数据挖掘-数据的输入和输出

在数据挖掘的流程中，数据的输入和输出也是必不可少的。

因为需要导入数据才可以进行后续的数据预处理、分析、建模等；以及将最后的结果数据，导出保存在指定的目标库。

所以Smartbi分别提供数据源和目标源节点，满足数据的输入和输出。

数据源

Smartbi提供了多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。

文本数据源

概述

文本数据源是指将HDFS读取的csv等数据文件导入到Smartbi中。

输入/输出

输入	没有输入端口。
输出	只有一个输出端口，用于输出数据到下一节点资源。

参数配置

设置文本数据源的参数：

设置说明如下：

参数	说明
地址	文本数据在HDFS的路径，其中： • <host>表示HDFS所在服务器IP地址； • <port>表示HDFS端口号； • <path>表示文本数据在HDFS服务中的路径； 示例：hdfs://10.10.202.26:9000/data/mllib/UnitTest.csv
数据格式	选择文本的数据格式：csv、json、parquet、apache.orc。
文件编码	选择当前数据文件的编码格式：GBK或UTF-8。
读取行数	选择用于当前工作流的数据量：测试1000条、全部。
文本分隔符	选择当前数据文件中的分隔符：逗号、分号、空格、tab、竖线。
自动推断数据类型	若需要自动判断数据源中字段的数据类型，则选true，否则选false。
自动生成表头	表示上传数据时是否生成表头：若上传数据时没有表头，则选ture，系统自动生成表头；否则选false。

Kafka数据源

概述

Kafka数据源是指从Kafka读取数据。

输入/输出

输入	没有输入端口。
输出	只有一个输出端口，用于输出数据到下一节点资源。

参数配置

设置说明如下：

参数	说明
Kafka服务地址	连接Kafka的地址。
Topic	订阅的主题，一个topic可以看做为Kafka中的一类消息。
偏移量	每条消息在文件中保存的位置被称为偏移量（offset），从指定的起点开始消费Kafka数据。注：必须选择或者输入数字 • 从头开始：从头开始消费Kafka数据。 • 继续上次：从上次结束作为起点开始消费Kafka数据。 • 自定义数字：指定某个位置作为起点开始消费Kafka数据。
消息格式	支持csv跟json格式，如果是csv格式，需要设置分隔符跟字段映射。

关系数据源

概述

关系数据源是指从Smartbi关系数据源中读取的库表数据。

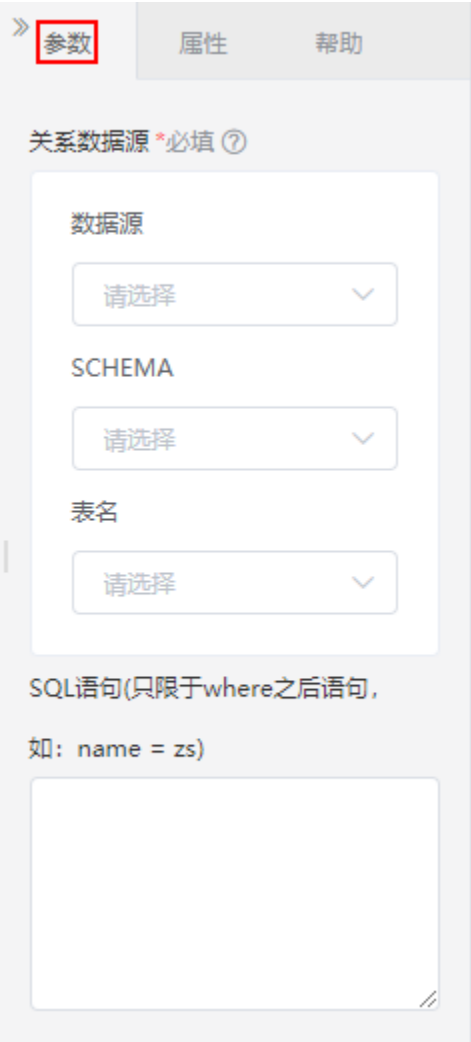
> **支持数据库**
>
> 目前支持Infobright、ClickHouse、Vertica、Oracle、MySQL、DB2、MSSQL、Presto+hive、Guass100、PostgreSQL、Greenplum（V9.5目前不支持Greenplum数据库，V9.7支持Greenplum数据库）、星环（V9.5目前不支持星环数据库，V9.7支持星环数据库）、达梦（V9.5目前不支持达梦数据库，V9.7支持6、7.1、7.6版本的达梦数据库）、GBase（V9.5目前不支持GBase数据库，V9.7支持8A、8S V8.4、8S V8.8版本的GBase数据库）、Aliyun AnalyticDB（2.7.8版本）、ODPS。

