直播介绍

SMARTBI 思迈特软件

01. 课程宣讲

02. 有奖竞答（奖品：188麦豆）

03. 互动答疑

04. 问卷抽奖（奖品：188麦豆）

SMARTBI
思迈特软件

自助ETL最佳实践（下）

Ai 上与数据聊天

https://demo.smartbi.com.cn/smartbi/vision/index.jsp

SMARTBI
思迈特软件

Contents
目录

SMARTBI
思迈特软件

01 自助ETL定位说明

什么是自助ETL?

SMARTBI 思迈特软件

自助ETL模块提供了可视化的操作界面，大量组件兼顾一般数据处理与高级数据处理，包括行列转换、合并行列、去重、聚合、Join、增加序列号等，通过简单的“拖拉拽”操作就可将多来源的异构数据加工成具备语义一致性与完整性的数据模型，为数据分析和报表开发提供了前提保障。

01

代替传统的SQL语句和存储过程

02

业务人员也可介入到数据处理环节

03

可视化流程设计模式实现数据处理

为什么需要自助ETL?

01

将**分散的数据整合**到一个中心化的位置（如数据仓库），**形成统一的数据视图，**便于分析和决策制定。

02

通过内置的数据清洗和转换功能，自助ETL能**有效处理数据，**显著**提高数据质量，**确保分析结果的准确性。

03

自助ETL支持将数据迁移至更高效的存储和分析平台，能够**高效查询和分析大量数据，提高查询性能。**

04

运行复杂的分析查询可能会对生产系统造成压力。自助ETL可以将数据加载到专门的分析系统，**减轻源系统的负担。**

05

原有的数据表可能已无法满足当前的分析需求。自助ETL能够对这些存量数据**进一步的加工和处理，**以满足新的分析要求。**提高数据的可用性。**

自助ETL的产品定位及核心场景

定位：专注于数据准备的工具，从各种数据源中提取、清洗和整合数据，用于后续的分析展示使用。
目标用户：业务人员、业务分析师、数据分析师、技术人员

目标用户的关键特征

存量数据不满足需求，在数据展示和分析前，需要进一步处理数据。

- **数据建设不完善：**未搭建专业规范的数据仓库，数据未达到可直接使用的状态
- **业务需求个性化：**数仓建设主要满足通用的数据使用场景，个性化的业务场景未能满足
- **敏捷开发的需求：**希望在做报表展示、数据分析之前或做的过程中，能够快速调整数据

典型场景

- **自助数据分析：**当前数据不满足需求，对数据进行二次加工；
- **大屏展示：**展示内容缺少对应数据表/字段，对数据进行二次加工；
- **轻量数据集成：**为了局部分析场景/业务流程自动化，进行数据整合。例如将金蝶云/钉钉/企微/SaaS应用的数据双向打通，进行关联分析。
- **轻量数据集市：**涉及的数据表不多，从不同的数据源中抽取数据，并将数据进行转换和清洗，然后加载到数据集市，满足更多分析场景。

模块功能特点

- ✓ 直观的可视化操作界面
- ✓ 丰富的数据接入及处理等节点
- ✓ 实时执行及预览数据
- ✓ 支持自定义脚本处理数据
- ✓ 通过作业流配置任务调度

Smartbi 产品图谱

SMARTBI 思迈特软件

应用商店

- **金融** 银行|保险|证券…
- **政府** 政数局|财政|工商…
- **教育** 高校|高职…
- **科技** 互联网|软件…
- **能源** 石油|电力|电网…
- **制造** 制药|汽车|冶金|…
- **运营商** 电信|联通|移动|广电…
- **零售** 线上|线下
- …

平台产品

- 一站式ABI平台
- 智慧数据运营平台
- 机器学习平台
- 电子表格软件
- 嵌入式BI
- SaaS BI云平台

分析工具

传统BI → 现代BI（自助BI+智能BI）

- **电子表格** Excel电子表格 Web电子表格
- **分析报告** Word分析报告 PPT分析报告
- **大屏可视化** 炫酷美观大屏 语音操控大屏
- **即席查询** 自助式 明细数据查询
- **透视分析** 无需建模的 OLAP分析
- **交互式仪表盘** 实时业务监控预警/ 数据大屏..
- **对话式分析** 基于自然语言的 智能数据分析
- **数据挖掘** 预测性分析 机器学习
- **分享协同** 帮助构建 数据文化

数据管理

- **数据接入** 丰富的数据接入
- **数据采集** 在线填报 批量导入
- **自助ETL** 可视化的数据处理
- **数据模型** 多源的数据整合 OLAP多维建模
- **指标模型** 统一指标口径 提供指标自增长能力
- **高速缓存** 提高数据查询性能

- 安全管理
- 运维监控
- 集群管理
- 元数据管理
- 用户行为分析

SMARTBI
思迈特软件

02

自助ETL功能介绍及示例

自助ETL功能全景图

SMARTBI 思迈特软件

自助ETL

抽取（Extract）

- 数据源
  - 文本数据源
  - FTP/SFTP数据源
  - Kafka数据源
  - 关系数据源
  - Mongo数据源
  - 示例数据源
  - 数据集
  - 数据查询
  - Excel文件
  - 读取Excel sheet
  - API取数
  - ES数据源

转换（Transform）

- 数据处理
  - 行处理
  - 列处理
  - 组合查询
  - 多表JOIN
  - JOIN
  - 合并行
  - 元数据编辑
  - 聚合
  - 行转列
  - 列转行
  - 字符串处理
  - 正则表达式
  - JSON解析
  - XML解析
  - 循环API
- 行处理
  - 行过滤
  - 去除重复值
  - 排序
  - 增加序列号
  - 数据清洗
- 列处理
  - 列选择
  - 派生列
  - 拆分列
  - 合并列
  - 日期时间
  - 日期计算
- 脚本模块
  - Spark SQL脚本
  - PYTHON脚本

加载（Load）

- 目标源
  - 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(追加)
  - 导出数据到HDFS(覆盖)

作业流

- 通用
  - 开始
  - 检查依赖
  - 检查挖掘评估
  - 检查字段值
  - 参数输出
  - 循环器
  - Foreach循环器
  - 检查文件存在
  - 源库SQL脚本
  - Shell脚本

数据模型

- 存储过程查询
- ETL高级查询
- 生成日期表

配置和运维

1.数据挖掘配置

2.实验监控

3.服务监控

4.作业流监控

销售分析示例演示

SMARTBI 思迈特软件

示例场景：某企业为了提升销售数据分析效率与直观性，计划构建一个先进的数据可视化大屏系统。希望通过多样化的图表类型深入剖析销售数据，并能根据特定时点及年度至今（YTD）的时间维度灵活展示一个销售看板与销售明细信息。为实现这一目标，首要步骤是对既有数据进行预处理，确保数据质量与分析需求的高度匹配，为后续的可视化分析奠定坚实基础。

