2019

“智采”系统

V8.0

目录

第1章　系统简介

“智采”系统，既可以用于科研支撑，帮助研究人员自主和高效的采集互联网中所需要的精确数据，构建自己所需要的研究数据库；也可以用于教学，培养相关专业的学生自主策划、采集和处理互联网上的大量数据，形成决策所需要的行业数据库，为科学的量化管理培养基于数据的管理和决策分析技能。

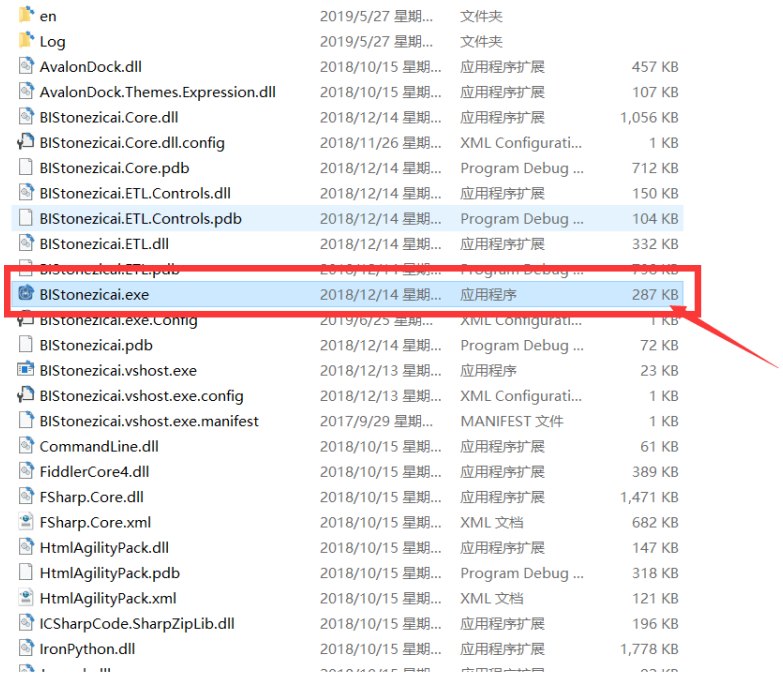
1.1　系统环境要求和安装部署

“智采”系统的设计初衷，是用于自动的采集、处理和提取网页类非结构化数据，以满足高等院校、科研机构对互联网信息进行周期性采集、监测和研究的使用要求。

既可以用于科研支撑，帮助研究人员自主和高效的采集和互联网中所需要的精确数据；也可以用于教学，培养相关专业的学生自主策划、采集和处理互联网上的大量数据，为科学的量化管理培养基于数据的管理和决策分析技能。

该系统支持 windows XP 到当前最新版本的 windows 操作系统，安装部署十分简单，需提前安装好.net framework4.5 版本，即可复制即运行。

下载压缩包，解压后，双击“BIStonezicai.exe”即可直接运行。



1.2　“智采”系统特点

“智采”系统作为一种数据采集和清洗的强大工具，有效地采集来自网页、数据库、文件的数据，并通过可视化地拖拽，快速地进行生成，过滤，转换等操作。
该系统的特点有如下几点：

1. 轻量级。

该系统是基于 windows 系统设计的轻量级信息监测服务系统，复制即用，无需安装，半天的培训即可上手，实现对于网页精确数据采集和提取的功用。

2. 功能强大。

该系统以工作流的方式，组织整个信息监测服务流程，内含 50 多种数据提取和清洗模块，覆盖了文本语义处理、数据存储、数据库投影等模块，变化无穷，功能强大。

3. 满足个性化采集要求。

大数据时代，对大数据的采集、分析和价值提取需求日益旺盛。人人都有个性化的数据采集需求。无需借助 IT 工程师的帮助，借助”智采”系统，即可快速实现对于个性化数据的采集。

注：附带案例。

一、 餐饮行业研究案例数据库

1. 北京餐饮门店评价数据库（来自大众点评网），包括：门店名称、评论、人均、类别、位置、地址、口味、环境、服务、优惠。
2. 北京餐饮门店评价数据的采集项目（XML 格式，可以基于该项目，变更为其他目标城市），包含四个子项目，可以自行运行采集北京所有餐饮门店数据。

二、 房地产行业研究案例数据库

1. 北京小区二手交易房价数据库（来自链家网），包括：小区名称、成交数、出租数、区域、特点、均价、正在出售房屋列表数据。
2. 北京在售二手房房价数据库（来自链家网），包括：房源标题、格局、面积、朝向、楼层、特点 1、特点 2、特点 3、总价、单价。
3. 北京小区二手交易房价数据库的采集项目（XML 格式，可以基于该项目，变更为其他目标城市），包含四个子项目，可以自行运行采集北京小区二手交易房价数据。
4. 北京在售二手房房价数据库的采集项目（XML 格式，可以基于该项目，变更为其他目标城市），包含四个子项目，可以自行运行采集北京在售二手房房价数据。

三、 旅游行业研究案例数据库

1. 西藏旅游线路数据库（来自携程网），包括：旅游线路 Id、旅游线路名称、旅游线路星级、旅行方式（跟团游、自由行）、详细信息链接、当前实时价格、出发城市、用户评分、出游人数、是否已经售罄、供应商、发团时间计划、所含节假日、是否首次发团、产品特色、线路服务内容、行程概要、行程详情、行程说明、产品经理推荐、产品概要、行程精华。
2. 西藏攻略数据库，包含所有攻略（合计 16000 多篇）的标题、采集链接、正文、游玩天数、游玩时间、和谁一起、花费、游玩方式。
3. 西藏旅游线路基本数据的采集项目（XML 格式，可以基于该项目，变更为其他目标城市），包

含两个子项目，可以自行运行采集西藏最新的旅游线路基本数据。

4. 西藏攻略数据的采集项目（XML 格式，可以基于该项目，变更为其他目标城市），包含两个子项目，可以自行运行采集西藏最新的旅游线路基本数据。

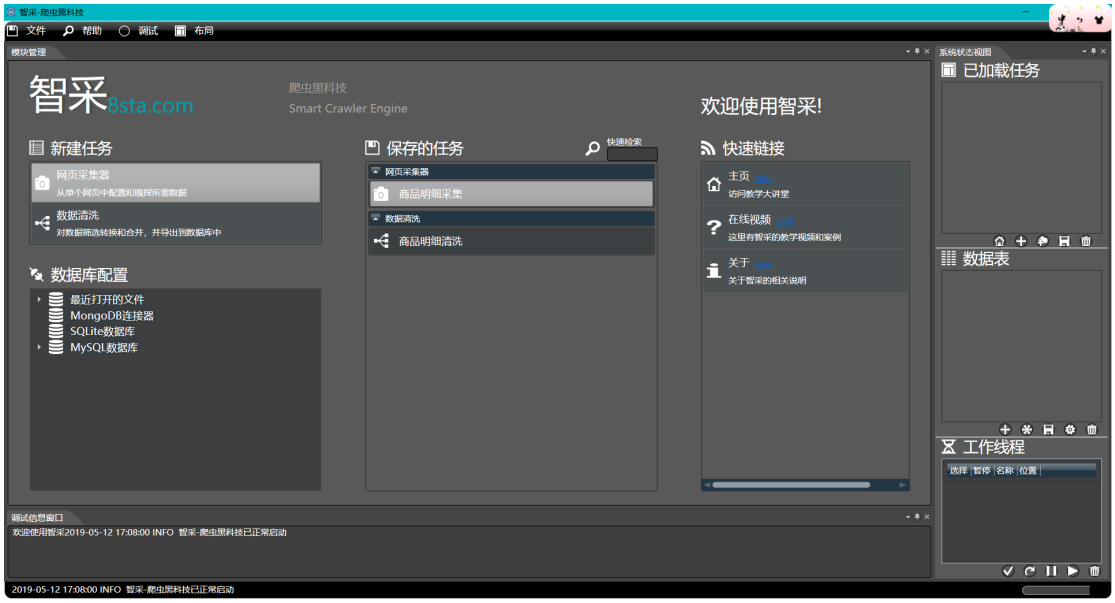
四、 教育行业研究案例数据库:

1. 北京、上海、广州的中小学、幼儿园的基本数据，字段包括：学校名称,所在地区（地市、区县）,重点级别（重点、非重点），性质（公办、民办），类别（中、小学、幼儿园）、点评数量、点评内容、地址、联系电话、地图经度、纬度（POI 信息）；
2. 北京、上海、广州的中小学教育机构基本数据的采集项目（XML 格式，可以基于该项目，变更为其他目标城市），包含两个子项目，可以自行运行采集国内最新的教育机构基本数据。

五、 互联网金融行业研究案例数据库:

1. 国内上千家互联网金融平台的档案数据，字段包括：平台名称、平台状态、平台拼音简称、平台首字母、平均收益、平台上线时间、Android APP 下载地址 、所在城市、注册资金、平台图标链接、投资期限、平台 ID、平台站岗指数、平台服务态度指数、平台提现指数、平台体验指数、微信二维码图标链接、IOS 下载链接、平台综合指数；
2. 互联网金融平台档案数据的采集项目（XML 格式），包含六个子项目，可以自行运行采集国内最新的互联网金融平台档案数据。

1.3 界面功能简介



系统主界面分为几大区域，任务区，用于管理所有的运行中任务；属性区，用于调整各项任务的属性；活动资源区：用于管理活动任务和数据；日志区：用于管理运行日志。下面对功能一一说明如下：

1. 任务区

- 模块列表：管理主模块，这是任务的入口。
  包含如下两大主模块：
  网页采集器：用于采集网页。
  数据清洗：用于对网页数据进行提取、清洗和存储。
- 保存的任务：用于管理当前项目的所有任务。
  项目下的任务列表，分为：数据采集任务和数据清洗任务两类。
- 数据源：用于管理所有的数据连接器。
  当前项目数据存储的目标方式分为：
  文件管理：连接本地文件系统中的文件。
  MongoDB 方式：连接到 MongoDB 数据库。
  MySql 方式：连接到 MySql 数据库。

2. 属性区

- 网页采集器任务的属性配置器
- 数据清洗任务的属性配置器

3. 活动资源区

- 算法视图：当前活动的网页采集、数据清洗任务列表。
- 数据视图：当前运行过程中的内存数据。

4. 日志区

- 任务管理视图：监测数据清洗任务的运行进度，可暂停、停止、删除任意任务和子任务。
- 调试信息窗口：记录各项操作日志。

第2章　任务区

任务区，以项目开始。每一个信息监测服务任务，在本系统中，统称为项目。每一个项目可以包含多个任务。用户可以新建一个项目，或者加载历史项目。

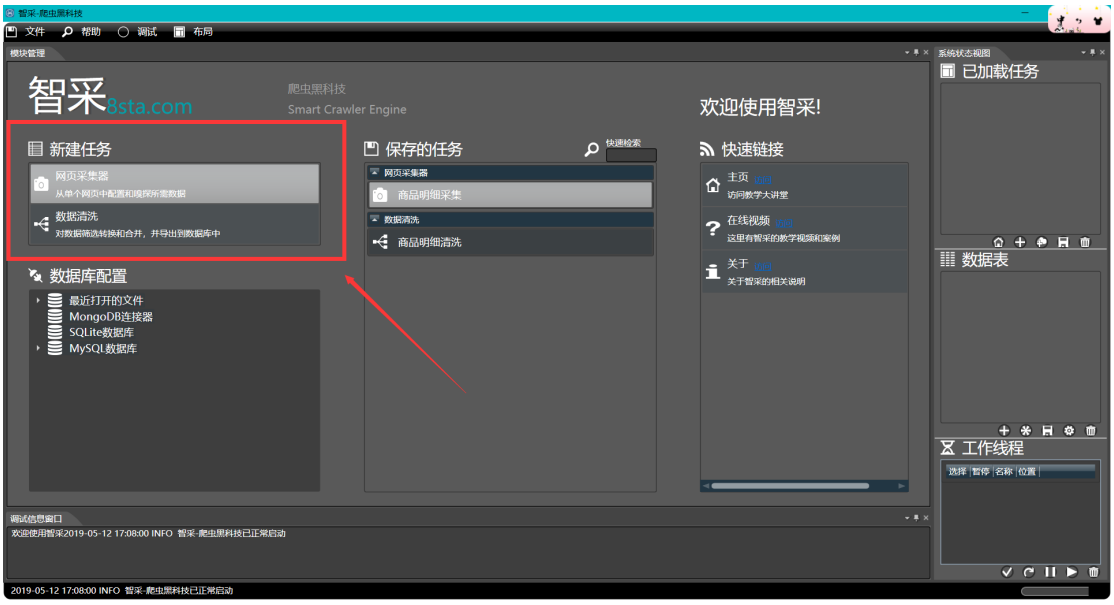
每次打开主界面，显示的界面就是任务区，任务区包含了三大部分：

- 模块列表：模块列表中，有两大主模块：网页采集器模块、数据清洗模块。
- 保存的任务：列表显示当前项目所包含的所有任务（网页采集器任务、数据清洗任务。）
- 数据源：列表显示当前项目所有可用的数据源。

2.1　模块列表

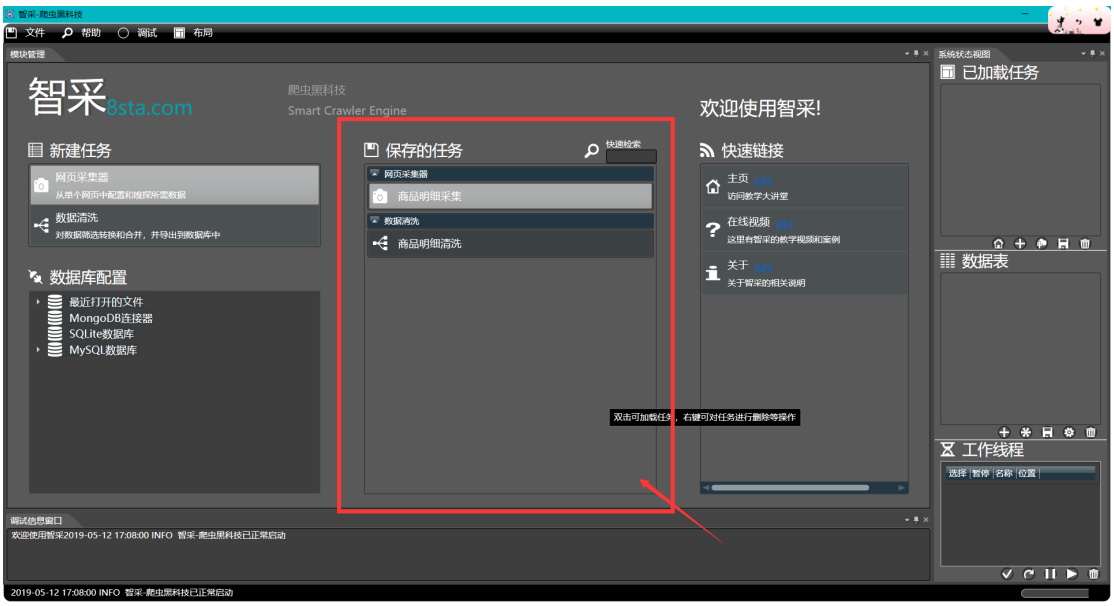
模块列表下的模块，是开展信息采集和数据清洗任务的主入口。当前主要包括如下两大主模块：

- 网页采集器：用于对于指定的网页链接，进行采集。同一个项目中，可以建立多个网页采集任务。
- 数据清洗：用于对网页采集器任务所采集的数据，进行数据清洗。同一个项目中，可以建立多个数据清洗任务。每个数据清洗任务，可以包含多个数据清洗子模块。数据清洗任务以任务流的方式，来串联每一个子模块。



2.2 保存的任务

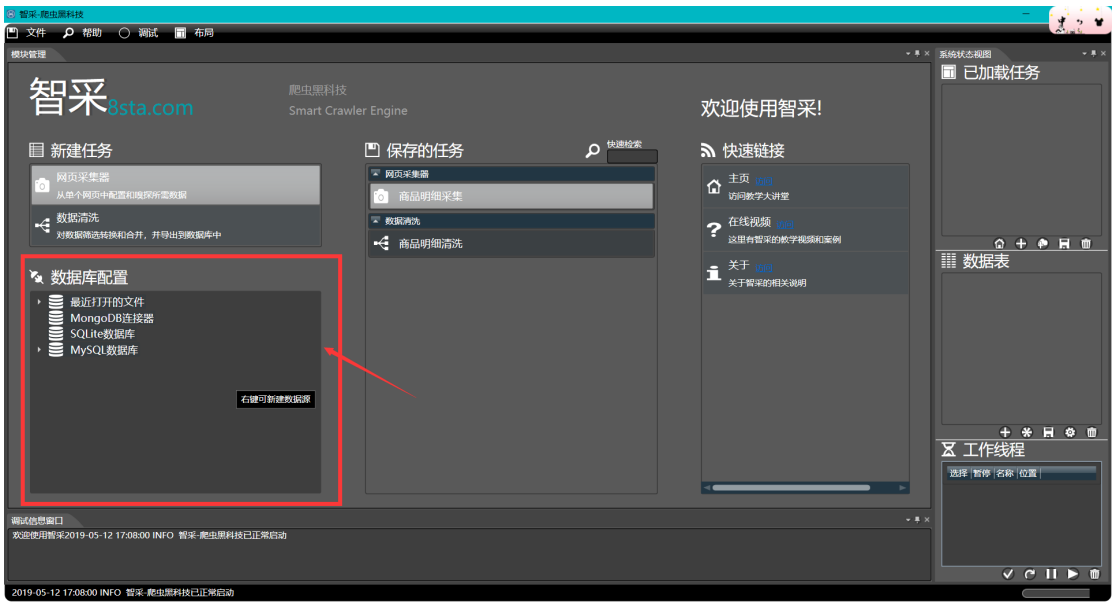
项目下的任务列表，分为：数据采集任务和数据清洗任务两类。



2.3 数据源

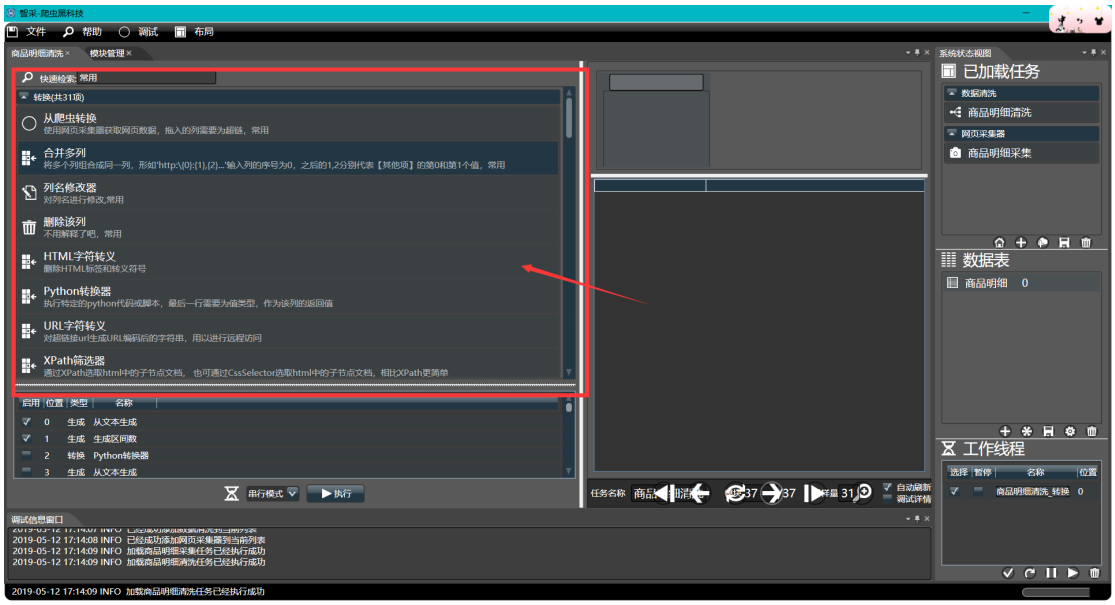
当前项目数据存储的目标方式。
分为：

- 文件管理
- MongoDB 方式
- MySql 方式



第3章　数据清洗任务中的子模块

数据清洗任务中，包含了 53 个非结构化数据清洗子模块，按照转换、执行、存储、生成和过滤进行分类，专门用于对采集下来的非结构化进行信息过滤、数据提取、格式转换、数据存储等操作，所有子模块都可以通过工作流的方式，进行串联执行，从而实现信息采集、监测任务的完全自动化。



子模块分为三个大类：

转换类子模块：一共包含 36 个数据转换类子模块，用于对非结构化数据的数据提取、处理和转换。子模块包括：搜索位置、获取 IP 的坐标、获取路径信息、检索附近、自然语言处理、语言翻译转换、添加新列、自增键生成、列名修改器、删除该列、从爬虫转换、数据库投影、重复当前值、获取请求详情、时间转字符串、URL 字符转义、HTML 字符转义、正则分割、字符串分割、转换为 Json、合并多列、Python 转换器、正则替换、正则转换器、提取数字、字符首尾抽取、启动并行、清除空白符、读取文件文本、读取文件数据、字典转换、子任务-转换、路径是否存在、重复项合并、延时、XPath 筛选器。

执行类子模块：一共包含 5 个任务执行类子模块，用于执行非结构化数据清洗子任务、数据存储任务。特别是通过“子任务-执行”类似的模块，实现对于非结构化数据清洗的多层嵌套，使得其功能组合变化无穷，可以满足几乎大多数的非结构化数据采集任务。子模块包括：写入文件文本、子任务-执行、数据库操作、保存超链接文件、写入数据库。

生成类子模块：一共包含 7 项生成类子模块，用于主动生成数据处理过程的过度数据。子模块有：获取文件夹文件、从连接器生成、生成区间数、从数据表生成、从文本生成、请求队列、子任务-生成。

过滤类子模块：一共包含 5 个过滤类子模块，用于对网页采集下来的结构化数据进行各种条件的数据过滤。子模块有：空对象过滤器、数字范围过滤器、正则筛选器、删除重复项、数

