

DataC 使用教程

V2025.1.0

1. 概述	3
2. HTTP 插件	3
2.1. HTTP 插件内容	3
2.2. 插件方法	4
3. 大模型	5
4. 任务流程编排	5
5. 数据表定义	6
6. 变量定义	8
7. 浏览器界面	8
8. 流程节点	9
8.1. 开始	9
8.2. HTTP 插件方法节点	9
8.3. 数据表节点	9
8.4. 浏览器方法节点	11
1.1.1. 加载页面	11
1.1.2. 内容提取	12
1.1.3. 等待元素	12
1.1.4. 元素赋值	12
1.1.5. 元素取值	13
1.1.6. 元素点击	13
8.5. 变量	13
8.5.1. 变量读取	13
8.5.2. 变量写入	14
8.6. IF 条件节点	14
8.7. 对象合并	15
8.8. 字符串拼接	15
8.9. 数字运算	16
8.10. 大模型	16
8.11. 用户输入	17
8.12. 正则表达式	17

1. 概述

dataC（中文名：灯塔）是一款基于流程定义的智能数据处理工具，其核心定位是解决异构系统间的数据转换问题。该工具支持对接多种异构数据源，包括 **Web API** 接口、网页数据以及本地文件系统等，并通过 **AI** 技术实现高效的数据采集与处理。

使用 **dataC** 采集网页数据非常方便。

在功能架构上，**dataC** 采用任务驱动的设计模式：

1. **任务管理**：每个任务可视为完整的数据处理单元，通过流程编排将复杂业务逻辑分解为可配置的标准化流程。
2. **数据处理**：支持在任务中定义数据表和变量，实现处理过程中的状态与数据暂存。
3. **浏览器自动化**：内置的浏览器插件提供 **DOM** 元素提取、内容修改及事件触发能力，支持动态网页数据采集。
4. **系统集成**：提供 **HTTP** 插件实现 **API** 调用，并可配置大模型接口进行智能数据处理。

技术特性：

- **流程可视化**：支持通过图形化界面配置数据处理流程。
- **多源适配**：统一对接各类异构数据源。
- **智能增强**：集成大模型能力提升数据处理智能化水平。
- **可扩展架构**：支持通过插件机制扩展系统功能。

2. HTTP 插件

HTTP 插件，包括一些列插件方法，插件方法与 **Web API** 访问方法对应，可在流程中使用。**HTTP** 插件方法是全局可见的，即所有任务都可以使用。

2.1. HTTP 插件内容

一个 **HTTP** 插件可以理解为 **Web API** 方法通用参数的提取，如名称、访问地址(域名通用部分)、授权。意味着具有相同参数的 **Web API** 方法可以放到一个插件中。

插件详情
保存退出

基本信息

名称

描述

访问地址

Header 列表

KEY	Value	操作
		+

授权配置

授权类型

授权参数位置

参数名

API key/token

2.2. 插件方法

插件功能详情
蓝伏豚
保存退出

基本信息

名称

描述

路径

请求方法 内容格式

输入参数

参数名	参数描述	数据类型	传入位置	必填	默认值	开启	操作
incomplete	1表示获取信息不完整的笔记本信息	整数数字	BODY	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	-
price_day	表示在此天数内未更新价格的笔记本	整数数字	BODY	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	-
buy_url_day	表示在此天数内未更新购买地址的笔记本记录	整数数字	BODY	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	-

+ 新增参数

输出参数

参数名	参数描述	数据类型	开启	操作
code		整数数字	<input checked="" type="checkbox"/>	-
message		字符串	<input checked="" type="checkbox"/>	-

