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提供了几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、数据查询和Excel文件。

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

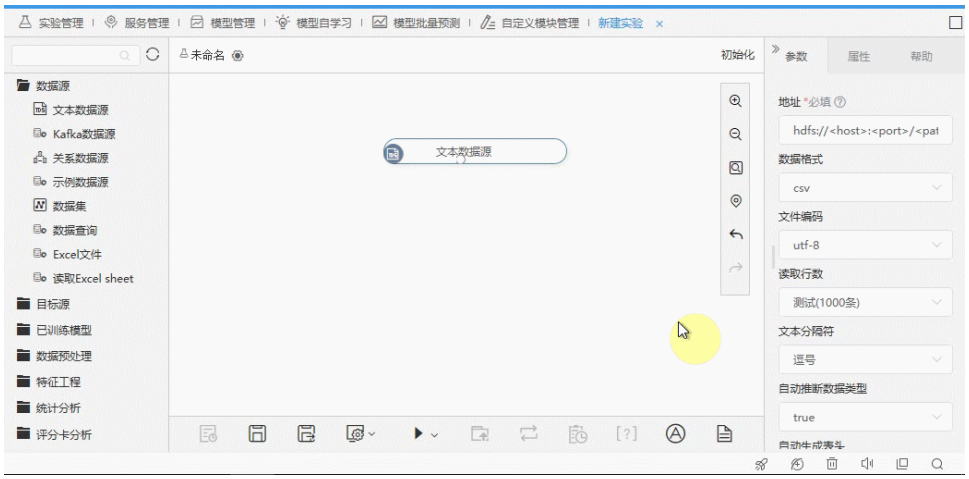
输入	没有输入端口。
输出	只有一个输出端口，用于输出数据到下一节点资源。

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，除了系统默认支持的逗号、分号、空格、tab、竖线，还支持用户自定义分隔符。 用户自定义分隔符的方法是：直接在文本框中输入分隔符，弹出该分隔符的下拉选项供选择即可。如下所示： 
自动推断数据类型	自动判断数据源中字段的数据类型，是则选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。

FTP数据源

概述

FTP数据源是指通过FTP方式读取数据。

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：

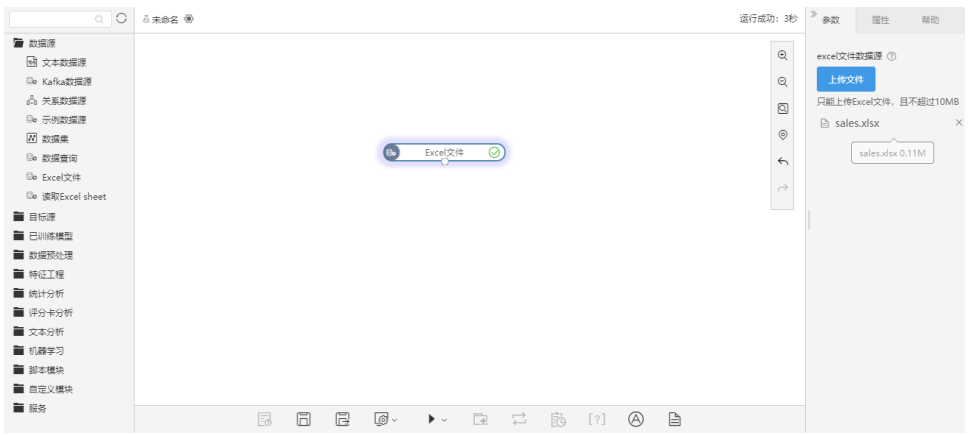
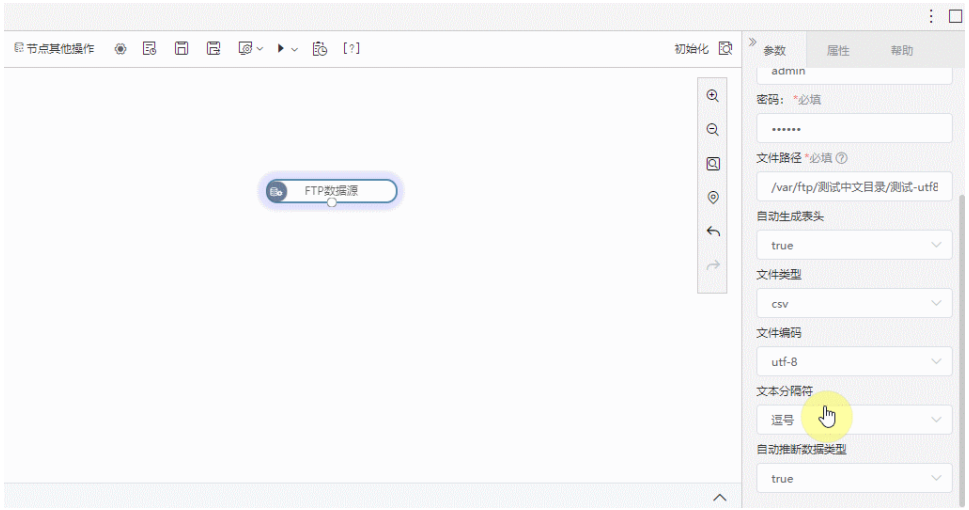
参数	说明
FTP服务器ip或主机名	连接FTP服务器的地址
FTP服务器端口	连接FTP服务器的端口
用户名	服务器用户名
密码	服务器密码
文件路径	填写读取文件的路径
自动生成表头	表示上传数据时是否生成表头：若上传数据时没有表头，则选ture，系统自动生成表头；否则选false。
文件类型	选择读取的文件类型，支持Excel文件、CSV文件

- 选择文件类型为“Excel”，还需设置以下设置项：

参数	说明

读取Sheet页名称	选择读取Sheet页的名称，只能输入一个Sheet页。 ⚠ 如果目录下的文件很多，则文件的Sheet页必须于名称一致，否则读取不了。
从第几行开始读取数据	设置从第几行开始读取数据，可输入大于0的整数。

- 选择文件类型为“CSV”，还需设置以下设置项：

参数	说明
文件编码	选择当前数据文件的编码格式：gbk或utf-8。
文本分隔符	 选择当前数据文件中的分隔符，除了系统默认支持的逗号、分号、空格、tab、竖线，还支持用户自定义分隔符。 用户自定义分隔符的方法是：直接在文本框中输入分隔符，弹出该分隔符的下拉选项供选择即可。如下所示： 
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。

