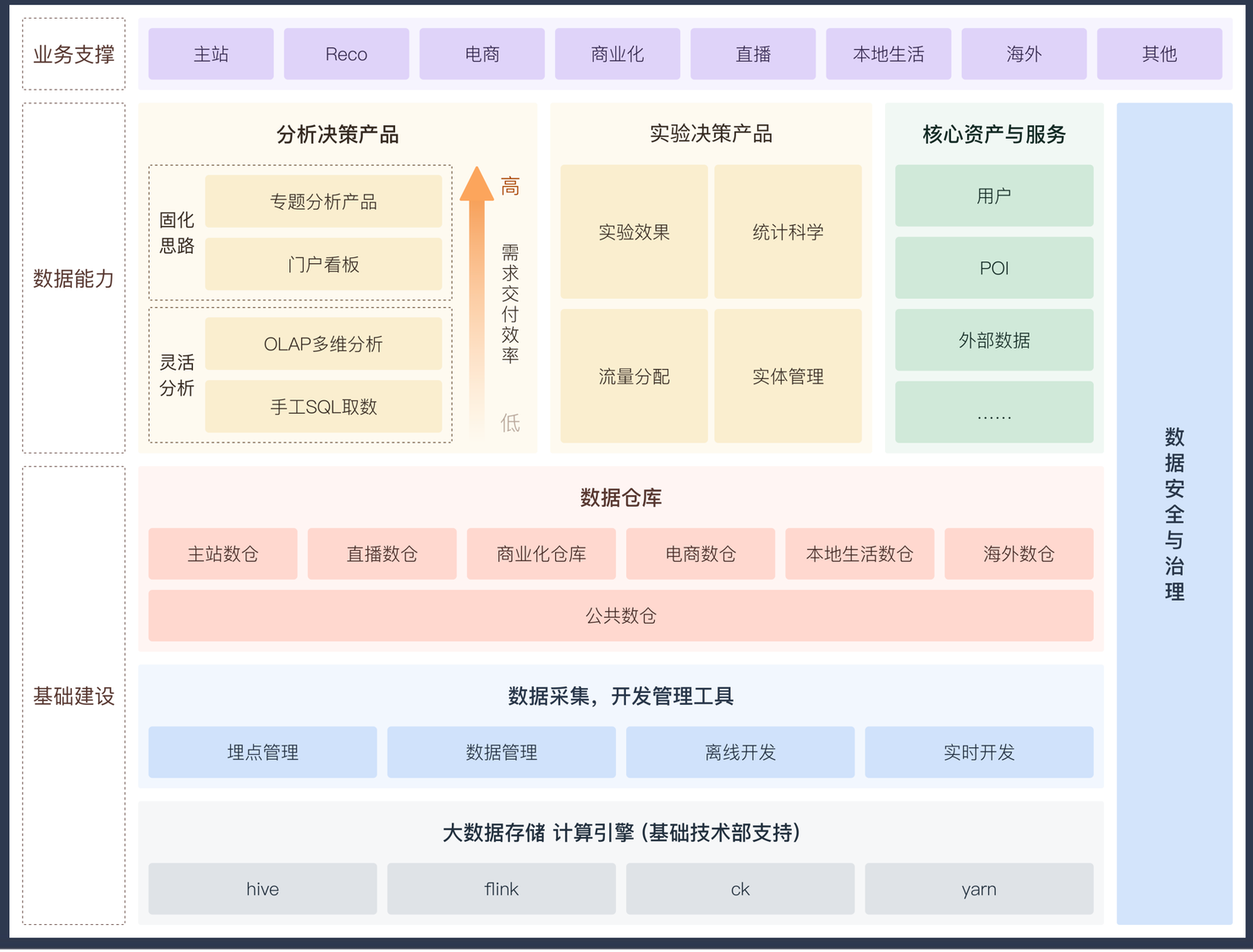
基于自动化治理实践驱动数据成本零增长

孙伟
快手 数据平台部 数据治理负责人&商业创新数据BP负责人

- 十多年大数据建设和应用经验，曾就职于百度、阿里，目前在快手负责数据治理和商业创新数据BP团队，专注打造高效的治理工具和可持续的治理机制，以及建设丰富易用的数据内容赋能业务