自助ETL示例：前期准备流程

➢ 第一步：

1、根据业务需求梳理分析，从业务目标或经营重点角度定义可度量单位以及业务口径；

➢ 第二步

2、根据业务流程确定数据来源表；

➢ 第三步

3、进行数据仓库分层架构的设计，包括事实表/维度表的设计；

自助ETL示例：数据仓库分层架构的设计

贴源层 ODS

- **数据采集：**从各种数据源（如业务库）中收集、识别和记录数据。
- **数据传输：**使用自助ETL工具将数据源中的数据传输至ODS层。在此过程中，几乎不对原始数据进行处理，保持数据的原始性和完整性。

明细层 DWD

- **数据清洗：**对ODS层的数据进行清洗，去除空数据、脏数据等，提高数据质量。
- **数据规范化：**对数据进行格式化处理，使其符合统一的规范和标准。
- **维度退化：**将维度退化至事实表中，减少事实表和维表的关联，提高数据的易用性。

服务层 DWS

- **数据聚合：**数据被进一步汇总和计算，形成多维数据分析模型。
- **宽表构建：**整合汇总成分析某一个主题域的数据服务层，一般是宽表。宽表化处理有助于提高数据查询和分析的效率。
- **数据服务化：**为后续的业务查询、OLAP分析、数据分发等提供统一的数据服务。

