数据集】节点右边的小“+”号，在弹出的加工节点选择【行列转换】，生成一个【行列转换】的节点。



行转列：选中【行列转换】节点，在右侧区域内，【转换方式】选择【行转列】，然后选择【要行转列的列】，点击“运行”。



行转列示例如下：

实际效果说明：

情况1：有一个度量值

举个例子

月份	省份	购买方式	销量
1月	福建省	全款	11123
1月	福建省	按揭	2324
1月	江苏省	按揭	54666
1月	江苏省	全款	68880
1月	河北省	按揭	44445
1月	河北省	全款	39057
2月	福建省	全款	66664
2月	福建省	按揭	23411
2月	江苏省	按揭	9754
2月	江苏省	全款	565732
2月	河北省	全款	6663356
2月	河北省	按揭	4532

月份	省份	全款	按揭
1月	福建省	11123	2324
1月	江苏省	68880	54666
1月	河北省	39057	44445
2月	福建省	66664	23411
2月	江苏省	565732	9754
2月	河北省	6663356	4532

情况2：有多个度量值

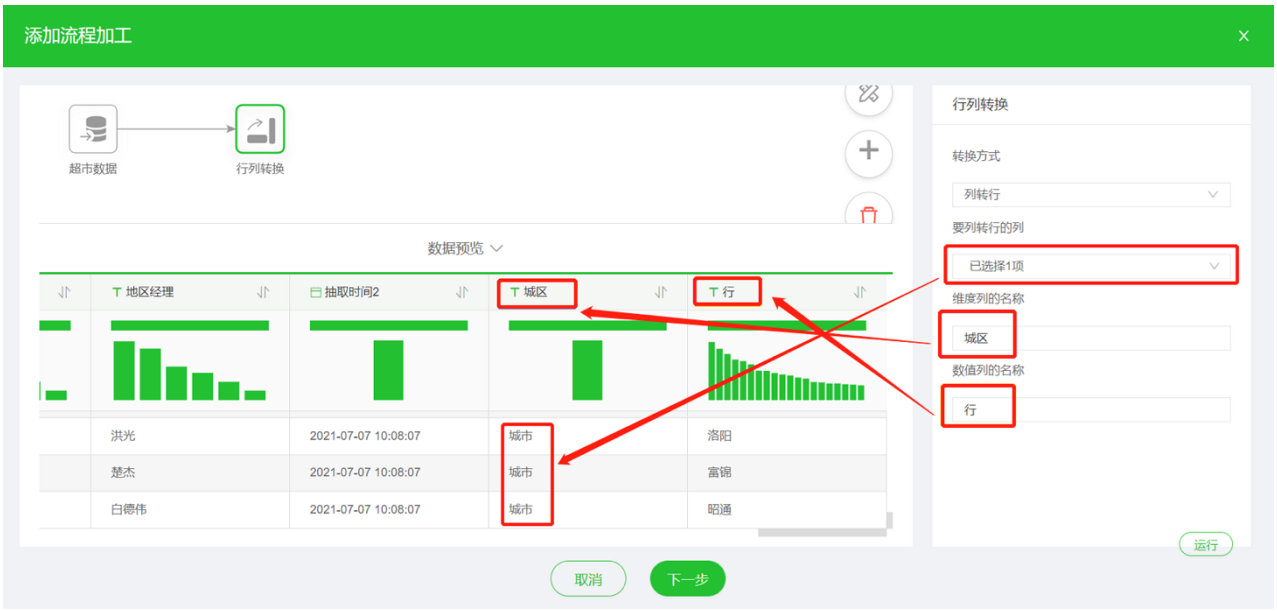
举个例子

月份	省份	购买方式	销量	利润
1月	福建省	全款	11123	A
1月	福建省	按揭	2324	B
1月	江苏省	按揭	54666	C
1月	江苏省	全款	68880	A
1月	河北省	按揭	44445	C
1月	河北省	全款	39057	B
2月	福建省	全款	66664	B
2月	福建省	按揭	23411	B
2月	江苏省	按揭	9754	A
2月	江苏省	全款	565732	C
2月	河北省	全款	6663356	B
2月	河北省	按揭	4532	B

列名自动是原列的值+度量字段名

月份	省份	全款-销量	全款-利润	按揭-销量	按揭-利润
1月	福建省	11123	A	2324	B
1月	江苏省	68880	A	54666	C
1月	河北省	39057	B	44445	C
2月	福建省	66664	B	23411	B
2月	江苏省	565732	C	9754	A
2月	河北省	6663356	B	4532	B

列转行：选中【行列转换】节点，在右侧区域内，【转换方式】选择【列转行】，然后选择【要列转行的列】，输入【维度列的名称】、【数值列的名称】，点击“运行”。



列转行示例如下：

实际效果说明：

原表

姓名	语文	数学	英语
张三	89	100	69
李四	80	92	76
王五	60	82	96

选择列转行的列：
语文、数学、英语

维度列的名称：
科目

数值列的名称：
成绩

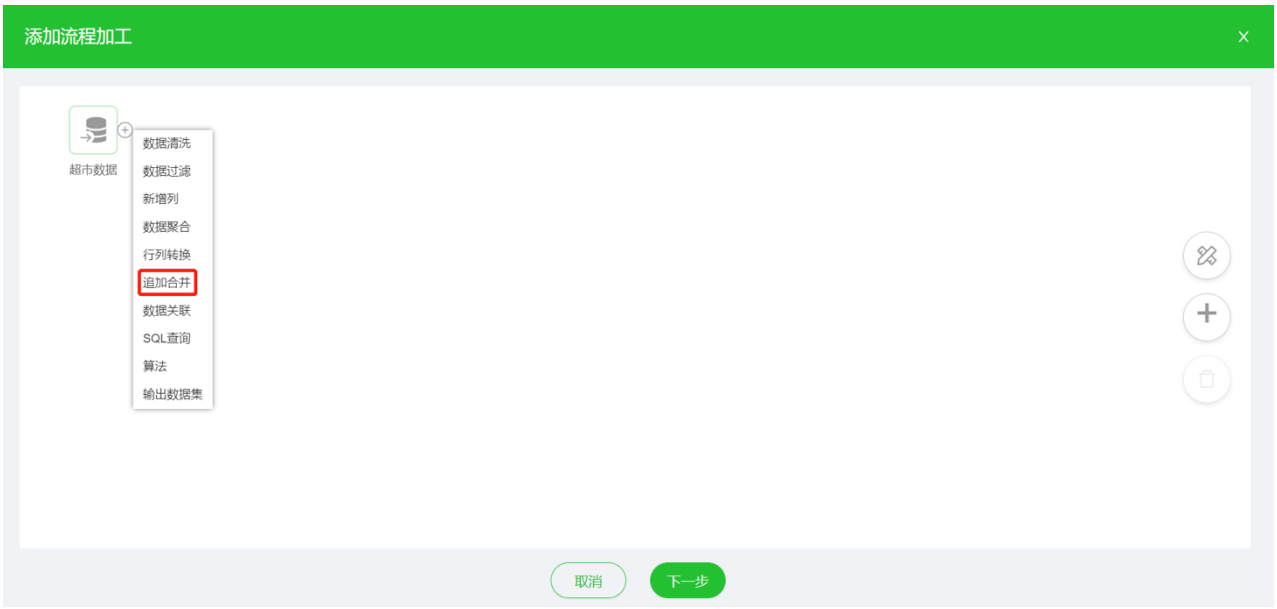
结果表

姓名	科目	成绩
张三	语文	89
李四	语文	80
王五	语文	60
张三	数学	100
李四	数学	92
王五	数学	82
张三	英语	69
李四	英语	76
王五	英语	96

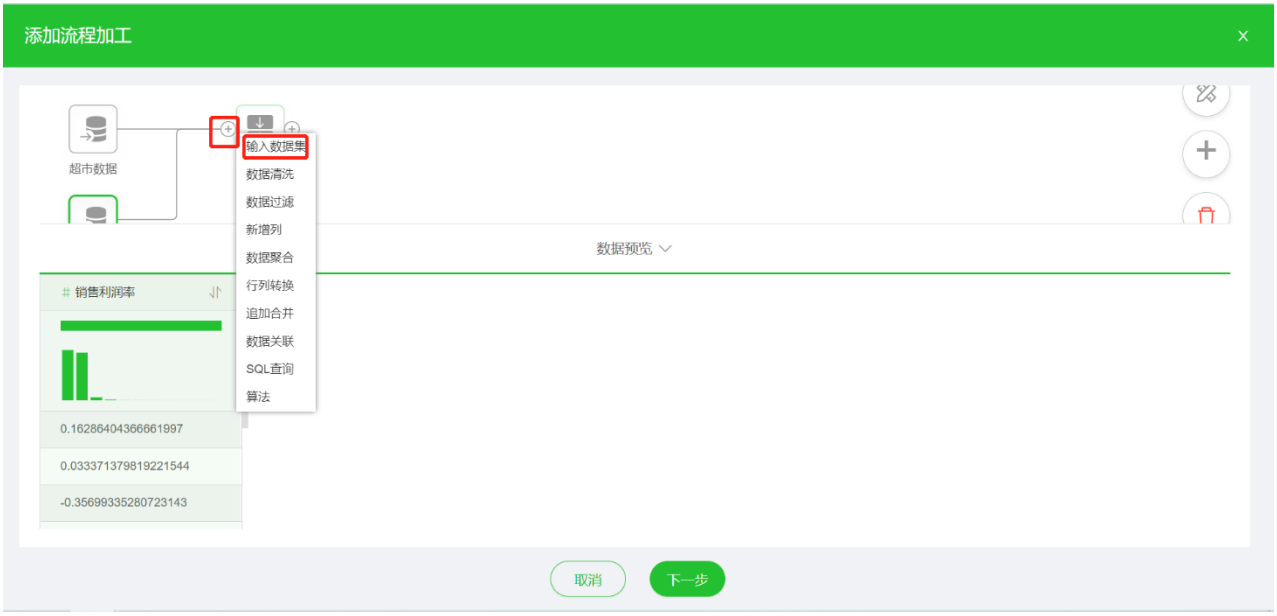
（6）追加合并

追加合并是对多个数据集的行数据进行合并，追加合并是多个数据集的操作。

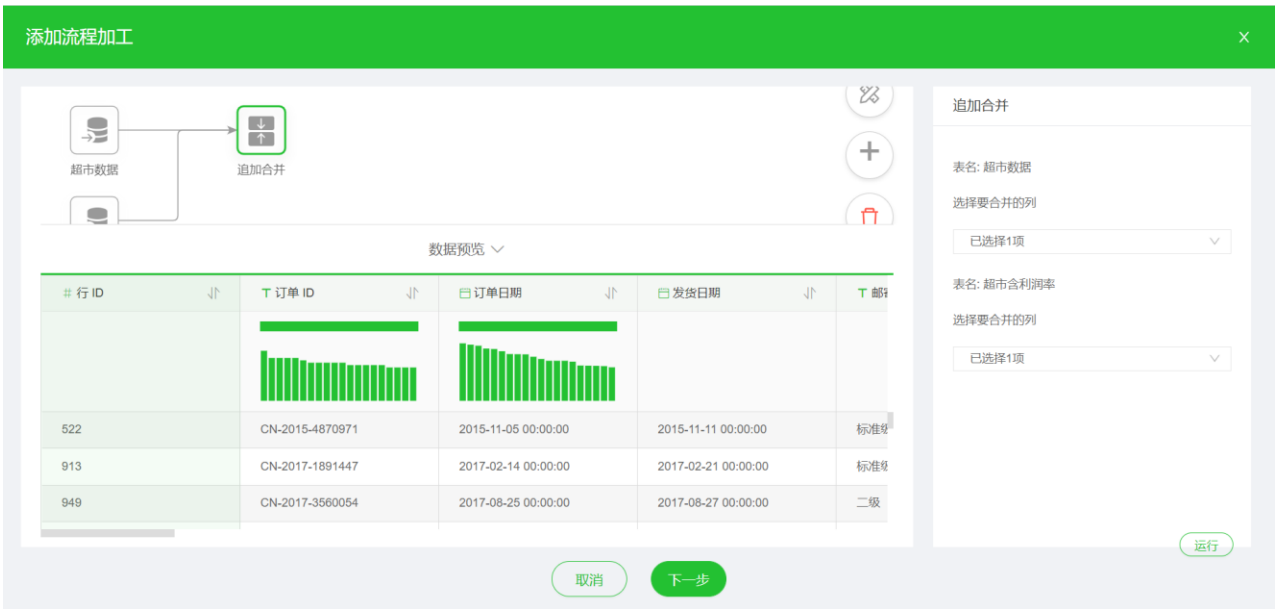
点击【数据集】节点右边的小“+”号，在弹出的加工节点选择【追加合并】，生成一个【追加合并】的节点。



因为追加合并是多个数据集的数据进行合并，所以需要点中【追加合并】节点左侧的“+”添加至少一个【输入数据集】。



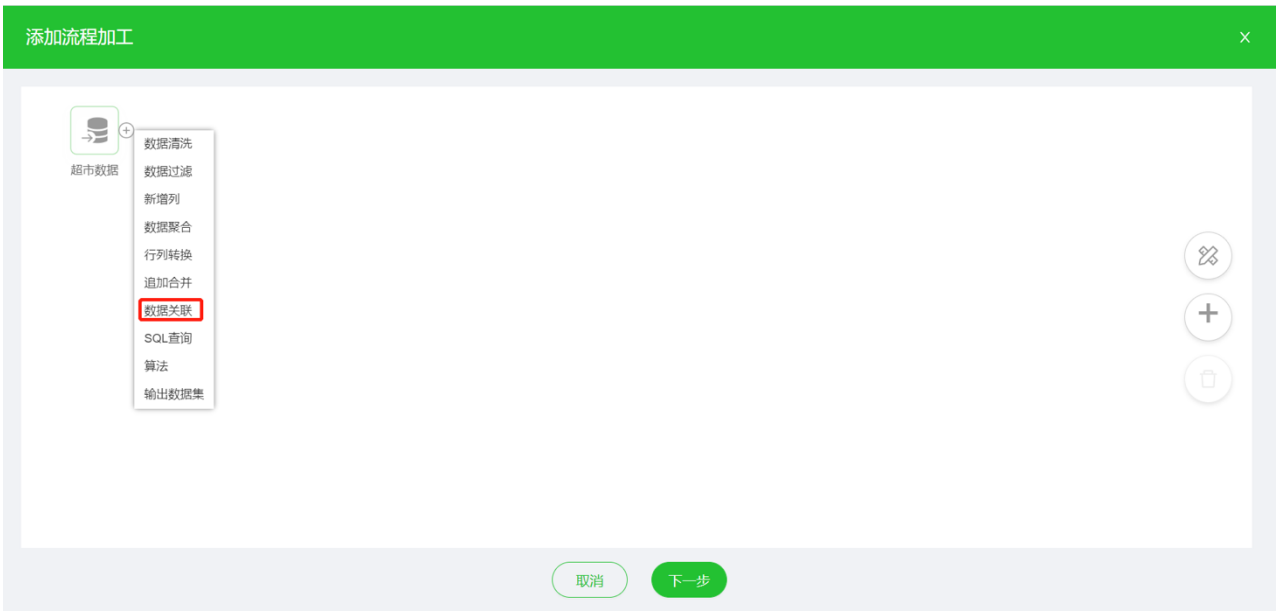
两个及两个以上输入数据集添加完成后，点中【追加合并】节点，右侧操作区域出现【数据集名称】和【选择要合并的列】选项，此时选择要合并的列，点击“运行”。合并的列字段类型需要一致。



