字节跳动一站式数据治理架构实践

基于数据驱动的分布式治理

王慧祥
字节跳动全域数据治理负责人

目录

- 机遇与挑战
- 字节数据治理理念
- 分布式数据治理架构及实践
- 数据驱动治理
- 智能化治理探索
- 总结&未来展望

01 机遇与挑战

数据治理挑战

01

治理效益与业务影响的矛盾

- 业务系统、生产流程改造影响业务
- 需求难统一，全局策略难落
- 保障治理大目标，无法顾及业务个性需求
- ROI评估：治理收益、时间周期、业务影响

02

治理涉及的组织和管理难度大

- 角色多、范围广、链路长
- 治理目标对齐、管理、跟进难度大
- 组织越复杂，数据治理难度越大

03

规范“人”的动作难度大

- 人员能力参差不齐，对齐目标和优先级困难
- 治理操作依靠人，规范对人的偏差操作容忍度低
- 组织文化差异，数据治理落地的方法、挑战、成效各异

04

缺乏适配性强的产品工具

- 现状、问题客观工具缺失
- 无全局视角工具，直接跳入治理细节
- 跨部门、跨系统治理目标对齐、协商缺乏治理全流程工具
- 平台工具不够灵活，只能解决通用治理问题

字节治理挑战

文化与效率、业务第一

业务要求

多业务齐发展
业务快速发展
快速响应业务需求
敏捷迭代

OKR文化

每个人都可参与规划与策略制定分解
主动寻找实现路径互相对齐
组织快速前进

高效治理

没有集团层面的数据治理委员会
各部门采取自决策自治的数据治理模式
决策与执行效率很高

规模大

业务场景丰富

- 互娱
- 电商
- 商业化

海量数据

数据驱动

资产数据盘点，体系建设

- 资产元数据，特征、标签
- 资源使用，存储、计算
- 工具，操作及收益

经验数据反哺，算法推荐

影响大

业务影响

- 数据延迟
- 质量问题
- 数据生命周期

02 字节治理理念

分布式数据自治

传统式治理

目标一刀切、自上而下、运动式

组织与制度

- 梳理业务与数据部门，设立公司级别数据治理委员会/部门

职权与管理

- 定期梳理公司数据资产，确保资产归属与治理权责明确

成果抽查

- 组织定期检查业务治理过程是否符合制度，定期检查治理结果

分布式治理

目标多元化、灵活自治、常态化

业务影响小

- 业务自决策，各级业务/个人都可自驱治理
- 工具灵活，业务根据自身发展按需，治理助力业务发展

周期短，见效快

- 以业务为目标对齐优先级
- 确认核心数据问题，聚焦投入，非“一刀切”

效率高，省人力

- 业务内治理目标对齐
- 实施、追踪、核算工具化
- 低门槛与算法推荐：业务自驱分析与诊断，自驱优化治理
- 产品横向沉淀业务治理经验，治理规则、策略共享

分布式数据治理平台

业务影响小、治理效率高、适配性强

业务影响小-灵活的自治模式

- 治理是不同业务与阶段的实践，在规范与组织上应足够灵活，业务可自身发展阶段制定治理内容，自行对齐与制定部分治理标准，互相对齐形成自驱组织
- “一个业务单元内的数据有效性提升为数据治理的范围和目标”

沉淀各业务治理经验，提升治理效率

- 产品辅助业务自驱，沉淀业务经验规则化、策略化、自动化进行持续的数据治理
- 低门槛与算法推荐：业务自驱进行分析与诊断能力，算法赋能治理提效
- 提供自上而下的规划式治理和自下而上的响应式治理

适配性强-产品建设覆盖治理全链路

- 从治理规划到执行诊断与复盘全流程进行治理把控。集成多种治理场景-稳定性、质量、安全、成本、报警
- 各模块可独立使用，按需组合，满足不同业务场景下的数据治理需求
- 产品提供完整的开放能力，业务根据自身特性和发展阶段进行接入

分布式数据治理平台-逻辑架构

治理用户层
- 管理角色
- 治理推动角色
- 治理执行角色

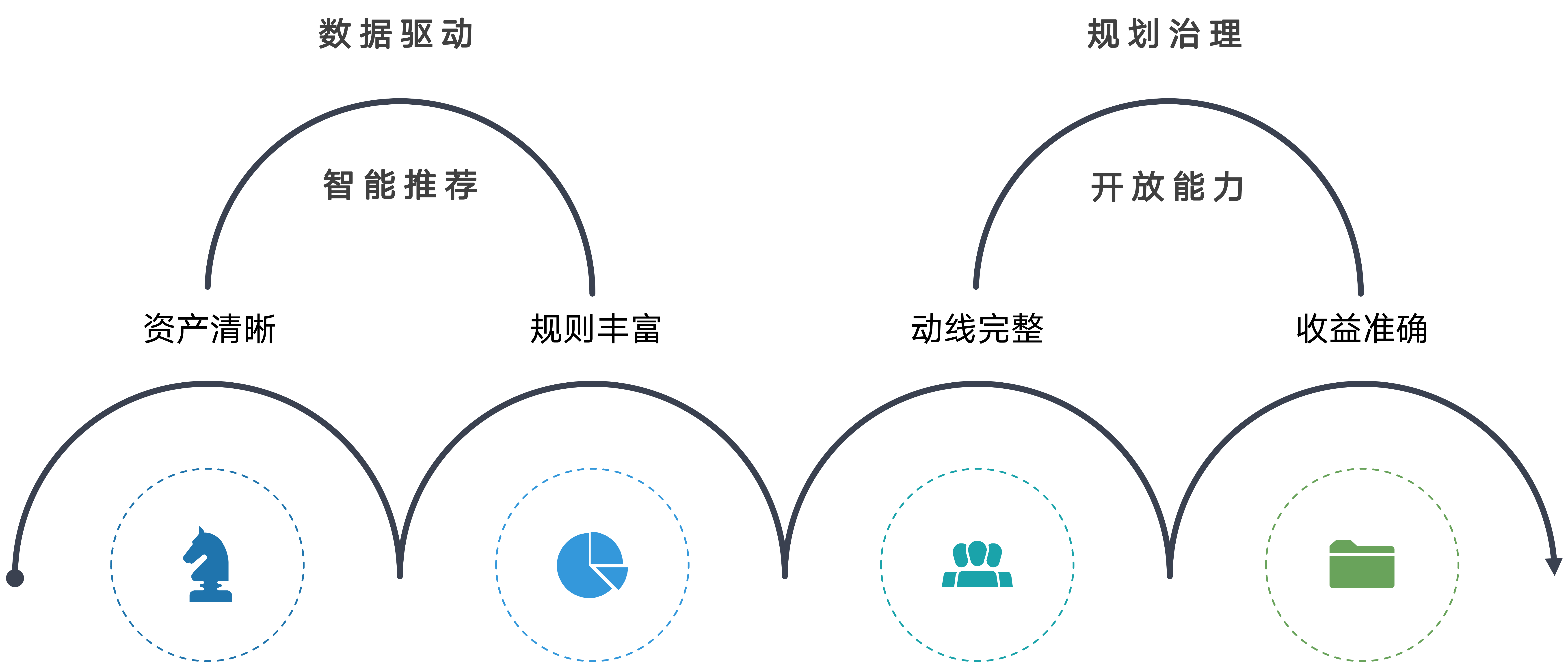
治理评估层
- **健康分体系**：存储/计算/质量…
- **SLA大盘**：就绪情况/延迟趋势…
- **资产大盘**：数量/资源用量/成本…
- **报警大盘**：趋势/起夜率/根因…