实操环境

实操环境链接：https://demo.smartbi.com.cn/smartbi/vision/index.jsp

用户名/密码：使用麦粉社区账号登录即可

资源存放目录：

在每个功能模块的根节点的“2025年麦粉社区线上直播资源-自助ETL最佳实践-实操资源保存目录”下新建一个自己姓名的目录，进行存放后续制作的资源

有奖竞答

SMARTBI 思迈特软件

规则：发布题目后，大家可以将答案在视频号的互动入口刷起来，最快答对的麦粉可以获得我们送出的礼品。

（一共2道题）

奖品：188麦豆

01 Q：以下哪项不属于自助ETL过程中的主要环节？

A、**数据抽取（Extract）**　　B、**数据转换（Transform）**

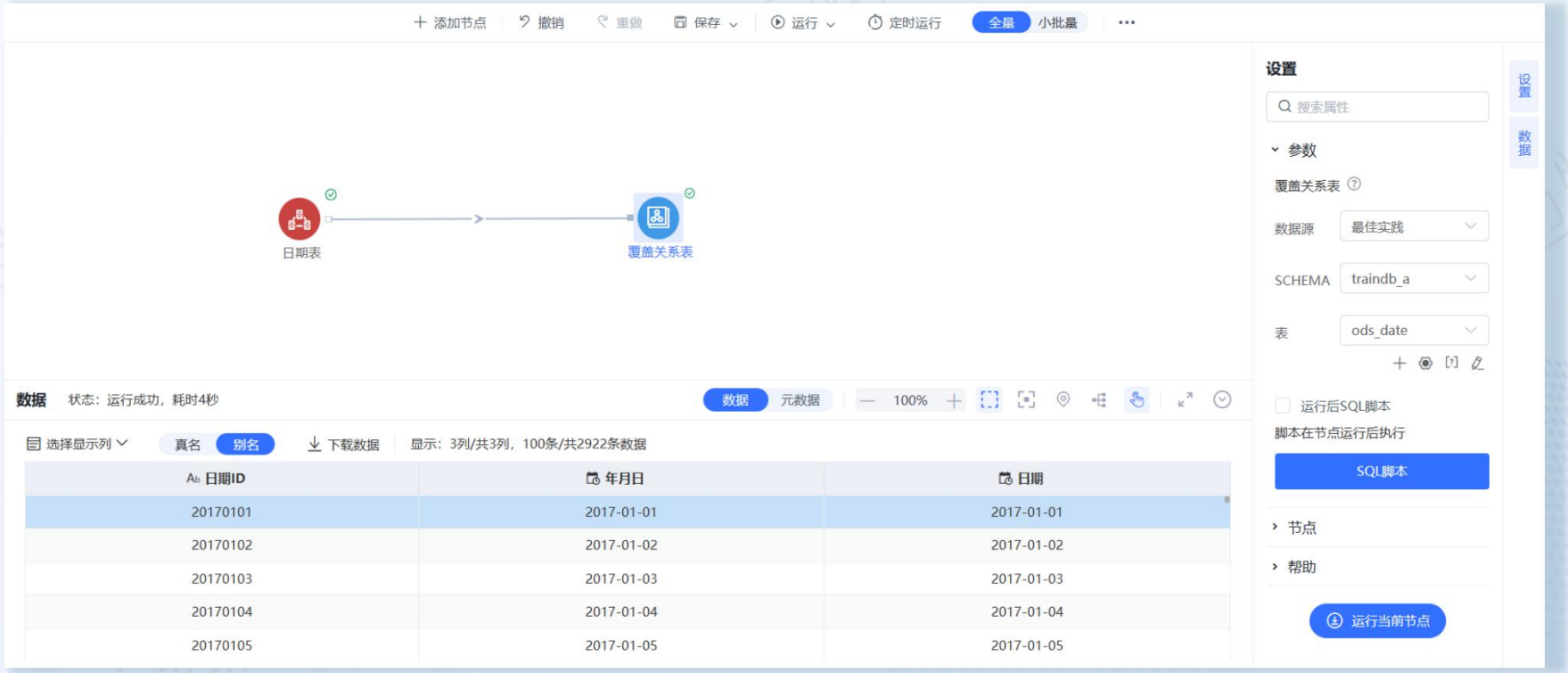
C、**数据加载（Load）**　　D、**数据备份（Backup）** √

02 Q：哪些不属于自助ETL的特点？

A、**可代替传统的SQL语句**　　B、**只需要使用简短的SQL语句** √

C、**可视化流程设计**　　D、**业务人员也可简单使用**

自助ETL示例演示：ODS层



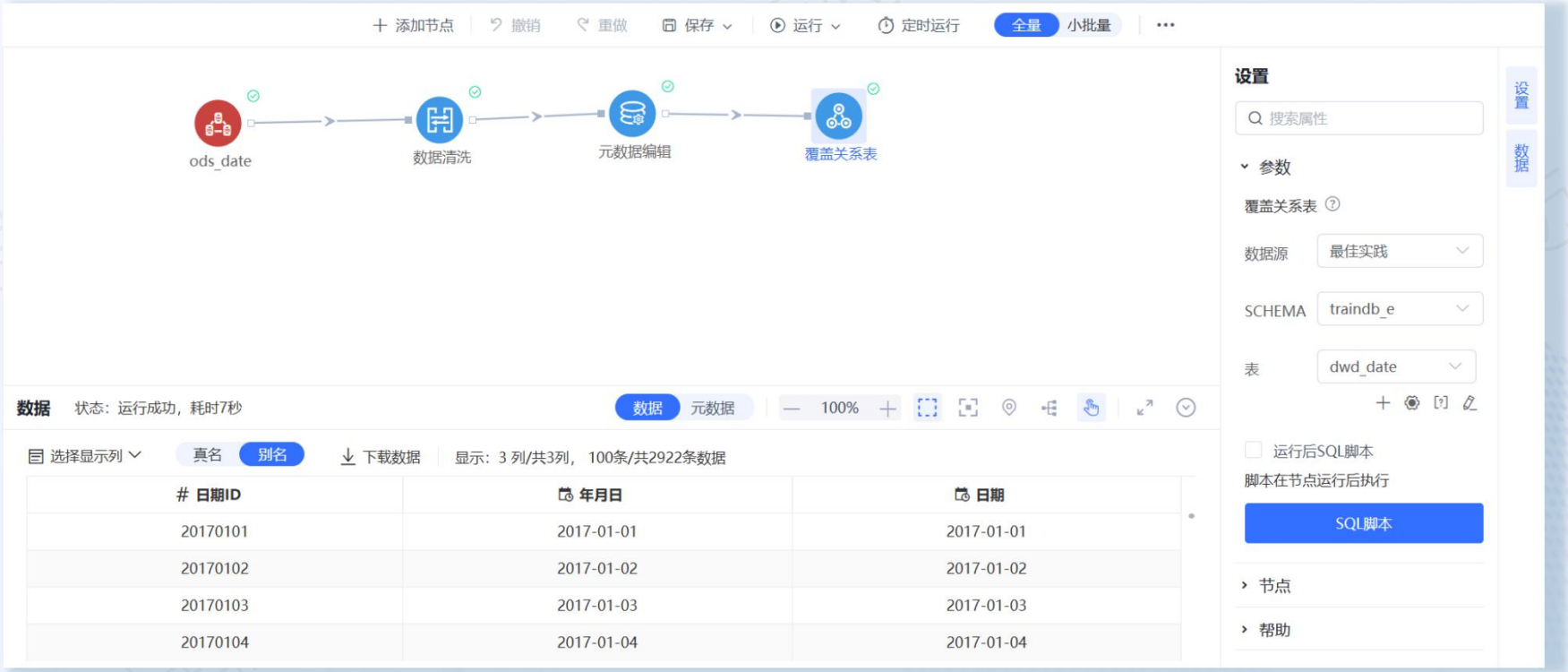
关系数据源

覆盖关系表

作用：

- 作为后续数据仓库层的准备区，为DWD层提供原始数据。
- 减少对业务系统的影响，避免在数据源层进行过多的数据操作。

自助ETL示例演示：DWD层



关系数据源

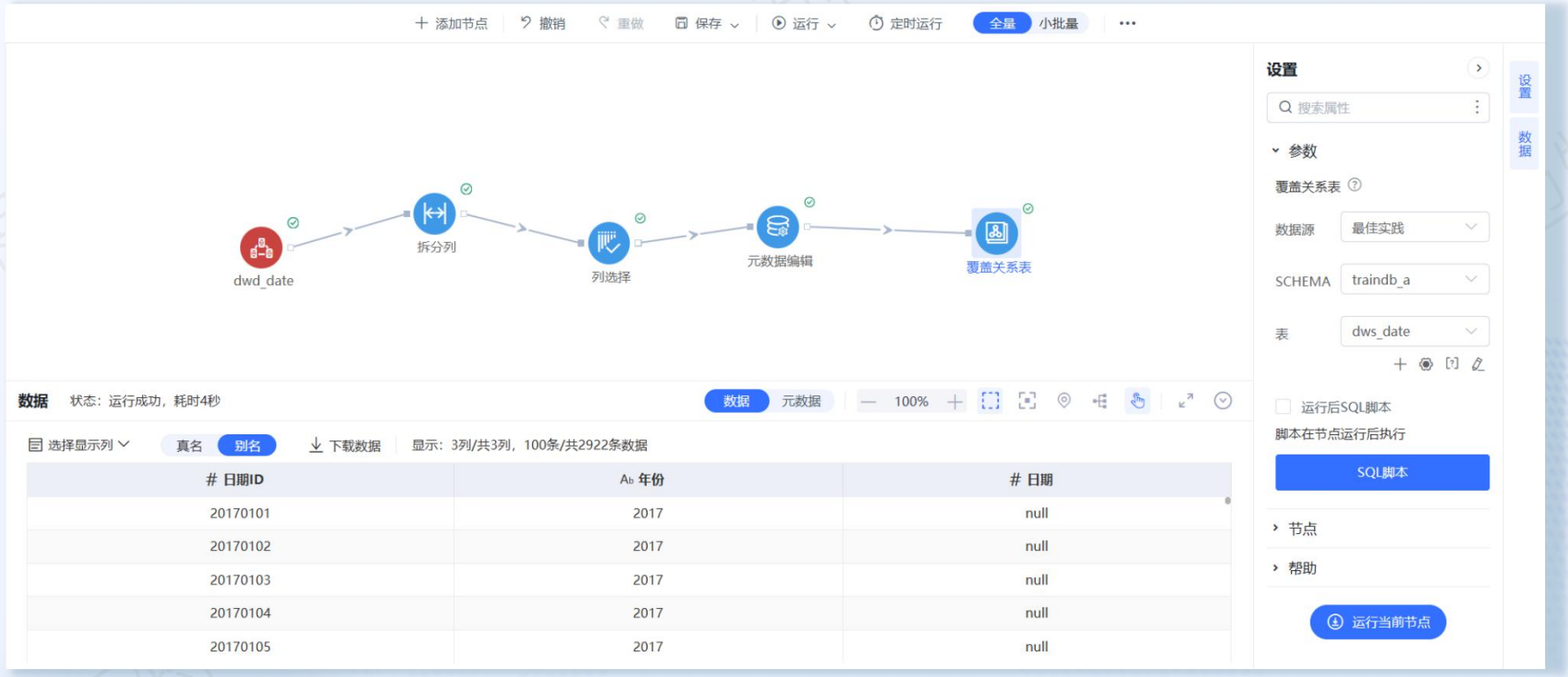
元数据编辑

覆盖关系表

作用：

- 作为业务层和数据仓库的隔离层，保持和ODS层一样的数据颗粒度。
- 提供清洗和规范化后的数据，为后续的数据处理和分析奠定基础。

自助ETL示例演示：DWS层



关系数据源

拆分列

列选择

元数据编辑

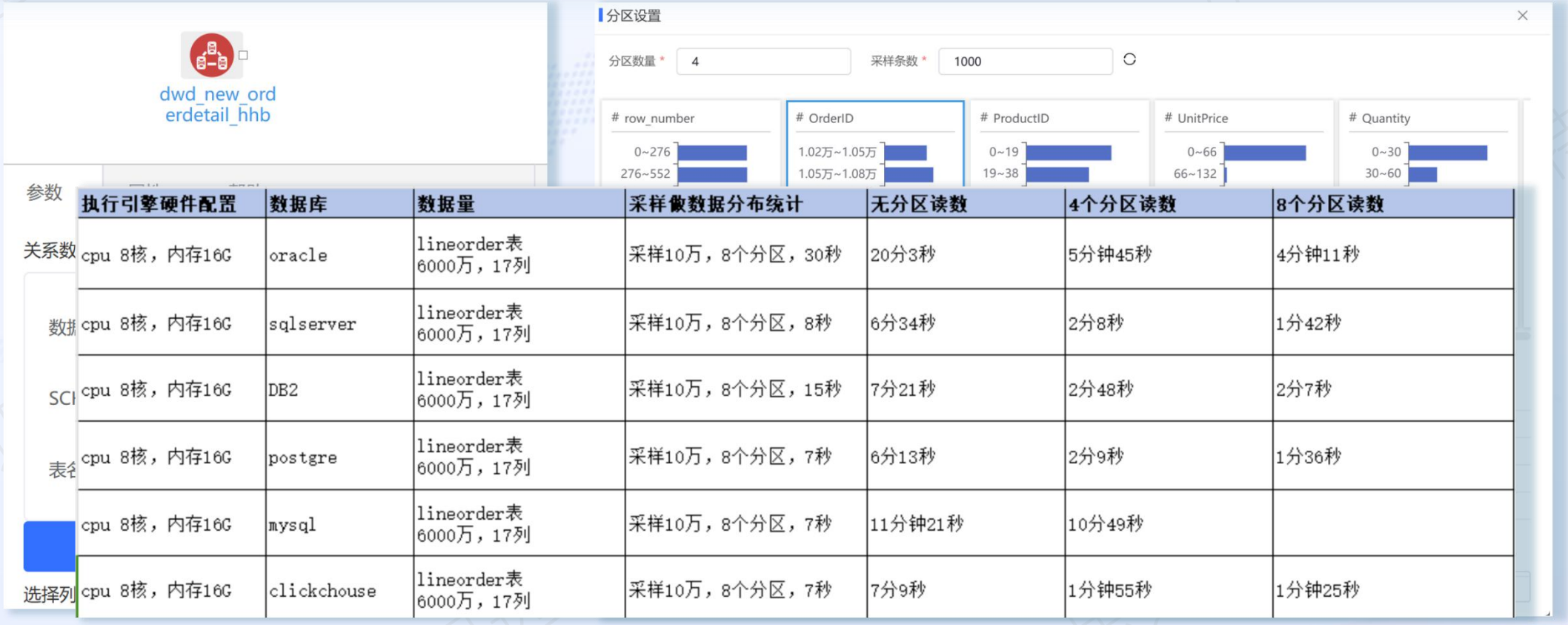
覆盖关系表

作用：

- 提供整合后的数据服务，满足业务查询和分析的需求。
- 通过宽表化处理，提高数据查询和分析的性能。

部分常用节点介绍：关系数据源