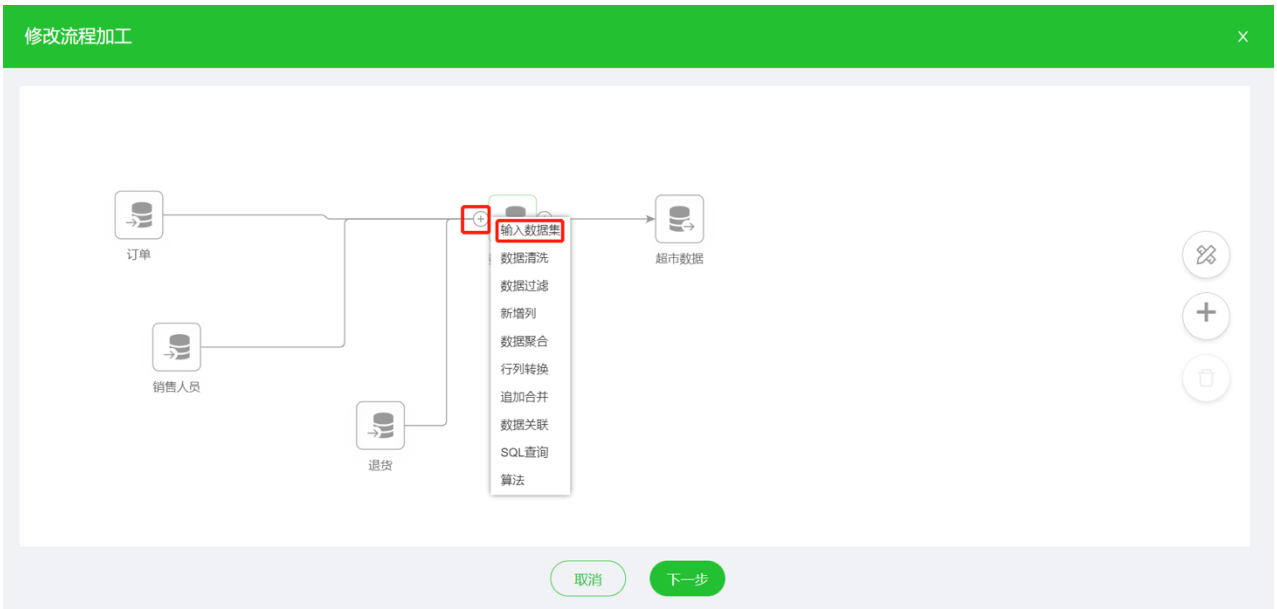
（7）数据关联

数据关联是多个数据集做数据关联，数据关联操作是多个数据集的操作。

点击【数据集】节点右边的小“+”号，在弹出的加工节点选择【数据关联】，生成一个【数据关联】的节点。

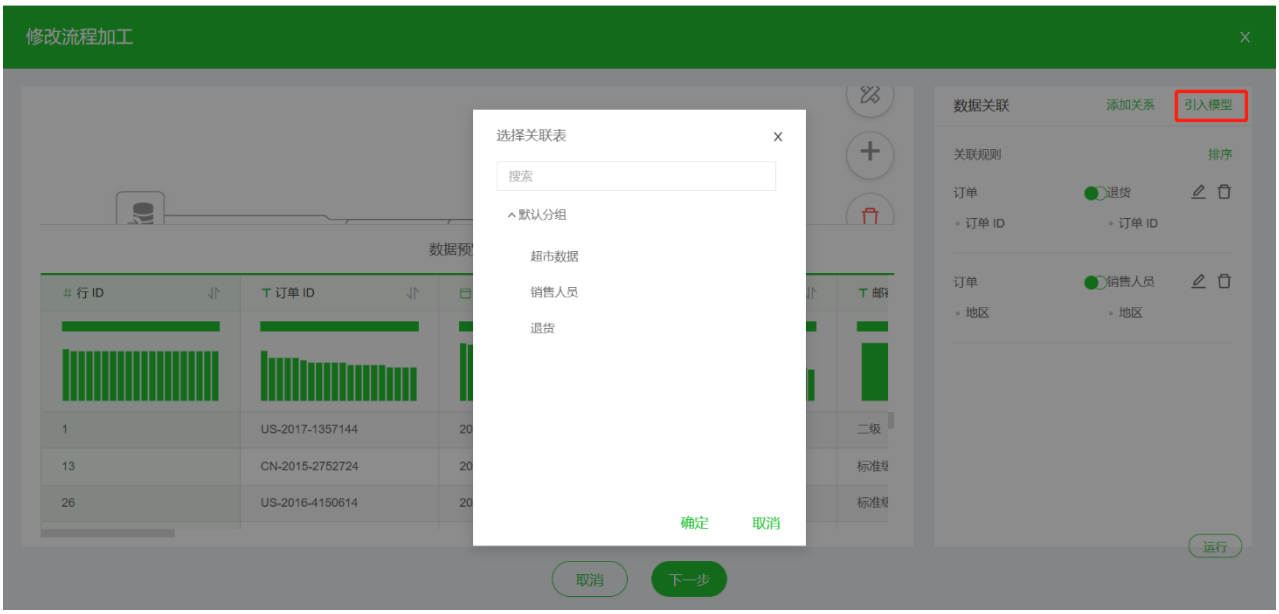


因为数据关联是多个数据集的数据进行关联，所以需要点中【数据关联】节点左侧的“+”添加至少一个【输入数据集】。



两个及两个以上输入数据集添加完成后，点中【数据关联】节点，页面右侧出现数据关联操作区域，区域中有两个按钮：【添加关系】和【引入模型】。点击“添加关系”，弹出【添加关联关系】弹窗，选择关联字段后，点击“确认”。创建好关联关系后，点击“运行”。

点击“引入模型”，弹出【选择关联表】弹窗，选择已经建立过关联模型的数据集，点击“确定”，系统直接将已建立的关联规则导入当前模型，这样可以达到快速复用已有模型的效果。



数据关联示例如下：

操作结果说明：
（维度1为关联字段）

A表：

维度1	度量1
A	3
B	2
D	6

B表：

维度1	度量2
A	8
B	8
C	7

A表和B表左关联：

维度1	度量1	度量2
A	3	8
B	2	8
D	6	

A表和B表右关联：

维度1	度量1	度量2
A	3	8
B	2	8
C		7

A表和B表全关联：

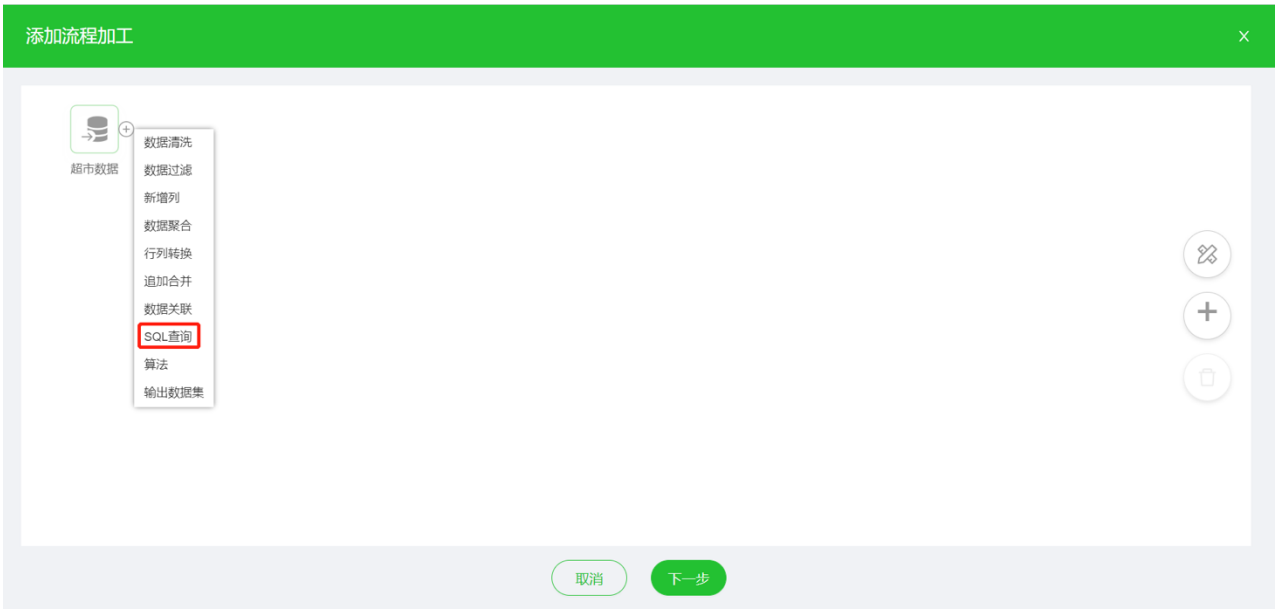
维度1	度量1	度量2
A	3	8
B	2	8
C		7
D	6	

A表和B表内关联：

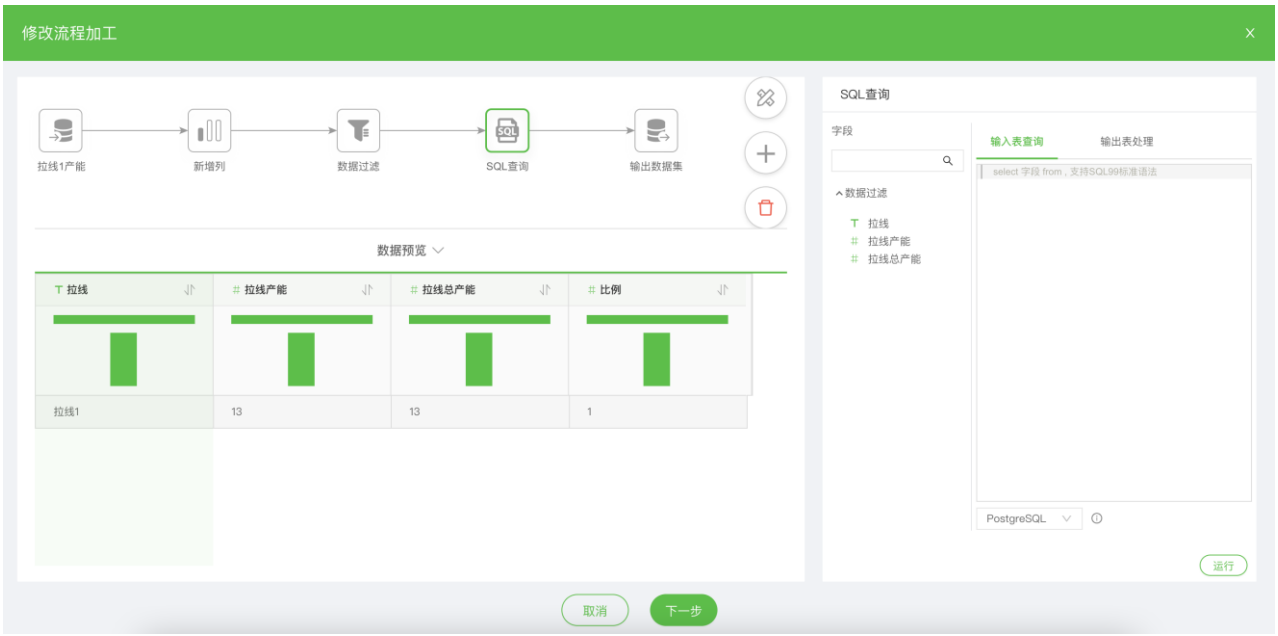
维度1	度量1	度量2
A	3	8
B	2	8

（8） SQL 查询

点击【数据集】节点右边的小“+”号，在弹出的加工节点选择【SQL 查询】，生成一个【SQL 查询】的节点。

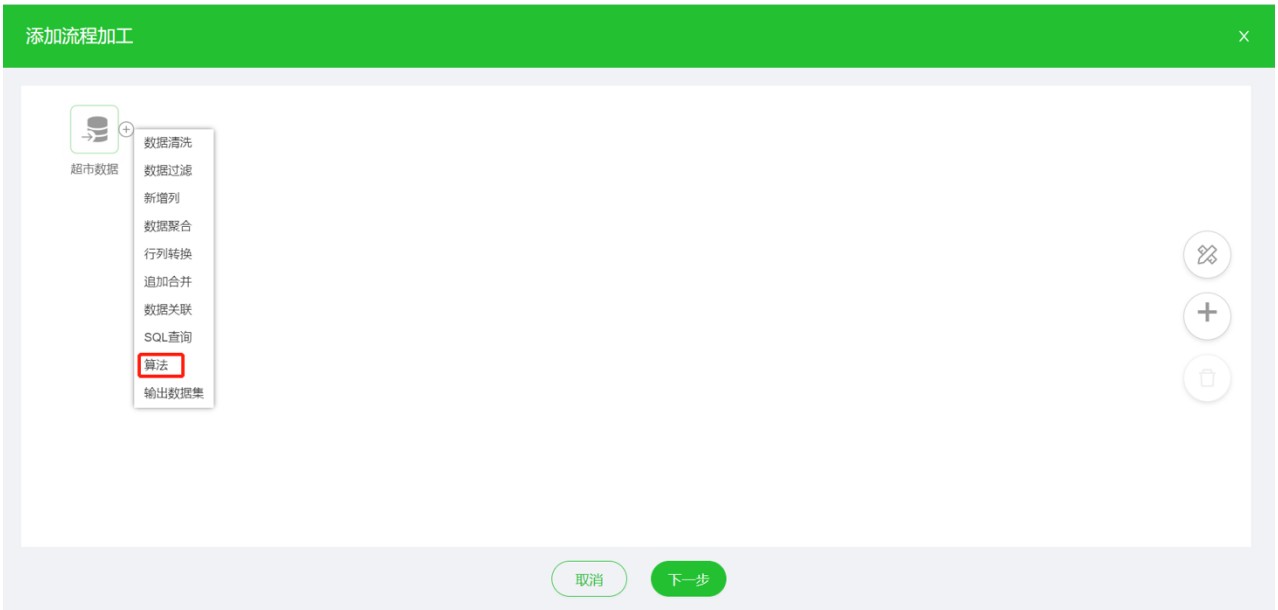


点中【SQL 查询】节点，页面右侧区域出现数据集字段和 SQL 查询语句输入框，输入 SQL 语句后，点击“运行”，运行结果展示在左侧的【数据预览】中。SQL 支持 99 标准语法，下方选择当前输入的 SQL 对应的数据库，若 SQL 语句报错，会有报错提示。



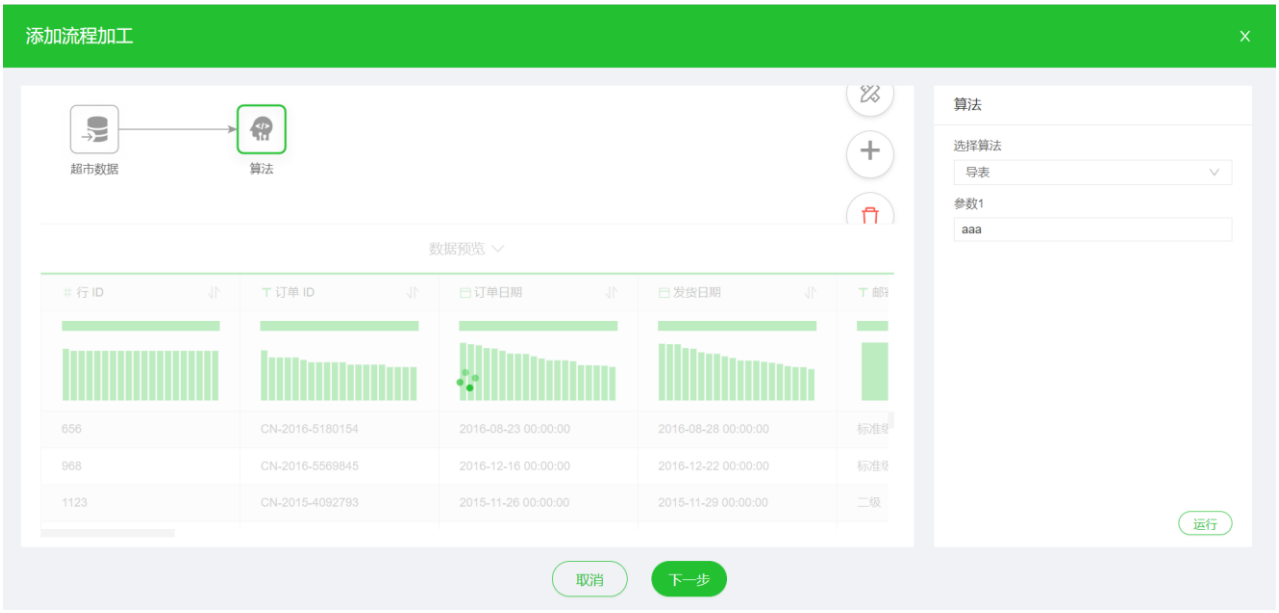
（9） 算法

点击【数据集】节点右边的小“+”号，在弹出的加工节点选择【算法】，生成一个【算法】的节点。



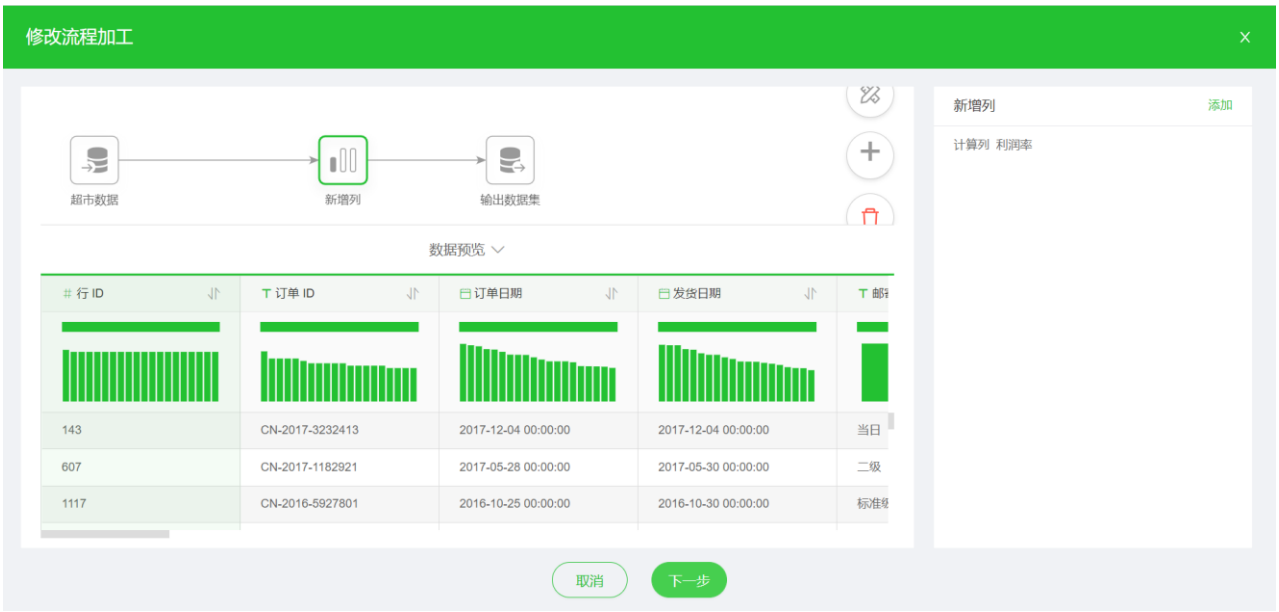
点中【算法】节点，页面右侧区域出现【选择算法】选择框，在【数据资产】-【算法】中创建的算法会显示在该选择框中。选中一个算法后，系统根据选中的算法显示需要填写的内容，若该算法只有一个输入数据集和一个输出数据集，则只需填写参数，点击“运行”即

