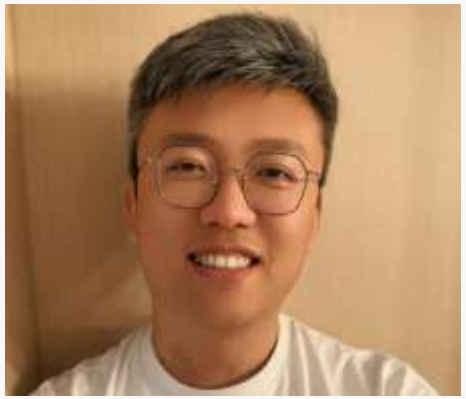
B站的 数据治理运营框架 实践

高隆 bilibili 数据仓库工程师

讲师简介



高隆

表哥
DAMA
B站
数据成本治理
数据质量

目录

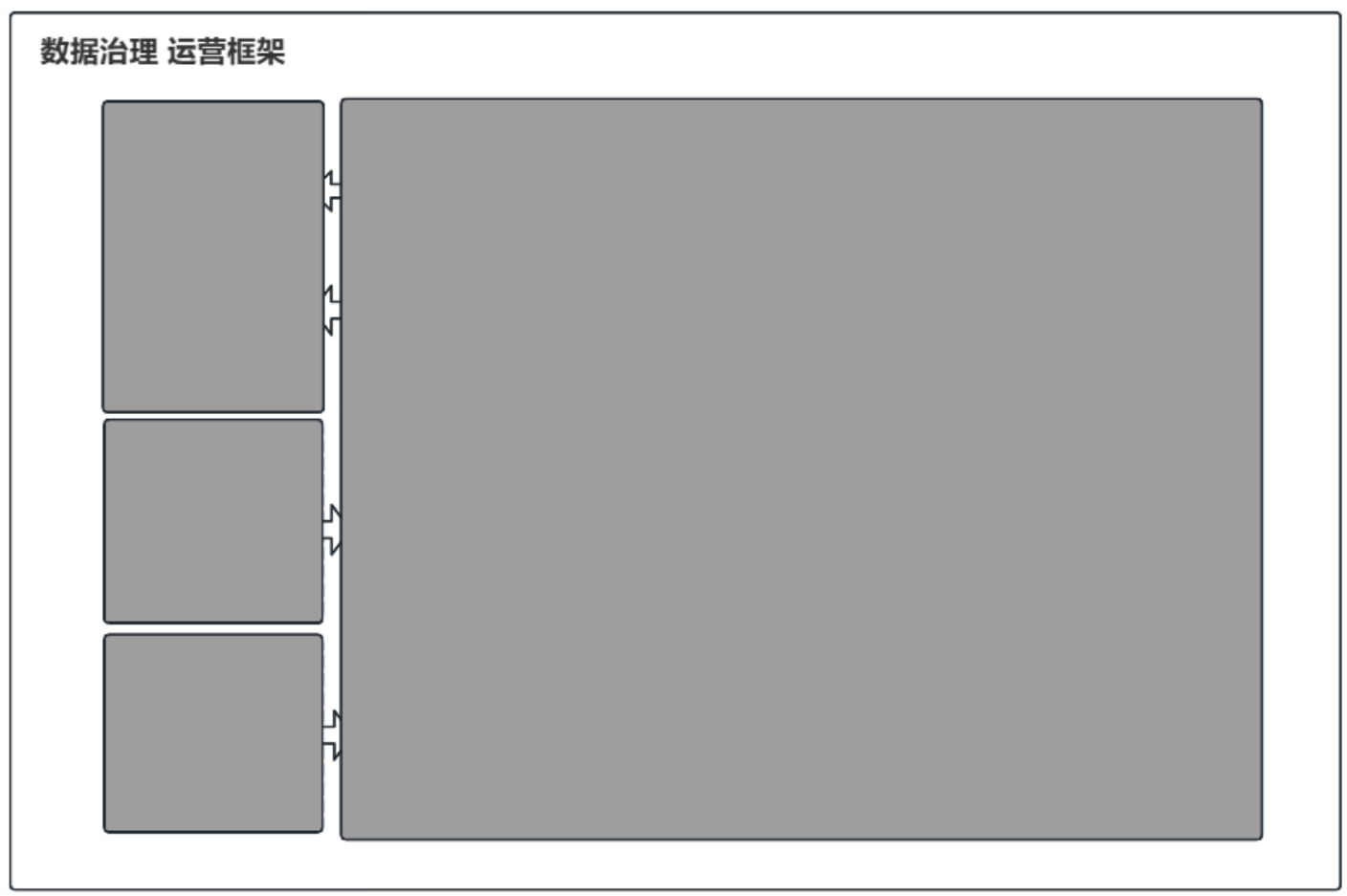
分析工具：DAMA-Bok

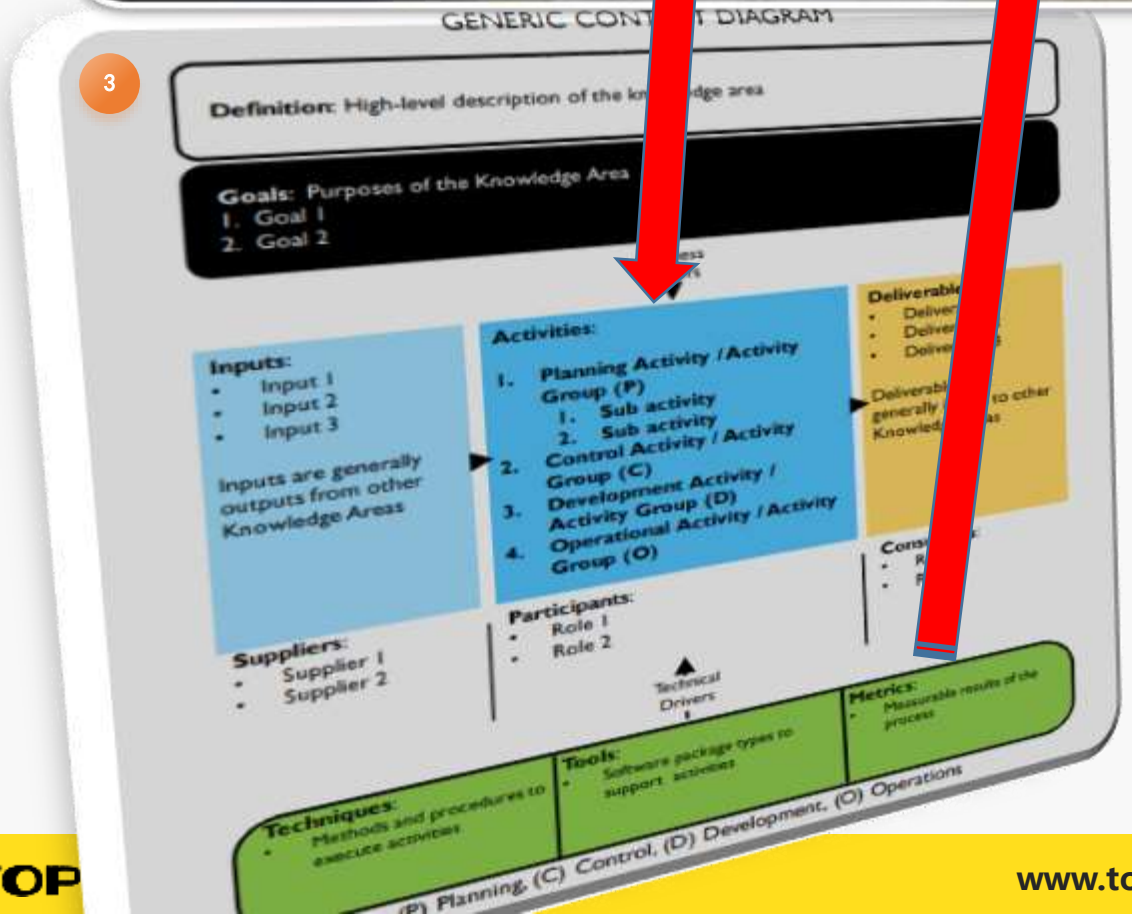
案例1--2022-05-11 存储水位风险

- 虚拟组织
- 嵌入治理
- 元数据管理

案例2--2023-10-30 数据丢失复盘

- 质量运营
- 质量的需求与满足
- 数据治理中的风险





数据管理知识体系
DAMA-DMBOK

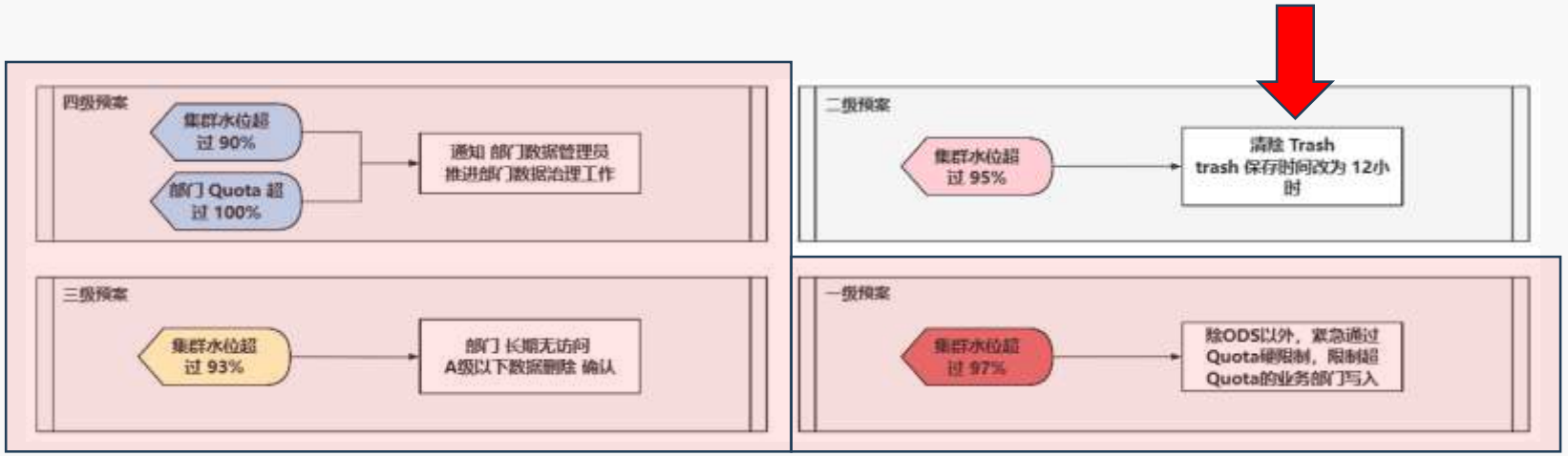
DAMA 成立于 1980年，是一个全球性数据管理和业务专业志愿人士组成的非营利协会，致力于数据管理的研究和实践