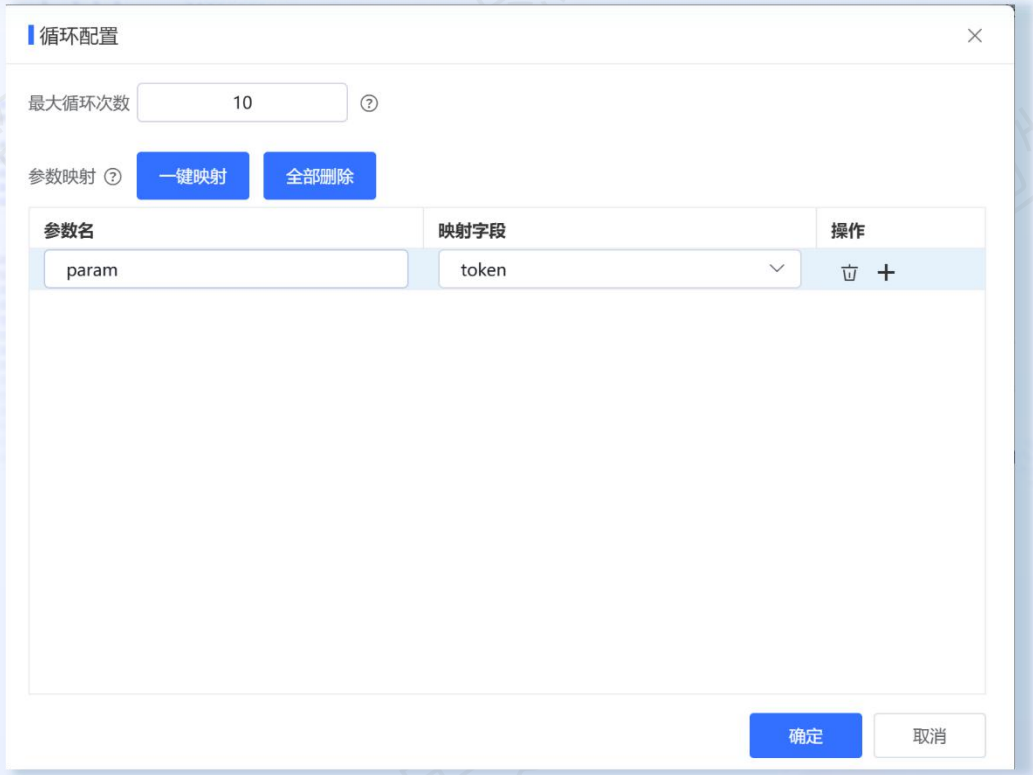
分区设置能显著提高关系数据源的抽取效率，自助ETL默认单线程抽取数据，通过分区设置，能够实现多线程抽取，显著加快抽取速度。例如Oracle数据库，1亿条数据单线程需要20分钟，8线程只需4分11秒，速度提升4.8倍。



执行引擎硬件配置	数据库	数据量	采样做数据分布统计	无分区读数	4个分区读数	8个分区读数
cpu 8核，内存16G	oracle	lineorder表 6000万，17列	采样10万，8个分区，30秒	20分3秒	5分钟45秒	4分钟11秒
cpu 8核，内存16G	sqlserver	lineorder表 6000万，17列	采样10万，8个分区，8秒	6分34秒	2分8秒	1分42秒
cpu 8核，内存16G	DB2	lineorder表 6000万，17列	采样10万，8个分区，15秒	7分21秒	2分48秒	2分7秒
cpu 8核，内存16G	postgre	lineorder表 6000万，17列	采样10万，8个分区，7秒	6分13秒	2分9秒	1分36秒
cpu 8核，内存16G	mysql	lineorder表 6000万，17列	采样10万，8个分区，7秒	11分钟21秒	10分49秒	
cpu 8核，内存16G	clickchouse	lineorder表 6000万，17列	采样10万，8个分区，7秒	7分9秒	1分钟55秒	1分钟25秒

部分常用节点介绍：循环API取数

从金蝶云/北森/钉钉/企业微信SaaS平台取数，一般分为两个步骤，先调一个api 获取登录token，然后再把token作为参数传给下一个api 进行取数；或者先由一个api获取列表数据，再根据这个列表数据循环api取数。此时需要用“循环API取数”节点。



部分常用节点介绍：派生列/spark脚本

派生列是个超级节点，借助Spark的函数，能够实现各种复杂的计算，例如滚动计算、移动计算、前期值、后期值等。更复杂的可以用Spark SQL脚本。

分析函数名(参数) over （子partition by 句 order by 字句 rows/range 字句)