治理方案层
- **范围域**：部门/项目/数据团队/个人；资源组/队列/库；数仓层级/优先级/成本Top…
- **目标域**：提升健康分；降低存储/计算资源；优化资产数量…
- **规则域**：存储/计算治理规则；质量治理规则；安全治理规则…
- **消息域**：SLA报警；任务运行报警；质量规则报警…

流程框架层
- **健康分驱动**：健康分 → 扣分分析 → 问题定位 → 实施治理 → 健康分更新
- **规划驱动**：确定范围 → 设定目标 → 选取规则 → 执行诊断 → 消息触达 → 实施治理 → 进展更新
- **响应驱动**：报警订阅 → 问题处置 → 根因登记 → 复盘总结 → 大盘分析

基础能力层
- **Metadata仓库**：表/任务/报警…
- **治理规则引擎**：统计规则/算法规则
- **优化工具集**：TTL/温存/申报SLA…
- **收益核算**：存储量/任务量/vcore…

分布式数据治理平台-核心能力



- 数据治理的基础
- 对全域数据进行维度划分、打标及相关指标计算

- 根据资产特征生成
- 根据资产指标生成
- 根据治理手段生成

- 管理者视角
- 推动者视角
- 实施者视角

- 评估体系达标
- 成本缩减
- 优化性能

03 分布式数据治理架构及实践

分布式数据治理-体系建设

最小的业务打扰

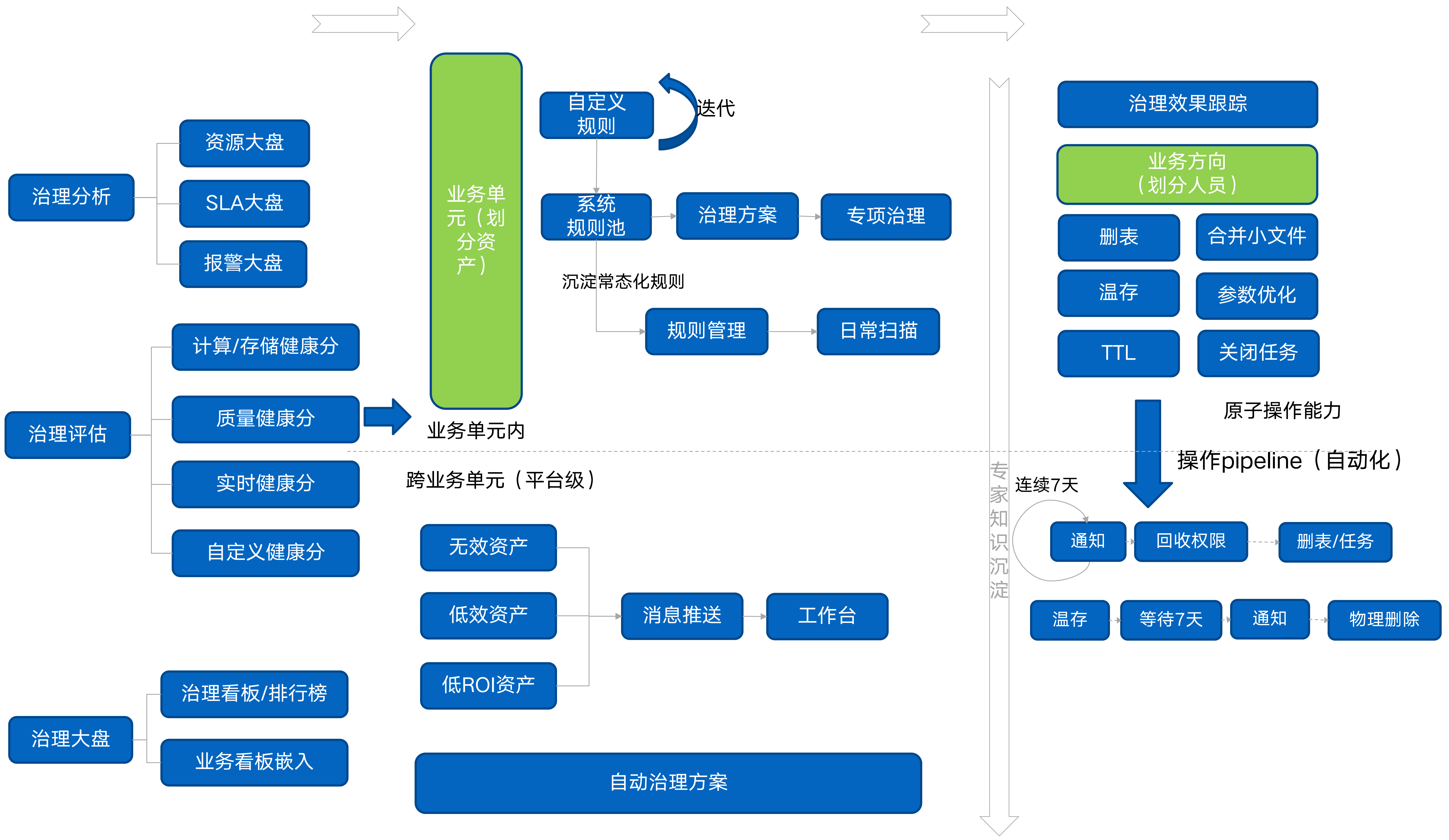
业务单元内制定目标，配合使用常态化及规划式诊断，构建业务自治体系