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、Hadoop_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（用户密码方式连接 V9.5目前不支持星环数据库，V9.7及以上版本支持星环数据库）、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）、Sybase、HANA、Aliyun AnalyticDB（2.7.8版本）、ODPS、华为Fusioninsight数据库、Kingbase、Kingbase_V8、Kingbase AnalyticsDB 、GaussDB 200、Teradata、Teradata_V12、神通、Obase、MonetDB、Informix、Kylin（用户密码方式连接）、Impala、starRocks(社区版2.2.2)、SelectDB(飞轮科技)、Rapids(博睿)、Spark SQL（用户密码方式连接）。

注意：

- V10.5版本开始支持：Kingbase、Kingbase_V8、Kingbase AnalyticsDB 、GaussDB 200、Teradata、Teradata_V12、神通、Obase、Informix、Kylin（用户密码方式连接）、Impala数据源、starRocks(社区版2.2.2)、Rapids(博睿)、SelectDB(飞轮科技)、Spark SQL（用户密码方式连接）。
- kingbaseV7数据源不支持大数据量运行。
- 关系数据源 KingbaseAnalytics、ShenTong集群暂不支持小批量运行功能。

输入/输出

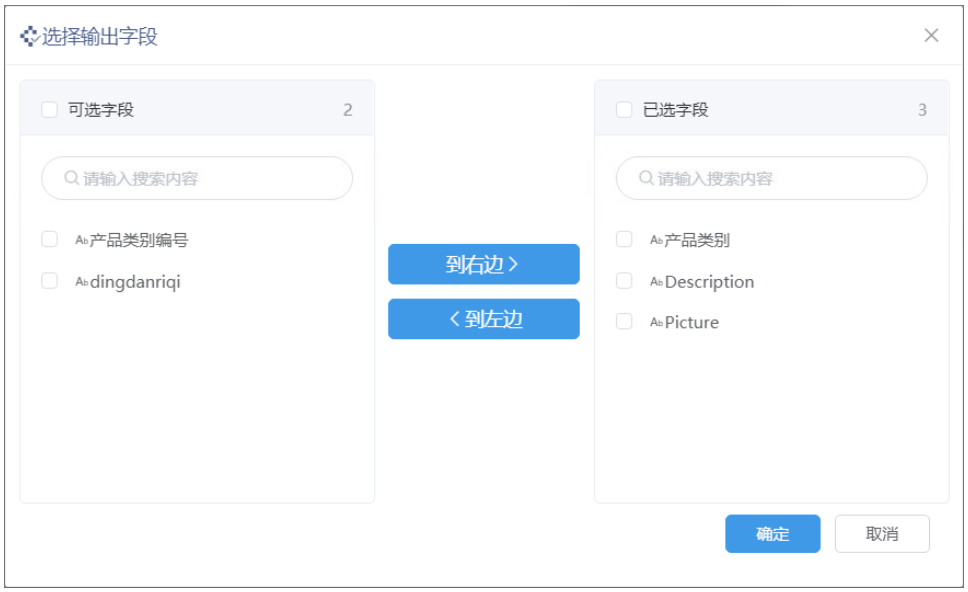
输入	没有输入端口。
输出	只有一个输出端口，用于输出数据到下一节点资源。

参数配置

设置关系数据源的参数：

设置说明如下：

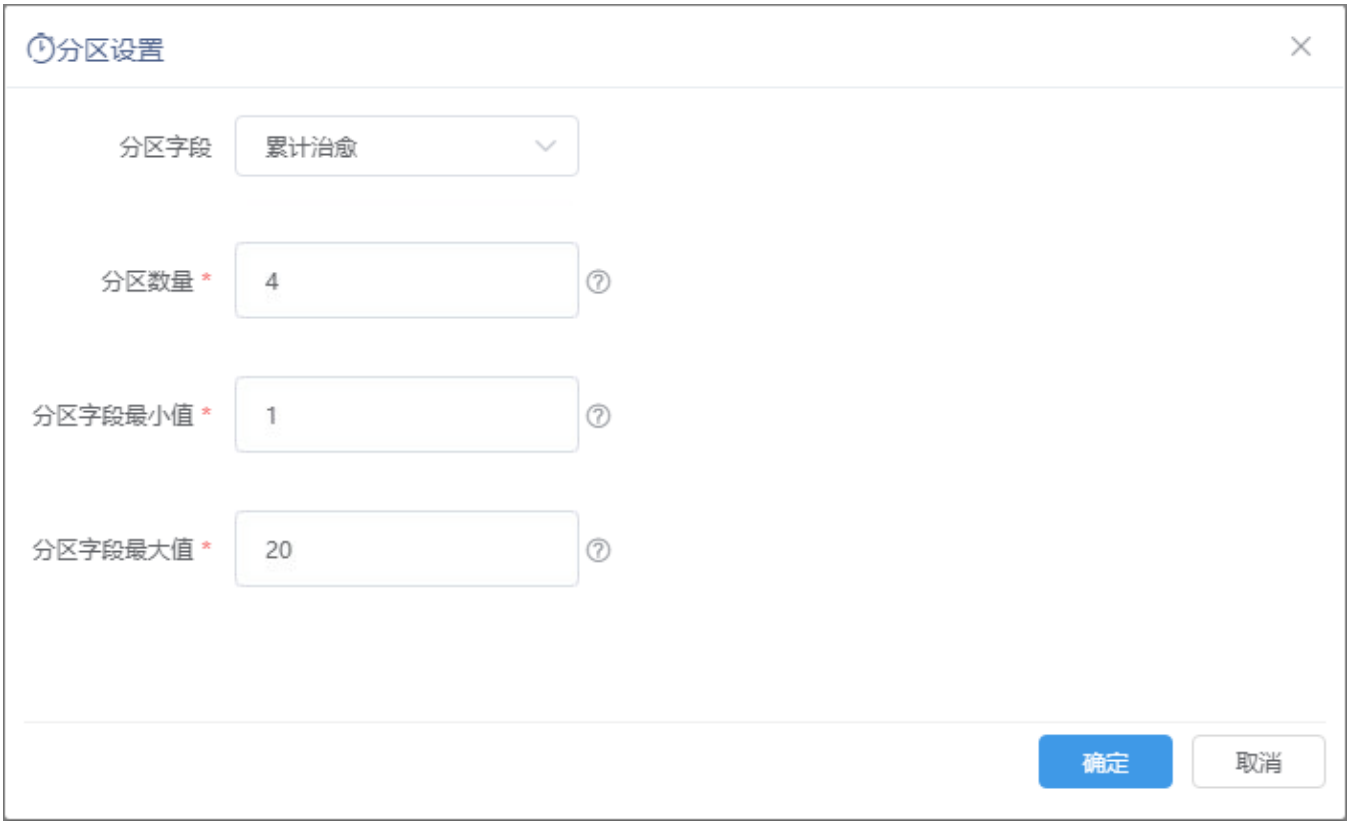
参数	说明
数据源	选择数据源，这些数据源是Smartbi中连接配置好的关系数据源，必填。
SCHEMA	选择SCHEMA，必填。