关于快手数据平台部



使命：提升数据决策效率，利用数据助力业绩提升

职责：通过大数据技术，对公司数据统一采集、存储、加工和挖掘形成高质量全域数据资产，以分析决策产品和服务的方式对外提供数据解决方案

集群规模	总数据量	日新增数据量	任务量
万级	EB级	PB级	十万级

目录

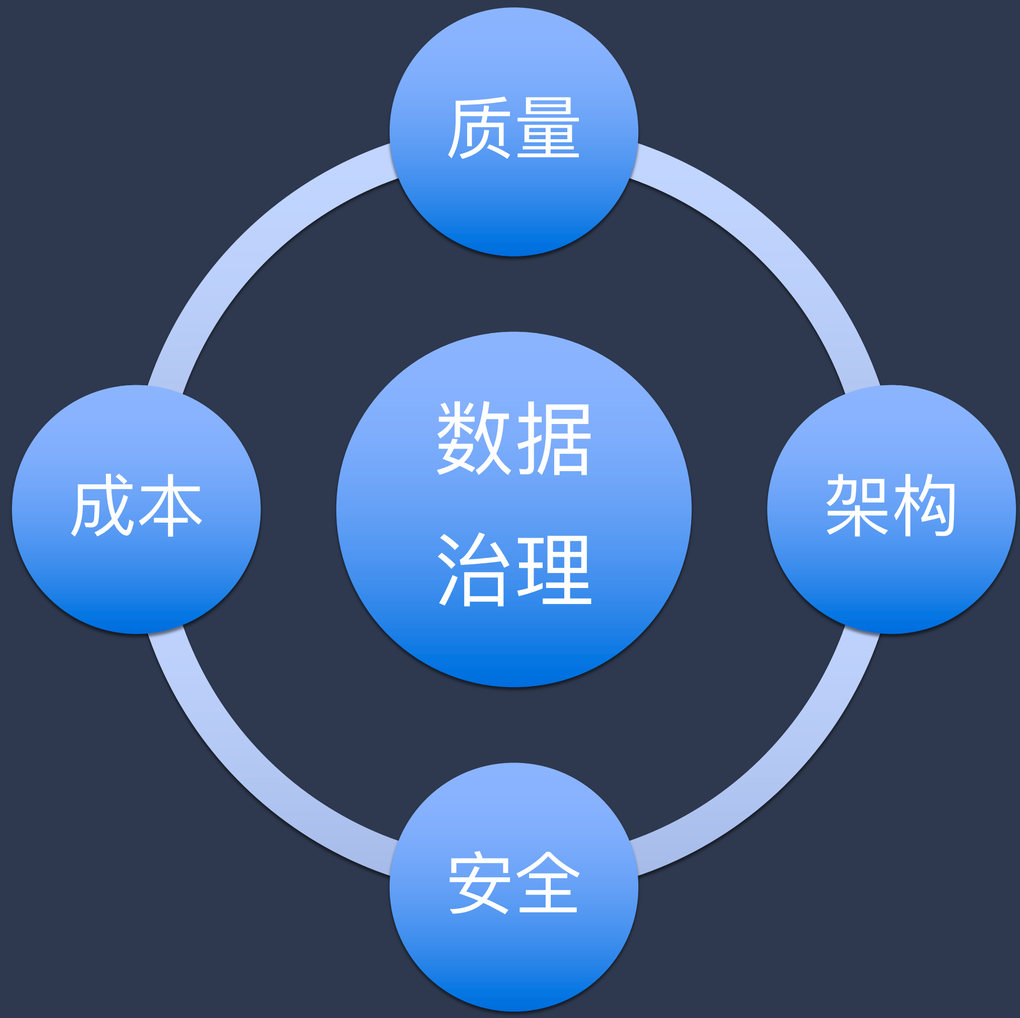
- 数据治理概述
- 成本治理方案
- 自动化治理实践
- 总结与展望