输入/输出

输入	没有输入端口。
输出	只有一个输出端口，用于输出数据到下一节点资源。

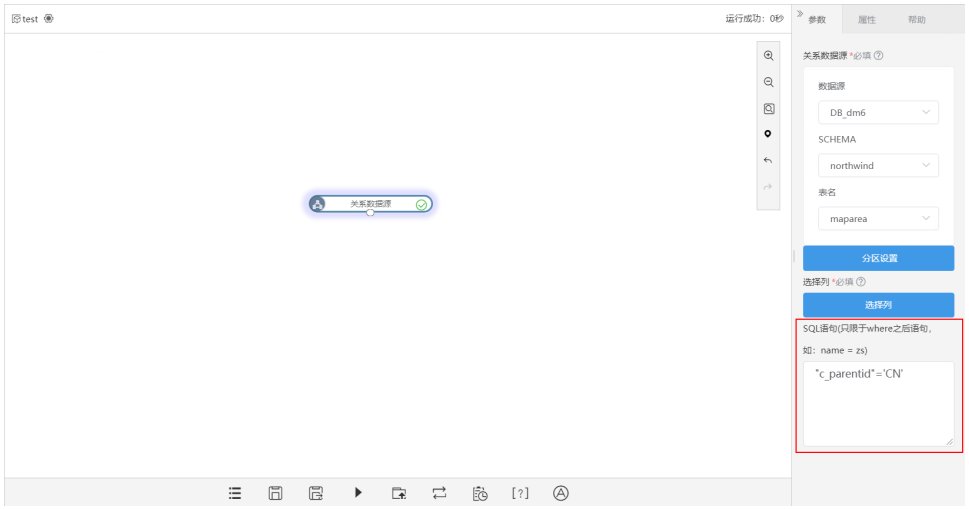
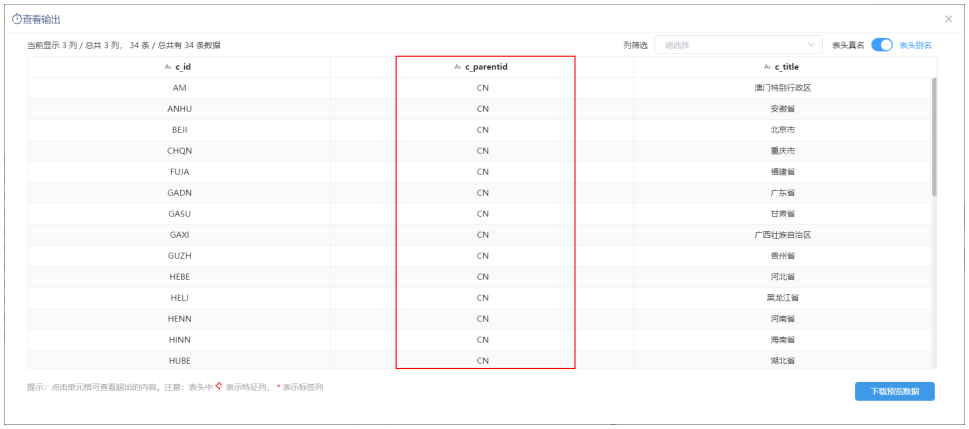
参数配置