名称：必填，可任意定义为容易理解的内容；

描述：可选，对方法的注解；

路径：必填，填写 Web API 地址除去插件定义的通用部分，所剩余的部分；

请求方法：POST 和 GET 可选。

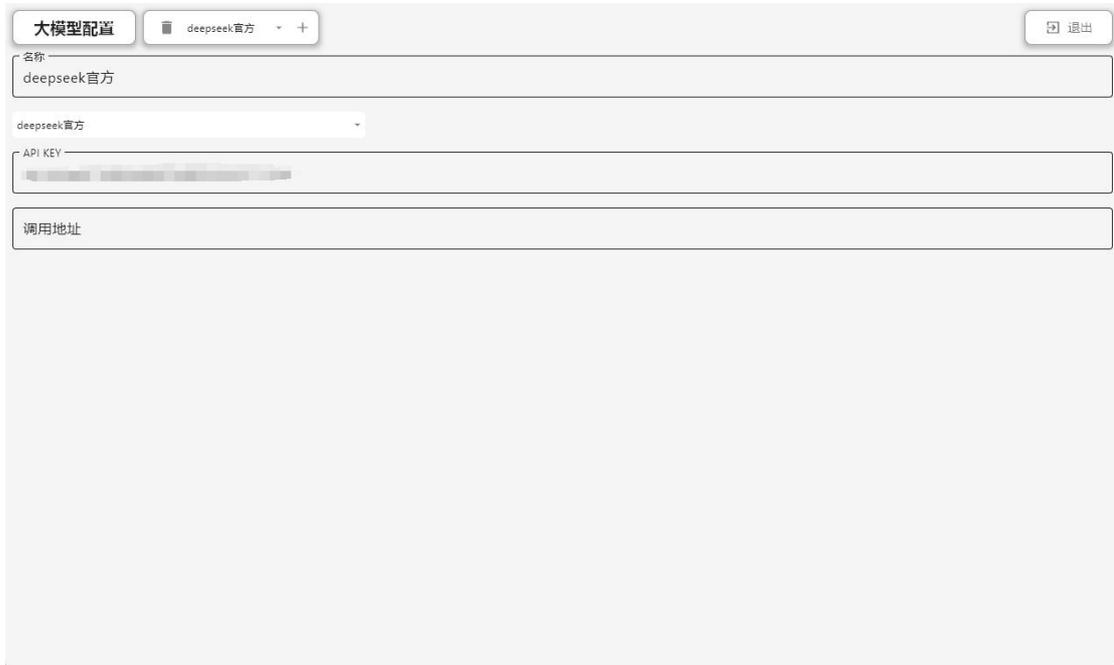
内容格式：支持 application/json、multipart/form-data、application/x-www-form-urlencoded；

