Smartbi V10.5-数据挖掘

注意：（新特性列表中：+表示新增；^表示增强）

具体改进点如下：

新增	增强
+【数据挖掘】文本分析新增LSH、相似集计算（LSH）节点	^【数据挖掘】关系目标表支持GaussDB 200数据库
+【数据挖掘】支持导出PMML模型文件	^【自助ETL/数据挖掘】关系数据源节点兼容更多数据源
+【数据挖掘】朴素贝叶斯、决策树、多层感知机等算法支持自动调参设置	^【数据挖掘】异常值处理节点新增删除异常行功能
+【数据挖掘】新增ETL和挖掘实验日志	
+【自助ETL/数据挖掘/ETL高级查询】数据源新增FTP数据源	

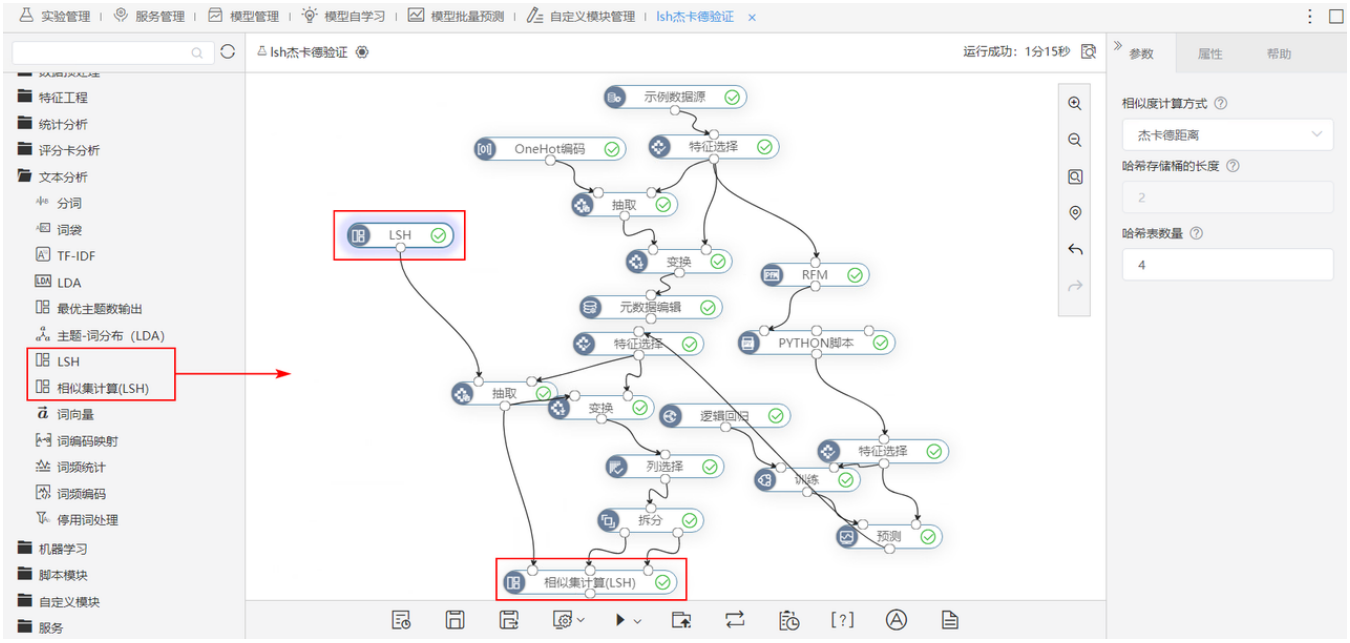
+【数据挖掘】文本分析新增LSH、相似集计算（LSH）节点

背景介绍

我们面对和需要处理的数据往往是海量并且具有很高的维度，怎样快速地从海量的高维数据集合中找到与某个数据最相似的一个数据或多个数据成为了一个难点和问题。如果是对一个海量的高维数据集采用线性查找匹配的话，会非常耗时，因此，为了解决该问题，我们可以使用LSH算法这类索引的技术来加快查找过程。