设置关系数据源的参数：



设置说明如下：

参数	说明

数据源	选择数据源，这些数据源是Smartbi中连接配置好的关系数据源。
SHEMA	选择SHEMA。
表名	选择表。
SQL语句	通过SQL语句设置where条件，过滤出表中的数据用于工作流。 注意：在使用达梦数据库时，如果字段名含有字母、数字或符号，则需要对字段名加双引号；如果判断条件的取值是字符串，则需要使用单引号。例如：  输出结果： 

示例数据源

概述

示例数据源是指从系统中读取内置的示例数据源。

输入输出

输入	没有输入端口。
输出	只有一个输出端口，用于输出数据到下一节点资源。

参数配置

设置示例数据源的参数：

设置说明如下：

参数	说明
数据源选择	选择平台内置的示例数据源

数据集

> 在V9.7版本中，数据集节点的参数设置界面新增了新建、编辑数据集的入口。

概述

数据集是指从Smartbi中读取数据集中的数据，包含：可视化数据集、SQL数据集、原生SQL数据集、Java数据集、存储过程数据集、多维数据集、自助数据集。

输入输出

输入	没有输入端口。
输出	只有一个输出端口，用于输出数据到下一节点资源。

参数配置

设置数据集的参数：

设置说明如下：

参数	说明
请选择数据集	用于单击按钮后，在“数据集选择”窗口中选择Smartbi中已定义的数据集。
新建数据集	用于新建指定类型的数据集，选择数据集后，跳转到指定数据集的新建界面；可选的数据集类型有：自助数据集、原生SQL数据集、可视化数据集、存储过程数据集、Java数据集、多维数据集。
编辑已选数据集	用于编辑选择的数据集，单击按钮后，会跳转到指定数据集的编辑界面。
数据更新设置	用于设置数据集是否需要重新抽取：“更新抽取数据”表示需要重新抽取；“使用已抽取数据”表示不需要重新抽取。