* 左上图是买书的时候送的鼠标垫

1. 车轮图：11个知识领域
2. 6边形图：每个知识领域的7件事情
3. 语境关系图：每个知识领域都可以展开成具体的活动、方法、目标、指标
4. 12原则：采取行动和判断时的依据

案例1 -- 2022-05-11 存储水位风险

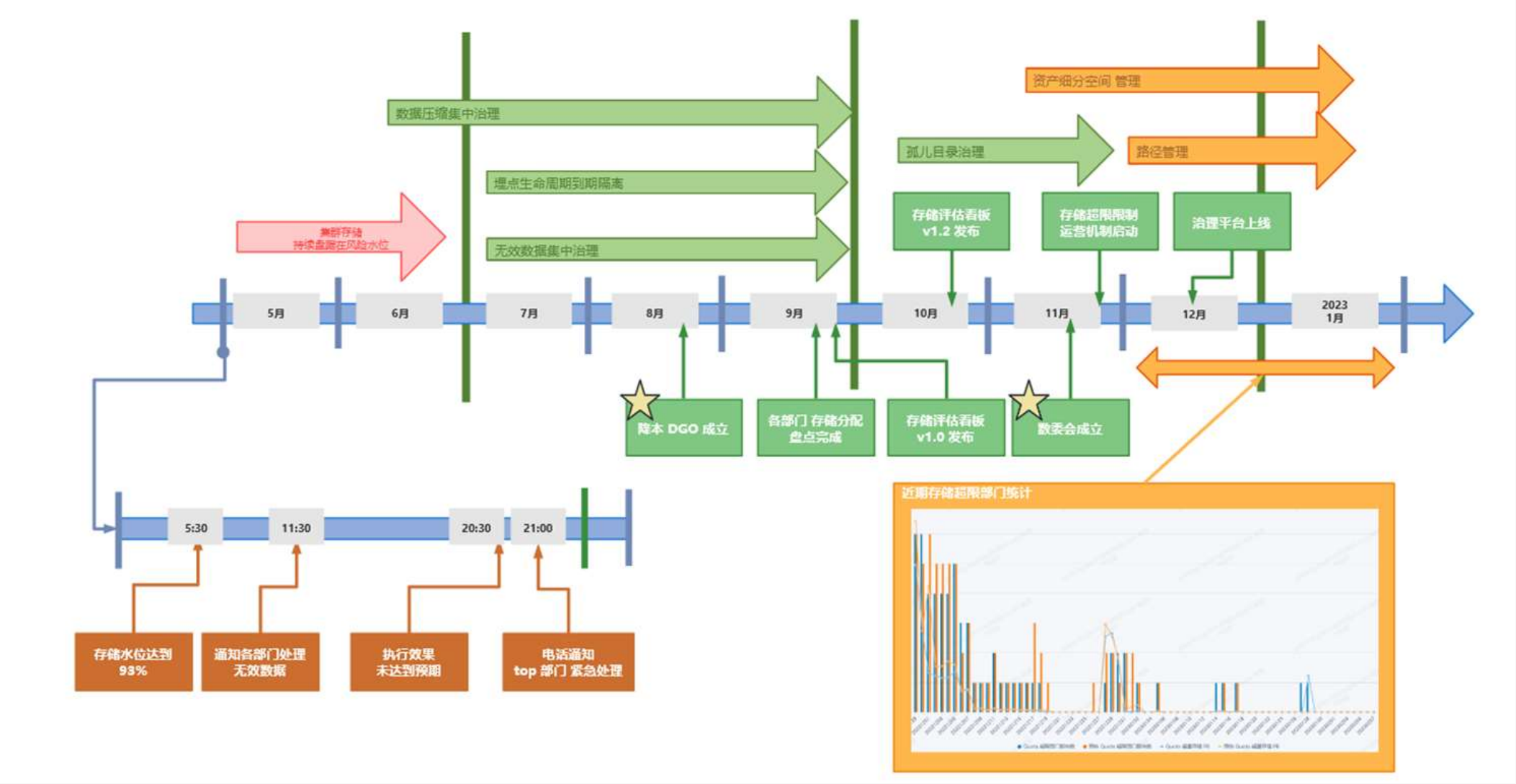
案例背景



名词解释：

- 存储水位：HDFS集群存储
- 部门数管：部门的“CDO”
- Quota：部门预算资源分配
- A级数据：多为跨部门使用数据
- Trash 数据：一般存7日

案例背景


数据压缩集中治理
埋点生命周期到期隔离
无效数据集中治理
资产细分空间 管理
孤儿目录治理
路径管理
存储评估看板 v1.2 发布
存储超限限制 运营机制启动
治理平台上线
5月
6月
7月
8月
9月
10月
11月
12月
2023 1月
降本 DGO 成立
各部门 存储分配 盘点完成
存储评估看板 v1.0 发布
数委会成立
5:30
11:30
20:30
21:00
存储水位达到 93%
通知各部门处理 无效数据
执行效果 未达到预期
电话通知 top 部门 紧急处理
近期存储超限部门统计

问题与挑战

预案	执行项	问题	DAMA Bok
• 4级（90%）	通知数管执行治理	• 组织变更 • 数管权责不明确	• 原则：数据管理需要领导力承诺 • 领域：数据治理
• 3级（93%）	删除长期无访问数据	• 没有执行驱动力 • 删除数据存在风险	• 原则：数据价值使用经济术语表达 • 领域：元数据
• 2级（95%）	删除 trash 调整冷数据容量	• trash类 数据如何归属 • 用户没有直接控制trash大小的能力	• 原则：数据管理需求驱动技术决策 • 领域：数据存储
• 1级（97%）	根据分配限制部门写入	• 组织变更预算归属变更 • 限制数据写入风险极高	• 原则：数据管理是数据生命周期的管理 • 领域：数据仓库与商务智能

破题思路

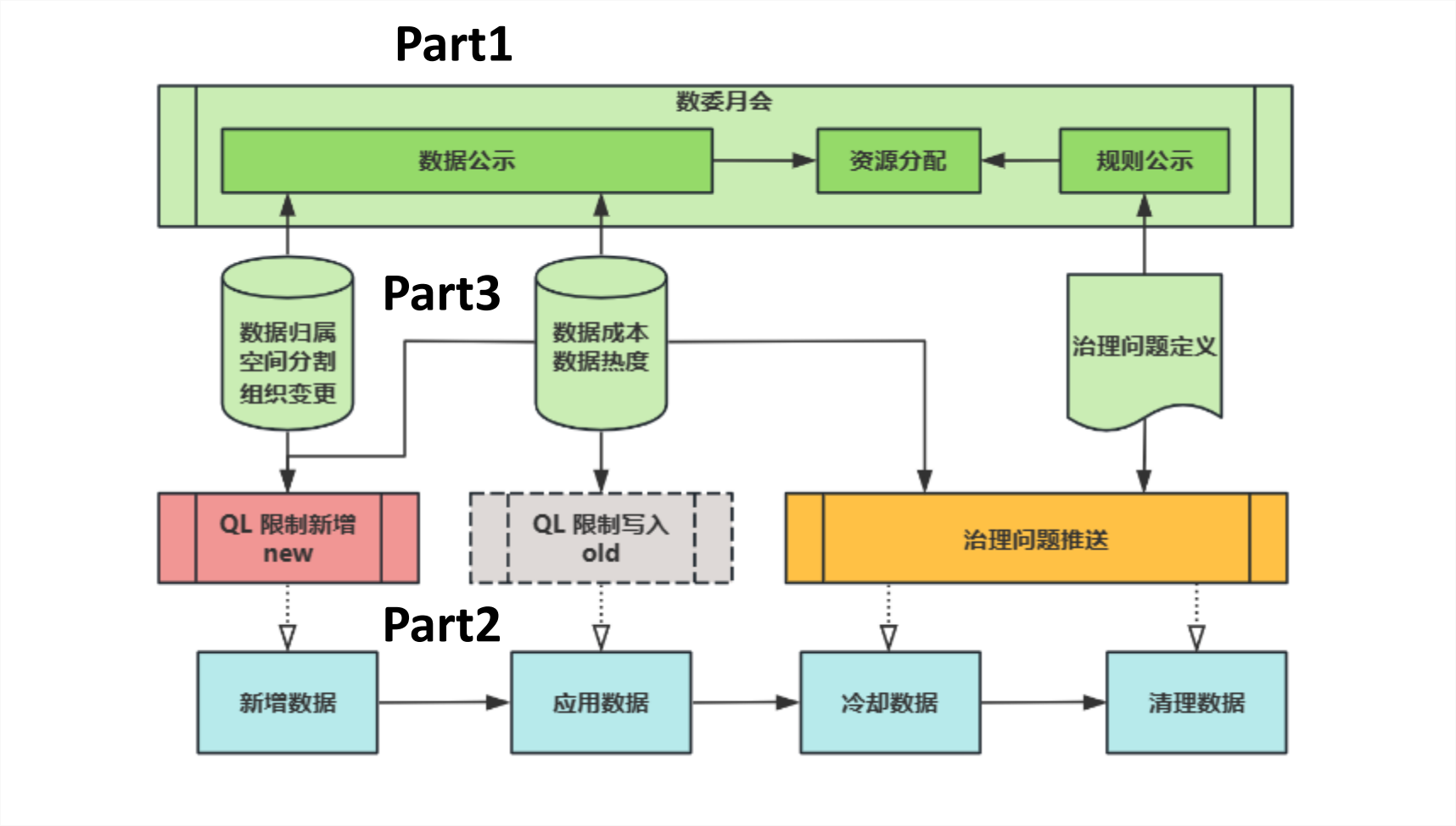
	问题	DAMA Bok	方案

Part1	组织变更	• 原则：数据管理需要领导力承诺
		• 领域：数据治理
			• 虚拟组织
			  • 数委会（人的虚拟组织）
			  • 资产空间（数据的虚拟组织）