高效的组织形式

灵活配置推进治理的业务单元，自下而上人人参与数据治理

最高的执行效率

沉淀专家知识及智能化工具，执行经验的传承与协同，不断提高自动化水平



分布式数据治理-治理动线

自定义治理、常态化治理

制定诊断方案

推动方案资产干系人治理

推动者/执行者视角

2 使用规划诊断能力，根据资产维度、标签、相关指标等信息圈选数据资产，预设目标，完成时间创建方案

4 通过治理操作明细，在治理过程中管控治理进度，并根据阶段性目标校准治理策略（激进OR稳妥）

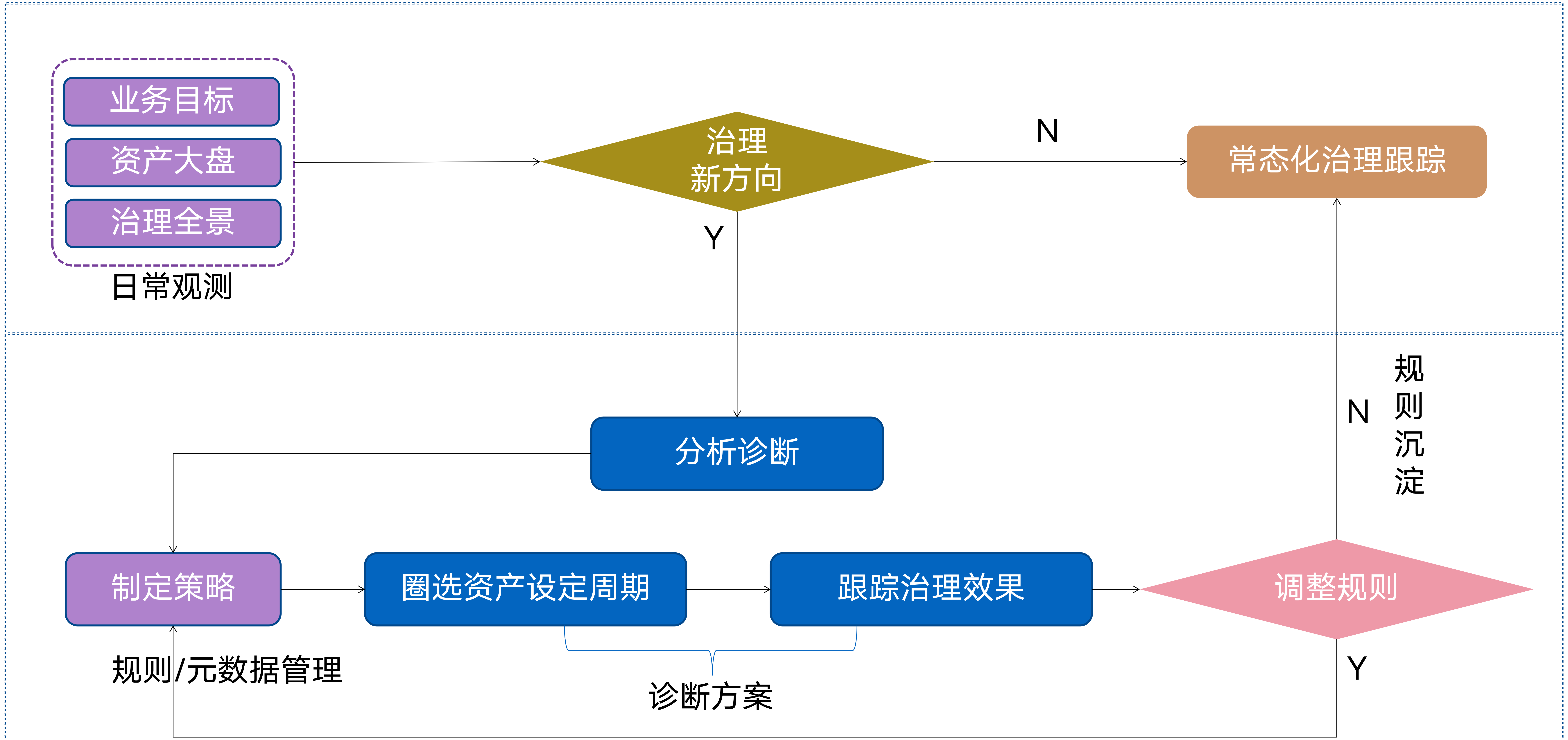
1 通过治理全景、业务目标、数据运营资产推送信息，发现资产问题