功能简介

新版本，文本分析中新增LSH、相似集计算（LSH）节点，支持计算大规模数据的相似度。



参考文档

详情请参考 数据挖掘 - LSH、数据挖掘 - 相似集计算(LSH) 。

+【数据挖掘】支持导入和导出PMML模型文件

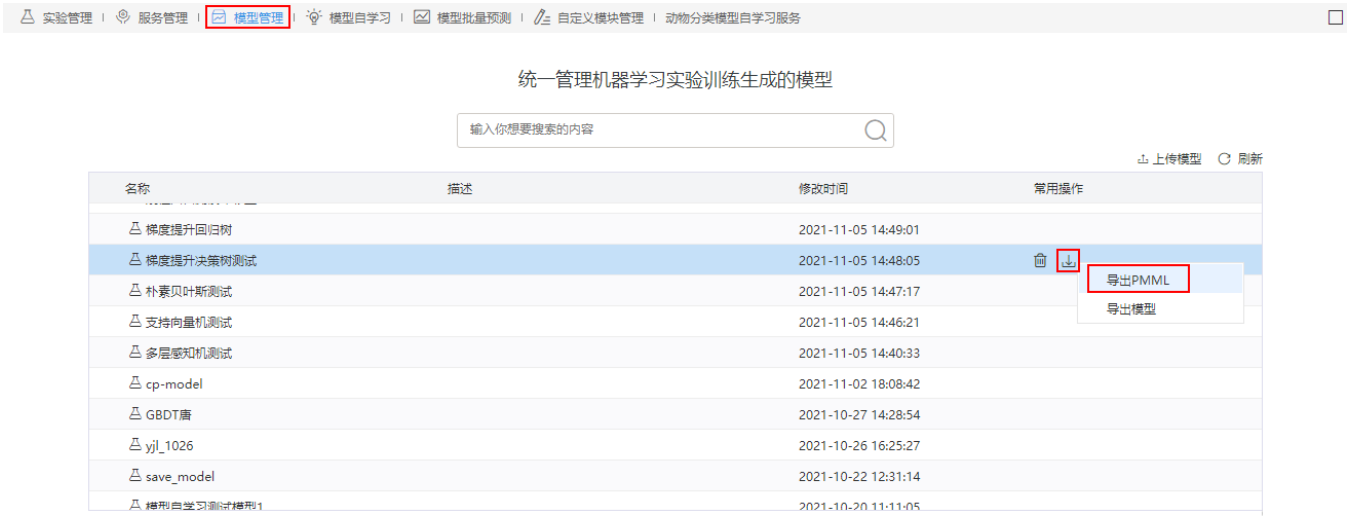
背景介绍

以前的版本，用户往往需要一次建立多个数据挖掘模型， 想要通过导出PMML文件快速进行跨平台的应用和部署。为了满足用户需求，新版本产品支持将训练得到的模型转化为PMML模型文件，用户可将PMML模型文件载入Python或其他平台中进行预测，适应更多环境；同时也支持上传其他服务器导出的模型文件，大大节省了等待时间。