数据治理概述

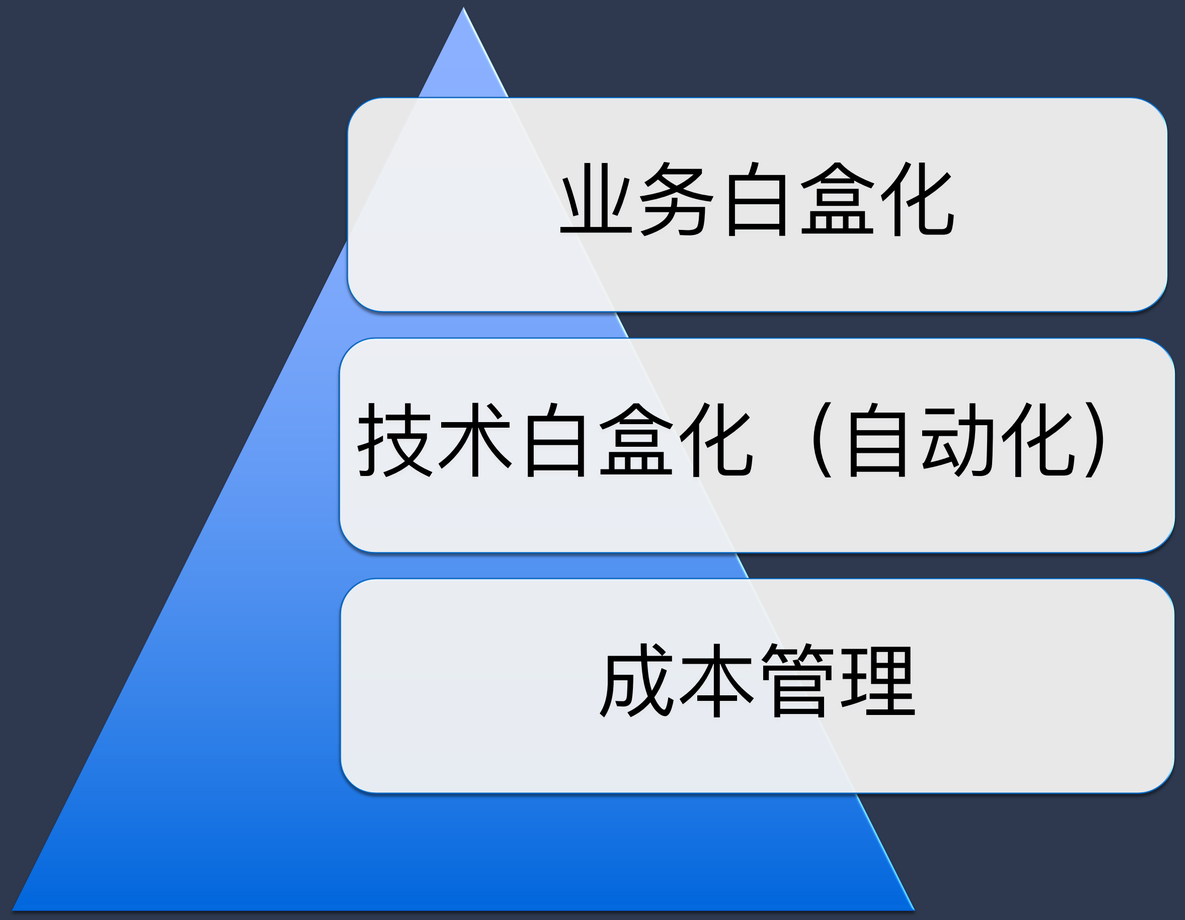
- **Why**：对抗大数据系统的熵增，让数据管理有序、可控，以及价值最大化

- **What**：保障数据质量，合理降低数据成本，守住安全红线，优化数据架构

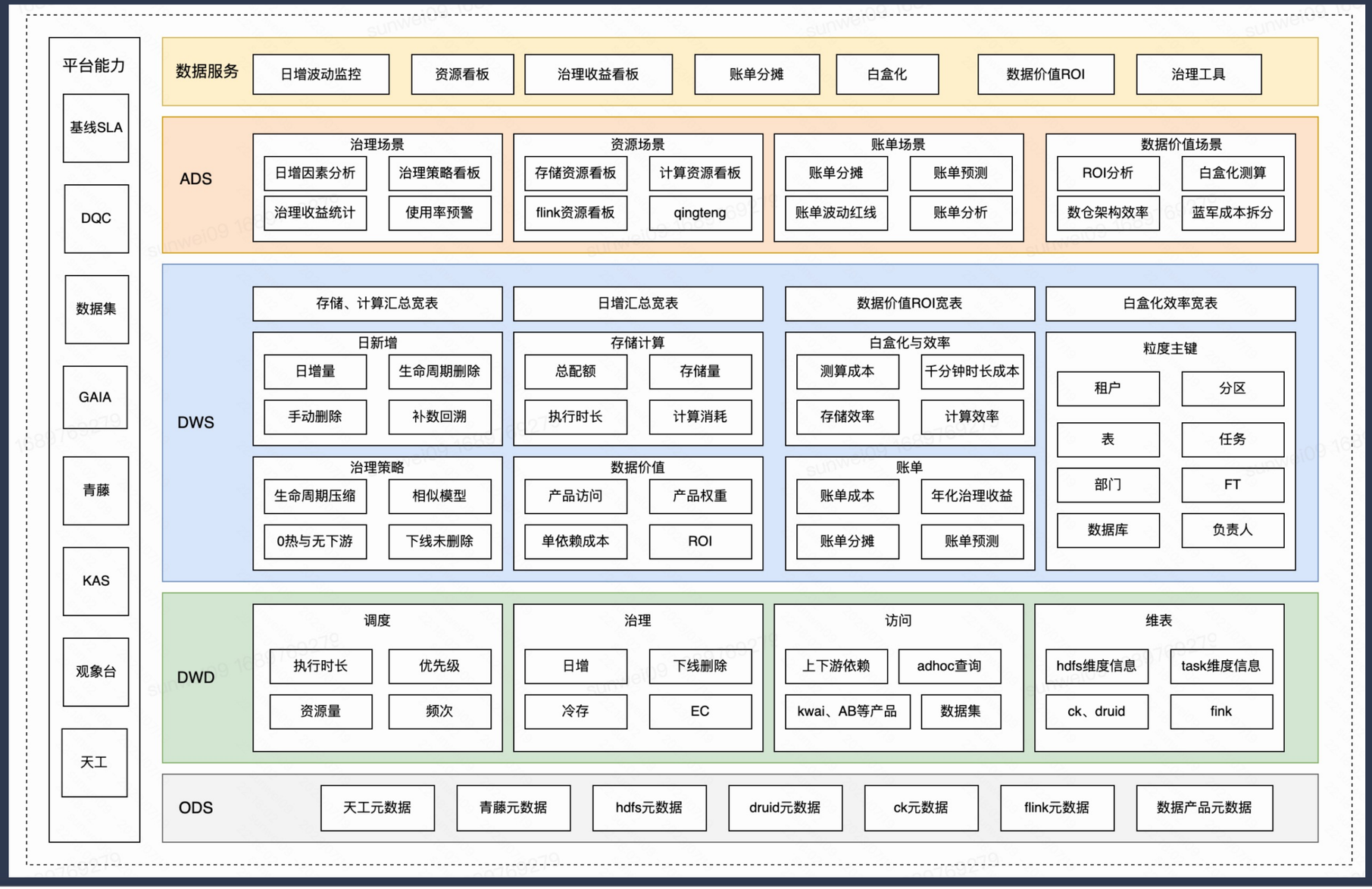
- **How**：管理+治理，通过有效的评估体系配合组织与流程机制，以及工具能力，驱动可持续治理