输入参数：传入位置可以是 HEADER、BODY、QUERY。若开启时，在流程节点中可以看见该参数，并设置值，否则流程节点不可见，使用默认值传递。

输出参数：未开启时，流程节点中该参数不见。输出参数可以自动解析，我们只需要定义好输入参数，点【自动解析】按钮即可。

3. 大模型

配置大模型的访问接口，当前版本仅支持 deepseek 官方接口，只需要填写官方的 api key 即可，调用地址空着。



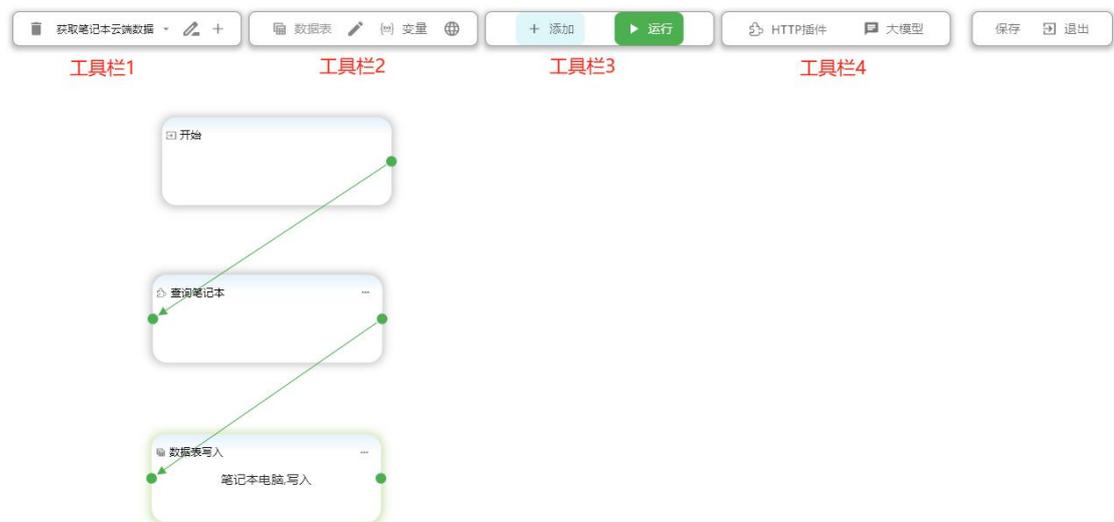
4. 任务流程编排

工具栏 1: 可以新增和删除流程，给流程改名。

工具栏 2: 可以打开关闭数据表窗口，打开变量窗口，打开关闭浏览器窗口。

工具栏 3: 添加节点到流程中，运行、暂停和继续流程操作。

工具栏 4: 打开 HTTP 插件窗口，打开大模型配置窗口。



5. 数据表定义

字段名	描述	数据类型	主键	操作
platform		字符串	<input type="checkbox"/>	<input type="checkbox"/>
product_id		字符串	<input type="checkbox"/>	<input type="checkbox"/>
title		字符串	<input type="checkbox"/>	<input type="checkbox"/>
source_url		字符串	<input type="checkbox"/>	<input type="checkbox"/>
buy_url		字符串	<input type="checkbox"/>	<input type="checkbox"/>
buy_url_updated_at		字符串	<input type="checkbox"/>	<input type="checkbox"/>
price		整数数字	<input type="checkbox"/>	<input type="checkbox"/>
price_updated_at		字符串	<input type="checkbox"/>	<input type="checkbox"/>
product_type_name		字符串	<input type="checkbox"/>	<input type="checkbox"/>
evaluation		字符串	<input type="checkbox"/>	<input type="checkbox"/>
product_value		小数数字	<input type="checkbox"/>	<input type="checkbox"/>
product_value_considerations		字符串	<input type="checkbox"/>	<input type="checkbox"/>
正确地配置表字段结构, 可以更好地管理数据				

工具栏 1: 新增、删除数据表, 修改数据表名称。

工具栏 2: 从已定义的 HTTP 插件导入数据表字段, 为当前数据表新增字段。

字段名: 最好与你需要读取或写入系统时的参数名一致。

主键: 在流程节点写入和读取数据时, 提供匹配依据。可以尽可能多地勾选关键字段, 流程节点只是在这些字段中选择更适合的使用。

dataC 将数据表分为本地数据和云端数据, 本地数据方便我们修改调整, 云端数据是下载到本地, 但值与云端相同的数据。这种结构方便我们核实对比。

一个任务中, 可以定义多个数据表。

需要清除数据表数据时, 点击表名称右侧的三横标记按钮。

获取笔记本云端数据 ✎ +

数据表 ✎ (xxx) 变量 🌐

数据表 ✕

笔记本电脑 ☰

本地数据	云端数据
	<p>【title】联想笔记本电脑小新Pro14超能本 高性能标压酷睿i5 14英寸轻薄本 1T硬盘大内存 2.8K 120Hz高刷屏</p> <p>【buy_url_updated_at】2024-12-14 16:03:44</p> <p>【price】4799</p> <p>【price_updated_at】2025-05-09 10:21:52</p> <p>【product_type_name】笔记本电脑</p> <p>【evaluation】适合商务办公和学生使用。其PPI达到242，显示效果细腻；32GB内存和1T硬盘容量充足，满足大多数需求；16:10的屏幕比例和100% DCI-P3色域适合设计工作；重量仅1.46kg，便于携带。但若主要用于游戏或专业图形处理，建议选择独立显卡的机型。</p> <p>【product_value】0.81</p> <p>【product_value_considerations】CPU</p>

