

淘宝数据治理介绍

郭进士

自我介绍

郭进士 高级数据技术专家

- 15年加入阿里巴巴，当前是淘宝数据平台负责人
- 参与淘宝、零售通、国际化、天猫精灵等多业务数据架构设计治理
- 关注大模型时代数据平台的演变升级

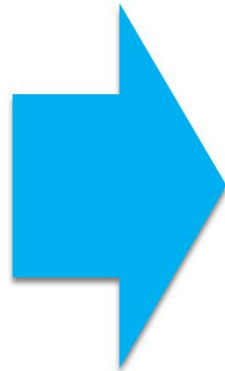
目录

- 治理背景
- 成本治理
- 模型治理
- 稳定性治理

数据治理背景

背景

- 1 组织策略强调降本增效
- 2 强竞争形态诉求更高效率
- 3 组织上浮数据保障标准需要统一



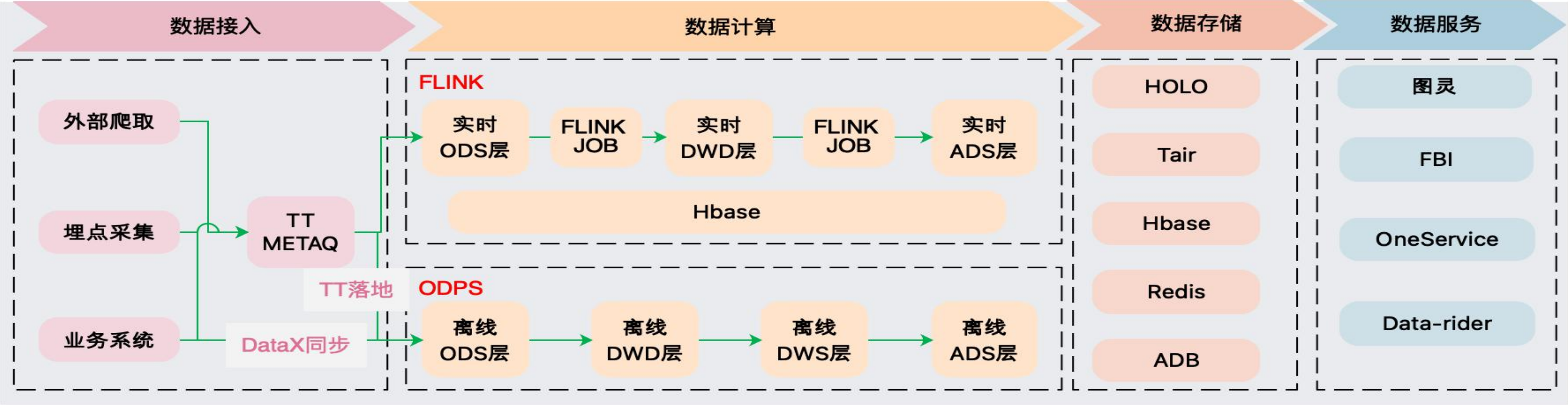
策略

- 成本治理
- 模型治理
- 稳定性治理

目录

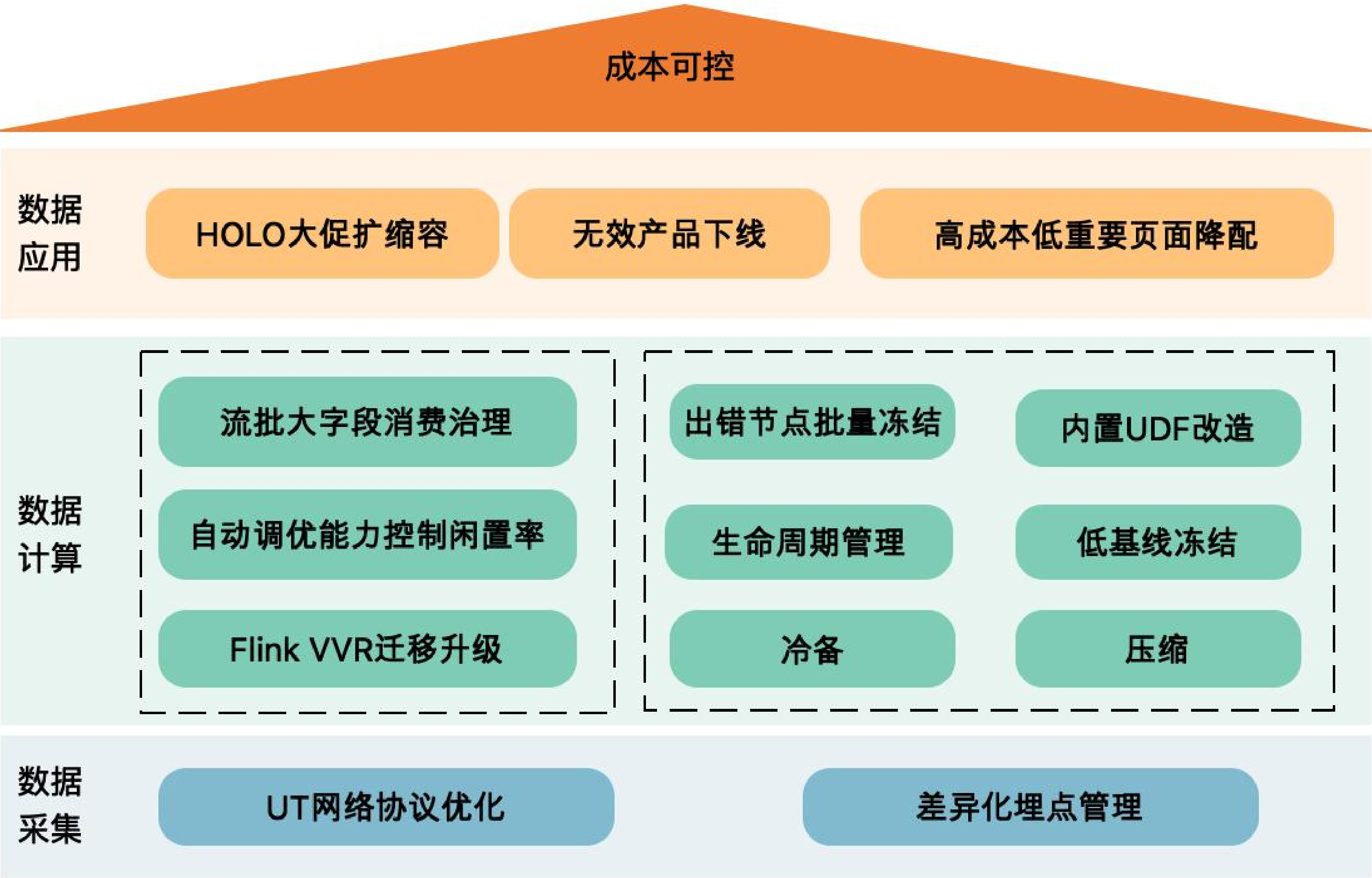
- 治理背景
- 成本治理
- 模型治理
- 稳定性治理

成本治理机会



- 网络压缩协议未更新以满足新埋点
- 埋点只上不下永久保存
- 实时: Flink集团版本非最新版本, 性能较弱; Flink Top任务消耗较高; 实时资源闲置率较高, 存在资源浪费;
- 离线: TOP表存储有压缩空间; 大量表未根据实际使用做存储生命周期管理; 任务只上不下; 冷数据未及时归档备份;
- HOLO: 基于额度计费, 大促波峰波谷利用率不高
- 数据产品: 不计成本追求时效性

