B站数据治理指标体系

椰子 哔哩哔哩 资深数仓开发工程师

目录
CONTENT

01 数据治理的背景

数据治理项目背景

各类管理问题

1. 数据爆发式增长，存储猛增，任务性能堪忧
2. 数据质量无保障，事故频繁，客诉多
3. 资产缺乏管理，成本无法评估
4. 数据权限体系混乱，存在数据安全隐患
5. 其余问题等等……

怎么做？

类目繁杂

- 产出超时、数据不一致、存储紧张、任务跑不动、找不到负责人、数据还有人在用吗……

怎么做完？

存量巨大

- 多年历史积累：孤岛数据、未压缩、有一部分无主数据、无人跟进……

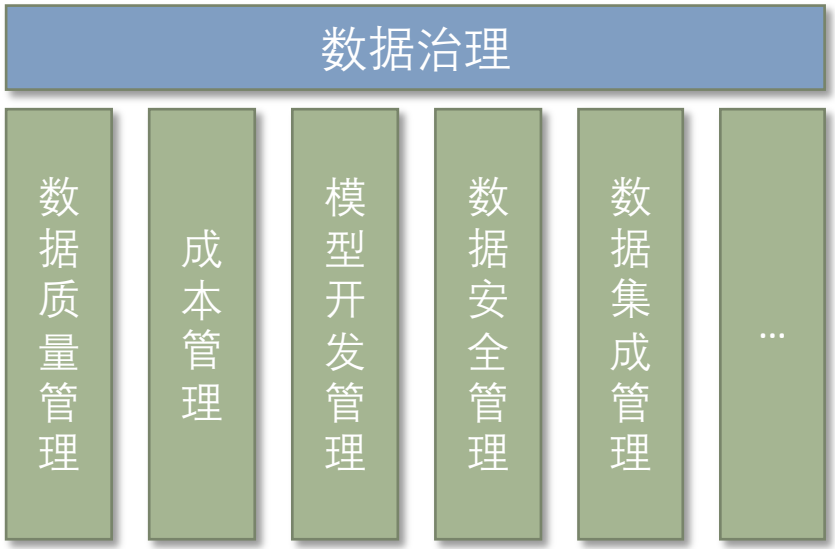
数据治理项目背景

◆ 引入数据化数据治理方法

要解决：

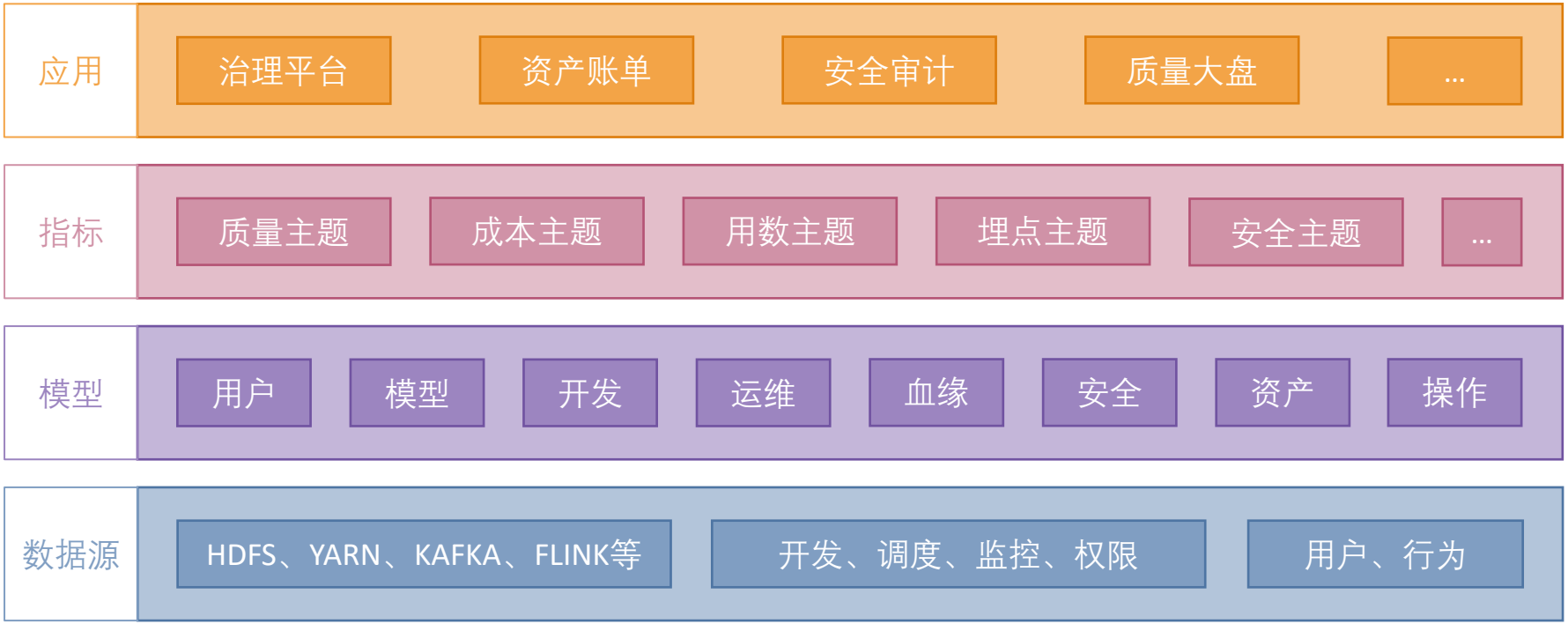
- ✓ 由谁
- ✓ 做些什么动作和内容
- ✓ 为什么要做
- ✓ 怎么做
- ✓ 做到什么程度

的问题



- ✦ 数据治理是数据管理框架中的横向骨干部分
  - ✦ 数据治理是数据管理的计划者、监督者、推进者
    - ✦ 数据治理促使人员按标准和最佳实践来管理数据