成本治理方案：思路



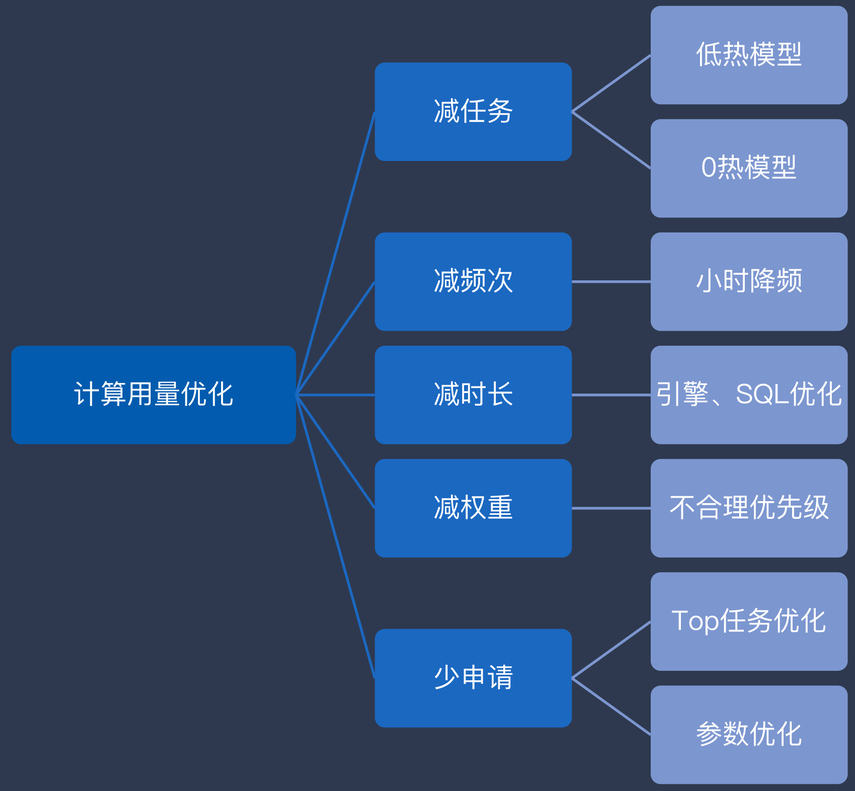
成本治理方案：成本元数仓



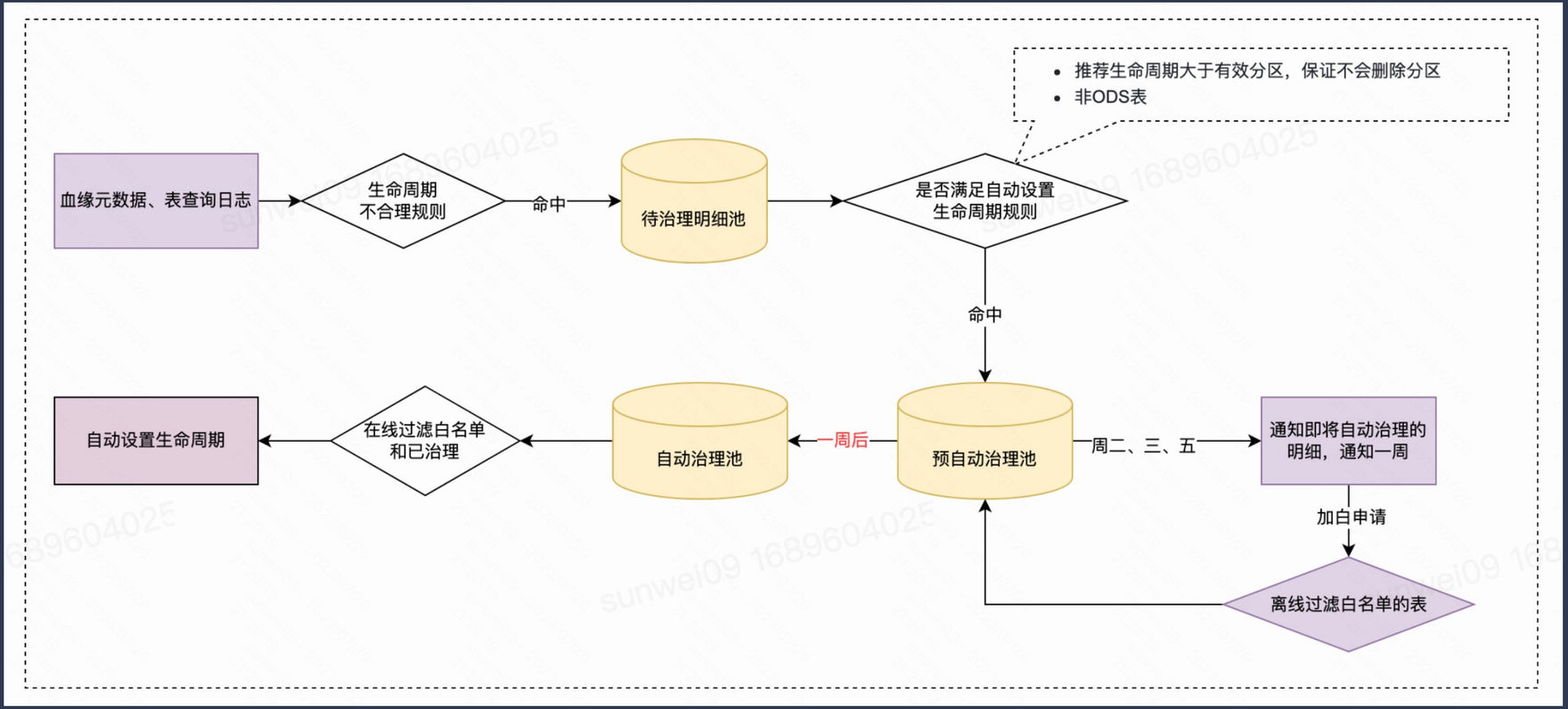
成本治理方案：技术白盒化

- 计算用量公式 = 任务数 X 调度频次 X 申请计算资源数 X 运行时长 X 优先级权重
- 存储用量公式 = 单行单列存储量 X 列数量 X 行数量 X 表数量 X EC副本数量