成本治理方案



数据采集：

- UT网络协议优化，升级压缩算法、映射字典升级；
- 差异化埋点管理，将性能和算法埋点标准化后，大促降级、差异化存储策略

数据计算：

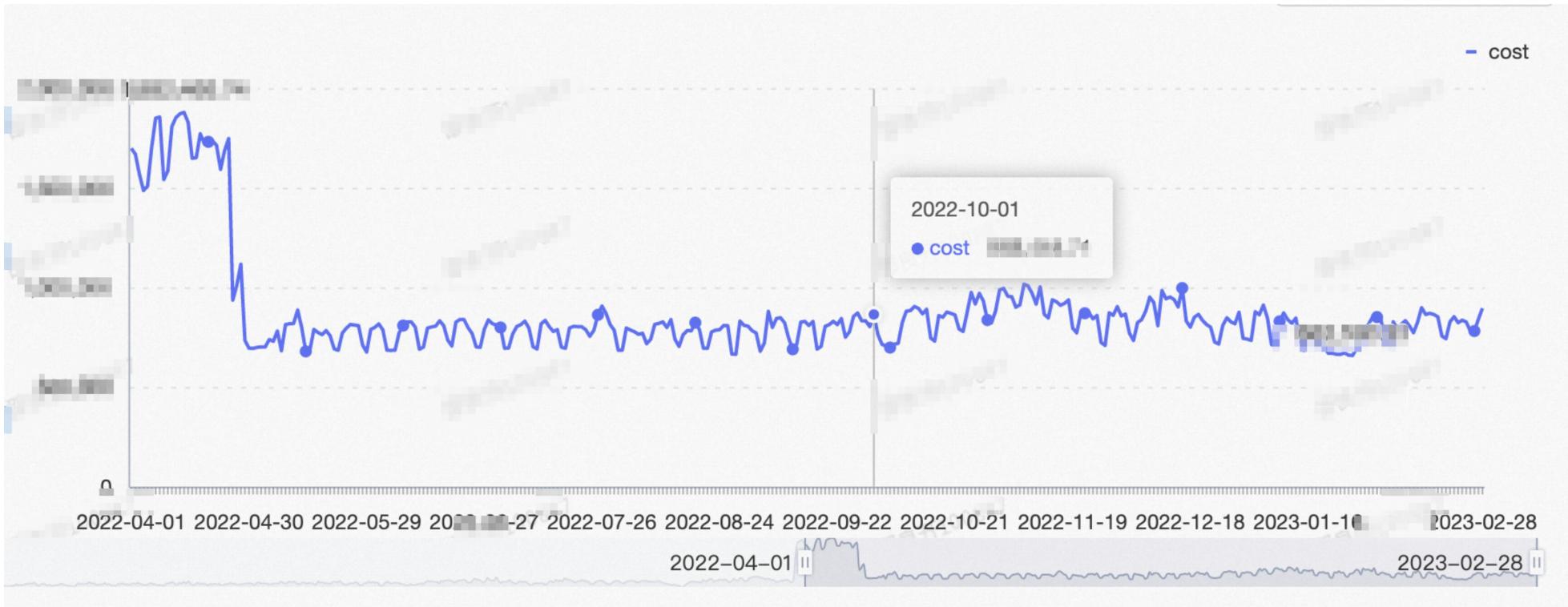
- 实时链路，升级flink版本提升引擎性能，引入自动参数调优能力控制任务闲置率，共性流任务合并消费治理降低重复消费；
- 离线链路，识别冷数据进入冷备、对大表进行重排压缩、基于消费调用进行生命周期治理、低重要任务的批量冻结、无人维护任务冻结

数据应用：

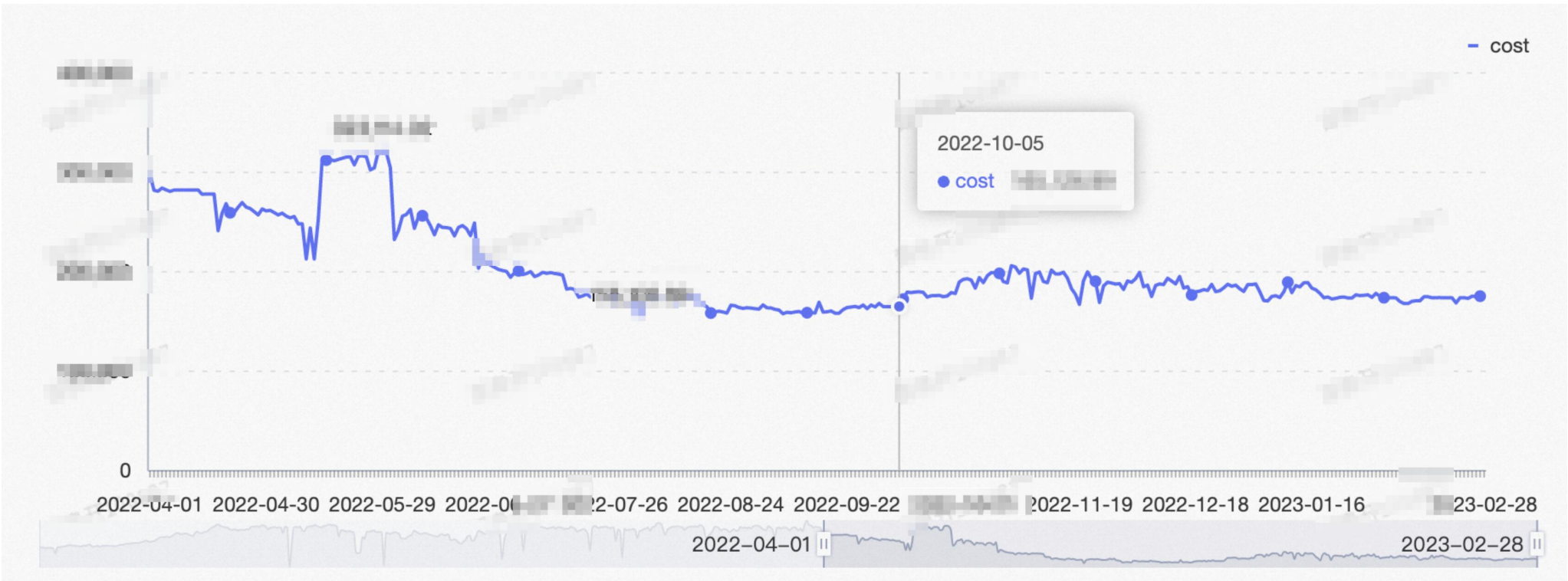
- HOLO，基于使用水位进行动态扩缩容；
- 数据产品，无效产品页面全链路下线、业务合理性的实时页面转离线或小时