Excel文件数据源

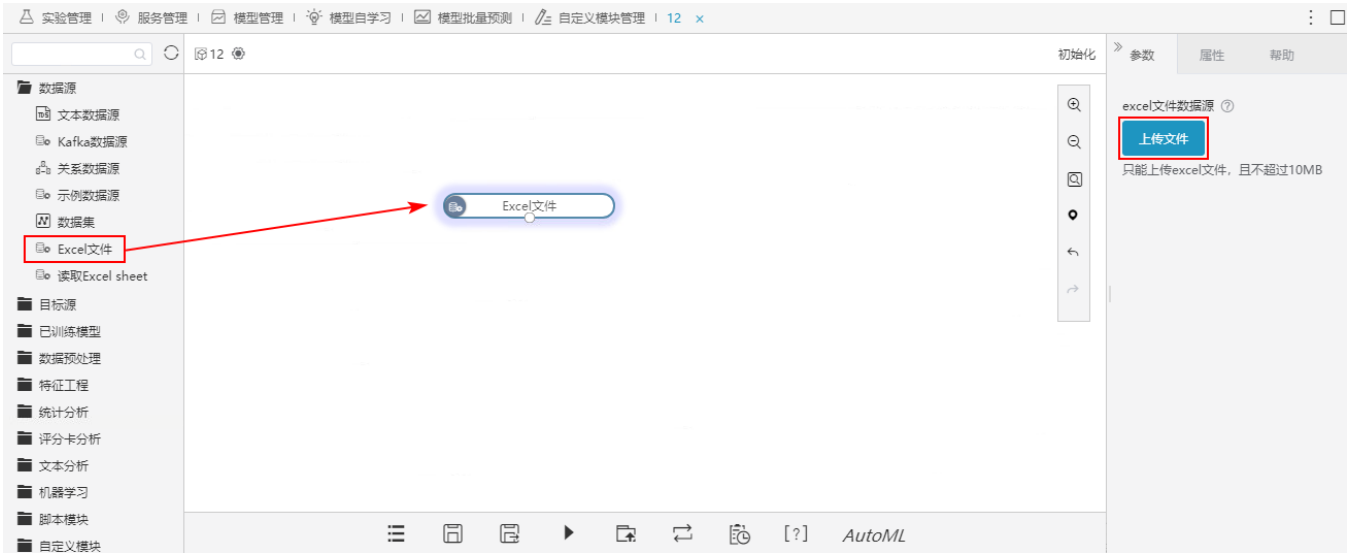
概述

Excel文件数据源是指将Excel文件中的数据导入到Smartbi中。

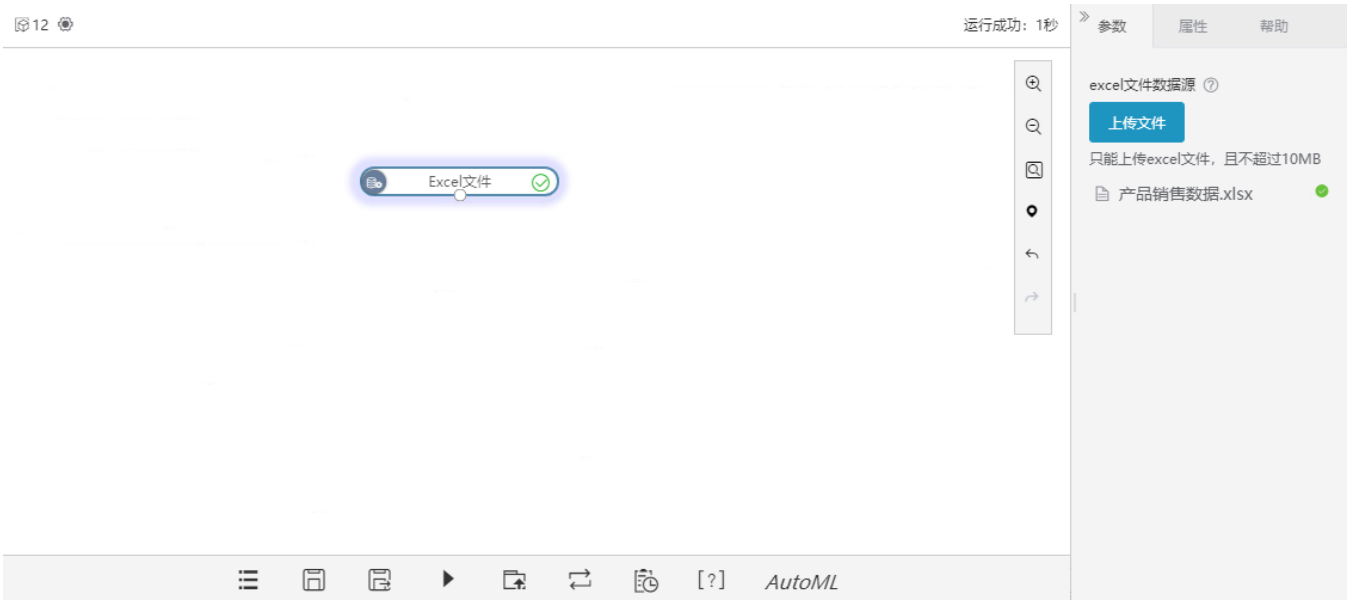
- 上传Excel文件：用于上传excel文件；
- 读取Excel sheet：用于读取指定sheet页的数据，只能接在上传Excel文件节点后面。