成本治理方案：技术白盒化策略



自动化治理实践：生命周期自动纠正



生命周期规范

- 根据不同数据等级、不同数据分层，结合数据是否可恢复以及恢复的成本制定标准生命周期规范

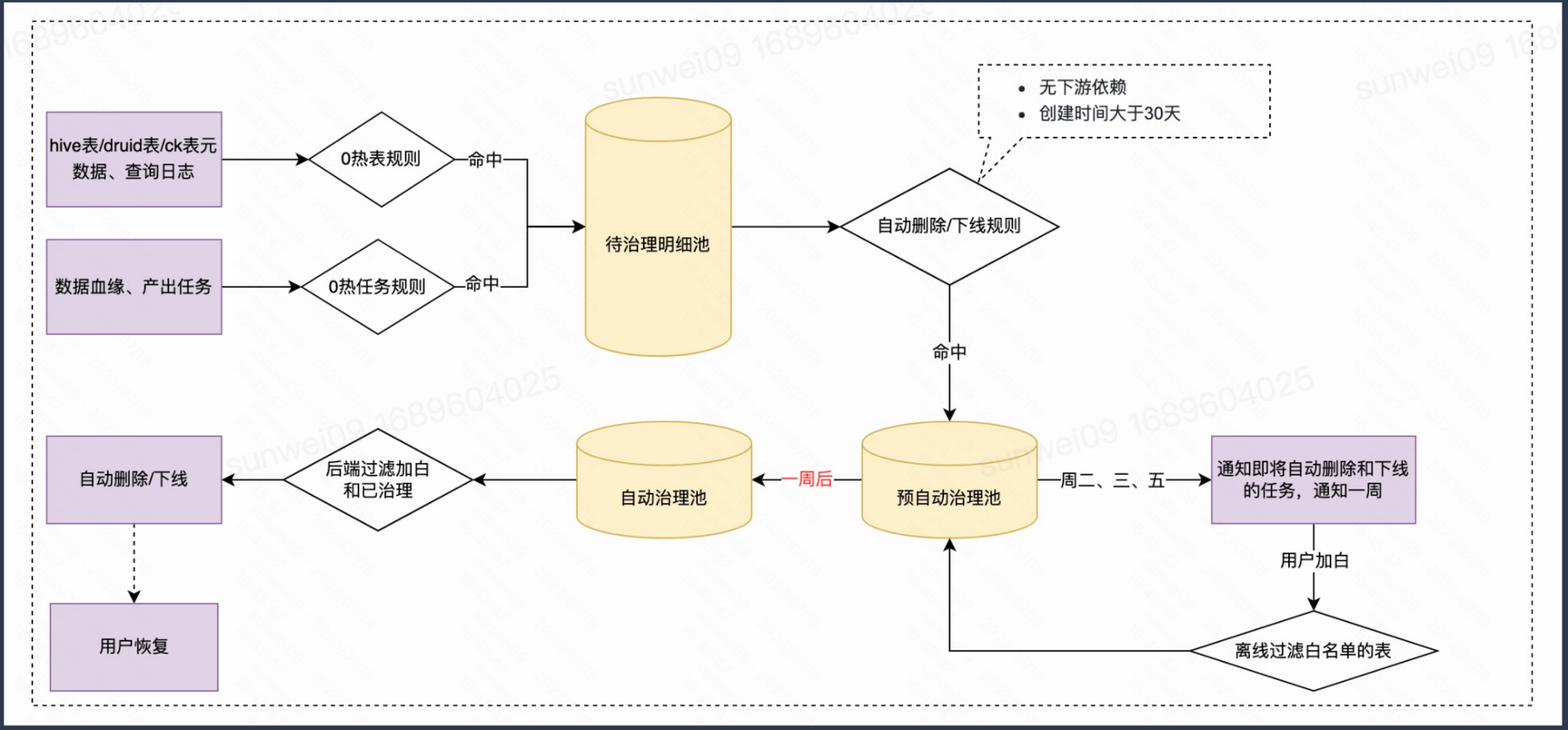
避免误删数据

- 通过基础的数据血缘，结合数据的查询访问日志来判断，并且取最早分区和推荐生命周期的最大值

通知机制

- 三轮通知，过程用户可以申请加白，无反馈后，一周后进行治理纠正

自动化治理实践：0热度表/任务自动下线



避免误删数据

- 通过基础的数据血缘判断下游依赖，结合数据的查询访问日志和创建时间判断该表和任务是否真实在用

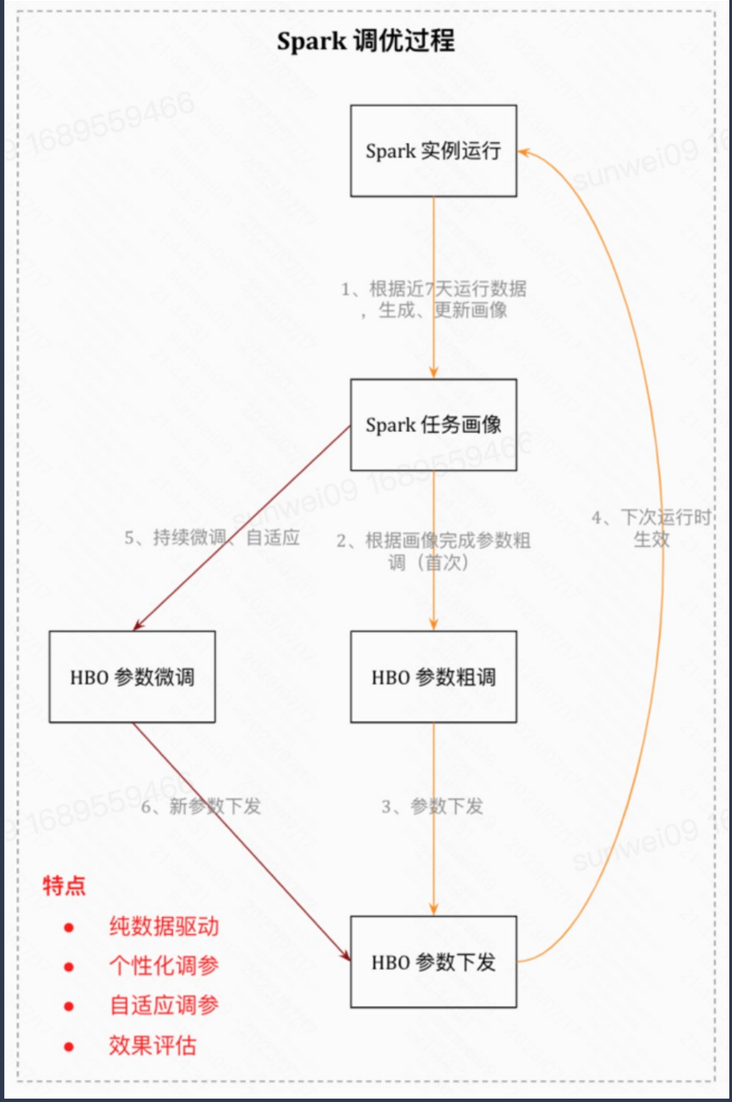
通知机制

- 三轮通知，过程用户可以申请加白，无反馈后，一周后进行治理删除和下线