3 使用平台工具，如TTL调整、删表、转温存、参数优化、任务下线等手段，实施治理操作

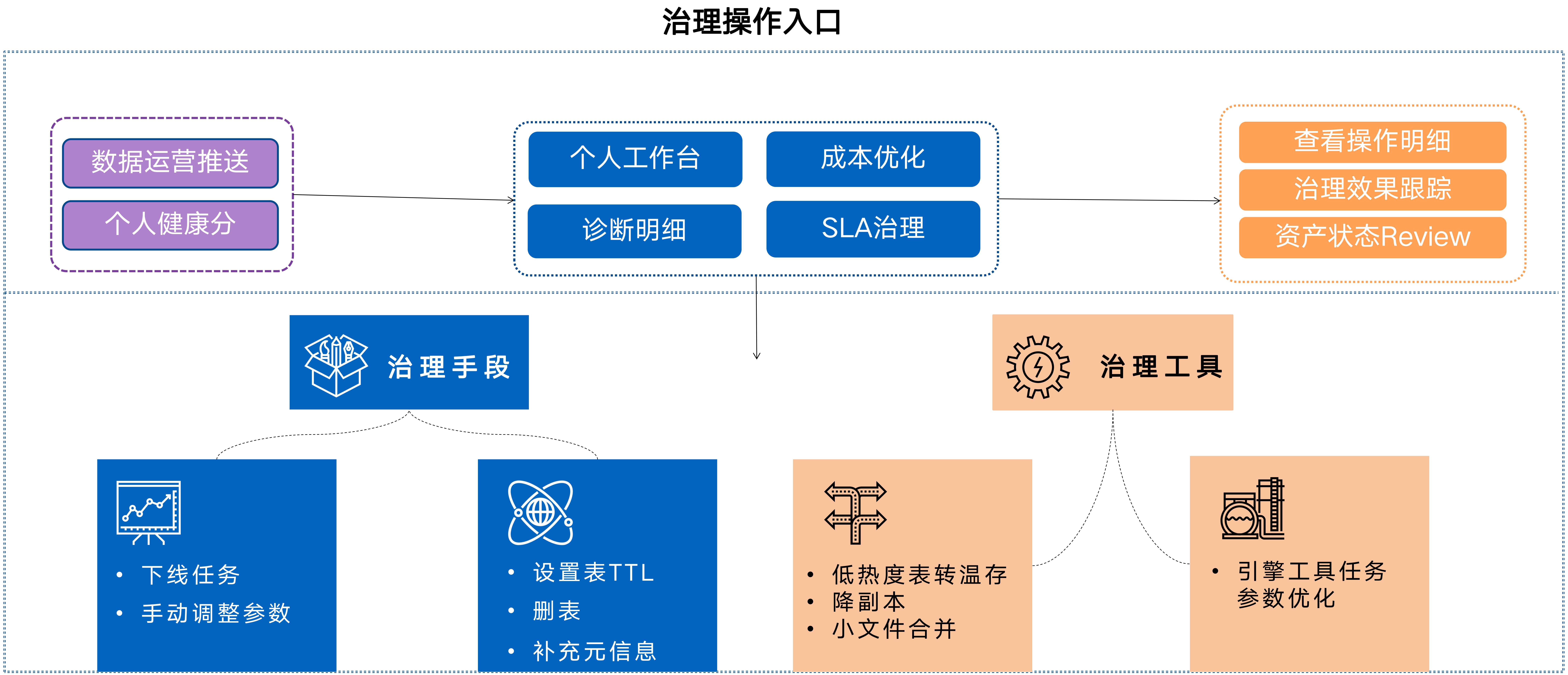
5 方案完结后，核对收益是否符合预期，评估治理工作，并复盘治理问题

实施者视角

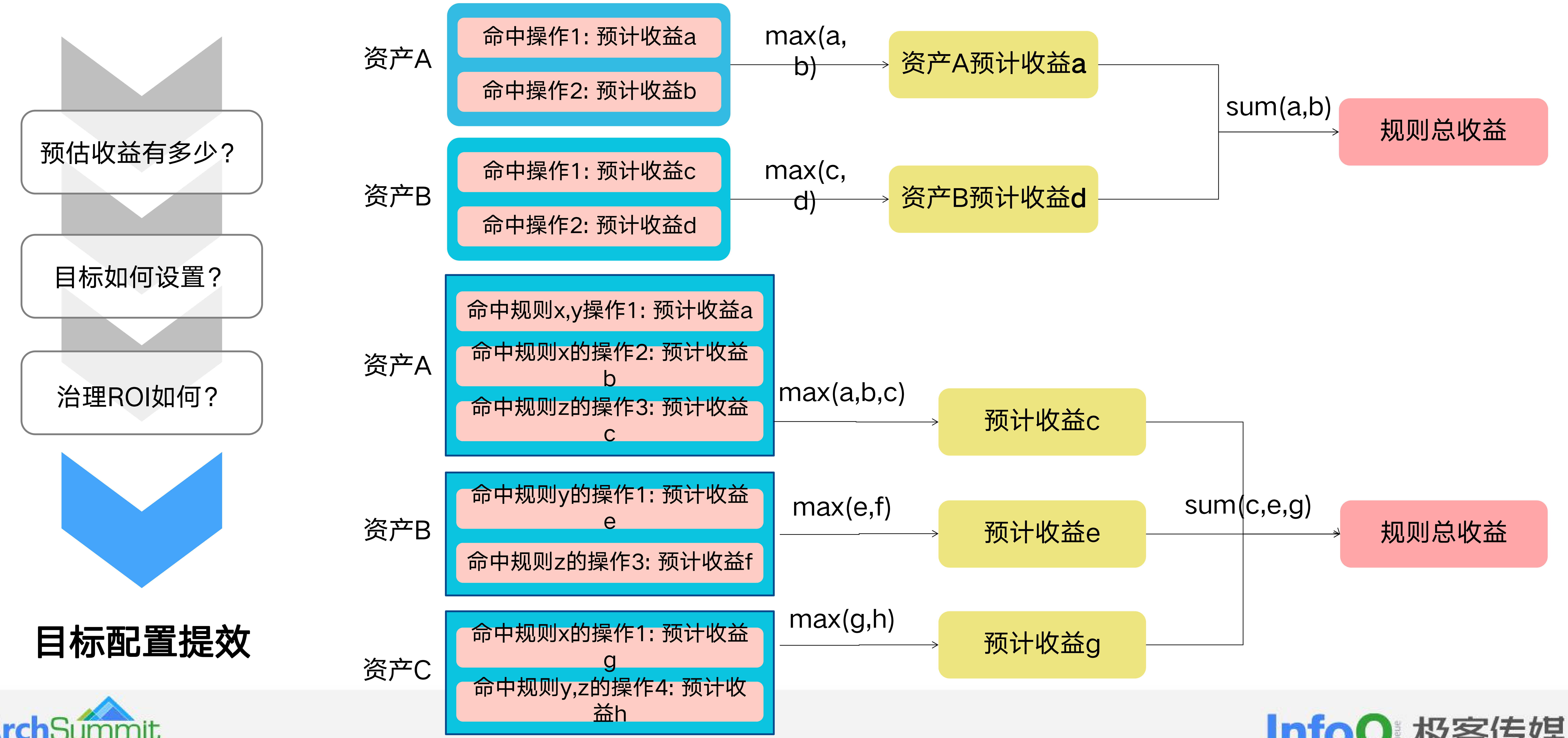
分布式数据治理-推动者动线



分布式数据治理-实施者动线



分布式数据治理-创建方案&目标



目标配置提效

分布式数据治理-治理实施&操作

集中式：平台集中建设规则数据及治理手段

- 研发人力投入成本高
- 很难匹配所有业务的需求

开放能力建设

分布式：数据开放、规则开放、治理操作开放