成本治理效果

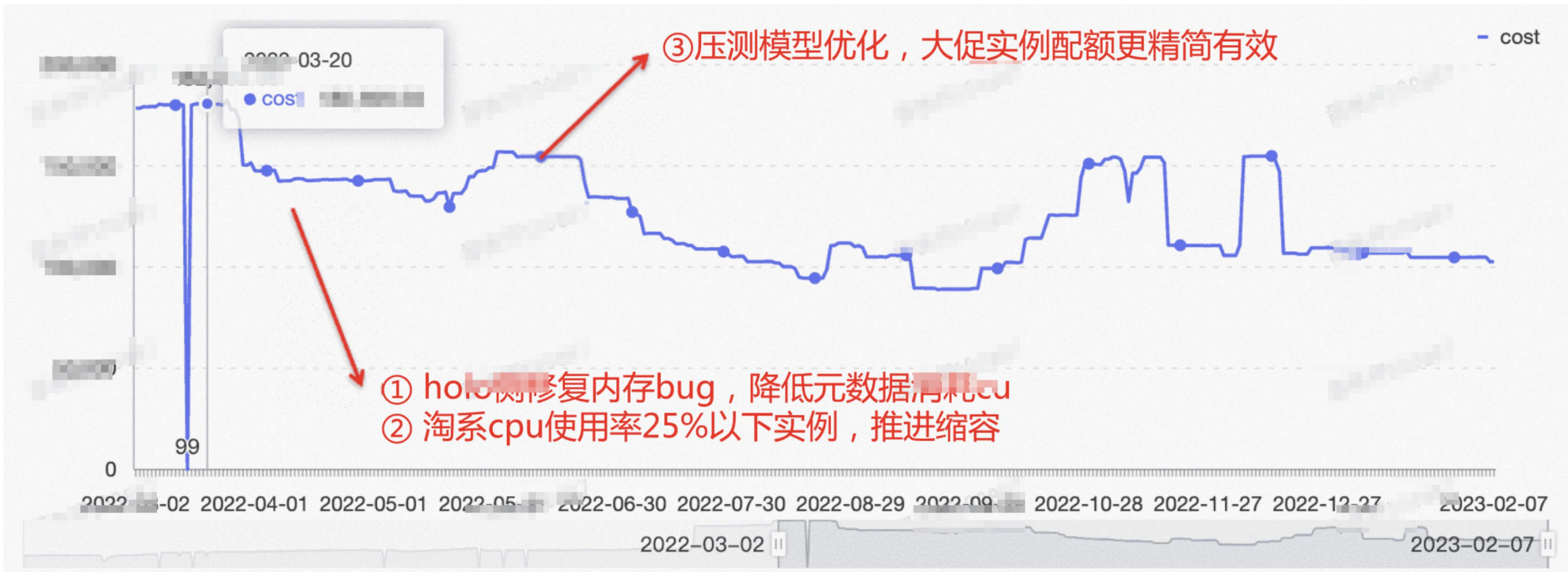
离线治理效果



实时治理效果



HOLO治理效果



成本治理思考

- 80%的成本治理收益通过技术手段实现的；
- 治理ROI的考量需要对数据分级治理；
- 业务合理性的成本治理才能确保数仓成本持续可控；

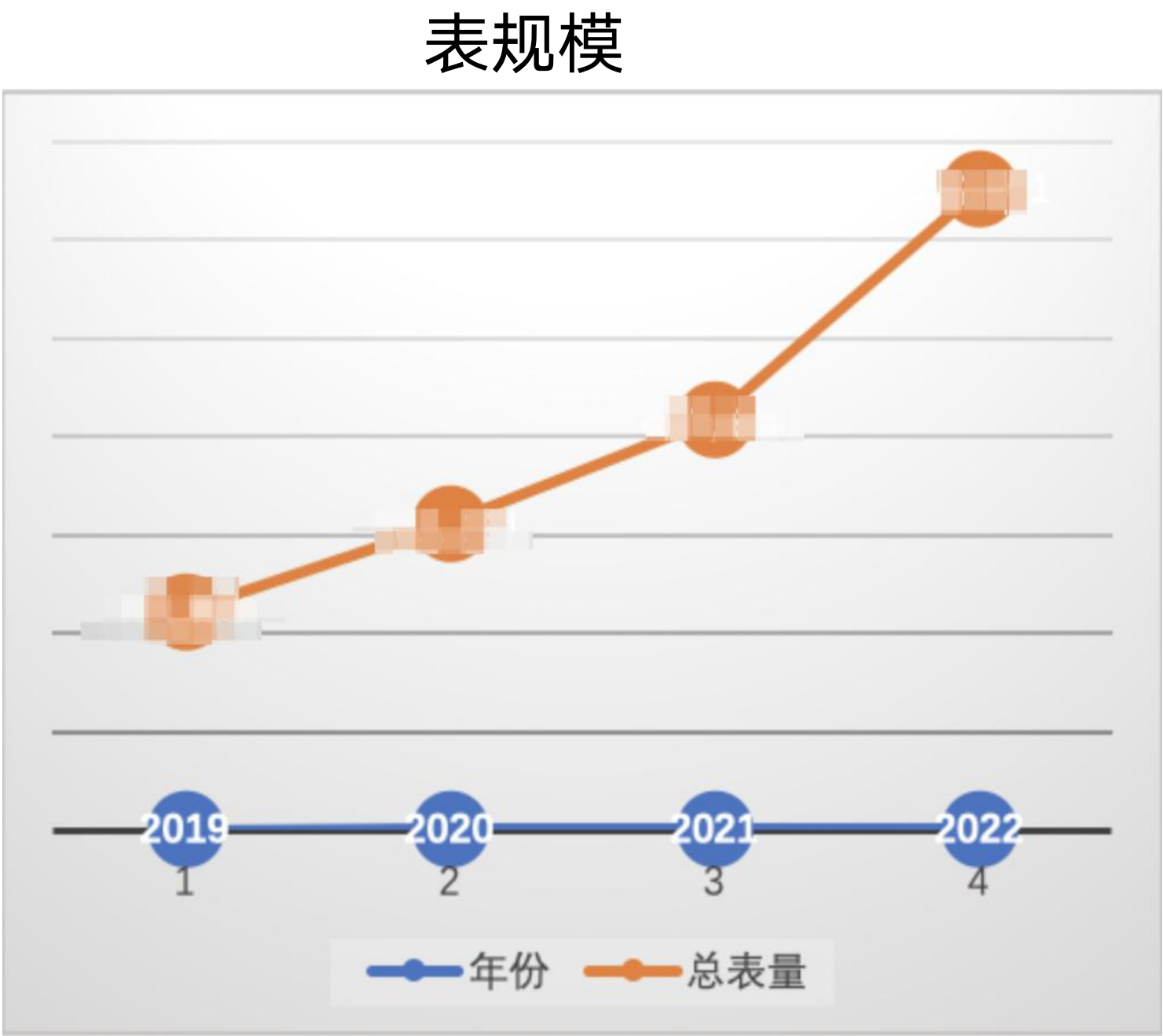
目录

- 治理背景
- 成本治理
- 模型治理
- 稳定性治理

模型治理问题-看规模

① 数据规模增长快

淘宝数据在2020年~2022年之间规模增长迅速



②无效表&无效节点占比较高

无效表占比较高，带来成本、运维和找数据效率问题