功能简介

1、导入PMML模型文件：在模型管理页面中，新增“上传模型”按钮，支持上传其他服务器模型导出模型文件，并生成对应的模型记录。

2、导出PMML模型文件：已训练的模型右键更多中和模型管理页面的常用操作中，新增“导出PMML模型”功能，支持导出选中模型的PMML模型文件。



参考文档

关于导入和导出PMML模型文件，详情请参考 模型管理 。

^【数据挖掘】异常值处理节点新增删除异常行功能

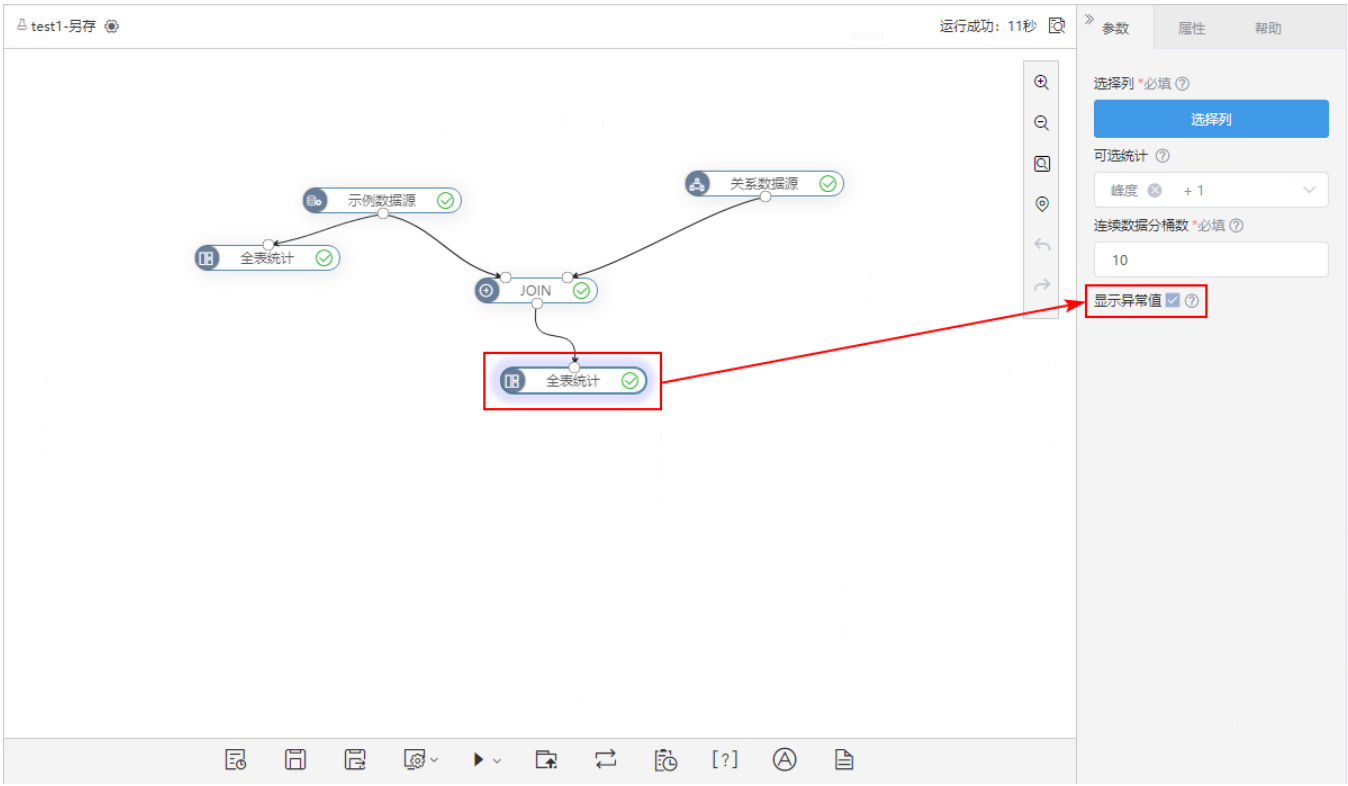
背景介绍

在实际应用中，用于分析的数据常常会存在大量的异常值，影响正常的分析结果。新版本，在产品异常值处理节点新增删除异常行处理策略，可用于直接删除一些异常值存在的行，提高数据质量。