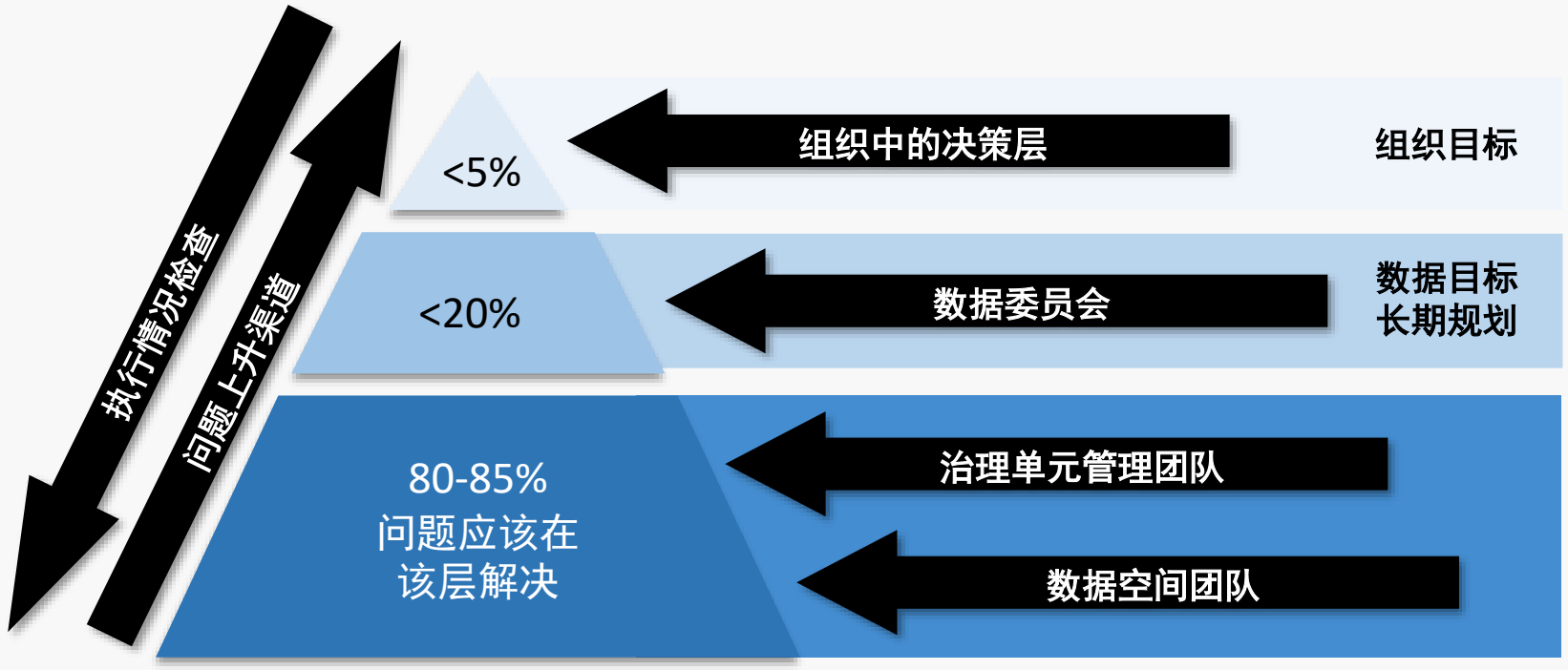
Part2	限制数据写入风险极高	• 原则：数据管理是数据生命周期的管理
		• 领域：数据仓库与商务智能
			• 嵌入治理
			  • 预算分配
			  • QuotaLimit（限制“新增”）

Part3	没有执行驱动力	• 原则：数据价值使用经济术语表达
		• 领域：元数据
			• 元数据的管理与应用
			  • 元数据数仓
			  • 资产账单
			  • 数据治理平台

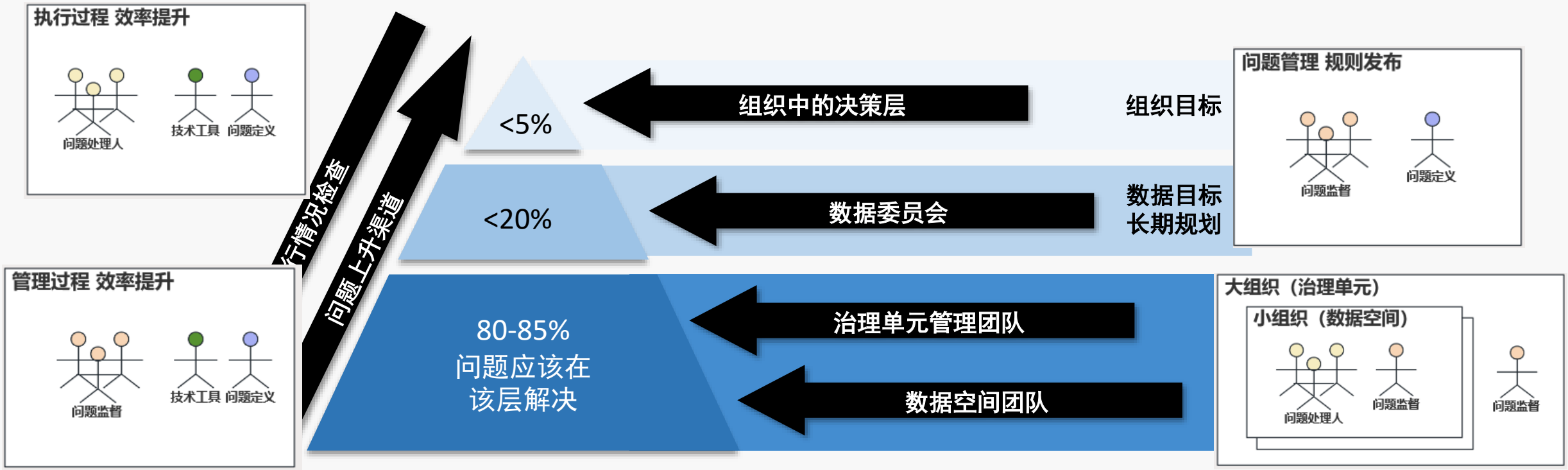
破题思路 – 变更方案


Part1
数委月会
数据公示
资源分配
规则公示
Part3
数据归属
空间分割
组织变更
数据成本
数据热度
治理问题定义
QL 限制新增
new
QL 限制写入
old
治理问题推送
Part2
新增数据
应用数据
冷却数据
清理数据

Part1：数委会 -- 组织

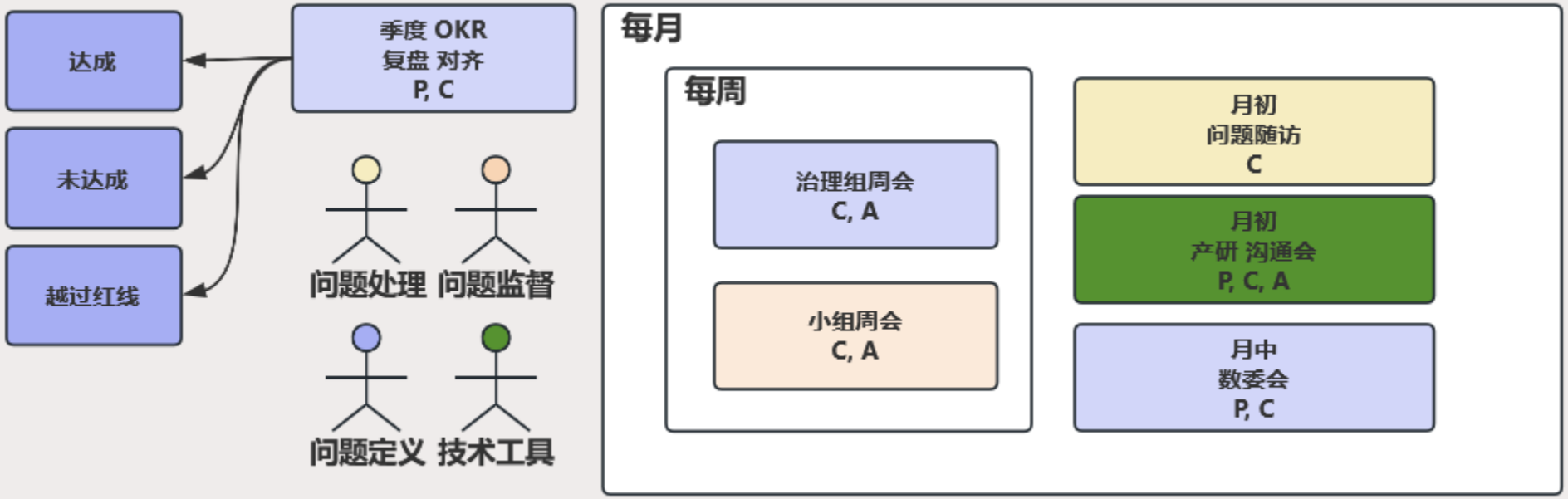

执行情况检查
问题上升渠道
<5%
组织中的决策层
组织目标
<20%
数据委员会
数据目标
长期规划
80-85%
问题应该在
该层解决
治理单元管理团队
数据空间团队

Part1：数委会 -- 角色


执行过程 效率提升
问题处理人
技术工具
问题定义
管理过程 效率提升
问题监督
技术工具
问题定义
行情况检查
问题上升渠道
<5%
组织中的决策层
组织目标
<20%
数据委员会
数据目标
长期规划
80-85%
问题应该在
该层解决
治理单元管理团队
数据空间团队
问题管理 规则发布
问题监督
问题定义
大组织（治理单元）
小组织（数据空间）
问题处理人
问题监督
问题监督