数据架构



数据治理指标集的视角

◇ 数据治理定义了数据管理中的：

行为的标准	角色的责任	实施的方法	执行的工具	数据的价值
模型建设标准	生产者责任	如何发现问题	通知型工具	成本评估
版本发布标准	运维者责任	如何制定标准	引导型工具	可靠性评估
质量保障标准	使用者责任	如何执行策略	操作型工具	信息量评估
运维操作标准	管理者责任	如何检验效果	拦截型工具	使用度评估
……	……	……	……	……

每一块定义，都具备建立指标集的视角和模式

数据治理指标集的视角

◆ 数据治理定义了数据管理中的：

行为的标准	角色的责任	实施的方法	执行的工具	数据的价值
模型建设标准	生产者责任	如何发现问题	通知型工具	成本评估
版本发布标准	运维者责任	如何制定标准	引导型工具	可靠性评估
质量保障标准	使用者责任	如何执行策略	操作型工具	信息量评估
运维操作标准	管理者责任	如何检验效果	拦截型工具	使用度评估
……	……	……	……	……

今日的视角

02 数据治理指标模型

治理指标体系模型

指标体系模型

基本运营方法

治理目标

要完成的目标指标

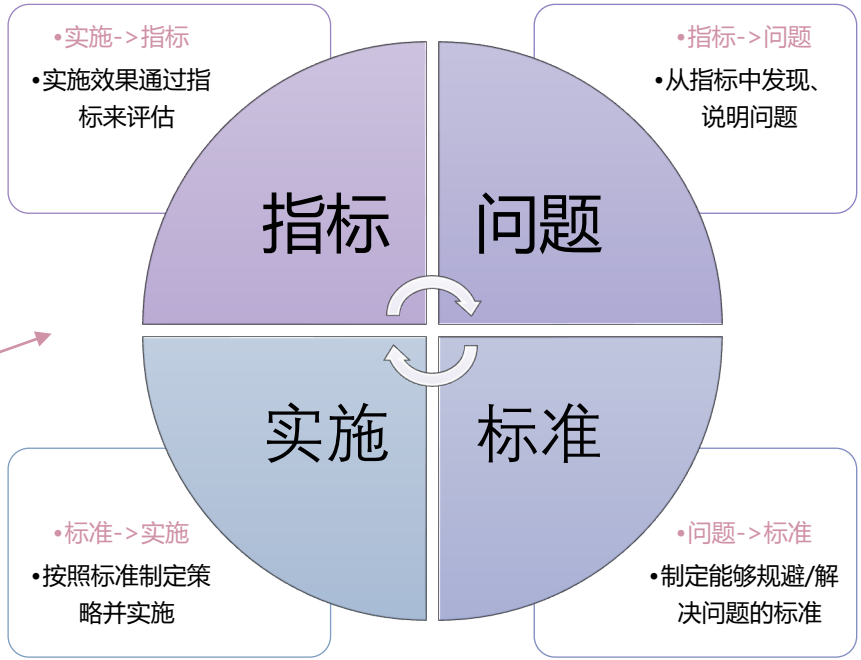
治理策略

从目标拆解实施策略

策略评估