功能简介

1、异常值处理节点的配置界面，新增“删除异常行”处理策略，可用于直接删除一些异常值存在的行，为生成训练集做准备。

2、全表统计节点新增“显示异常值”选项。



支持在输出结果的箱线图中显示异常值。

参考文档

详情请参考 数据挖掘-异常值处理 、数据挖掘-全表统计 。

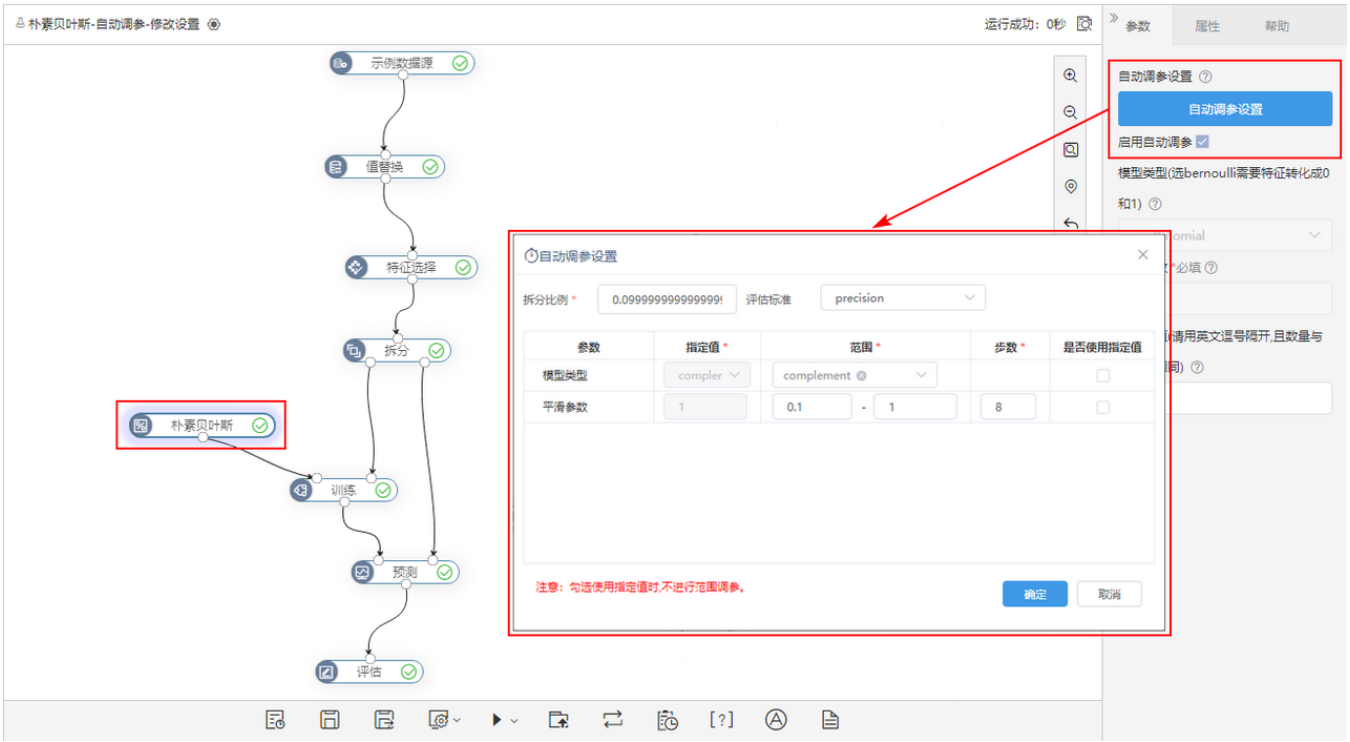
+【数据挖掘】朴素贝叶斯、决策树、多层感知机等算法支持自动调参设置

背景介绍

在实际应用中，用户在进行机器学习时需要对选择的模型进行调参，从而得到最优参数进行匹配，而手动调参往往需要耗费大量时间和人力，可以采取系统自动调参的方式解决问题。此前产品部分机器学习算法已实现了自动调参设置，新版本产品更多的算法支持自动调参设置，帮助用户节约时间和人力，提升数据分析效率。

