

网易严选指标异常检测与 诊断分析实践

李隆熙-网易-资深数据挖掘工程师

DataFunSummit 2023



目录 Contents

01

背景介绍

02

指标异常检测

03

指标异常诊断

04

指标问答机器人

01

背景介绍

DataFunSummit 2023



为何要做指标异常检测和诊断

息息相关

公司财务



GMV/成本/毛利

库存计划



库存/周转/缺货

用户运营



拉新/留存/复购

价值所在



1、发现问题

- 🚩 定位问题
- 🚩 分析问题
- 🚩 解决问题



2、发现亮点

- 🏆 定位原因
- 🏆 分析原因
- 🏆 推广亮点

背景介绍

目标与挑战

挑 战

指标 数量多
指标 差异大
指标 变化快
定位 不精准
发现 有滞后

实 现

目 标

自动化：无需依赖用户输入
通用性：适应多种指标分布
时效性：天&小时检测归因
准确性：严格统计检验计算
主动性：群机器人主动播报

流程设计



02

指标异常检测

DataFunSummit 2023



指标异常检测 | 问题定义



网易严选 | DataFun.

□ 异常有哪些?

绝对值异常

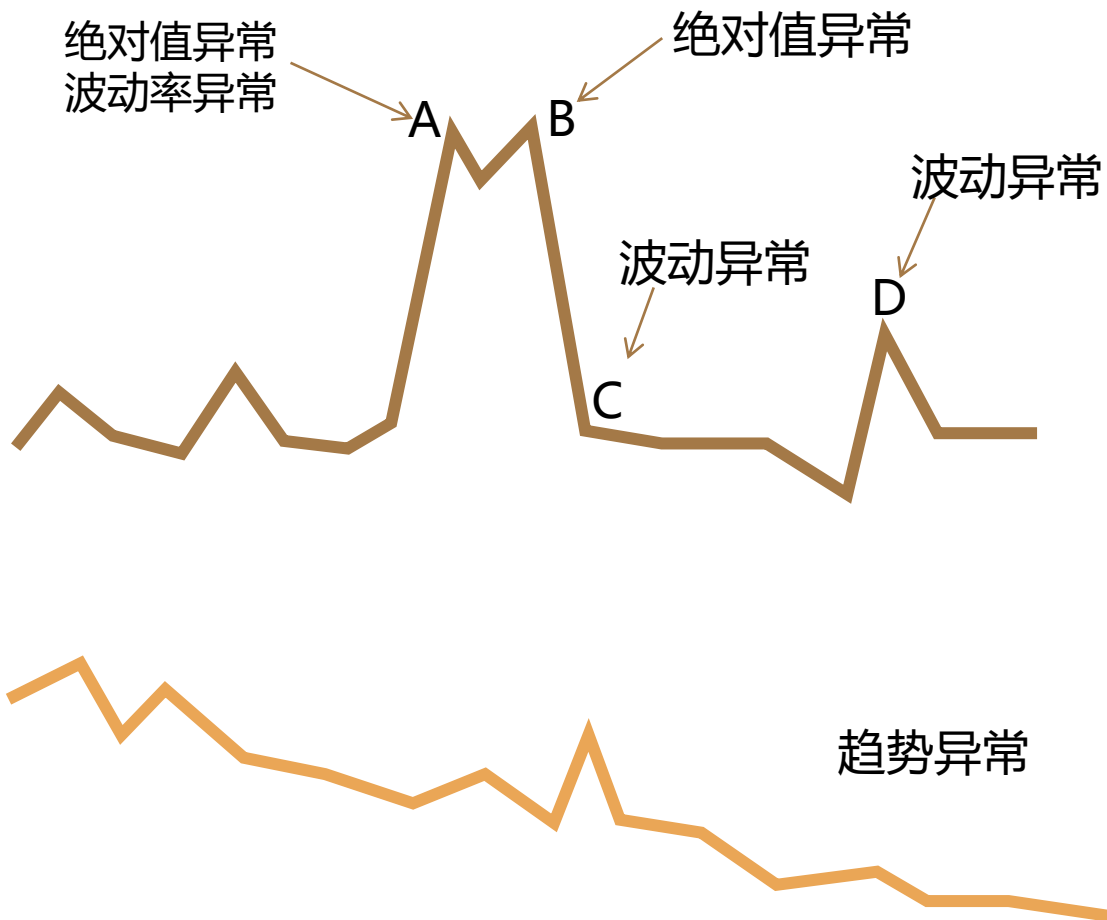
不遵循指标分布，统计学上的离群点
反应业务状态

波动异常

环比过大的突增点或者突降点
反应业务变化

趋势异常

中长期呈现出确定性上升或下降趋势
预示潜在风险



$$R_i = \frac{\max_i |x_i - \bar{x}|}{s} \quad i = 1, 2, \dots, r$$

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{(n-i-1+t_{p,n-i-1}^2)(n-i+1)}}$$

$$p = 1 - \frac{\alpha}{2(n-i+1)}$$

outlier = where($R_i > \lambda_i$)