表名	选择表，必填。
分区设置	把表数据相对平均分成多个分区，抽取程序会尽可能一个分区分配一个线程进行并行抽取，能够极大的提高大数据量情况下的数据抽取性能。详情请参考 分区设置 。
选择列	用于筛选节点输出的列，适用于大数据量情况下，提升数据导出速度。 
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。

分区设置

把表数据相对平均分成多个分区，抽取程序会尽可能一个分区分配一个线程进行并行抽取，能够极大的提高大数据量情况下的数据抽取性能。

如图设置分区字段“累计治愈”的分区数量为4，最小值为1，最大值为20，则系统会以(-∞,5)、[5,10)、[10,15)、[15,+∞)这4个区间来并行读取数据，提升数据抽取性能。



- 分区字段（必填）：分区字段为数值型（不支持浮点型）、日期类型。
- 分区数量（必填）：设置抽取分区的数量，正整数。
- 分区字段最小值/分区字段最大值（必填）：在设置的最小值和最大值的区间中抽取数据。

分区字段的选取：

1. 选择的字段尽可能把数据按照不同区间，相对平均分成多个分区。
2. 在分区表中，可以选择创建分区表时选择的字段作为分区字段。如果不是分区表，建议选取一些比较有区分度的字段。例如在一张用户表中，“年龄”比“性别”字段更具有区分度，因此可以选择“年龄”作为分区字段。