量范围选择。

3.1 转换类子模块

- 搜索位置：指定城市范围内，可以对该城市内的地址，自动搜索出其准确的经纬度。系统可以自动去调度百度地图 API，获取当前地标的经纬度坐标。对于需要自动和批量精确采集 POI 的场景，非常强大。
- 获取 IP 的坐标：可以依据 IP 地址自动搜索该 IP 地址的经纬度。用于批量获取 IP 地址的经纬度坐标。
- 获取路径信息：指定目标位置，可以自动批量计算出某地址到目位置的时间，交通方式可以选择：公交、驾车和步行三种方案。
- 检索附近：获取当前经纬度某一半径范围内的所有地物，需要拖入的为代表经度的列
- 自然语言处理：通过语言云获取的自然语言处理功能，包括分词，词性标注，命名实体识别、依存句法分析、语义依存分析、语义角色标注等，输出文字格式包括：text、Json、XML 和 byte。
- 语言翻译转换：从当前语言翻译为目标语言（支持十三种语言）。可以对包括英语、日语、文言文在内的语种之间进行自动转换。
- 添加新列：为数据集添加新列，值为某固定值。
- 自增键生成：添加一个自增列。
- 列名修改器：修改列名。
- 删除该列：删除指定列。
- 从爬虫转换：使用网页采集器获取网页数据，拖入的列需要为超链接。
- 数据库投影：利用该模块，可以快速的查询数据的数据，投影到内存中。
- 重复当前值：对当前行进行重复性生成。
- 获取请求详情：使用网页采集器获取网页数据，拖入的列需要为超链接。
- 时间转字符串：将时间类型转换为特定格式的字符串。
- URL 字符转义：对超链接 url 生成 URL 编码后的字符串，用以进行远程访问。
- HTML 字符转义：删除 HTML 标签和转义符号。
- 正则分割：使用正则表达式分割字符串。
- 字符串分割：对字符串进行分割，可以分割为一列或者多列。
- 转换为 Json：从字符串转换为 json（数组或字典类型）
- 合并多列：将多个列组合成同一列，形如：http://…..{0}{1},本列的序号为 0，其他项为 1,2,3
- Python 转换器：执行特定的 python 代码。文档列表：[{}]，转换为多个数据行构成的列表；单文档：{}，将结果的键值对附加到本行；不进行转换：直接将值放入本列。
- 正则替换：可以通过正则表达式，对列的内容进行替换。
- 正则转换器：通过正则表达式，对列的内容进行转换。
- 提取数字：提取当前列中的数字。
- 字符首尾抽取：提取字符串中，从首串到尾串中间的文本内容
- 启动并行：对多个种子合并为一个任务执行，这对于小型种子任务可以有效提升效率。该模块在执行时，会将本模块之前的流实例化，转换为实际的 list，用于提高并行性能。
- 清除空白符：清除字符串前后和中间的空白符。
- 读取文件文本：获取文件中的文本。
- 读取文件数据：从文件中读取内容。

- 字典转换：将两列数据，转换为一行数据，拖入的列为 key。
- 子任务-转换：从其他数据清洗模块中生成序列，用以组合大模块。
- 路径是否存在：判断某一个文件是否已经在指定路径上。
- 重复项合并：对重复的数据行，进行合并操作。
- 延时：在工作流中插入延时，单位为 ms，值为拖入列的值。
- XPath 筛选器：通过 XPath 选取 html 中的子节点文档。

3.1.1 搜索位置

指定城市范围内，可以对该城市内的地址，自动搜索出其准确的经纬度。系统可以自动去调度百度地图 API，获取当前地标的经纬度坐标。对于需要自动和批量精确采集 POI 的场景，非常强大。

基本选项：

- 标签：自定义模块标签。
- 类型：搜索位置
- 原列名：待搜索的地址列名。

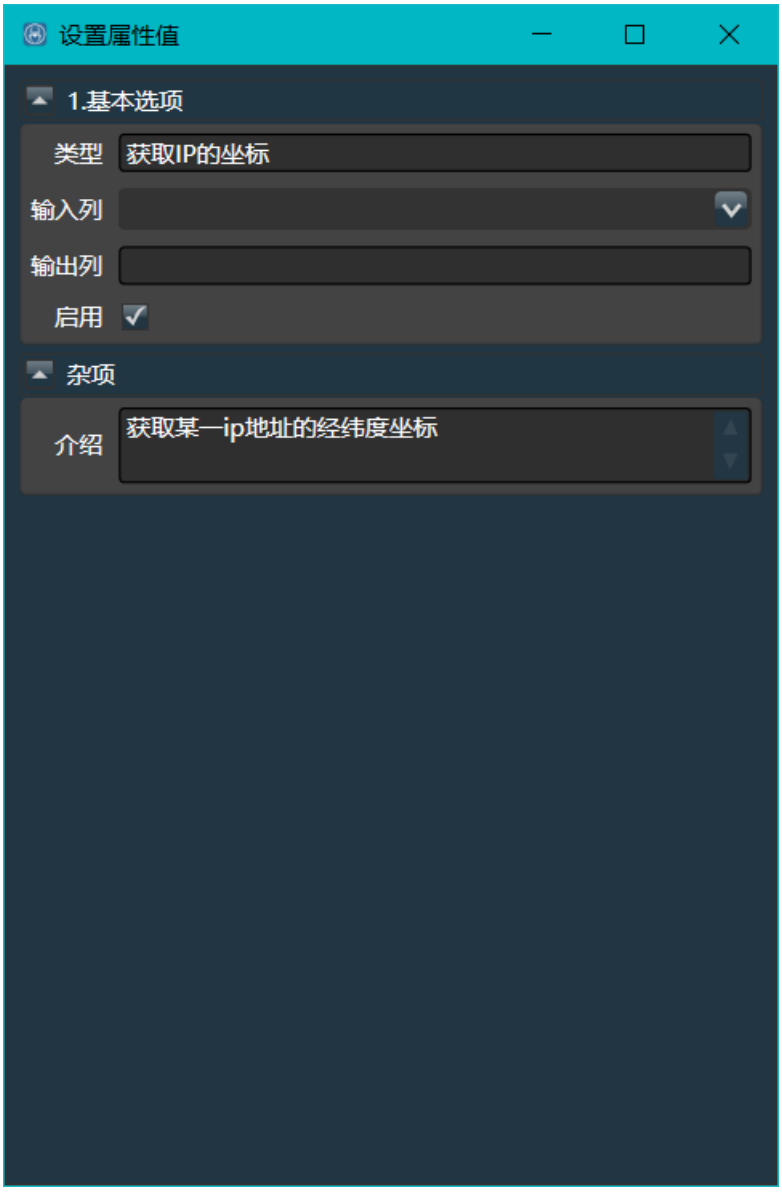
- 新列名：自定义经纬度坐标的列名。
- 启用：是否启用该模块。

杂项：

- 所属地市：指定地址所属的地市，以便缩小查找范围。
- 介绍：该子模块的介绍。

3.1.2 获取 IP 的坐标

可以依据 IP 地址自动搜索该 IP 地址的经纬度。用于批量获取 IP 地址的经纬度坐标。



基本选项：

- 标签：自定义模块标签。
- 类型：获取 IP 的坐标。
- 原列名：待搜索的 IP 地址列名。
- 新列名：自定义经纬度坐标的列名。
- 启用：是否启用该模块。

杂项：

- 介绍：该子模块的介绍。

3.1.3 获取路径信息

指定目标位置，可以自动批量计算出某地址到目位置的时间，交通方式可以选择：公交、驾车和步行三种方案。

基本选项：

- 标签：自定义模块标签。
- 类型：获取路径信息。
- 原列名：出发地点。
- 新列名：从出发地点运动到目标位置所需的时间。
- 启用：是否启用该模块。

杂项：

- 目标位置：计算的目标位置。
- 目标城市：目标位置所属的城市。

- 源城市：出发地点所属的城市。
- 运动方案：运动的方式，包括：公交、驾车和步行三种方案。

3.1.4 检索附近

获取当前经纬度某一半径范围内的所有地物，需要拖入的为代表经度的列。

基本选项：

- 标签：自定义模块标签。
- 类型：检索附近。
- 原列名：待搜索的地址列名。
- 新列名：存储附近地物。
- 启用：是否启用该模块。

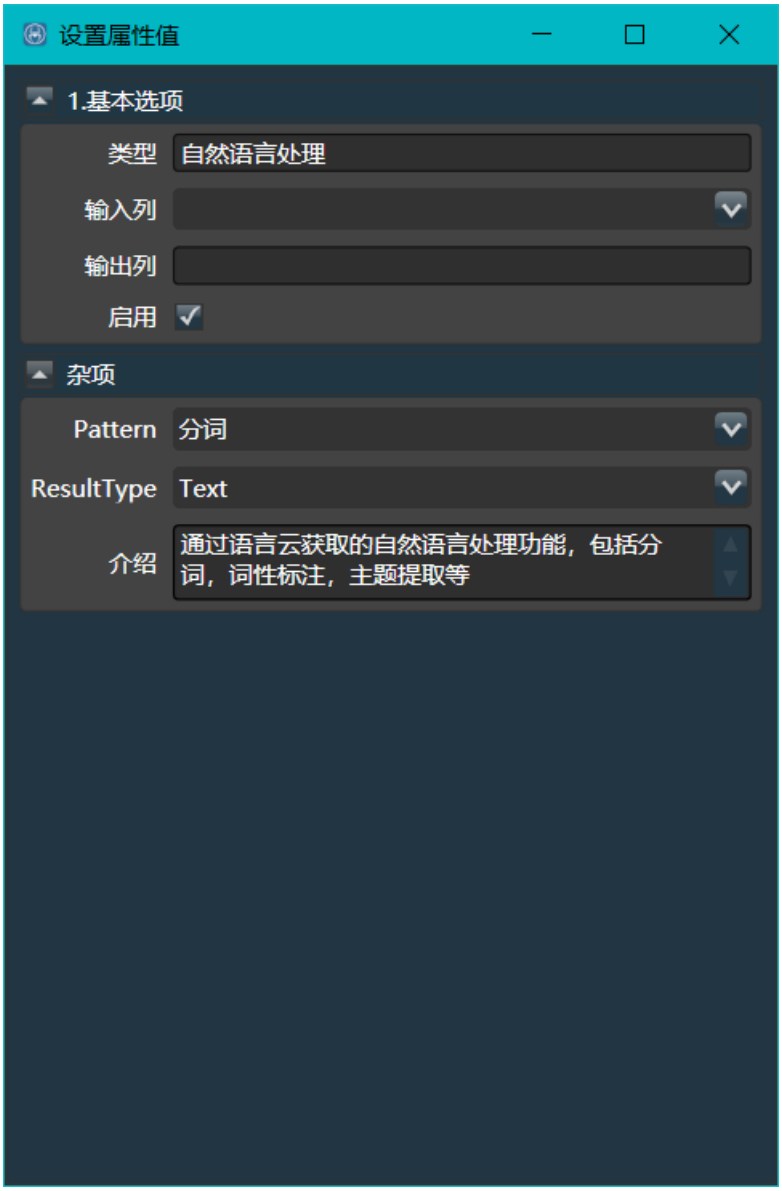
杂项：

- 所有结果：勾选记录所有结果。
- 纬度列：纬度所在列。
- 查询地物：比如：公园、加油站等。

- 搜索半径：半径单位为米。

3.1.5 自然语言处理

通过语言云获取的自然语言处理功能，包括分词，词性标注，命名实体识别、依存句法分析、语义依存分析、语义角色标注等，输出文字格式包括：text、Json、XML 和 byte。



基本选项：

- 标签：自定义模块标签。
- 类型：自然语言处理。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- Pattem（自然语言处理方法）：分词，词性标注，命名实体识别、依存句法分析、语义

依存分析、语义角色标注。

- ResultType（输出结果格式）：text、Json、XML 和 byte。

分词示例：

content	new
005000020141231734948	005000020141231734948
客户姓名:张三	客户 姓名 : 张三
民生卡号: 1111111111111111	民生卡号 : 1111111111111111
联系方式: 11111111111	联系 方式 : 11111111111
乐收银号码: 11111111111	乐收银 号码 : 11111111111
机具经办行: 广州天河北支行	机具 经办行 : 广州 天河北 支行
来电号码: 11111111111	来电 号码 : 11111111111
商户编号: 111111111111111	商户 编号 : 111111111111111
交易时间及金额: 11月11号的1111元+11号的1111元+11号的1111元	交易 时间 及 金额 : 11月 11号 的 1111 元 + 11 号 的 1111 元 + 11 号 的 1111 元
商户名称: 广州市荔湾区青藤百草药行"	商户 名称 : 广州市 荔湾区 青藤百草药 行 "
2015-01-06 10:46:59 分行处理 黄茵茵 您好，联系商户，商户称已到账，建议结束工	2015-01-06 10 : 46 : 59 分行 处理 黄 茵茵 您好 ， 联系 商户 ， 商户 称 已 到 账 ， 建议 结束 工单 。 | 投诉 认定 结果 ： 无效 投诉 | 客户 满意度 ： 满意
2015-01-06 10:56:09 分行分发 郑毓婷 结束。"	2015-01-06 10 : 56:09 分行 分发 郑毓婷 结束 。 "
003000020150104741271	003000020150104741271
2015-01-08 14:16:08 分行处理 陈赛1 清机无长款，流水核实客户金额已冲正，已联	2015-01-08 14 : 16 : 08 分行 处理 陈赛 1 清机 无 长款 ， 流水 核实 客户 金额 已 冲正 ， 已 联系 客户 核实 已经
2015-01-08 14:21:22 分行处理 苏鼎新 结束	2015-01-08 14 : 21 : 22 分行 处理 苏鼎新 结束
2015-01-08 14:25:54 分行分发 苏鼎新 结束"	2015-01-08 14 : 25 : 54 分行 分发 苏 鼎新 结束 "
003000020150105741290	003000020150105741290
2015-01-07 16:38:16 分行处理 鹿飞 分行已确认此笔错帐，今日安排调账，工单办结	2015-01-07 16 : 38 : 16 分行 处理 鹿 飞 分行 已 确认 此 笔 错帐 ， 今日 安排 调账 ， 工单 办 结 。
2015-01-07 16:43:29 分行分发 彭开宇 办结。"	2015-01-07 16 : 43 : 29 分行 分发 彭 开宇 办 结 。 "
001000020150105741292	001000020150105741292

词性标注示例：

content	new
005000020141231734948	005000020141231734948_m
客户姓名:张三	客户_n 姓名_n :_wp 张三_nh
民生卡号: 1111111111111111	民生卡号_n :_wp 1111111111111111_m
联系方式: 11111111111	联系_n 方式_n :_wp 11111111111_m
乐收银号码: 11111111111	乐收银_j 号码_n :_wp 11111111111_m
机具经办行: 广州天河北支行	机具_n 经办行_n :_wp 广州_ns 天河北_ns 支行_n
来电号码: 11111111111	来电_n 号码_n :_wp 11111111111_m
商户编号: 111111111111111	商户_n 编号_n :_wp 111111111111111_m
交易时间及金额: 11月11号的1111元+11号的1111元+11号的1111元	交易_n 时间_n 及_c 金额_n :_wp 11月_nt 11号_nt 的_u 1111_m 元_q +_wp 11_m 号_q 的_u 1111_m 元_
商户名称: 广州市荔湾区青藤百草药行"	商户_n 名称_n :_wp 广州市_ns 荔湾区_ns 青藤百草药_n 行_n "_wp
2015-01-06 10:46:59 分行处理 黄茵茵 您好，联系商户，商户称已到	2015-01-06_m 10_m :_wp 46_m :_wp 59_m 分行_q 处理_v 黄_nh 茵茵_nh 您好_i ，_wp 联系_v 商户_n |_wp 投诉_v 认定_v 结果_n ：_wp 无效_v 投诉_v |_wp 客户_n 满意度_n ：_wp 满意_v
2015-01-06 10:56:09 分行分发 郑毓婷 结束。"	2015-01-06_m 10_m :_wp 56:09_m 分行_n 分发_v 郑毓婷_nh 结束_v 。_wp "_wp
003000020150104741271	003000020150104741271_m
2015-01-08 14:16:08 分行处理 陈赛1 清机无长款，流水核实客户金	2015-01-08_m 14_m :_wp 16_m :_wp 08_m 分行_n 处理_v 陈赛_nh 1_m 清机_n 无_v 长款_n ，_wp 流水
2015-01-08 14:21:22 分行处理 苏鼎新 结束	2015-01-08_m 14_m :_wp 21_m :_wp 22_m 分行_q 处理_v 苏鼎新_nh 结束_v
2015-01-08 14:25:54 分行分发 苏鼎新 结束"	2015-01-08_m 14_m :_wp 25_m :_wp 54_m 分行_n 分发_v 苏_j 鼎新_d 结束_v "_wp
003000020150105741290	003000020150105741290_m
2015-01-07 16:38:16 分行处理 鹿飞 分行已确认此笔错帐，今日安排	2015-01-07_m 16_m :_wp 38_m :_wp 16_m 分行_q 处理_v 鹿_n 飞_v 分行_n 已_d 确认_v 此_r 笔_q 错帐
2015-01-07 16:43:29 分行分发 彭开宇 办结。"	2015-01-07_m 16_m :_wp 43_m :_wp 29_m 分行_n 分发_v 彭_nh 开宇_nh 办_v 结_v 。_wp "_wp
001000020150105741292	001000020150105741292_m

语义角色标注示例：

content	new
2015-01-06 10:46:59 分行	2015-01-06 10 : 46 : 59 分行 [处理]v [黄 茵茵 您好 ， 联系 商户 ， 商户 称 已 到 账 ， 建议 结束 工单]A1 。 2015-01-06 10 : 46 : 59 分行 处理 [黄 茵茵]A0 [您好]i ， 联系 商户 ， 商户 称 已 到 账 ， 建议 结束 工单 。 2015-01-06 10 : 46 : 59 分行 处理 黄 茵茵 您好 ， [联系]v [商户]A1 ， 商户 称 已 到 账 ， 建议 结束 工单 。 2015-01-06 10 : 46 : 59 分行 处理 黄 茵茵 您好 ， 联系 商户 ， [商户]A0 [称]v 已 到 [账]A1 ， 建议 结束 工单 。 2015-01-06 10 : 46 : 59 分行 处理 黄 茵茵 您好 ， 联系 商户 ， 商户 称 [已]ADV [到]v [账]A1 ， 建议 结束 工单 。 2015-01-06 10 : 46 : 59 分行 处理 黄 茵茵 您好 ， 联系 商户 ， 商户 称 已 到 账 ， [建议]v [结束 工单]A1 。 2015-01-06 10 : 46 : 59 分行 处理 黄 茵茵 您好 ， 联系 商户 ， 商户 称 已 到 账 ， 建议 [结束]v [工单]A1 。
2015-01-06 10:56:09 分行	2015-01-06 10 : [56:09 分行 分发 郑[illegible]A0 [结束]v 。
00300002015010474127	
2015-01-08 14:16:08 分行	2015-01-08 14 : 16 : [08 分行]A0 [处理]v 陈赛 1 清机 无 长款 ， 流水 核实 客户 金额 已 冲正 ， 已 联系 客户 核实 已经 到 账 ， 请 2015-01-08 14 : 16 : 08 分行 处理 [陈赛 1 清机]A0 [无]v [长款]A1 ， 流水 核实 客户 金额 已 冲正 ， 已 联系 客户 核实 已经 到 账 2015-01-08 14 : 16 : 08 分行 处理 陈赛 1 清机 无 长款 ， [流水]A0 [核实]v 客户 金额 已 冲正 ， 已 联系 客户 核实 已经 到 账 ， 请 2015-01-08 14 : 16 : 08 分行 处理 陈赛 1 清机 无 长款 ， 流水 核实 [客户 金额]A0 [已]ADV [冲正]v ， 已 联系 客户 核实 已经 到 账 2015-01-08 14 : 16 : 08 分行 处理 陈赛 1 清机 无 长款 ， 流水 核实 客户 金额 已 冲正 ， [已]ADV [联系]v [客户]A1 [核实 已经 到 2015-01-08 14 : 16 : 08 分行 处理 陈赛 1 清机 无 长款 ， 流水 核实 客户 金额 已 冲正 ， 已 联系 客户 [核实]v 已经 到 [账]A1 ， 请 2015-01-08 14 : 16 : 08 分行 处理 陈赛 1 清机 无 长款 ， 流水 核实 客户 金额 已 冲正 ， 已 联系 客户 核实 [已经]ADV [到]v [账]A 2015-01-08 14 : 16 : 08 分行 处理 陈赛 1 清机 无 长款 ， 流水 核实 客户 金额 已 冲正 ， 已 联系 客户 核实 已经 到 账 ， [请]v [结
2015-01-08 14:21:22 分行	2015-01-08 14 : 21 : [22 分行 处理 苏鼎新]A0 [结束]v
2015-01-08 14:25:54 分行	2015-01-08 14 : 25 : [54 分行]A0 [分发]v 苏 鼎新 结束 " 2015-01-08 14 : 25 : 54 分行 分发 苏 [鼎新]ADV [结束]v "
00300002015010574129	
2015-01-07 16:38:16 分行	2015-01-07 16 : 38 : 16 分行 [处理]v [鹿]A1 飞 分行 已 确认 此 笔 错帐 ， 今日 安排 调账 ， 工单 办 结 。 2015-01-07 16 : 38 : [16 分行 处理 鹿 飞 分行]A0 [已]ADV [确认]v [此 笔 错帐]A1 ， 今日 安排 调账 ， 工单 办 结 。 2015-01-07 16 : 38 : 16 分行 处理 鹿 飞 分行 已 确认 此 笔 错帐 ， [今日]TMP [安排]v 调账 ， 工单 办 结 。 2015-01-07 16 : 38 : 16 分行 处理 鹿 飞 分行 已 确认 此 笔 错帐 ， 今日 安排 调账 ， [工单]A0 [办]v 结 。
2015-01-07 16:43:29 分行	2015-01-07 16 : 43 : [29 分行]A0 [分发]v [彭 开宇]A1 办 结 。 2015-01-07 16 : 43 : 29 分行 分发 彭 开宇 [办]v [结]A1 。
00100002015010574129	

3.1.6 语言翻译转换

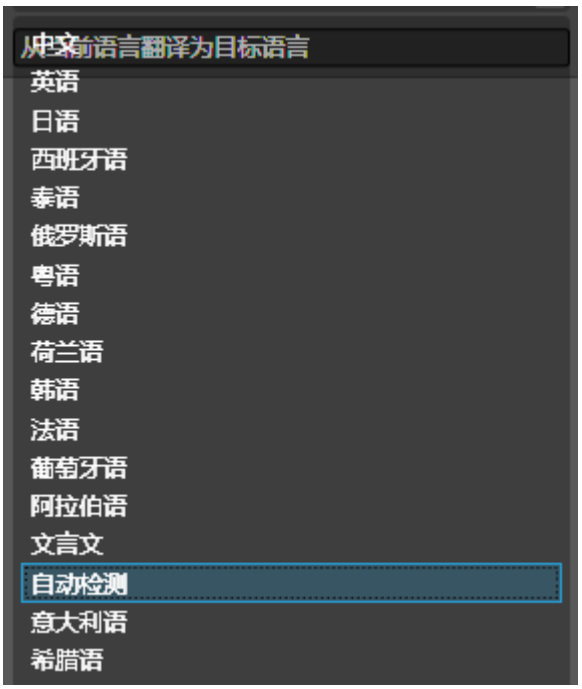
从当前语言翻译为目标语言。可以对包括英语、日语、文言文在内的语种之间进行自动转换。

基本选项：

- 标签：自定义模块标签。
- 类型：语言翻译转换。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 应用中心账号：云翻译中心的登录账户。
- Source（当前语言）：支持多种语言（十六种语言）。
- Target（目标语言）：支持多种语言（十六种语言）。



3.1.7 添加新列

为数据集添加新列，值为某固定值。

基本选项：

- 标签：自定义模块标签。
- 类型：添加新列。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 生成值：为数据集添加新列，此为新列的固定值。