实施命中及效果评估

实施策略和评估指标建立起一个循环



治理策略怎么定？

治理目标 — 要完成的目标指标

治理策略 — 从目标拆解实施策略

策略评估 — 实施命中及效果评估

✓ 策略从目标拆解，所以要先确定目标指标

目标指标是一个周期内要完成的，不是一个无限期目标

正例：本季度目标是存储下降500PB

反例：成本治理目标是存储下降500PB

目标指标是明确的度量，不是一个抽象的概念

正例：本季度目标是P0事故数=0

反例：本季度目标是不发生重大事故

✓ 策略制定分为策略方向和实施项

策略方向是通过目标指标或上层策略方向直接获得的

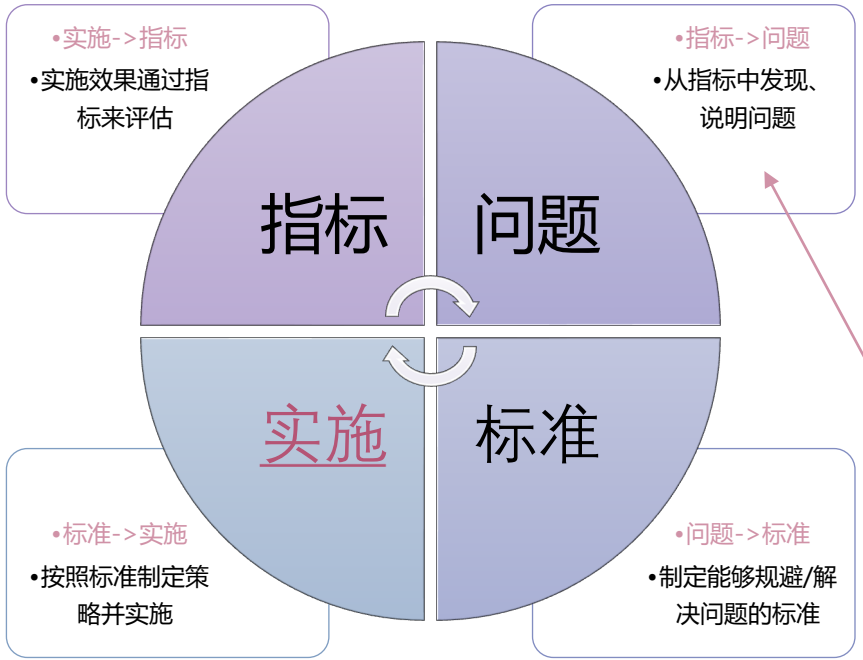
策略方向是一个基于因果、包含等直接关联的拆解

实施项是基于上层策略方向，探索数据而得的

实施项有时不易于从因果性、包含性等关联关系的思考中直接获得

基本运营方法

实施项是基于上层策略方向，探索数据而得的



上层策略方向

找出与策略方向大相径庭的数据资产清单

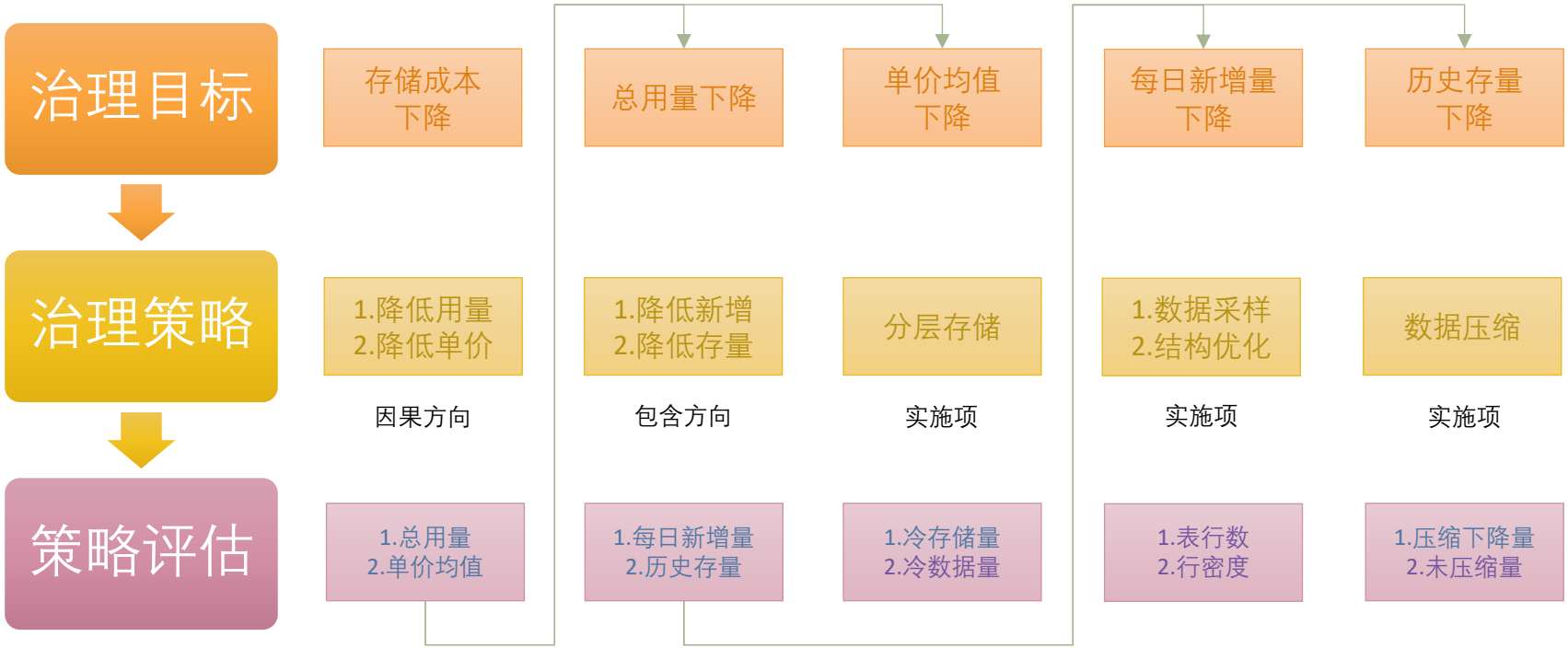
探索该清单中的共性特点
探索目标：从中发现问题、**说明问题**

“从xx指标来看，这个数据存在的问题是xxxxx。”