1-绝对值异常检测

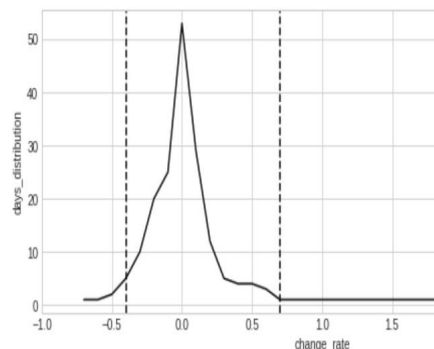
主要基于GESD检验算法

原理:

计算广义极端学生化偏差统计量寻找异常点

适用指标分布:

正态分布



2-波动异常检测

主要基于波动率分布拐点阈值

原理:

基于二阶导数和距离寻找MBP(最大弯曲点)为拐点

适用指标分布:

拐点存在的正态、非正态分布

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(X_j - X_i)$$

$$\text{sgn}(X_j - X_i) = \begin{cases} 1 & (X_j - X_i) > 0 \\ 0 & (X_j - X_i) = 0 \\ -1 & (X_j - X_i) < 0 \end{cases}$$

$$Z = \begin{cases} \frac{S-1}{\sqrt{\text{Var}(S)}} & S > 0 \\ 0 & S = 0 \\ \frac{S+1}{\sqrt{\text{Var}(S)}} & S < 0 \end{cases}$$

3-趋势异常检测

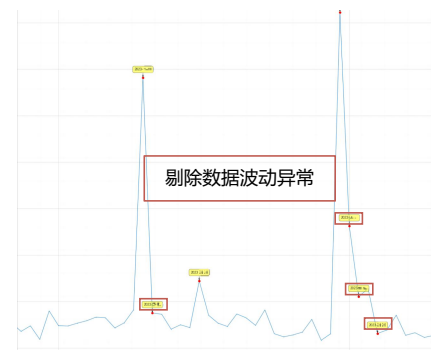
基于Man-Kendall检验

原理:

计算指标序列M-K统计量, 基于p值判断异常趋势

适用指标分布:

所有分布



4-后处理

减少不必要预警, 降低打扰

数据异常:

上周期异常导致的本周期波动异常, 基于规则剔除

S级大促:

基于数据协同剔除

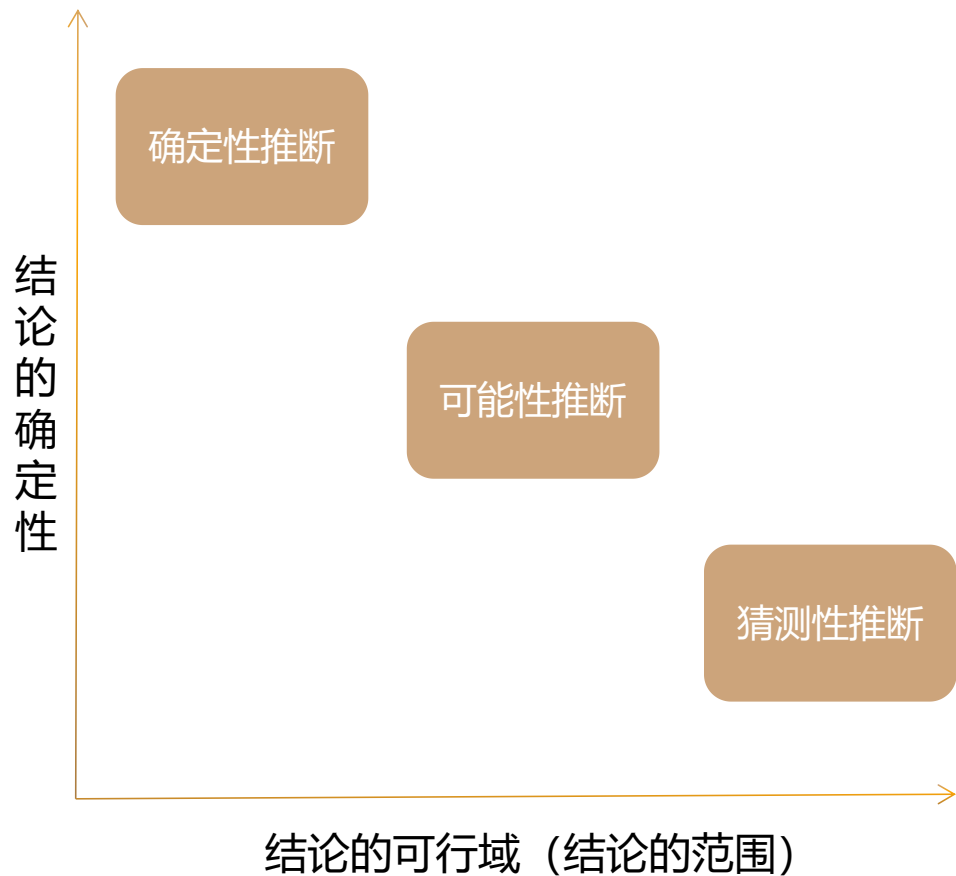
03

指标异常诊断

DataFunSummit 2023



推断的三个层次



确定性推断:

- 🔖 GMV上涨30%，各类商品产生多大的贡献？
- 🔖 新客户、老客户各起到了怎样的作用？

可能性推断:

- 🔖 DAU下降30%，跟用户满意度下降有多大关系？

猜测性推断:

- 🔖 GMV下降30%，因为政策变化、竞品活动还是天气？

可能性推断

01

机器学习拟合

- **原理：**面向指标数据做回归预测，计算特征重要性
- **缺点：**不能解释单次异常

02

机器学习 + SHAP

- **原理：**用shap value计算特征对目标的贡献值
- **缺点：**不够精准，仅得出相关性非因果性

03

贝叶斯网络

- **原理：**构建贝叶斯网络，衡量指标关系
- **缺点：**计算复杂、黑盒

确定性推断

01

拆解贡献度

加法

GMV

=

GMV_1

+

GMV_2

+

...

乘法

GMV

=

流量

X

转化率

X

客单价

除法

毛利率

=

毛利_1

+

毛利_2

+

...

GMV_1

+

GMV_2

+

...

原理

指标拆解下钻，
计算各部分贡献度

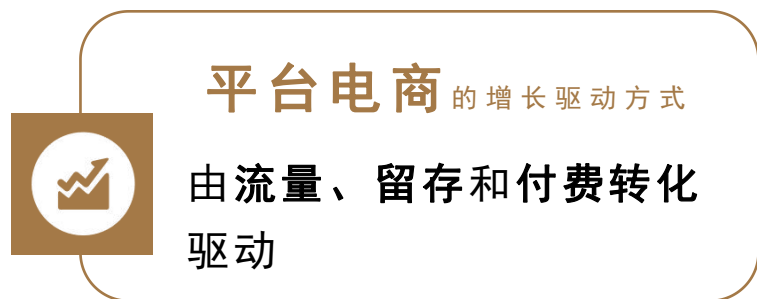
缺点

拆解维度多样，
组合维度爆炸

指标异常诊断 | 业务现状

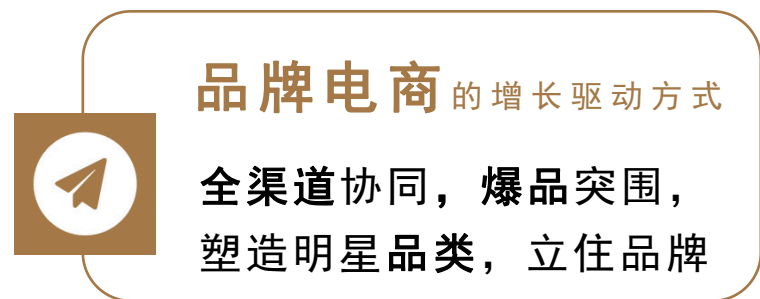
网易严选

DataFun.



网易严选

战略升级



品牌电商的
指标分级



拆解



加法拆解

$$Y = \sum_i X_i$$
$$\Delta Y\% = \frac{Y^1 - Y^0}{Y^0}$$
$$C_{X_i} = \frac{X_i^1 - X_i^0}{Y^0} = \frac{\Delta X_i}{Y^0}$$

乘法LMDI拆解

$$Y = \prod_i X_i$$
$$\Delta Y\% = \frac{Y^1 - Y^0}{Y^0}$$
$$L(Y^1, Y^0) = \frac{Y^1 - Y^0}{\ln(Y^1) - \ln(Y^0)}$$
$$C_{X_i} = \frac{L(Y^1, Y^0) \cdot \ln(\frac{X_i^1}{X_i^0})}{Y^0}$$

除法双因素拆解

$$Y = \frac{S}{P} = \frac{\sum_i s_{X_i}}{\sum_i p_{X_i}},$$
$$P_i = \frac{p_{X_i}}{\sum_i p_{X_i}},$$
$$S_i = \frac{s_{X_i}}{\sum_i s_{X_i}},$$
$$Y_i = \frac{s_{X_i}}{p_{X_i}}$$