功能简介

新版本，朴素贝叶斯、决策树、多层感知机、随机森林、梯度提升回归树算法支持自动调参设置，系统可对设置指定或范围内的参数值循环调参，匹配出最优的组合。



参考文档

详情请参考　数据挖掘-朴素贝叶斯、数据挖掘-决策树、数据挖掘-多层感知机、数据挖掘-随机森林、数据挖掘-梯度提升回归树　。

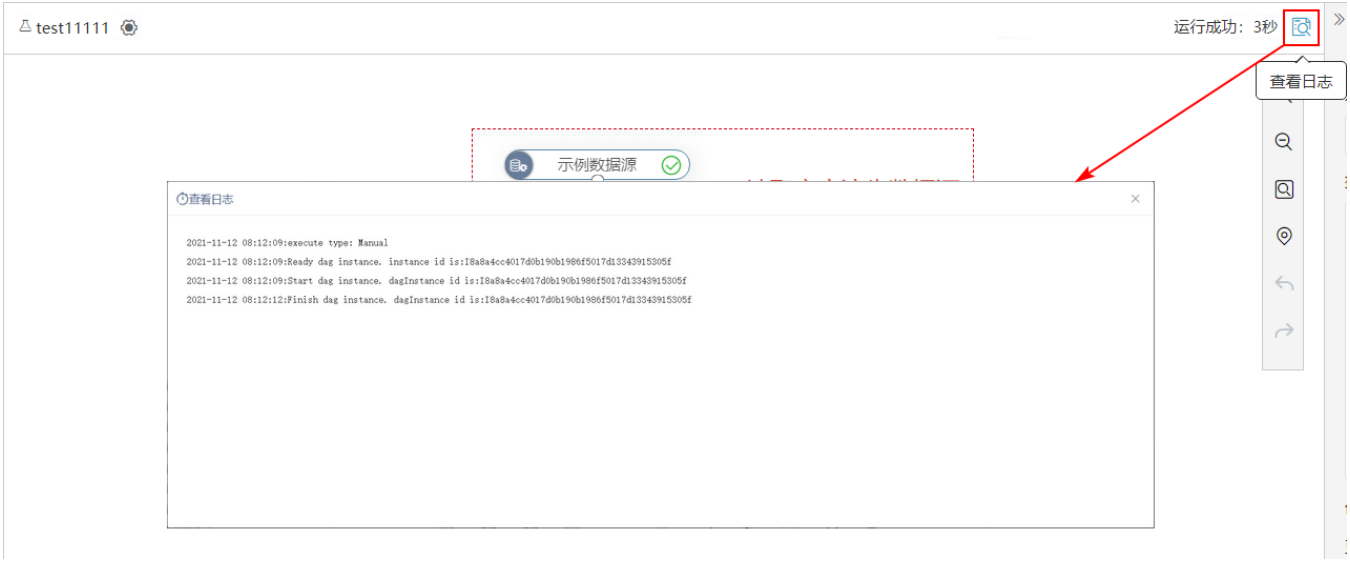
+【数据挖掘】新增ETL和挖掘实验日志

背景介绍

在实际应用中，用户使用数据模型的高级ETL查询时，进行某次数据抽取失败，右上角只显示“初始化”等运行状态，因此想要查看更加详细的运行状态信息，便于及时排查、定位问题。为了满足用户需求，新版本在高级ETL查询、自助ETL、数据挖掘中，新增ETL和挖掘实验日志功能，用于记录ETL和挖掘实验运行状态信息。