循环的起点

治理指标体系模型

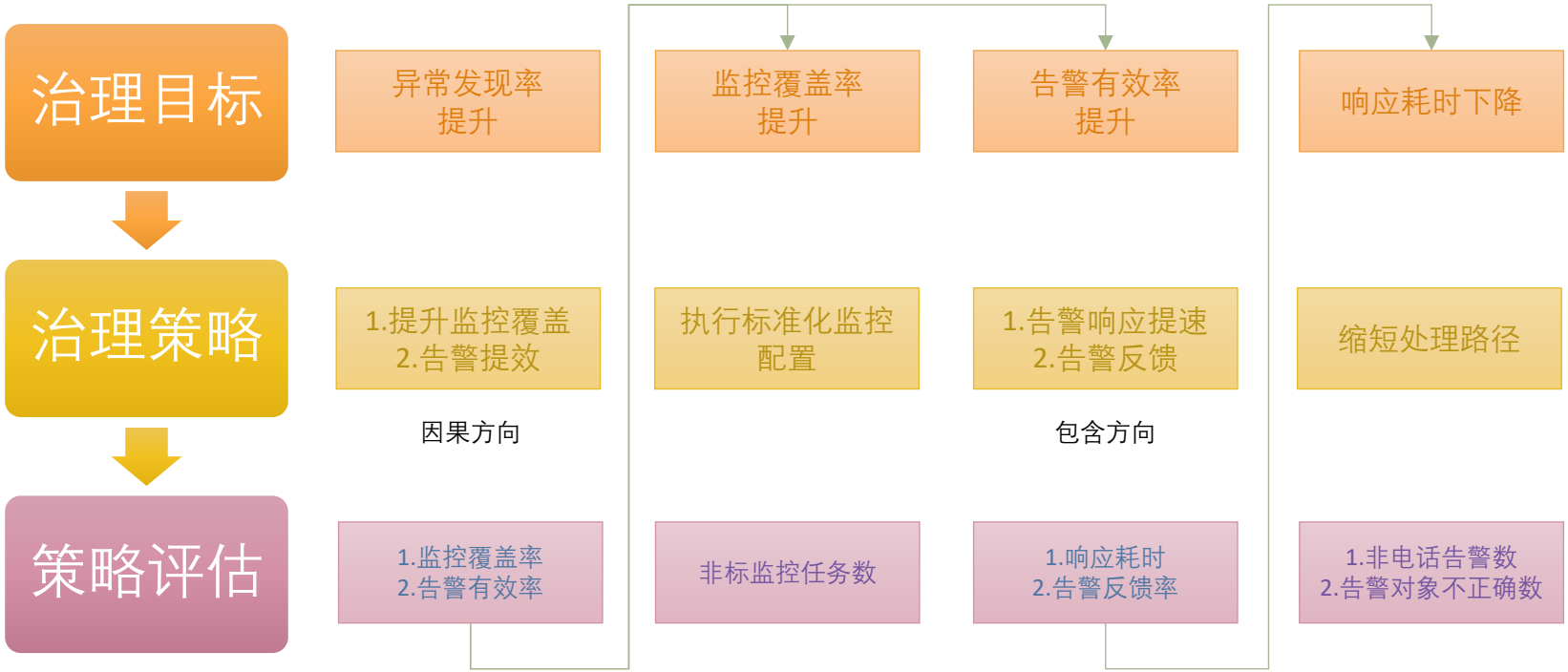
一个体系，全盘适用 | 成本管理 | 存储成本治理



指标拆解：先列策略效果指标，再列实施项命中指标

治理指标体系模型

一个体系，全盘适用 | 数据质量管理 | 监控治理



指标拆解：先列策略效果指标，再列实施项命中指标

治理指标体系模型

指标拆解：先列策略效果指标，再列实施项命中指标

建设的过程

1.明确治理场景与目标
2.根据治理目标，制定策略大纲
3.根据策略，**预定义指标体系**

1.明确业务过程与事实度量
2.明确维度与维度属性标准
3.建立治理主题数据标准库

1.模型设计与创建
2.任务链路开发、发布上线
3.治理工具、报表等透出

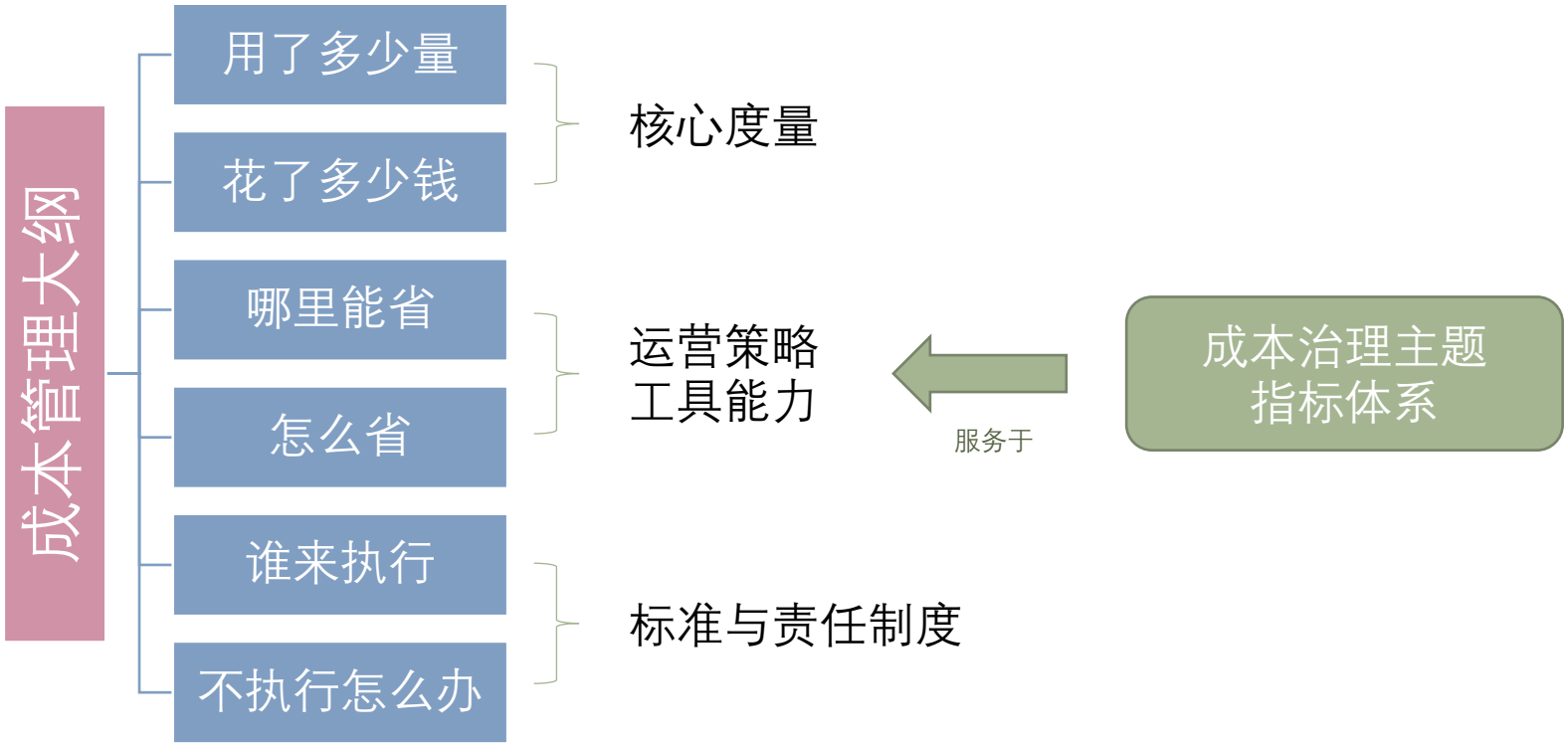
场景分析 → 数据源调研 → 数据标准定义 → 指标口径定义 → 维度建模 → 指标汇总