可。系统将运行结果显示在【数据预览】区域。若该算法有多个输入数据集或者多个输出数据集，则需要在流程中添加多个输入数据集节点或者多个输出数据集节点。

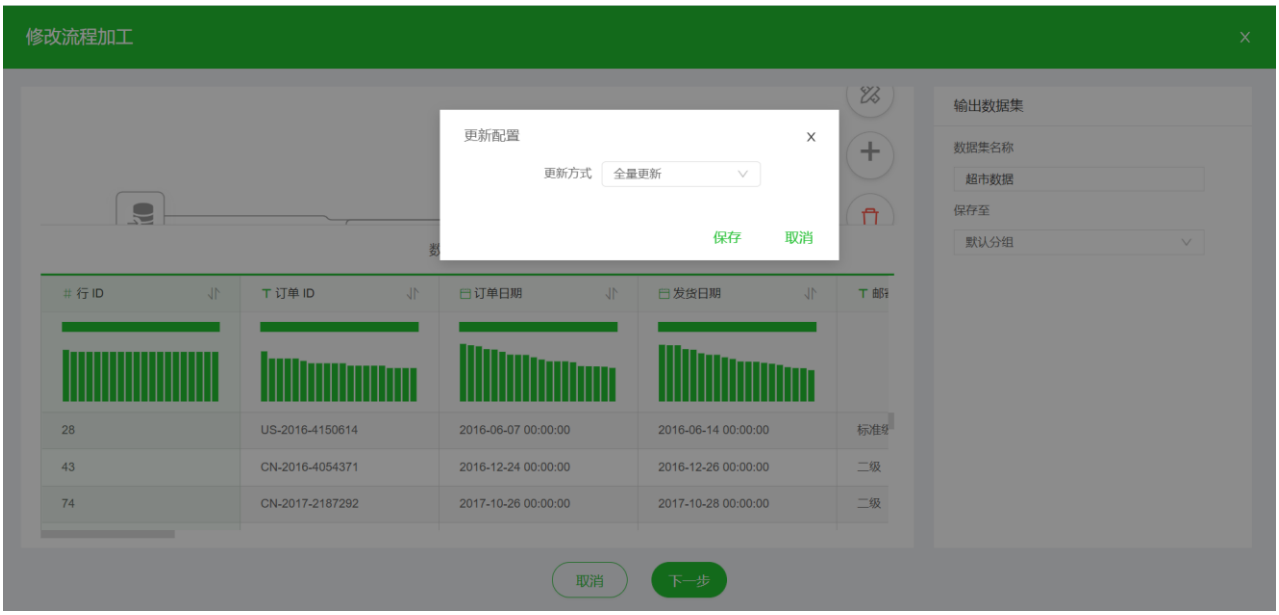


（10）输出数据集

所有的数据流程加工完成后，若要做数据输出，都需要在流程的最后添加一个【输出结果集】节点。



然后点击“下一步”，弹出【更新配置】弹窗。在【更新配置】弹窗选择更新配置后，点击“保存”，流程加工完成。



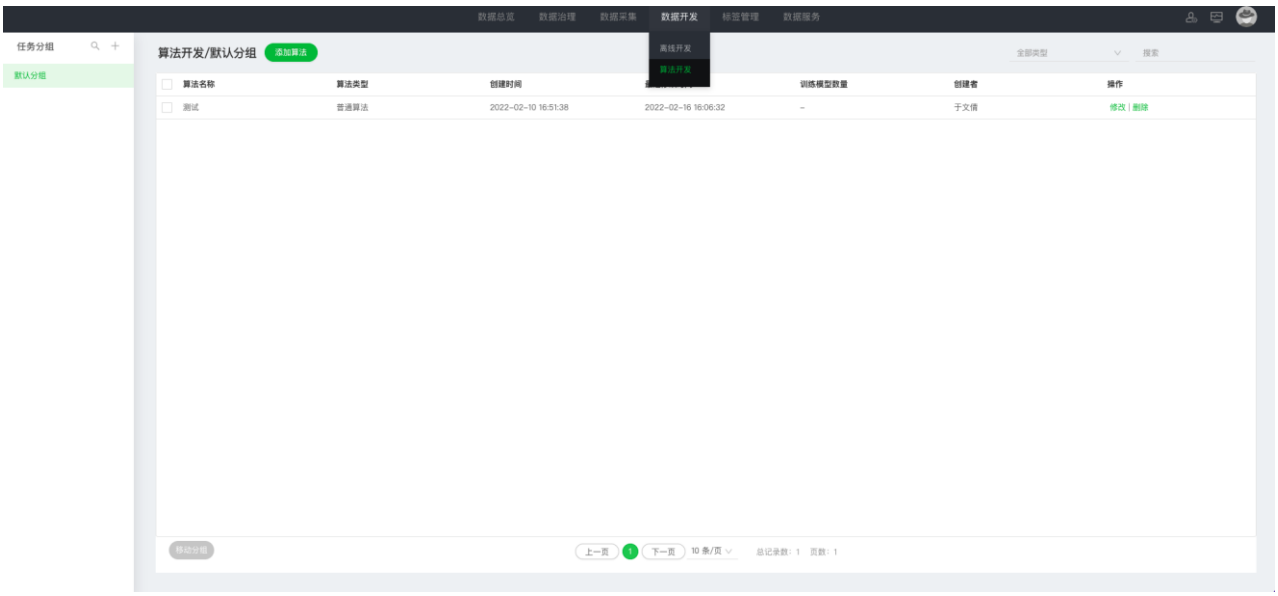
二、加工任务列表操作

在加工任务列表中，可以对添加完成的加工任务进行查看、复用、修改、删除操作。

2.4.2 算法开发

算法模块支持用户自定义算法，Data Formula 目前支持的算法脚本语言为 Python。用户可以将自定义的 Python 算法引入 Data Formula，对数据进行加工处理，也可以使用 Data Formula 中的数据对机器学习（AI）算法进行训练优化。

模块支持的算法包括：普通算法、机器学习算法。



一、添加算法

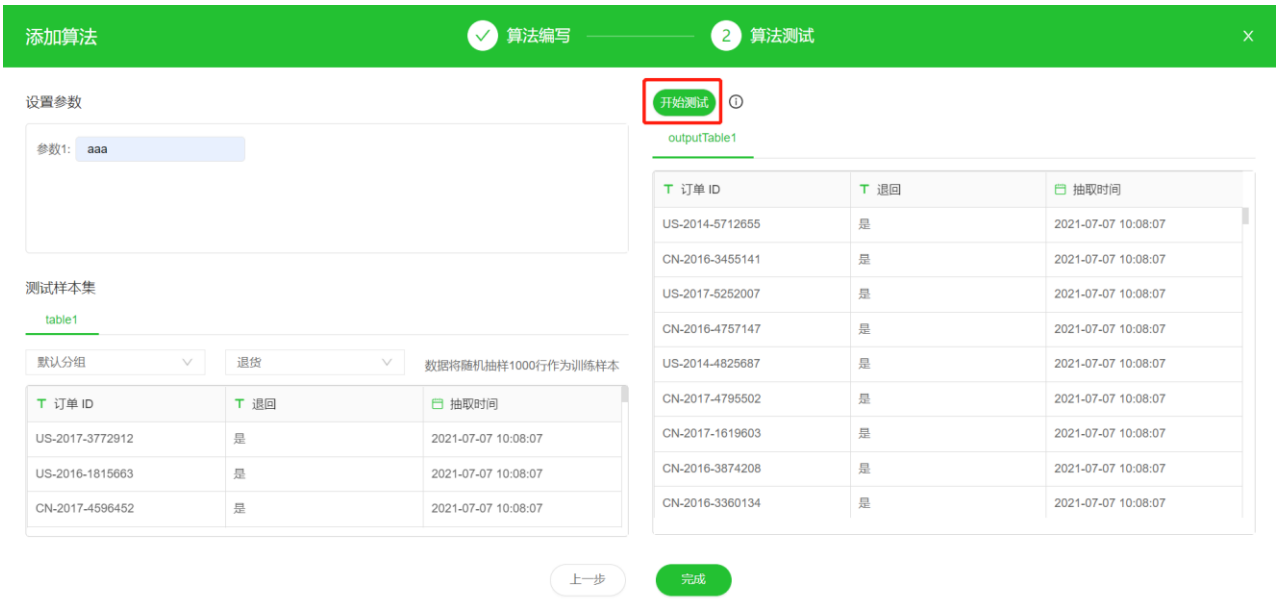
（1） 添加普通算法

普通算法的添加分为两个步骤：算法编写、算法测试。

点击“添加算法”，选择“普通算法”，跳转到【算法编写】页面。

在【算法编写】页面，录入算法脚本，输入【设置参数】、【设置输入数据集】、【设置输出数据集】，点击“下一步”，跳转到【算法测试】页面。【设置参数】输入与算法脚本对应的参数。【设置输入数据集】要与算法脚本中的输入数据集名称一致。【设置输出数据集】要与算法脚本中的输出数据集名称一致。具体格式见下图：

在【算法测试】页面，输入【参数】，选择【测试样本集】，然后点击“开始测试”，系统在执行完算法脚本后，将执行结果在右侧显示出来。【测试样本集】要选择系统中已有的数据集。在选择时，先选择数据集所在分组，再选数据集。



（2）添加机器学习算法

机器学习算法的添加分为三个步骤：训练模型编写、算法编写、算法测试。

点击“添加算法”，选择“机器学习算法”，跳转到【训练模型编写】页面。

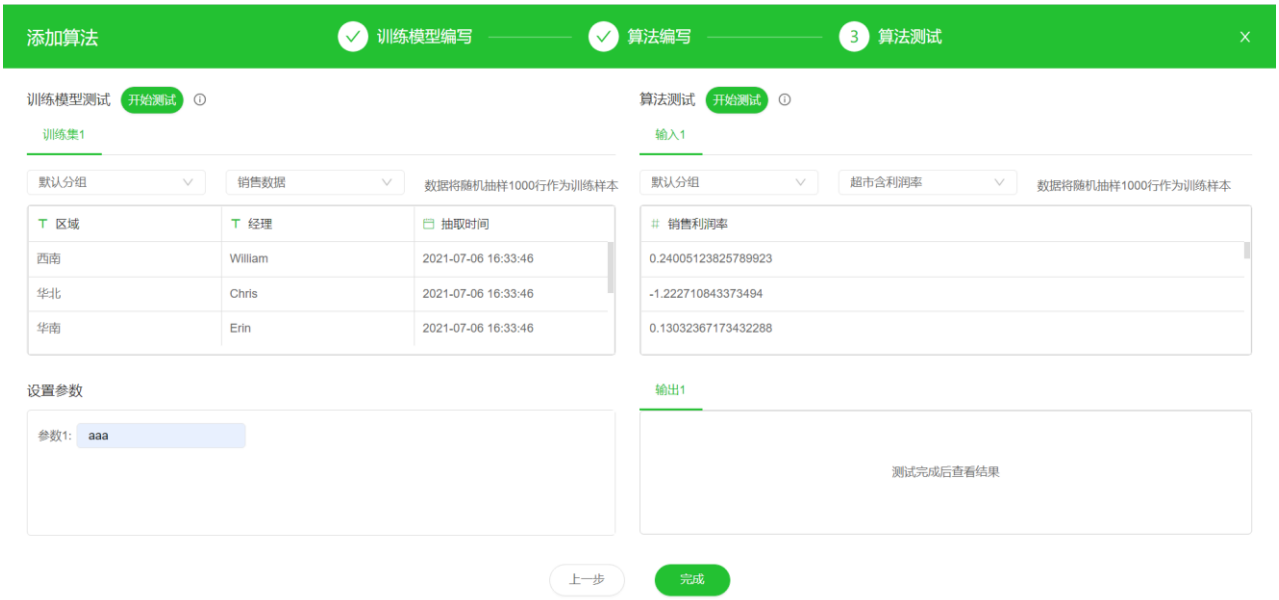
在【训练模型编写】页面，录入训练模型脚本，输入【设置训练集】、【设置参数】，

点击“下一步”，跳转到【算法编写】页面。【设置训练集】需要与训练模型脚本中的保持一致。【设置参数】输入与机器学习算法脚本对应的参数。具体格式见下图：

在【算法编写】页面，录入机器算法脚本，输入【设置输入数据集】、【设置输出数据集】，点击“下一步”，跳转到【算法测试】页面。【设置输入数据集】要与机器算法脚本中的输入数据集名称一致。【设置输出数据集】要与机器算法脚本中的输出数据集名称一致。

在【算法测试】页面的左侧【训练模型测试】模块中，选择【测试样本集】，输入【参数】，然后点击“开始测试”，系统在执行完训练模型脚本后，显示执行完成。然后在右侧

的【算法测试】模块中，选择【测试样本集】，然后点击“开始测试”，系统在执行完算法脚本后，将执行结果在右下角【输出】里面显示出来。



二、算法列表操作

在【算法列表】页面，可以对算法进行搜索、修改、删除操作。

2.5 标签管理

2.5.1 标签开发

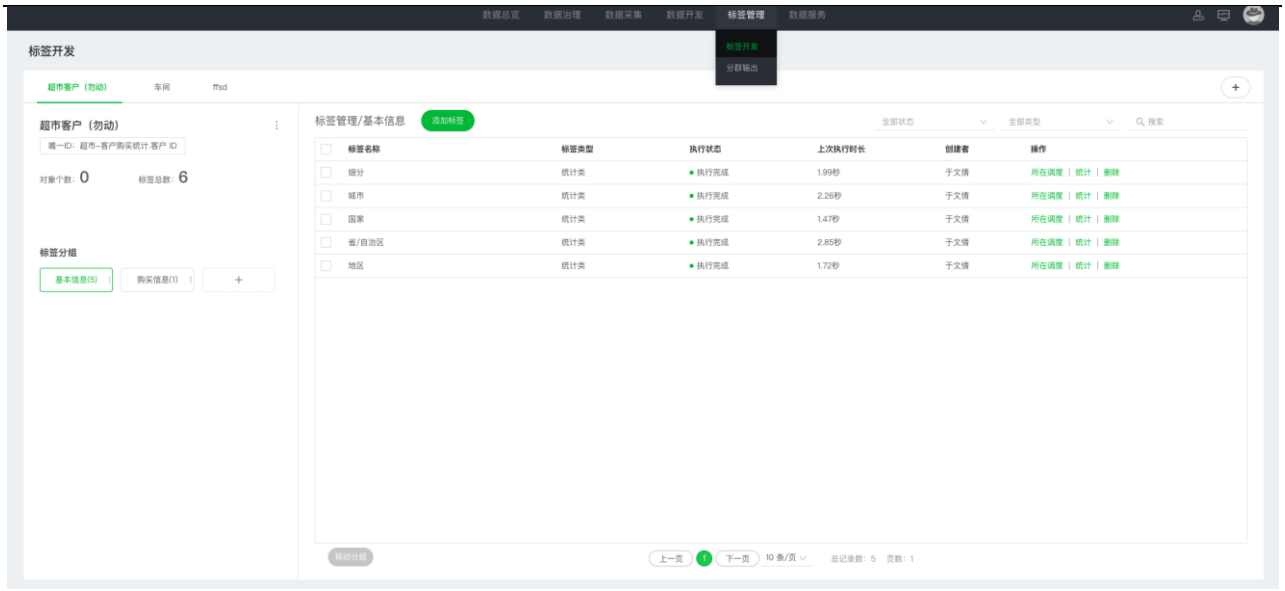
通过此模块，用户可以使用标签进行业务数据标签化。

标签的作用：某主体中单个个体的关键词描述，便于快速查找和定位。

标签使用场景：

a）根据不同的业务需求，将已有的主体信息，通过计算、分析等，获得新的关键词，便于业务推荐、营销等。

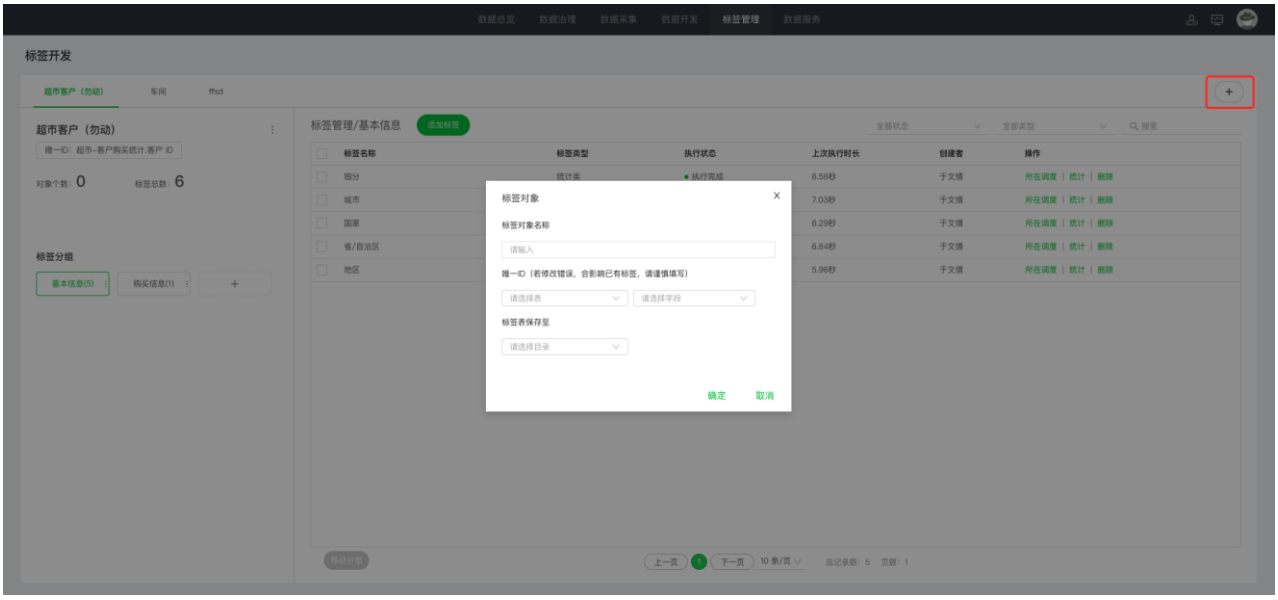
b）根据已有的目标群体标签特征，去全量群体中，查找有相似特征的用户群体，用于潜在群体挖掘。