③大量表无人负责或非本团队负责

未归属表占比：16%，其中活跃表12%

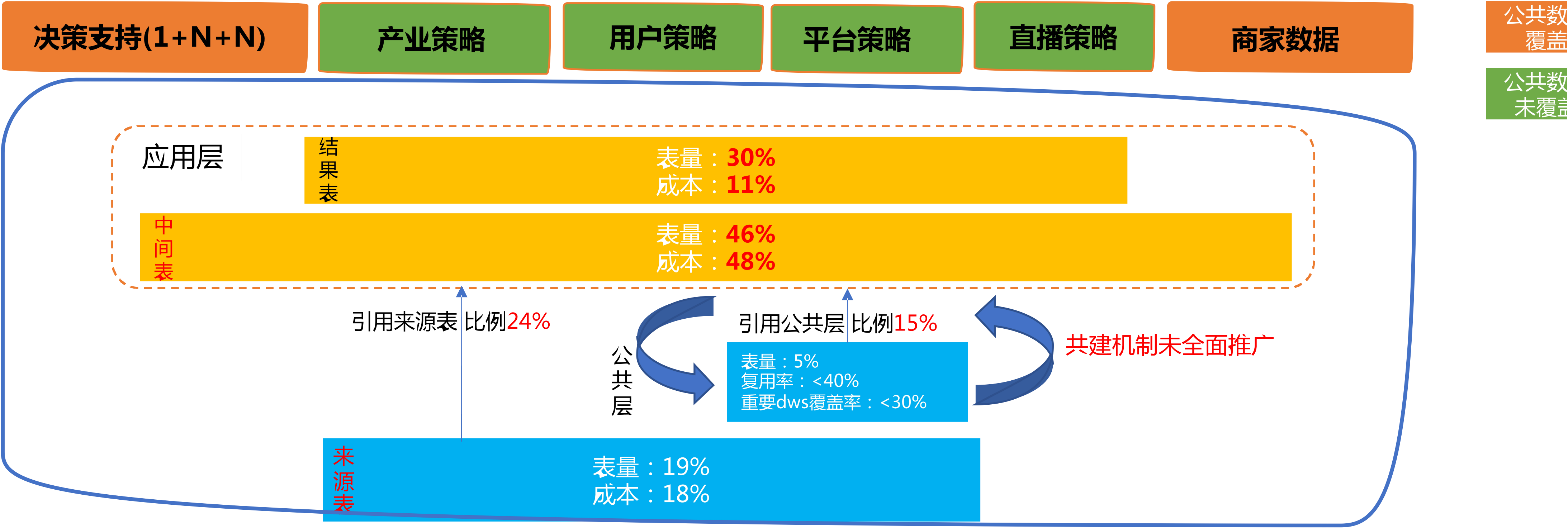
团队未归属人员类型分布

人员类型	占比
淘宝业务	56.9%
其他BU	22.7%
离职员工	10.4%
外包岗位	3.8%
其他	6.2%

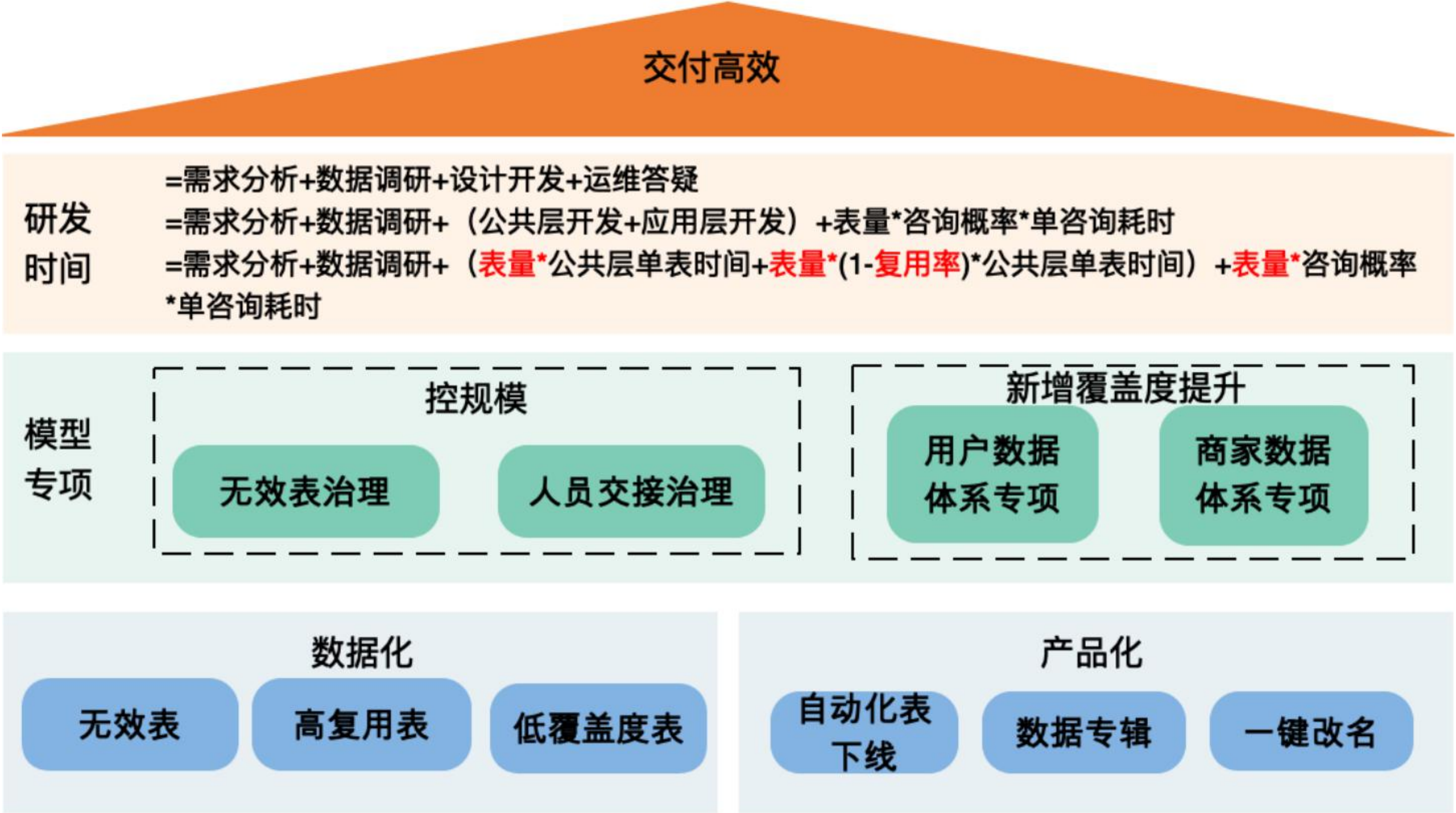
模型治理问题-看结构

公共层被引用不足，应用层自建大量中间表

- 公共层
复用率：
重要dws覆盖率：
- 应用层
存量-不足40%
新增-不足20%
覆盖率：15%
引用占比：ods-24% 公共层-15% 自建中间表占比：46%