1.按所需指标，调研相关数据源
2.确认既存数据源并探查质量
3.提炼新数据源需求

1.明确原子指标与派生指标
2.定义指标的统计口径
3.**重塑指标体系**，评估策略匹配度

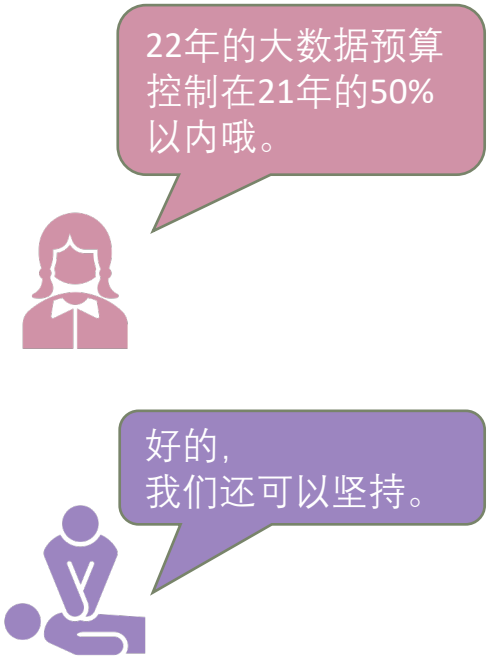
03 成本治理实践

成本管理大纲

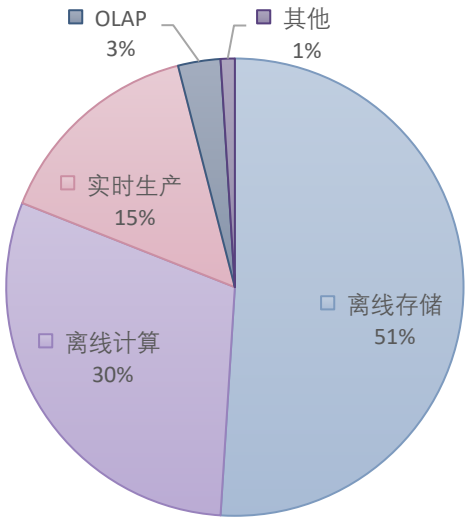


目标的确定

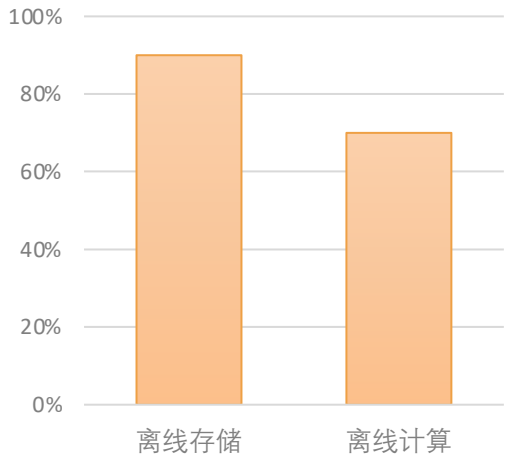
找到瓶颈



成本分布

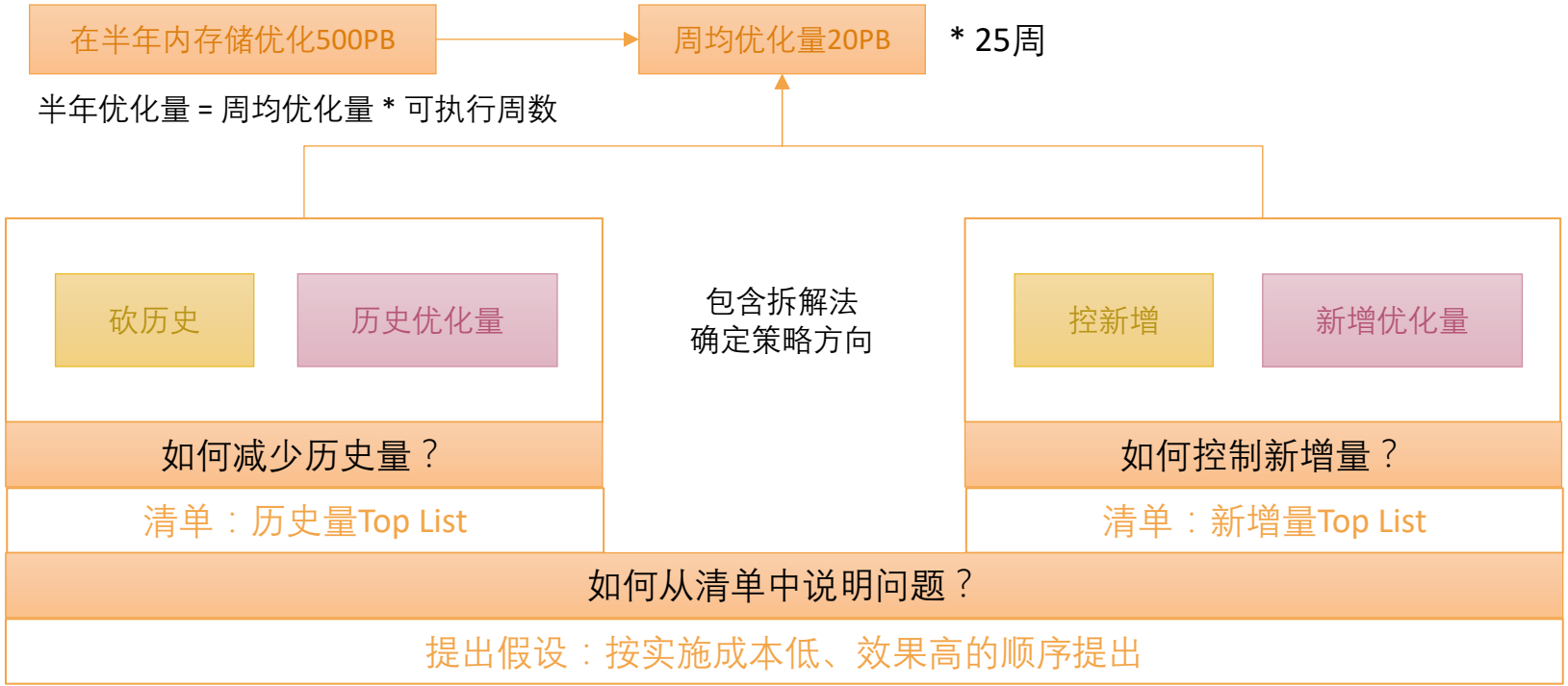


利用率

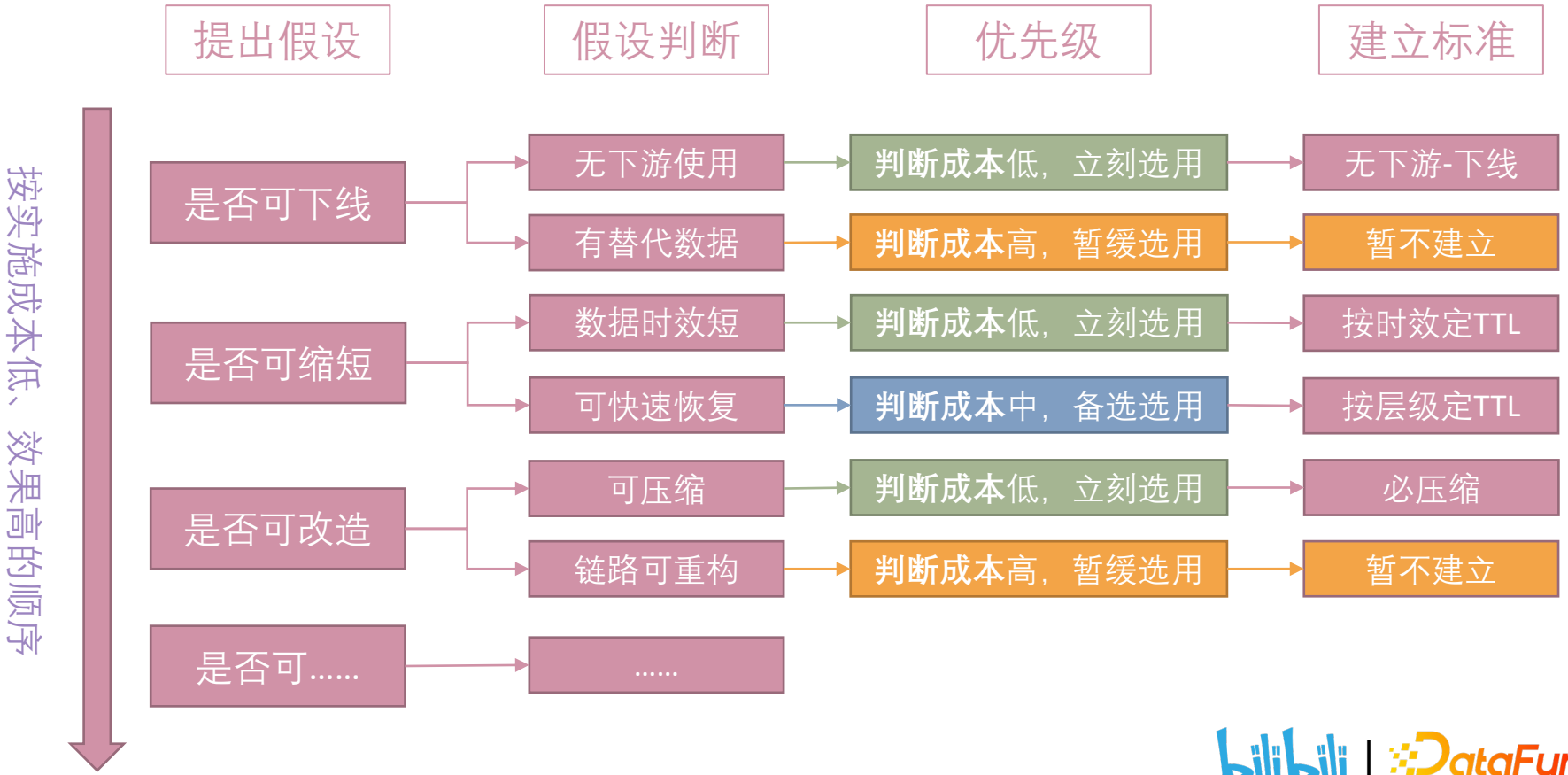


结论：成败的关键在于存储。

策略的确定

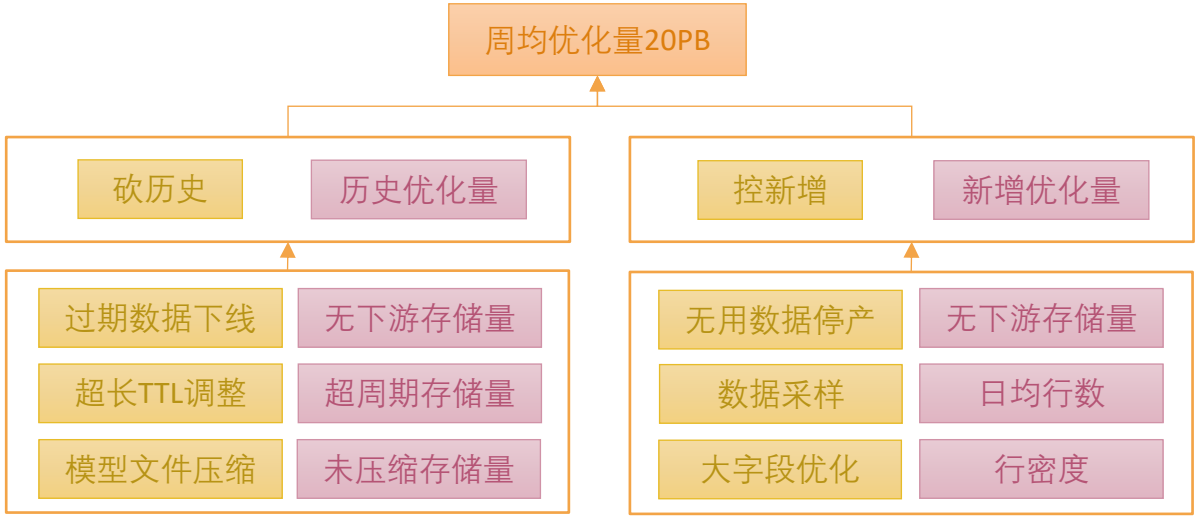