Part1：数委会 -- 活动

Plan, Do, Check, Activity



Part1：数委会 -- 议题

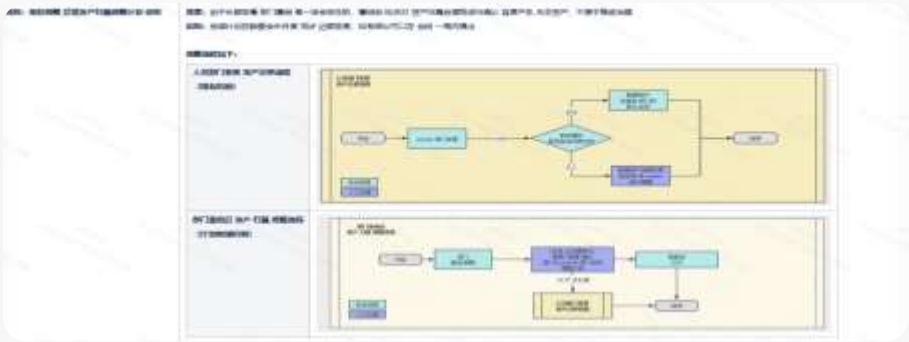
季度目标

月度审计

资源分配

概念说明

组织调整

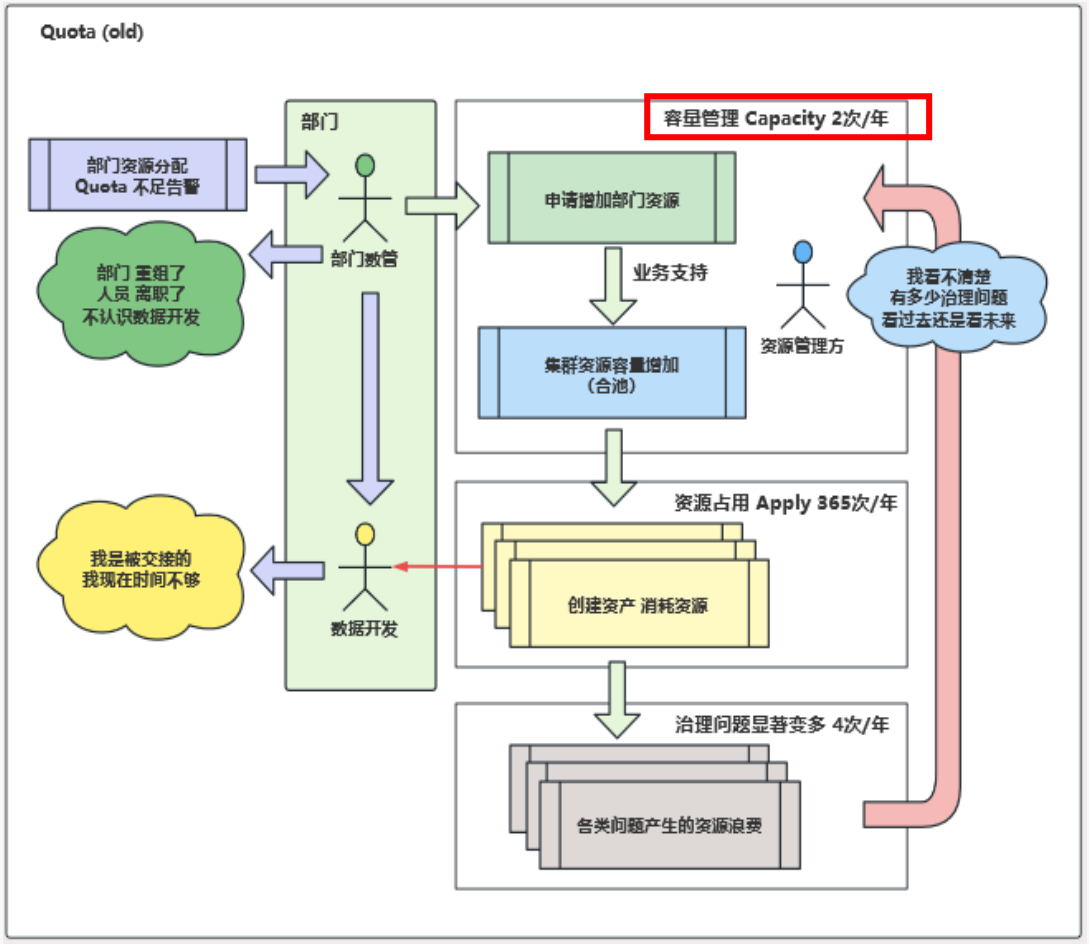


红线规则

Part2：嵌入治理

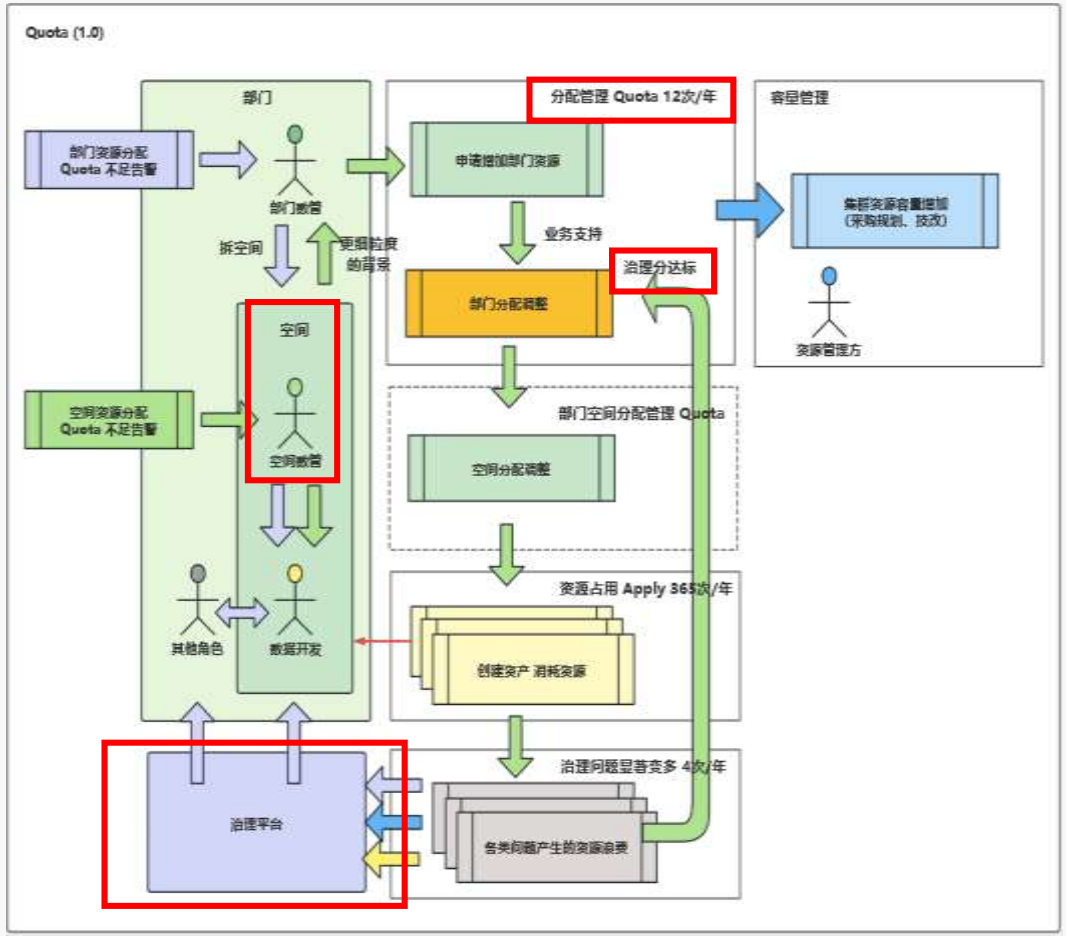
Quota – old (till 2022.H1)