数据恢复

- 对于自动下线任务和删除表，用户可以在工具上进行一键恢复

自动化治理实践：任务参数自动调优（HBO）



通过分析作业历史运行指标，以**数据驱动**的方式，**自动化**为每一个DAG推断最优的运行参数，以**减少资源开销**、提升运行效率

优化资源配额

- 通过自适应扩缩容CPU/MEM，解决资源不足和分配过大的问题

优化任务分片

- 通过自适应调整Map/Shuffle分片，解决分片不够、过多的问题

优化功能参数

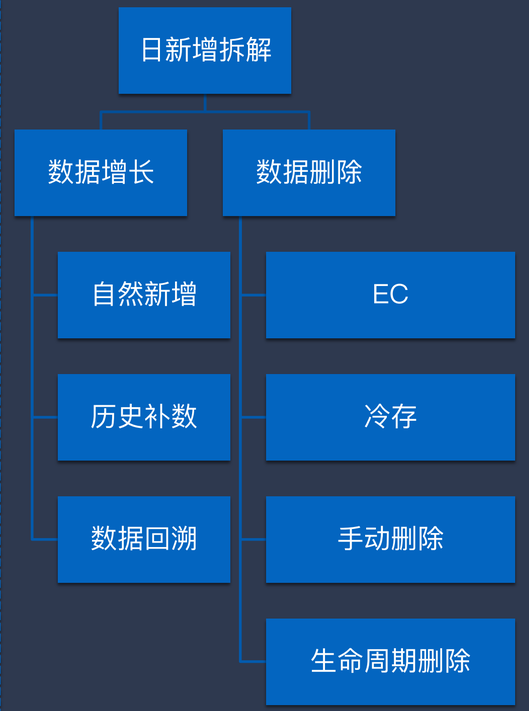
- 通过小文件合并等参数调整，提升性能

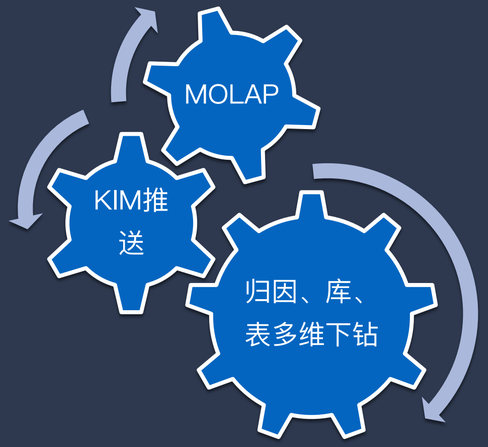
自动化治理实践：增量自动化归因

