Smartbi V9-数据准备

【数据准备】模块**加强功能，提高数据抽取效率，缩减数据准备工作的时间**，具体优化如下：

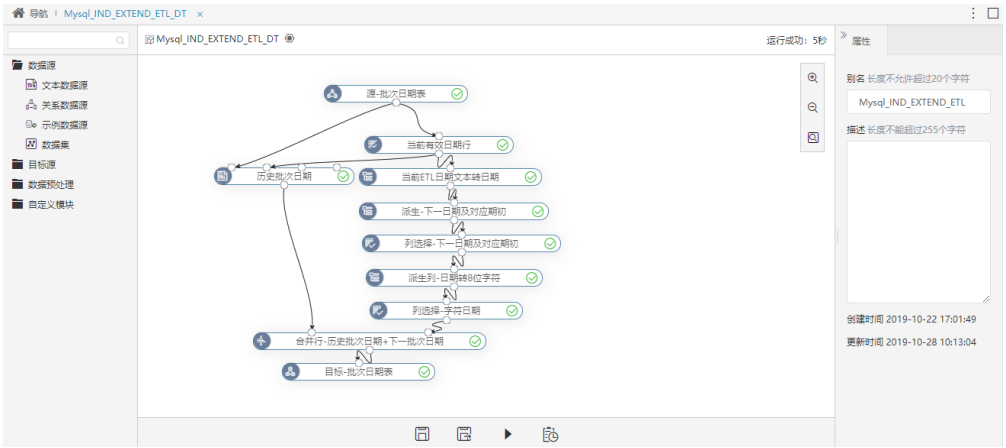
- **功能加强**
  a. 自助数据集支持更多的数据来源及支持参数
  b. 自助数据集的计算字段使用统一的函数
  c. 数据集支持增量抽取
  d. SmartbiMPP支持集群抽取
  e. 抽取支持自定义表名
  f. 业务主题增加多个时间维度层次
- **性能提升**
  采取分批次读数据和多线程并发抽取，从而提高抽取效率

新增	+【自助ETL】增加自助ETL +【自助数据集】自助数据集支持参数 +【自助数据集】自助数据集支持更多的数据来源 +【数据抽取】分批次读数据和多线程并发抽取从而提高抽取效率
增强	^【自助数据集】统一计算字段使用的函数 ^【数据集】支持增量抽取 ^【数据抽取】SmartbiMPP支持集群抽取 ^【数据抽取】抽取支持自定义表名 ^【业务主题】时间维度管理增加"半年""旬""周" ^【数据准备】小优化

+【自助ETL】增加自助ETL

功能简介

V9及之后版本增加自助ETL。自助ETL代替传统的SQL语句和存储过程，使用可视化流程设计模式实现数据处理，大大降低了数据处理的难度，让业务人员也能介入到数据处理环节。



详情参考

关于自助ETL的说明，详情请参考 自助ETL。

+【自助数据集】自助数据集支持参数

背景介绍

自助数据集的数据来源为带参数的hana数据源时，期望自助数据集支持参数。因此，我们对其进行优化，V9及之后版本自助数据集支持参数。

功能简介

V9及之后版本自助数据集的数据来源为“hana数据源”时，自助数据集支持其所带参数。

自助数据集增加“设置参数（[?]）”设置项。

注意事项

参数是否含有默认值对自助数据集抽取的影响：

- 参数没有默认值，自助数据集进行抽取，则抽取全部数据。
- 参数含有默认值，自助数据集进行抽取，则只会抽取参数默认值相关的数据。

+【自助数据集】自助数据集支持更多的数据来源

功能简介

V9及之后版本所有关系数据源都可作为自助数据集的数据来源；增加“多维数据集”作为自助数据集的数据来源。

注意事项

“kingbase、神通、达梦6、达梦7”数据源暂不支持跨库。

详情参考

关于自助数据集的数据来源，详情请参考 自助数据集-数据来源。

+【数据抽取】分批次读数据和多线程并发抽取从而提高抽取效率

背景介绍

数据抽取存在如下两种情况：

1、之前的版本，一次性将所有数据从数据库写入到缓存库中，在数据量较大的情况下，会出现占用内存太大的问题。因此，我们对其进行优化，V9及之后版本支持分批次将数据写入到缓存库中，合理设置既能保证查询速度又避免占用太大内存，达到最高效率。