一、创建标签对象

标签对象是指被打标签的对象。要创建标签，需要先指定标签对象，然后在标签对象上面创建标签。

点击右侧的“+”号，弹出【标签对象】弹窗。输入【标签对象名称】、【唯一 ID】，点击“确定”，标签对象创建完成。



二、创建标签

Data Formula 系统中包含四种标签类型，分别是统计类标签、规则类标签、SQL 计算

类标签、算法类标签。

统计类标签：是最为基础的标签类型，该类标签构成了用户画像的基础，是用户分析和数据分析最常用的指标及分类规则。

规则类标签：该类标签基于用户行为及确定的规则产生。

SQL 计算类标签：通过 SQL 脚本计算得出的标签。

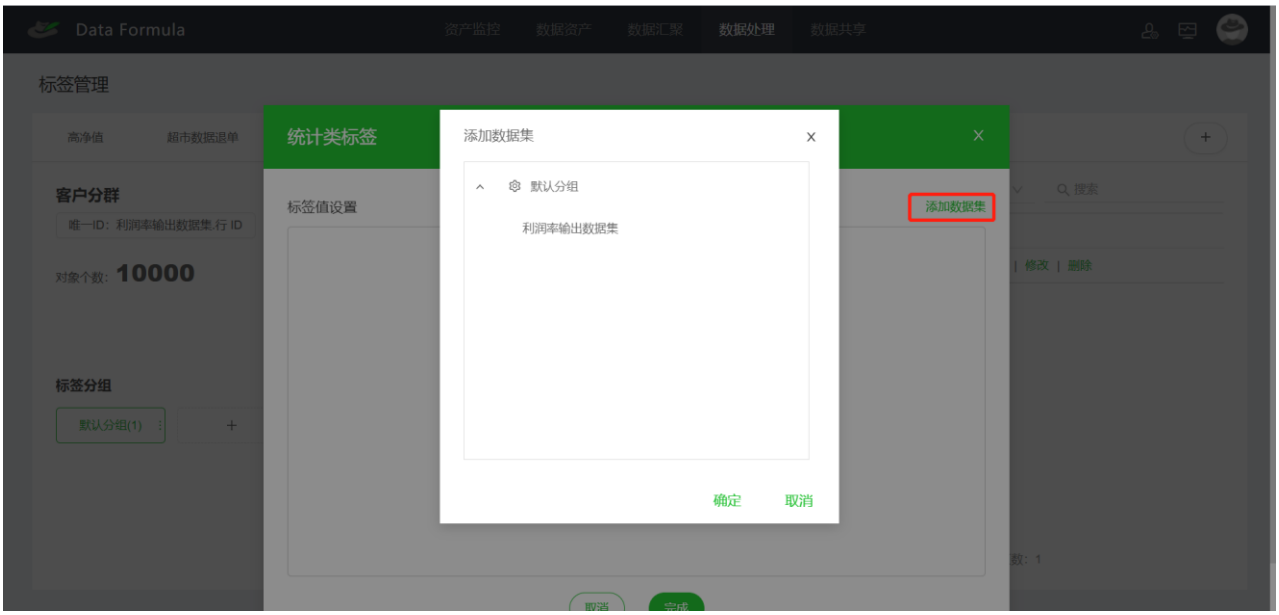
算法类标签：通过算法程序的运算得出的标签，可以对用户的某些属性或者某些行为进行预判。

选中一个创建好的标签对象，然后在右侧点击“添加标签”。

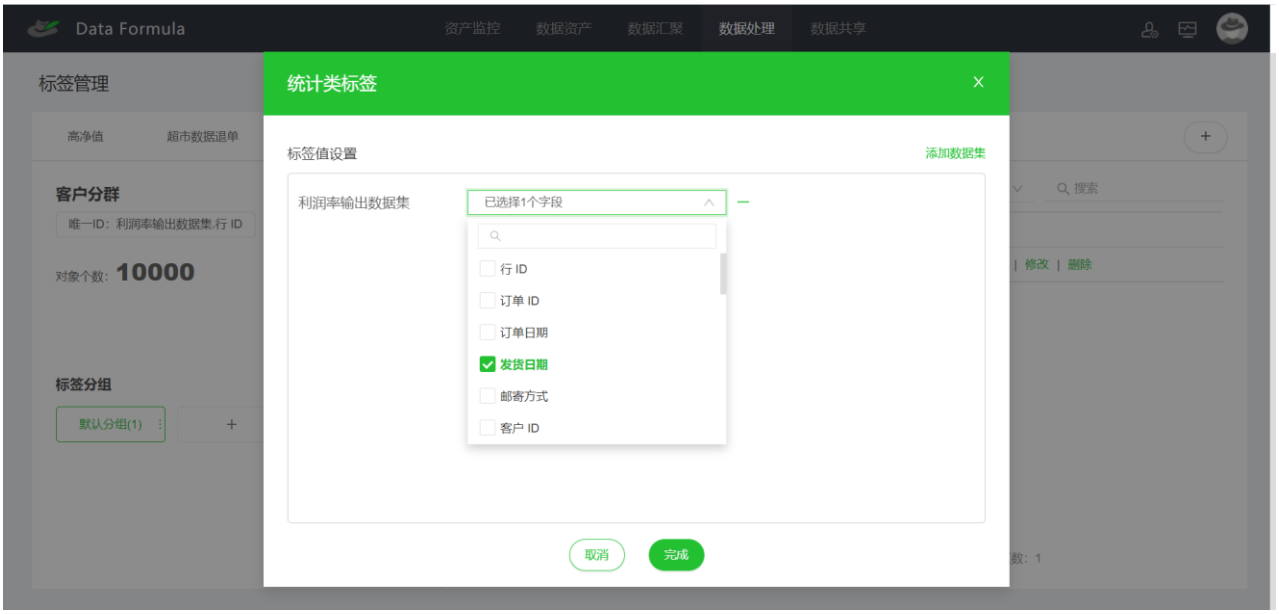
（1）统计类标签

在弹出的标签类型选项中，选择【统计类】，系统弹出【统计类标签】页面。

在【统计类标签】页面，点击“添加数据集”，在【添加数据集】弹窗中选在对应的数据集，然后点击“确定”。添加数据集完成。



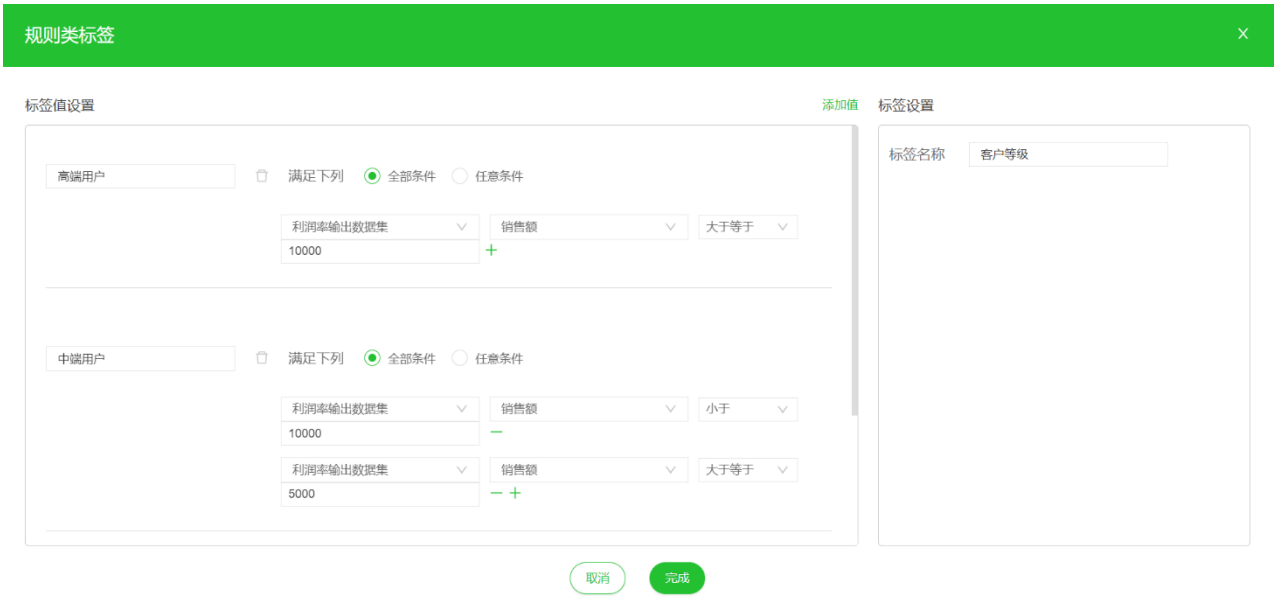
在添加的数据集中选择标签字段，支持多选，选中后点击“完成”，统计类标签添加完成。



（2）规则类标签

在弹出的标签类型选项中，选择【规则类】，系统跳转到【规则类标签】页面。

在【规则类标签】页面输入标签名称、标签值、满足条件，点击“完成”，规则类标签添加完成。系统支持添加多个标签值和多个满足条件。



（3）SQL 计算标签

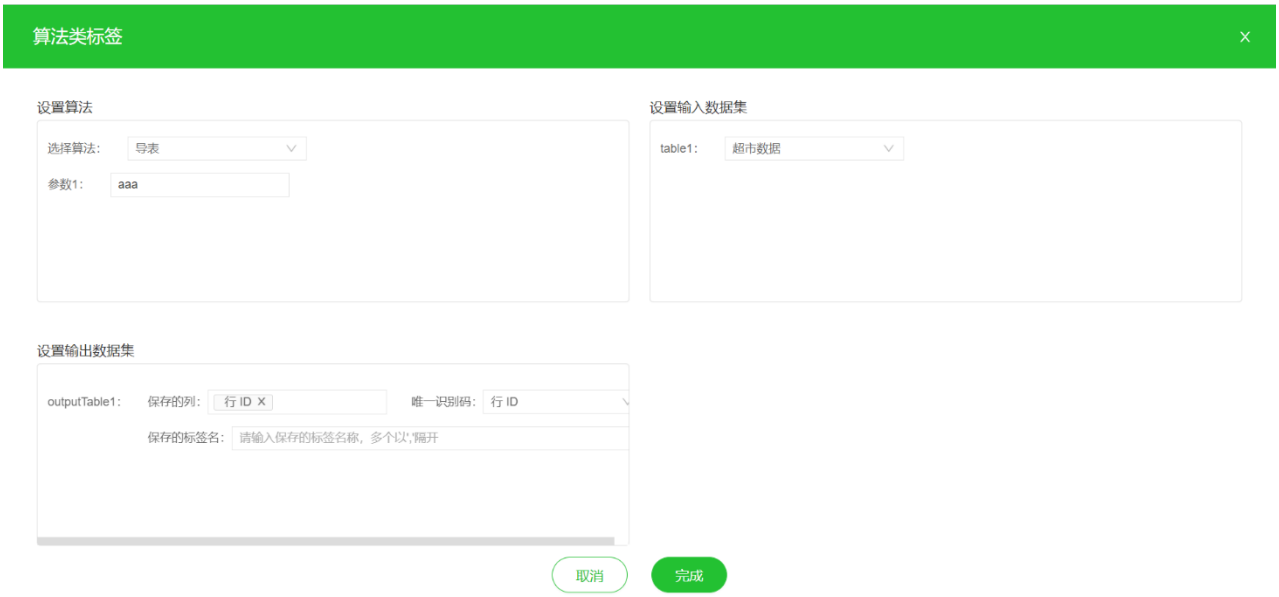
在弹出的标签类型选项中，选择【SQL 计算】，系统跳转到【SQL 计算标签】页面。

在【SQL 计算标签】页面输入 SQL 脚本，在右侧输入脚本对应的标签名称、唯一识别码、保存的列，点击“完成”。【唯一标识码】要填入数据集唯一标识字段的名称。【保存的列】要填入查询结果字段的名称。

（4）算法类标签

在弹出的标签类型选项中，选择【算法类】，系统跳转到【算法类标签】页面。

【算法类标签】页面包括三个设置区域，分别是：设置算法、设置输入数据集、设置输出数据集。设置算法区域：在【算法开发】中创建的算法会显示在【选择算法】下拉框中，选择算法后，在输入算法对应的参数。设置输入数据集区域：选择算法对应的输入数据集。设置输出数据集区域：选择保存的列、唯一标识码、保存的标签名。然后点击“完成”。

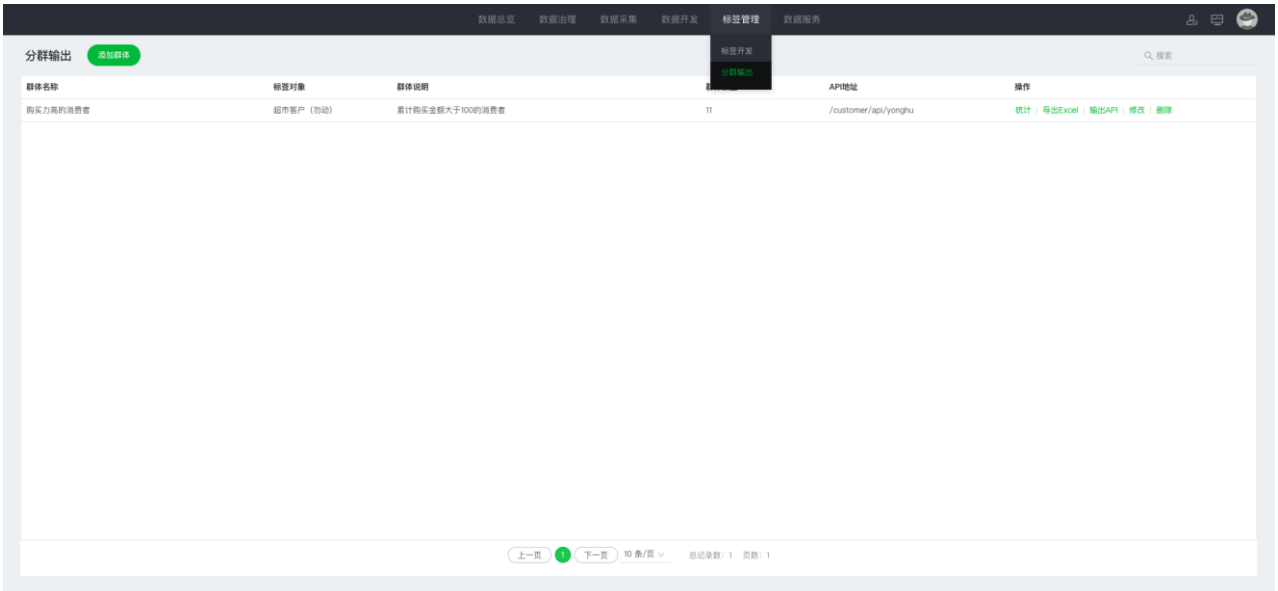


三、标签列表操作

点击单个标签对象，能看到该标签对象下面创建的标签列表。在标签列表中，可以进行标签的统计、修改、删除操作。

2.5.2 分群输出

客户分群是将【标签开发】模块中创建的客户标签进行分群管理，并将分群后的数据输出。

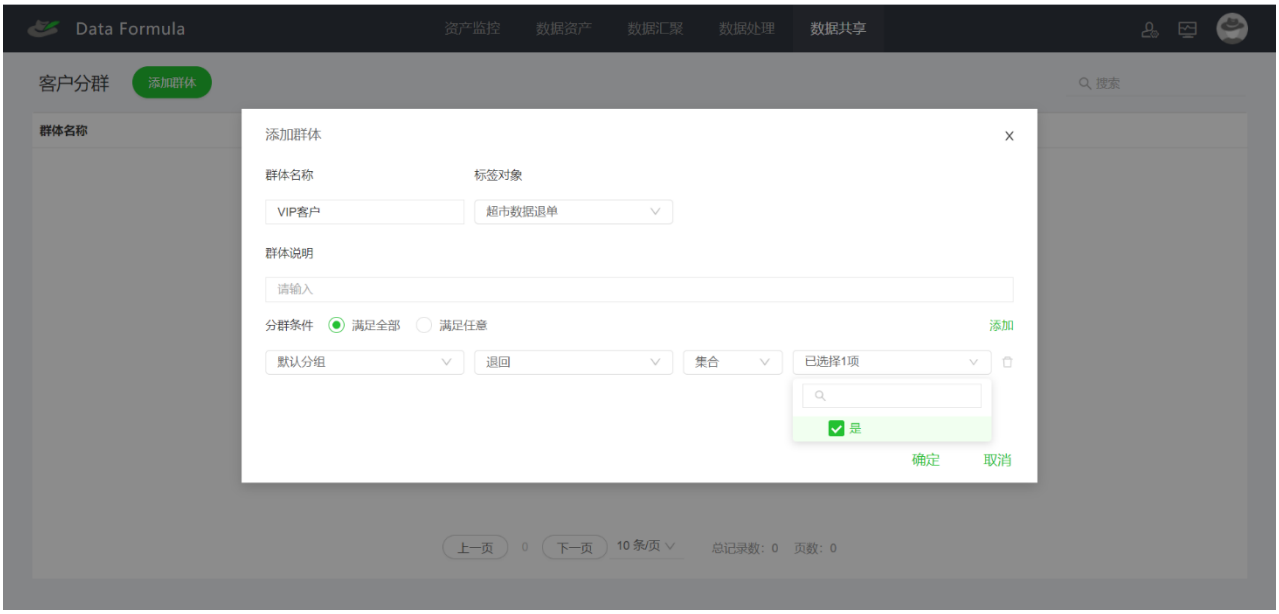


一、添加群体

点击“添加群体”，弹出【添加群体】弹窗。

在【添加群体】页面，输入群体名称、选择标签对象、输入群体说明、配置分群条件，

然后点击“确定”，群体创建完成。【标签对象】是从标签管理中已创建的标签对象中选择。【群体说明】可以为不填。【分群条件】是从标签管理中已创建的标签中选择过滤条件。系统支持添加多个分群条件，多个分群条件之间可以选择【全部满足】或者【满足任意一个】。