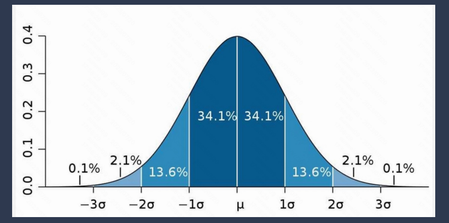
面临的问题

- 降本背景下，存储水位处于高位(95%)
- 日新增波动较大，缺少合理监控，每次发现很被动
- 波动原因无法快速定位，每次排查工作量大，不能快速修复问题

思路与方案



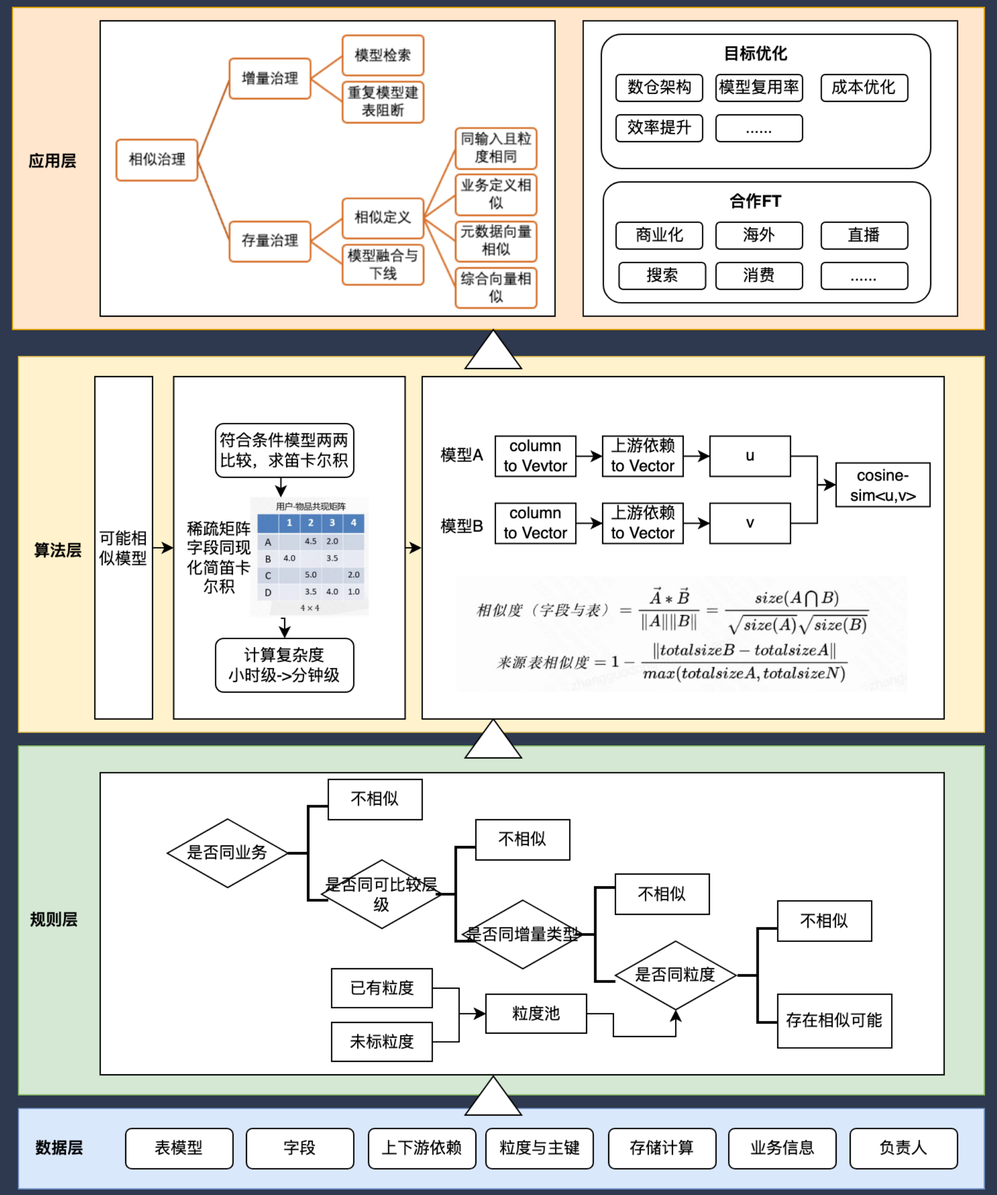




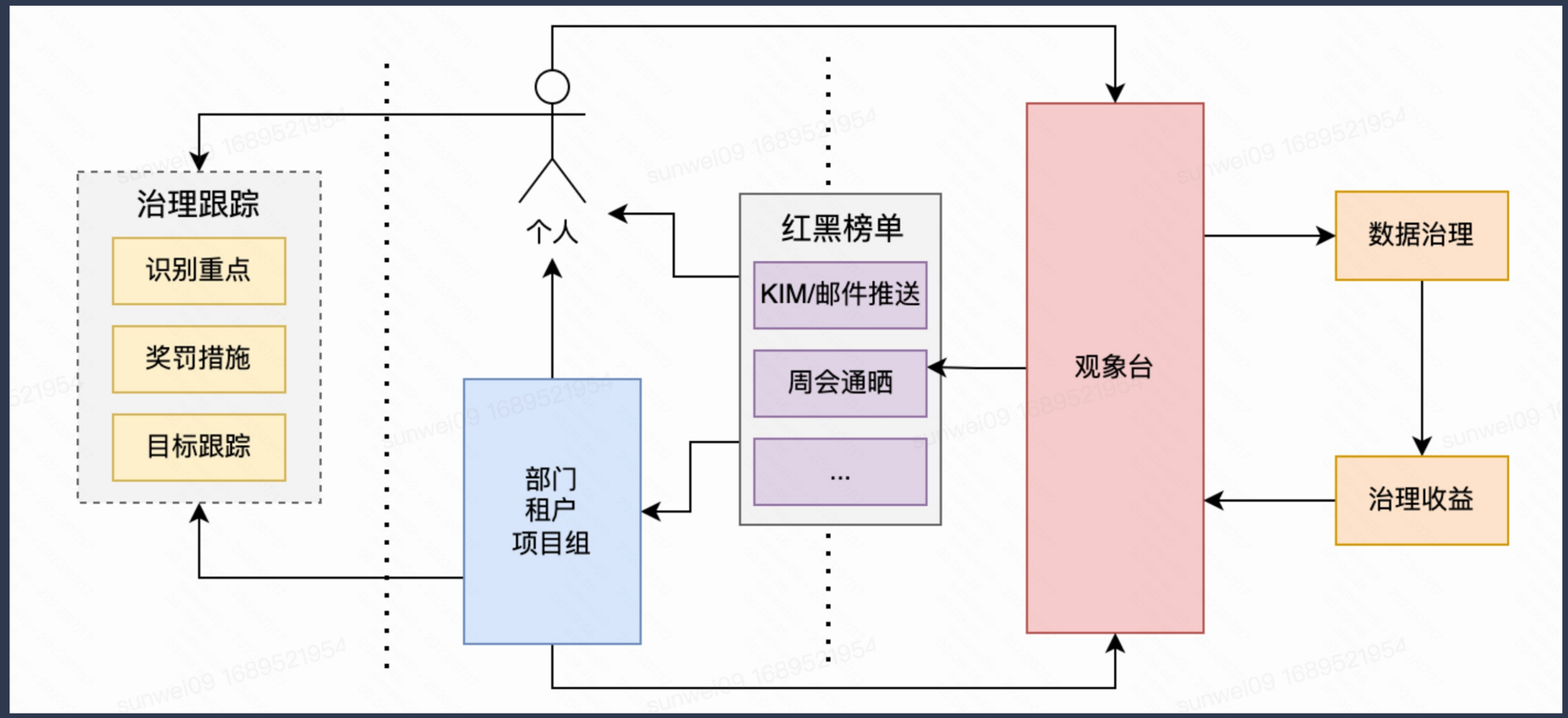
智能化相似模型检测

面临的问题

- 业务烟囱建设导致大量相似模型
- 难以定义相似模型
- 难以计算相似模型



成本治理运营机制



治理收益

成本 • 大数据成本节约上亿元

效率 • 治理效率提升N倍

总结

成本优化思路

- 成本管理（评估、流程、组织）
- 技术白盒化
- 业务白盒化

