自助ETL-节点资源区介绍

节点资源区用于显示当前流程可拖拽使用的资源，最顶端的文本框支持输入资源名称关键字模糊匹配搜索结果。

数据源

Smartbi提供了四种数据源用于数据输入，分别是文本数据源、Kafka数据源、关系数据源、示例数据源、数据集，支持从这几个数据来源中导入数据。

名称	使用说明
文本数据源	文本数据源是指将HDFS读取的csv等数据文件导入到Smartbi中。
关系数据源	关系数据源是指从Smartbi关系数据源中读取的库表数据。
示例数据源	示例数据源是指从系统中读取内置的示例数据源。
数据集	数据集是指从Smartbi中读取数据集中的数据。
Kafka数据源	Kafka数据源是指从kafka读取数据。

目标源

Smartbi提供了2种方式用于数据的输出，分别是关系目标源（追加）、关系目标源（覆盖）、关系目标源（插入或更新）、导出数据到HDFS。

名称	使用说明
关系目标源	关系目标源通过追加、覆盖、插入或更新的方式将结果数据保存到Smartbi的关系数据源中。 ⚠ 覆盖、插入或更新的方式为Smartbi V97以上版本提供。 目前支持 Infobright、ClickHouse、Vertica、Oracle、MySQL、DB2、MSSQL、PostgreSQL、GuassDB 100、GuassDB 200、达梦（V9.5不支持达梦数据库，V9.7支持6、7.1、7.6版本的达梦数据库）、GBase（V9.5不支持GBase数据库，V9.7支持8A、8S V8.4、8S V8.8版本的GBase数据库）、Sybase、HANA。
导出数据到HDFS	导出数据到HDFS是指将结果数据保存到HDFS中。

数据预处理

自助ETL拥有强大的数据处理功能，对各种结构化数据，可进行排序、去重、映射、行列合并等处理，满足客户日常数据处理的需要。

使用数据预处理可以：

1、提高数据的质量。

2、让数据更好地适应特定的挖掘技术或工具。

名称	使用说明
采样	按照某种规则从数据集中挑选样本数据。
拆分	将原始样本集按照训练集和测试集的方式拆分为两个子集。
过滤	根据用户需求，通过写SQL语句(片段)的方式，对数据集中指定字段进行条件筛选过滤。
列选择	列选择节点用于从输入数据集中选取指定的数据字段。
空值处理	空值处理节点是将空值替换为均值、最大频数或者用户自定义的值等，实现空值的填充或者过滤。
合并列/合并行	将两张表的数据按列或按行合并，组成新表。
元数据编辑	元数据编辑支持对数据集中的字段进行重新命名或者修改数据类型。
JOIN	JOIN是基于连接字段和给定的连接方式，进行两个数据集字段的组合后得到新的数据表。

行选择	行选择是根据不同的筛选或者删除条件，选择不同数量的行。
去除重复值	去除重复值是用于删除数据集中的重复行（假如有两行相同，保留其中一行）。
排序	排序节点可实现对单个字段或多个字段组合的升序或降序排序。
增加序列号	增加序列号节点是在数据表第一列追加ID列。
聚合	聚合可根据用户的需求对数据进行各种聚合运算。
分列	将字符串字段的内容进行分割。
派生列	派生列节点是用于在数据集中生成可行的新特征字段。
行转列/列转行	将数据表中的行转换成列或将列转换成行。

自定义模块

自定义模块是通过手动输入SQL或Python语言对数据进行数据处理、分析或查询。

名称	使用说明
SQL脚本	SQL脚本支持手动输入SQL语言完成对数据进行处理和查询的任务。
PYTHON脚本	支持用Python语言编程实现数据处理、数据分析等功能。 注：该功能为Smartbi V97以上版本提供。