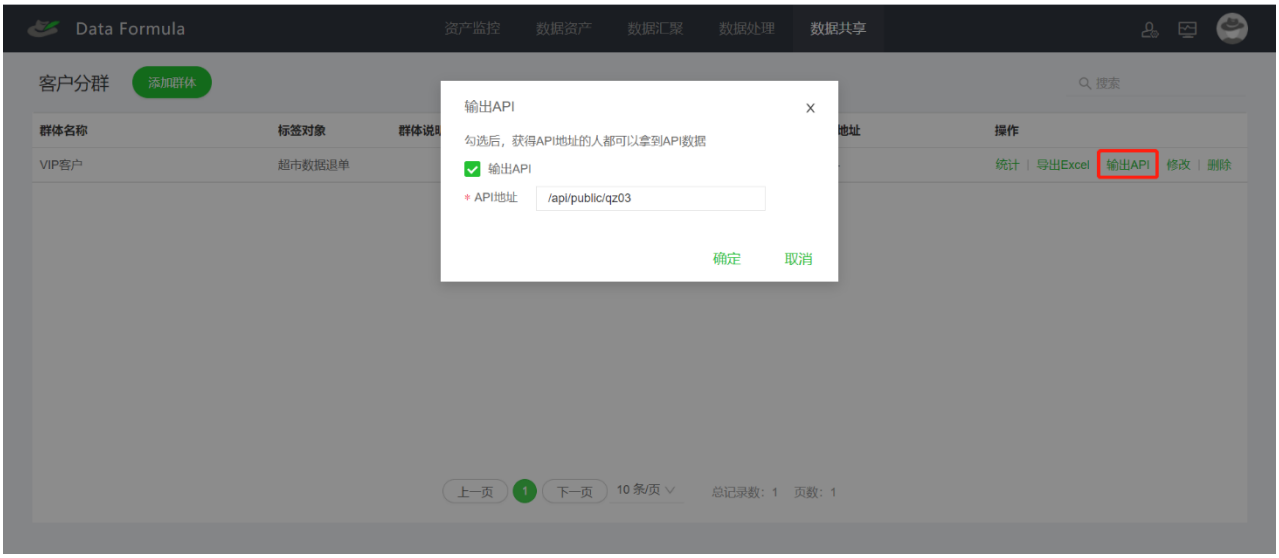


二、群体列表操作

添加完成群体之后，自动跳转到【群体列表】页面，在该页面可以对单个群体进行统计、导出 Excel、输出 API、修改、删除操作。

（1）输出 API

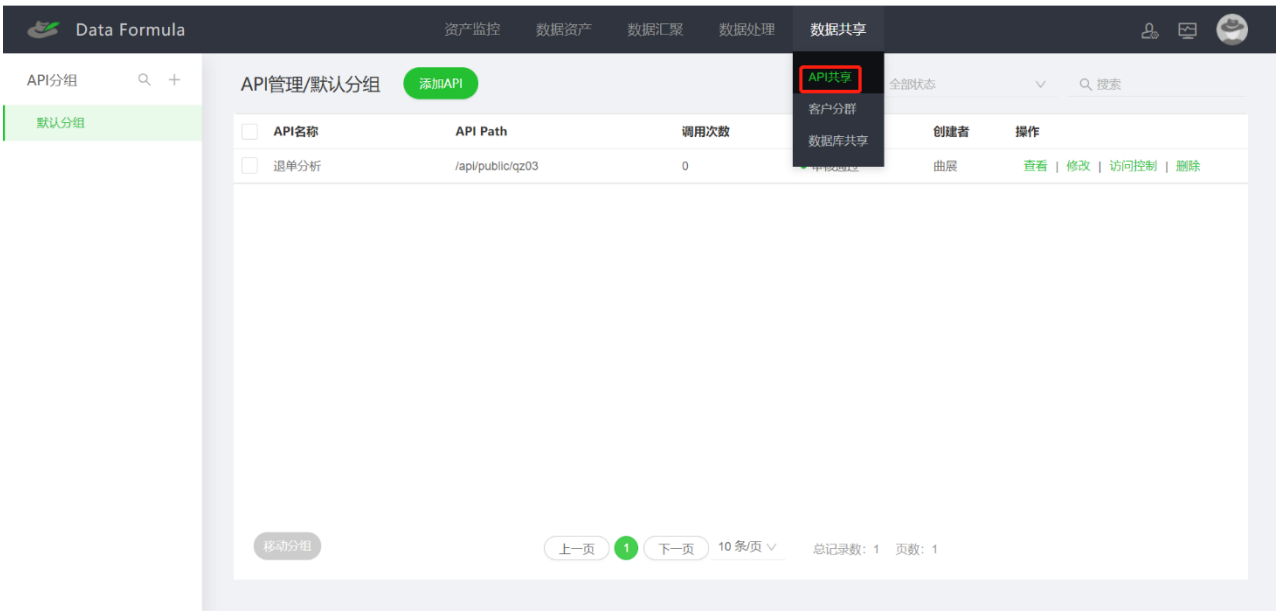
在【群体列表】页面，点击“输出 API”，弹出【输出 API】弹窗。在弹窗中输入 API 地址，点击“确定”，系统提示“输出 API 成功”。



2.6 数据服务

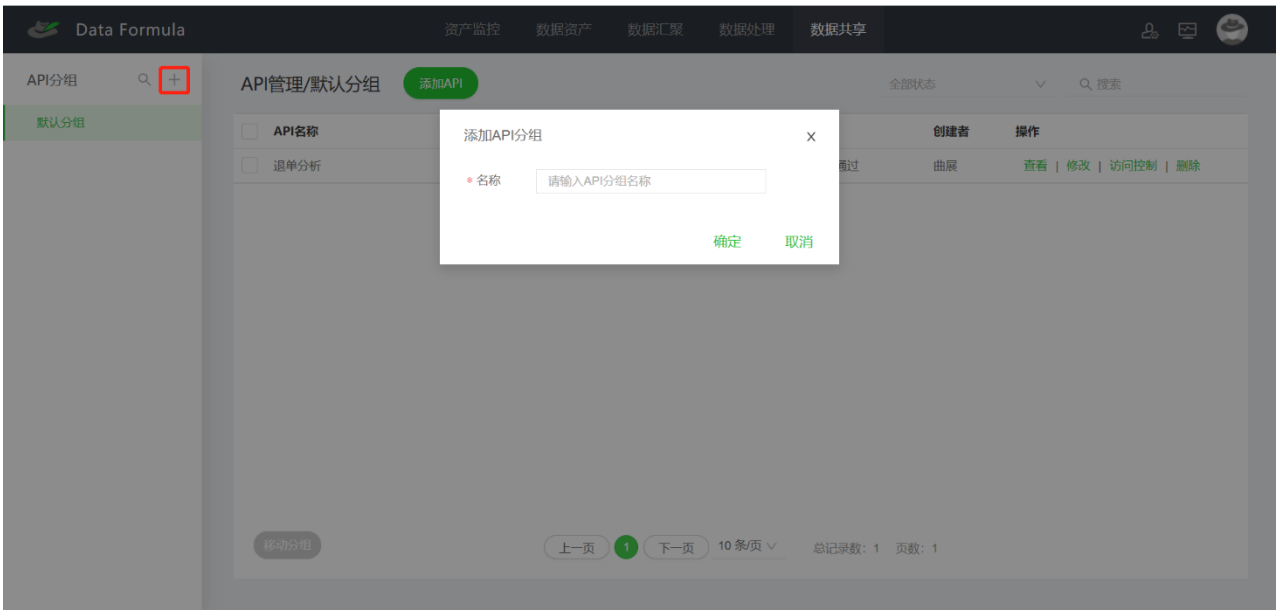
2.6.1 API 共享

Data Formula 系统加工完成的数据集，可以通过 API 共享，供其他业务系统读取和使用。



一、API 分组

点击左侧“+”号，弹出【添加 API 分组】弹窗，输入分组名称，可以创建一个 API 分组。API 分组，相当于 API 目录结构，可以在不同的 API 分组中创建 API。



二、添加 API

添加 API 共分为三个步骤，分别是：API 基础信息配置、API 参数配置、API 测试。

先选择一个分组，然后点击“添加 API”，跳转到【API 基础信息】页面。

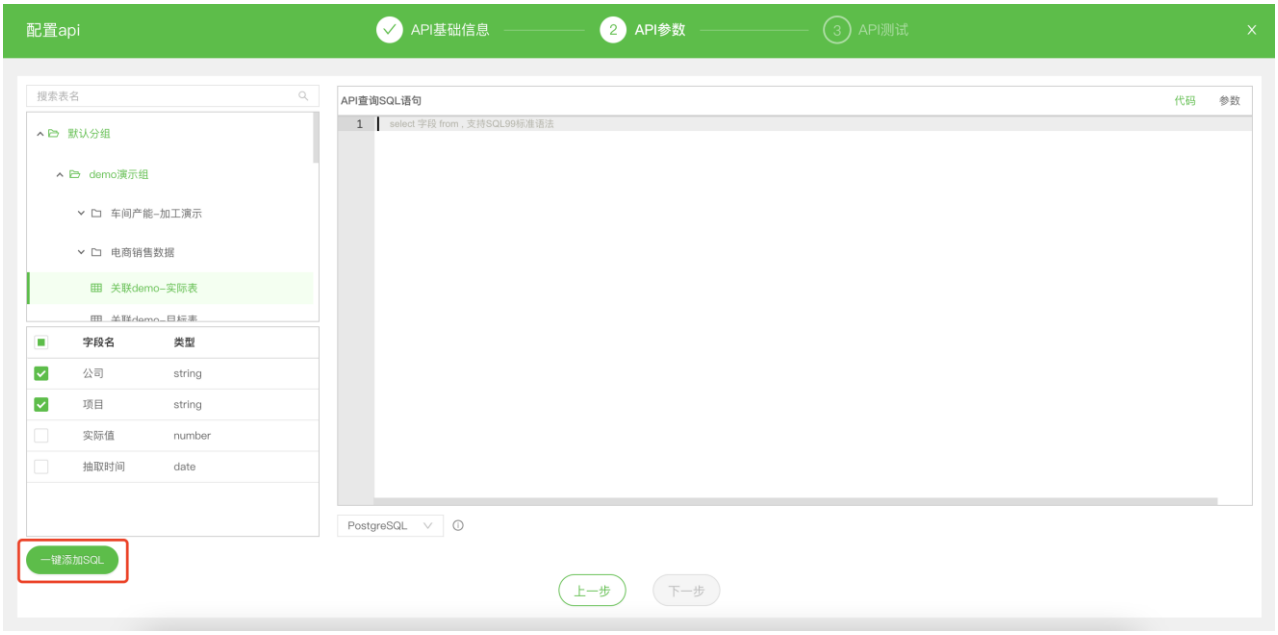
（1）API 基础信息设置

在【API 基础信息】页面，输入 API 名称、API path、请求方式、返回类型、描述，点击“下一步”，跳转到【API 参数】页面。【API 名称】用于命名 API。【API path】必须以/开头。【请求方式】和【返回类型】建议采用默认选项。【描述】可以不填。

（2）API 参数设置

在【API 参数】页面，选择需要共享的数据集并选中希望共享的字段，此时可以在右侧【API 查询 SQL 语句】区域直接手动输入 SQL 脚本，也可以点击“一键添加 SQL”将系统

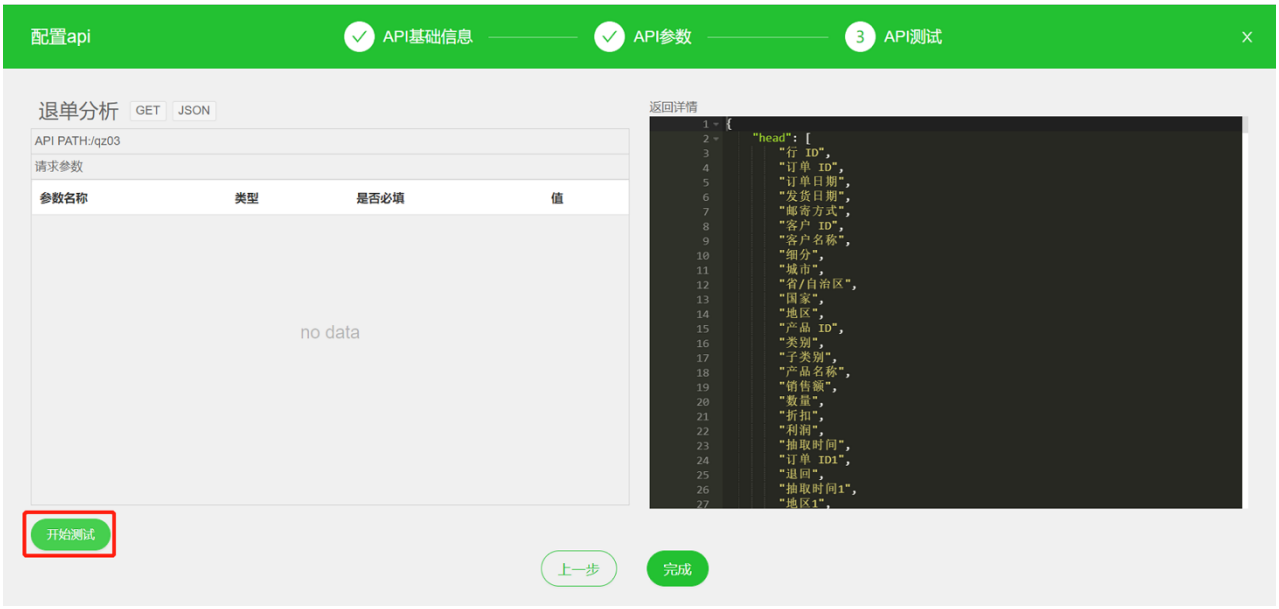
自动生成 SQL 脚本填充到【API 查询 SQL 语句】区域。SQL 支持 99 标准语法。



在【API 参数】页面，点击右侧的“参数”按钮，切换到参数展示，此时系统根据【代码】tab 页中的 SQL 脚本，自动填充返回参数字段，然后点击“下一步”，跳转到【API 测试】页面。