2、当抽取亿级别数据量时，如果为单线程抽取，容易出现抽取缓慢，抽取不成功甚至环境崩溃的情况。因此，V9及之后版本增加多线程并发抽取，在用户抽取较大数据量时，可设置多线程并发抽取。

以SmartbiMPP抽取1亿条数据为例，单线程抽取需要耗时2.5小时；设置10个线程进行抽取时，仅需48分钟，速度提升了3倍。

功能简介

1、在“**系统选项 > 查询设置**”界面增加“数据抽取时每次读取数据条数（JDBC Fetchsize）”设置项，用于设置一次从数据库读取的数据条数。

2、在“**系统选项 > 查询设置**”界面增加“数据抽取并发线程数”和“数据抽取时每页记录数”这两个设置项，支持多线程并发抽取。

数据抽取的处理机制如下：

1）数据抽取并发线程数等于1，抽取不做分页导出。

2）数据抽取并发线程数不等于1，抽取做分页导出，分为两种情况：

- 抽取总数量/数据抽取时每页记录数<=数据抽取并发线程数，抽取与“数据抽取时每页记录数”相关。

  抽取页数＝抽取总数量/数据抽取时每页记录数，有余数加1取整数。

- 抽取总数量/数据抽取时每页记录数>数据抽取并发线程数，抽取与“数据抽取并发线程数”和“数据抽取时每页记录数”都有关。

  抽取分批次抽取：

  一批的抽取数量=数据抽取并发线程数×数据抽取时每页记录数。

  批数=抽取总数量/一批的抽取数量，有余数加1取整数。

^【自助数据集】统一计算字段使用的函数

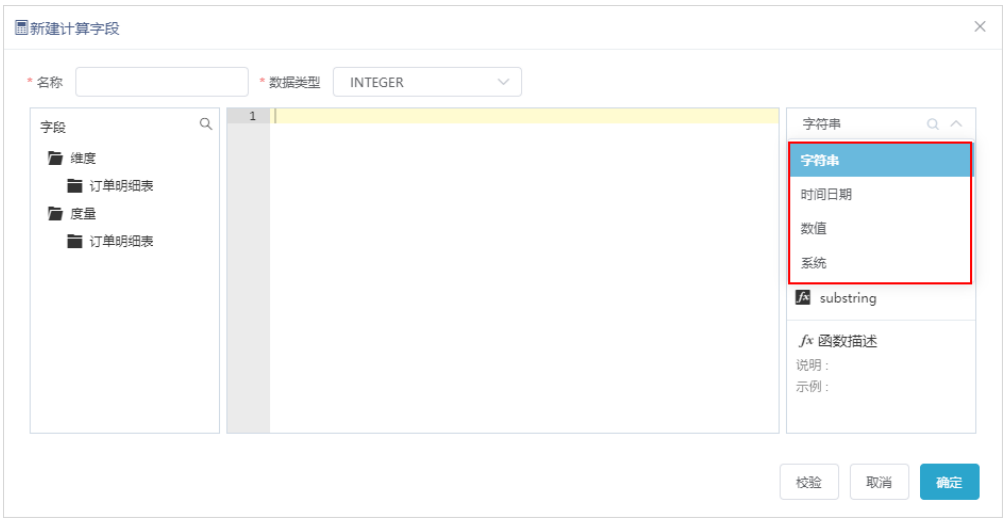
背景介绍

之前的版本自助数据集计算字段支持的函数是根据其所属数据库类型决定的，这种方式有个弊端在于当我们切换高速缓存库时会存在函数不兼容的问题，导致在抽取时报SQL错误。针对这一弊端，也结合产品的使用，V9及之后版本我们基于SQL92为标准，封装一套Smartbi自身的函数语法，用于适配Smartbi所支持的数据库，暂不包括“Teradata_v12”和“aliyun AnalyticDB”这两个数据库。