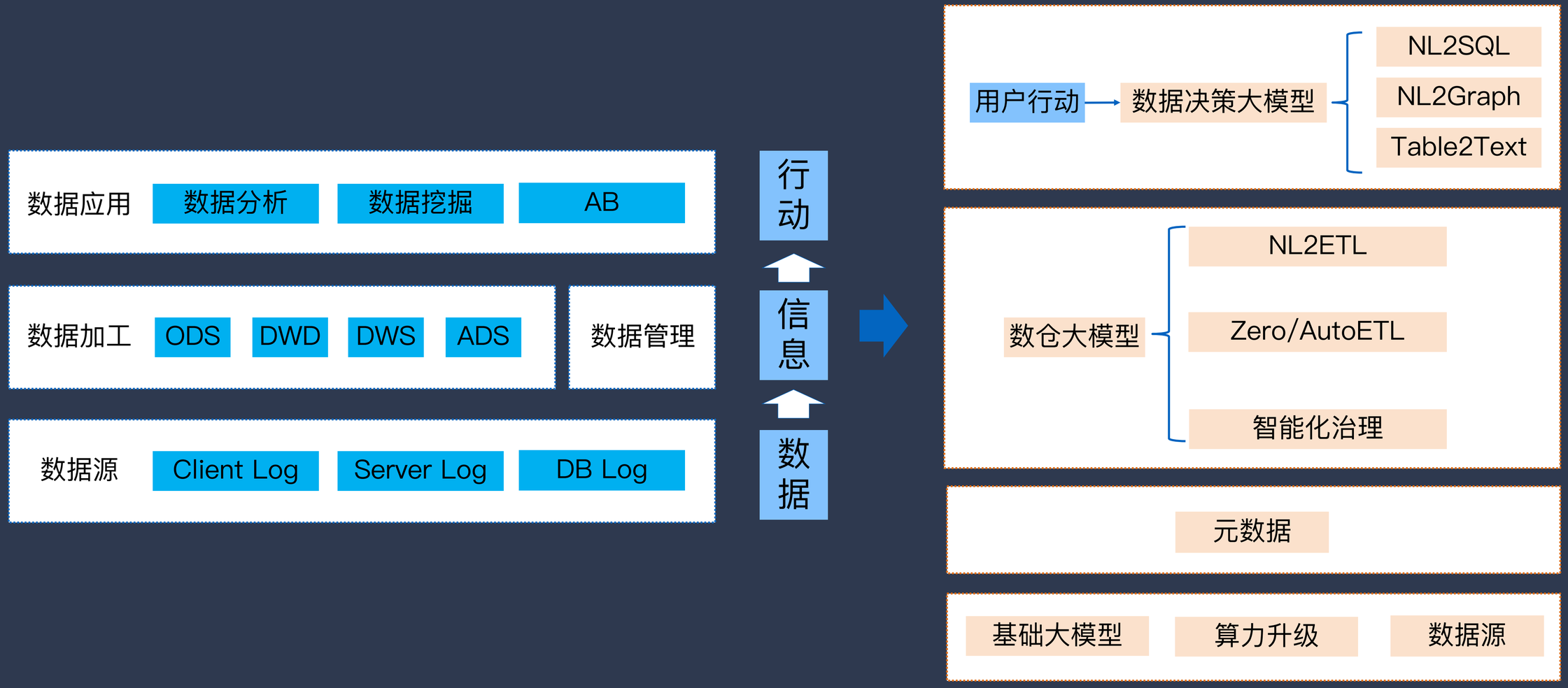
自动化治理方案

- 自动化生命周期纠正（标准规范、血缘准确率）
- 自动化下线任务
- 自动化删除表
- 自动化参数优化（HBO）

展望未来：规划

- 自动化治理覆盖提升
- 实时资源HBO
- 业务白盒化治理自动化诊断
- 数据湖治理

展望未来：思考


数据应用
数据分析
数据挖掘
AB
数据加工
ODS
DWD
DWS
ADS
数据管理
数据源
Client Log
Server Log
DB Log
行动
信息
数据
用户行动
数据决策大模型
NL2SQL
NL2Graph
Table2Text
数仓大模型
NL2ETL
Zero/AutoETL
智能化治理
元数据
基础大模型
算力升级
数据源

想一想，我该如何把这些技术应用在工作实践中？

THANKS