6. 变量定义

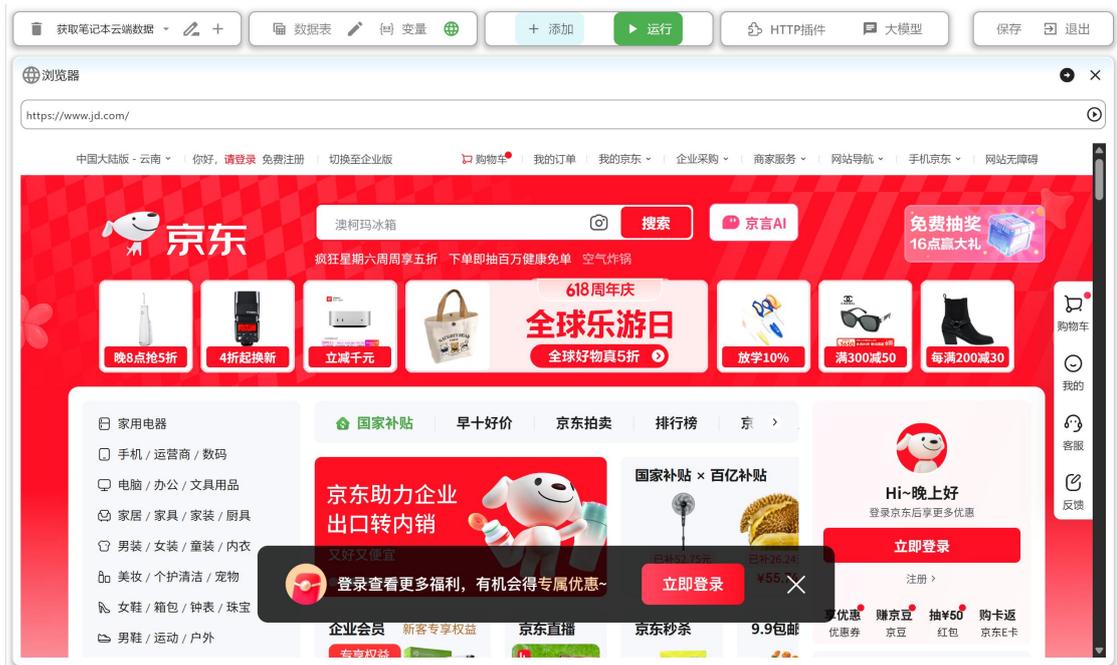


变量用于暂存流程数据，可在变量界面手动添加/修改，或在流程节点中动态写入和更新。

7. 浏览器界面

点击浏览器按钮，可以打开或关闭内置浏览器窗口。

可以使用浏览器流程节点，与内置浏览器交互。



8. 流程节点

8.1. 开始

流程新建的时候，开始节点自动建立，它是流程执行的第一个节点，必不可少。

8.2. HTTP 插件方法节点

该类节点，按照定义的参数发起 Web API 请求。

查询笔记本 ✎

输入

变量名	变量值
incomplete int 🗨	<input type="text"/>
price_day int 🗨	<input type="text"/>
buy_url_day int 🗨	<input type="text"/>

输出

- output 对象
 - code 整数数字
 - message 字符串
 - data 对象
 - current_page 整数数字
 - data 对象数组
 - platform 字符串
 - product_id 字符串
 - title 字符串
 - source_url 字符串
 - buy_url 字符串
 - buy_url_updated_at 字符串
 - price 整数数字
 - price_updated_at 字符串
 - product_type_name 字符串
 - evaluation 字符串

8.3. 数据表节点

读取：从数据表读取数据。

可以单条读取，也可以一次性读取整个数据表。

可以决定是否移动游标，默认情况，每读取一次，游标自动移动到下一条记录。

dataC 会根据数据表的定义，同时建立本地数据和云端数据，两者字段一一对应，便于比较修正数据。我们可以决定读取的是本地数据还是云端数据。

读取输出是根据绑定的数据表定义输出的。

数据表写入 ✎ ✕

操作

读取 写入 重置游标(指向记录第一条)

读取选项

单条读取 读取所有

不移动游标 移动游标

读写位置

本地数据 云端数据

输出参数

- ↑ output 对象数组
 - └─ platform 字符串
 - └─ product_id 字符串
 - └─ title 字符串
 - └─ source_url 字符串
 - └─ buy_url 字符串
 - └─ buy_url_updated_at 字符串
 - └─ price 整数数字
 - └─ price_updated_at 字符串
 - └─ product_type_name 字符串
 - └─ evaluation 字符串

写入：将数据写入数据表。

写入位置，指定将数据写入到本地数据还是云端数据位置。

写入数据，指定需要写入的数据。

匹配字段，使用所指定的字段，在数据表中查询，若有匹配记录，则数据覆盖该记录，否则新增。匹配时，同时匹配本地数据和云端数据字段名，只要有一端满足，则匹配该记录。可以选择的匹配字段，需要在数据表中勾选为关键字段。