3.1.8 自增键生成

添加一个自增列。

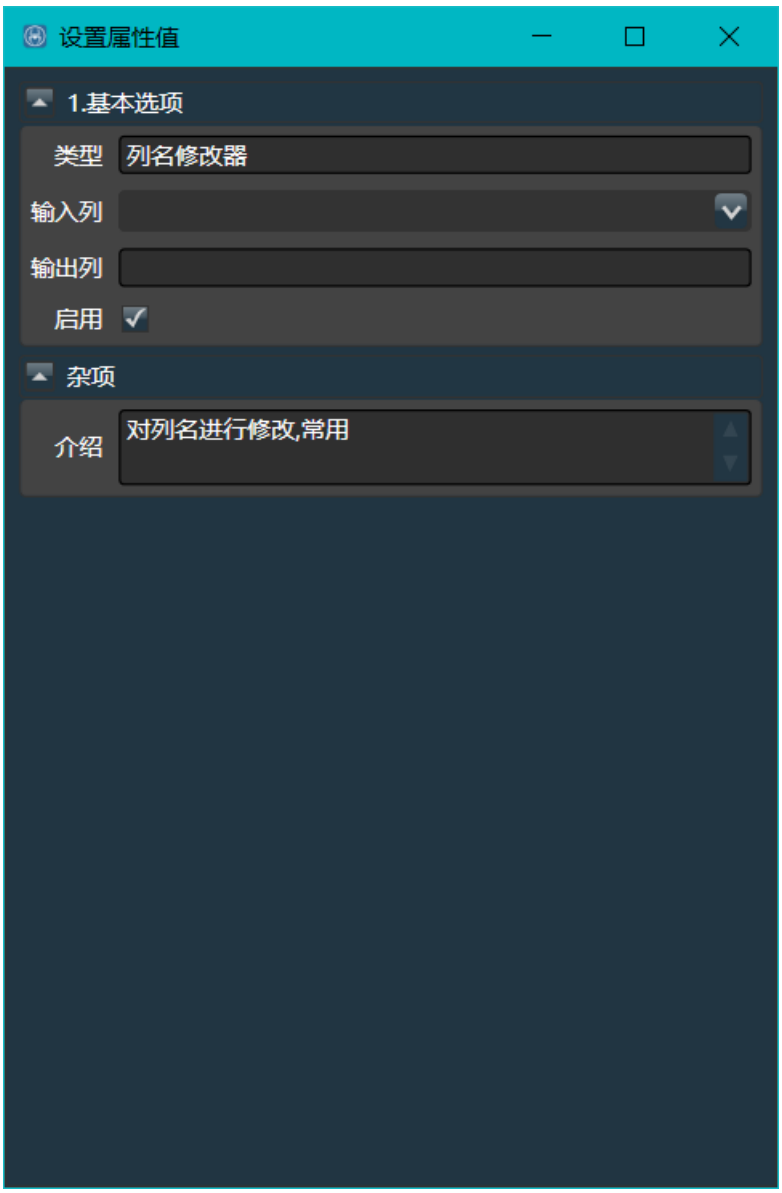
基本选项：

- 标签：自定义模块标签。
- 类型：自增键生成。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 起始索引：一般从 0 开始，每次自增 1。

3.1.9 列名修改器



基本选项：

- 标签：自定义模块标签。
- 类型：列名修改器。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

3.1.10 删除该列

基本选项：

- 标签：自定义模块标签。
- 类型：删除该列。
- 原列名：待删除列。
- 新列名：结果列。（没用）
- 启用：是否启用该模块。

3.1.11 从爬虫转换

此为一个重要的模块，唯一的依靠该模块，可以与网页采集器联合使用。该模块使用网页采集器获取网页数据。因此，拖入的列需要为超链接。
该模块可以自动依据超链接列，自动采集数据（指定的网页采集器名），生成多列（通过指定的网页采集器定义）。

基本选项：

- 标签：自定义模块标签。
- 类型：从爬虫转换。
- 原列名：待处理列（超链接列）。
- 新列名：结果列（当网页采集器是单列模式【one】还是【list】模式，都会制定列的内容）。
- 启用：是否启用该模块。

杂项：

- 爬虫选择：填写指定的网页采集器名称。
- 延时时间：延时多长时间开始采集（ms）。
- 错误延时时间：延时多长时间重试（ms）。
- 最大重复次数：出错重复抓取的次数。
- Post 数据：通过 Post 提交的数据。

请求队列（暂可不用，用在高度并发的环境下）。

示例：

URL 为超链接列，拖拽本模块进入该列，自动生成相关数据。

URL	city	属性0	p
http://you.ctrip.com/sight/linzhi126.html	林芝	1-15 / 164条	11
http://you.ctrip.com/sight/ritu120147.html	日土		1
http://you.ctrip.com/sight/anduo120385.html	安多		1
http://you.ctrip.com/sight/dingri2450.html	定日		1
http://you.ctrip.com/sight/chayu120394.html	察隅		1
http://you.ctrip.com/sight/suoxian120391.html	索县		1
http://you.ctrip.com/sight/geer120143.html	噶尔		1
http://you.ctrip.com/sight/rikaze100.html	日喀则	1-15 / 94条	7
http://you.ctrip.com/sight/jiangzi2437.html	江孜		1
http://you.ctrip.com/sight/yadong120090.html	亚东		1
http://you.ctrip.com/sight/chaya120402.html	察雅		1
http://you.ctrip.com/sight/luozha120411.html	洛扎		1
http://you.ctrip.com/sight/ali99.html	阿里	1-15 / 44条	3
http://you.ctrip.com/sight/naqu120387.html	那曲		1
http://you.ctrip.com/sight/jiacha120409.html	加查		1

3.1.12 数据库匹配

利用该模块，可以快速的查询数据的数据，投影到内存中。

基本选项：

- 类型：数据库匹配。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 导入列：选择表后，这里会自动显示可用的表名，供选择投影进入内存的列。
- 查询多数据：选择该项时，会同时查询所有满足条件的项，将同一列保存为数组。
- 表主键：指定表的主键。
- 匹配方式：包含：Contains，Like，Match，Initials 四种方式（与代处理列通过这四种关系进行匹配）。
- 连接器：选择数据源列表下的所有连接器。
- 表名：连接器下的表。

3.1.13 重复当前值

对当前行进行重复性生成。

基本选项：

- 标签：自定义模块标签。
- 类型：重复当前值。
- 原列名：待处理列（超链接列）。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 重复次数：指定重复次数。
- 重复模式：指定重复的方式。

3.1.14 获取请求响应

使用网页采集器获取网页数据，拖入的列需要为超链接。

基本选项：

- 标签：自定义模块标签。
- 类型：重复当前值。
- 原列名：待处理列（超链接列）。
- 新列名：结果列。
- 启用：是否启用该模块。

头数据：

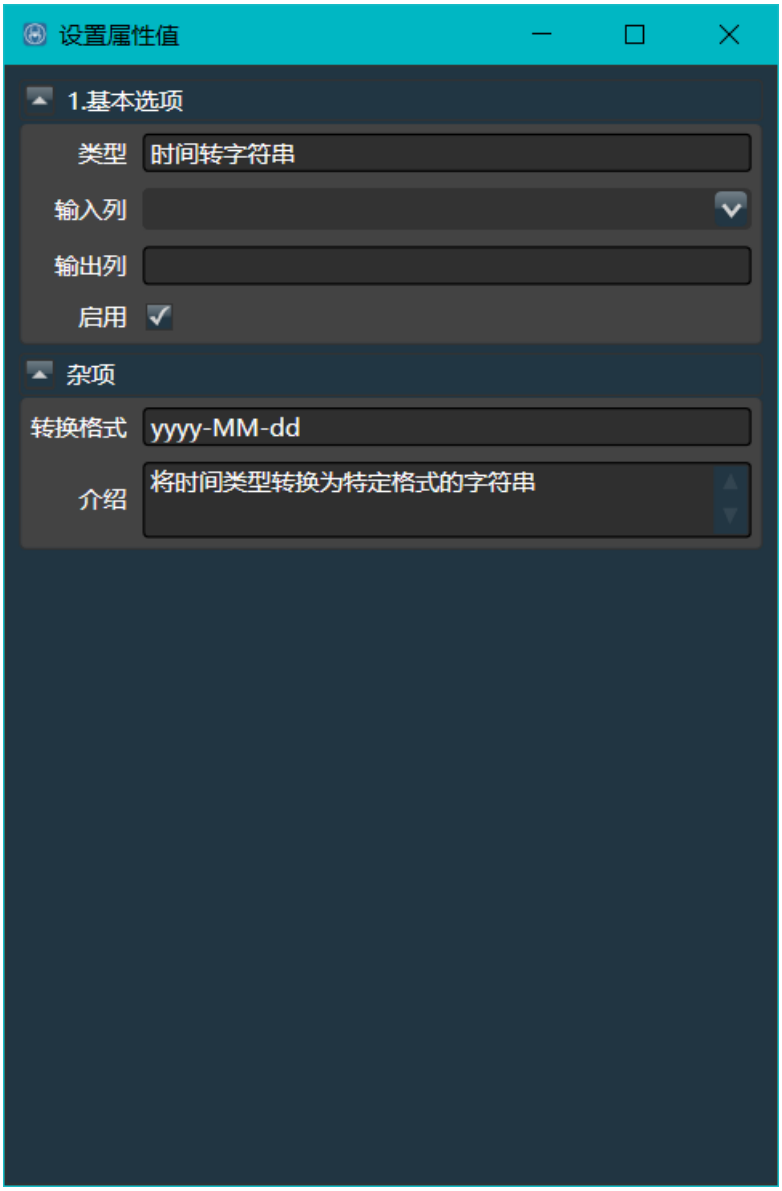
头数据过滤文件。

杂项：

- 爬虫选择：指定网页采集器。

3.1.15 时间转字符串

将时间类型转换为特定格式的字符串。



基本选项：

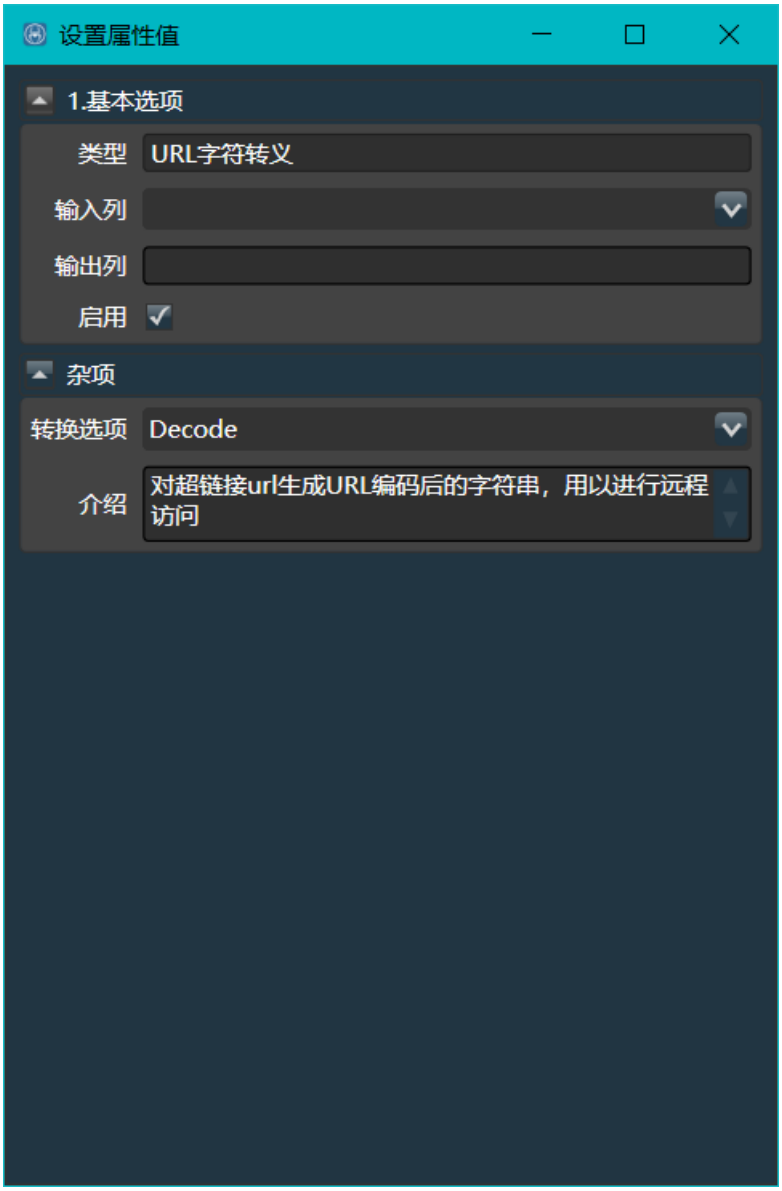
- 标签：自定义模块标签。
- 类型：时间转字符串。
- 原列名：待处理列（时间列）。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 转换格式：y 代表年，M 代表月，d 代表日，可自由组合格式。

3.1.16 URL 字符转义

对超链接 url 生成 URL 编码后的字符串，用以进行远程访问。



基本选项：

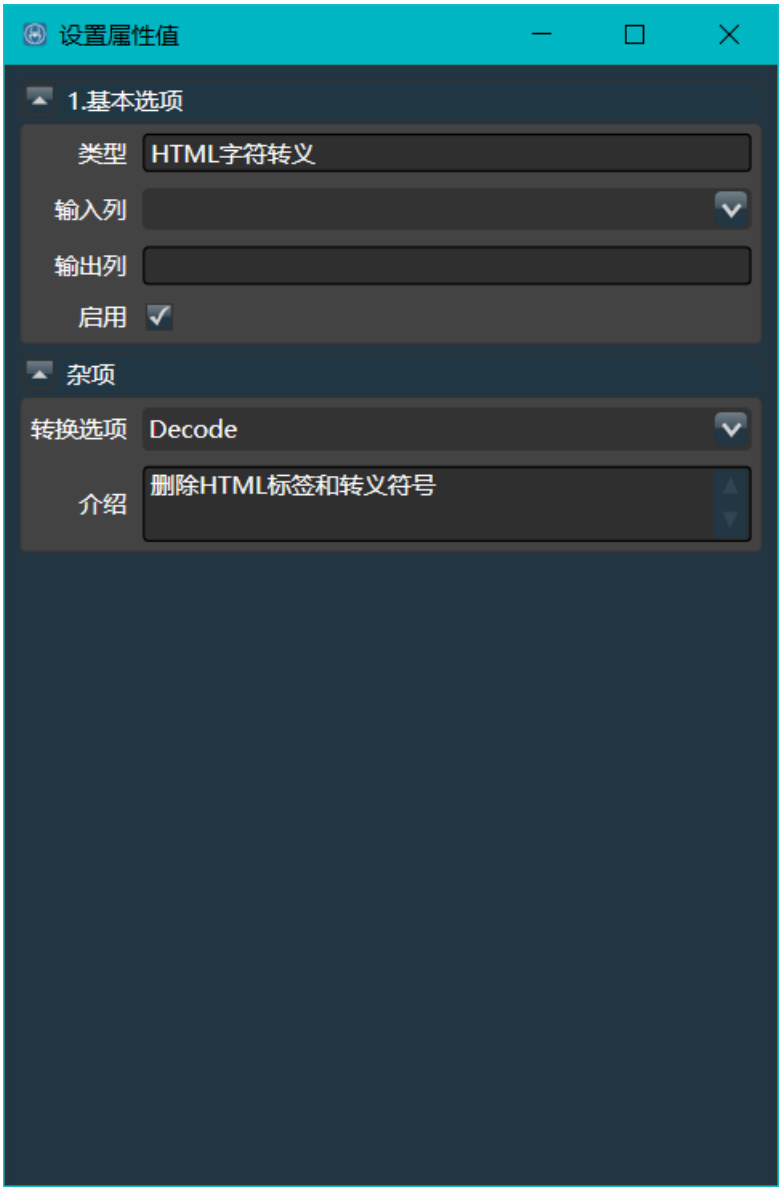
- 标签：自定义模块标签。
- 类型：URL 字符转义。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 转换选项：Decode，Encode。

3.1.17 HTML 字符转义

删除 HTML 标签和转义符号。



基本选项：

- 标签：自定义模块标签。
- 类型：HTML 字符转义。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 转换选项：Decode，Encode。

3.1.18 正则分割

使用正则表达式分割字符串。

基本选项：

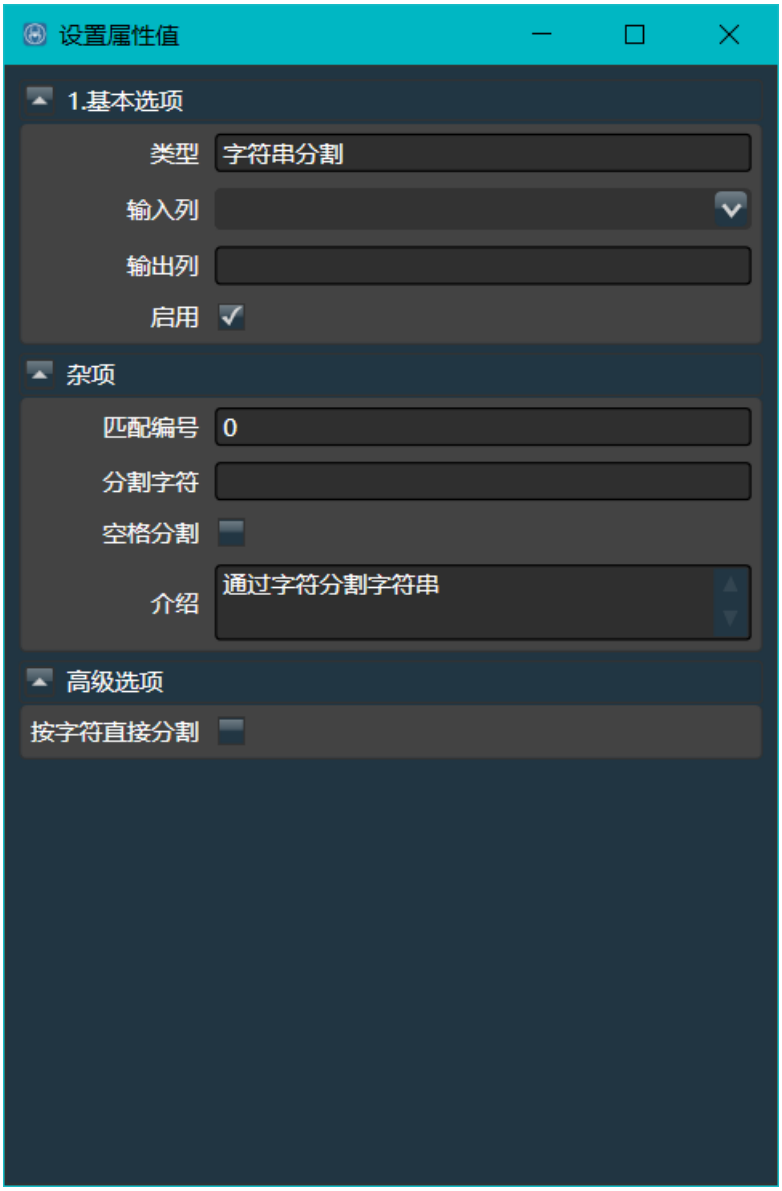
- 标签：自定义模块标签。
- 类型：正则分割。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 倒序：选择此项，对字符串的处理顺序为倒序处理。
- 匹配编号：0 代表保留第一个匹配到的，1 代表保留第二个，以此类推。
- 表达式：正则表达式。

3.1.19 字符串分割

对字符串进行分割，可以指定风格字符，并通过匹配编号，保留指定匹配位置的信息。



基本选项：

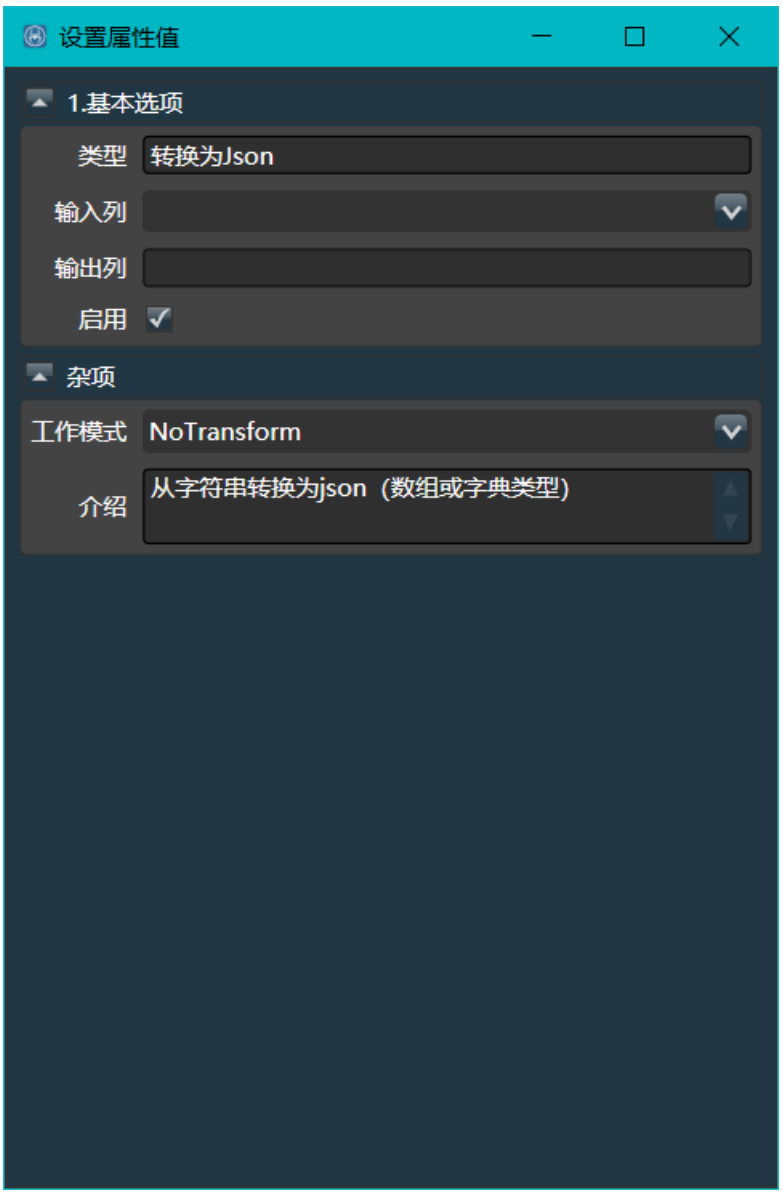
- 标签：自定义模块标签。
- 类型：字符串分割。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 倒序：选择此项，对字符串的处理顺序为倒序处理。
- 匹配编号：0 代表保留第一个匹配到的，1 代表保留第二个，以此类推。
- 分割字符：勾选表示选择指定字符进行分割。
- 换行分割：勾选表示通过换行符进行分割。
- 空格分割：勾选表示通过空格分割。
- 表达式：正则表达式。