- 满足个性化诊断治理需求
- 规则迭代稳定后沉淀到平台，实现共赢
- 操作开放，业务自定义组合pipeline，满足精细化治理

80+

默认规则

存储、计算、质量、安全

治理场景

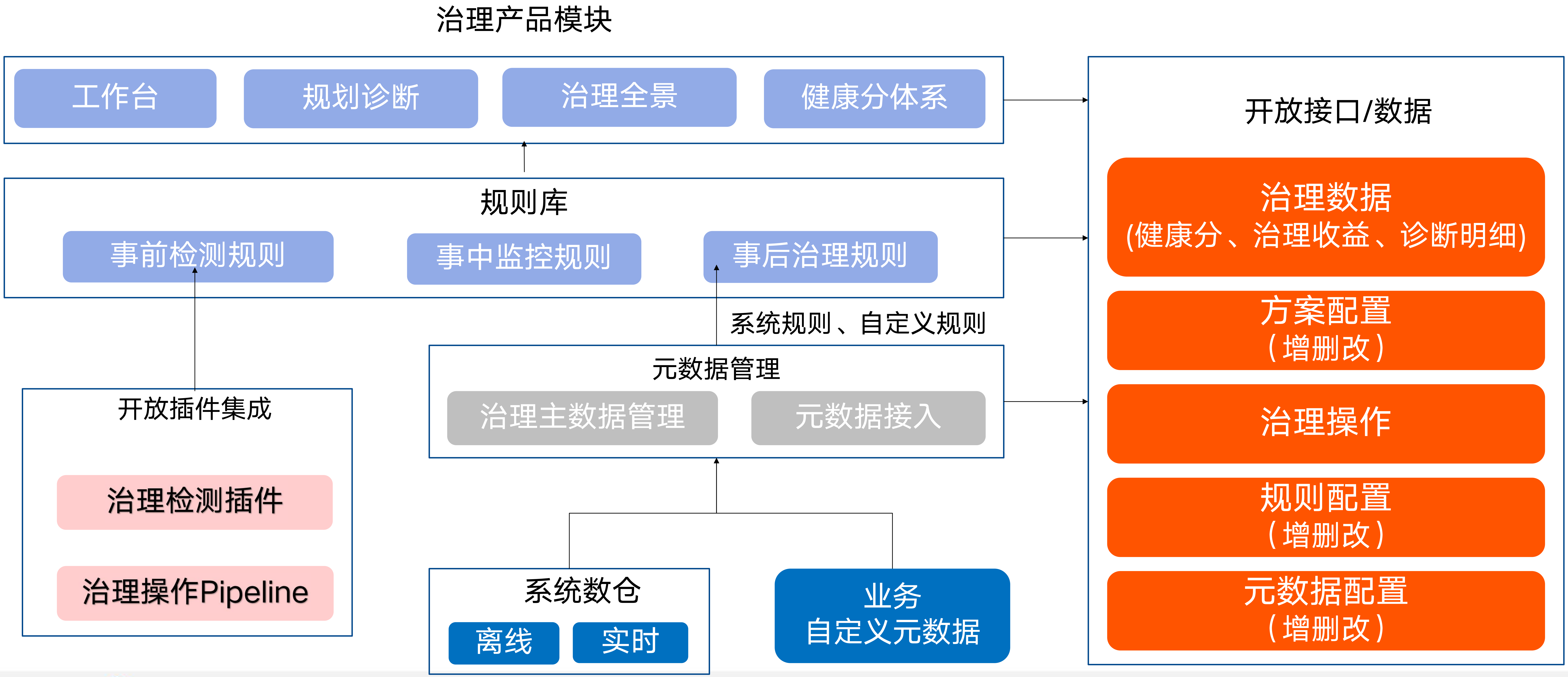
自定义元数据、规则逻辑

个性化需求

自助接入方法、灵活定义参数

精细化治理

分布式数据治理-治理实施&操作（开放性建设）



分布式数据治理-收益统计&结果验收

收益数据自动化收集

结果（评估/收益）标准化：

计算

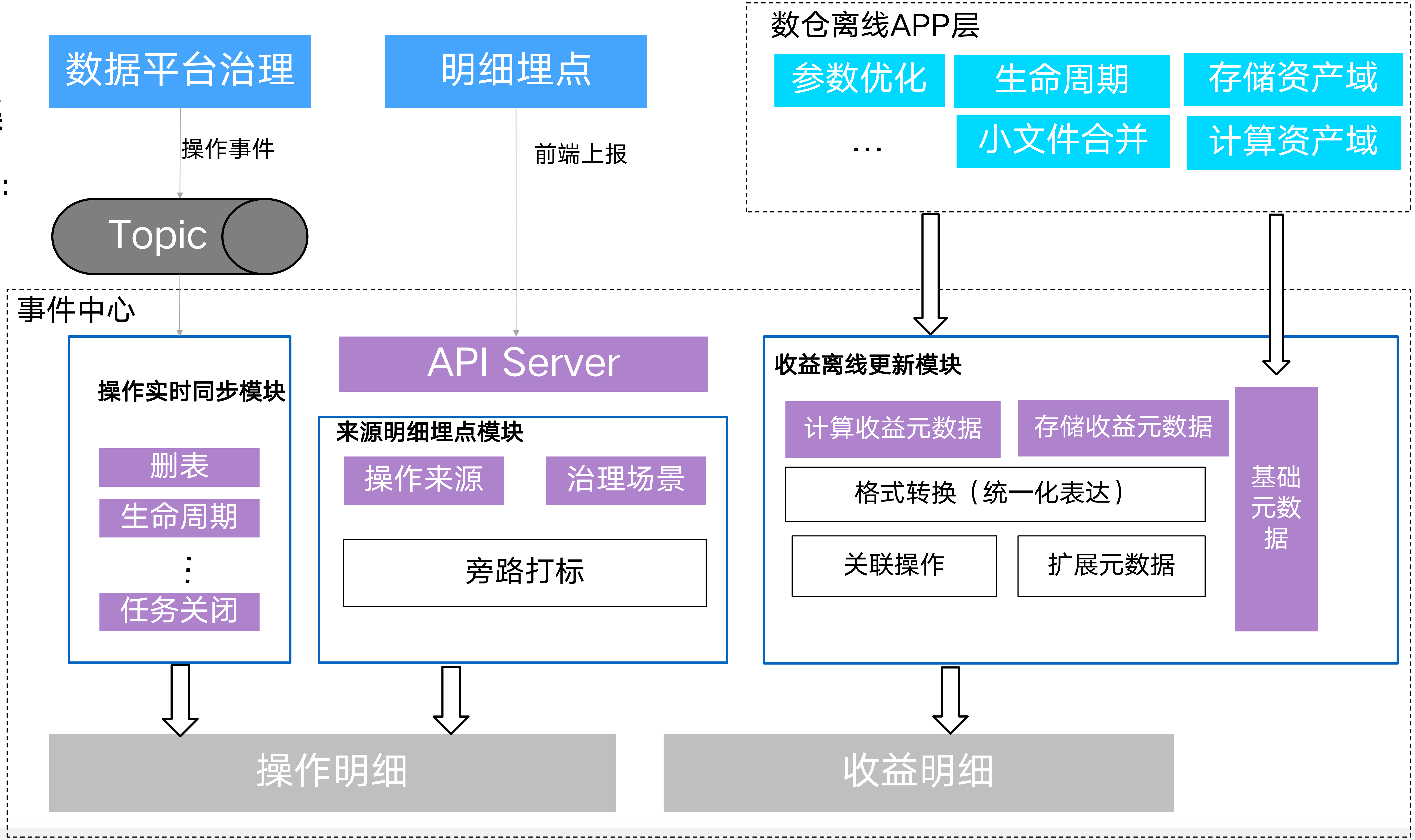
- 内存节约量/利用率
- CPU节约量/利用率