1、分析函数名：sum、max、min、count、avg等聚合函数

lead、lag等比较函数

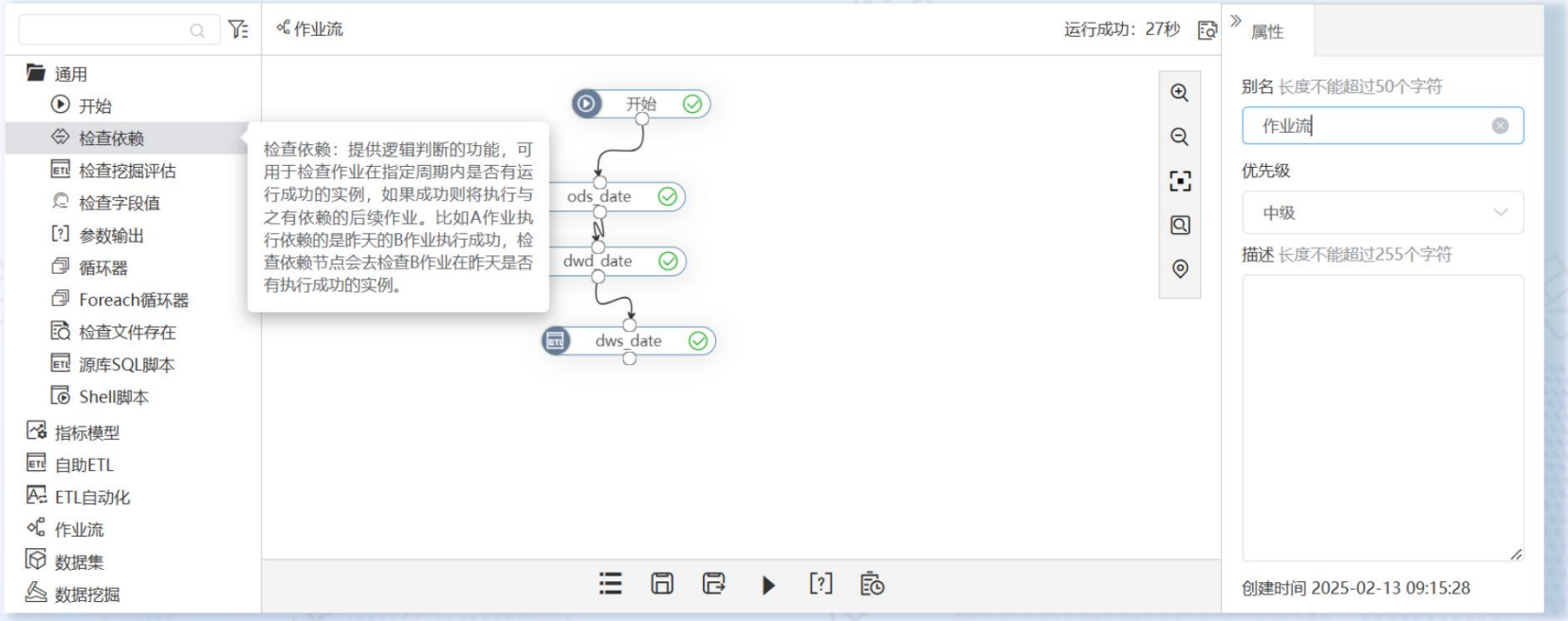
rank 等排名函数

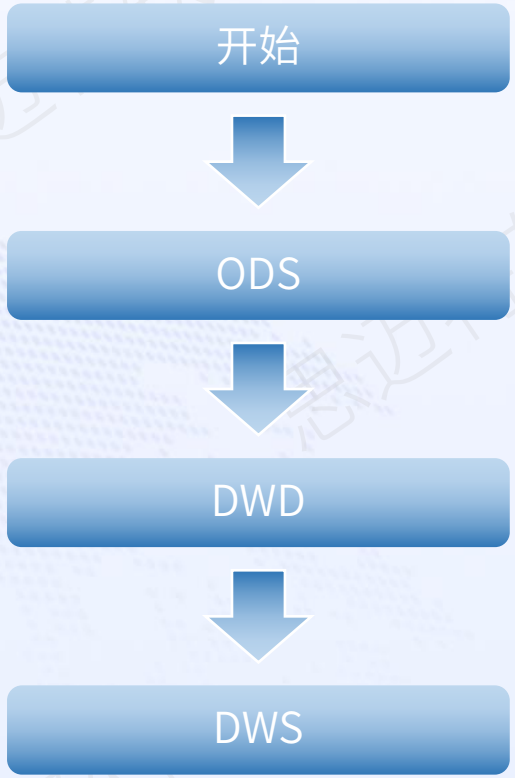
2、over：关键字，表示前面的函数是分析函数，不是普通的聚合函数

3、分析字句：over关键字后面括号内的内容为分析子句，包含以下三部分内容

- partition by ： 分组子句，表示分析函数的计算范围，各组之间互不相干
- order by：排序子句，表示分组后，组内的排序方式
- rows/range：窗口子句，是在分组(partition by)后，表示组内的子分组(也即窗口)