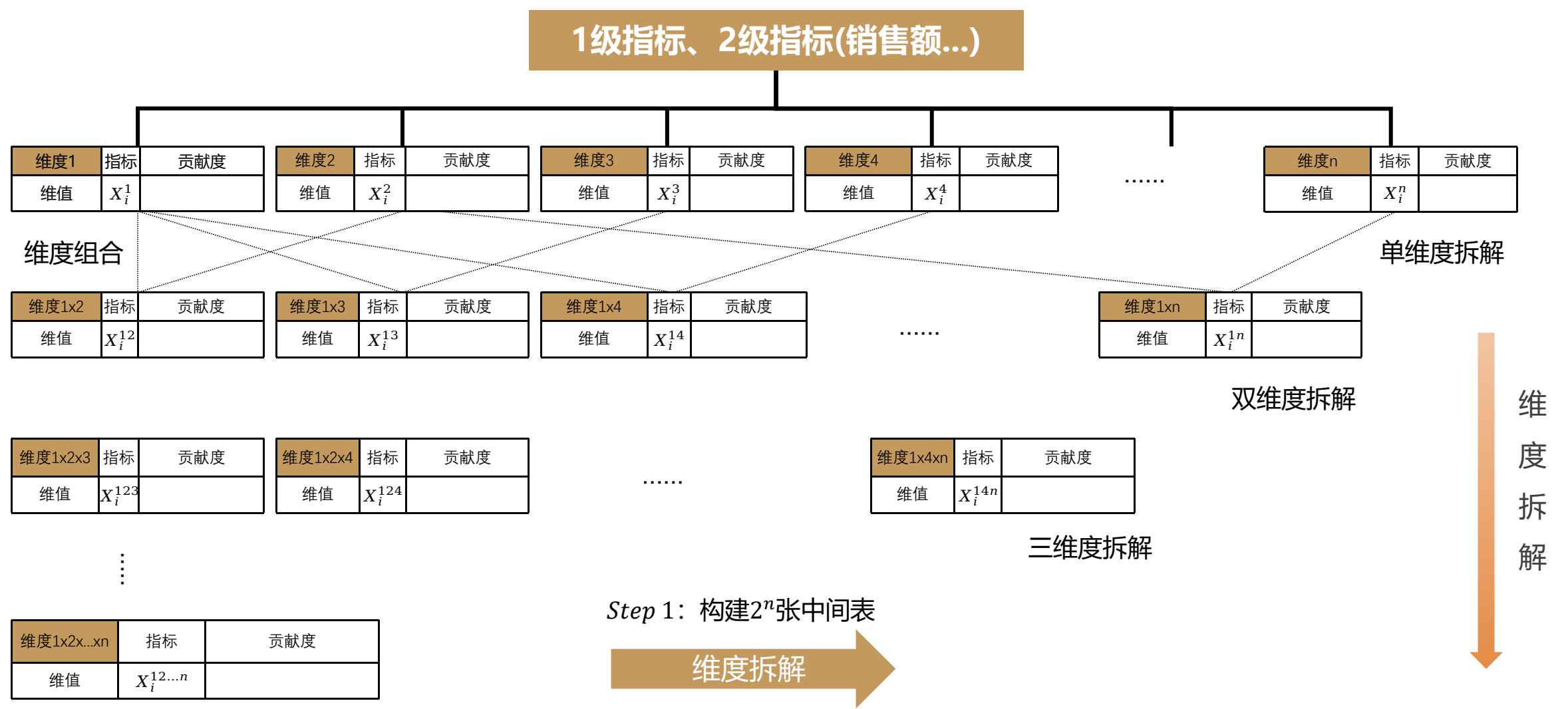
A: 波动贡献
B: 结构变化贡献

$$A_{X_i} = (Y_i^1 - Y_i^0) \times P_i^0$$
$$B_{X_i} = (P_i^1 - P_i^0) \times (Y_i^1 - Y^0)$$
$$C_{X_i} = \frac{A_{X_i} + B_{X_i}}{Y^0}$$