使用场景：在一家互联网类企业中，用户使用产品的日志表按天或按季度做成的分区表，可以通过分区抽取数据，提升抽取性能。

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

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。

数据查询

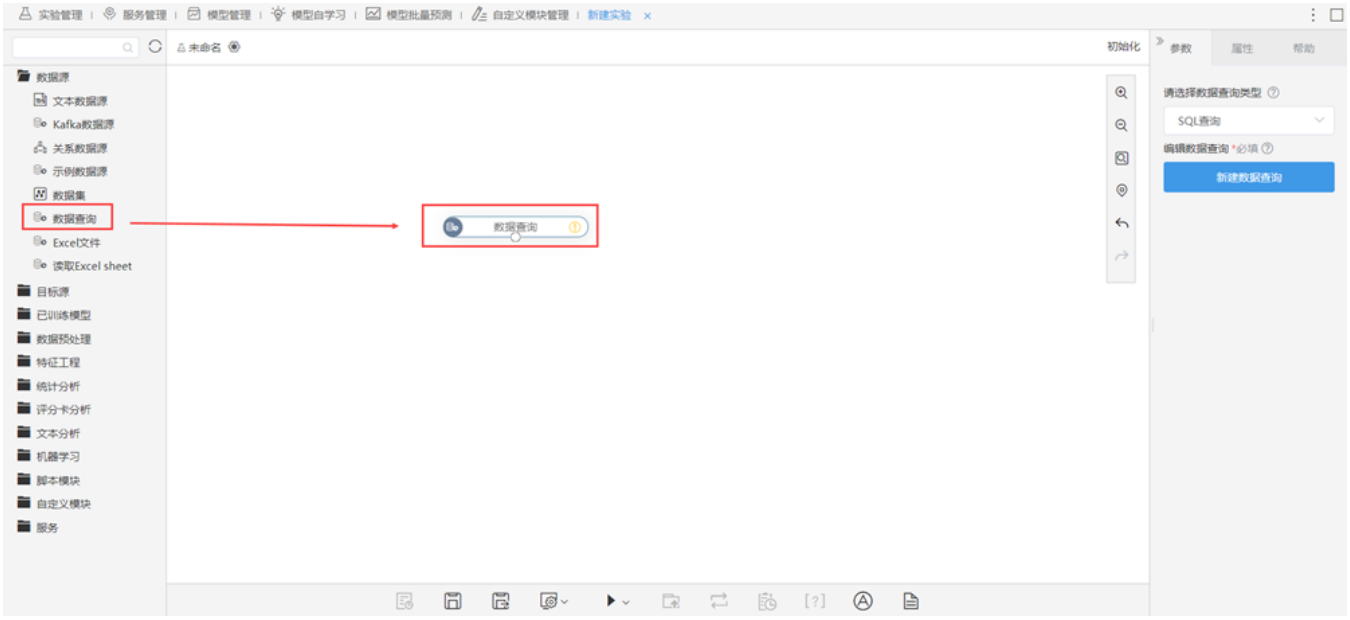
概述

数据查询是指将选择的数据查询转换成二维宽表。

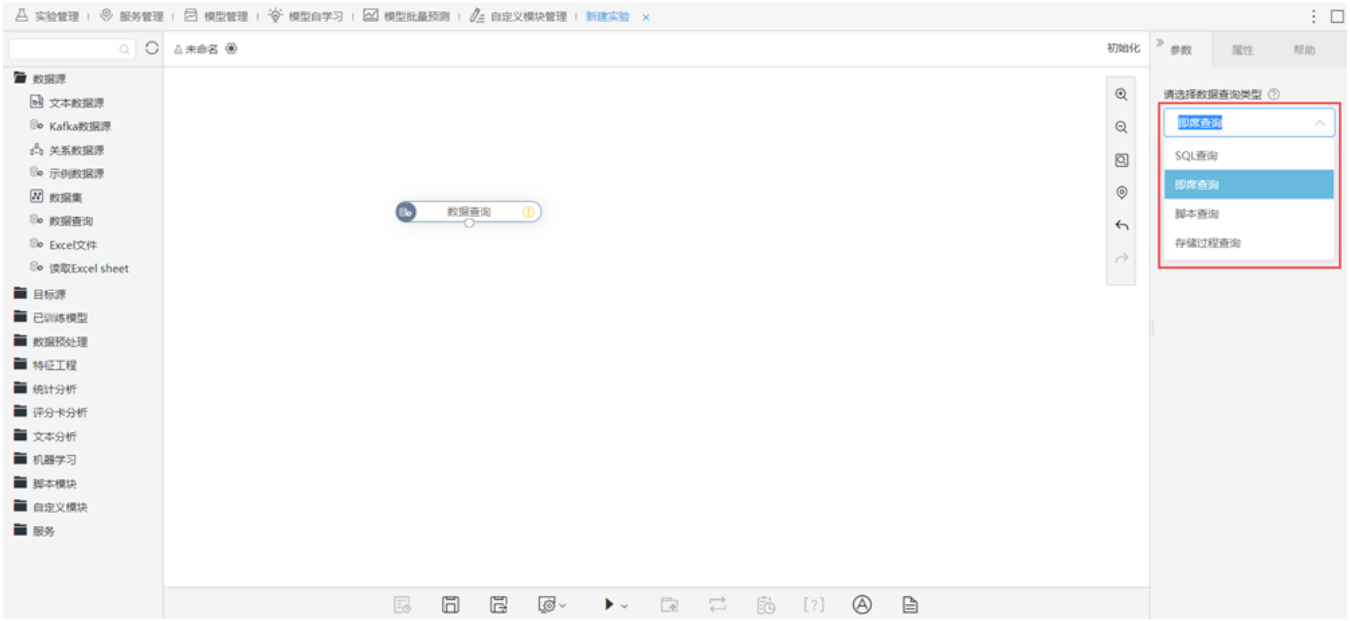
> ⚠ 数据查询依赖于高速缓存库，高速缓存库配置的url建议使用ip的方式连接，不推荐使用域名的方式连接。

操作步骤

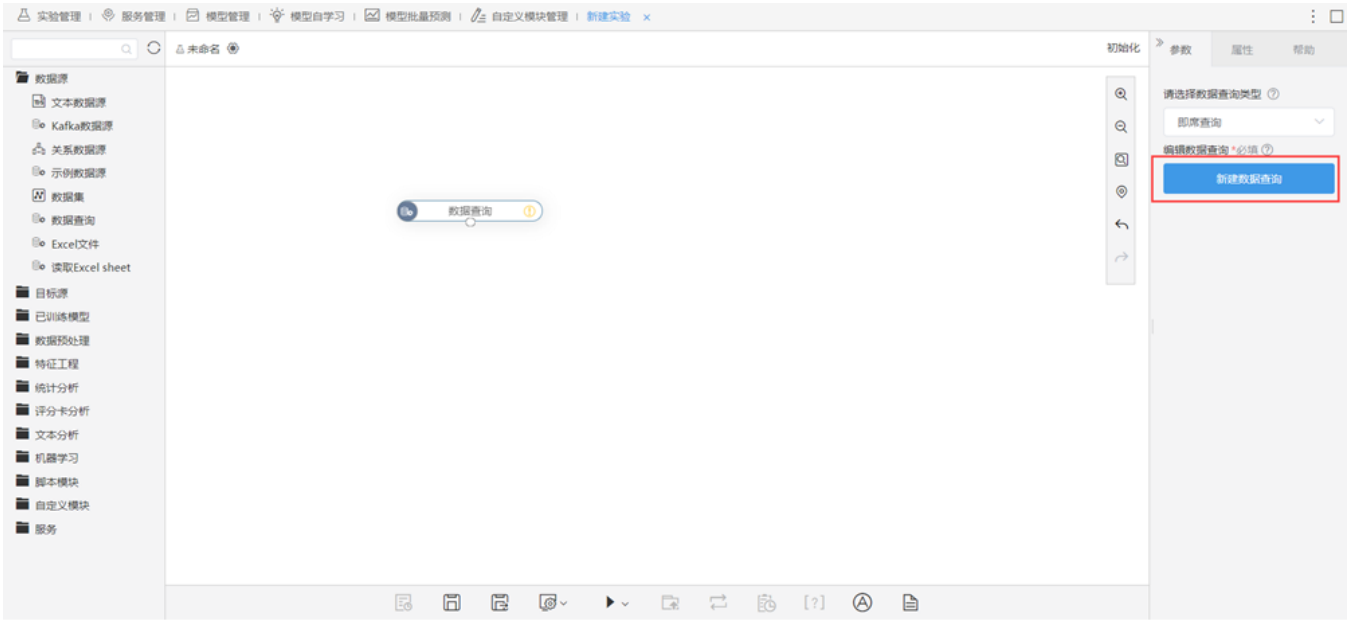
1、将数据查询拖入画布区。



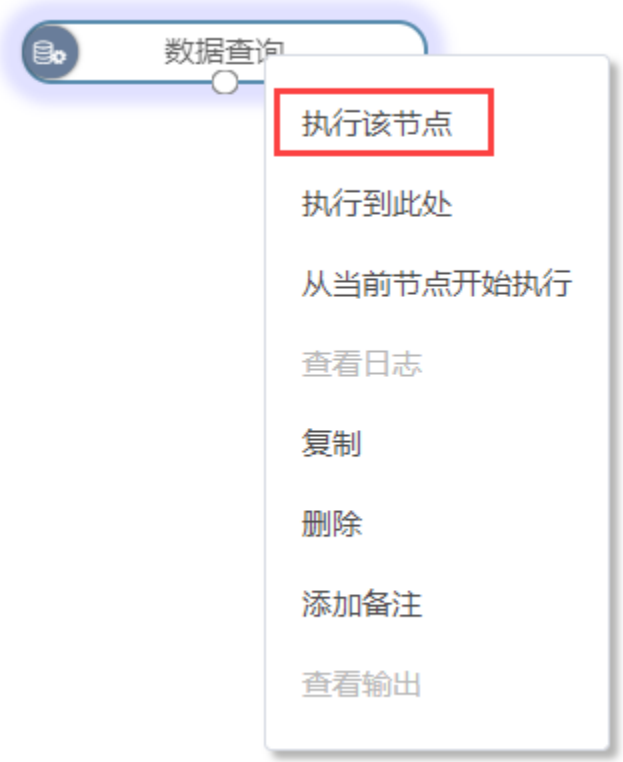
2、选择所需的数据查询类型。



3、点击新建数据查询，在弹出页面进行编辑。



4、新建完成后，右键执行该节点。



5、执行成功后，右键查看输出可浏览相关数据。

设置项说明

数据查询各设置项说明如下：

设置项	说明

请选择数据查询类型	可选择类型为SQL查询、即席查询、脚本查询、存储过程查询的数据查询类型。 SQL查询相关操作可参考文档SQL查询； 即席查询相关操作可参考文档即席查询； 脚本查询相关操作可参考文档脚本查询； 存储过程查询相关操作可参考文档存储过程查询；
新建数据查询	根据所选类型，新建一个新的数据查询。
编辑数据查询	对已创建的数据查询进行编辑。