3.1.20 转换为 Json

从字符串转换为 json（数组或字典类型）。



基本选项：

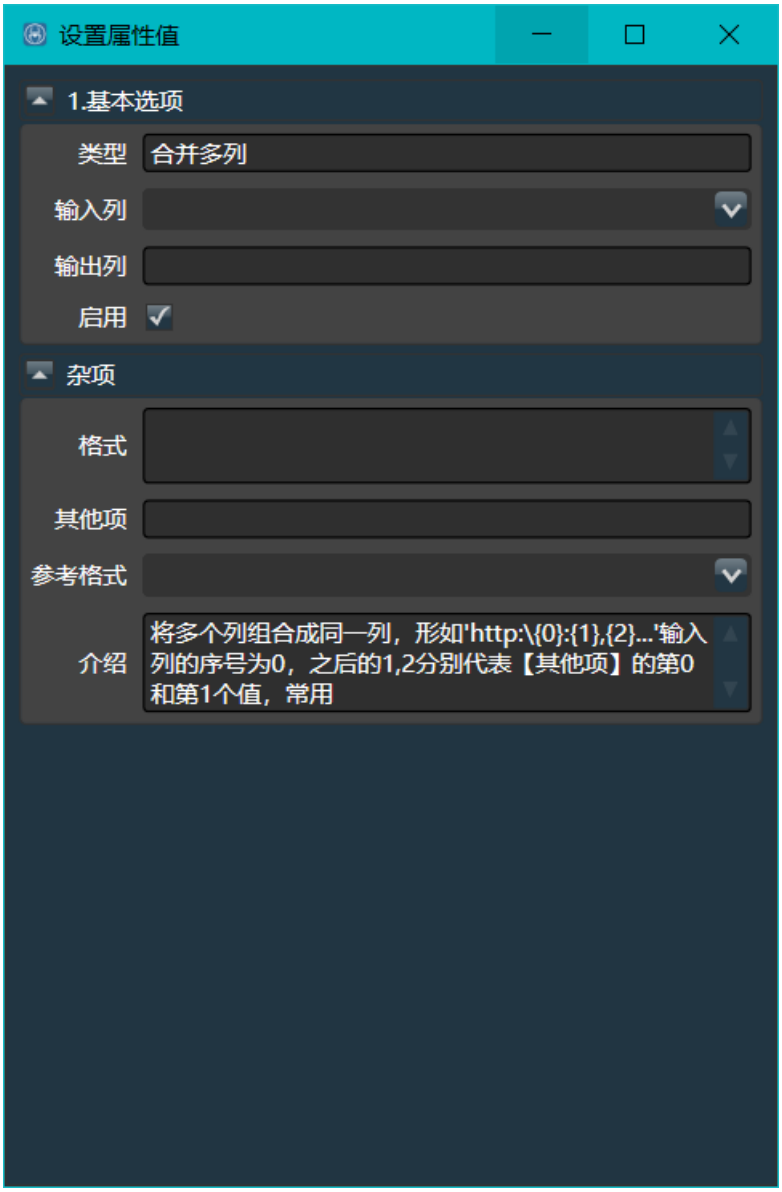
- 标签：自定义模块标签。
- 类型：转换为 Json。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 工作模式：可以转换为多种模式。

3.1.21 合并多列

将多个列组合成同一列，形如：http://.....{0}{1},本列的序号为 0，其他项为 1,2,3。



基本选项：

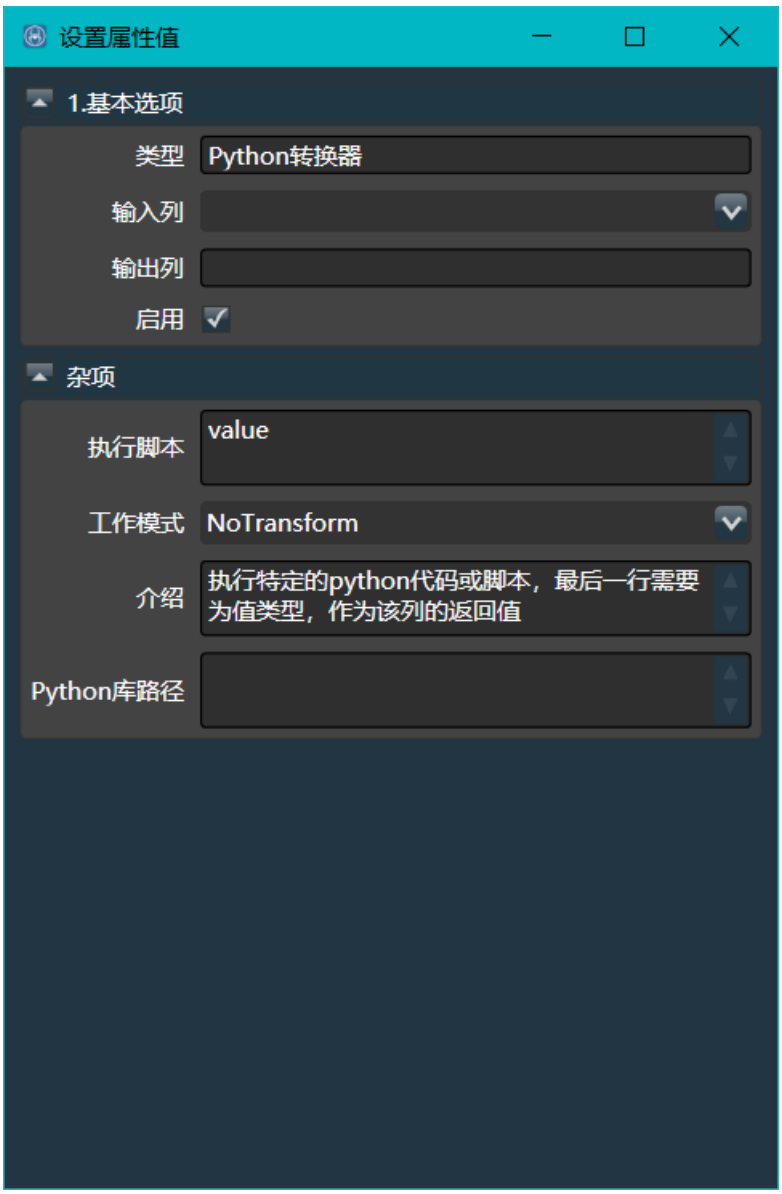
- 标签：自定义模块标签。
- 类型：合并多列。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- Format：指定新列格式，形如：http://.....{0}{1},本列的序号为 0，其他项为 1,2,3。
- 其他项：当前列之外的其他列，空格分割。

3.1.22 Python 转换器

执行特定的 python 代码。文档列表：[{}]，转换为多个数据行构成的列表；单文档：{}，将结果的键值对附加到本行；不进行转换：直接将值放入本列。



基本选项：

- 标签：自定义模块标签。
- 类型：Python 转换器。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 执行脚本：用于转换的自定 Python 脚本。
- 工作模式：可指定多种模式。

3.1.23 正则替换

基本选项：

- 标签：自定义模块标签。
- 类型：正则替换。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 匹配编号：0 代表替换第一个匹配到的，1 代表替换第二个，以此类推。
- 替换为：需要替换的数值。
- 表达式：正则表达式。

3.1.24 正则转换器

通过正则表达式，对列的内容进行转换。

基本选项：

- 标签：自定义模块标签。
- 类型：正则转化器。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 匹配编号：若编号小于 0 且匹配出多个新列，列名可用空格分割，若该列不需要添加，可用_表示，如“_匹配 1_”。通过正则表达式提取内容。
- 表达式：正则表达式。

3.1.25 提取数字

提取当前列中的数字。

基本选项：

- 标签：自定义模块标签。
- 类型：提取数字。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 匹配编号：若编号小于 0 且匹配出多个新列，列名可用空格分割，若该列不需要添加，可用_表示，如“_匹配 1_”。通过正则表达式提取内容。
- 表达式：正则表达式。

3.1.26 字符首尾抽取

提取字符串中，从首串到尾串中间的文本内容。

基本选项：

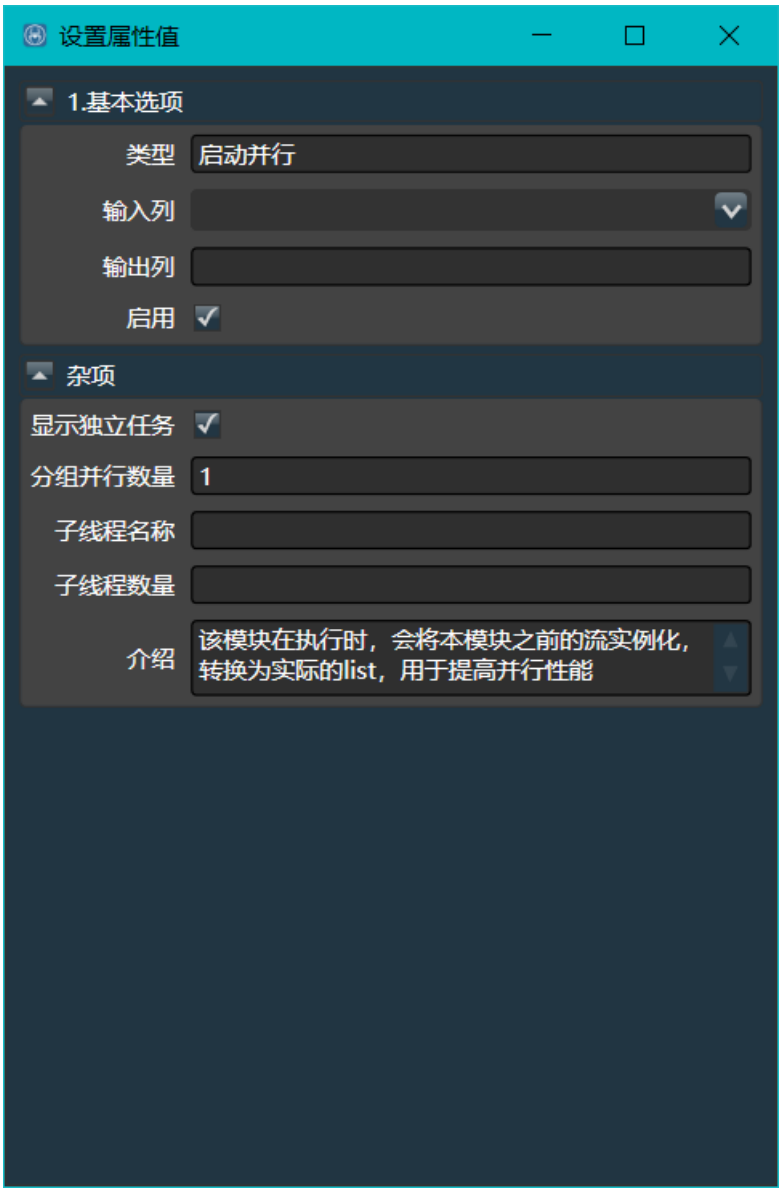
- 标签：自定义模块标签。
- 类型：字符首尾抽取。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 尾串：以该字符结尾。
- 首串：以该字符开始。
- 包含首尾串：勾选此项，提取串要包含首尾串。

3.1.27 启动并行

对多个种子合并为一个任务执行，这对于小型种子任务可以有效提升效率。该模块在执行时，会将本模块之前的流实例化，转换为实际的 list，用于提高并行性能。



基本选项：

- 标签：自定义模块标签。
- 类型：启动并行。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

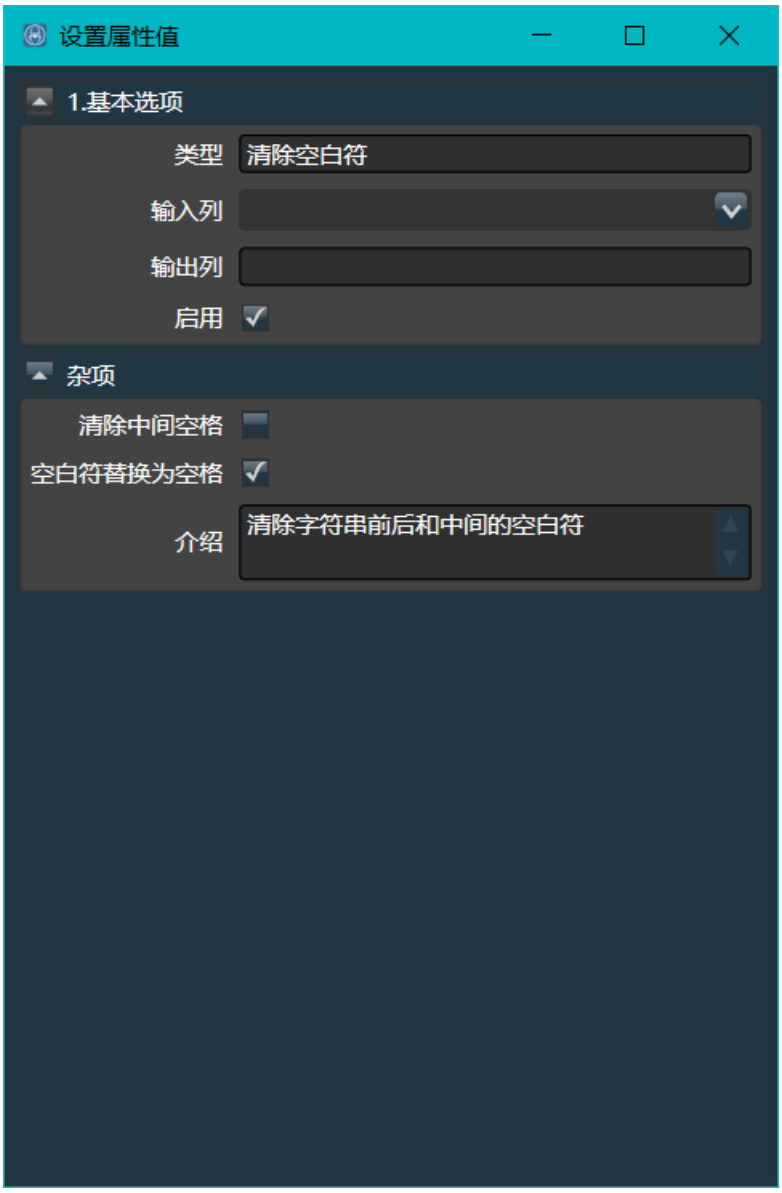
杂项：

- 显示独立任务：是否将每个子线程插入到任务队列中，从而显示进度。
- 分组并行数量：并行线程数量。
- 子线程名称：指示子线程名称。

- 子线程数量：并行的数量。

3.1.28 清除空白符

清除字符串前后和中间的空白符。



基本选项：

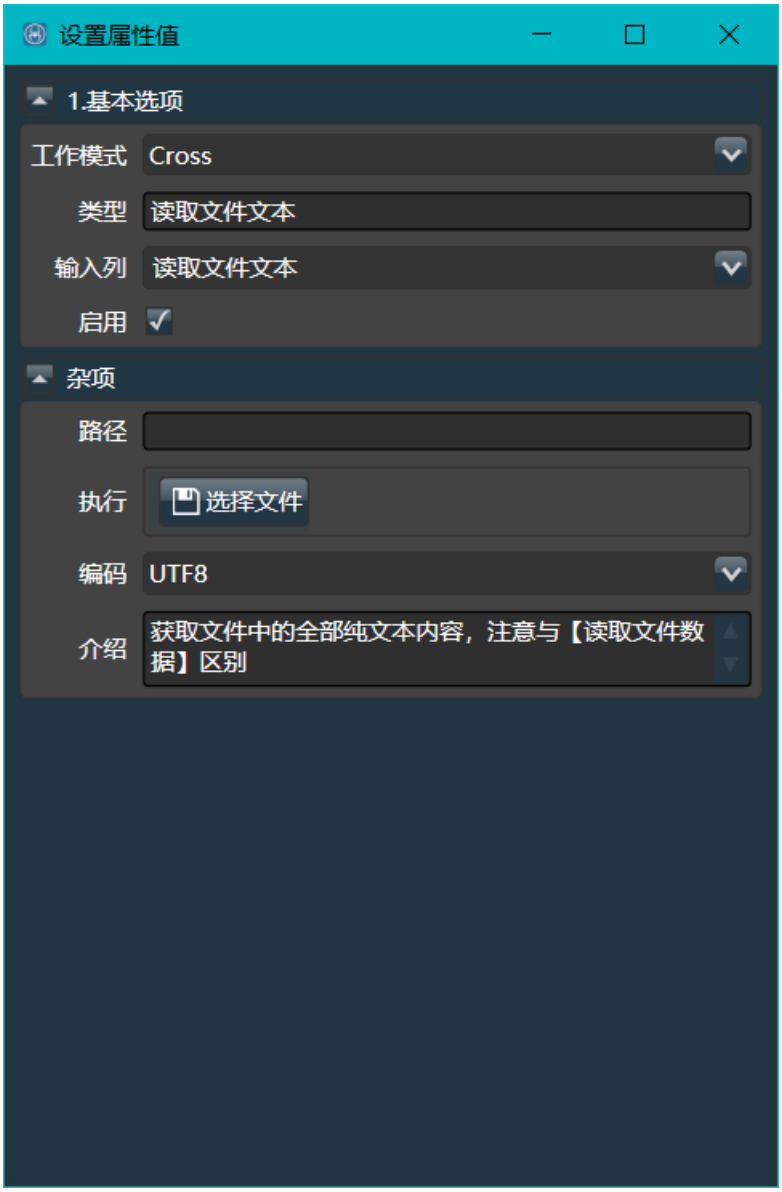
- 标签：自定义模块标签。
- 类型：清除空白符。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

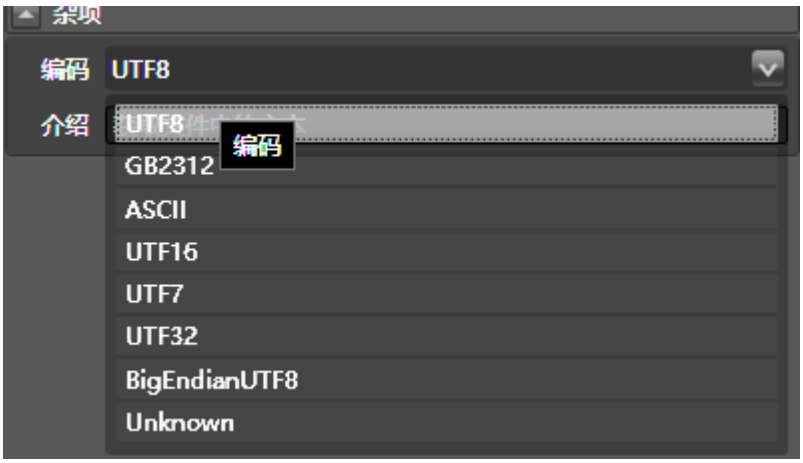
杂项：

- 清除中间空格：勾选此项，清除中间空格。
- 空白符替换为空格：将空白符替换为空格。

3.1.29 读取文件文本

获取文件中的文本。





基本选项：

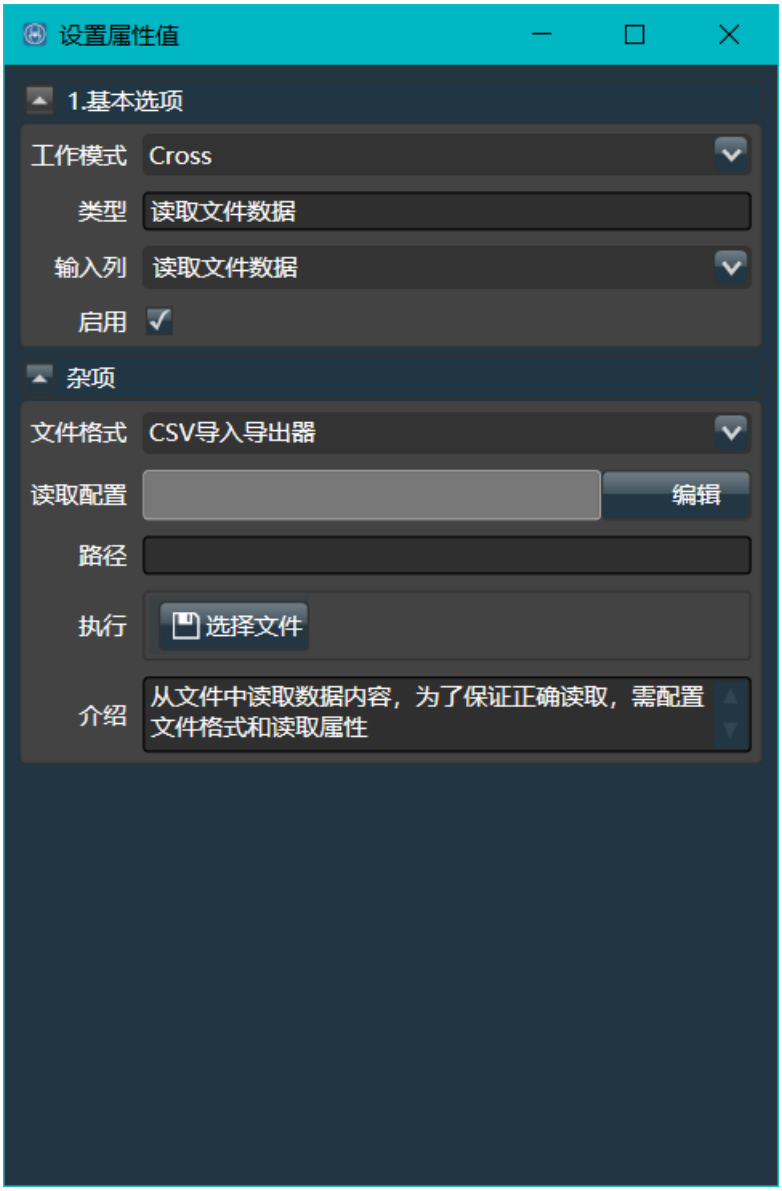
- 标签：自定义模块标签。
- 类型：读取文件文本。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

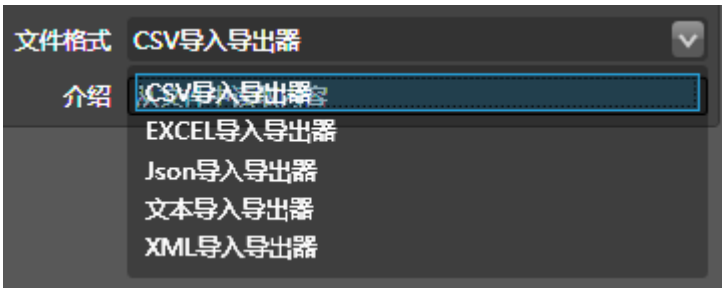
杂项：

- 编码：选择文本中的字符格式。

3.1.30 读取文件数据

从文件中读取内容。





基本选项：

- 标签：自定义模块标签。
- 类型：读取文件数据。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 文件格式：提供包含 5 种文件格式的导入导出器（CSV,EXCEL,Json,文本,XML 五种格式）。