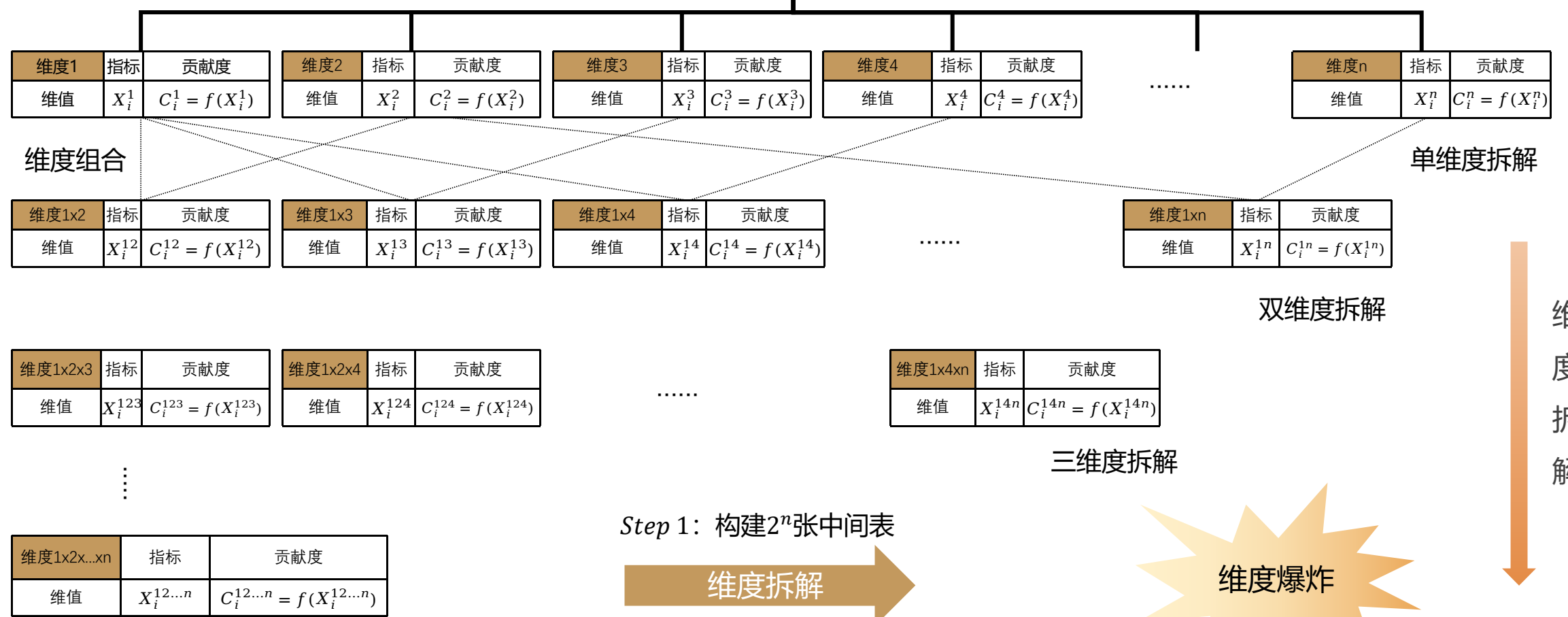
贡献度具有可加性，满足**MECE**原则
整体变化等于各部分变化贡献之和

指标异常诊断 | 拆解贡献度



指标异常诊断 | 拆解贡献度

1级指标、2级指标(销售额...)



维度拆解

指标异常诊断 | 拆解贡献度

维度1	指标	贡献度	维度2	指标	贡献度	维度3	指标	贡献度	维度4	指标	贡献度	维度n	指标	贡献度
维值	X_i^1	$C_i^1 = sum^1(C_i^{12...n})$	维值	X_i^2	$C_i^1 = sum^2(C_i^{12...n})$	维值	X_i^3	$C_i^1 = sum^3(C_i^{12...n})$	维值	X_i^4	$C_i^1 = sum^4(C_i^{12...n})$		维值	X_i^n	$C_i^1 = sum^n(C_i^{12...n})$

单维度聚合

维度1x2	指标	贡献度	维度1x3	指标	贡献度	维度1x4	指标	贡献度	维度1xn	指标	贡献度
维值	X_i^{12}	$C_i^{12} = sum^{12}(C_i^{12...n})$	维值	X_i^{13}	$C_i^{13} = sum^{13}(C_i^{12...n})$	维值	X_i^{14}	$C_i^{14} = sum^{14}(C_i^{12...n})$		维值	X_i^{1n}	$C_i^{1n} = sum^{1n}(C_i^{12...n})$

双维度聚合

维度1x2x3	指标	贡献度	维度1x2x4	指标	贡献度	维度1x4xn	指标	贡献度
维值	X_i^{123}	$C_i^{123} = sum^{123}(C_i^{12...n})$	维值	X_i^{124}	$C_i^{124} = sum^{124}(C_i^{12...n})$		维值	X_i^{14n}	$C_i^{14n} = sum^{14n}(C_i^{14...n})$

三维度聚合

维度1x2x4xn	指标	贡献度
维值	X_i^{124n}	$C_i^{124n} = sum^{124}(C_i^{124...n})$

四维度聚合

贡献度聚合

求和

维度1x2x...xn	指标	贡献度
维值	$X_i^{12...n}$	$C_i^{12...n} = f(X_i^{12...n})$

末级指标(销售额...)