Excel文件数据源

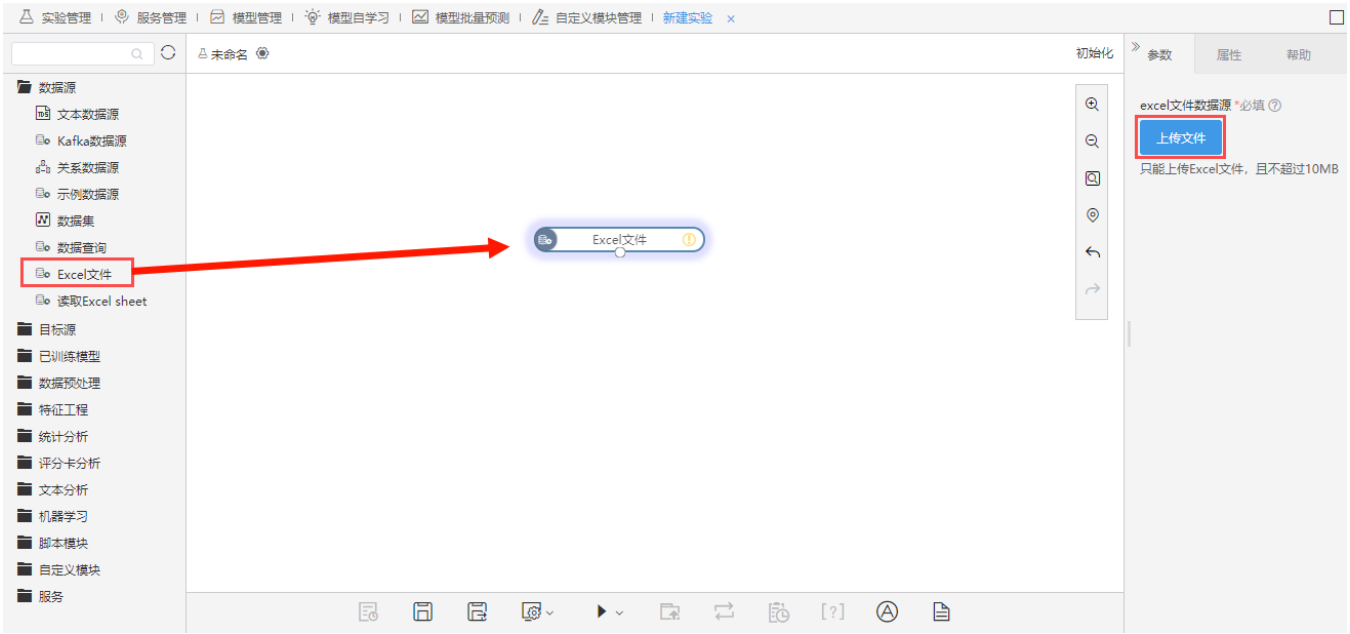
概述

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。

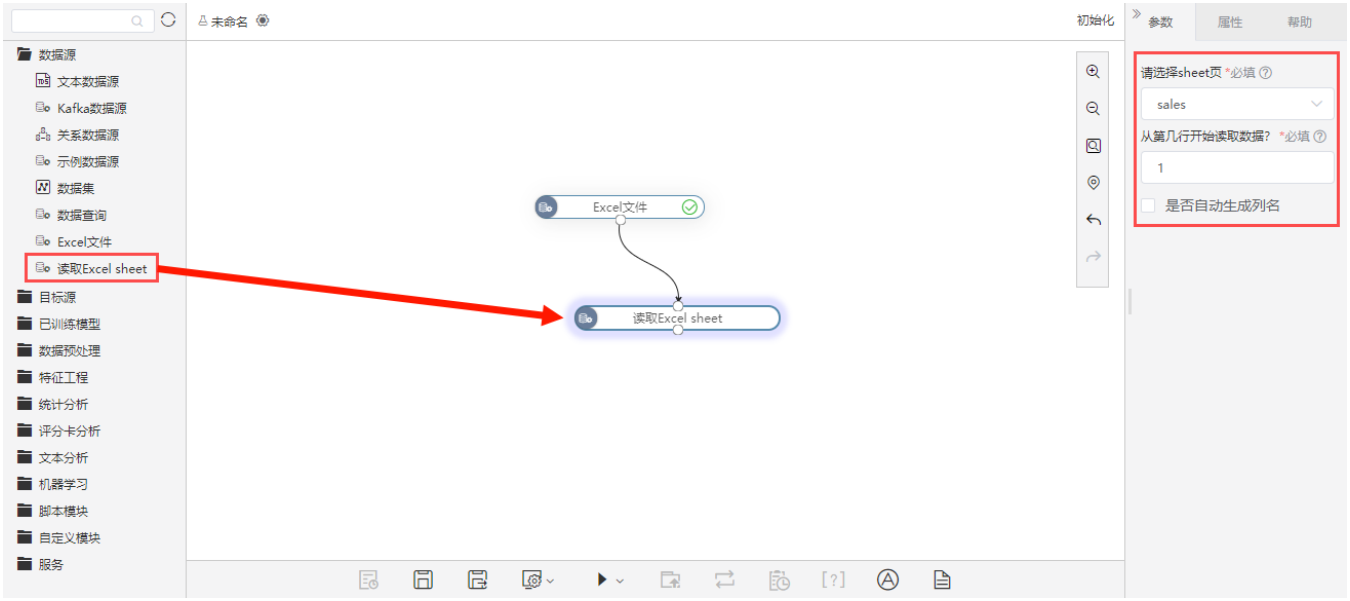
操作步骤

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。

不支持读取Excel文件中的图像、图表、公式和宏。

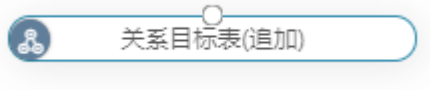
目标源

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。 ⚠ 暂不支持HANA。
	用新的数据对原数据进行覆盖。 ⚠ 暂不支持HANA。
	根据数据库表主键进行数据的插入或更新，若主键相同，则更新数据，否则插入数据。 ⚠ 暂不支持达梦、GBase数据库、HANA。