2次/年，组织架构变更

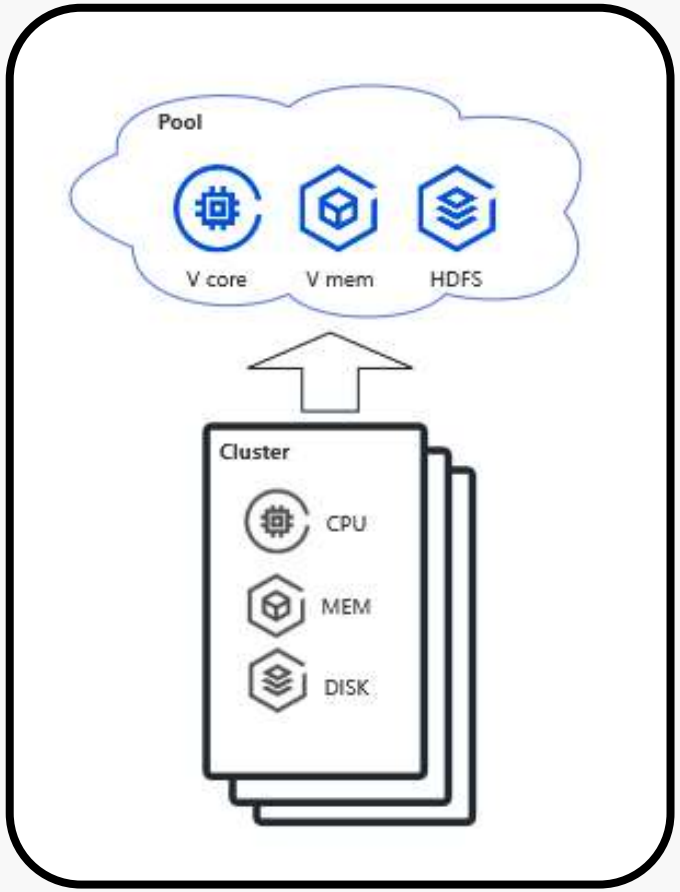


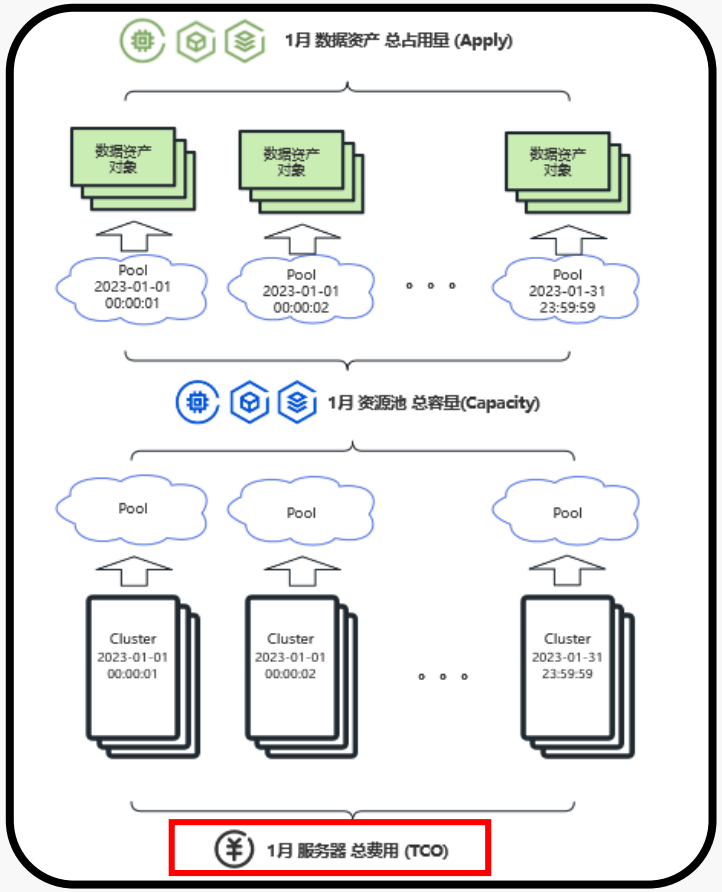
Quota – v1.0

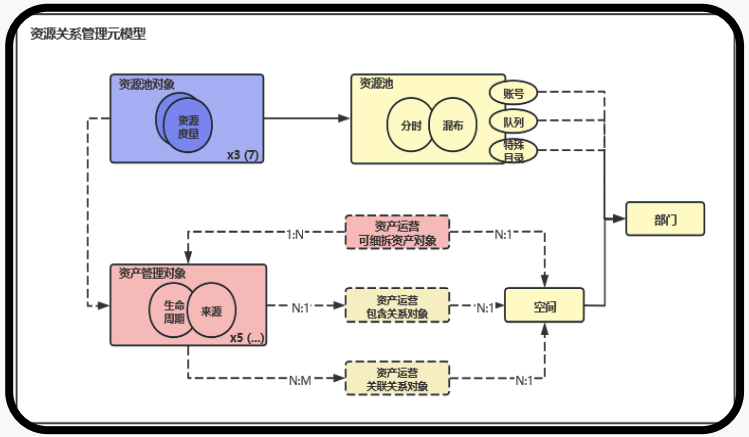
数委月会=12次/年，空间=中间组织，待治理问题=备用资产

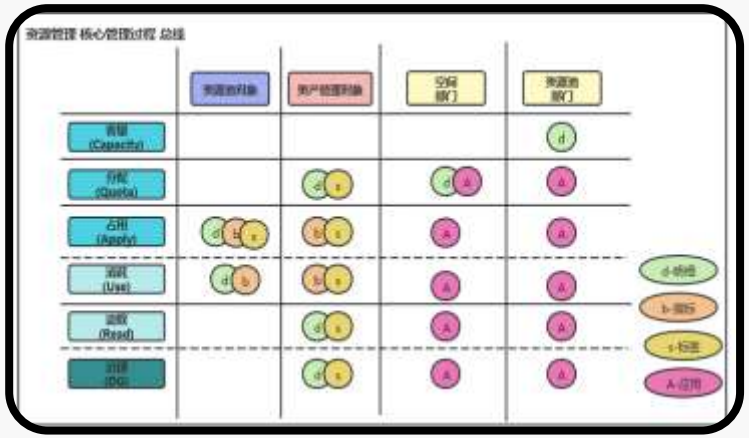


Part3：元数据的应用 -- 资产账单







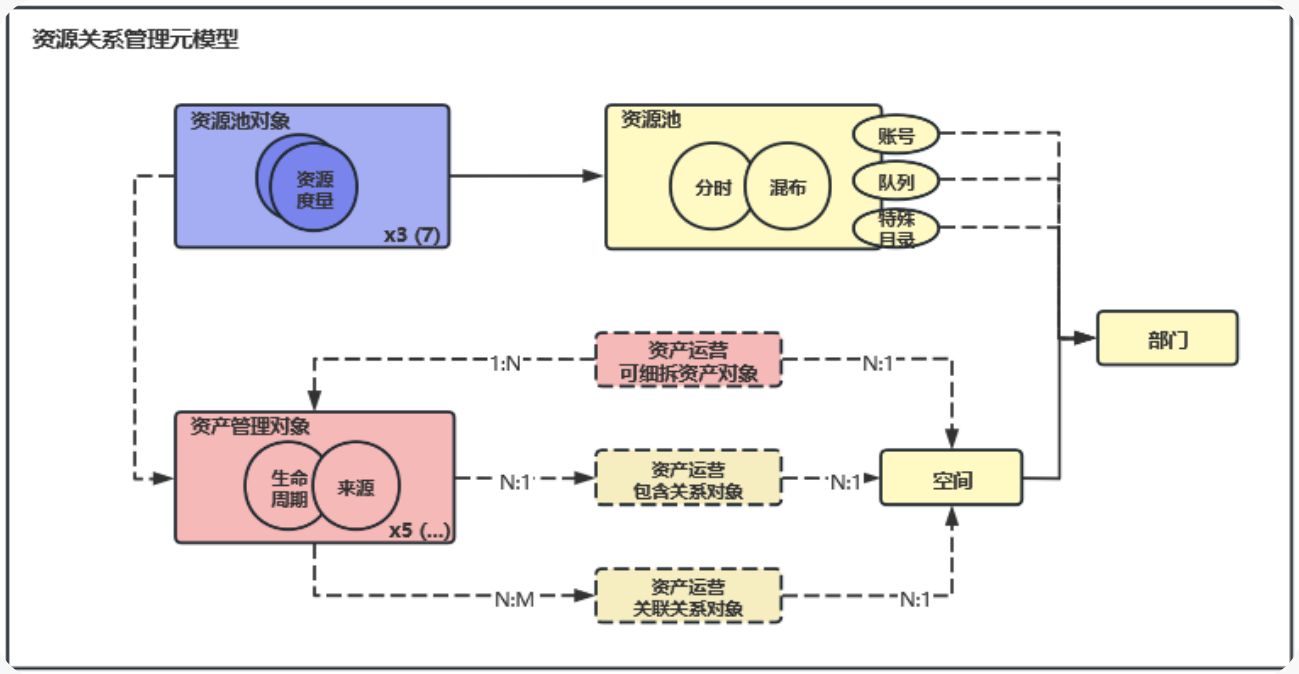


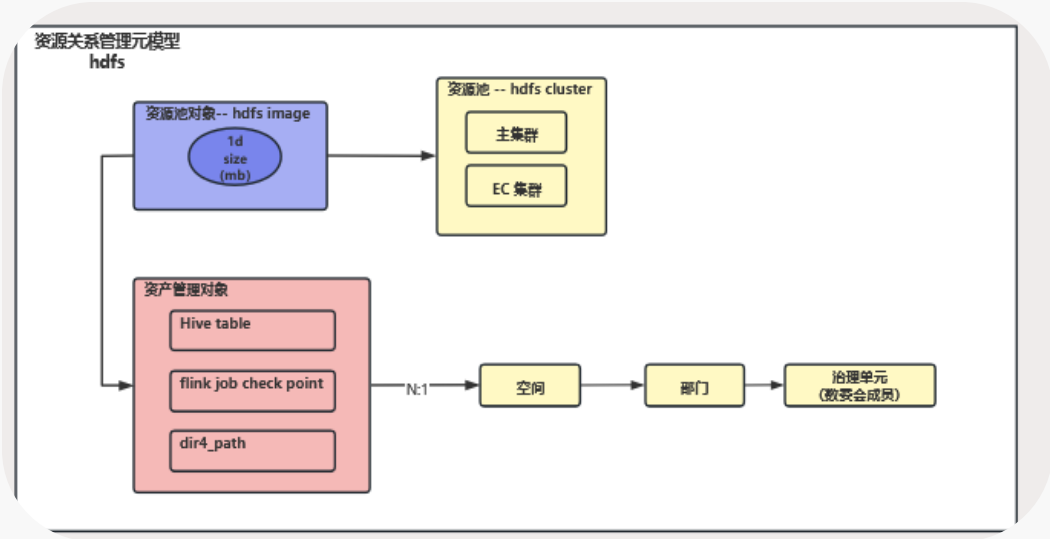
01 物理资源->虚拟资源

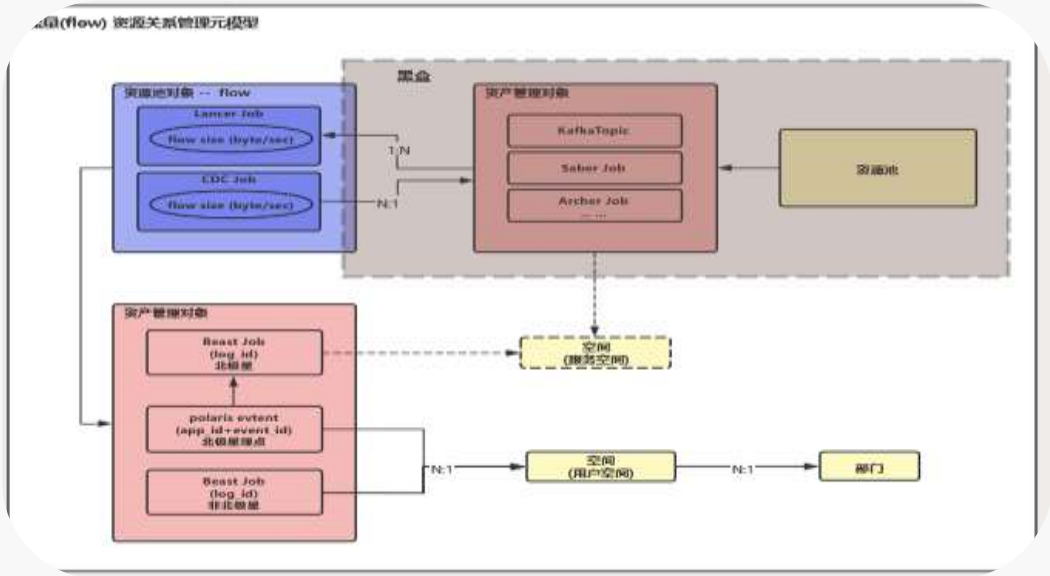
02 资源分配 + 运营成本

03 资源管理 元数据

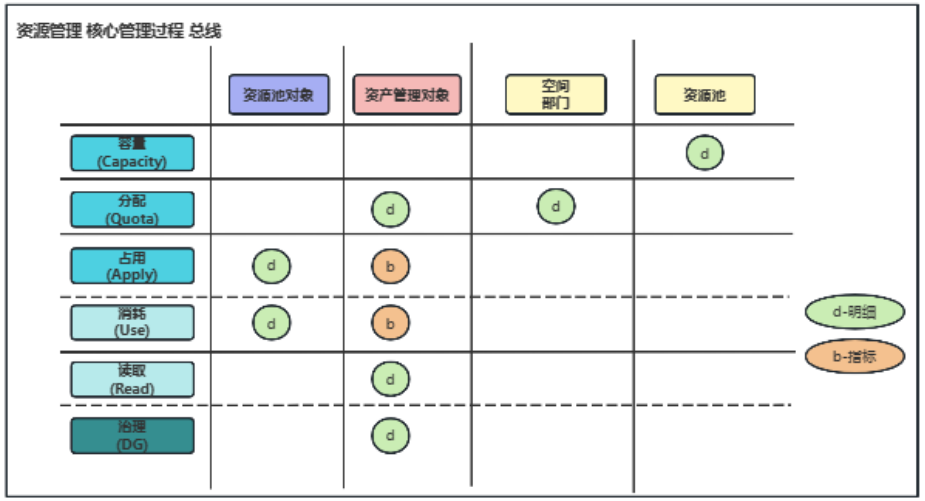
Part3：元数据的管理 -- 元模型

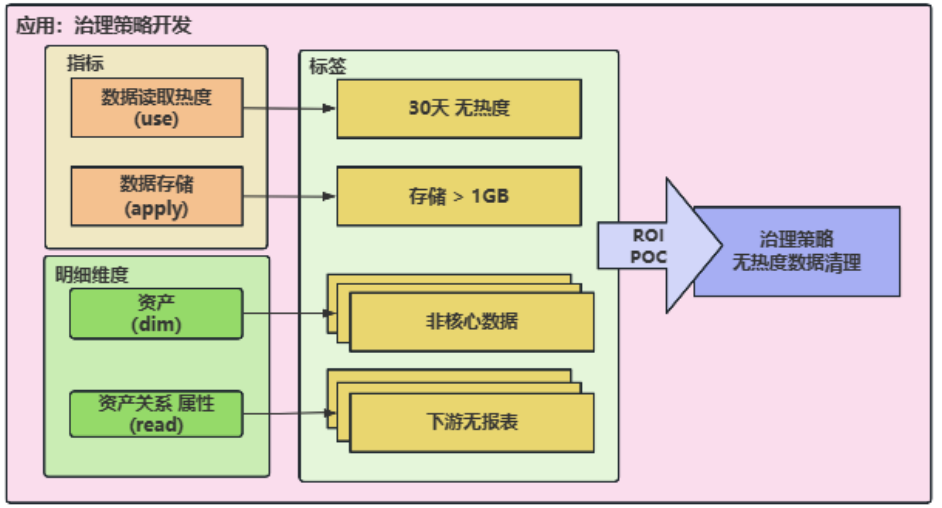


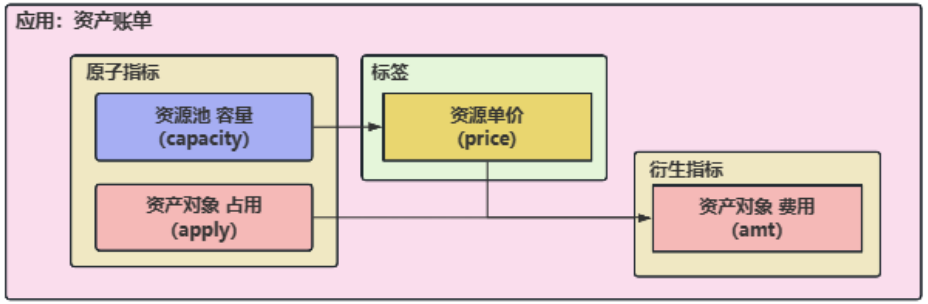


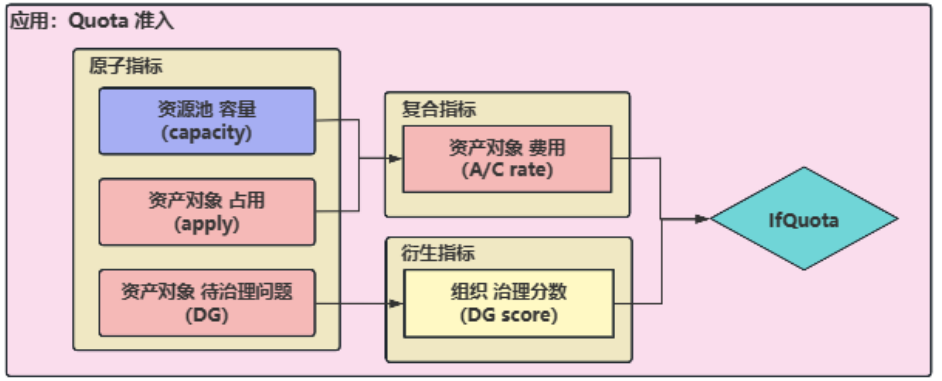


Part3：元数据的管理 -- 元数据的指标与标签

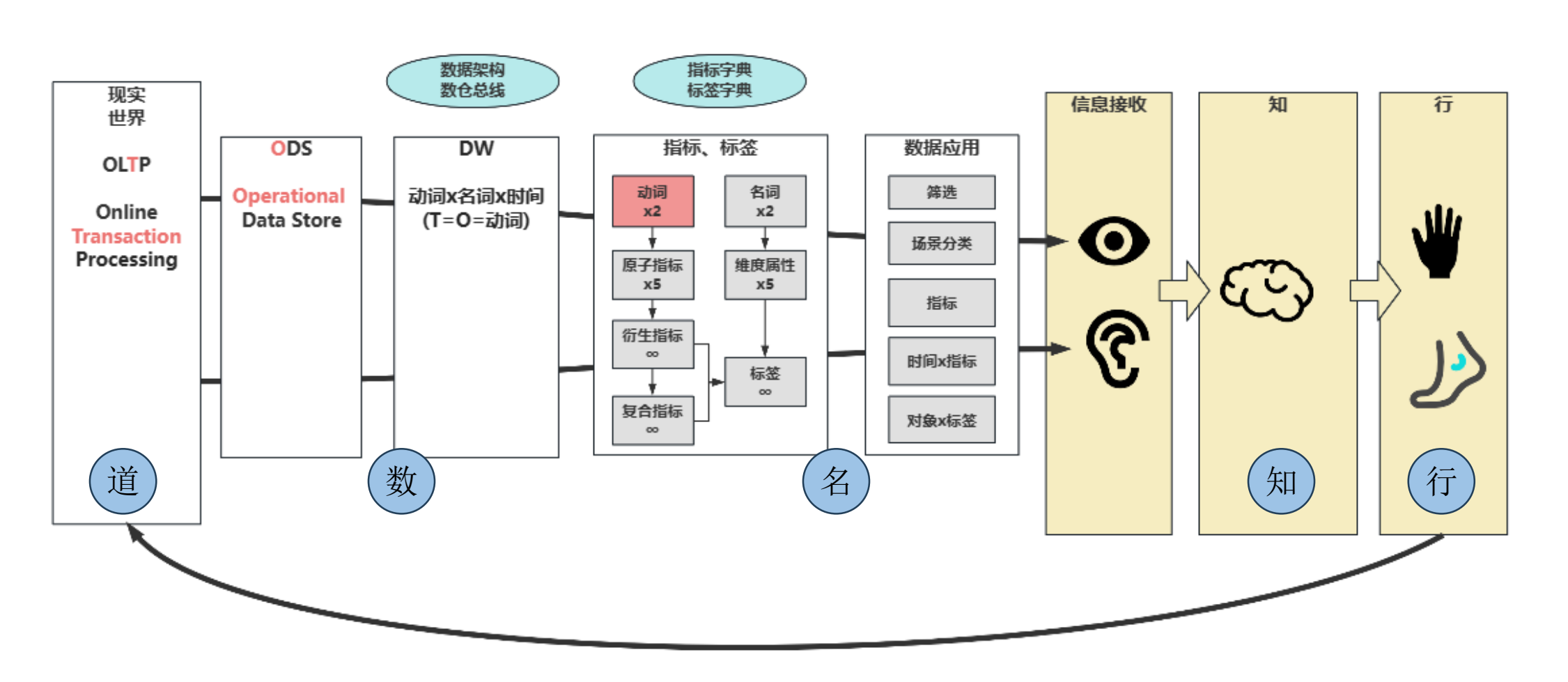


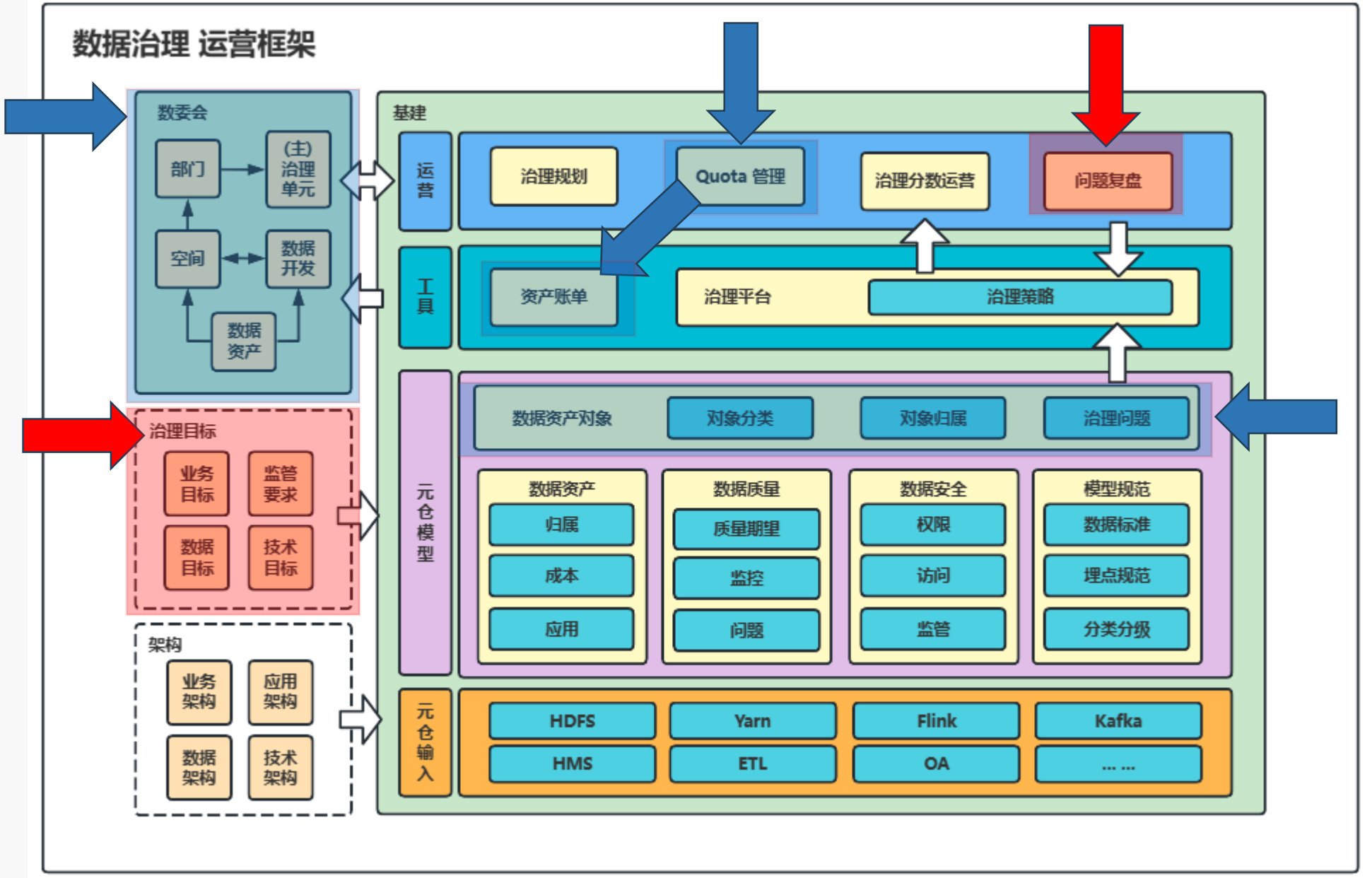






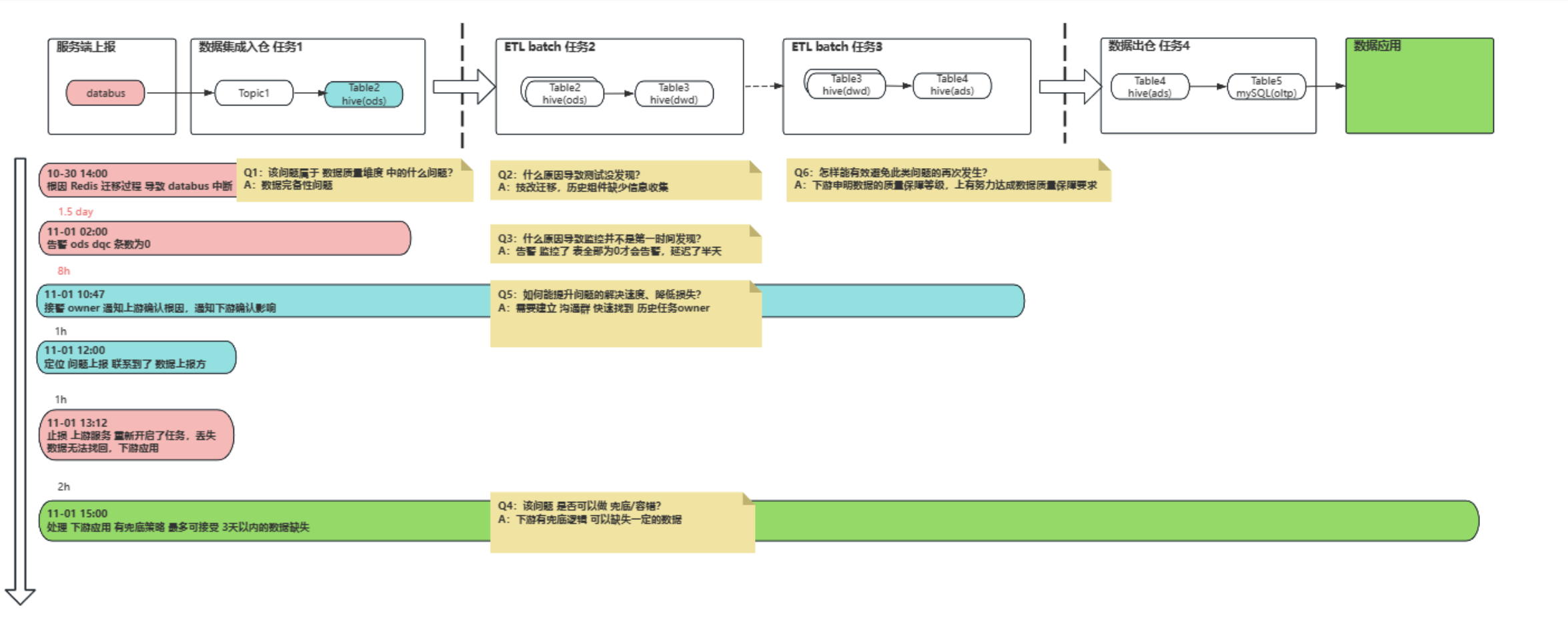