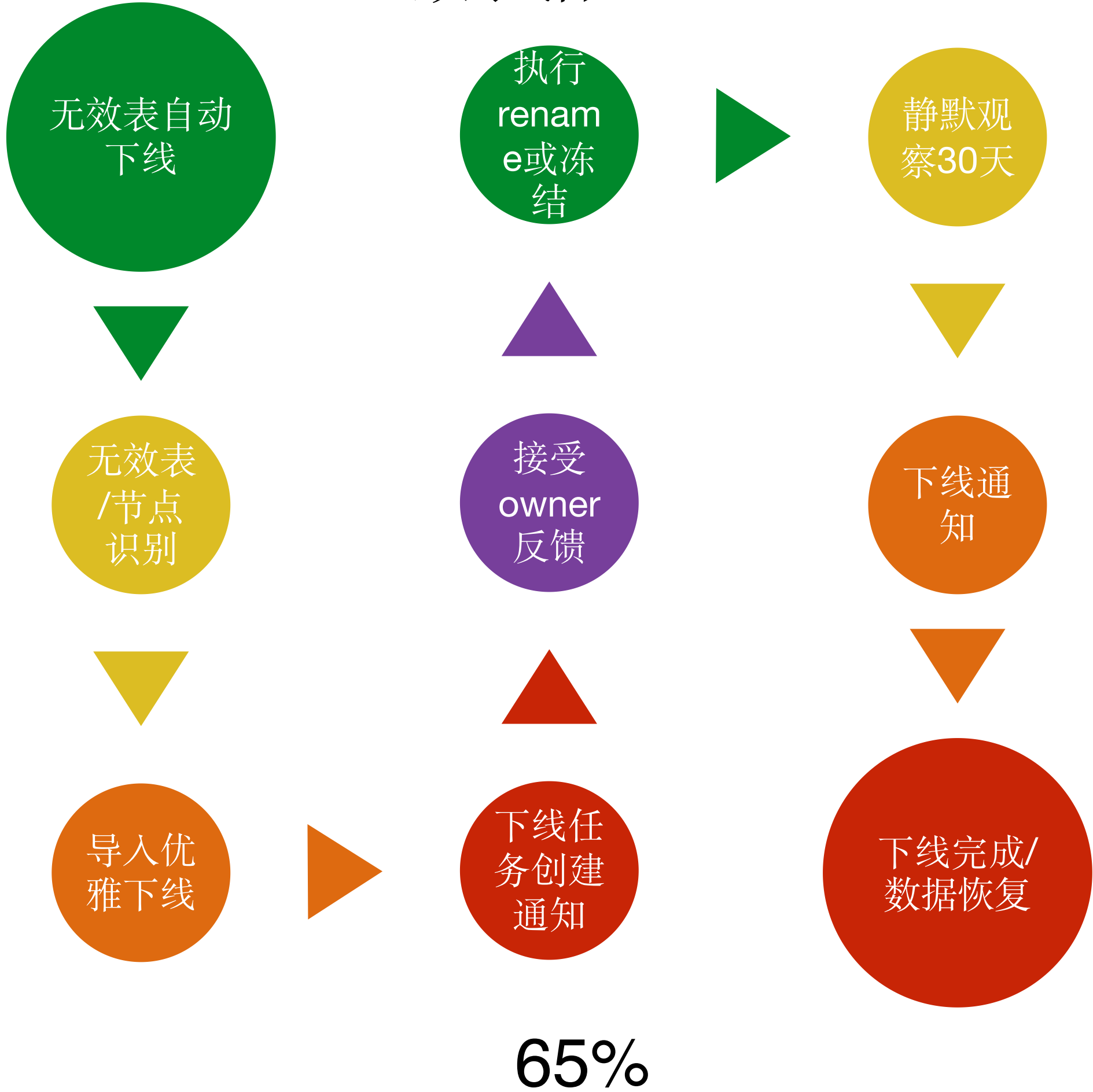


模型治理方案

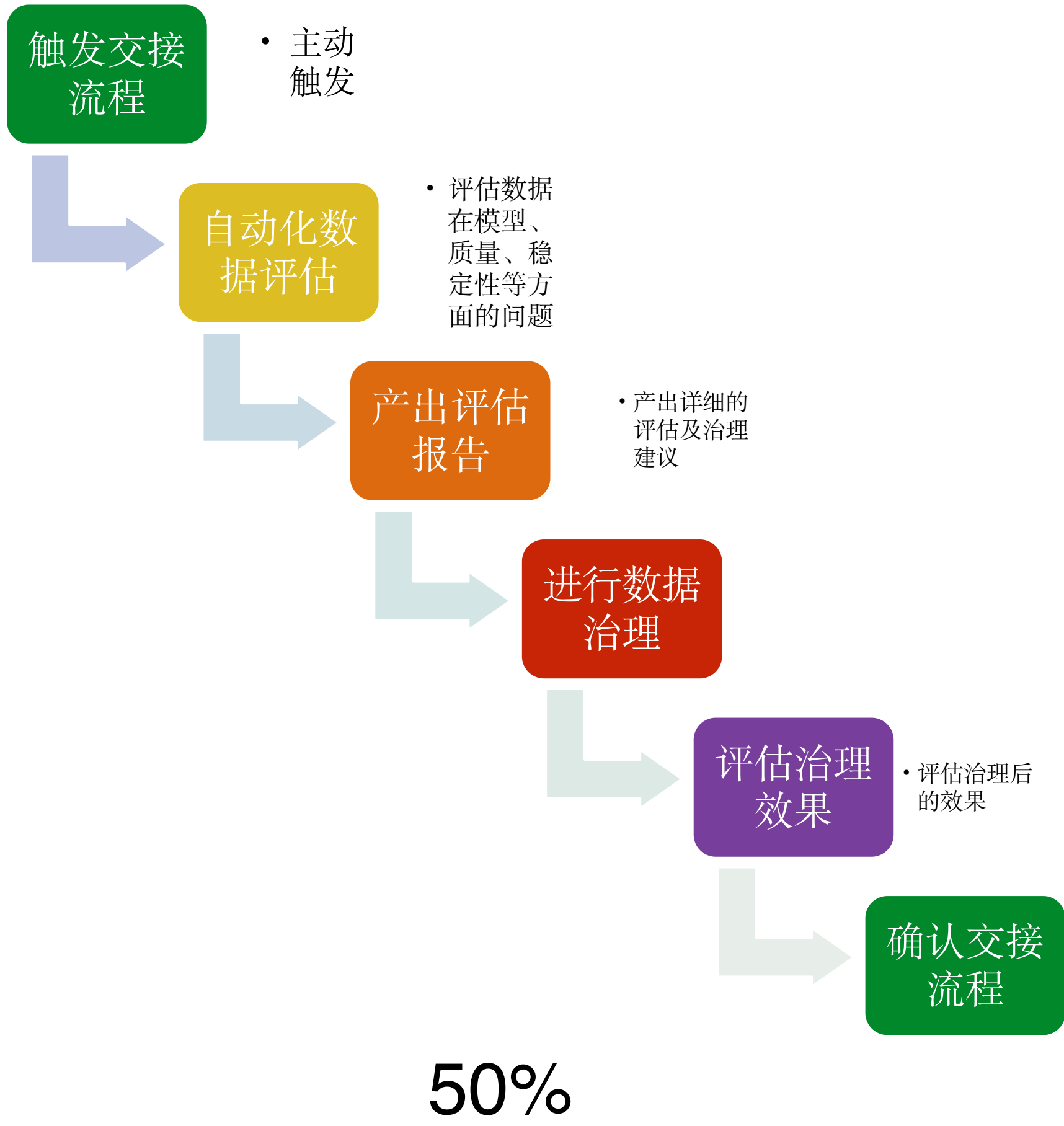


模型治理方案-控规模

无效表治理

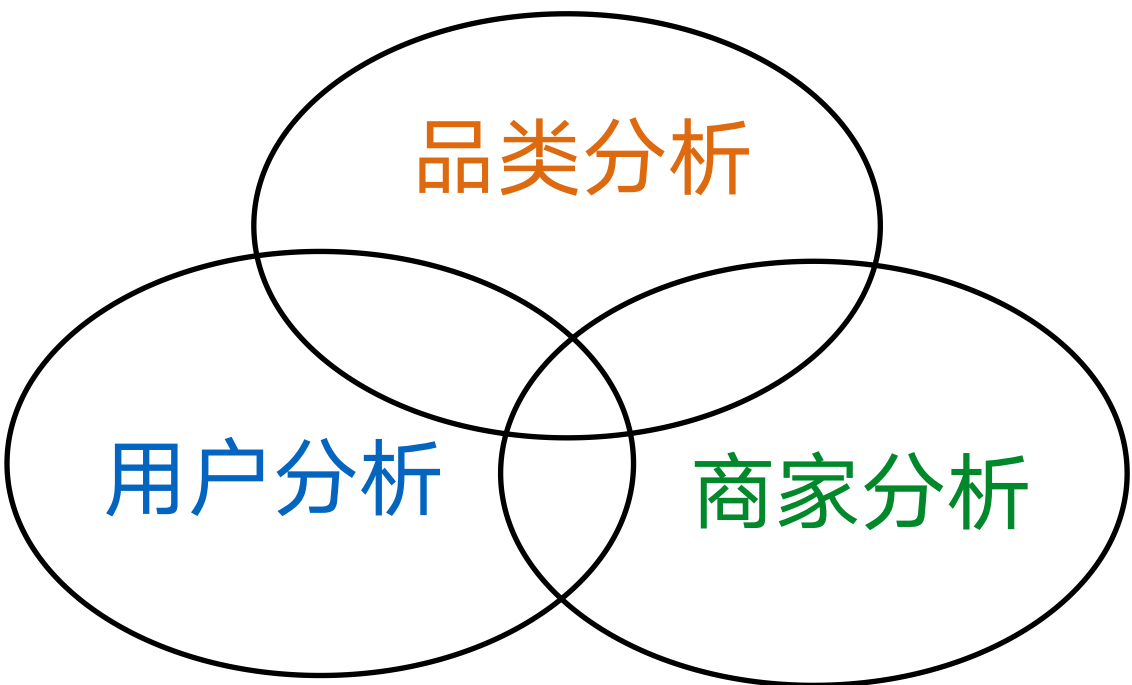


人员交接治理



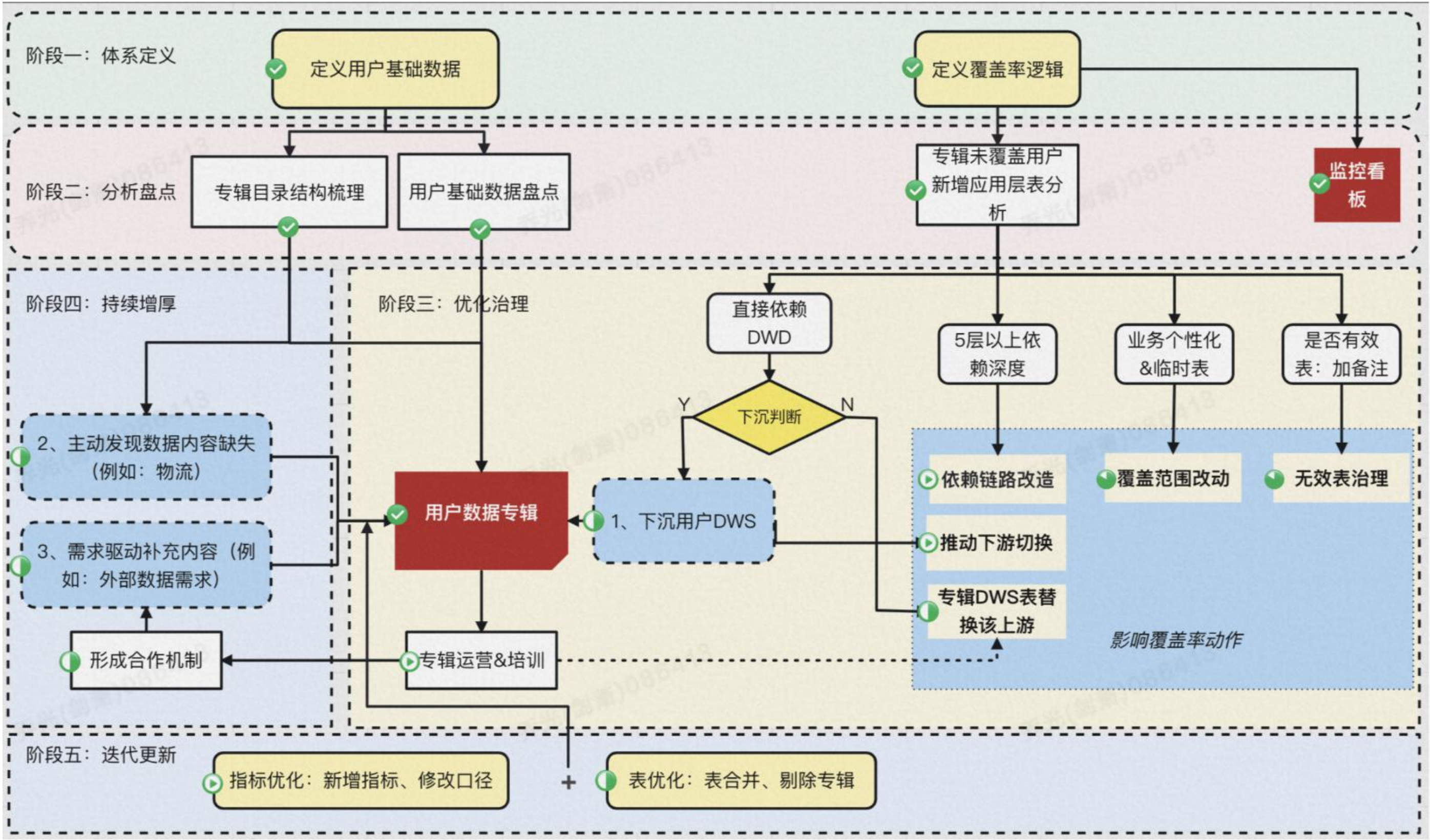
模型治理方案-覆盖度提升