ⓘ 目前支持Infobright、ClickHouse、Vertica、Oracle、MySQL、DB2、MSSQL、PostgreSQL、GuassDB 100、GuassDB 200、Greenplum（V9.5目前不支持Greenplum数据库）、星环（用户名密码方式　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）、Sybase、MariaDB、MonetDB。

⚠
- 从V10版本开始，新用户ETL目标数据源默认不能选择高速缓存库，如需要选择高速缓存库，需在系统选项/高级设置项中，添加或开启设置项： DISABLE_WRITE_TO_SMARTBI_CACHE=false（true 是不能选）
- 旧版本升级到10版本及以上，默认自动加设置项，保证旧用户能继续在ETL目标源中选择高速缓存库
- 新用户使用高速缓存库流程：
  1.系统选项->高级设置->添加:DISABLE_WRITE_TO_SMARTBI_CACHE=false
  2.使用任意数据集完成一次抽取（若未使用抽取时，ETL节点无法选择对应schema）

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SCHEMA。
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，详情请参考 目标表。
回退模式	回退模式用于在插入数据前先把满足条件的数据删除，可实现增量删除，详情请参考 回退模式。

目标表

1.新建表时，支持添加字段别名到数据库

支持的数据库有：MYSQL、INFOBRIGHT、DB2、ORACLE、POSTGRESQL、GREENPLUM、SYBASE、GBASE（8a版本）、DAMENG（7版本）、CLICKHOUSE、GAUSS100。

CLICKHOUSE、GAUSS100支持添加字段别名到数据库，但在数据源表不显示已有的注释。

2.当用户选中的目标表的order by字段跟主键字段不一致时，用户可通过该节点重新建表指定主键，或者通过数据库中将该表的主键字段更改为order by字段一致。