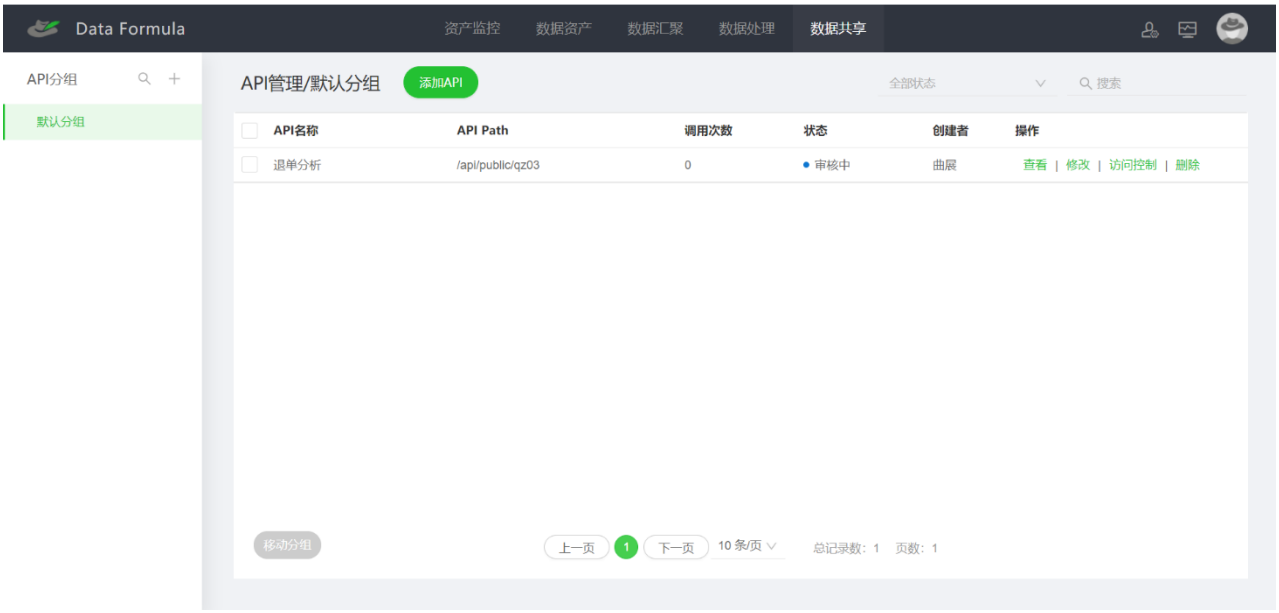
（3）API 测试

在【API 测试】页面，点击左下角“开始测试”，系统自动生成测试结果，并在【返回详情】中展示出来，若报错，会有报错提示。点击“完成”，API 创建成功。



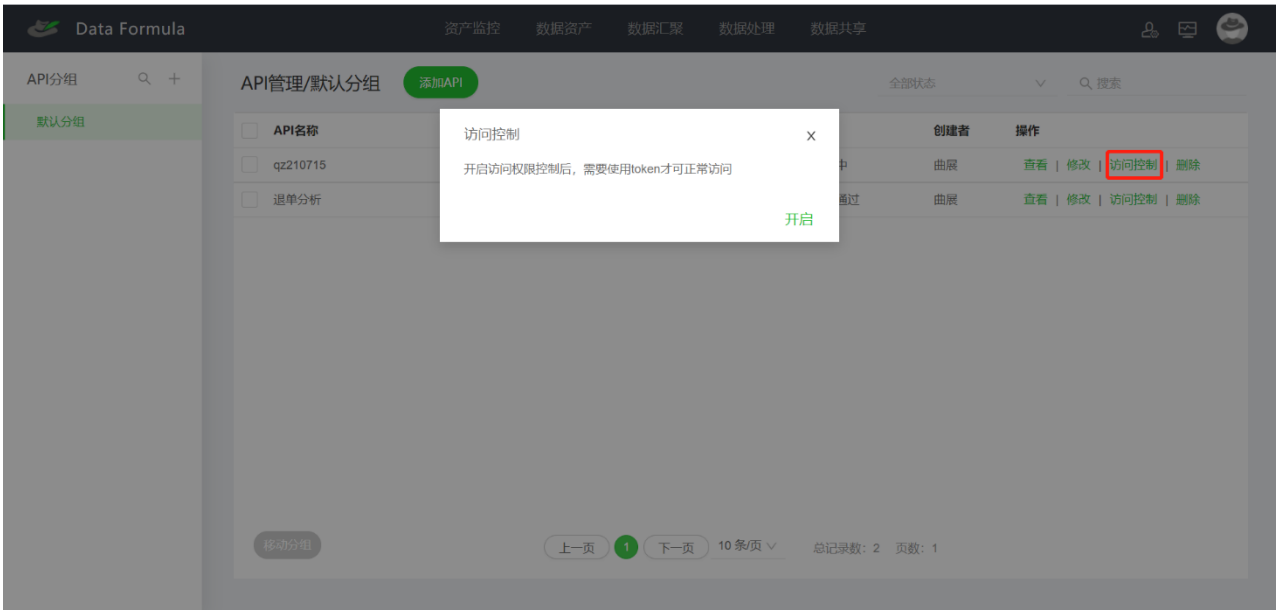
三、API 列表操作

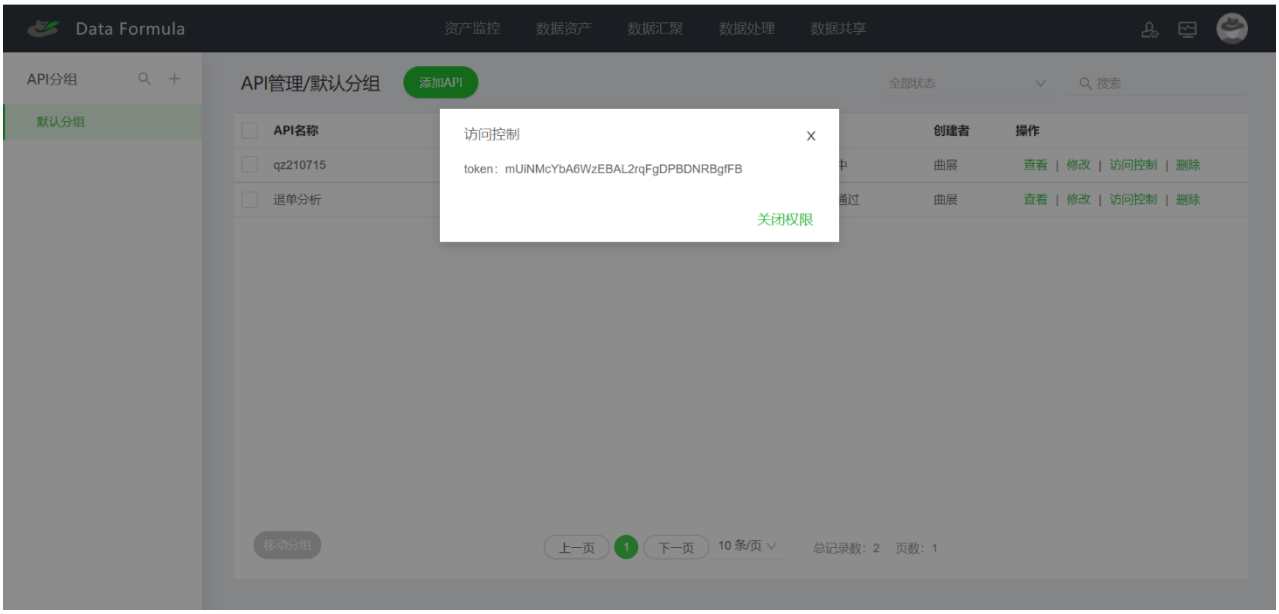
可以查看已创建的 API 列表，可以对 API 进行查看、修改、删除、访问控制、移动分组操作。



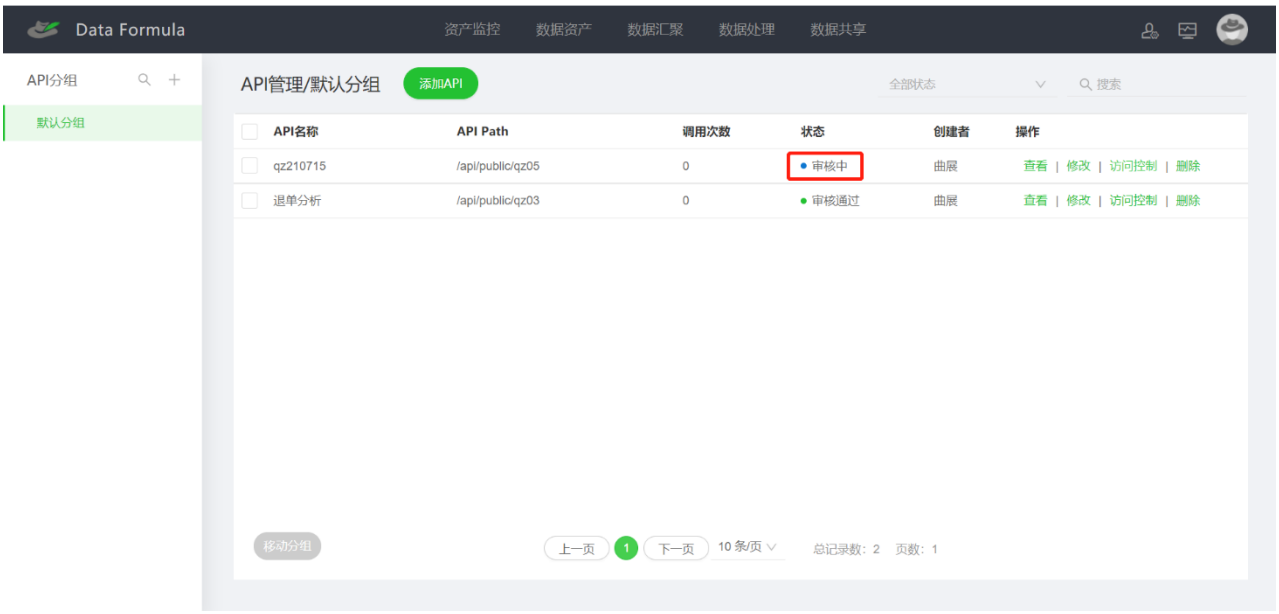
在 API 列表页面，点击单个 API 右侧的“访问控制”，弹出【访问控制】弹窗，点击弹窗右下角的“开启”，弹窗生成 token 字符串，此时点击“关闭”，则该 API 启用访问控

制，访问者需要提供正确的 token 才能访问该 API。





创建完成的 API，其状态变更为【审核中】，需要在【API 服务审计】模块中审核通过后，API 才正式生效。



2.6.2 数据库推送

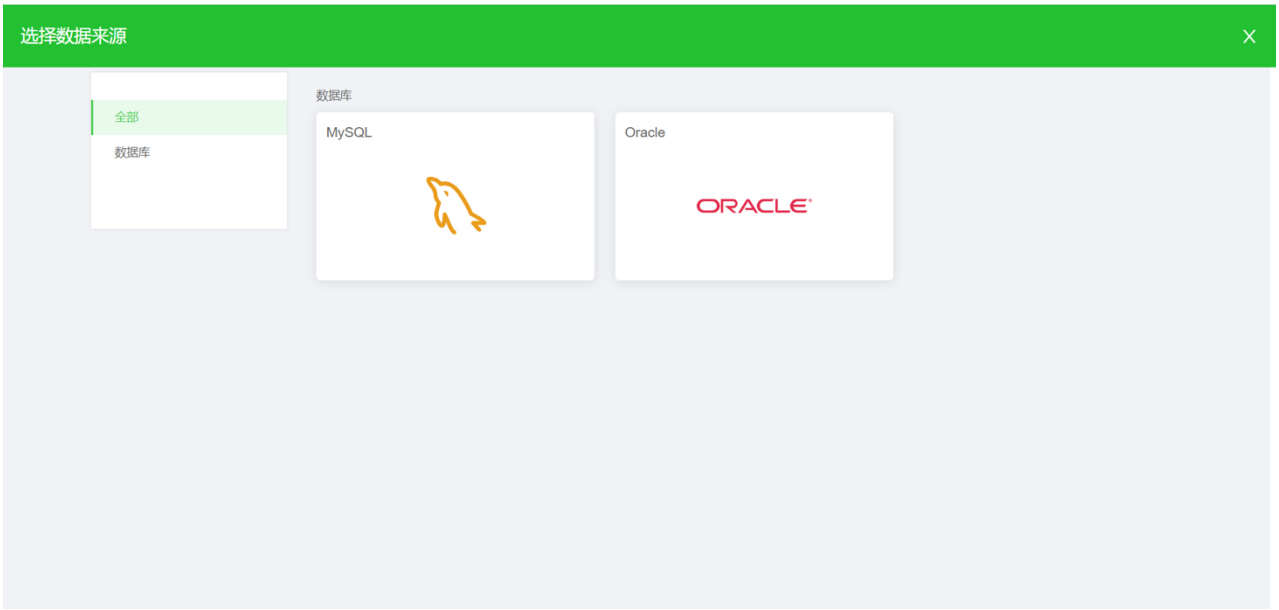
可以将 Data Formula 系统加工处理完的数据集写入数据库，供其他业务系统以访问数据库的方式使用数据。

一、添加共享任务

点击“添加共享任务”，跳转到【选择数据来源】页面。

（1）选择数据来源

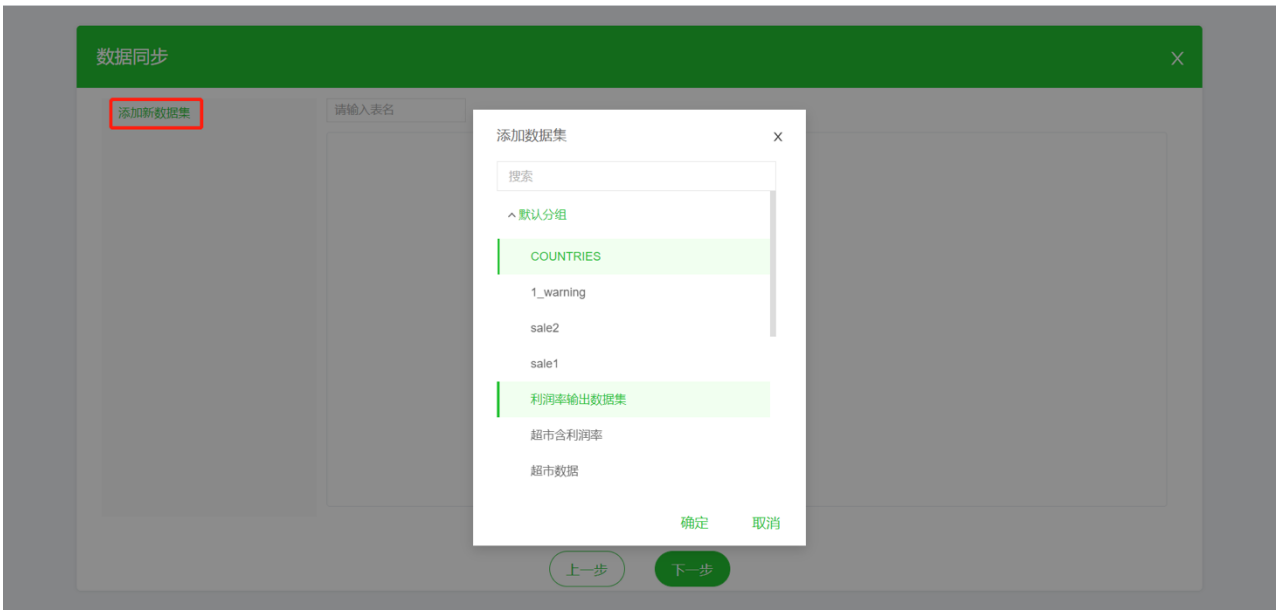
在【选择数据来源】页面，点击希望共享的数据库类型，弹出该数据库类型的连接信息配置页面。



（2）输入目标数据库信息

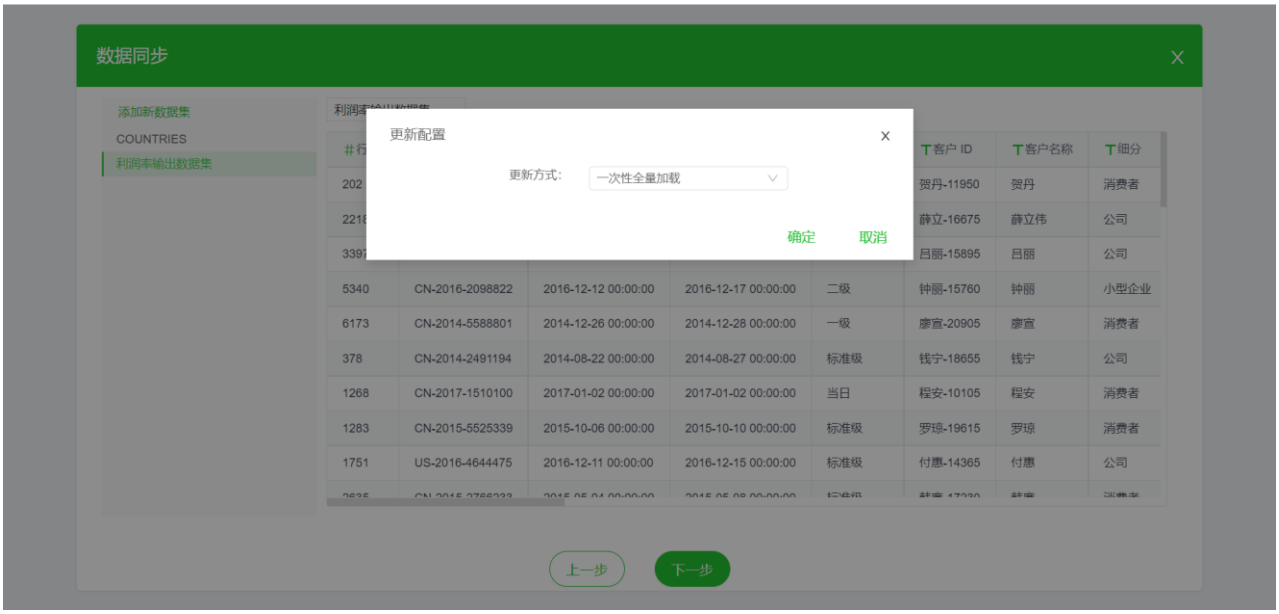
在【连接信息配置】页面，输入数据库别名、服务器、端口、数据库、用户名、密码，点击“连接”，若连接成功，跳转到【数据同步】页面。【数据库别名】、【服务器】、【端口】、【数据库】、【用户名】、【密码】需按照数据库信息填写。