操作步骤

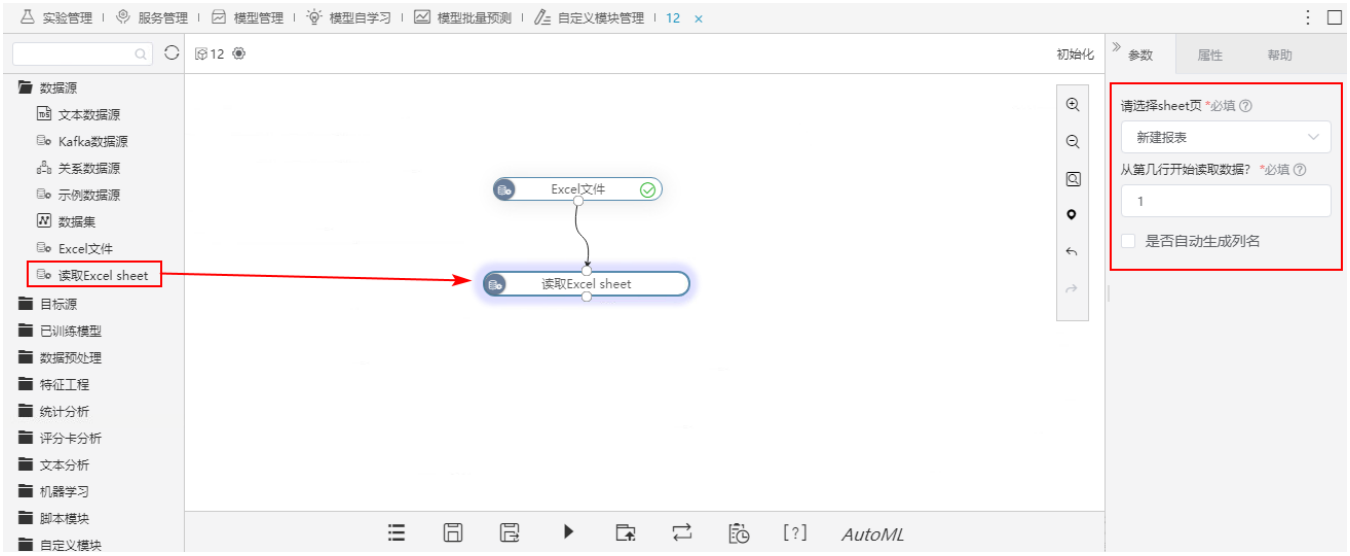
1、将Excel文件节点拖入画布区，点击 **上传文件** 按钮，上传Excel文件。



2、上传完成后，右键运行此节点。



3、将读取Excel sheet节点拖入画布区，配置参数，完成后运行此节点。



4、查看输出的数据如下：

设置说明如下：

节点	参数	说明
Excel文件	上传文件	上传Excel文件到服务器引擎，文件大小不超过10M。
读取Excel sheet	请选择sheet页	选择需要的sheet页。
	从第几行开始读取数据？	设置从第几行开始读取数据。
	是否自动生成列名	选择是否自动生成列名。

目标源

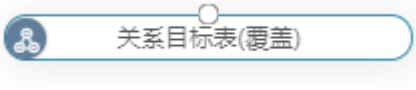
Smartbi提供了4种方式用于数据的输出，分别是关系目标表（追加）、关系目标表（覆盖）、关系目标表（插入或更新）、导出数据到HDFS，支持将数据导出到目标库中。

关系目标表

概述

关系目标表通过追加、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。

类型	说明

	在原数据的基础上增加新的数据。
	用新的数据对原数据进行覆盖。
	对数据库原有的数据进行更新，对数据库不存在的数据进行插入。

关系目标表（覆盖、插入或更新）适用于V97及以上版本，V95版本不支持。

支持数据库

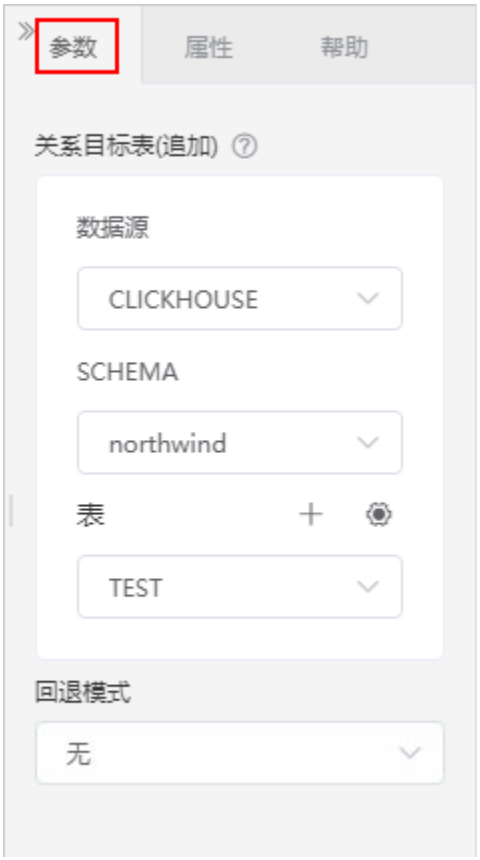
目前支持Infobright、ClickHouse、Vertica、Oracle、MySQL、DB2、MSSQL、PostgreSQL、Guass100、Greenplum（V9.5目前不支持Greenplum数据库）、星环（V9.5目前不支持星环数据库）、达梦（V9.5目前不支持达梦数据库，V9.7支持6、7.1、7.6版本的达梦数据库）、GBase（V9.5目前不支持GBase数据库，V9.7支持8A、8S V8.4、8S V8.8版本的GBase数据库）。