策略的确定



指标体系预定义与数据源调研

预定义指标体系



调研数据源

优化量	优化前后存储量	HDFS文件大小
无下游存储量	模型间关系	数据血缘
超周期存储量	SQL日期范围	全量任务信息

等等……

指标定义的改善与关注的维度

举个改善的栗子

无下游存储量 → 模型间关系 → 数据血缘

指标作用：想找到没有下游使用的模型，命中可下线的策略方向实施。

判断逻辑：通过数据平台的血缘信息，没有任务（包含调度及查询）使用。

遇到问题：个别团队有非标访问（野生客户端），不能被平台的血缘收录。

无下游存储量 → 所有的访问 → HDFS审计日志

数据源调整：由【数据血缘】调整为【HDFS审计日志】

业务过程调整：由【任务引用】调整为【HDFS的读/写】

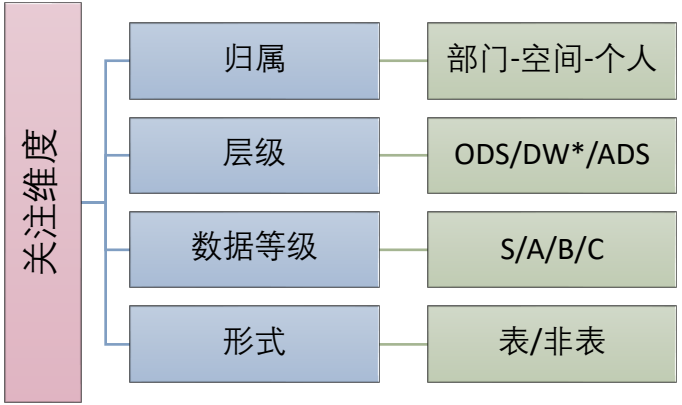
事实调整：由【任务数】调整为【读/写次数】