Step 1: 计算最细粒度贡献度

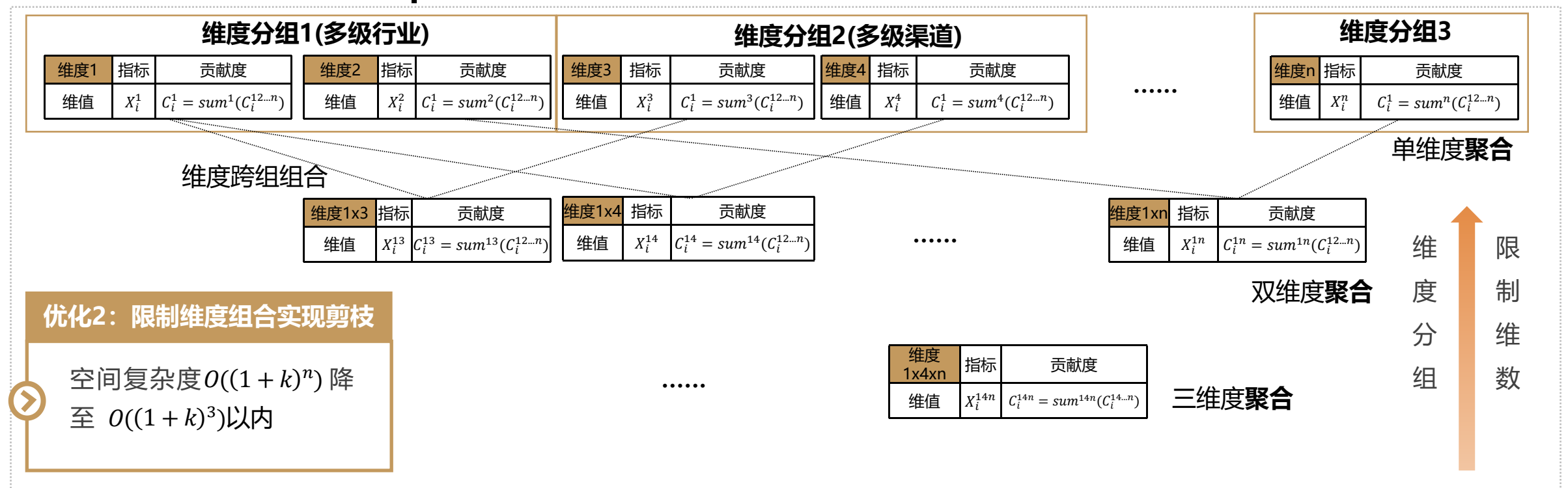
基于贡献度可加

Step 2: 各维度贡献度聚合

优化1: 维度拆解->贡献度聚合

省去中间表的I/O过程
仅需1次算法调用

指标异常诊断 | 拆解贡献度



$Gini(D) = 1 - \sum C_i^2$

优化3: 基于维度基尼系数排序

确定最佳诊断维度
实现异常精准定位

举例 →

指标	拆解维度	维值	贡献度	基尼系数	拆解维度	维值	贡献度	基尼系数
销量	商品	item_1	8%	0.99	二级渠道	渠道1	50%	0.65
		item_2	5%			渠道2	30%	
		item_3	3%			
		item_4	1%			渠道10	10%	
		item_5	1%					
		item_6	0%		一级行业 x 二级渠道	行业1 x 渠道1	60%	0.54
				行业2 x 渠道2	30%	
		item_100	-1%			
						行业8 x 渠道1	-10%	

指标异常诊断 | 效果展示

主动式播报

2023-05-29 指标绝对值播报

指标	数值	参考正常区间	诊断维度	主要原因
毛利额	-3 万	[万, 万]	二级渠道x商品	-7 万 (贡献度:110.7%)

文本解读

备注: 播报指标绝对值, 说明当前指标过大或过小

毛利额 昨日为-3 万, 参考正常区间[万, 万]。从二级渠道x商品维度来看, 主要原因为 -7 万, 贡献度为 110.7%

查看指标趋势

天级别负毛利异常

2023-08-24 15时 指标波动率播报

指标	日环比	参考正常区间	诊断维度	主要原因
销量	%↑	[- %, %]	四级渠道x行业	银行积分兑换x居家部 (贡献度:63.7%)

文本解读

备注: 播报指标波动率, 说明当前指标日环比(波动)过大

销量 环比昨日 %, 参考正常区间[- %, %]。从四级渠道x行业维度来看, 主要原因为 银行积分兑换x居家部 销量, 贡献度为 63.7%

查看指标趋势

小时级销量波动

问题与不足

查指标:

在2023年5月29号, 天猫网易严选旗舰店的销售额是多少?——异常时非主因

查贡献:

在2023年5月29号, 食品部对毛利额变化的影响是多少?——异常时非主因

查原因:

在2023年8月19号, 销售额变化的主要原因是?——指标未异常

定位

1. 播报仅在**异常**发生时提供**主要原因**
2. 但**95%**时间无指标异常

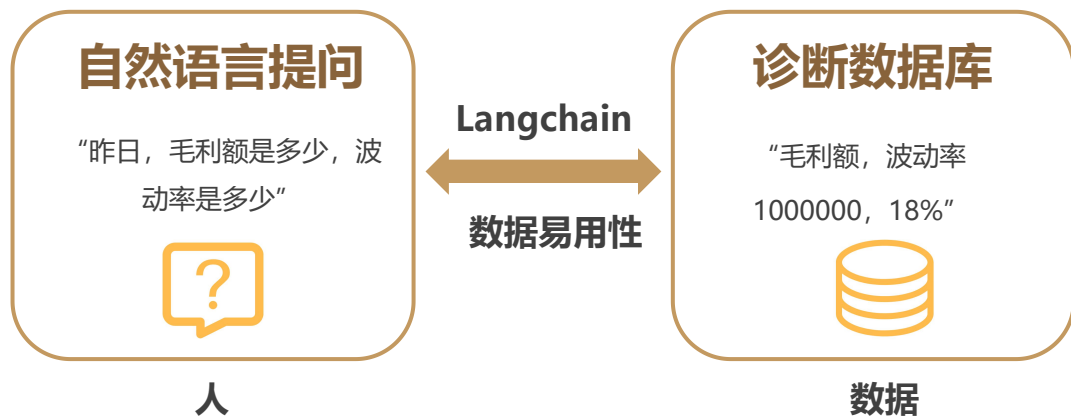


04


指标问答机器人

DataFunSummit 2023

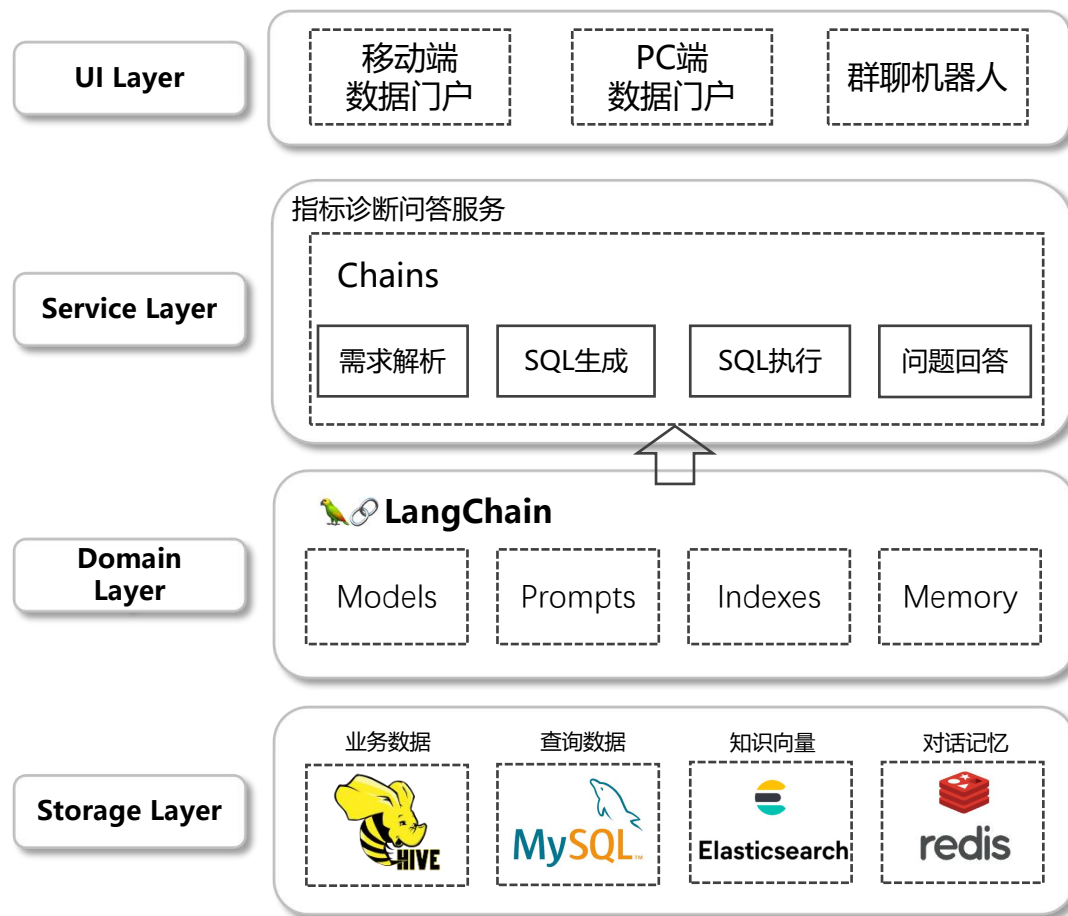
指标问答机器人



定位

 快速实现常用指标查询与诊断功能

 作为主动式播报形式的内容补充



挑战

输入的非标准化

影响

结果的有效性

在2023年8月19号，食品部1对销售额波动的影响是？

```
SQLQuery:SELECT sub_dimension_contribution
FROM metric_volatility_en
WHERE date = date('2023-08-19')
AND anomaly_indicator = '销售额'
AND sub_dimension_name = '食品部1'
;
```

SQLResult: []

Answer:结果为空，因为没有名为“食品部1”的项。

Case 1.维值不存在

食品部对销售额波动的影响是多少？

```
SQLQuery:SELECT indicator_fluctuation_rate * sub_dimension_contribution
FROM metric_volatility_en
WHERE anomaly_indicator = '销售额'
AND sub_dimension_name = '食品部'
;
```

SQLResult: [(0.005827831048926929,)]

Answer:食品部对毛利率波动的影响是0.005827831048926929。

Case 2.缺少必填项

需求确认模块

**你是一名电商行业的数据科学专家，能够根据用户的输入，解析用户的意图，填充需求参数，并针对缺失的参数进行提问补全。

**你必须使用下面的function来响应用户的意图：

- 1.查询指标数据："get_indicator", args: "indicator_name": "指标名称", "query_date": "查询时间", "query_dim": "查询维度".
- 2.查询指标贡献度或影响度："get_contribution", args: "indicator_name": "指标名称", "query_date": "查询时间", "query_dim": "诊断维度".
- 3.查询指标变化的诊断原因："get_reason", args: "indicator_name": "指标名称", "query_date": "查询时间".
- 4.无法执行前3个命令："reply_info", args: "input": "用户的原始输入".