在【数据同步】页面，点击左侧“添加新数据集”，弹出【添加数据集】弹窗，在弹窗中选择要共享的数据集，点击“确定”，页面将选中数据集的详细字段记录信息展示出来。此处支持一次选择多个数据集。



在【数据集展示】页面，点击“下一步”，弹出【更新配置】弹窗。

在【更新配置】弹窗中，选择更新方式，点击“确定”。

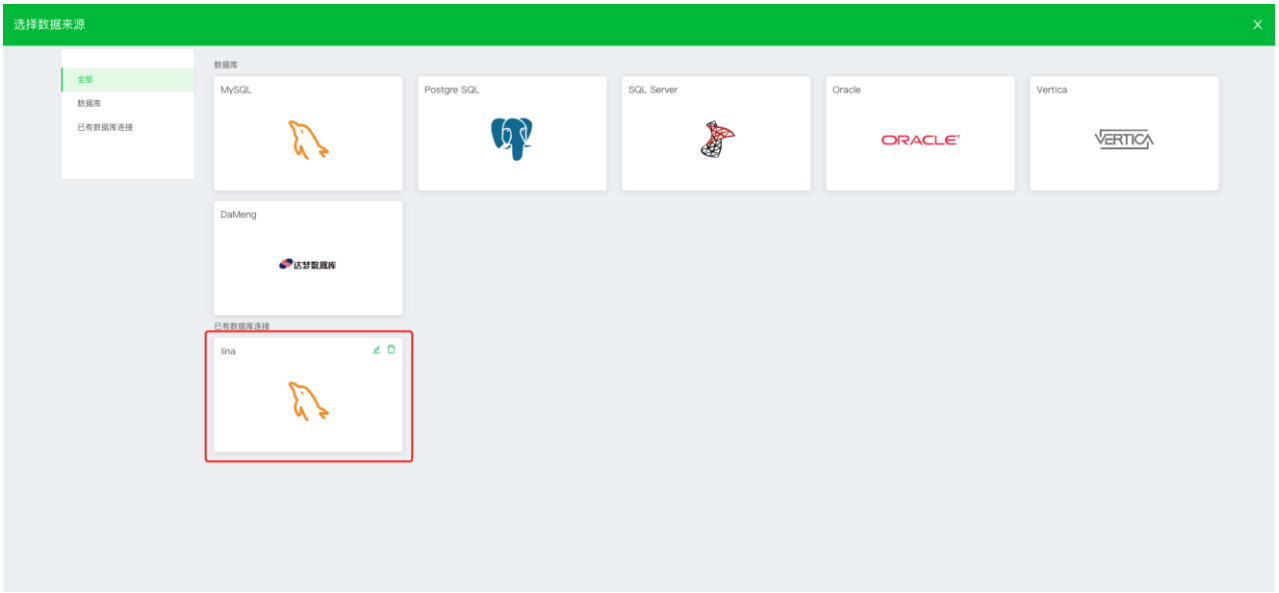


二、共享任务操作

添加成功的共享任务，会在数据共享列表中展示出来，选中单个数据共享任务，可以对该任务进行修改、删除、查看表详情的操作。

三、数据库连接操作

已连接的数据库，可修改连接信息或删除。

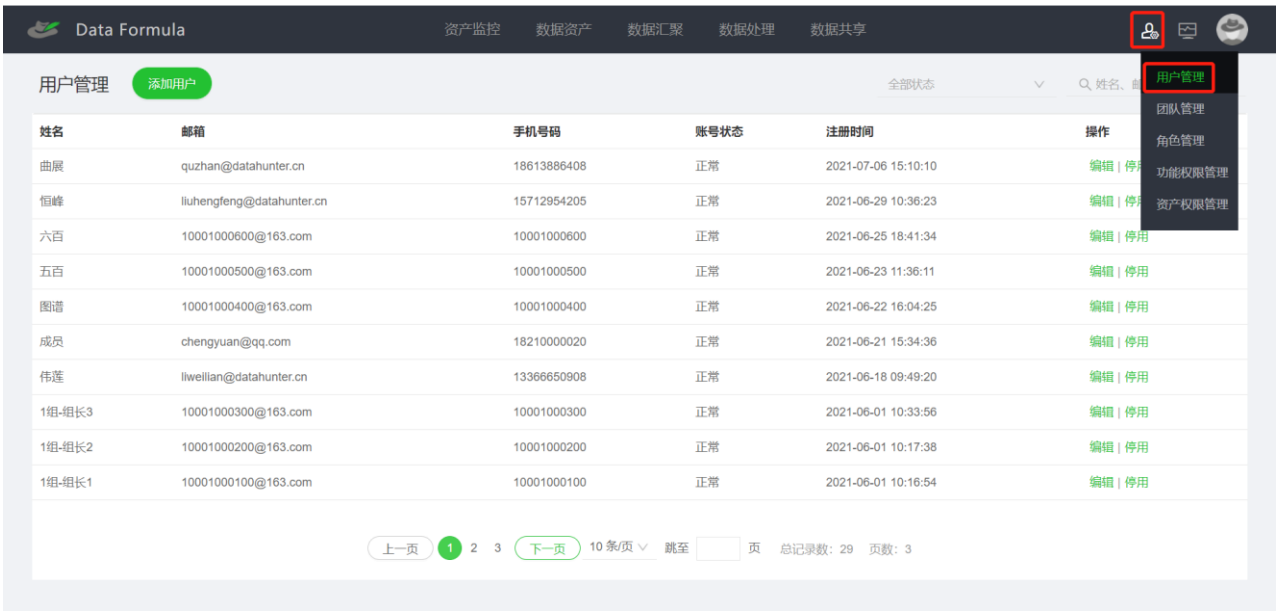


2.7 用户权限管理

2.7.1 用户管理

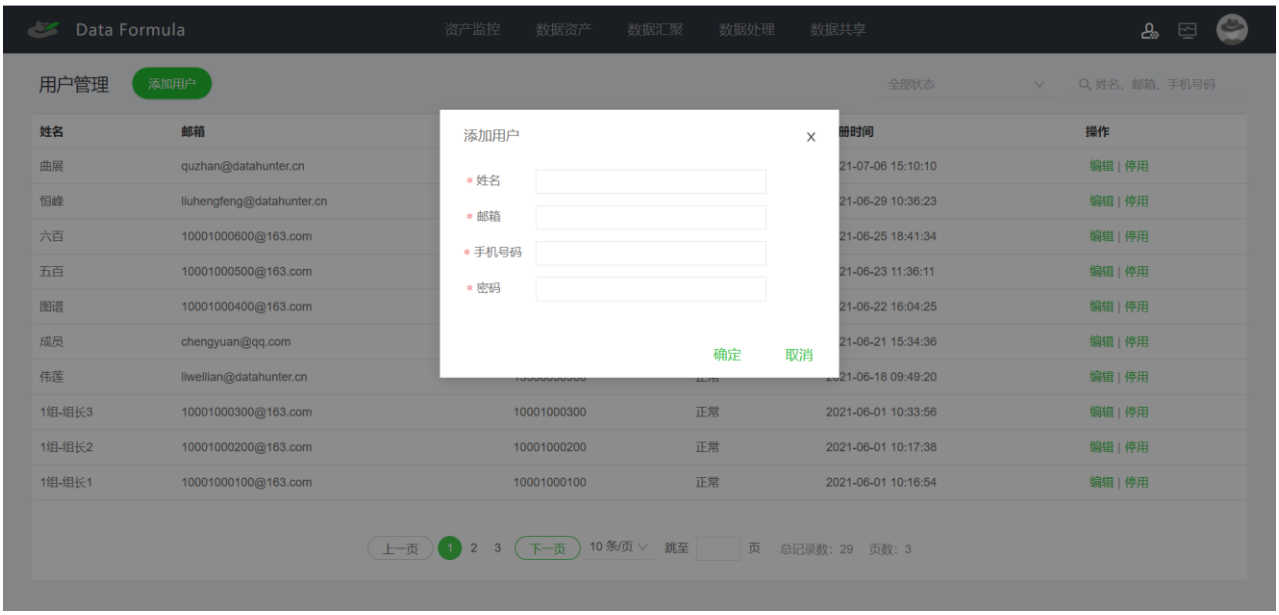
Data Formula 系统采用标准的用户角色权限管理模式。所有的功能、数据都可以根据角色/权限进行分配调整。