TOP3场景分析



提升空间大，业务价值大

- 商家专项
6%→56%
- 用户专项
18%→63%
- 直播专项
- 短视频专项
-



模型治理思考

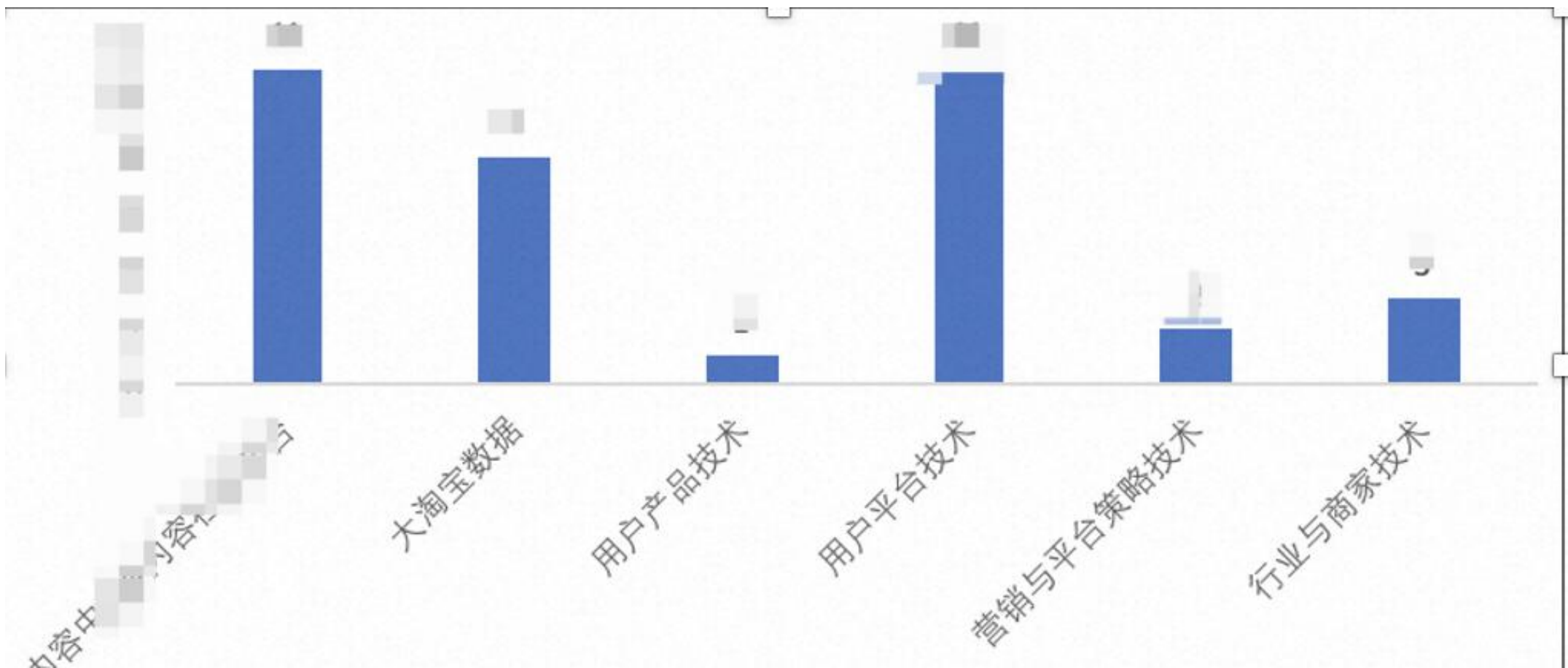
- 模型治理在于如何控制数据复杂度（表规模、表关系对规模）
- 产品化+数据驱动思维做数据主动运营，是提升公共层覆盖度的有效手段
- 湖仓一体的架构下，如何打造模型生命周期管控的机制