3.1.31 矩阵转置

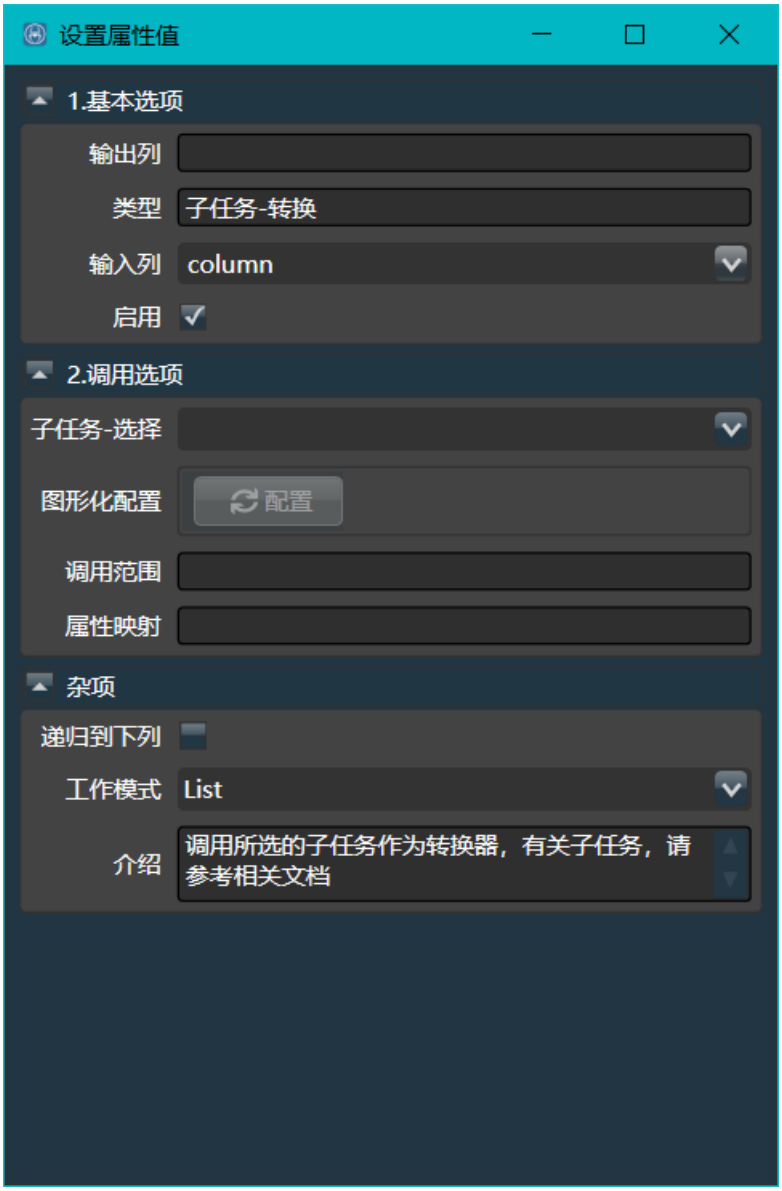
将两列数据，转换为一行数据，拖入的列为 key。

基本选项：

- 标签：自定义模块标签。
- 类型：矩阵转置。
- 原列名：待处理列（key 列）。
- 新列名：结果列。
- 启用：是否启用该模块。

3.1.32 子任务-转换

从其他数据清洗模块中生成序列，用以组合大模块。



基本选项：

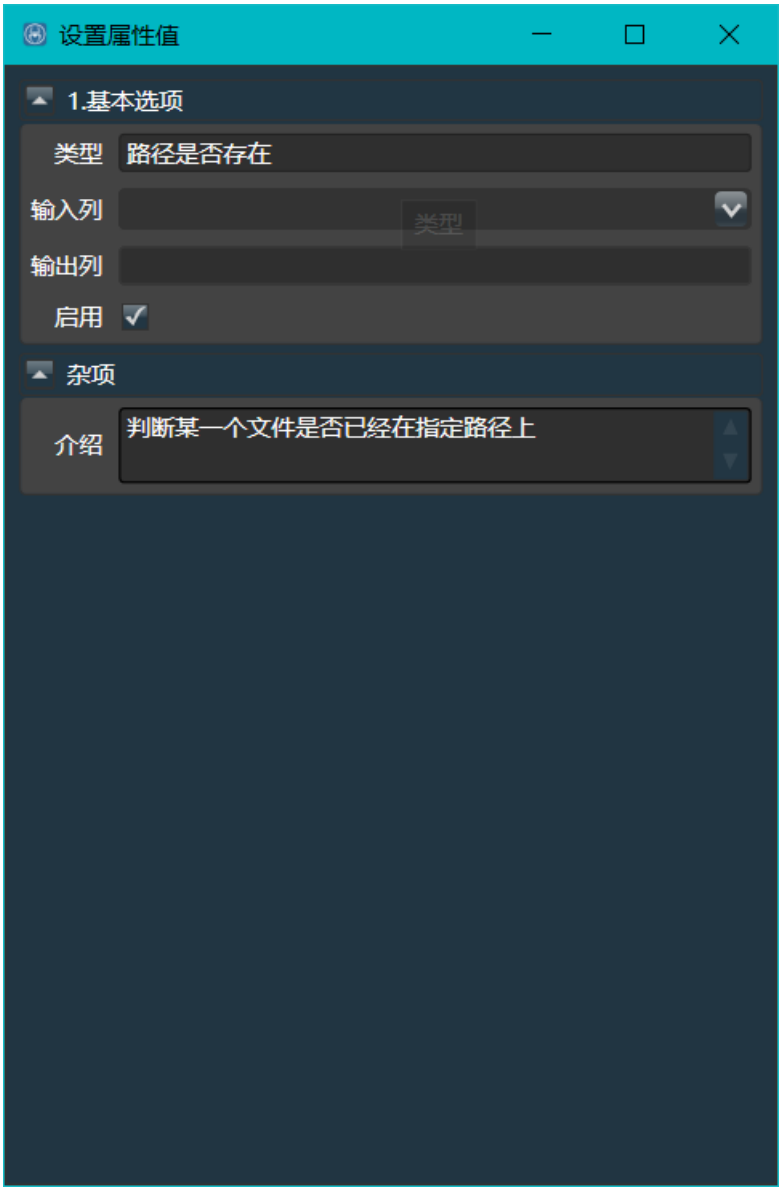
- 标签：自定义模块标签。
- 类型：子任务-转换。
- 原列名：待处理列。
- 启用：是否启用该模块。

杂项：

- 子任务-转换：输入 ETL 的任务名称。（子任务，指的是 ETL 数据清洗任务。）
- 递归到下列：勾选将递归到下列。
- 返回多个数据：勾选可以返回多个数据。
- 新列名：从原始数据中传递到子执行流中的列，多个列用空格分割。

3.1.33 路径是否存在

判断某一个文件是否已经在指定路径上。



基本选项：

- 标签：自定义模块标签。
- 类型：路径是否存在。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

3.1.34 重复项合并

对重复的数据行，进行合并操作。

基本选项：

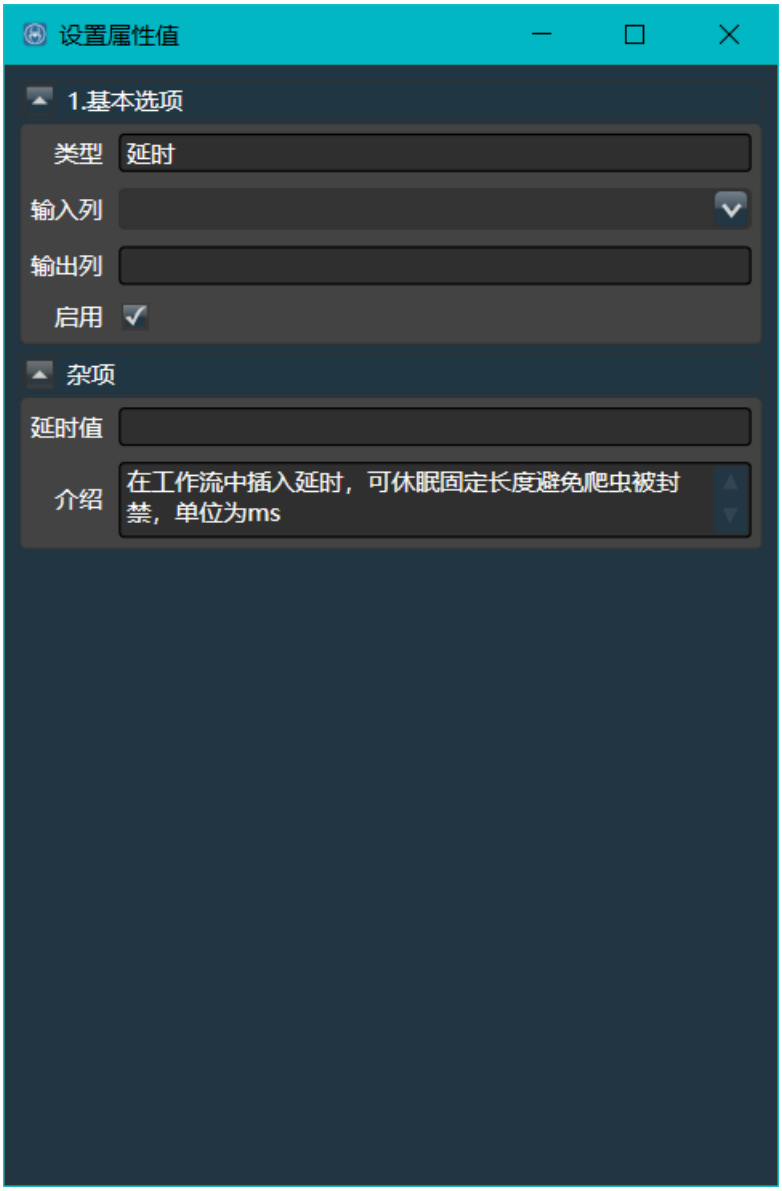
- 标签：自定义模块标签。
- 类型：重复项合并。
- 原列名：待处理列。
- 新列名：结果列。
- 启用：是否启用该模块。

杂项：

- 集合式合并键：空格分割的列名，键相同的写入同一个集合。
- 求和式合并键：空格分割的列名，键相同的所有项进行求和。

3.1.35 延时

在工作流中插入延时，单位为 ms，值为拖入列的值。



基本选项：

- 标签：自定义模块标签。
- 类型：延时。
- 原列名：待处理列（指定待延迟的时间）。
- 新列名：结果列。
- 启用：是否启用该模块。

3.1.36 XPath 筛选器

通过 XPath 选取 html 中的子节点文档。

- 获取节点数量：提取符合条件的节点数量。
- 获取正文：勾选此项，会提取新闻中的正文， XPath 路径可为空。
- 获取正文 HTML：勾选此项，会提取新闻中的 HTML， XPath 路径可为空。
- 插入空行：勾选此项，每个页面后会插入一个空行。
- 获取多个数据：当要获取符合 XPath 语法的多个节点时，勾选此项。

3.2 执行类子模块

- 写入文件文本：写入文件中的文本。
- 子任务-执行：从其他数据清洗模块中生成序列，用以组合大模块。
- 数据库操作：进行数据库操作，包括写入和更新，拖入的列为表的主键。
- 保存超链接文件：目标列需要为超链接类型，会保存链接的文件，如图片，视频等。
- 写入数据库：将数据保存至软件的数据管理器中，之后可方便进行其他处理，拖入到任意一列皆可。

3.2.1 写入文件文本

写入文件中的文本。

- 编码：写入时的文本编码。
- 要写入文件的完整路径：注意要提前建好文本文件，给出完整路径。

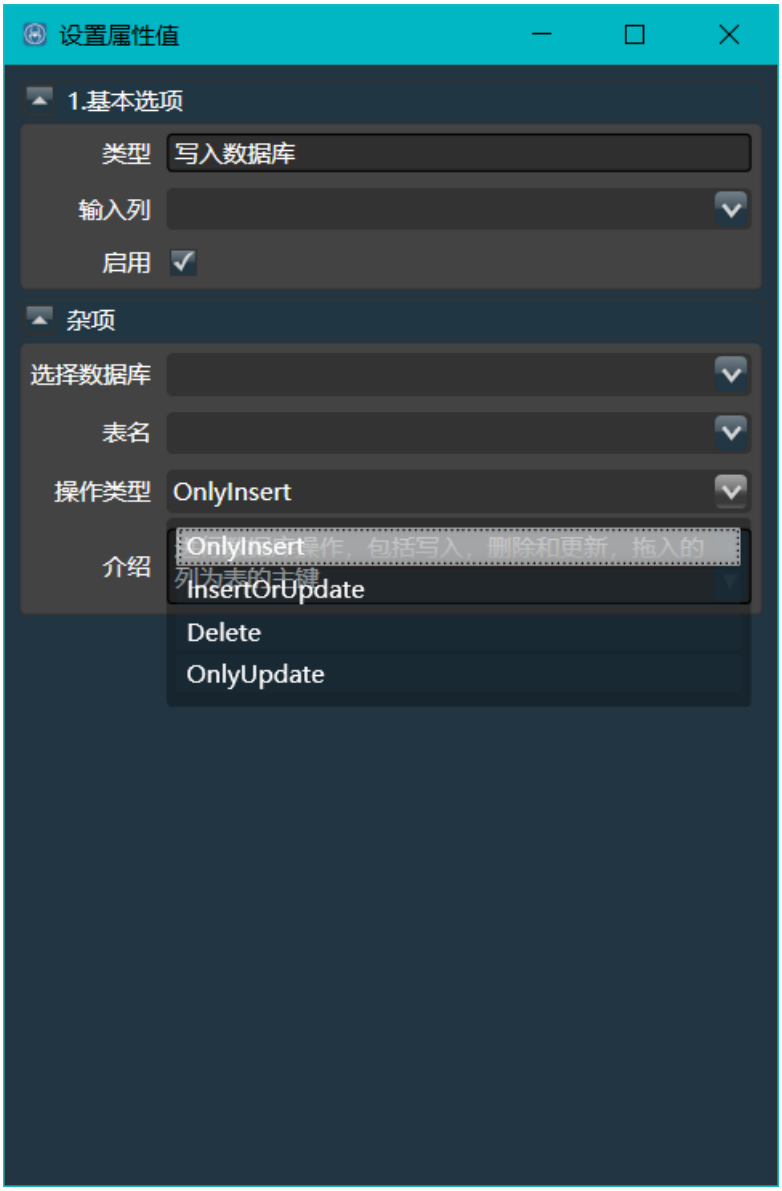
3.2.2 子任务-执行

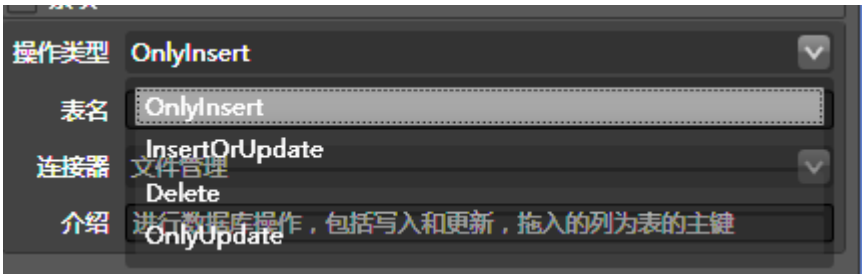
从其他数据清洗模块中生成序列，用以组合大模块。

- 子任务-选择：选择子任务任务。
- 新列名：从原始数据中传递到子执行流中的列，多个列用空格分割。

3.2.3 数据库操作

进行数据库操作，包括写入和更新，拖入的列为表的主键。





- 操作类型：包括对数据库的增删改查。
- 连接器：选择数据源的连接器。
- 表名：拟操作的表名。

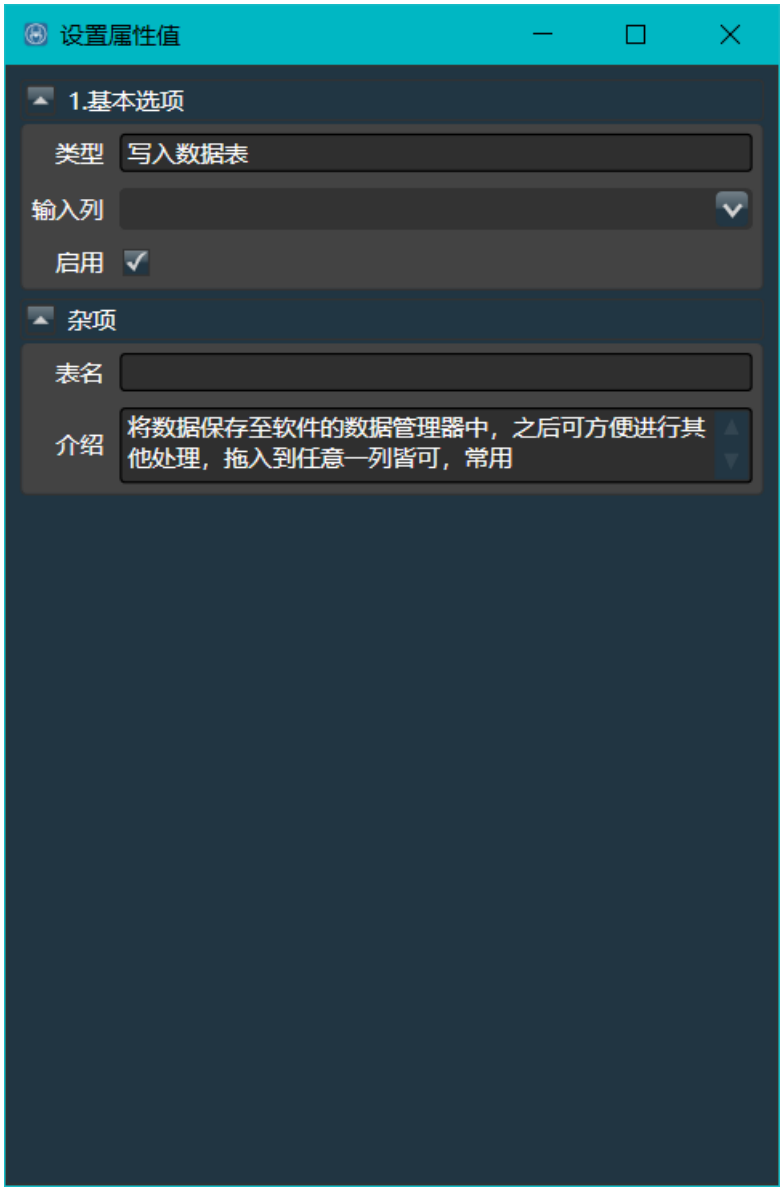
3.2.4 保存超链接文件

目标列需要为超链接类型，会保存链接的文件，如图片，视频等。

- 爬虫选择：要下载的文件的链接。目标列需要为超链接类型，会保存链接的文件，如图片，视频等。
- 是否异步：通常选择异步进行。
- 保存位置：可以指定字段，存放保存路径。

3.2.5 写入数据库

将数据保存至软件的数据管理器中，之后可方便进行其他处理，拖入到任意一列皆可。



- 写入内存数据库中，可在数据视图浏览。

3.3 生成类子模块

- 获取文件夹文件：获取文件夹下的所有文件。
- 从连接器生成：从数据管理的连接器中生成序列。
- 生成区间数：生成某范围内的数值列。
- 从数据表生成：从数据管理中已有的数据表中生成。
- 从文本生成：每行一条数据，生成列。
- 请求队列
- 子任务-生成：从其他数据清洗模块中生成序列，用以组合大模块