自助ETL示例演示：作业流



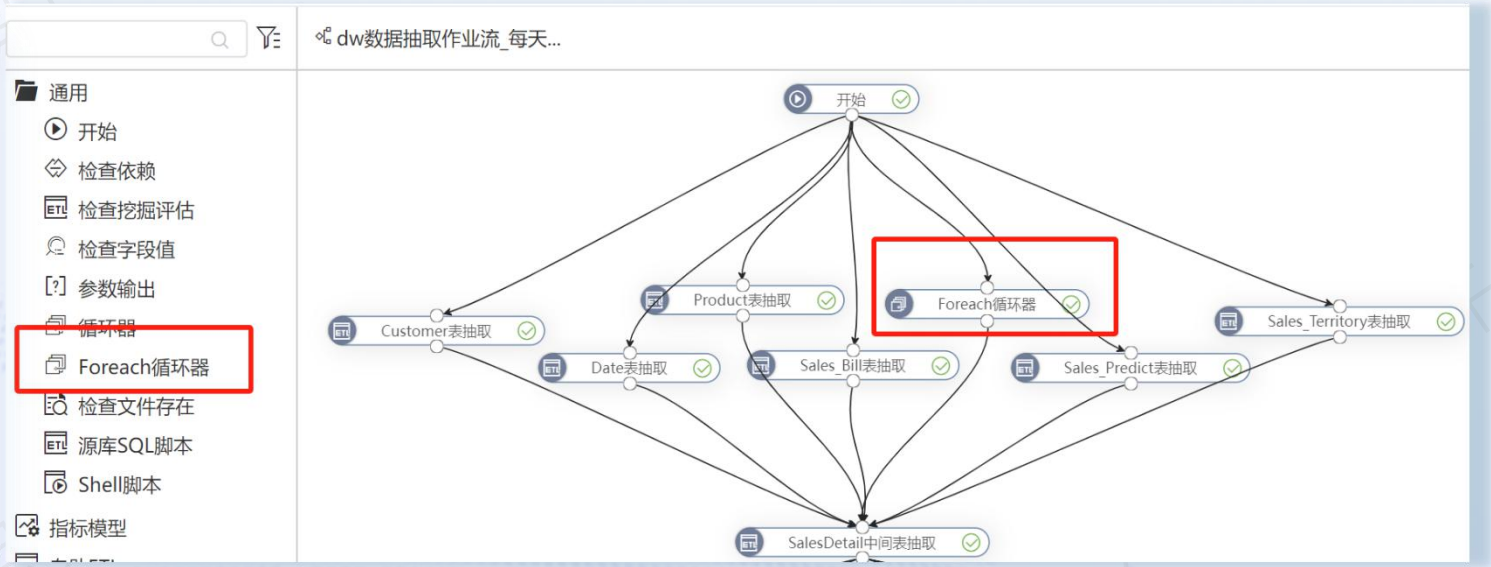


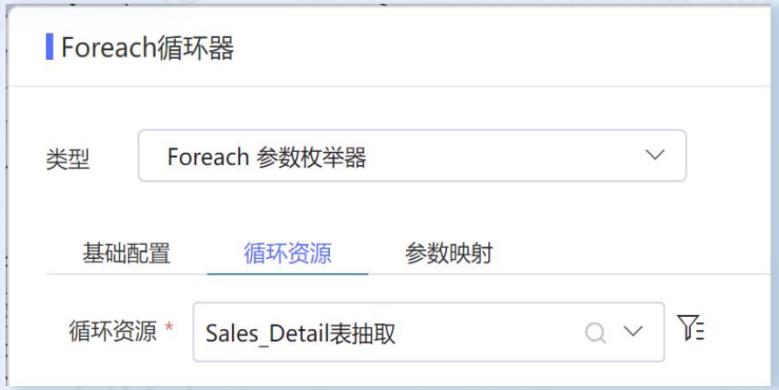
作用：

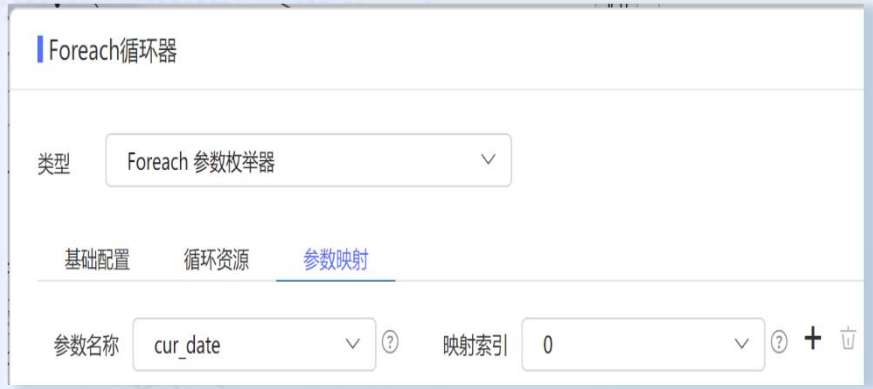
- 作业流可以自动运行ETL任务，减少人工干预。
- 支持按计划执行ETL任务，确保数据按时更新。
- 管理任务间的依赖关系，确保任务按正确顺序执行。
- 支持并行执行独立任务，提升效率。

部分常用节点介绍：Foreach循环器

- 在作业流中使用“Foreach循环器”，通过它把运维设置中配置的参数控件中的备选值逐个传入自助ETL，递归处理自助ETL数据。
- 增量、全量，只是换“Foreach循环器”中绑定的参数控件而已。