功能简介

新版本，在高级ETL、自助ETL、数据挖掘中新增“查看日志”功能，用于记录ETL和挖掘实验运行状态信息。



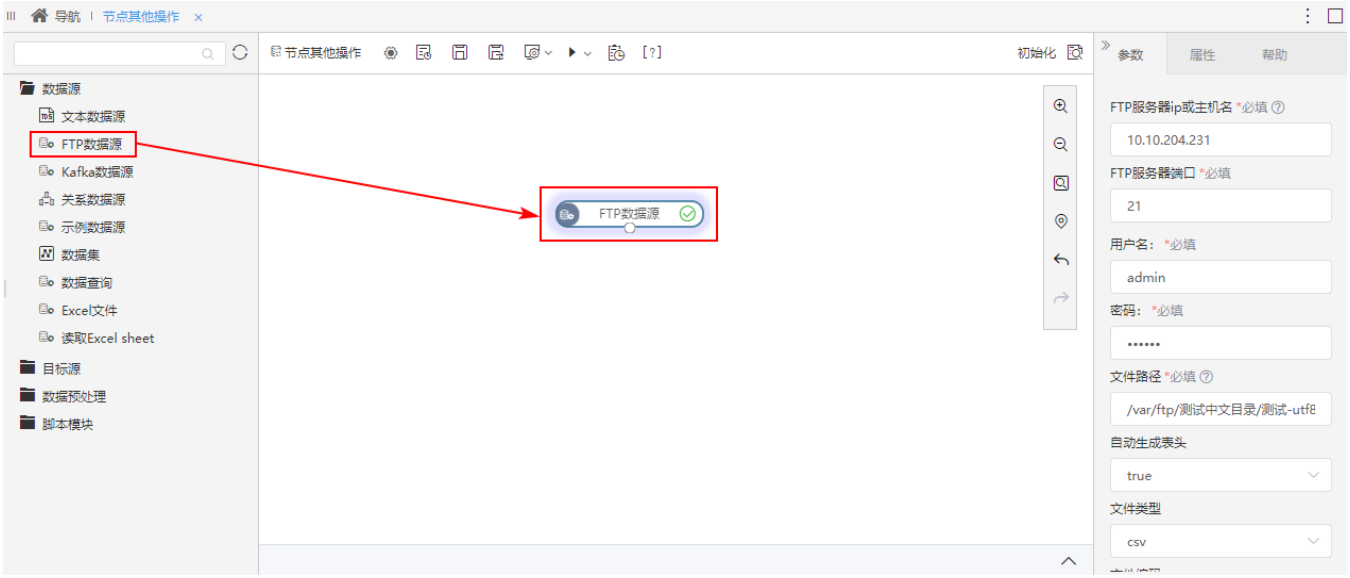
参考文档

详情请参考　ETL高级查询-界面介绍、自助ETL-界面介绍、　数据挖掘-实验界面介绍　。

+【自助ETL/数据挖掘/ETL高级查询】数据源新增FTP数据源

功能简介

新版本，在自助ETL、数据挖掘、数据模型的ETL高级查询中，数据源新增FTP数据源，支持通过FTP方式读取Excel、CSV文件数据。



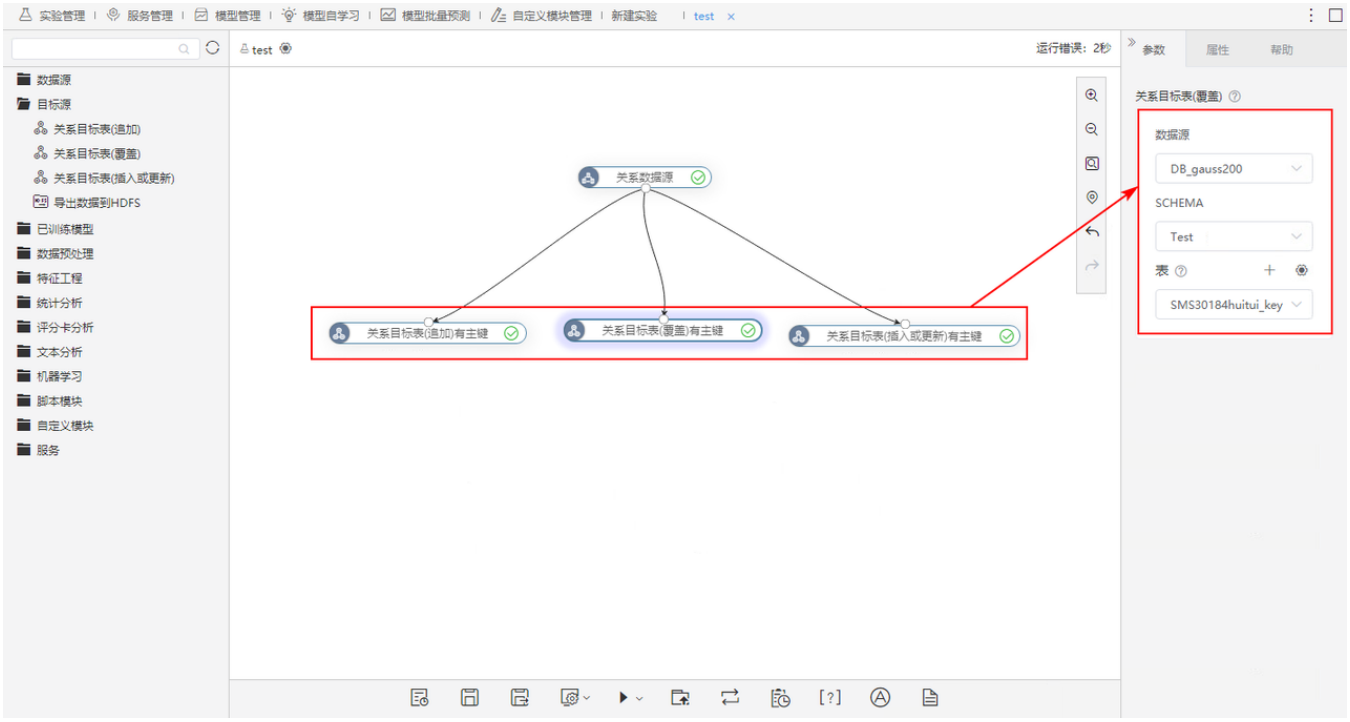
参考文档

关于FTP数据源设置说明，详情请参考 数据挖掘-数据的输入和输出 。