重置游标: 将游标指向第一条记录，该方法没有其他选项。

8.4. 浏览器方法节点

通过浏览器节点方法，与内置浏览器交互。

1.1.1.1. 加载页面

当页面加载成功，`success` 的值为 `true`。



1.1.2.内容提取

DOM 元素路径，可以填写 XPath 路径，或 CSS 选择器。

内容提取 ✎ ×

设置

元素路径 字符串

内容类型: Inner Html 文本 Outer Html

1.1.3.等待元素

该节点等待指定的 DOM 元素出现就返回，路径可以使用 XPath 或 CSS 选择器。
可以指定超时时间，当超时发生时，流程停在该节点处，由用户决定是否继续执行。

等待元素 ✎ ×

设置

元素路径 字符串

超时(秒):

1.1.4.元素赋值

元素赋值 ✎ ×

设置

元素路径

元素值

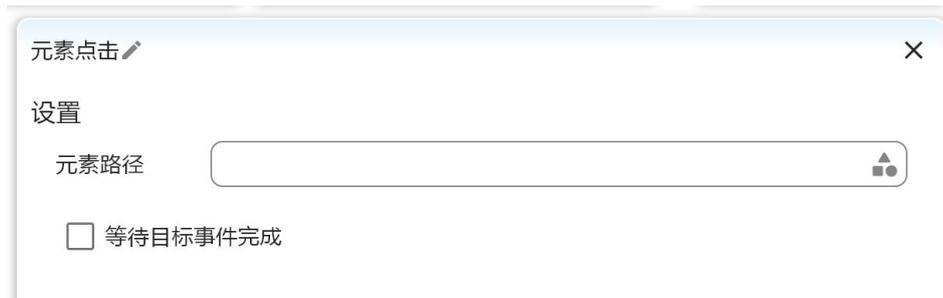
1.1.5.元素取值



1.1.6.元素点击

触发指定元素的 click 事件，比如点击按钮。

若 click 事件是一个网络请求相对耗时的事件，可以勾选“等待目标事件完成”，会等到 click 触发的请求完成才返回。若不勾选，则立刻返回，继续执行下一个节点。