* 关于“指标 标签”与“道行数知名”




数据治理 运营框架
数委会
部门
(主)治理单元
空间
数据开发
数据资产
治理目标
业务目标
监管要求
数据目标
技术目标
架构
业务架构
应用架构
数据架构
技术架构
基建
运营
治理规划
Quota 管理
治理分数运营
问题复盘
工具
资产账单
治理平台
治理策略
元仓模型
数据资产对象
对象分类
对象归属
治理问题
数据资产
归属
成本
应用
数据质量
质量期望
监控
问题
数据安全
权限
访问
监管
模型规范
数据标准
埋点规范
分类分级
元仓输入
HDFS
Yarn
Flink
Kafka
HMS
ETL
OA
... ...

案例2 -- 2023-10-30 数据丢失复盘

案例背景（2线6问）

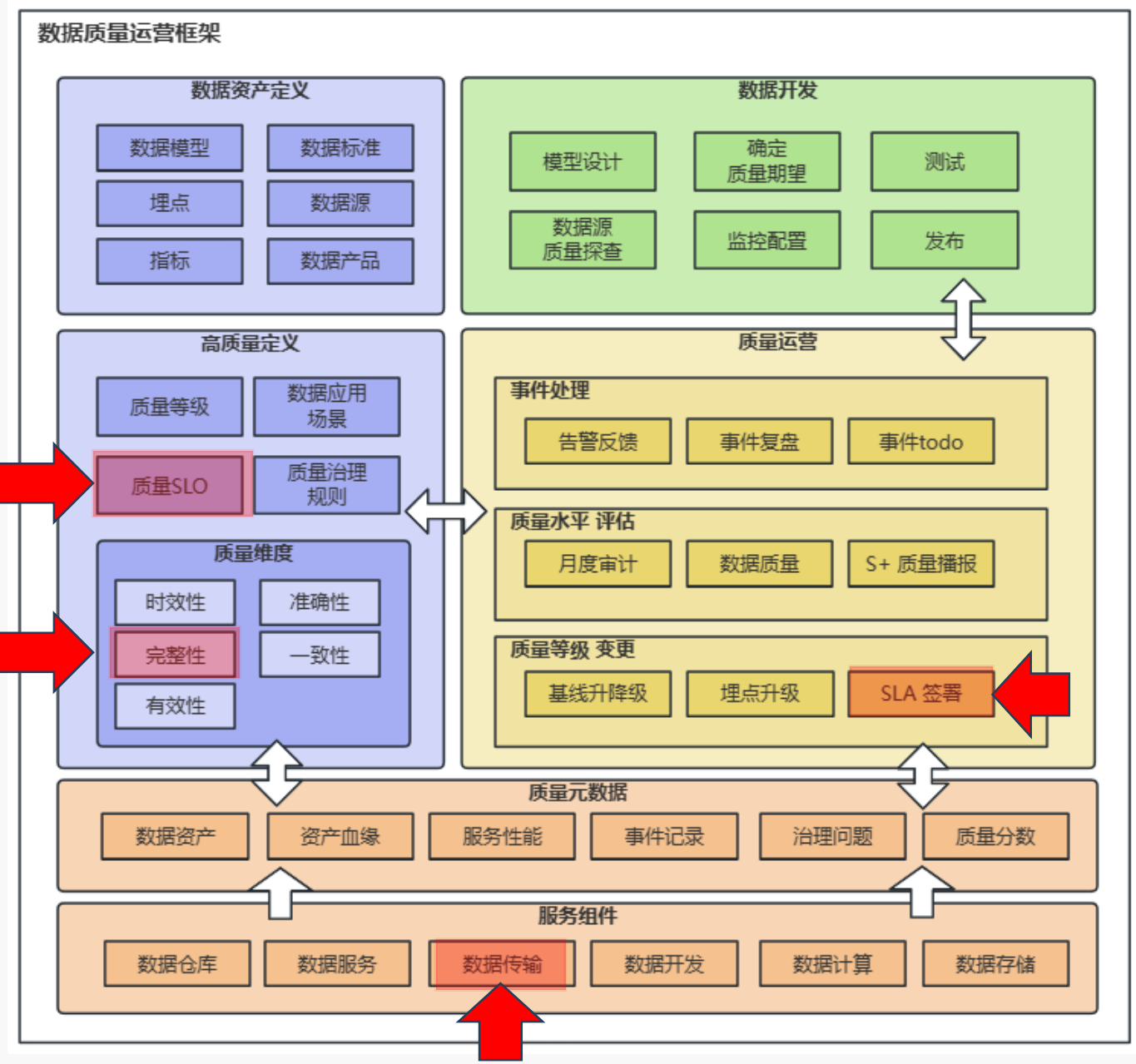

服务端上报
databus
数据集成入仓 任务1
Topic1
Table2 hive(ods)
ETL batch 任务2
Table2 hive(ods)
Table3 hive(dwd)
ETL batch 任务3
Table3 hive(dwd)
Table4 hive(ads)
数据出仓 任务4
Table4 hive(ads)
Table5 mySQL(oltp)
数据应用
10-30 14:00
根因 Redis 迁移过程 导致 databus 中断
Q1：该问题属于 数据质量维度 中的什么问题?
A：数据完备性问题
Q2：什么原因导致测试没发现?
A：技改迁移，历史组件缺少信息收集
Q6：怎样能有效避免此类问题的再次发生?
A：下游申明数据的质量保障等级，上有努力达成数据质量保障要求
1.5 day
11-01 02:00
告警 ods dqc 条数为0