功能简介

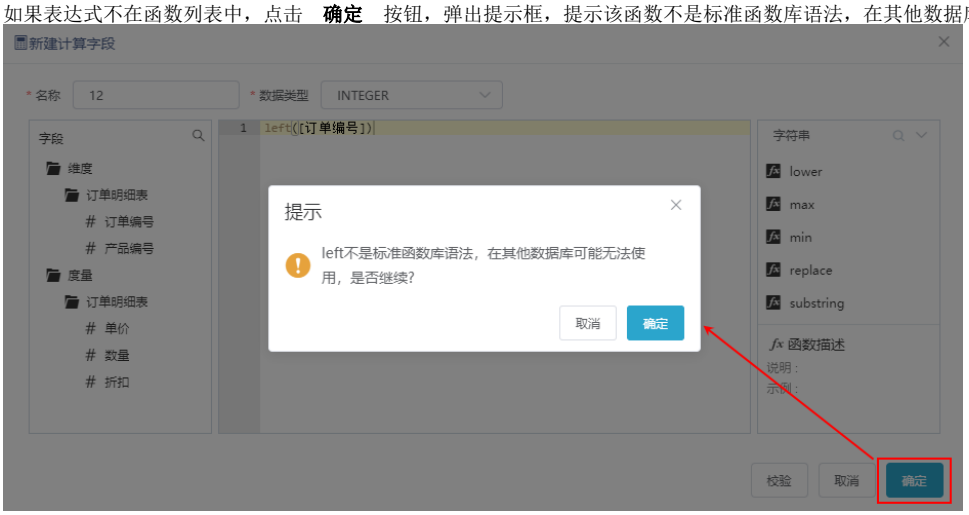
1、V9及之后版本封装一套Smartbi自身的函数语法，用于适配Smartbi所支持的数据库，暂不包括“Teradata_v12”和“aliyun AnalyticDB”这两个数据库。

函数分为四种类型：字符串、时间日期、数值、系统。

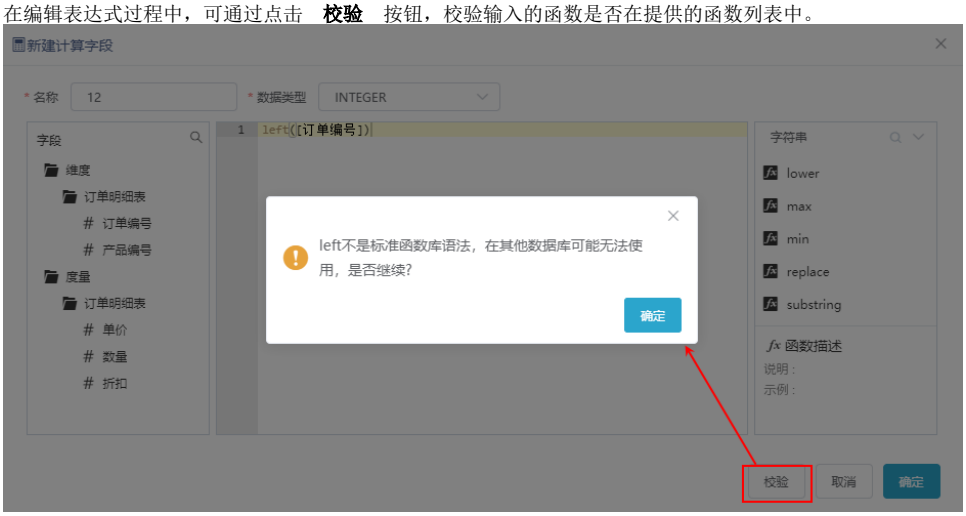


2、新增函数校验功能，校验输入的函数是否在提供的函数列表中。

- 如果表达式不在函数列表中，点击 **确定** 按钮，弹出提示框，提示该函数不是标准函数库语法，在其他数据库可能无法使用。



- 在编辑表达式过程中，可通过点击 **校验** 按钮，校验输入的函数是否在提供的函数列表中。



注意事项

1、对于ClickHouse数据库，用该自助数据集创建自助仪表盘时，对于一个组件要求：设置了聚合方式的字段与用该字段创建的计算字段不能同时使用。

2、对于HadoopHive数据库，用该自助数据集创建透视分析时，要求透视分析只能使用一个包含getdate函数的计算字段。

3、在创建计算字段时使用了getdate函数，且数据类型选择为“TIME”类型，不支持应用于透视分析和电子表格，原因为：元数据是DATE类型，不支持转换为TIME类型。