建模与仪表盘

使用处理好的数据进行数据建模

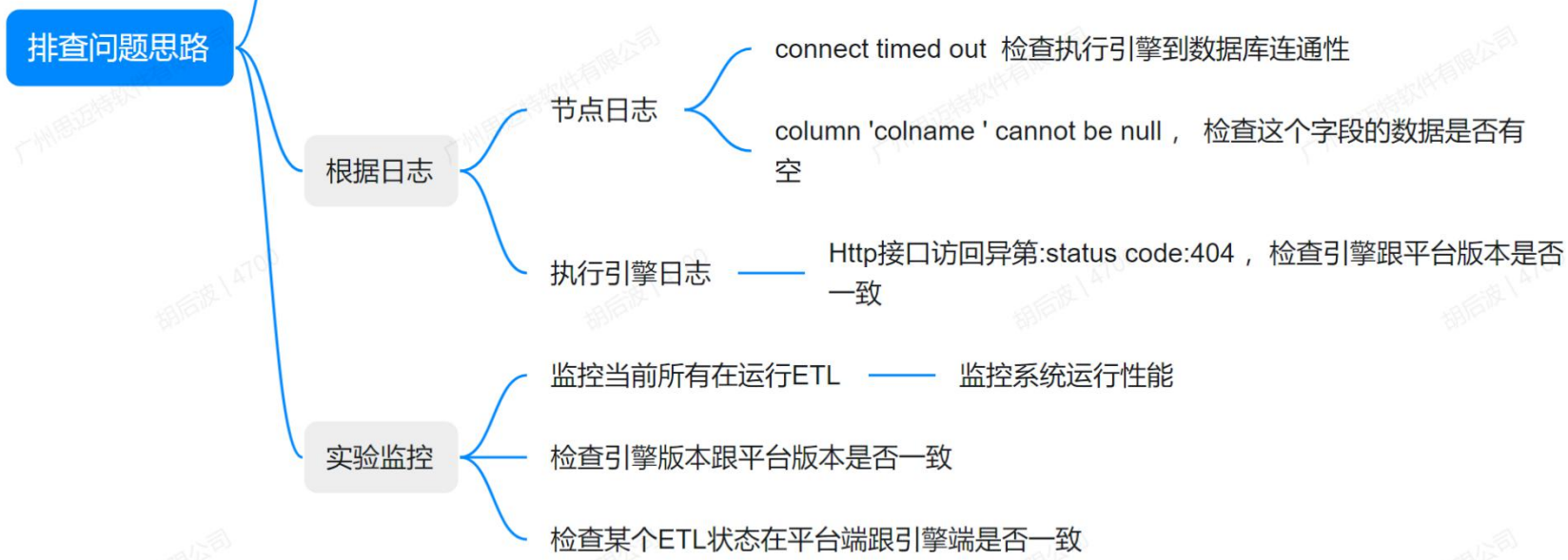
使用数据模型创建交互式仪表盘

SMARTBI
思迈特软件

03

自助ETL问题排查

自助ETL报错后的排查



自助ETL报错后的排查：问题提示

学会利用wiki文档帮助自己解决日常功能使用层面的问题以及报错、弹窗提示问题。wiki文档包含了**产品功能文档说明、常见问题FAQ文档。**

产品功能文档说明：详细介绍了每个模块、每个功能点要如何使用。

常见问题FAQ文档：根据用户咨询的问题整理出来的问题集和解决方法库。

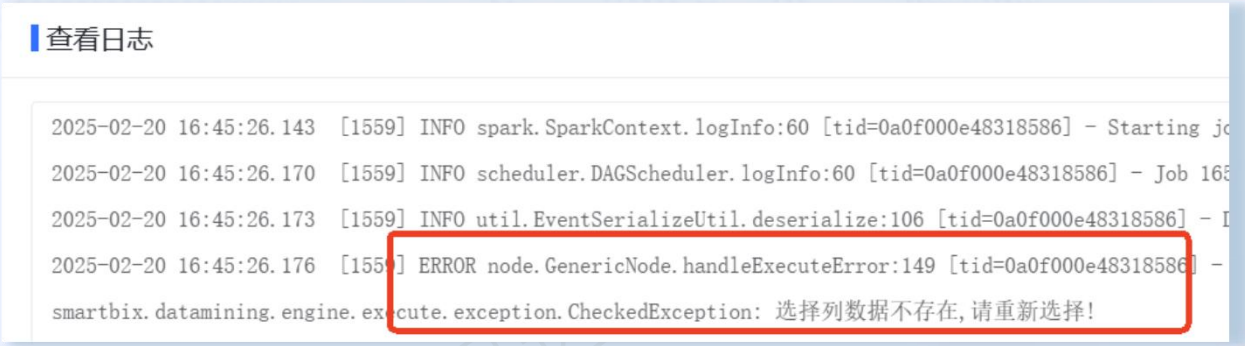
wiki链接：https://wiki.smartbi.com.cn/

错误提示

- 节点运行时，右上角出现红色感叹号，根据提示调整设置，或者复制相关提示信息，在wiki进行搜索。

节点日志

- 可以查看节点日志找到报错原因。根据提示调整设置，或者复制Caused by部分内容或者ERROR错误信息，在wiki上进行搜索。



自助ETL报错后的排查示例：问题提示

问题描述：客户在使用ETL数据处理后，进行存储时，提示：Data too long for column 'DESCRIPTION' at row 1

排查步骤：

1、查看提示信息

2、选择节点-->鼠标右键-->查看日志

3、查看Caused by部分内容或者ERROR错误信息

4、复制Caused by部分内容或者ERROR错误信息，在wiki进行搜索

```
at com.mysql.jdbc.Util.handleNewInstance(Util.java:420)
at com.mysql.jdbc.Util.getInstance(Util.java:408)
at com.mysql.jdbc.SQLError.createBatchUpdateException(SQLError.java:1154)
at com.mysql.jdbc.PreparedStatement.executeBatchSerially(PreparedStatement.java:1832)
at com.mysql.jdbc.PreparedStatement.executeBatchInternal(PreparedStatement.java:1316)
at com.mysql.jdbc.StatementImpl.executeBatch(StatementImpl.java:954)
at smartbix.datamining.engine.util.SparkFunctionUtil$5.call(SparkFunctionUtil.java:180)
at org.apache.spark.sql.Dataset.$anonfun$foreachPartition$2(Dataset.scala:2930)
at org.apache.spark.sql.Dataset.$anonfun$foreachPartition$2$adapted(Dataset.scala:2930)
at org.apache.spark.rdd.RDD.$anonfun$foreachPartition$2(RDD.scala:1020)
at org.apache.spark.rdd.RDD.$anonfun$foreachPartition$2$adapted(RDD.scala:1020)
at org.apache.spark.SparkContext.$anonfun$runJob$5(SparkContext.scala:2236)
at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:90)
at org.apache.spark.scheduler.Task.run(Task.scala:131)
at org.apache.spark.executor.Executor$TaskRunner.$anonfun$run$3(Executor.scala:498)
at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1439)
at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:501)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
Caused by: com.mysql.jdbc.MysqlDataTruncation: Data truncation: Data too long for column 'DESCRIPTION' at row 1
at com.mysql.jdbc.MysqlIO.checkErrorPacket(MysqlIO.java:3869)
at com.mysql.jdbc.MysqlIO.sendCommand(MysqlIO.java:2524)
at com.mysql.jdbc.MysqlIO.sqlQueryDirect(MysqlIO.java:2675)
at com.mysql.jdbc.ConnectionImpl.execSQL(ConnectionImpl.java:2465)
at com.mysql.jdbc.PreparedStatement.executeInternal(PreparedStatement.java:1912)
at com.mysql.jdbc.PreparedStatement.executeUpdateInternal(PreparedStatement.java:2133)
at com.mysql.jdbc.PreparedStatement.executeBatchSerially(PreparedStatement.java:1810)
... 16 more
```