目录

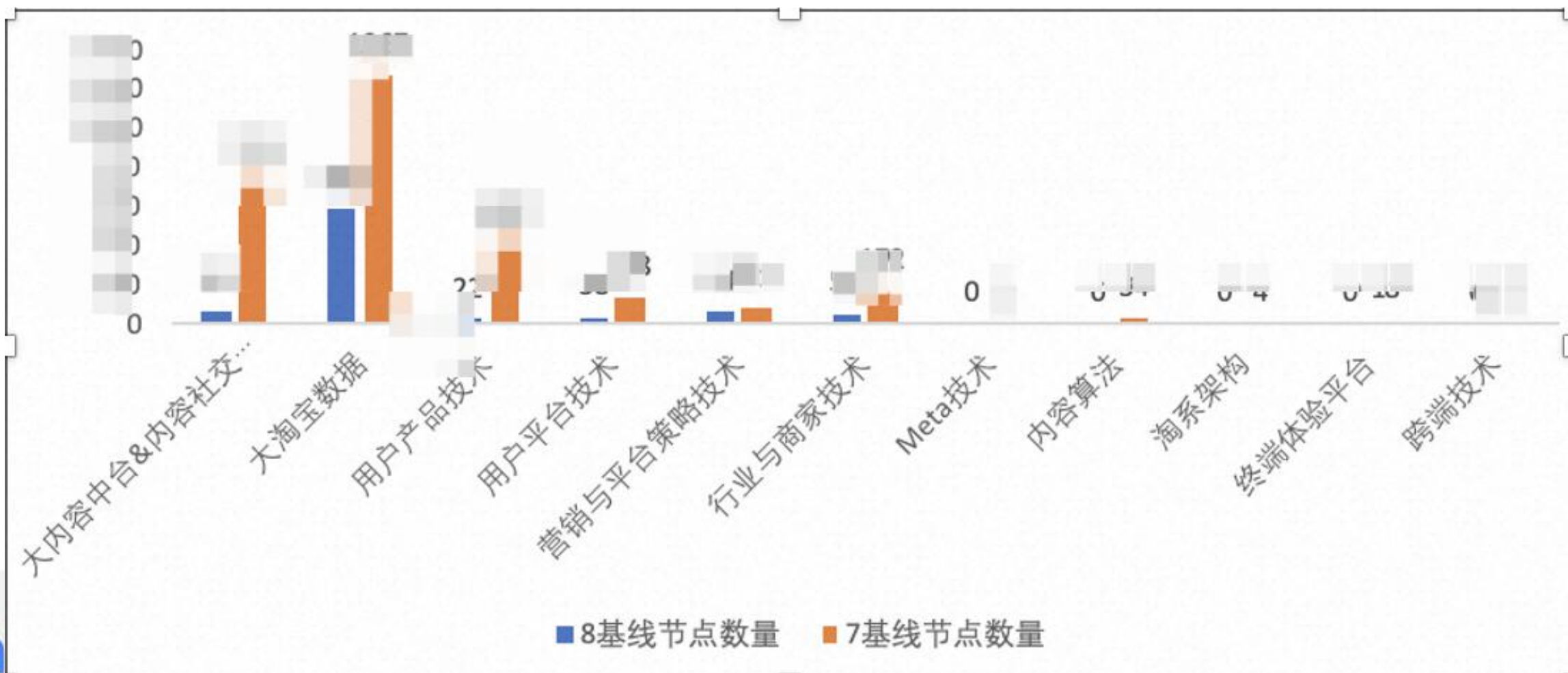
- 治理背景
- 成本治理
- 模型治理
- 稳定性治理

稳定性问题

高基线多、缺少准入规范



高基线节点数多、团队相对集中



重要基线稳定性问题突出严峻

基线等级	月破线情况	月节点预警量 (电话告警出错或变慢)
8	0	28
7	64	179
5	287	265
3	1084	97

稳定性保障方案

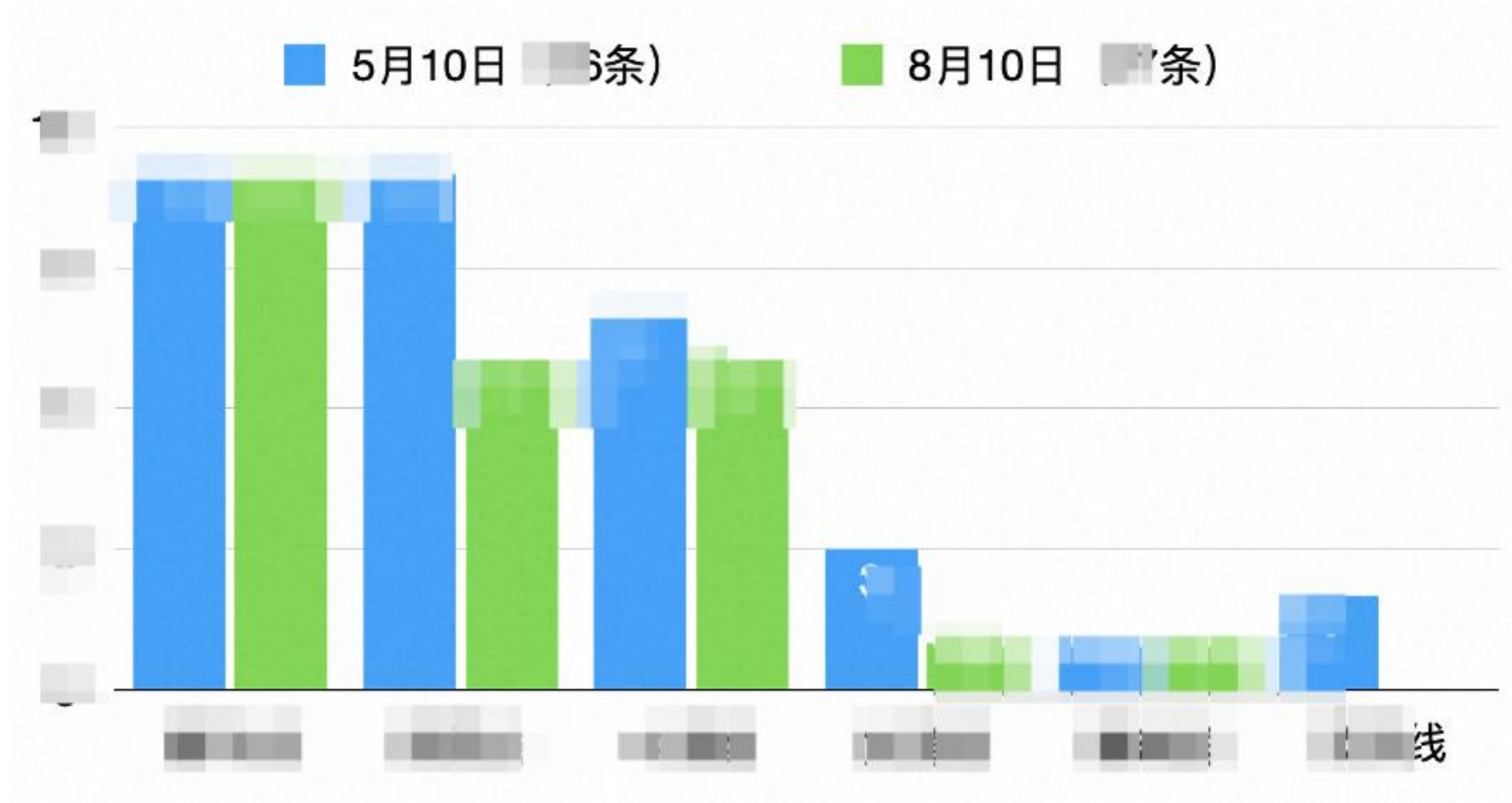


运维值班机制建设

- 值班机制说明
- 摩萨德使用说明
- 问题分析诊断方法
- 问题处理经验分享
- 大促运维经验分享
- 值班运维考试

稳定性保障效果

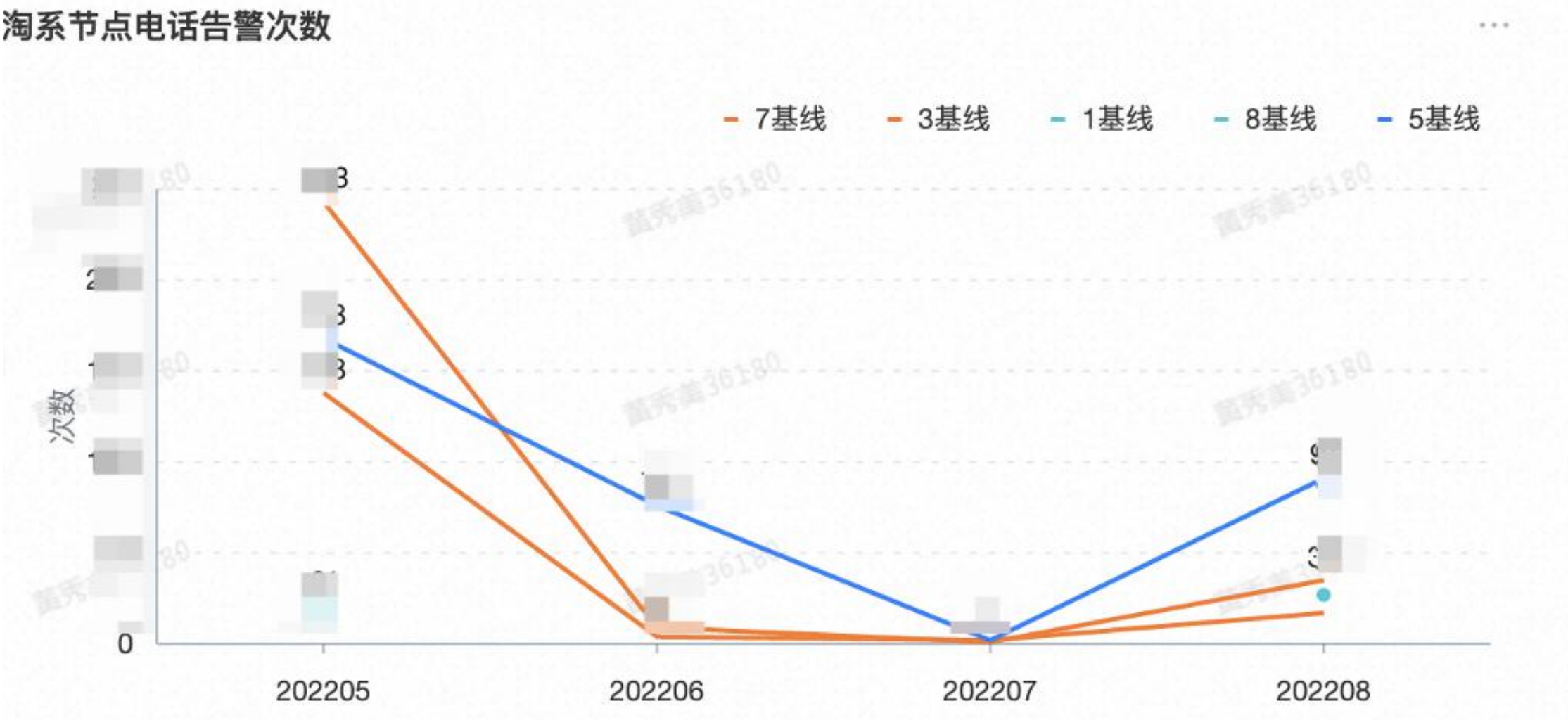
高基线数量对比



通过基线合并，无强时效要求基线降级等方式推进重保高基线合理性治理。治理后整体基线规范而内聚，整体高基线数下降30%。

治理后基线破线&预警下降明显

基线等级	治理前 基线破线情况	治理后 基线破线情况
8	0	0
7	64	21
5	287	204
3	1084	155



稳定性治理思考

- 稳定性在于在有限的机器资源下，确保核心应用的产出
- 成本和稳定性在一定程度下是相对立的，需要关注两者的平衡
- 稳定性的保障核心在于真正核心应用的识别，并基于稳定性问题不断迭代升级保障策略

数据治理的未来

- 架构升级：湖仓一体->Data mesh，分布式的数据治理？
- 治理效率：数据驱动治理->产品驱动治理->智能化？

想一想，我该如何把这些
技术应用在工作实践中？

THANKS