详情参考

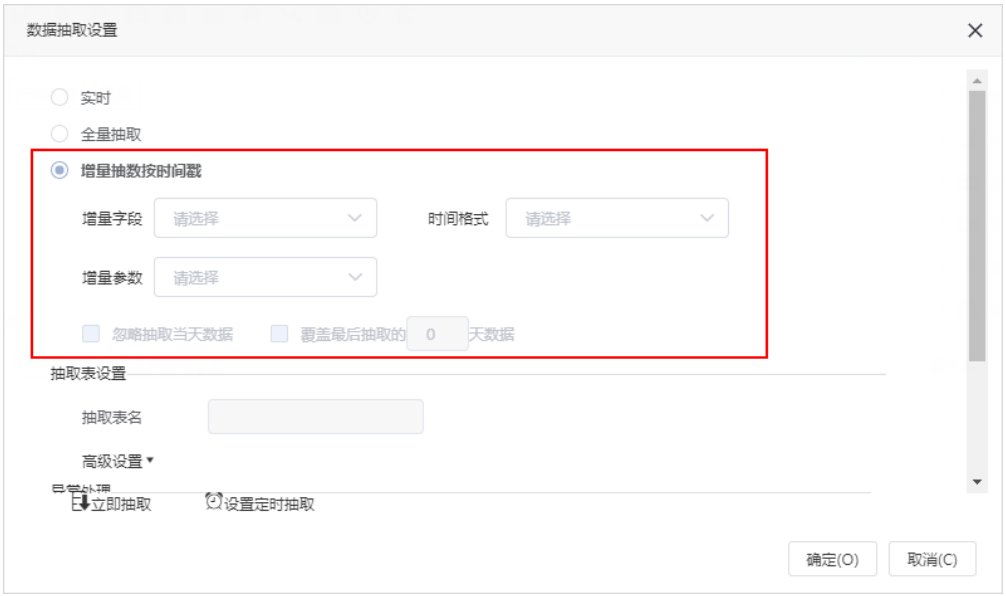
关于自助数据集的计算字段，详情请参考 自助数据集-计算字段。

^【数据集】支持增量抽取

功能简介

“可视化数据集、原生SQL数据集、SQL数据集、存储过程数据集、Java数据集”支持增量抽取。

在“抽取设置界面”增加“增量抽取按时间戳”选项，如图：



详情参考

关于增量抽取的说明，详情请参考 数据抽取 。

^【数据抽取】SmartbiMPP支持集群抽取

背景介绍

在实际运用中，期望SmartbiMPP支持集群抽取。结合产品的使用，我们对其进行优化，V9及之后版本SmartbiMPP支持集群抽取。

功能简介

V9及之后版本高速缓存库连接的驱动程序类型是“SmartbiMPP”时，支持集群抽取。

在“抽取设置”界面增加“高级设置”，对“分区字段”和“分区类型”进行设置。其中，分区字段要求为日期类型。

注意事项

在不清空数据的情况下，SmartbiMPP多次集群抽取以首次选择的集群分区类型为准。如首次抽取分区类型选择“年”，再次抽取且不清空数据时，分区字段选择“月/日”，抽取后表数据追加，仍按年分区。

详情参考

关于数据抽取，详情请参考 数据抽取。

^【数据抽取】抽取支持自定义表名

背景介绍

之前的版本，数据集和即席查询抽取保存在高速缓存库的表，默认以“数据集ID”作为表名称，“数据集名称”作为表别名，在数据库查看表时，以数据集ID作为表名称，不利于用户直观查找需要的表，因此我们对其进行优化，V9及之后版本，数据集和即席查询抽取保存在高速缓存库的表支持自定义表名。