原子指标：存储量

派生指标：

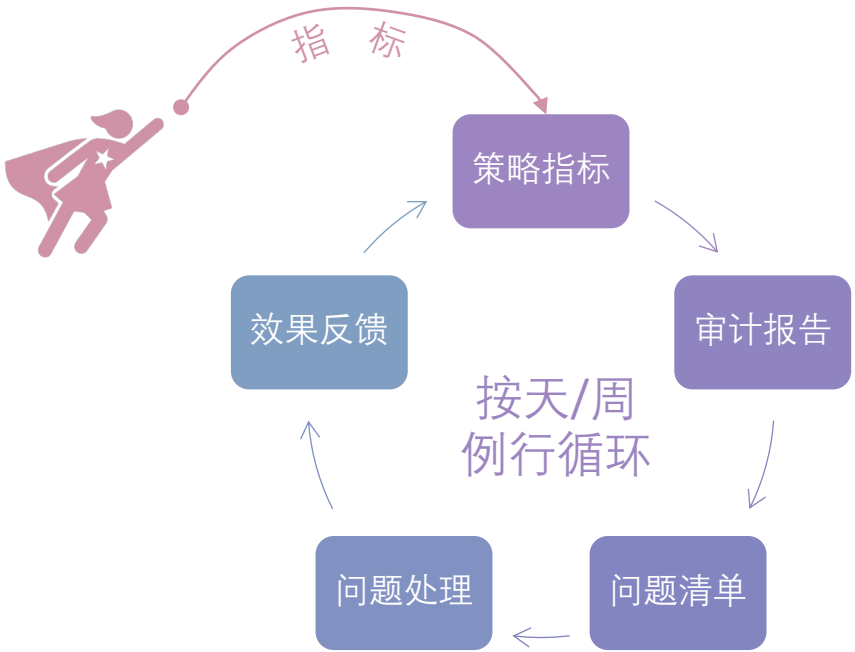
- 无下游存储量 = 过滤系统账号（如dqc账号）的访问后，读次数为0的模型所使用的存储量总和
- 30天/60天/90天无下游存储量 = 连续30天/60天/90天无下游的模型所使用存储量总和

实施过程中的关注维度



不同维度属性下，实施细则可微调

将指标投入运营



小循环：持续解决已确认问题

每周一早上

- 数据任务运行，更新审计报告

截止当前，问题项统计

dept	总存储量	无下游剩余	待压缩剩余	...
A	100PB	10PB	20PB	...
B	150PB	15PB	30PB	...

- 更新问题清单

dept	user	table	无下游	待压缩	...
A	甲	db.tbl_1	是	否	...
B	乙	db.tbl_2	否	是	...