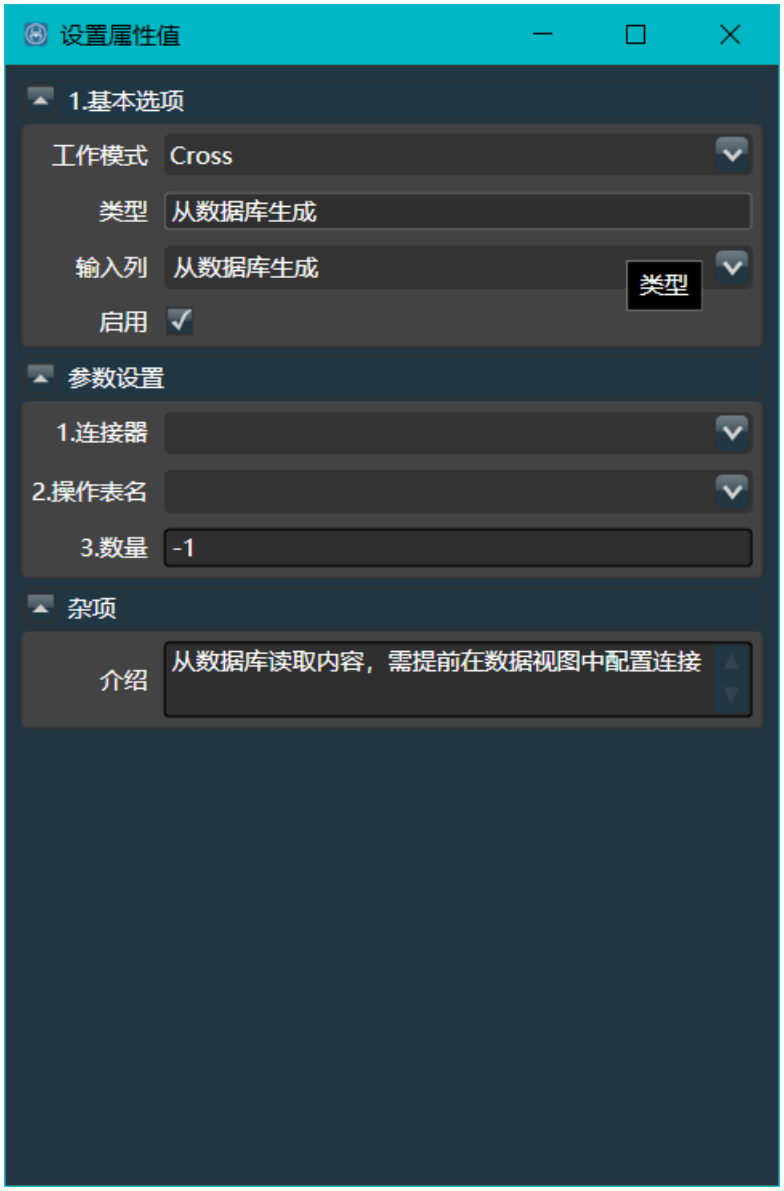
3.3.1 获取文件夹文件

获取文件夹下的所有文件。

- 路径：指定路径，可以通过遍历，获取文件。
- 生成模式：提供多种生成模式。

3.3.2 从数据库生成

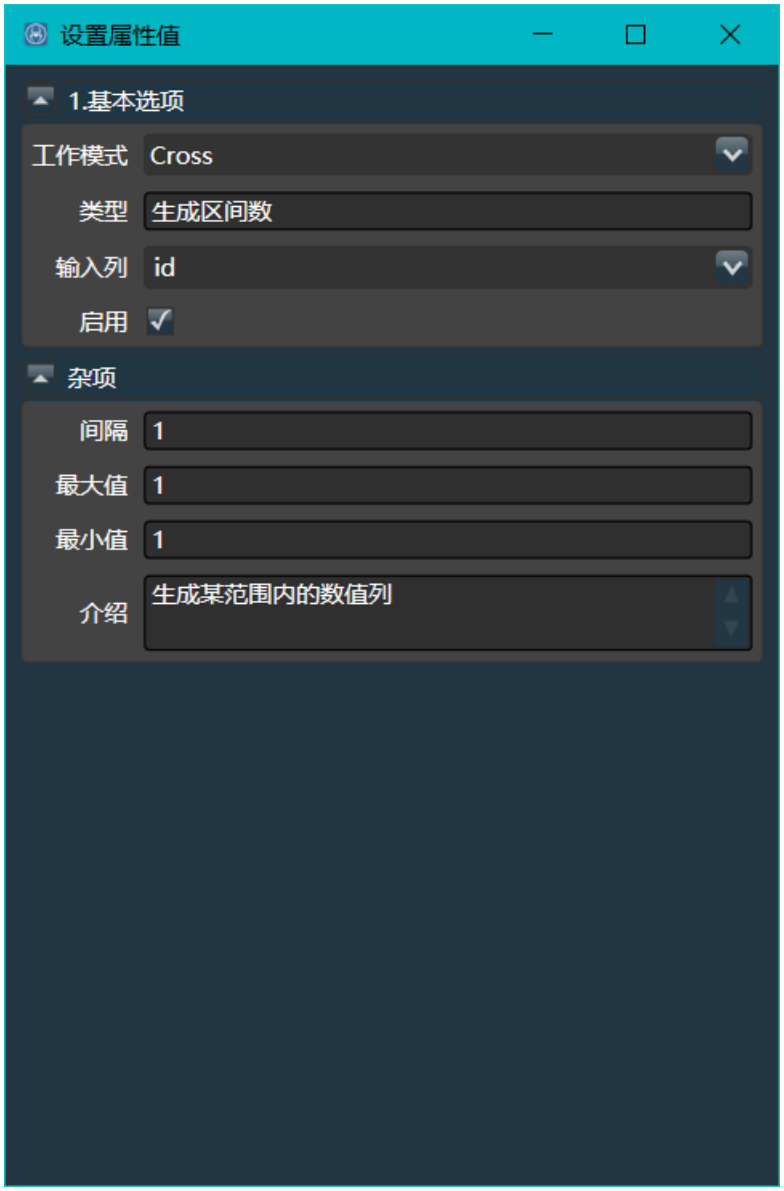
从数据管理的连接器中生成序列。

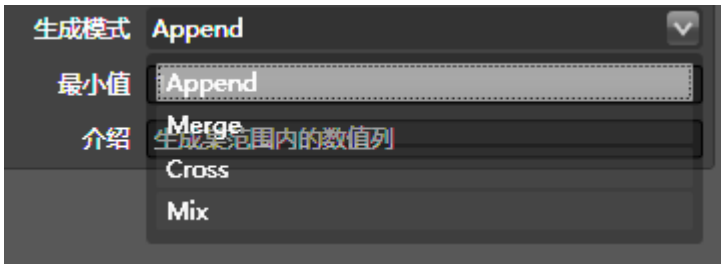


- 操作表名：指的是连接器选定后，所列出的表名。

3.3.3 生成区间数

生成某范围内的数值列。

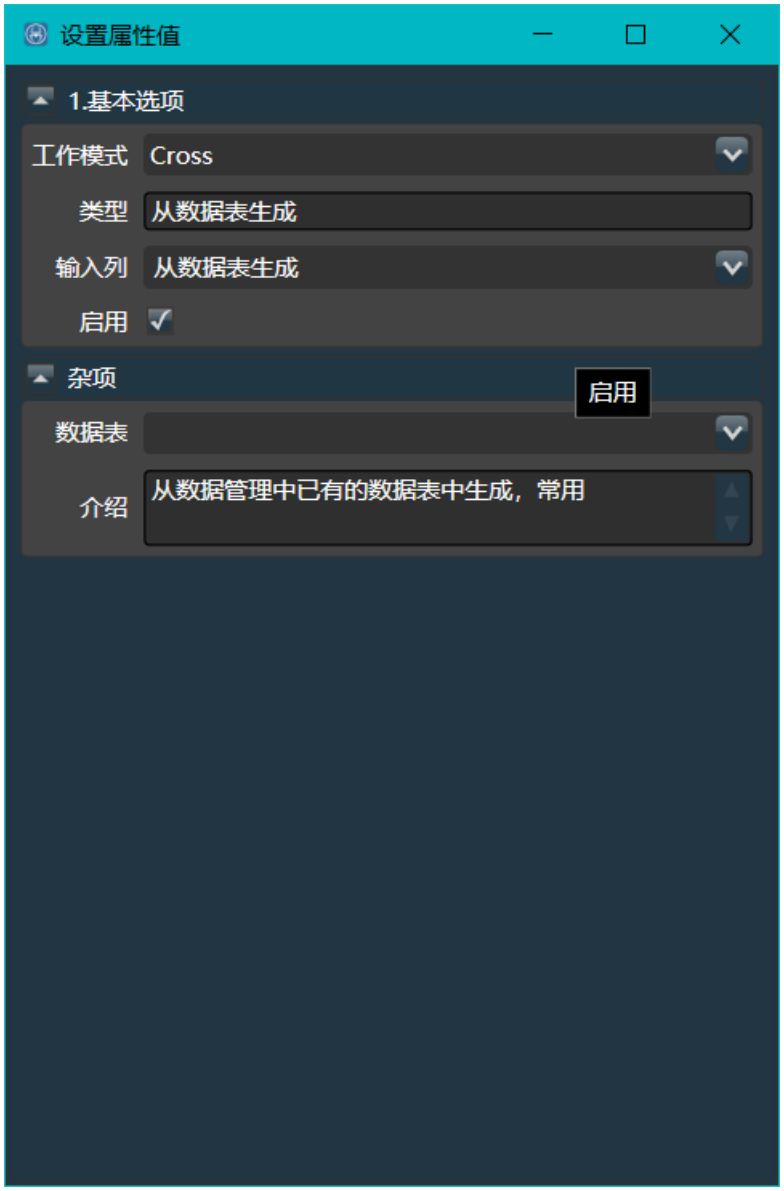




- 最小值、最大值、间隔：生成最小值至最大值时间的数据，中间间隔可以自定义。

3.3.4 从数据表生成

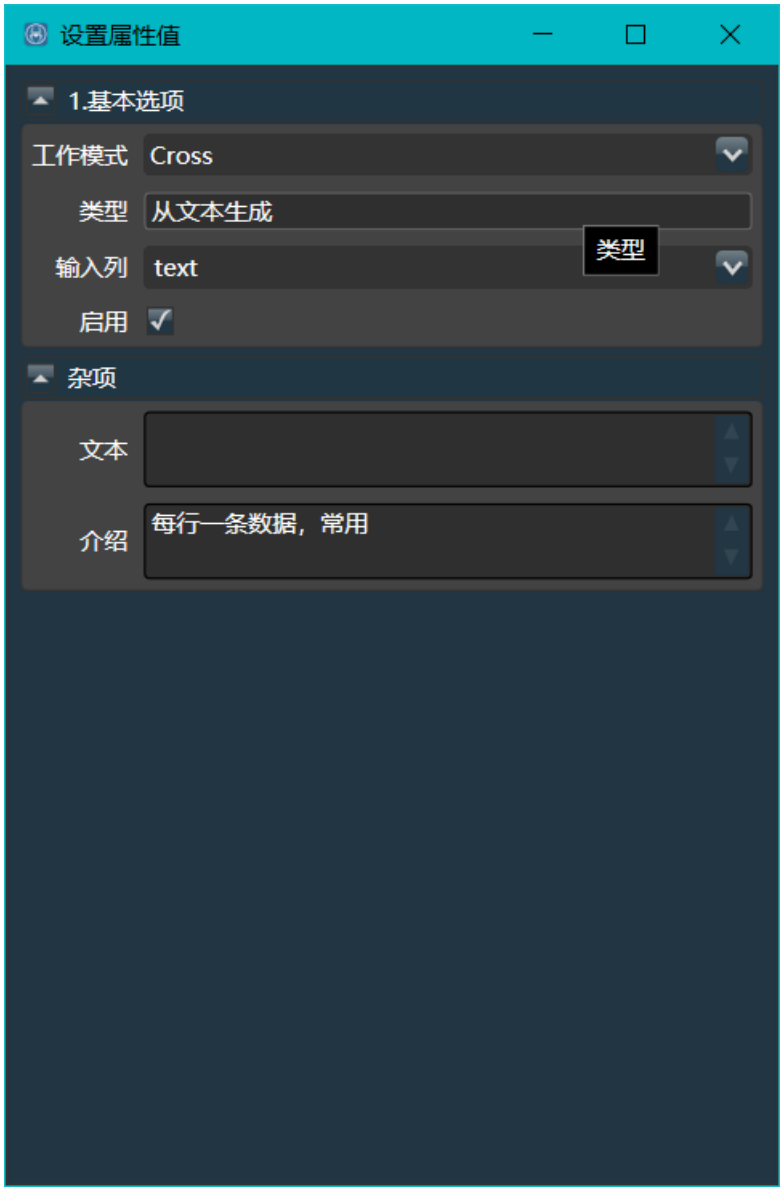
从数据管理中已有的数据表中生成。



- 从数据管理中已有的数据表中生成。提供数据表。

3.3.5 从文本生成

每行一条数据，生成列。



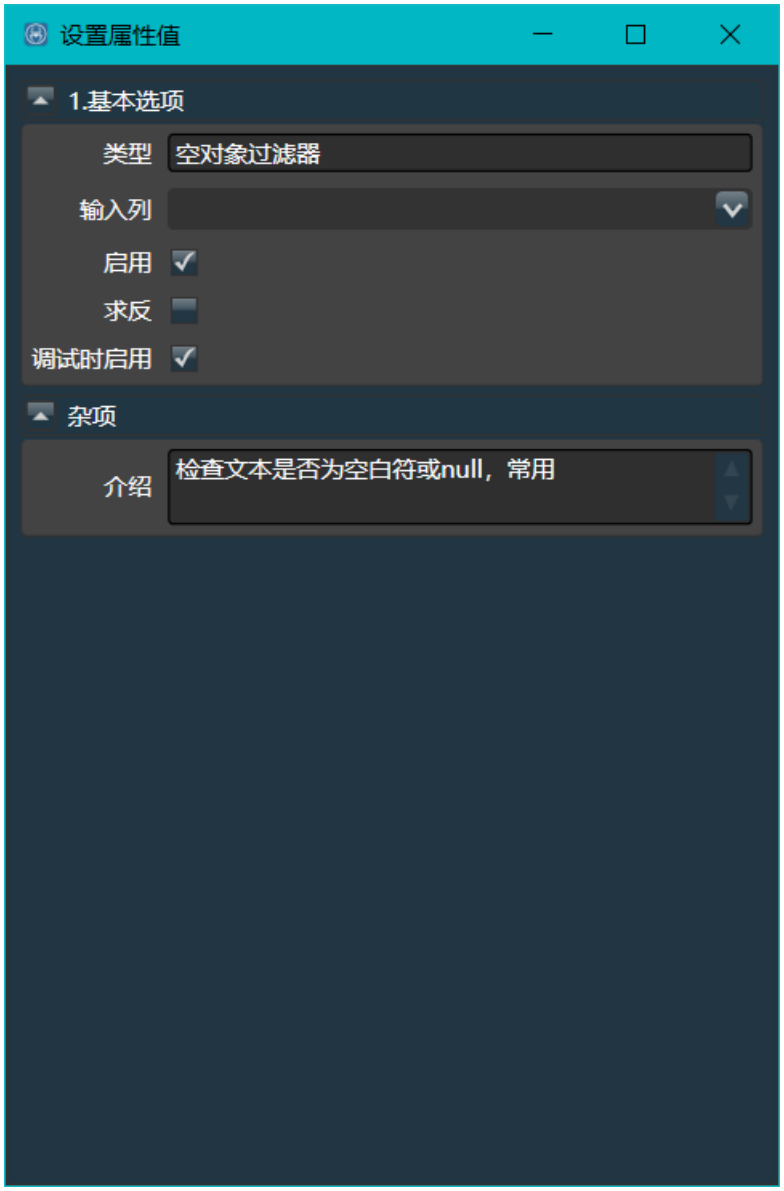
- 每行一条数据，生成列。

3.4 过滤类子模块

- 空对象过滤器：检查文本是否为空白符或 null。
- 数字范围过滤器：从数值列中筛选出从最小值到最大值范围的文档。
- 正则筛选器：编写正则表达式来过滤文本。
- 删除重复项：以拖入的列为键，自动去重，仅保留重复出现的第一项。
- 数量范围选择：选择一定数量的行，如跳过前 100 条，选取 50 条。

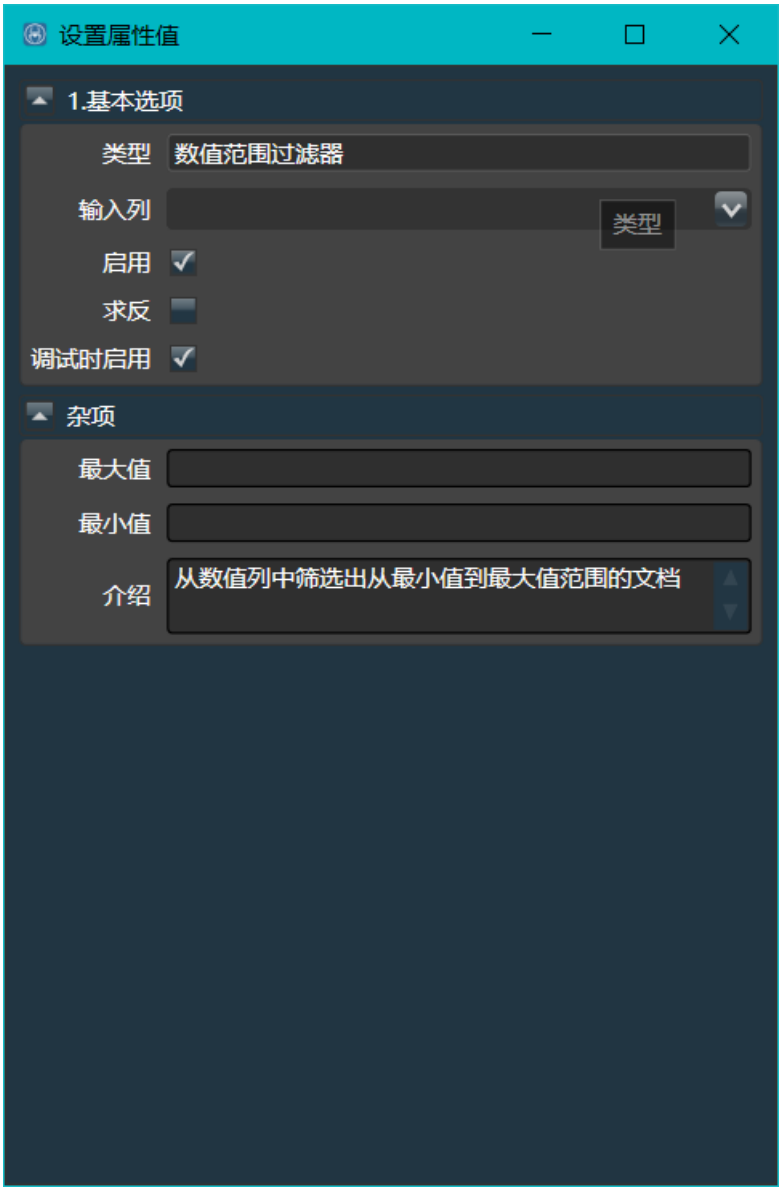
3.4.1 空对象过滤器

检查文本是否为空白符或 nul。



3.4.2 数字范围过滤器

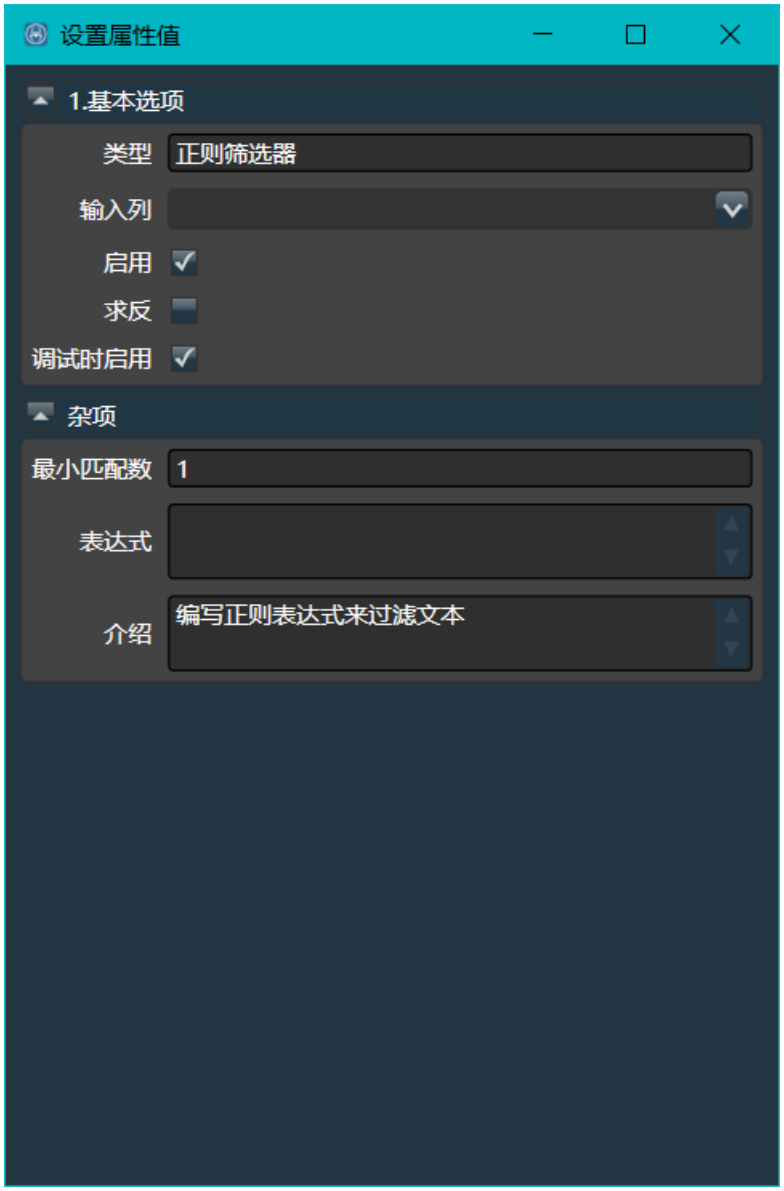
从数值列中筛选出从最小值到最大值范围的文档。



- 最大值、最小值：分别制定范围。从数值列中筛选出从最小值到最大值范围的文档。

3.4.3 正则筛选器

编写正则表达式来过滤文本。



3.4.4 删除重复项

以拖入的列为键，自动去重，仅保留重复出现的第一项。

3.4.5 数量范围选择

选择一定数量的行，如跳过前 100 条，选取 50 条。

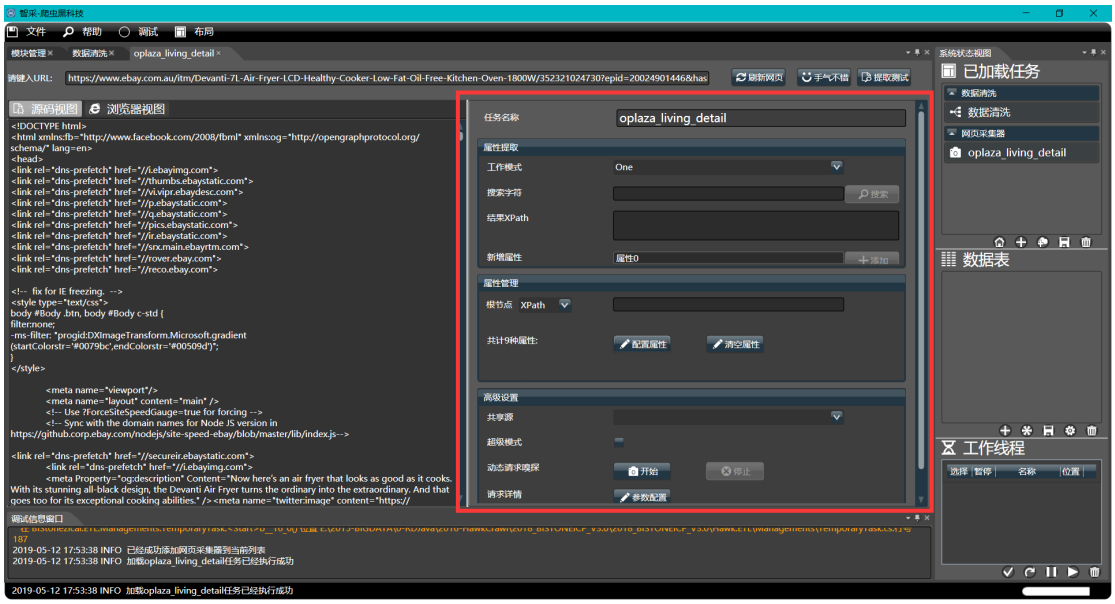

设置属性值
1.基本选项
类型
数量范围选择
输入列
启用
求反
调试时启用
杂项
跳过
0
获取
0
介绍
选择一定数量的行，如跳过前100条，选取50条

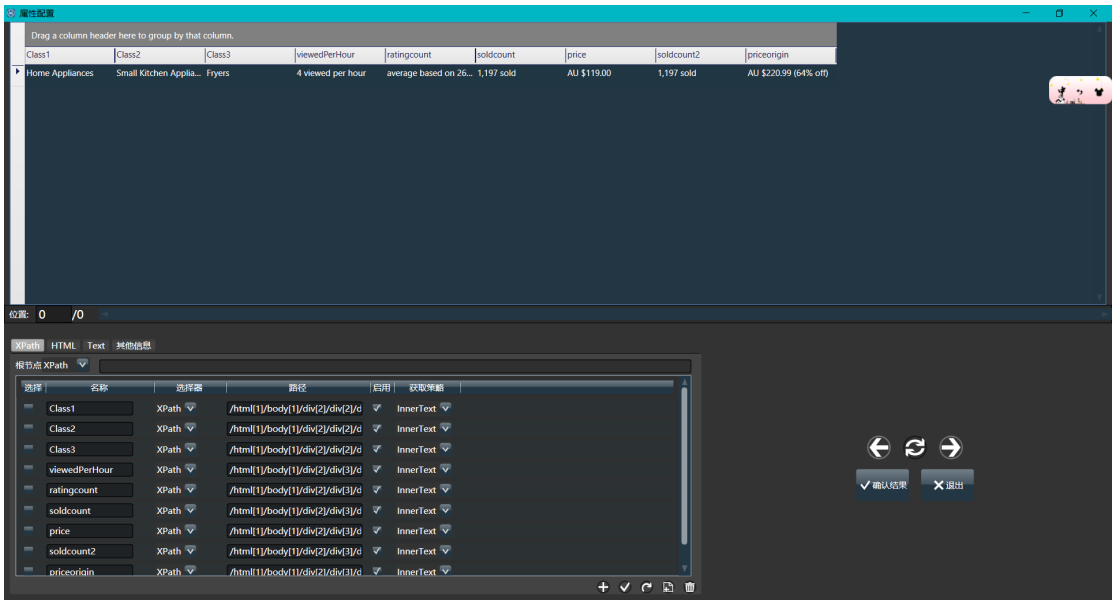
第4章　属性区

属性区，是配置两类主要任务属性的界面。网页采集任务的属性界面，用于指定链接的网页结构中，哪些内容是需要提取的。而数据清洗任务的属性界面，则提供了拖拽式工作流的画面，供组装工作流，预览结果，以及批量执行等功能。