Q3：什么原因导致监控并不是第一时间发现?
A：告警 监控了 表全部为0才会告警，延迟了半天
8h
11-01 10:47
接警 owner 通知上游确认根因，通知下游确认影响
Q5：如何能提升问题的解决速度、降低损失?
A：需要建立 沟通群 快速找到 历史任务owner
1h
11-01 12:00
定位 问题上报 联系到了 数据上报方
1h
11-01 13:12
止损 上游服务 重新开启了任务，丢失
数据无法找回，下游应用
2h
11-01 15:00
处理 下游应用 有兜底策略 最多可接受 3天以内的数据缺失
Q4：该问题 是否可以做 兜底/容错?
A：下游有兜底逻辑 可以缺失一定的数据

问题与挑战

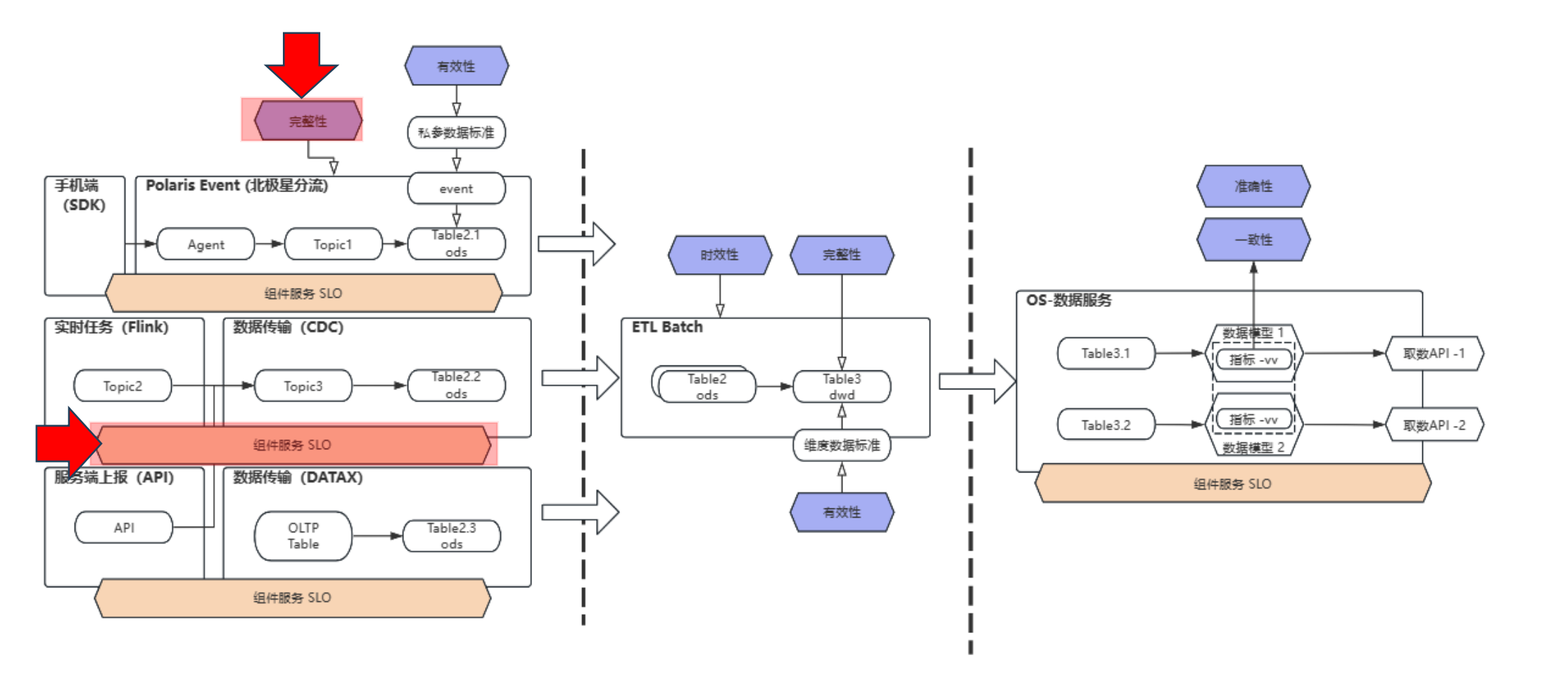
	问	答	DAMA Bok
Q1	该问题属于 数据质量维度 中的什么问题?	数据完整性问题	• 原则：数据管理即使对数据的质量进行管理 • 领域：数据质量
Q2	什么原因导致测试没发现?	技改迁移，历史组件缺少信息收集	• 原则：数据管理需要元数据 • 领域：数据架构
Q3	什么原因导致监控并不是第一时间发现?	告警 监控了，表全部为空才会告警，延迟了半天	• 原则：数据价值使用经济术语表达 • 领域：数据质量
Q4	该问题 是否可以做 兜底/容错?	下游有兜底逻辑 可以缺失一定的数据	• 原则：数据管理需要全景视角 • 领域：数据质量
Q5	如何能提升问题的解决速度、降低损失?	数据提供方 应该感知下游数据应用的重要等级	• 原则：数据管理需要全景视角 • 领域：数据质量
Q6	怎样能有效避免此类问题的再次发生?	数据服务方 应该感知下游数据应用的重要等级	• 原则：数据管理是跨功能的 • 领域：数据质量