**你给function输入的args只能为以下几种格式：

- 1.indicator_name: 需求必要参数，参数取值为{indicator_list}，如果问题中没有提及，设为空.
- 2.query_date: 需求必要参数，具体的年月日，值用2023-08-19这种类型表示，当前时间为{cur_date}，可作为时间锚点，推算用户的时间，如果问题中没有提及，设为空.
- 3.query_dim: 参数取值为{dim_list}，如果问题中没有提及，设为空.

**你所有的回答必须是JSON格式，返回如下所述响应格式：

```
{
  "function": {
    "name": "function name",
    "args": {
      "arg name": "value"
    }
  }
}
```

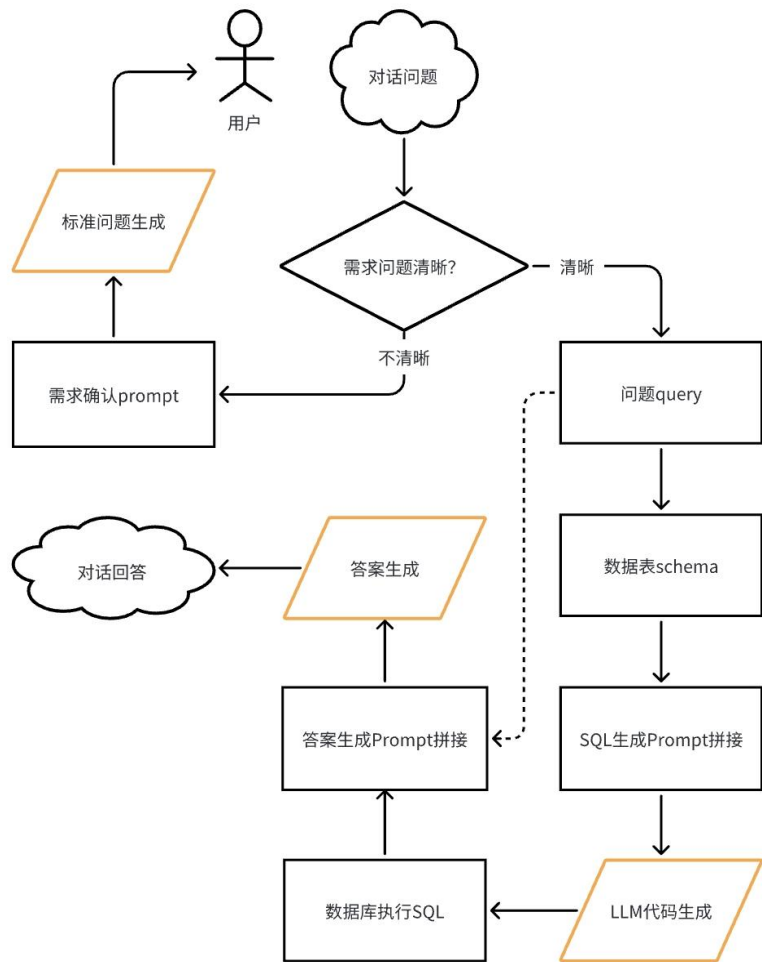
```
{
  "function": {
    "name": "get_indicator_data",
    "args": {
      "indicator_name": "销售额",
      "query_date": "2023-08-19",
      "query_dim": "食品部"
    }
  }
}
```

标准问题query:2023年8月19日，食品部的销售额是多少？

时间、指标、维值

指标问答机器人

业务流程



效果展示

The screenshots show the user interface of the indicator Q&A robot. The interface includes a search bar, a chat area with user questions and system responses, and feedback buttons (thumbs up/down). The responses are generated by the LLM and provide detailed information about the sales data.

Query 1: 昨日 淘系销售额

Response 1: 您的问题已经包含了指标名称（销售额）、查询维度（日期）、请向您关注的日期是？

Query 2: 2023年8月19号，淘系的销售额数据

Response 2: 2023年8月19号，淘系的销售额为 1.2 亿，环比下降 10%。

Query 3: 为什么销售额环比下降这么多？

Response 3: 2023年8月19号，销售数据的波动诊断

• 销售额 环比昨日 -10%，参考正常区间[-4%，4%]。从二级渠道维度来看，主要原因为 渠道销售额 下降。



感谢观看