3. 当数据源为 ClickHouse ，且当前选中的目标表无主键时，则需要用户手动指定更新依据字段。

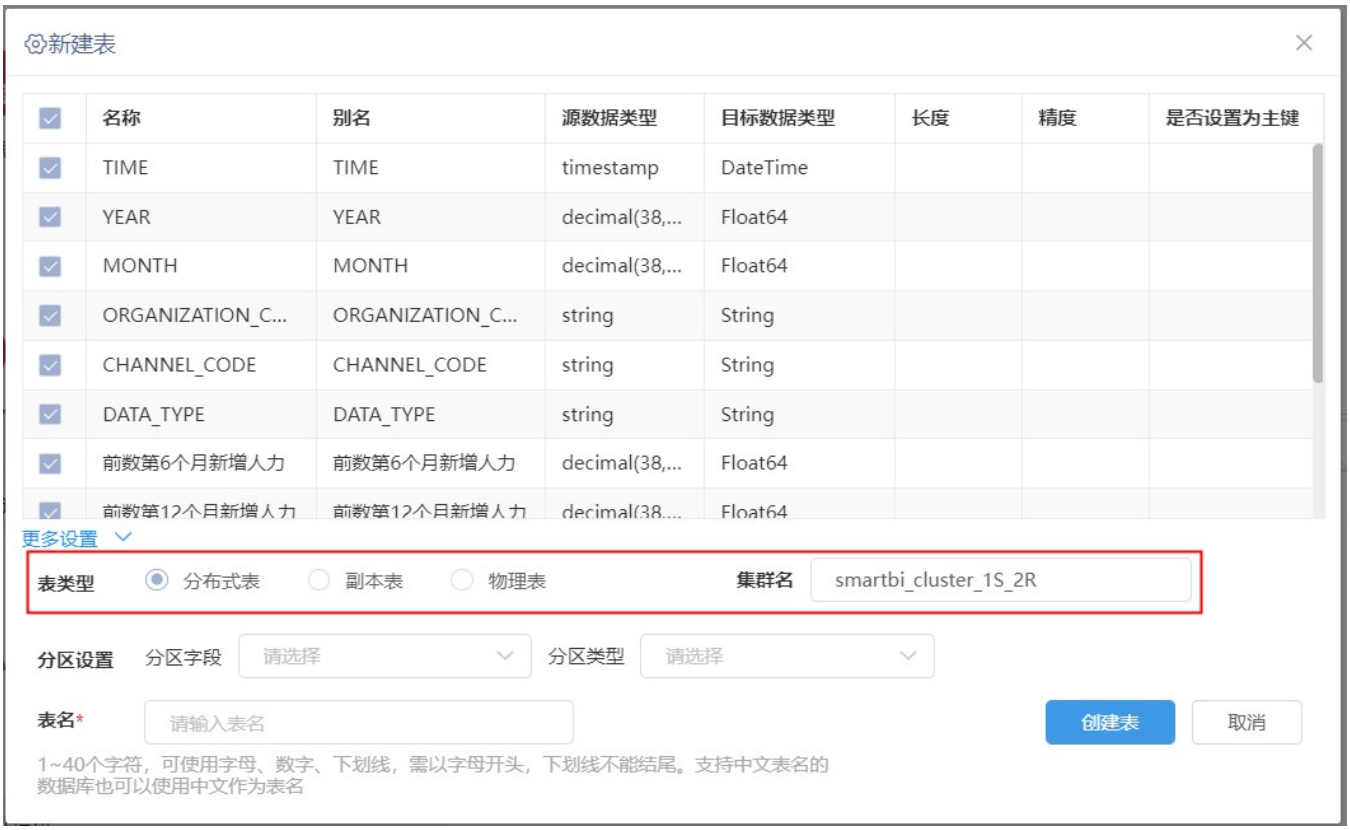
否则节点将会执行失败，提示信息如下：

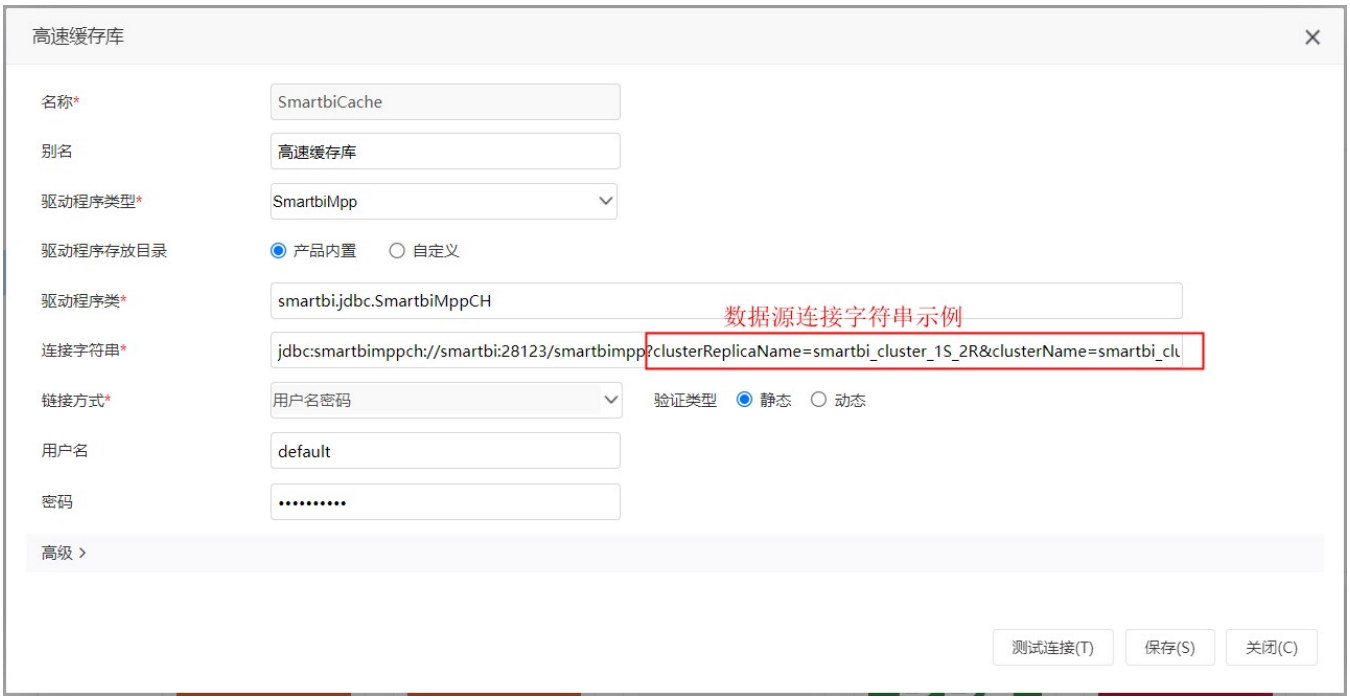
4. 如果数据源连的是ClickHouse集群环境，支持新建 分布式表、副本表、物理表。

分布式表和副本表的集群名，默认是读数据源连接字符串中的集群名称，用户可以手动修改：

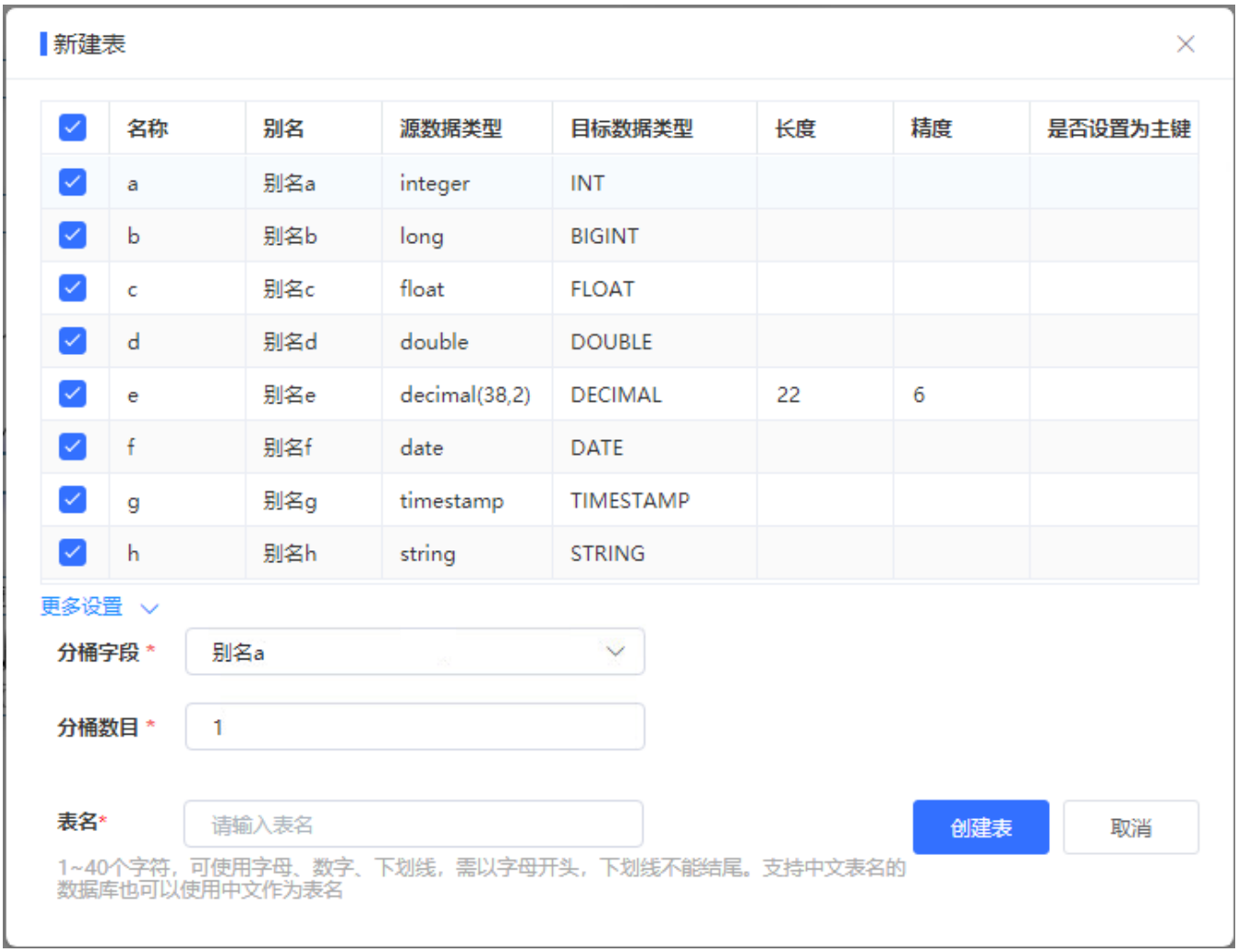
分片集群的数据源连接字符串的属性名：clusterName

副本集群的数据源连接字符串的属性名：clusterReplicaName





5. 如果数据源连的是星环数据库时，支持建orc非分区事务表。



分桶字段默认为第一个非decimal类型的字段，分桶数目默认为1，两者用户均可根据实际情况，自行修改调整。

- 产品从10.5.15版本及以上才支持星环建表功能。
- 星环数据库认证方式必须是用户名和密码方式。
- 如果是数据量特别大的表，建议客户建表时要根据实际业务场景手工调整分桶字段和分桶数，不要使用默认值。分桶表一旦建立，后续星环数据库不支持再对该表进行分桶处理。