8.5. 变量

8.5.1. 变量读取

若指定的变量在变量表中不存在，返回空值。



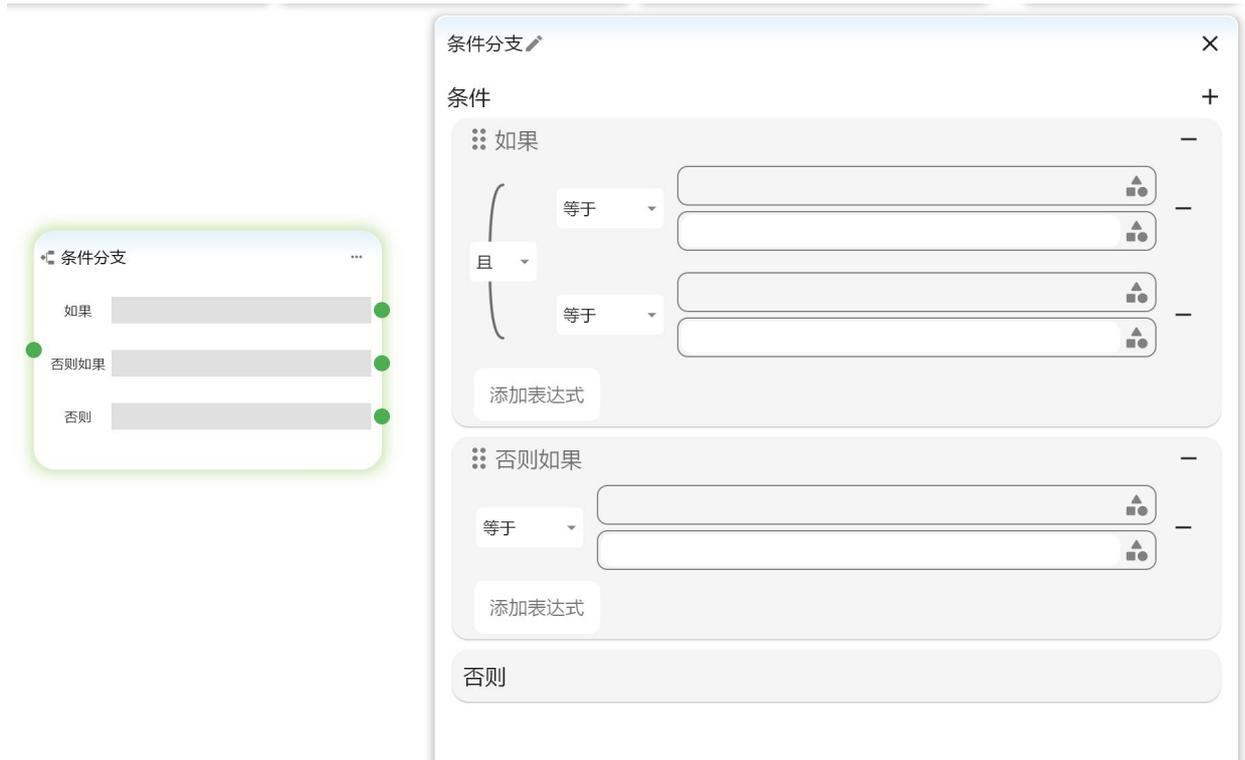
8.5.2. 变量写入

若参数名在变量表中不存在，则会自动建立。



8.6. IF 条件节点

可以灵活地定义多个条件，形成多个流程分支。每个条件可以组合多个判断。



8.7. 对象合并

当我们需要将不同节点产生的数据，一次性写入到数据表时，可以使用对象合并节点，合并前节点的输出。



8.8. 字符串拼接

可以使用拼接模板，将多个字符串拼接成我们需要的格式。在模板中使用{{参数名}}可以引入参数值。

字符串拼接

待组合参数

参数名	参数值	操作
<input type="text" value="v1"/>	<input type="text"/>	-
<input type="text" value="v2"/>	<input type="text"/>	-

[+ 新增](#)

拼接模板

{{v1}}的目标是{{v2}}

8.9. 数字运算

可以对数字进行加减乘除运算。

数字运算

运算符	操作数
<input type="text" value="加"/>	<input type="text"/>
	<input type="text"/>

8.10. 大模型

大模型节点，可以将提示词中的{{变量}}，替换成对应的值后，发送到大模型接口，并按照提示词说描述的格式返回。

我们可以填写完成输入参数和输出参数相关信息，点提示词后面的按钮自动生成提示词。

输入参数的参数描述，可以描述该变量值所包含的内容。输出参数的参数描述，必须准确地描述该参数的值的意义。这样才能自动生成准确的提示词。

大模型 ✎ ✕

大模型

deepseek官方 ▼ +

输入参数 +

参数名	参数描述	参数值	操作
<input type="text" value="title"/>	<input type="text" value="标题, 包含品牌、型号信息"/>	<input type="text"/>	▲ ▼ —
<input type="text" value="price"/>	<input type="text" value="包含价格信息"/>	<input type="text"/>	▲ ▼ —

提示词 🔄

商品标题, 包含品牌、型号信息:
{{title}}

包含价格信息:
{{price}}

根据上面的内容说明, 提取以下数据, 以json格式输出:
title:<标题>
price:<价格>
bland:<品牌>

输出参数 +

参数名	参数描述	参数类型	操作
<input type="text" value="title"/>	<input type="text" value="标题"/>	<input type="text" value="字符串"/>	—
<input type="text" value="price"/>	<input type="text" value="价格"/>	<input type="text" value="字符串"/>	—
<input type="text" value="bland"/>	<input type="text" value="品牌"/>	<input type="text" value="字符串"/>	—

8.11. 用户输入

该节点执行时, 会弹出对话框, 要求用户输入, 并按照输出数据类型指定的类型输出。

用户输入 ✎ ✕

输出数据类型

8.12. 正则表达式

我们可以使用正则表达式, 从提取源中提取数据, 并按照结果重组的格式输出。可以定义输出字段名称和类型。默认情况下, 输出结果为对象, 若需要输出为数组, 勾选“输出组装为数组”。

正则表达式提取 

正则表达式

提取源

结果重组 

使用\$1,\$2,\$3...表示匹配分组

输出字段名和类型

输出组装为数组