- 产出小文件数量
- …

存储

- 节约物理存储量

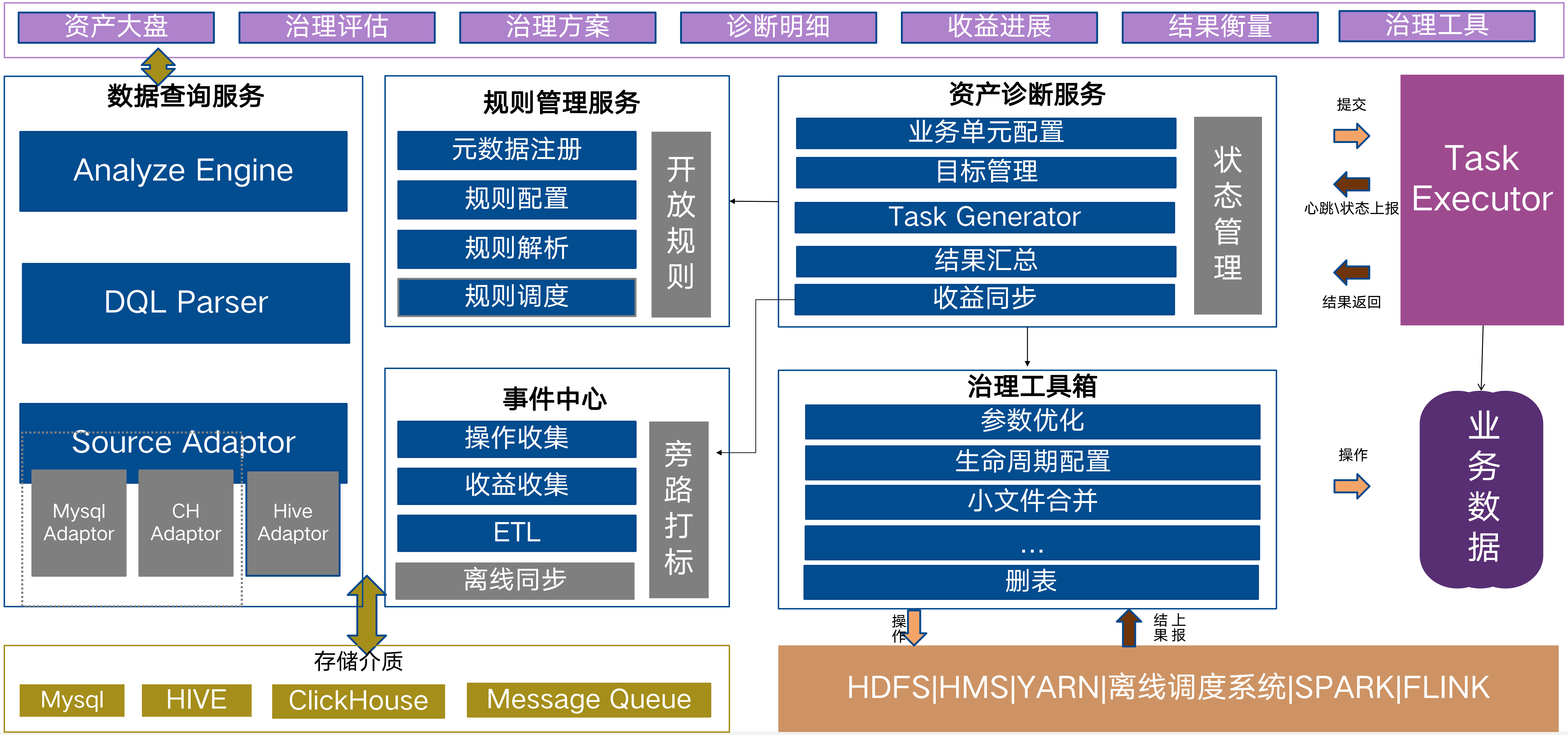
质量&安全：

- 质量监控治理数
- 安全风险处置数



思路：操作实时同步、收益离线更新、埋点旁路打标

分布式数据治理-平台架构



04 数据驱动治理

数据驱动治理

如何高效定位资产问题

盘点资产数据，构建完备的元数据组织方式，
通过特征、标签描述元数据，根据不同场景设计治理策略（存储、计算等）

如何高优治理业务资产数据

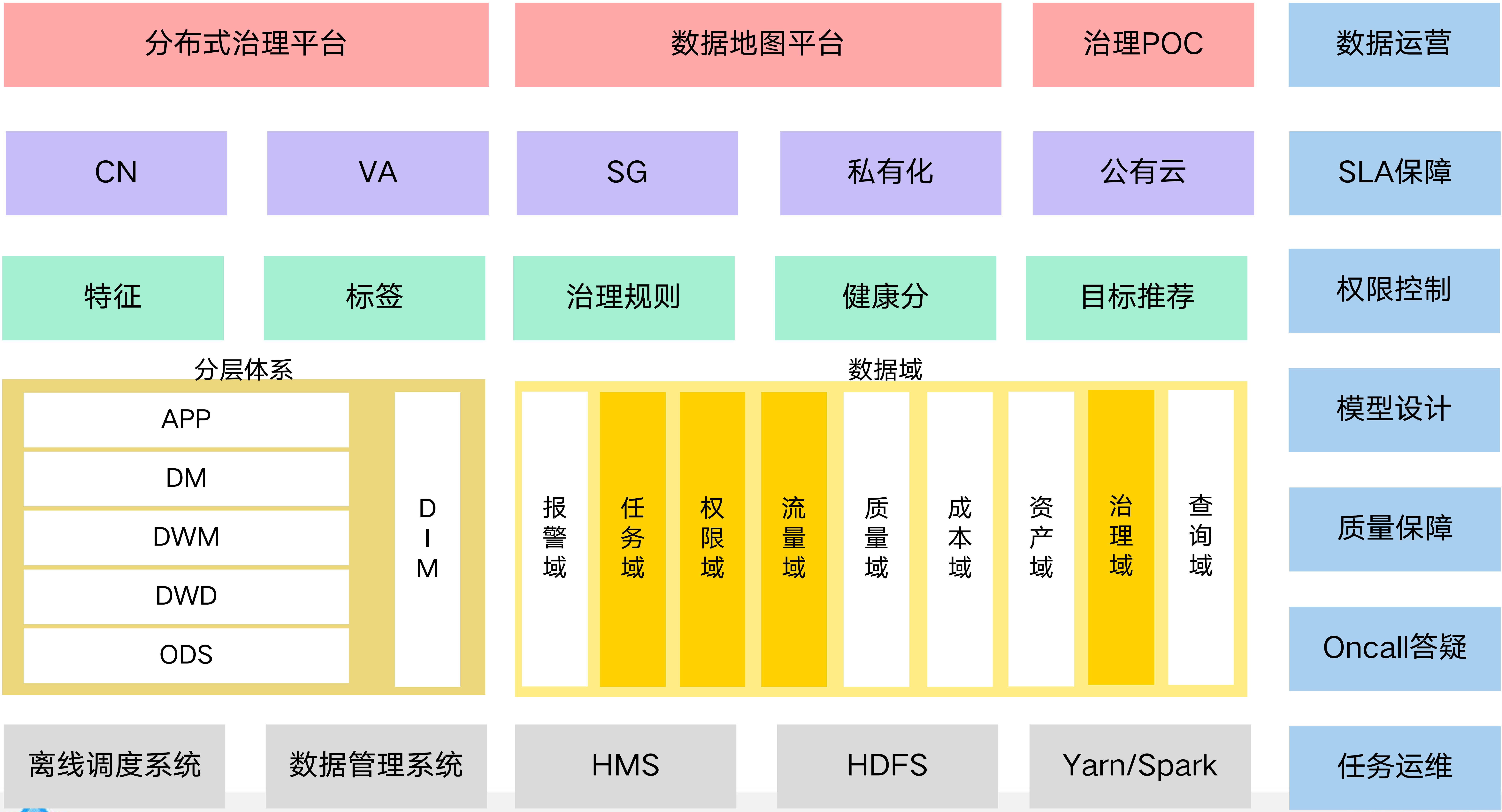
构建全公司的治理评估体系，提炼核心资产问题项
通过对资产打标，可快速定位高优待治理资产