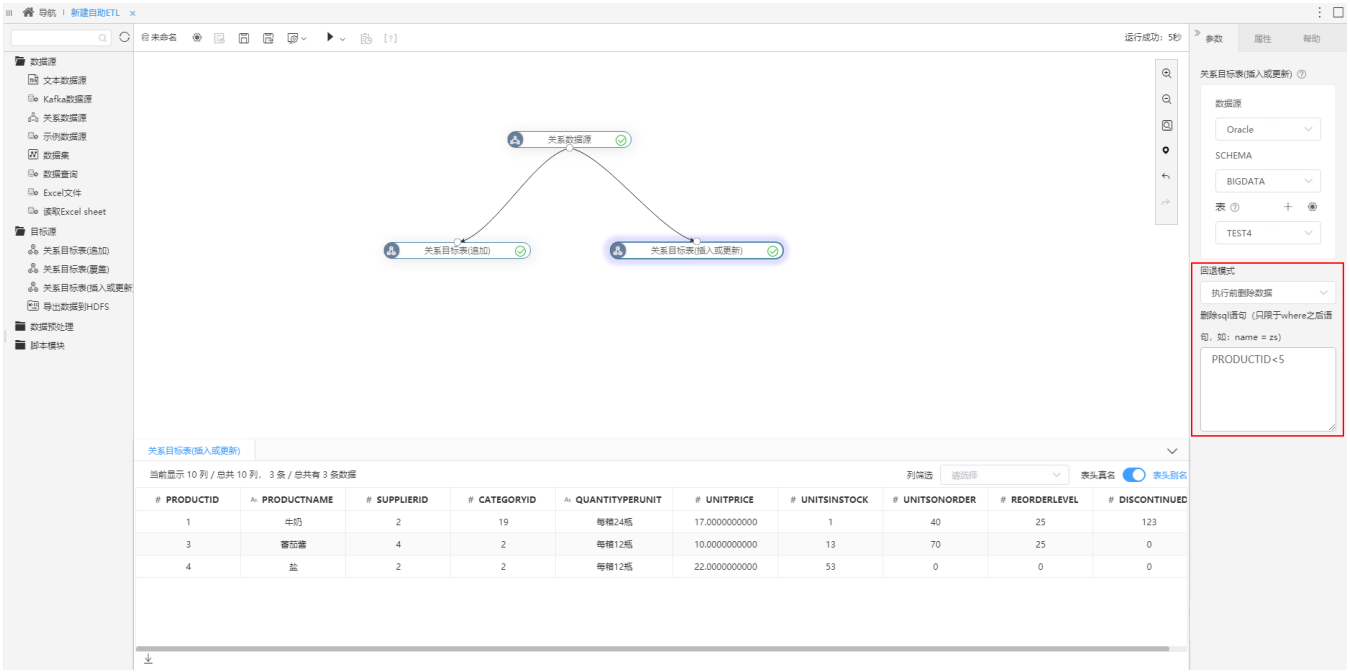
回退模式

回退模式用于在插入数据前先把满足条件的数据删除，可实现增量删除。

- 无（默认）；
- 执行前删除数据：根据填写的删除sql语句条件，可实现在插入数据前先把满足条件的数据删除掉。

关系目标表（追加）、关系目标表（插入或更新）节点支持回退模式功能。

在参数设置区中，回退模式选择”执行前删除数据“，在删除sql语句框中填写删除语句（条件SQL使用表头真名）。



如上图，根据条件会先删除”PRODUCTID<5“的数据，然后根据节点功能更新、追加数据：关系目标表（追加）节点会直接追加新增的数据；关系目标表（插入或更新）节点会先更新原有的数据，然后再插入新增的数据。

> ⚠ 在回退模式填写SQL语句时，由于GuassDB 200数据库中默认字段为小写，所以字段为大写需要添加双引号才能生效。

应用场景：用户在进行ETL调度时，发现某天调度的数据有问题，需要进行重跑（把之前已经入库的数据删除再插入），可以使用此功能可以先把入库的数据删除，再将新数据追加到目标表中。

导出数据到HDFS

概述

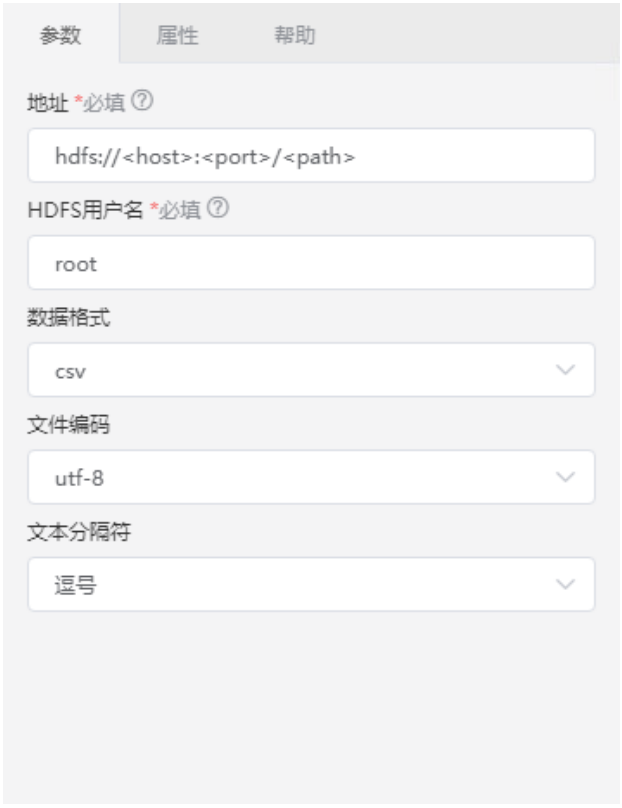
导出数据到HDFS是指将结果数据保存到HDFS中。

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
地址	目标HDFS的路径的IP和端口以及导入的目录：hdfs://<host>:<port>/<path>； 示例：hdfs://10.10.204.130:9000/home/hadoop。
HDFS用户名	HDFS用户名。
数据格式	CSV（CSV文件格式请指定文件的编码格式：utf-8/gbk，以及分隔符）、 Parquet、Orc