自助ETL报错后的排查：系统监控

系统监控支持对系统网络，服务器等全面监控。便于用户优化系统参数，定位性能瓶颈。

1 监控当前所有在运行ETL

2 检查引擎版本跟平台版本是否一致

3 检查某个ETL状态在平台端跟引擎端是否一致

...更多

自助ETL报错后的排查示例：系统监控

问题描述：通过计划任务执行ETL任务时间过长或一直没有结束

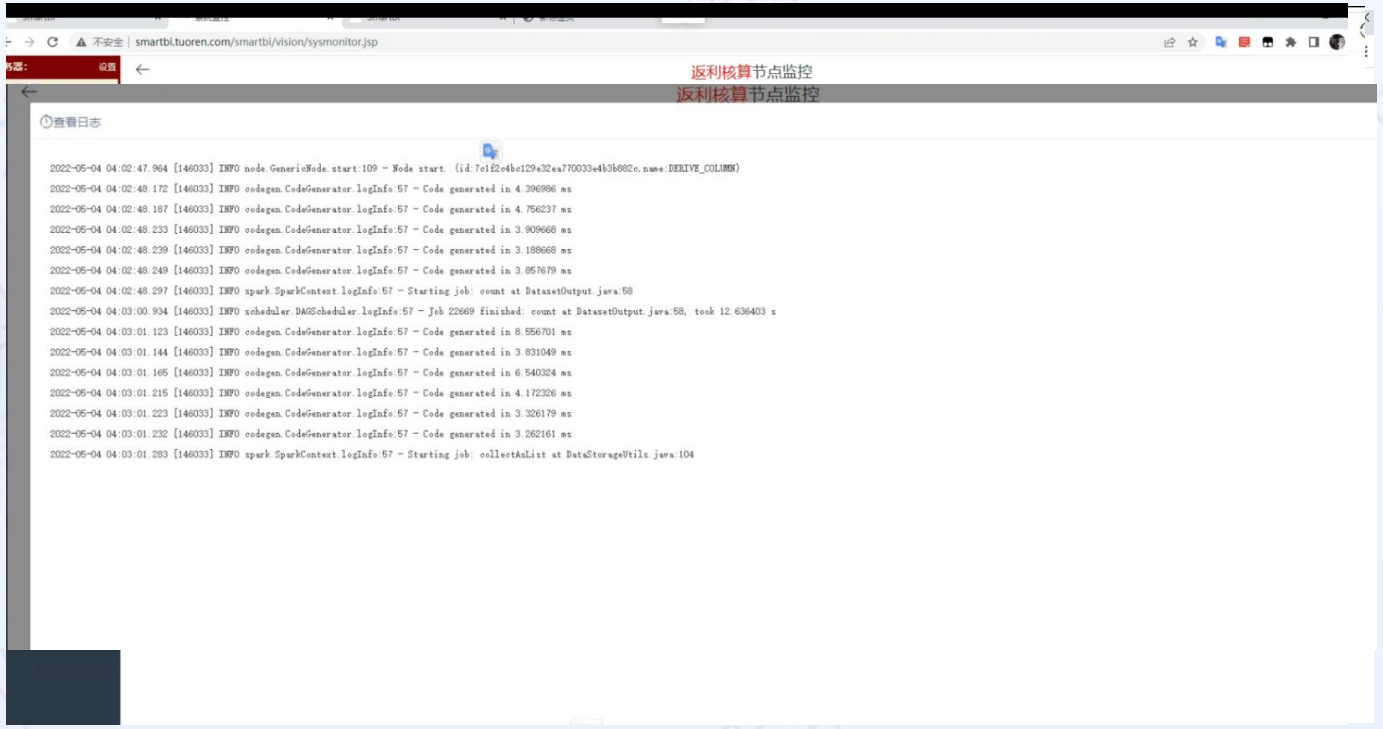
排查步骤：

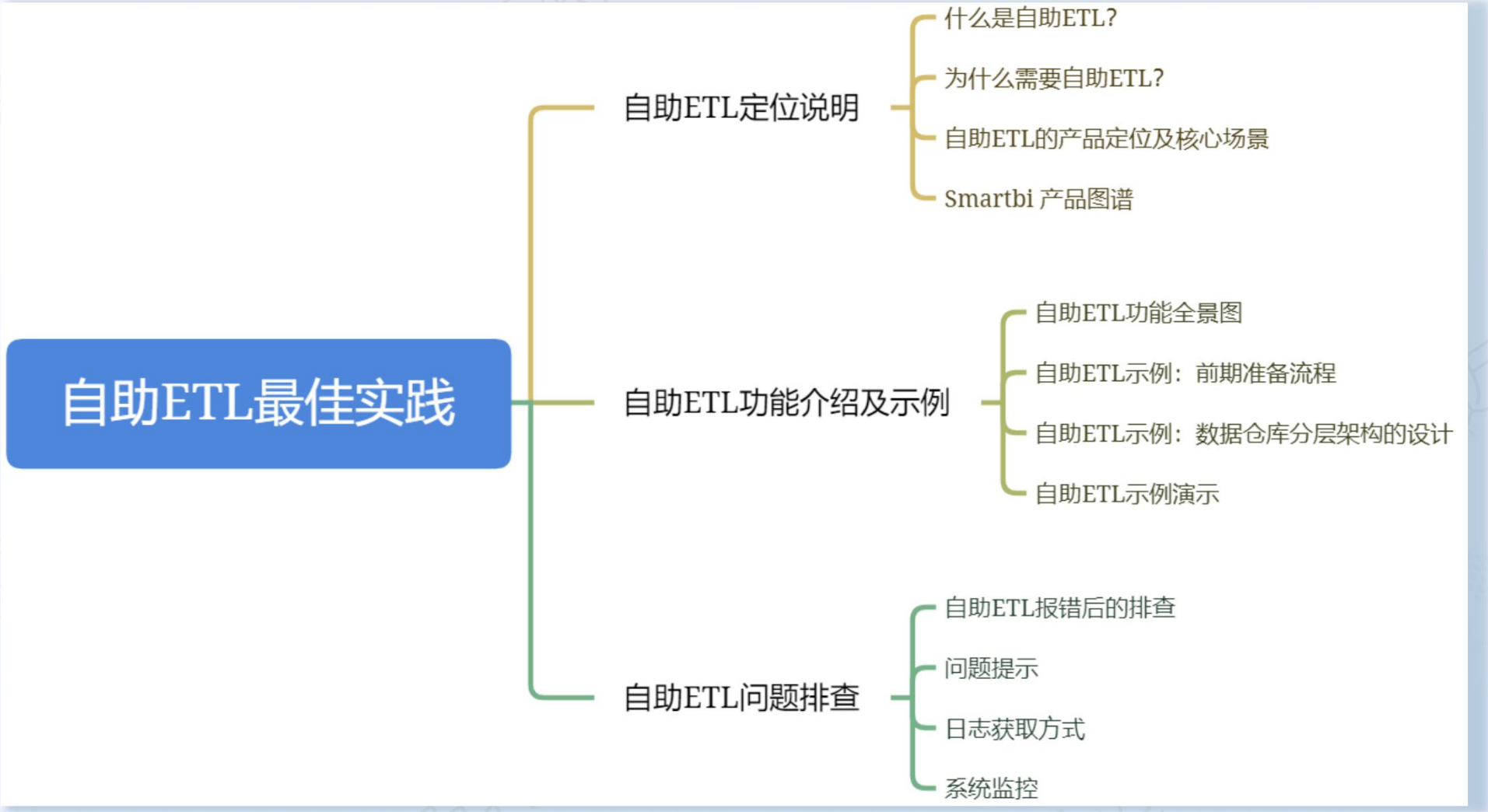
1、查看计算任务耗时分析

2、打开系统监控----实验监控-----详细内容

3、选择对应的实验名称----->点击节点列表

4、点击执行日志查看具体的节点运行日志




自助ETL最佳实践
自助ETL定位说明
什么是自助ETL?
为什么需要自助ETL?
自助ETL的产品定位及核心场景
Smartbi 产品图谱
自助ETL功能介绍及示例
自助ETL功能全景图
自助ETL示例：前期准备流程
自助ETL示例：数据仓库分层架构的设计
自助ETL示例演示
自助ETL问题排查
自助ETL报错后的排查
问题提示
日志获取方式
系统监控

实操环境

实操环境链接：https://demo.smartbi.com.cn/smartbi/vision/index.jsp

用户名/密码：使用麦粉社区账号登录即可

资源存放目录：

在每个功能模块的根节点的“2025年麦粉社区线上直播资源-自助ETL最佳实践-实操资源保存目录”下新建一个自己姓名的目录，进行存放后续制作的资源

课后答疑

问答版块答疑

可通过发帖进行提问，选择“其它”的分类，且在标题处加上【直播】的前缀

问答版块链接：https://my.smartbi.com.cn/wenda

多个任务活动持续上线

更多活动详情：

https://my.smartbi.com.cn/forum.php?mod=viewthread&tid=3183

新手任务	不限周期	1	288
日常任务	不限周期	不限次数	100
进阶任务	不限周期	不限次数	50-500
挑战任务	不限周期	不限次数	268-800
心得分享	不限周期	不限次数	10-100

问卷调查

SMARTBI 思迈特软件

填写调查问卷，get抽奖机会！

问卷二维码

奖品：188麦豆

麦粉社区 - 认证 - 个人用户认证体系

公众号

小麦微信

https://my.smartbi.com.cn/plugin.php?id=exam

SMARTBI
思迈特软件

AI上与数据聊天

www.smartbi.com.cn