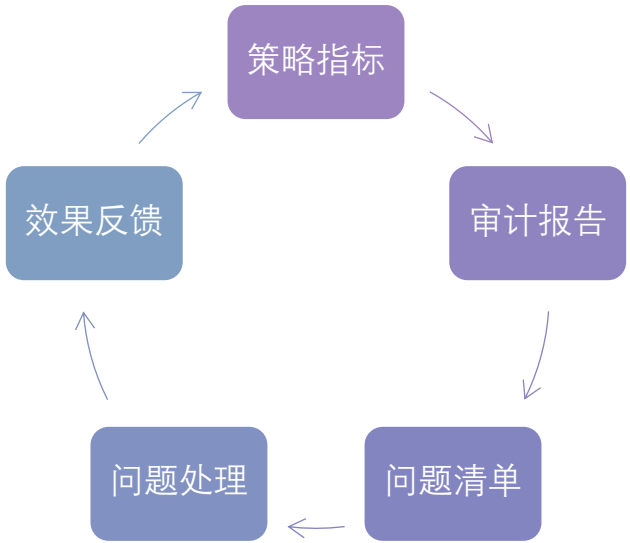
- 通知owner处理问题

请@甲 @乙 关注，在周五前完成处理。

每周五下午

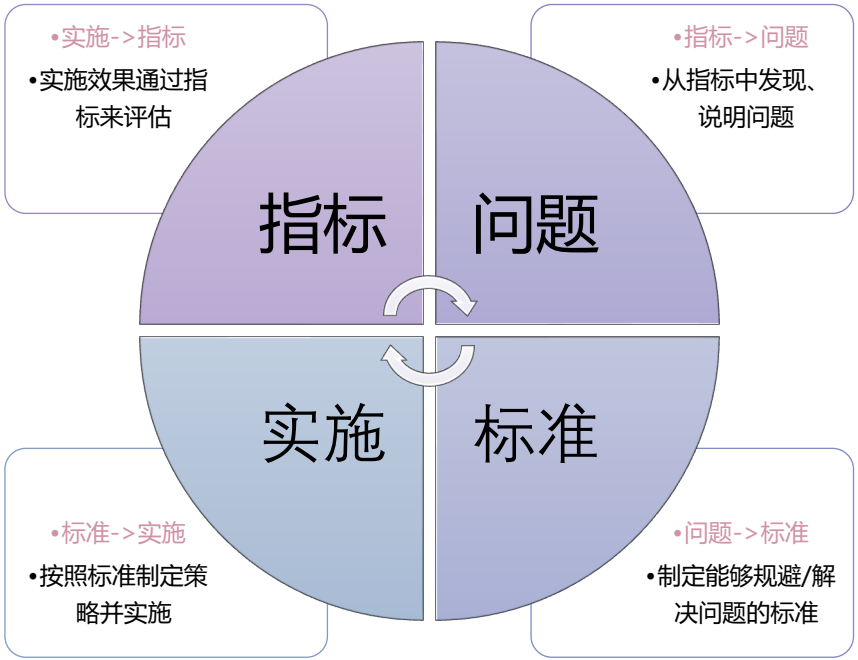
dept	本周优化量	无下游实施	压缩实施	...
A	10PB	1PB	2PB	...
...	...	...	...	...

持续化运营



解决了已有问题
无下游、待压缩...

新的问题与策略
游离目录、大字段...



小循环与大循环的来回往复

治理成效

实施量 远超 计划量

达成时间比预计提早近 一个月

下半年存储增长 1%（近0增长）

存储年增长量下降 66%

04 题外

题外

◆ 数据治理定义了数据管理中的：

行为的标准	角色的责任	实施的方法	执行的工具	数据的价值
模型建设标准	生产者责任	如何发现问题	通知型工具	成本评估
版本发布标准	运维者责任	如何制定标准	引导型工具	可靠性评估
质量保障标准	使用者责任	如何执行策略	操作型工具	信息量评估
运维操作标准	管理者责任	如何检验效果	拦截型工具	使用度评估
……	……	……	……	……

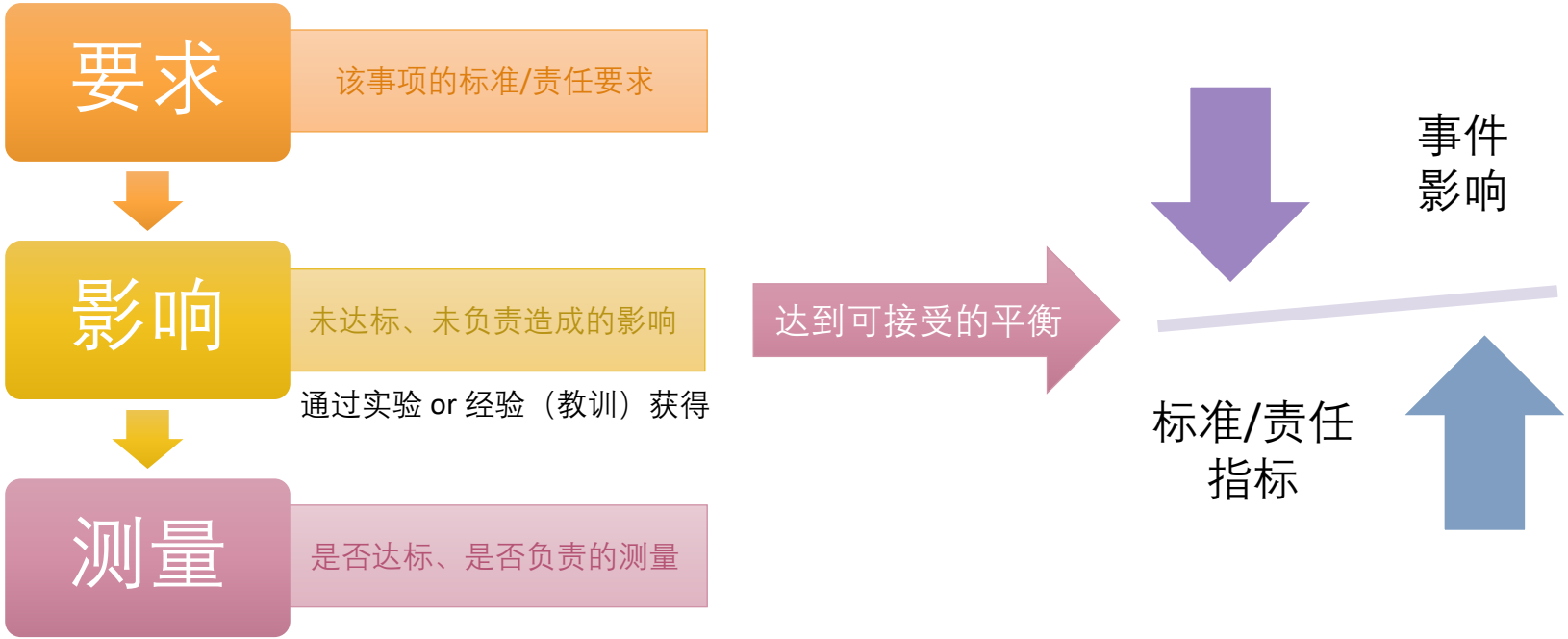
看看不同的视角

运营两字，听起来较为温和

但数据治理并不只有温和的一面

题外

标准/责任-指标体系 ——通常标准/责任类指标体系，是连带指标值一起定义的。



bilibili | DataFun.

欢迎关注

哔哩哔哩技术

微信扫描二维码，关注我的公众号

非常感谢您的观看

bilibili | DataFun.