【用户权限管理】-【用户管理】模块支持添加、编辑、查看、停用/启用用户。



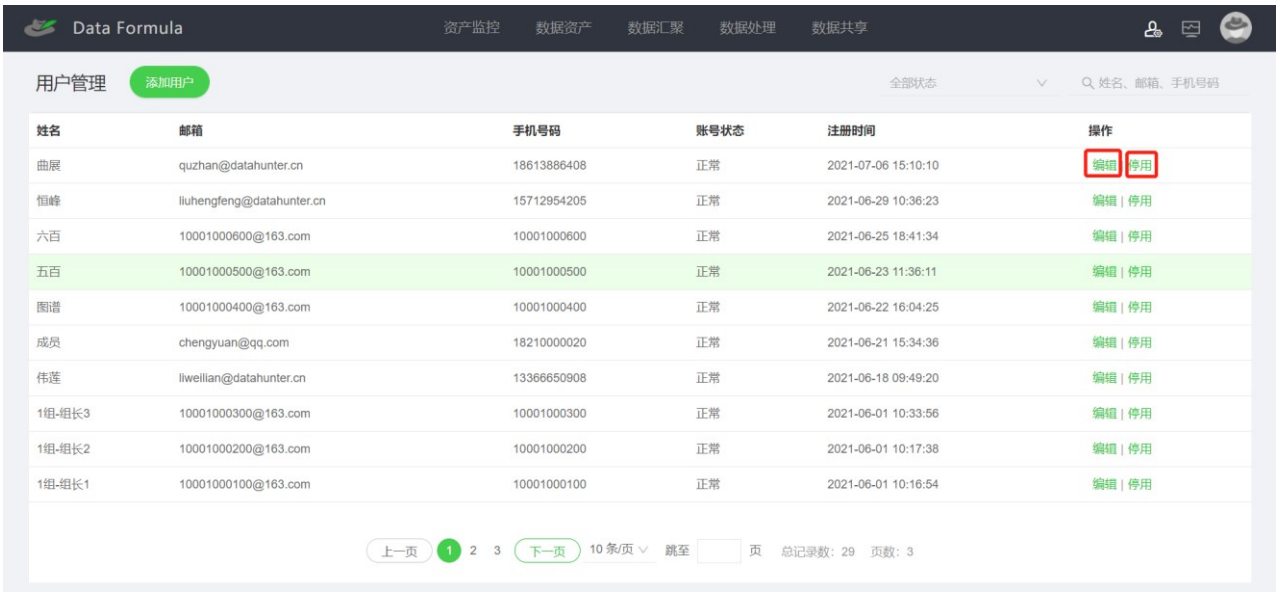
一、添加用户

在【用户权限管理】-【用户管理】页面，点击“添加用户”，弹出【添加用户】弹窗。在【添加用户】弹窗，输入姓名、邮箱、手机号码、密码，点击“确定”，添加用户完成。



二、用户列表操作

在【用户权限管理】-【用户管理】页面，单个用户的右侧，可以点击“编辑”对用户信息进行修改。可以通过点击“停用/启用”来控制用户权限。用户列表支持搜索和翻页操作。

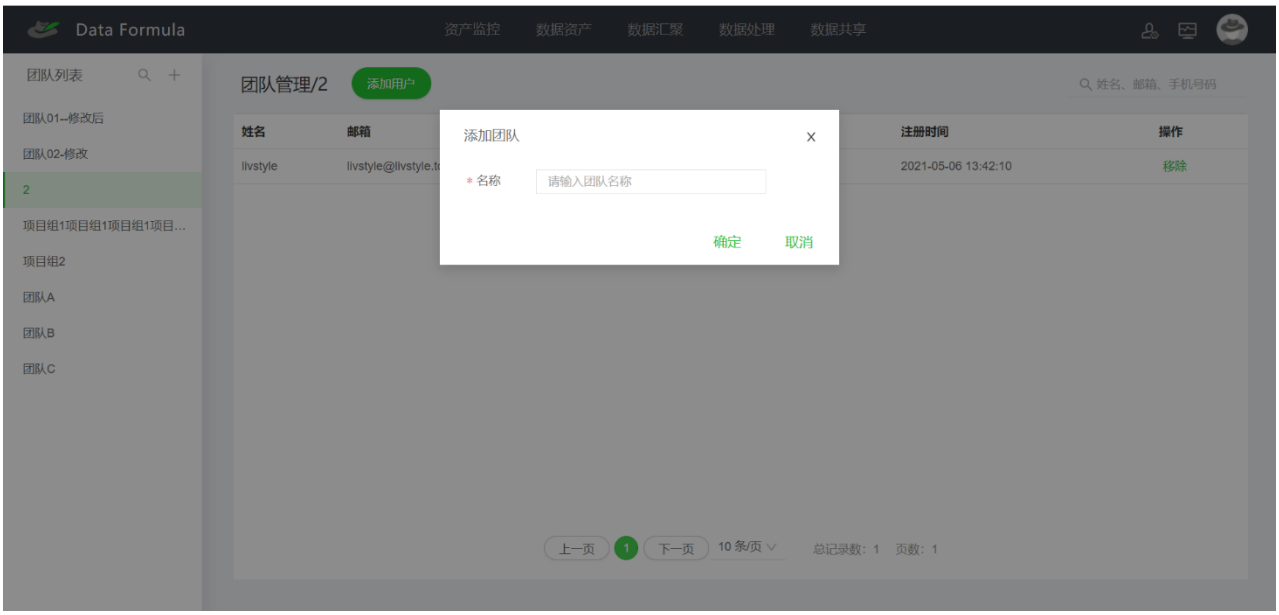


2.7.2 团队管理

【用户权限管理】-【团队管理】模块可以创建多个团队，并将用户划分到不同的团队，然后就可以按照团队来给用户分配不同的数据权限。

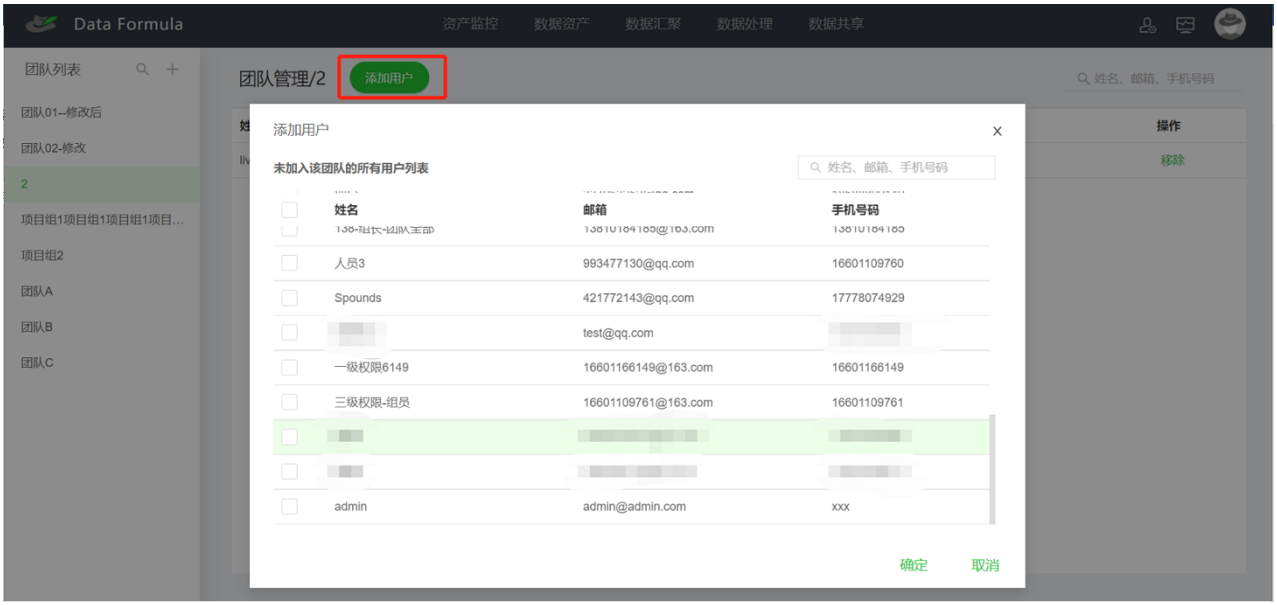
一、添加团队

点击页面左侧的“+”号，弹出【添加团队】弹窗，输入团队名称，点击“确认”。



二、为团队划分用户

在左侧团队的目录中选中一个团队，然后再点击“添加用户”，弹出【添加用户】弹窗，在弹窗用户列表中选择希望划入团队的用户，点击“确定”。



三、移除用户

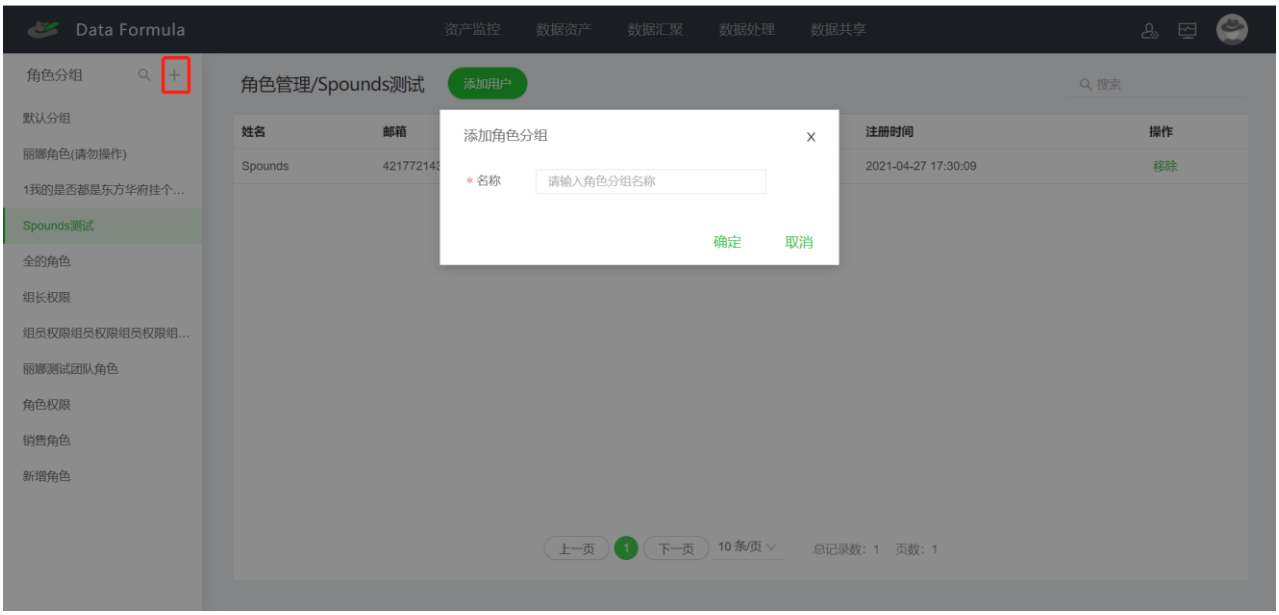
在左侧团队列表中选中一个团队，右侧会显示出该团队的所有用户，点击单个用户后面的“移除”按钮，可以将该用户移出团队。

2.7.3 角色管理

在【用户权限管理】-【角色管理】模块可以先创建角色分组，然后将用户添加到不同的角色分组，以达到为用户分配角色的目的。

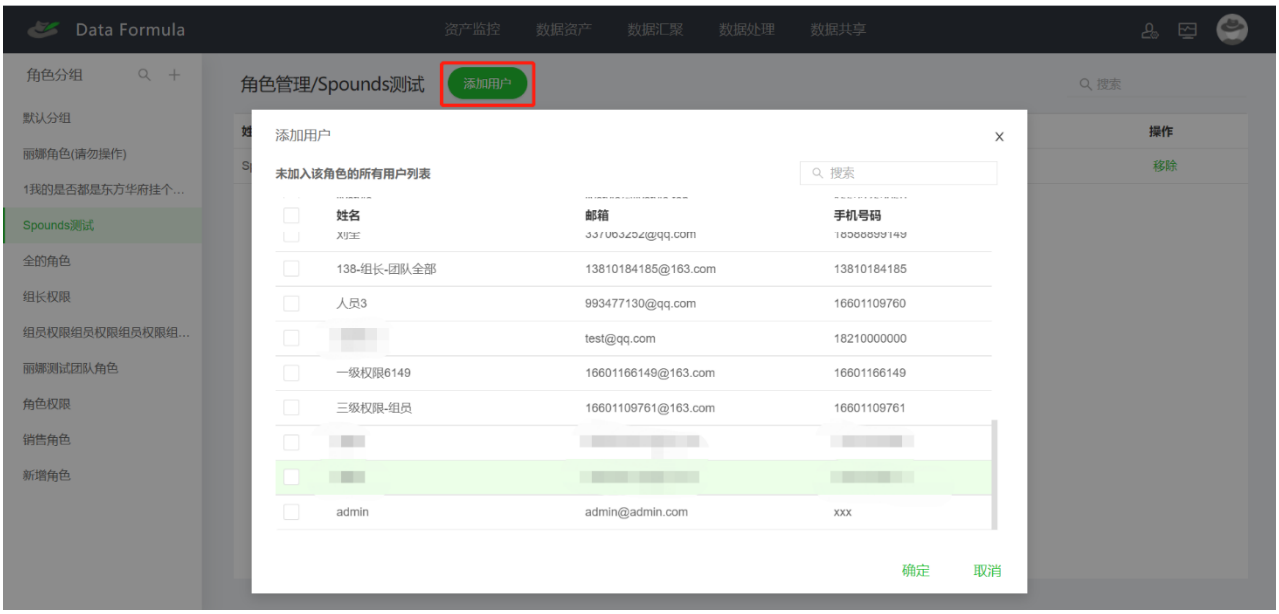
一、创建角色分组

点击页面左侧的“+”号，弹出【添加角色分组】弹窗，输入角色分组名称，点击“确认”。



二、为分组添加用户

在左侧角色分组目录中选中一个角色分组，然后再点击“添加用户”，弹出【添加用户】弹窗，在弹窗用户列表中选择希望分配该角色的用户，点击“确定”。

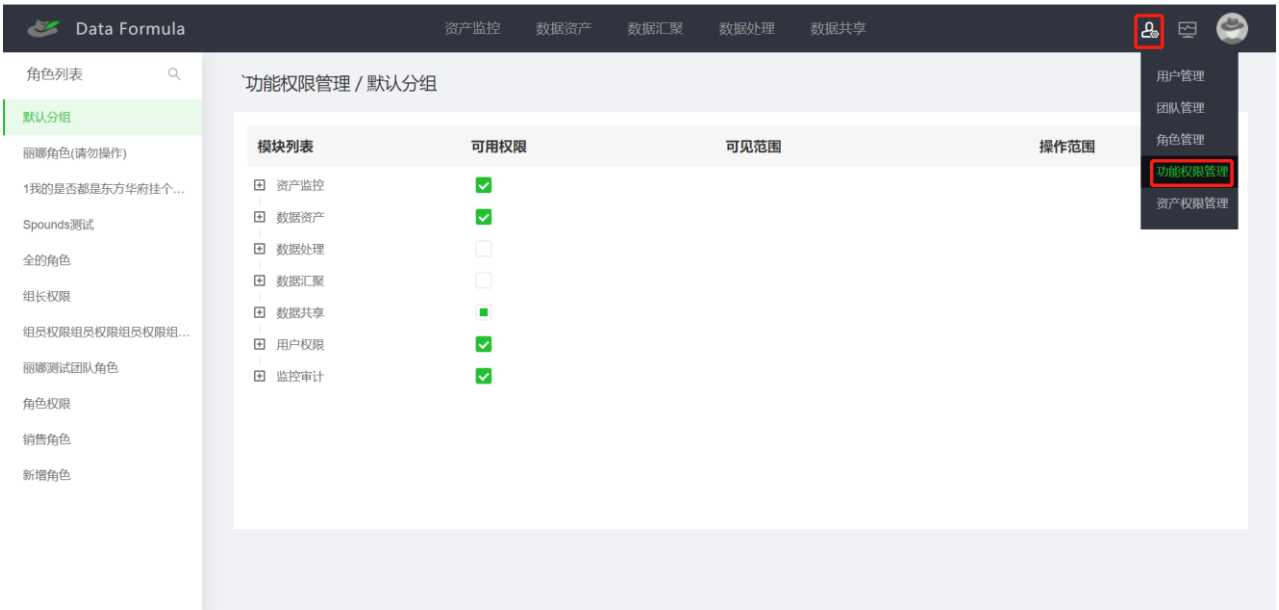


三、移除用户

在左侧角色分组目录中选中一个角色分组，右侧显示出该分组的所有用户，点击单个用户后面的“移除”按钮，可以将该用户移出角色分组。

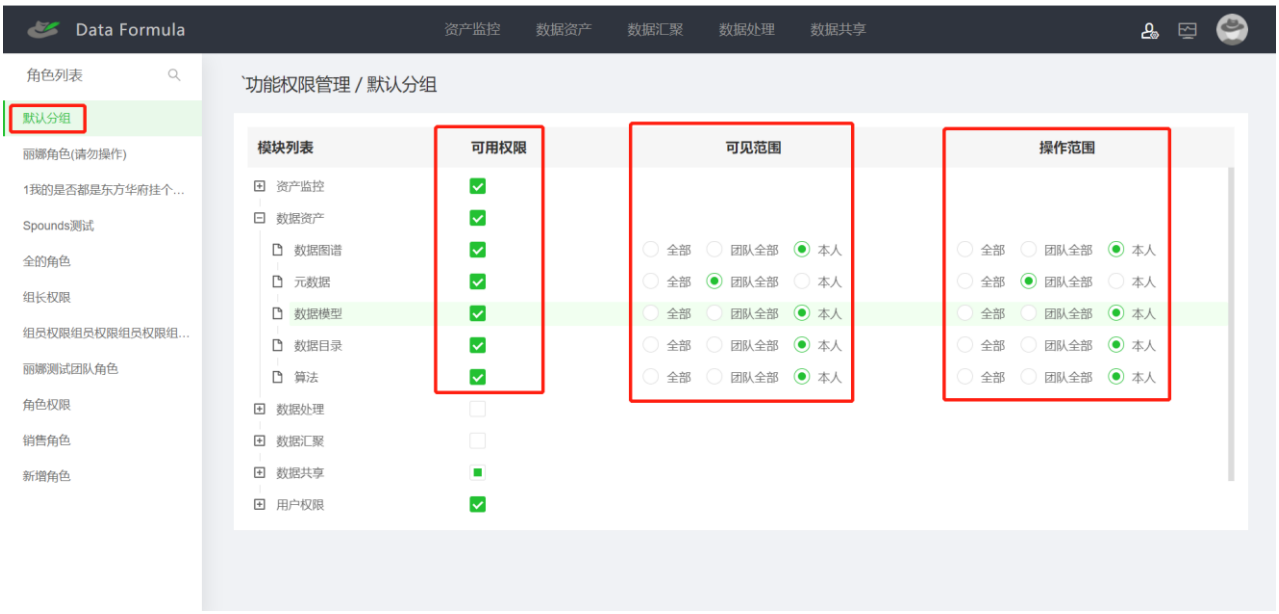
2.7.4 功能权限管理

【用户权限管理】-【功能权限管理】模块可以为角色分组，分配对应的功能权限。



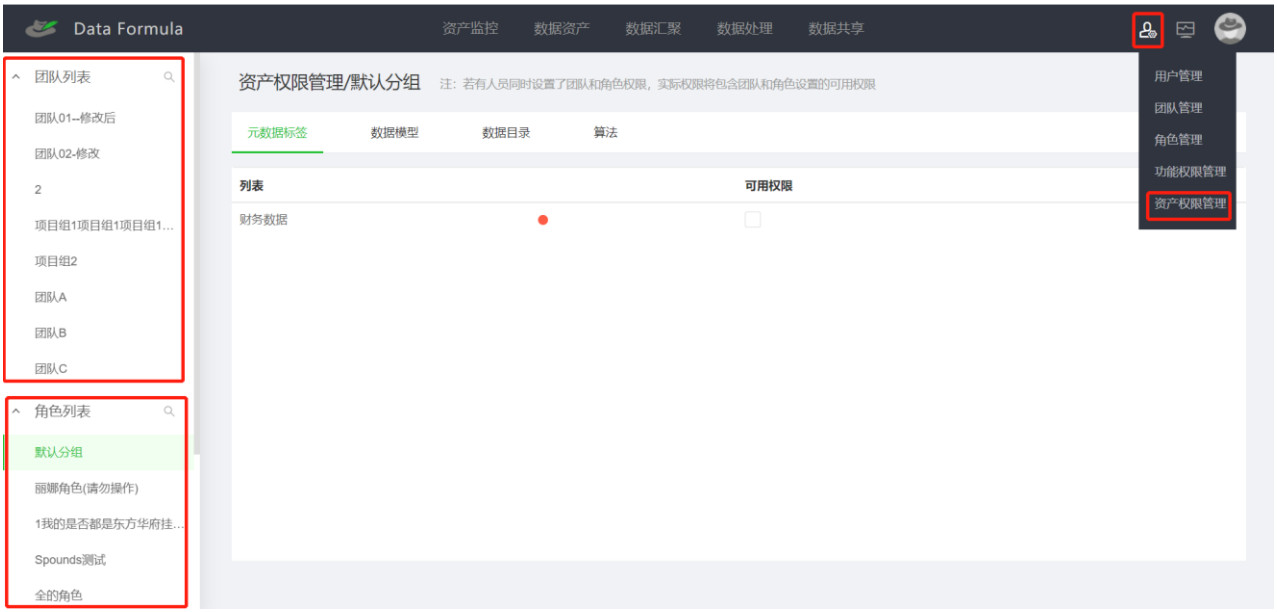
一、权限分配

点击左侧【角色分组目录】，页面右侧区域会将系统功能菜单以树形结构展示出来供用户勾选，并且可以针对每个功能设置可见范围、操作范围。



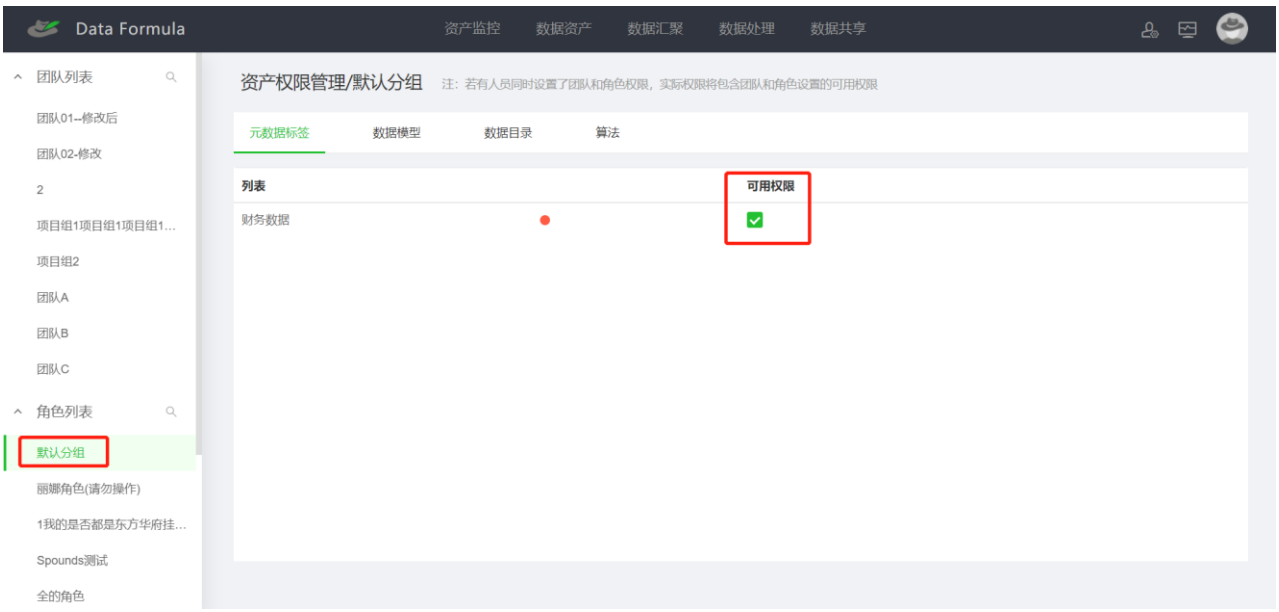
2.7.5 资产权限管理

【用户权限管理】-【资产权限管理】模块可以为团队或者角色分组，分配不同的数据权限。



一、资产权限分配

点击左侧【团队分组】或者【角色分组】，页面右侧展示出当前登录用户创建的数据资产信息，通过勾选方式，可以将当前登录用户创建的数据资产共享给相应的团队或者角色。



2.8 日志审计管理

2.8.1 API 服务审核

为了确保 Data Formula 系统数据的安全，在【数据共享】-【API 共享】中添加的 API 接口，必须通过 API 服务审核通过后才能对外共享。

点击屏幕右侧“API 服务审核”按钮，跳转到【API 服务审核】页面。该页面将【数据共享】-【API 共享】中创建的 API 接口进行展示，点击单个 API 右侧的“审核”按钮，审核通过。

该模块支持批量审核操作，可以通过选中多个 API，点击左下方的“批量审核”按钮实现。

2.8.2 操作日志

操作日志模块记录了用户在 Data Formula 系统中的所有操作记录，包括用户姓名、操作行为、操作内容、操作时间等。

点击屏幕右侧“操作日志”按钮，跳转到【操作日志】页面。