破题思路

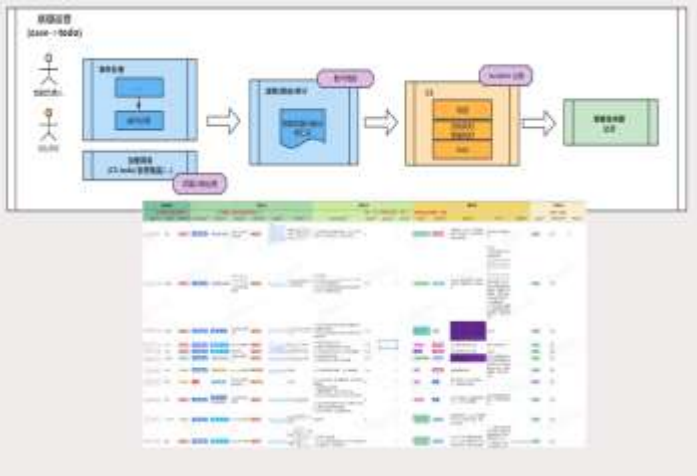
问题	DAMA Bok	方案
历史组件缺少信息收集?	• 原则：数据管理需要元数据 • 领域：数据架构	• 各类组件都应具备自回收能力，基于该能力补充需要的元数据与功能更流程
能否不等表全部为空就可以发出告警	• 原则：数据价值使用经济术语表达 • 领域：数据质量	• 数据传输链路中需要 增加完整性监控方案
是否下游都有兜底逻辑?	• 原则：数据管理需要全景视角 • 领域：数据质量	• 重要数据下游使用方需要明确提供数据质量的问题识别与容忍区间
数据提供方如何感知下游数据的重要性?	• 原则：数据管理需要全景视角 • 领域：数据质量	• 数据下游的使用场景，质量容忍度需要传递给数据提供方
数据服务方是否提供了足够的质量保障?	• 原则：数据管理是跨功能的 • 领域：数据质量	• 数据服务方需要定期进行质量服务水平的审计


数据质量运营框架
数据资产定义
数据模型
数据标准
埋点
数据源
指标
数据产品
数据开发
模型设计
确定
质量期望
测试
数据源
质量探查
监控配置
发布
高质量定义
质量等级
数据应用
场景
质量SLO
质量治理
规则
质量维度
时效性
准确性
完整性
一致性
有效性
质量运营
事件处理
告警反馈
事件复盘
事件todo
质量水平 评估
月度审计
数据质量
S+ 质量播报
质量等级 变更
基线升降级
埋点升级
SLA 签署
质量元数据
数据资产
资产血缘
服务性能
事件记录
治理问题
质量分数
服务组件
数据仓库
数据服务
数据传输
数据开发
数据计算
数据存储

数据质量监控

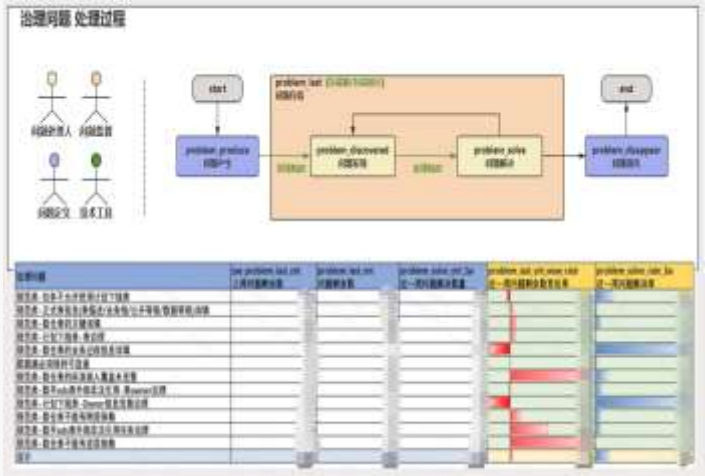

有效性
完整性
私参数据标准
手机端
(SDK)
Polaris Event (北极星分流)
event
Agent
Topic1
Table2.1
ods
组件服务 SLO
实时任务（Flink）
数据传输（CDC）
Topic2
Topic3
Table2.2
ods
组件服务 SLO
服务端上报（API）
数据传输（DATAX）
API
OLTP
Table
Table2.3
ods
组件服务 SLO
时效性
完整性
ETL Batch
Table2
ods
Table3
dwd
维度数据标准
有效性
准确性
一致性
OS-数据服务
Table3.1
数据模型 1
指标 -vv
取数API -1
Table3.2
指标 -vv
数据模型 2
取数API -2
组件服务 SLO

从 CS 到 TODO



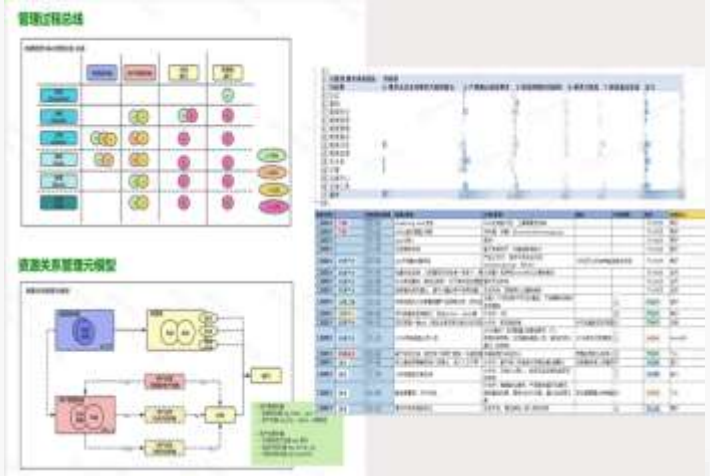
事前 分析问题

通过 各种渠道收集问题
通过 复盘寻找问题处理方法
通过 治理策略、开发基建 控制问题的影响



事中 处理问题

需要观察发布的治理项目是否有在被顺利执行
需要观察问题新增速度是否远远大于处理速度
需要评估使用人力参与数治理的工作是否合理



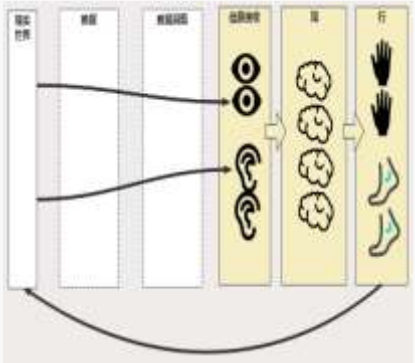
事后 沉淀基建

开发 元数据增强 数据管理能力
开发 治理策略 提高问题处理 效率
开发 基建降低问题产生的几率

TakeAway

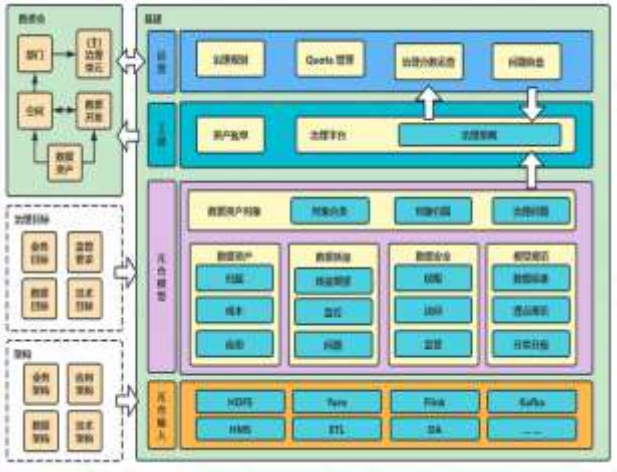
DAMA 车轮图

11 个数据管理知识领域



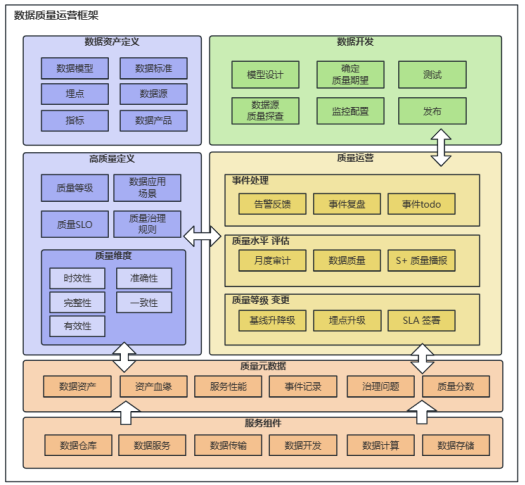
道-行-数-知-名

分析数据问题往往
就是分析管道问题



数据治理运营框架

实施数据治理需要操作系统



数据质量运营框架

处理 复盘 执行 沉淀

TOP1

主办方 msup

微信官方公众号：壹佰案例
关注查看更多年度实践案例

微信官方公众号：哔哩哔哩技术
关注查看更多B站技术分享