功能简介

1、数据集和即席查询抽取增加“抽取表名”设置项，支持自定义表名。

2、抽取表名的处理逻辑如下：

1）抽取表名分两种情况：

- 不设置抽取表名：默认以“数据集ID”作为表名称，“数据集名称”作为表别名。
- 设置抽取表名：以“输入的表名”作为表名称和表别名。

2）再次抽取表名分两种情况：

- 第一次不设置抽取表名：再次抽取不支持设置抽取表名。
- 第一次设置抽取表名：再次抽取默认以第一次抽取设置的表名，不支持修改。

3）设置抽取表名重复时，提示“当前抽取名称已存在，请修改。”。如不修改执行抽取会提示其他抽取资源已经占有该表。

注意事项

1、输入的表名不支持“@#$%^&*{}[]/”等特殊字符。

2、抽取到“星环”和“Presto+Hive”高速缓存库，输入的表名不支持“中文”。

详情参考

关于数据抽取，详情请参考 数据抽取。

^【业务主题】时间维度管理增加“半年”“旬”“周”

背景介绍

业务人员经常需要做统计分析报表，如月汇报、周汇报等。为了满足更多维度的统计分析，V9及之后版本维度管理增加支持“半年”“旬”“周”。

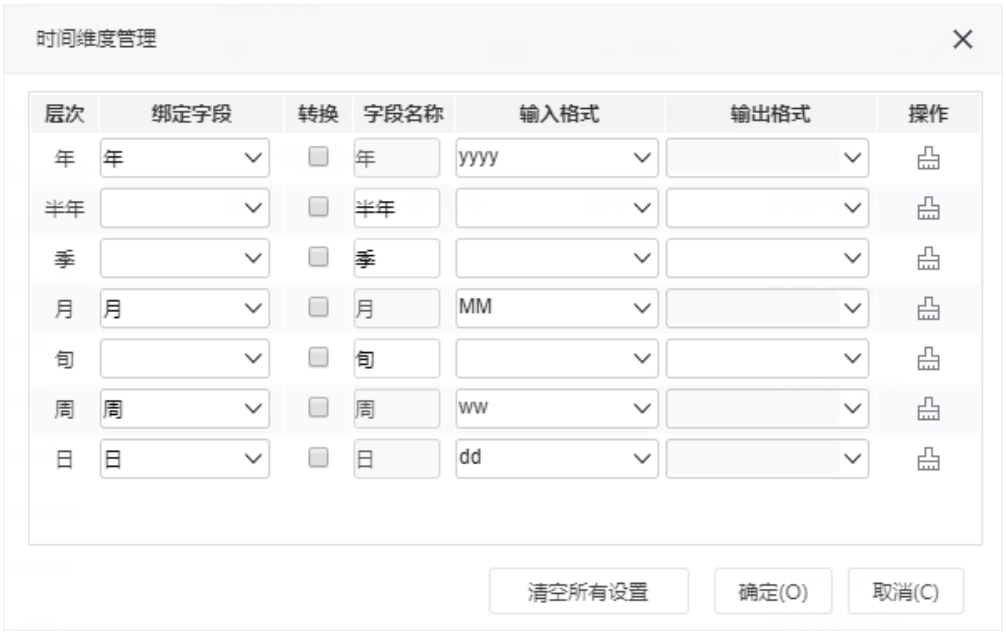
功能简介

1、V9及之后版本业务主题的时间字段生成时间层次增加“年半年季月旬日”“年周”两种。

其中，“半年”“旬”“周”的显示格式为：

- 半年：上半年、下半年。
- 旬：上、中、下。
- 周：Www（W不变，就是一个字母；ww标识第几周。如2019年第一周为：2019-W01。）

2、在透视分析的待选列区的时间维度管理增加“半年”“旬”“周”层次。



^【数据准备】小优化

1、自助数据集的预览数据界面提示文字改为“刷新实时数据”，表明数据是数据库的实时数据。

2、完善自助数据集的筛选器操作符，具体如下：

- 筛选器字段类型为“integer”、“datetime”，增加“模糊匹配”和“不匹配”操作符。
- 筛选器字段类型为“string”，增加“开头为”和“结尾为”操作符。

3、数据集的抽取日志采取分页加载，默认先加载前30条数据，按照时间降序排序。当滚动条移动到底部时会自动加载下一页数据。