4.1　网页采集器任务的属性配置器

网页采集器任务的属性配置界面如下。





基本信息

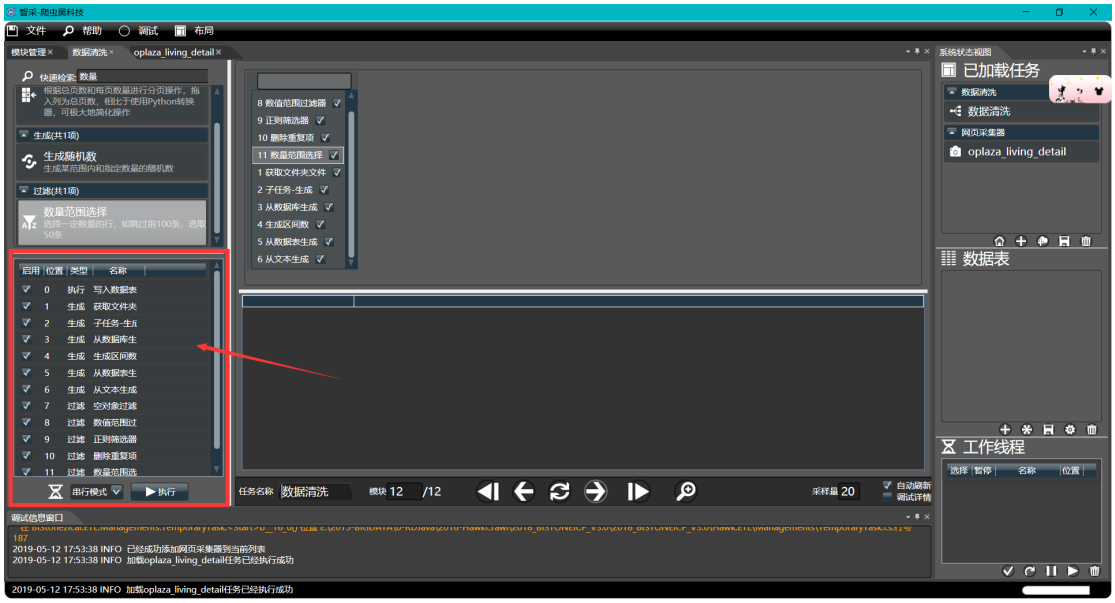
- 模块名称：可以自定义每个模块的名称。

属性提取

- 读取模式：分为 list 模式和 one 模式。前者，是指的一个网页可以生成多行数据；one 模式，指的一个网页对应一条数据。
- 搜索字符：输入想要提取的字符，系统自动搜索其 XPath 语法标记。
- 提取标签：勾选此项，系统可以进入到属性中进行字符搜索。
- 获取的 XPath：系统自动生成的 Xpath。
- 属性名称：对于提取的内容，一一命名，对应最终表格的字段，可以点击添加属性按钮进行添加，也可进行属性列表的编辑和清空操作。
- 手气不错：点击此按钮，系统可以自动搜索信息，节省 XPath 的配置时间。
- 提取测试：可以预览提取的结果。
- 父节点 XPath：所有节点公共的父节点 XPath。
- 请求详情：当前请求的头文件信息，可以更改，比如请求的字符格式等。

自动嗅探：系统提供的自动嗅探工具，帮助更好的了解结构。

4.2 数据清洗任务的属性配置器



基本信息

- 模块名称：可以自定义每个模块的名称。

清洗流程

- 已加载：模块列表，列出了所有模块所组成的工作流。可以编辑和修改。

调试

- 采样量：当前符合工作流的数据，进行采样。
- 命令：刷新当前工作流的运行态，弹出样例，查看目标数据。

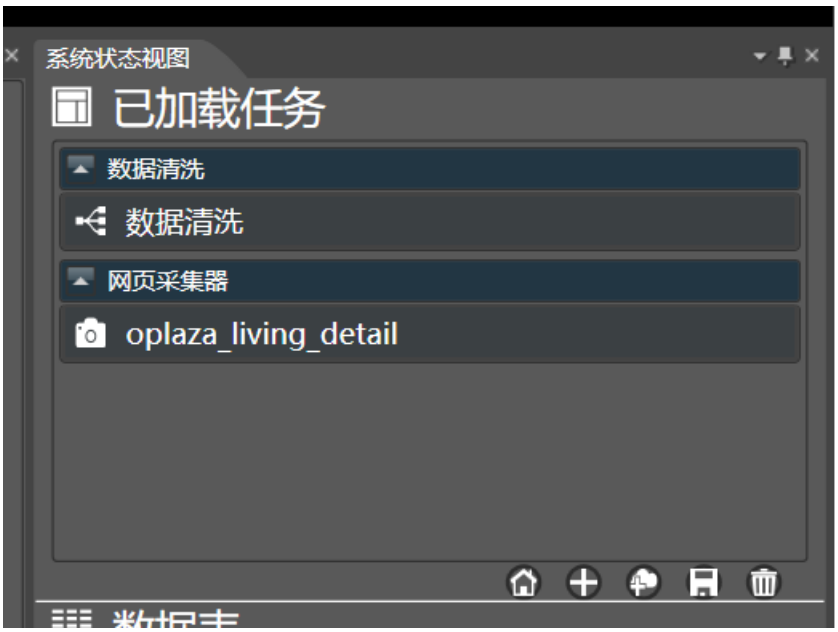
执行

- 工作模式：可以分为并行模式和串行模式。
- 最大线程数：可以自定义最大的线程数量。

第5章　活动资源区

5.1　算法视图

当前活动的网页采集、数据清洗任务列表，双击可以直接运行加载。右键鼠标，可以选择清空加载的任务，可以保存当前的任务。



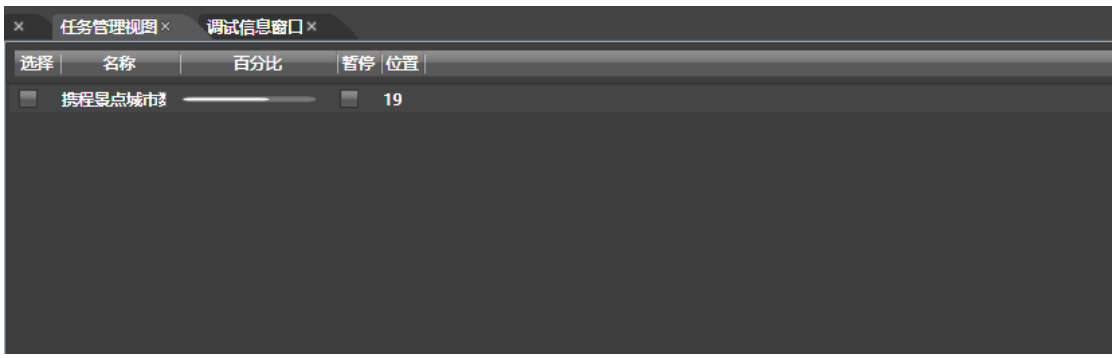
5.2　数据视图

当前运行过程中的内存数据，可以鼠标右键点击任意一个数据，进行导出保存、预览等操作。也可以保存到数据源中。

第6章　日志区

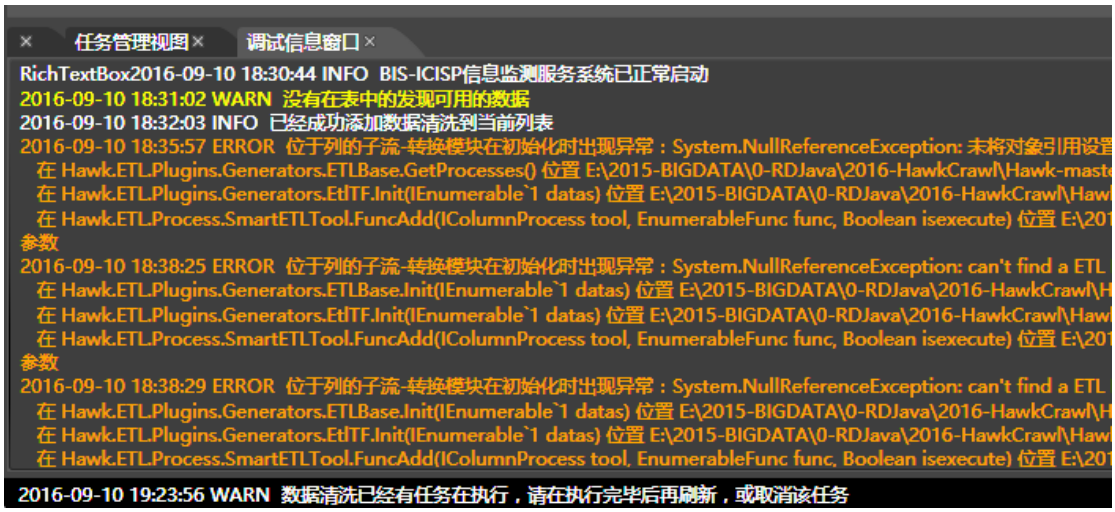
6.1　任务管理视图

用于管理当前的任务，可以进行任务的暂停、删除等操作。



6.2 调试信息窗口

可以实时了解各项任务的运行情况。



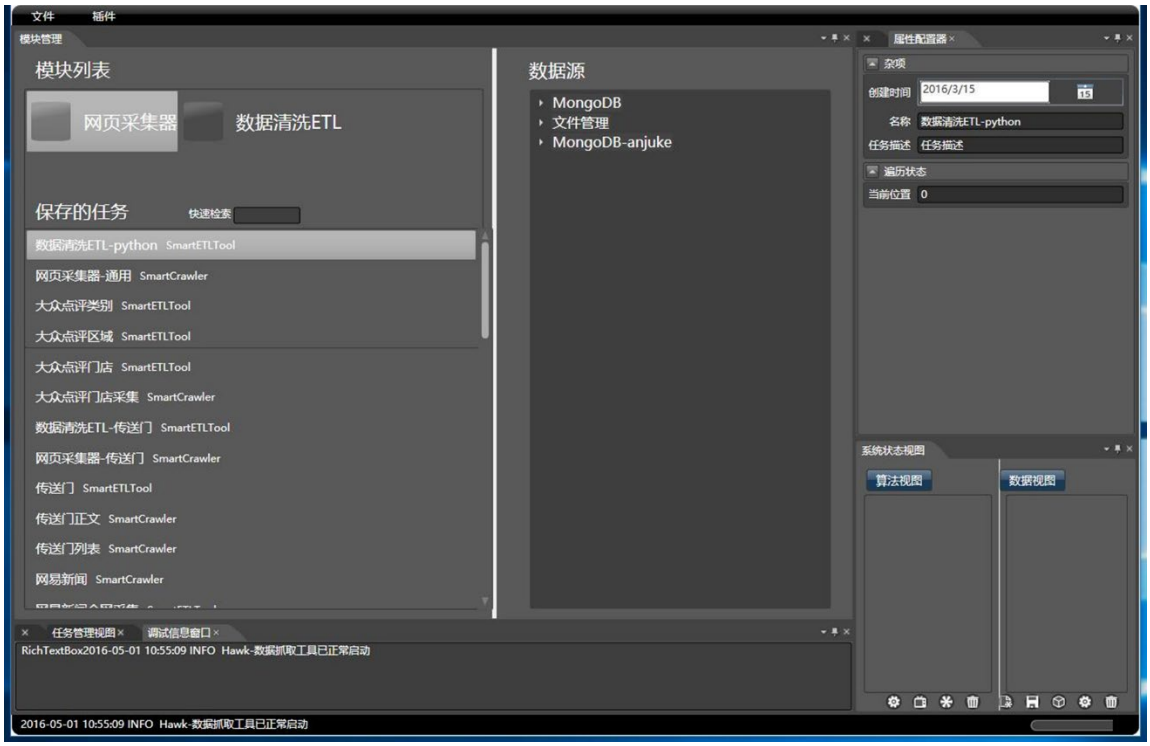
第7章　信息监测服务示例

本文以抓取房地产网站链家为例，介绍了软件的整体使用流程，使用本软件可在 10 分钟内完成配置，在 1 小时之内自动并行抓取全部内容，并能监视子线程工作情况。而手工编写代码，即使是使用 python，一个熟练的程序员也可能需要一天以上：

7.1　界面和组件介绍

7.1.1 界面介绍

软件采用类似 Visual Studio 和 Eclipse 的 Dock 风格，所有的组件都可以悬停和切换。包括以下核心组件：



- 左上角区域：主要工作区，可模块管理。
- 下方： 输出调试信息，和任务管理，监控一项任务完成的百分比。
- 右上方区域： 属性管理器，能对不同的模块设置属性。
- 右下方区域： 显示当前已经加载的所有数据表和模块。

7.1.2 数据管理

能够添加来自不同数据源的连接器， 并对数据进行加载和管理：

在空白处，点击右键，可增加新的连接器。在连接器的数据表上，双击可查看样例，点击右键，可以将数据加载到内存中。

也可以选择加载虚拟数据集，此时系统会维护一个虚拟集合， 当上层请求分页数据时， 动态地访问数据库， 从而有效提升性能。

7.1.3 模块管理

目前系统仅仅提供了两个模块： 网页采集器和数据清洗 ETL， 双击即可加载一个新的模块。

之前配置好的模块，可以保存为任务， 双击可加载一个已有任务：

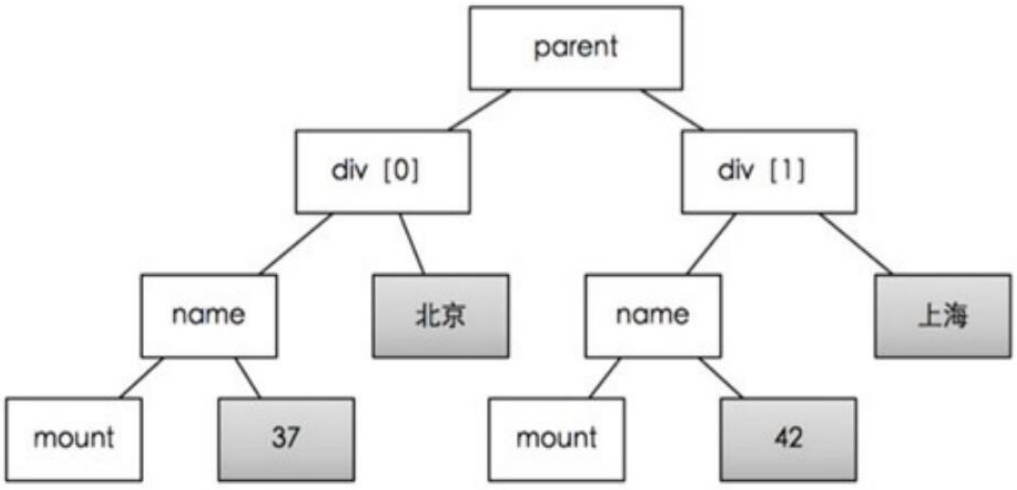
7.1.4 系统状态管理

当加载了数据集或模块时，在系统状态管理中，就可对其查看和编辑：
点击右键，可以对数据集进行删除，修改名称等。也可以将数据集拖拽到下方的图标上，如拖到回收站，即可删除该模块。
双击数据集或模块，可查看模块的内容。 将数据集拖拽到数据清洗（ 数据视图的下方第一个图标），可直接对本数据集做数据清洗。

7.2 网页采集器

7.2.1 原理（建议阅读）

网页采集器的功能是获取网页中的数据。通常来说，目标可能是列表（如购物车列表），或是一个页面中的固定字段（如 JD 某商品的价格和介绍，在页面中只有一个）。因此需要设置其读取模式。传统的采集器需要编写正则表达式，但方法过分复杂。如果认识到 html 是一棵树，只要找到了承载数据的节点即可。XPath 就是一种在树中描述路径的语法。指定 XPath，就能搜索到树中的节点。



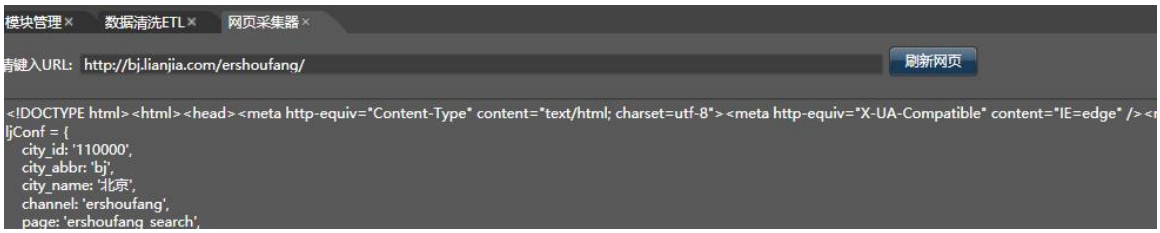
手工编写 XPath 也很复杂，因此软件可以通过关键字，自动检索 XPath，提供关键字，软件就会从树中递归搜索包含该数据的叶子节点。因此关键字最好是在页面中独一无二的。

如上图所示，只要提供“北京”和“42”这两个关键字，就能找到 parent 节点， 进而获取 div[0]和 div1 这两个列表元素。通过 div[0]和 div1 两个节点的比较，我们就能自动发现相同的子节点（name,mount）和不同的节点（北京:上海,37:42）。相同的节点会保存为属性名，不同的节点为属性值。但是，不能提供北京和 37，此时，公共节点是 div[0]， 这不是列表。

软件在不提供关键字的情况下，也能通过 html 文档的特征，去计算最可能是列表父节点（如图中的 parent）的节点，但当网页特别复杂时，猜测可能会出错，所以需要至少提供两个关键字（ 属性）。

7.2.2 基本列表

我们以爬取链家二手房为例，介绍网页采集器的使用。首先双击图标，加载采集器：

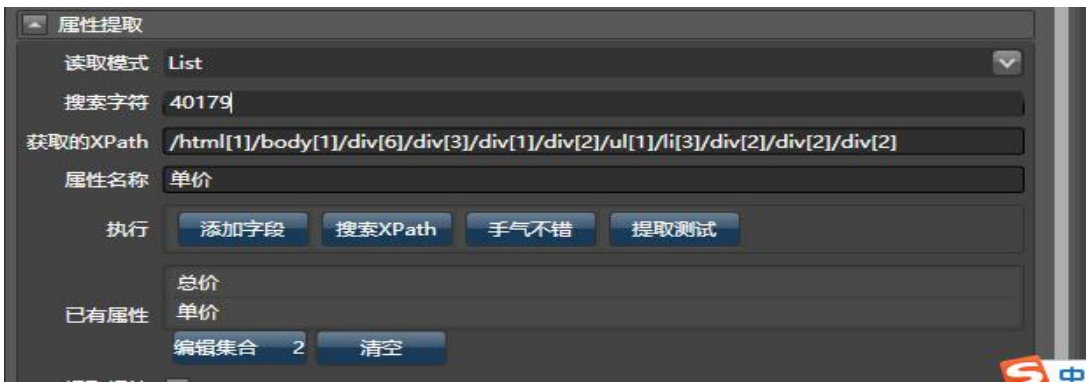


在最上方的地址栏中，输入要采集的目标网址，本次是：
http://bj.lianjia.com/ershoufang/
并点击刷新网页。此时，下方展示的是获取的 html 文本。

原始网站页面如下：
由于软件不知道到底要获取哪些内容，因此需要手工给定几个关键字，让系统搜索关键字，并获取位置。

以上述页面为例，通过检索 820 万和 51789（单价，每次采集时都会有所不同），我们就能通过 DOM 树的路径，找出整个房源列表的根节点。

下面是实际步骤：

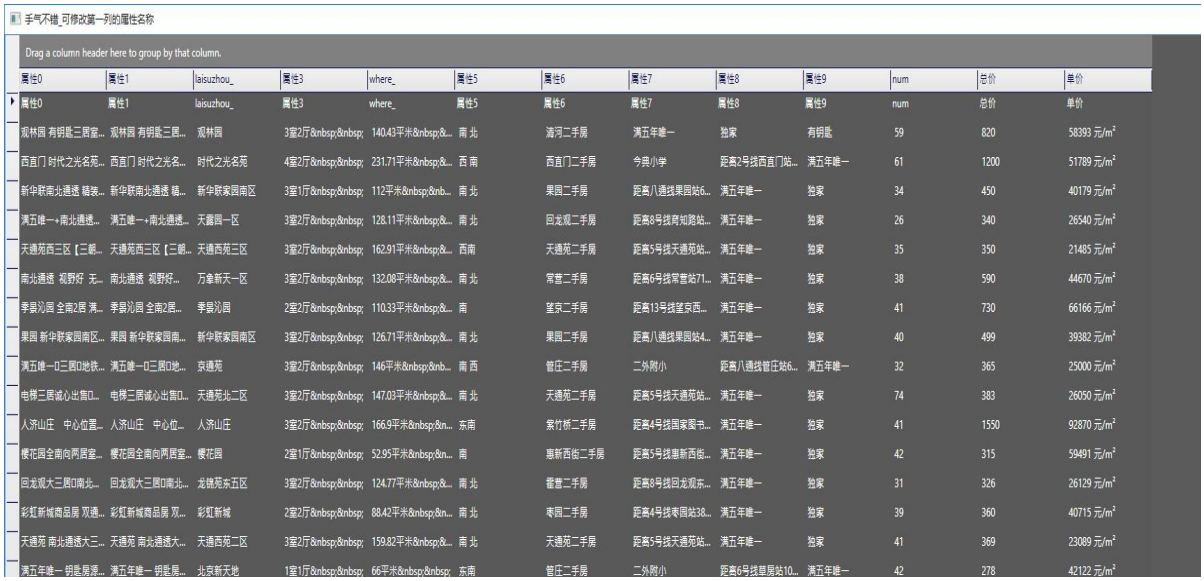


由于要抓取列表，所以**读取模式**选择 List。 填入搜索字符 700， 发现能够成功获取 XPath，

编写属性为“总价” ，点击**添加字段**，即可添加一个属性。类似地，再填入 30535，设置属性名称为“单价”，即可添加另外一个属性。

如果发现有错误，可点击**编辑集合**， 对属性进行删除，修改和排序。

你可以类似的将所有要抓取的特征字段添加进去，或是直接点击**手气不错**，系统会根据目前的属性，推测其他属性：


手气不错_可修改第一列的属性名称

Drag a column header here to group by that column.