^【数据挖掘】关系目标表支持GaussDB 200数据库

功能简介

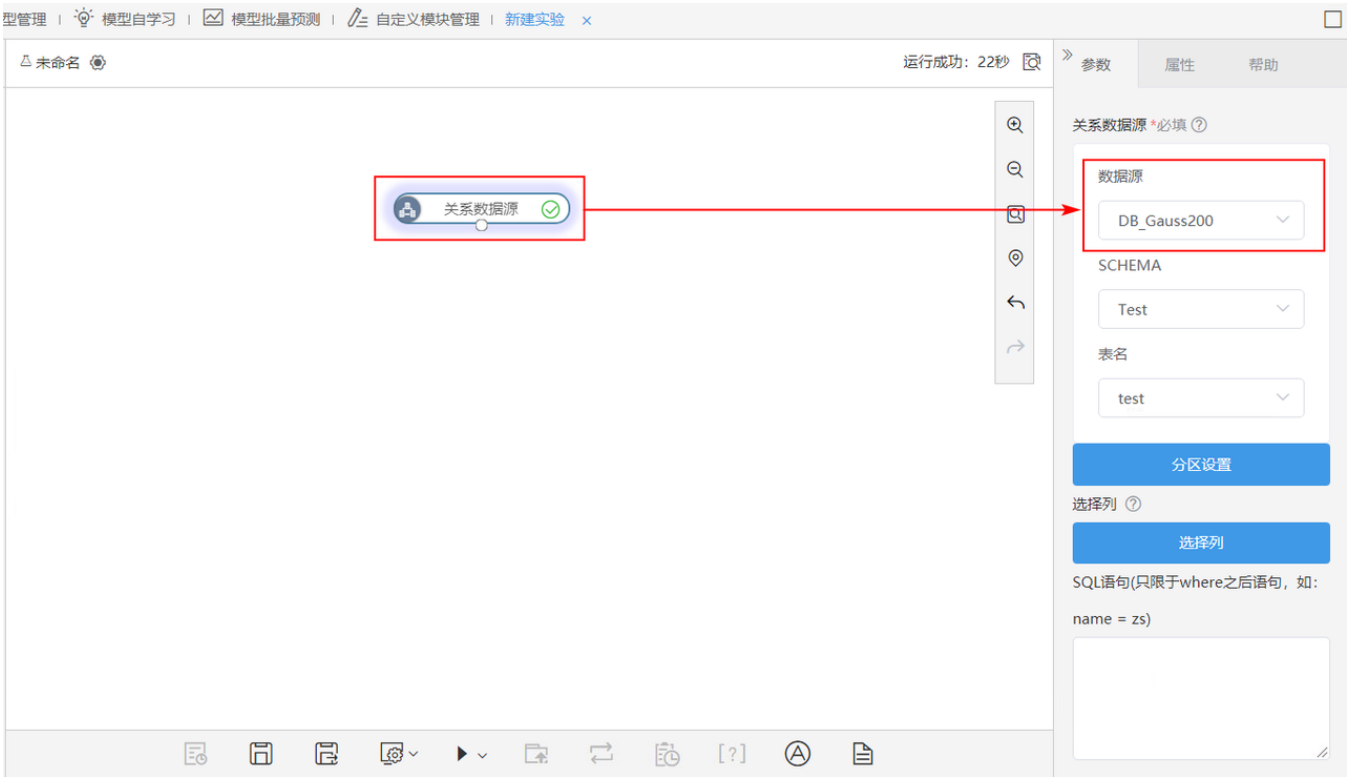
新版本，关系目标表（追加）、关系目标表（覆盖）、关系目标表（插入或更新）支持GaussDB 200数据库。



^【自助ETL/数据挖掘】关系数据源节点兼容更多数据源

功能简介

新版本，关系数据源节点兼容更多数据源，包括：Kingbase、Kingbase_V8、Kingbase AnalyticsDB 、GaussDB 200、Teradata、Teradata_V12、神通、Obase、Informix、Kylin、Impala。



注意事项

其中， 关系数据源 KingbaseAnalytics、ShenTong集群暂不支持小批量运行。