输入输出

输入	只有一个输入端口，用于将接收到的结果数据存储到指定库中。
输出	没有输出端口。

参数配置

关系目标源（追加）的参数：

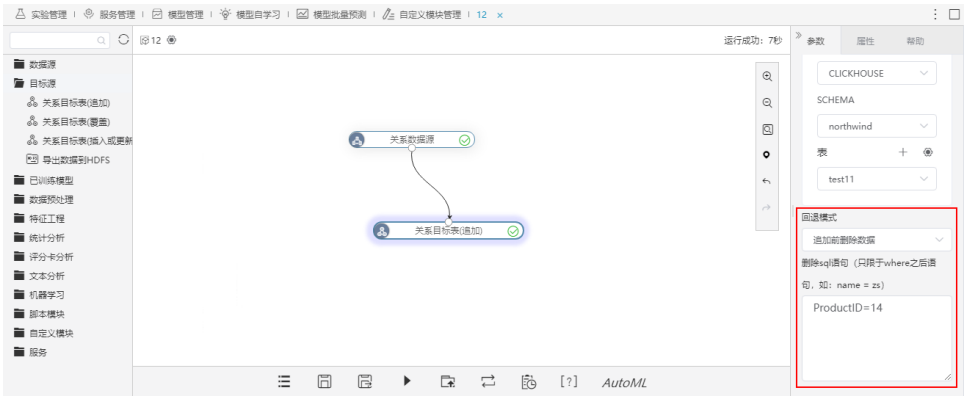


关系目标源（覆盖）的参数：

关系目标源（插入或更新）的参数：

参数说明如下：

参数	说明
数据源	选择数据源，这些数据源是在Smartbi中连接的关系数据源。
SCHEMA	在选择的数据源中选择SHEMA。

表	选择表。选择数据源和SCHEMA之后，可以选择 新建一张表，也可以在下拉框中选择已有的表，也可以修改表的映射关系 。 注意： 1、在GBase数据库中，不支持大写表名建表；8A版本支持大写列名，8S V8.8 和8S V8.4不支持大写列名和数字开头列名。 使用时如果源节点字段有大写字母，需要重新配置字段映射关系再执行节点。 2、对于关系目标源（插入或更新）节点，表中必须包含不重复的主键，且至少含有一个非主键列。 如果表中有相同主键的数据行， 可以先去除重复数据再执行节点。
回退模式	• 无（默认）； • 追加前删除数据：先删除表中部分或全部的数据，再将新数据追加到表中。 在删除SQL语句框中，填写where之后的删除语句（条件SQL使用表头真名）：  注意： 1、目前只有ClickHouse数据源（19.4.2.7版本及以上）支持此功能。 2、如果是新建的表，则表中不能有值为NULL的数据。

导出数据到HDFS

概述

导出数据到HDFS是指将结果数据保存到HDFS中。

输入输出

输入	只有一个输入端口，用于将接收到的结果数据存储到HDFS中。
输出	没有输出端口。

参数配置

设置导出数据到HDFS的参数：

设置说明如下：

参数	说明
IP和端口	目标HDFS的路径的IP和端口：<ip>:<port>； 示例：10.10.202.26:9000。
文件名	存储到HDFS的数据文件名。
HDFS用户名	HDFS用户名。
HDFS web端口	HDFS web端口，默认是50070。