属性0	属性1	laisuzhou_	属性3	where_	属性5	属性6	属性7	属性8	属性9	num	总价	单价
属性0	属性1	laisuzhou_	属性3	where_	属性5	属性6	属性7	属性8	属性9	num	总价	单价
观林园 有钥匙三居室...	观林园 有钥匙三居...	观林园	3室2厅	140.43平米 &...	南 北	滹河二手房	满五年唯一	独家	有钥匙	59	820	58393 元/m²
西直门 时代之光名苑...	西直门 时代之光名...	时代之光名苑	4室2厅	231.71平米 &...	西 南	西直门二手房	今典小学	距离2号线西直门站...	满五年唯一	61	1200	51789 元/m²
新华联南北通透 精装...	新华联南北通透 精...	新华联家园南区	3室1厅	112平米 &nb...	南 北	果园二手房	距离八通线果园站6...	满五年唯一	独家	34	450	40179 元/m²
满五唯一+南北通透...	满五唯一+南北通透...	天露园一区	3室2厅	128.11平米 &...	南 北	回龙观二手房	距离8号线育知路站...	满五年唯一	独家	26	340	26540 元/m²
天通苑西三区【三顺...	天通苑西三区【三顺...	天通西苑三区	3室2厅	162.91平米 &...	西南	天通苑二手房	距离5号线天通苑站...	满五年唯一	独家	35	350	21485 元/m²
南北通透 视野好 无...	南北通透 视野好...	万象新天一区	3室2厅	132.08平米 &...	南 北	常营二手房	距离6号线常营站71...	满五年唯一	独家	38	590	44670 元/m²
季景沁园 全南2居 满...	季景沁园 全南2居...	季景沁园	2室2厅	110.33平米 &...	南	望京二手房	距离13号线望京西...	满五年唯一	独家	41	730	66166 元/m²
果园 新华联家园南区...	果园 新华联家园南...	新华联家园南区	3室2厅	126.71平米 &...	南 北	果园二手房	距离八通线果园站4...	满五年唯一	独家	40	499	39382 元/m²
满五唯一0三居0地铁...	满五唯一0三居0地...	京通苑	3室2厅	146平米 &nb...	南 西	管庄二手房	二外附小	距离八通线管庄站6...	满五年唯一	32	365	25000 元/m²
电梯三居诚心出售0...	电梯三居诚心出售0...	天通苑北二区	3室2厅	147.03平米 &...	南 北	天通苑二手房	距离5号线天通苑站...	满五年唯一	独家	74	383	26050 元/m²
人济山庄 中心位置...	人济山庄 中心位...	人济山庄	3室2厅	166.9平米 &n...	东南	紫竹桥二手房	距离4号线国家图书...	满五年唯一	独家	41	1550	92870 元/m²
樱花园全南向两居室...	樱花园全南向两居室...	樱花园	2室1厅	52.95平米 &n...	南	惠新西街二手房	距离5号线惠新西街...	满五年唯一	独家	42	315	59491 元/m²
回龙观大三居0南北...	回龙观大三居0南北...	龙腾苑东五区	3室2厅	124.77平米 &...	南 北	霍营二手房	距离8号线回龙观东...	满五年唯一	独家	31	326	26129 元/m²
彩虹新城商品房 双通...	彩虹新城商品房 双...	彩虹新城	2室2厅	88.42平米 &n...	南 北	枣园二手房	距离4号线枣园站38...	满五年唯一	独家	39	360	40715 元/m²
天通苑 南北通透大三...	天通苑 南北通透大...	天通西苑二区	3室2厅	159.82平米 &...	南 北	天通苑二手房	距离5号线天通苑站...	满五年唯一	独家	41	369	23089 元/m²
满五年唯一 钥匙房源...	满五年唯一 钥匙房...	北京新天地	1室1厅	66平米	东南	管庄二手房	二外附小	距离6号线草房站10...	满五年唯一	42	278	42122 元/m²

属性的名称是自动推断的，如果不满意，可以**修改列表第一列的属性名**， 在对应的列中敲键盘**回车**提交修改。之后系统就会自动将这些属性添加到属性列表中。

工作过程中，可点击**提取测试** ，随时查看采集器目前的能够抓取的数据内容。这样，一个链家二手房的网页采集器即可完成。可属性管理器的上方，可以修改采集器的模块名称，这样就方便**数据清洗**模块调用该采集器。

7.3 数据清洗

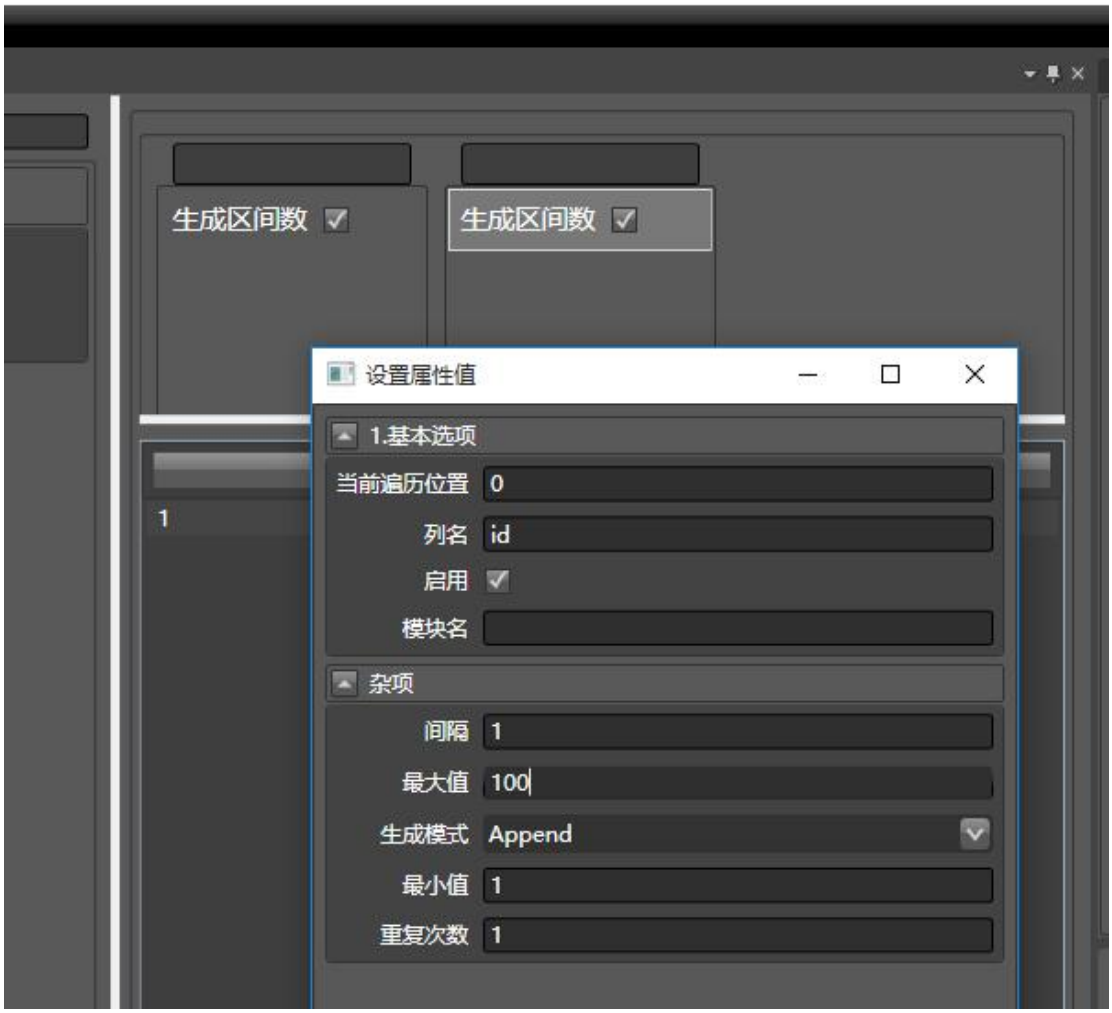
数据清洗模块，包括几十个子模块，这些子模块包含四类：生成，转换，过滤和执行。

7.3.1 构造 url 列表

在前面介绍了如何实现一个页面的采集，但如何采集所有二手房数据呢？这涉及到翻页。

还是以链家为例，翻页时，我们会看到页面是这样变换的：
http://bj.lianjia.com/ershoufang/pg2/
http://bj.lianjia.com/ershoufang/pg3/
…
因此，需要构造一串上面的 url. 聪明的你肯定会想到， 应当先生成一组序列， 从 1 到 100（假设我们只抓取前 100 页）。

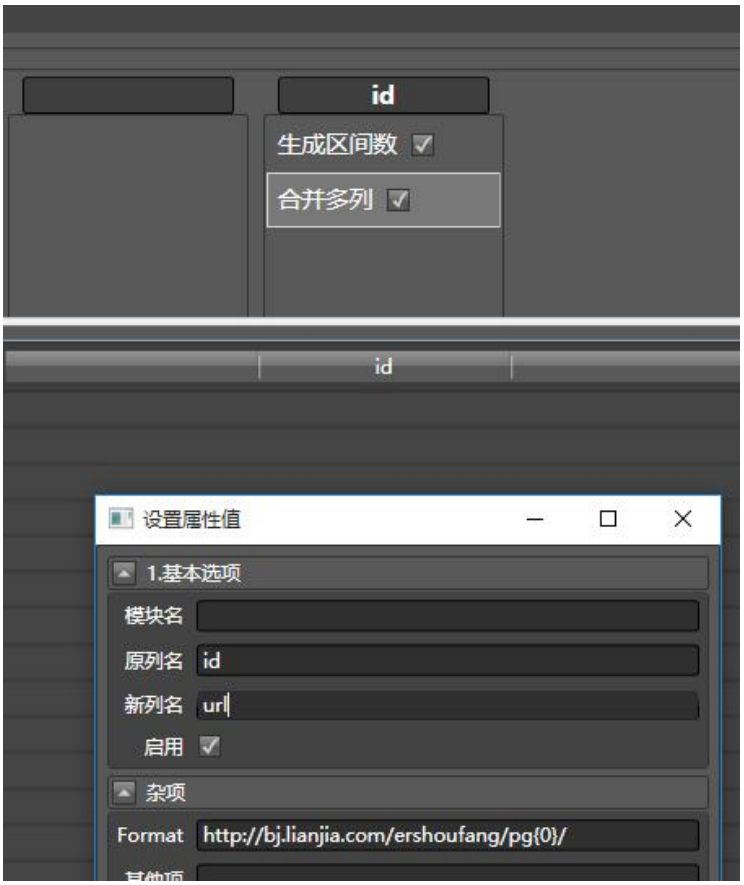
双击**数据清洗** ETL 左侧的搜索栏中搜索**生成区间数**， 将该模块拖到右侧上方的栏目中：



在右侧栏目中双击生成区间数，可弹出设置窗口， 为该列起名字(id)， 最大值填写为 100, 生成模式默认为 Append:

*为什么只显示了前 20 个？ 这是程序的虚拟化机制， 并没有加载全部的数据，可在 ETL 属性的***调试***栏目中，修改***采样量*** *（默认为 20）。*

将数字转换为 url， 熟悉 C#的读者，可以想到 string.format， 或者 python 的%符号：搜索**合并多列**，并将其拖拽到刚才生成的 id 列， 编写 format 为下图的格式，即可将原先的数值列变换为一组 url。



(如果需要多个列合并为一个列，则在“其他项” 栏目中填写其他列的列名,用空格分割，并在 format 中用{1},{2}..等表示)

(由于设计的问题，数据查看器的宽度不超过 150 像素，因此对长文本显示不全，可以在右侧属性对话框点击**查看样例**， 弹出的编辑器可支持拷贝数据和修改列宽。

7.3.2 使用配置好的网页采集器

生成这串 URL 之后，我们即可将刚才已经完成的网页采集器与这串 url 进行合并。

拖拽**从爬虫转换**到当前的 url，双击该模块：将刚才的网页采集器的名称， 填入**爬虫选择**栏目中。

之后，系统就会转换出爬取的前 20 条数据：

属性0	属性1	laisuzhou_	属性3	where_	属性5	属性6
观林园 有钥匙三居室 南北	观林园 有钥匙三居室 南北	观林园	3室2厅	140.43平米	南 北	清河二手房
西直门 时代之光名苑 西南	西直门 时代之光名苑 西南	时代之光名苑	4室2厅	231.71平米	西 南	西直门二手房
新华联南北通透 精装 免税	新华联南北通透 精装 免税	新华联家园南区	3室1厅	112平米	南 北	果园二手房
满五唯一+南北通透+总价	满五唯一+南北通透+总价	天露园一区	3室2厅	128.11平米	南 北	回龙观二手房
天通苑西三区【三朗南明卫	天通苑西三区【三朗南明卫	天通西苑三区	3室2厅	162.91平米	西南	天通苑二手房
南北通透 视野好 无遮挡大	南北通透 视野好 无遮挡大	万象新天一区	3室2厅	132.08平米	南 北	常营二手房
季景沁园 全南2居 满五年唯	季景沁园 全南2居 满五年唯	季景沁园	2室2厅	110.33平米	南	望京二手房
果园 新华联家园南区 精装	果园 新华联家园南区 精装	新华联家园南区	3室2厅	126.71平米	南 北	果园二手房
满五唯一，三居，地铁房，	满五唯一，三居，地铁房，	京通苑	3室2厅	146平米	南 西	管庄二手房
电梯三居诚心出售，送全新	电梯三居诚心出售，送全新	天通苑北二区	3室2厅	147.03平米	南 北	天通苑二手房
人济山庄 中心位置 东南	人济山庄 中心位置 东南	人济山庄	3室2厅	166.9平米	东南	紫竹桥二手房
樱花园全南向两居室，满五	樱花园全南向两居室，满五	樱花园	2室1厅	52.95平米	南	惠新西街二手房
回龙观大三居，南北通透，	回龙观大三居，南北通透，	龙腾苑东五区	3室2厅	124.77平米	南 北	霍营二手房
彩虹新城商品房 双通透格局	彩虹新城商品房 双通透格局	彩虹新城	2室2厅	88.42平米	南 北	枣园二手房
天通苑 南北通透大三居 高	天通苑 南北通透大三居 高	天通西苑二区	3室2厅	159.82平米	南 北	天通苑二手房
满五年唯一 钥匙房源方便看	满五年唯一 钥匙房源方便看	北京新天地	1室1厅	66平米	东南	管庄二手房
龙乡小区低楼层两居室 楼层	龙乡小区低楼层两居室 楼层	龙乡东区	2室1厅	67平米	南 北	西三旗二手房

可以看到， 数据中“属性 3” 列包含 html 转义符， 拖拽 **html 字符转义**，到属性 3 列，即可自动转换所有转义符。

属性0	属性1	laisuzhou_	属性3	where_	属性5	属性6
观林园 有钥匙三居室 南北	观林园 有钥匙三居室 南北	观林园	3室2厅	140.43平米	南 北	清河二手房
西直门 时代之光名苑 西南	西直门 时代之光名苑 西南	时代之光名苑	4室2厅	231.71平米	西 南	西直门二手房
新华联南北通透 精装 免税	新华联南北通透 精装 免税	新华联家园南区	3室1厅	112平米	南 北	果园二手房
满五唯一+南北通透+总价	满五唯一+南北通透+总价	天露园一区	3室2厅	128.11平米	南 北	回龙观二手房
天通苑西三区【三朗南明卫	天通苑西三区【三朗南明卫	天通西苑三区	3室2厅	162.91平米	西南	天通苑二手房
南北通透 视野好 无遮挡大	南北通透 视野好 无遮挡大	万象新天一区	3室2厅	132.08平米	南 北	常营二手房
季景沁园 全南2居 满五年唯	季景沁园 全南2居 满五年唯	季景沁园	2室2厅	110.33平米	南	望京二手房
果园 新华联家园南区 精装	果园 新华联家园南区 精装	新华联家园南区	3室2厅	126.71平米	南 北	果园二手房
满五唯一，三居，地铁房，	满五唯一，三居，地铁房，	京通苑	3室2厅	146平米	南 西	管庄二手房
电梯三居诚心出售，送全新	电梯三居诚心出售，送全新	天通苑北二区	3室2厅	147.03平米	南 北	天通苑二手房
人济山庄 中心位置 东南	人济山庄 中心位置 东南	人济山庄	3室2厅	166.9平米	东南	紫竹桥二手房
樱花园全南向两居室，满五	樱花园全南向两居室，满五	樱花园	2室1厅	52.95平米	南	惠新西街二手房

如果要修改列名，在最上方的列名上直接修改， 点击回车即可修改名字。

where（面积）列中包含数字，想把数字提取出来，可以将**提取数字**模块拖拽到该列上，所有数字即可提取出来。

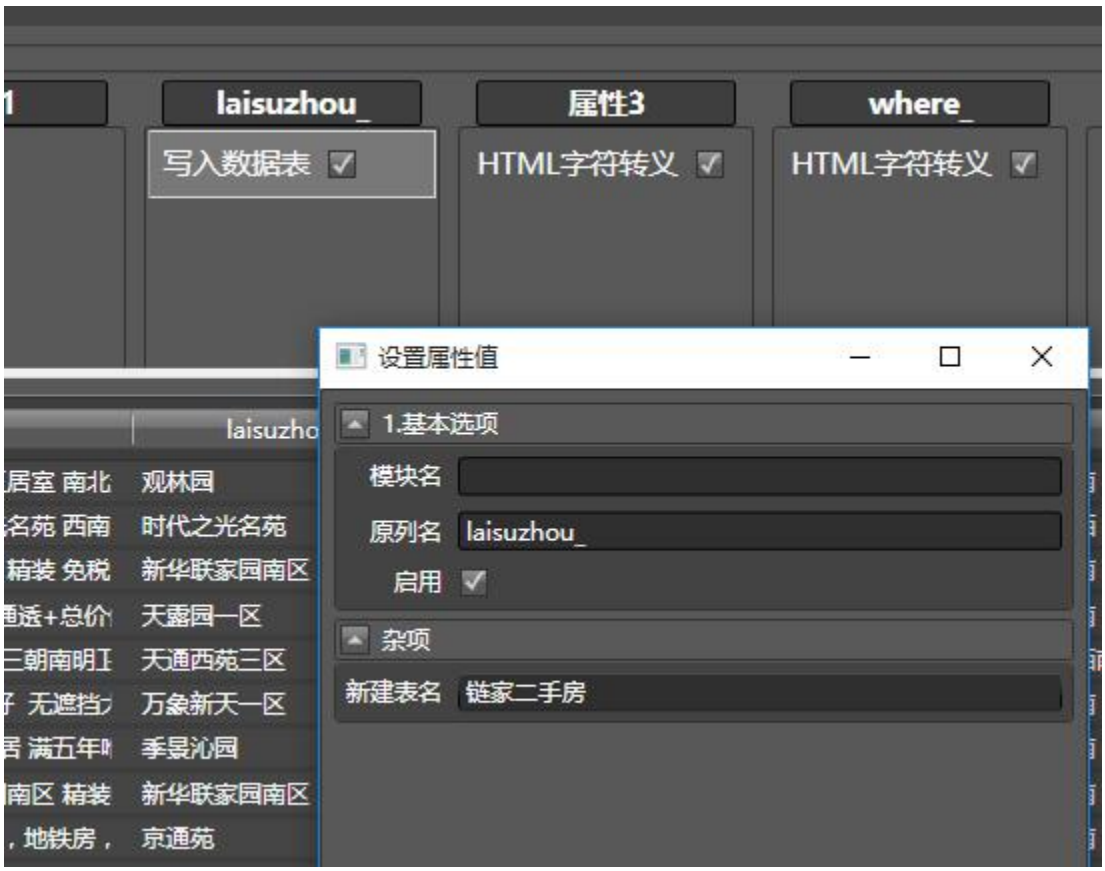
类似地，可以拖拽**字符分割**或**正则分割** 到某一列，从而分割文本和替换文本。此处不再赘述。

有一些列为空，可以拖拽**空对象过滤器** 到该列，那么该列为空的话会自动过滤这一行数据。

7.3.3 保存和导出数据

需要保存数据时，可以选择写入文件，或者是临时存储（本软件的数据管理器），或是数据库。因此可以将“执行” 模块， 拖入清洗链的后端：

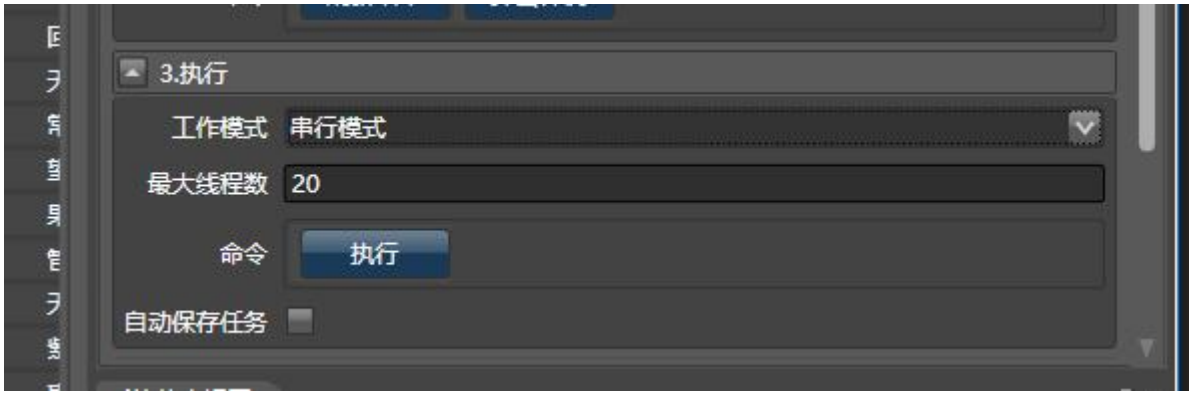
拖**写入数据表**到任意一列， 并填入**新建表名**(如链家二手房)。



下图是这次操作的所有子模块列表：

之后，即可对整个过程进行操作：

选择**串行模式**或**并行模式**， 并行模式使用线程池， 可设定最多同时执行的线程数（最好不要超过 100）。推荐使用并行模式，



点击**执行**按钮，即可在任务管理视图中采集数据。

之后，在**数据管理**的数据表**链家二手房**上点击右键， 选择另存为， 导出到 Excel, Json 等，即可将原始数据导出到外部文件中。

类似的，你可以在清洗流程中拖入执行器，则保存的是中间过程，也可以在结尾拖入多个执行器，这样就能同时写入数据库或文件，从而获得了极大的灵活性。

7.3.4 保存任务

在右下角的**算法视图**中的任意模块上点击右键，保存任务，即可在任务视图中保存新任务（任务名称与当前模块名字一致），下次可直接加载即可。如果存在同名任务， 则会对原有任务进行覆盖。

在**算法视图**的空白处，点击**保存所有模块**, 会批量保存所有的任务。

你可以将一批任务，保存为一个工程文件(xml)，并在之后将其加载和分发。

7.4 总结

上文以抓取房地产网站链家为例，介绍了软件的整体使用流程。当然，本系统功能远远不限于这些。

- 智采新版本特点：

序号	功能描述
1	增加 Mysql 数据库连接，可以支持读写 Mysql 数据库。
2	优化“手气不错”。增加排序策略，分为“按面积排序”、“按列数排序”、“按行数排序”、“按分数排序”。不同的排序策略，可以主动搜索到的数据内容不同。
3	优化从数据表导出到 Excel 功能，大大提升数据导出的效率。
4	布局大幅度优化和调整，更为美观和易用。
5	增加全局登录，为未来的线上模板大小基础
6、	为每一个节点增加文字解释