如何快速完成资产治理

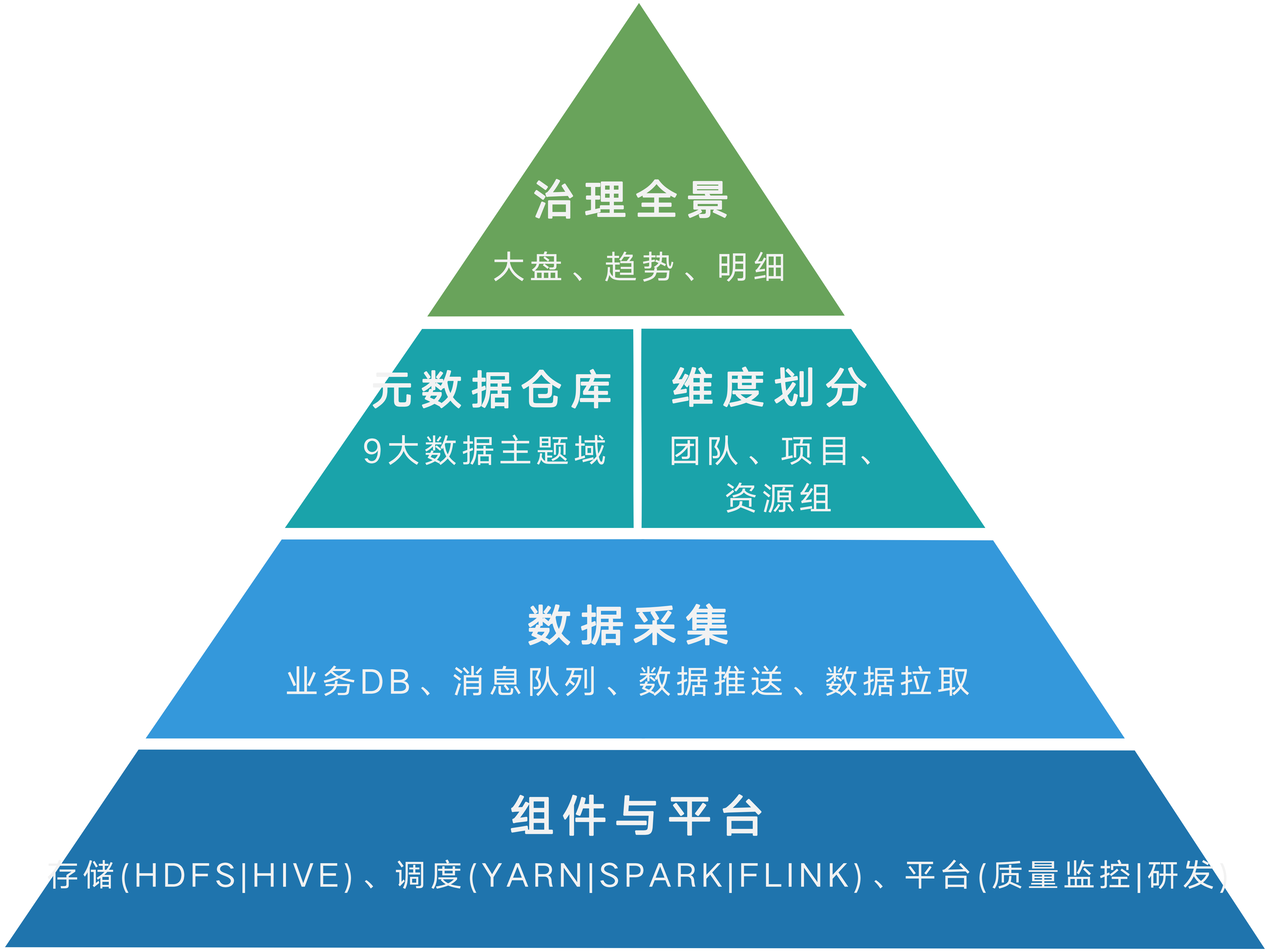
挖掘、沉淀并复用治理经验
通过往期治理经验，并对行为埋点数据分析，智能推荐治理目标

数据驱动闭环

整体数据架构


分布式治理平台
数据地图平台
治理POC
数据运营
CN
VA
SG
私有化
公有云
SLA保障
特征
标签
治理规则
健康分
目标推荐
权限控制
分层体系
APP
DM
DWM
DWD
ODS
DIM
数据域
报警域
任务域
权限域
流量域
质量域
成本域
资产域
治理域
查询域
模型设计
质量保障
Oncall答疑
离线调度系统
数据管理系统
HMS
HDFS
Yarn/Spark
任务运维

数据驱动-资产体系建设



- **01.** 数据分析与展示能力，解释性强，功能丰富
- **02.** 根据数据特征划分不同主题域，提供稳定可靠的维度、指标等
- **03.** 全链路保障数据采集，做到数据无丢失、可监控、质量稳定
- **04.** 从源头对资产数据打标，血缘脉络清晰，可追踪、可优化

数据驱动-评估体系建设

从完备的数据域建立资产评估体系，将资产问题具像化，并提炼高优问题项
根据资产类型进行分数加权计算，形成健康分，在公司层面达成共识，进行推广运行

全局视角 | 团队视角 | 个人视角

整体评估
单指标分数
治理成效
同环比值
趋势曲线
资产问题数
成本收益

多维分析
成员排名
资源消耗

催办
诊断

资产评估体系
- 存储
  - 无效存储
    - TTL不合理
    - 数据无查询
  - 异常存储
    - 小文件数过多
    - 存储格式不合理
- 计算
  - 无效计算
    - 产出表为空
    - 产出表无热度
    - 连续失败
  - 异常计算
    - 资源申请不合理
    - 数据倾斜
    - 低优任务调度在高优队列
- 质量
  - 数据及时性
    - SLA buffer不足
  - 数据内容质量
    - 核心任务缺失质量监控
    - 质量监控不达标
  - 监控告警
    - 核心任务缺失告警配置
    - 非核心任务配置夜间值班计划
  - 任务稳定性
    - 高优任务运行在低优队列
  - 元信息完整性
    - 层级信息缺失
    - 描述或中文名缺失

数据驱动-规则体系建设

1

数据模型

- 数据建设
- 分析挖掘
- 规则建模

2

规则体系

- 存储规则
- 计算规则
- 质量规则
- 报警规则

3

资产圈选

- 资产维度
- 特征细节
- 指标范围

- 完备的治理规则能力
- 存储、计算、质量、报警4大维度（80+）
- 全局规则 & 自定义规则
  - ➢ 生命周期永久 /近7天产出为空 / 暴力扫描任务
  - ➢ 生命周期xxx天 / 近xxx天产出为空
- 统计类规则 & 挖掘类规则
  - ➢ 近90天无访问表 / 数据倾斜任务
  - ➢ 相似库表 / 相似任务
- 根据规则圈选资产范围
- 用户自定义规则

数据驱动-智能提效

TTL推荐
合理设置表生命周期

温存推荐
减少存储层压力

治理目标推荐
精细化推进资产治理

阶梯分层推荐TTL

- 访问热度
- 表分层
  - ODS
  - DWD
- 表类型
  - 全量表
  - 增量表

通过打分机制推荐

- 访问得分
  - 访问周期
  - 访问次数
- 总文件大小得分
  - 目录总大小
  - 文件平均大小
- 元数据平台目录得分
  - 基础库，核心目录减少进入温存得分
  - 跨机房访问状况，越频繁使用，越不应导入温存

根据治理经验数据预测治理收益

- 单资产多操作收益预估
  - Max(O1,O2)
- 多规则的目标计算
  - Max(R1,R2,R3)
- 考虑治理整体完成度，初步将总目标值计算最后 * 40%

05 智能化治理探索

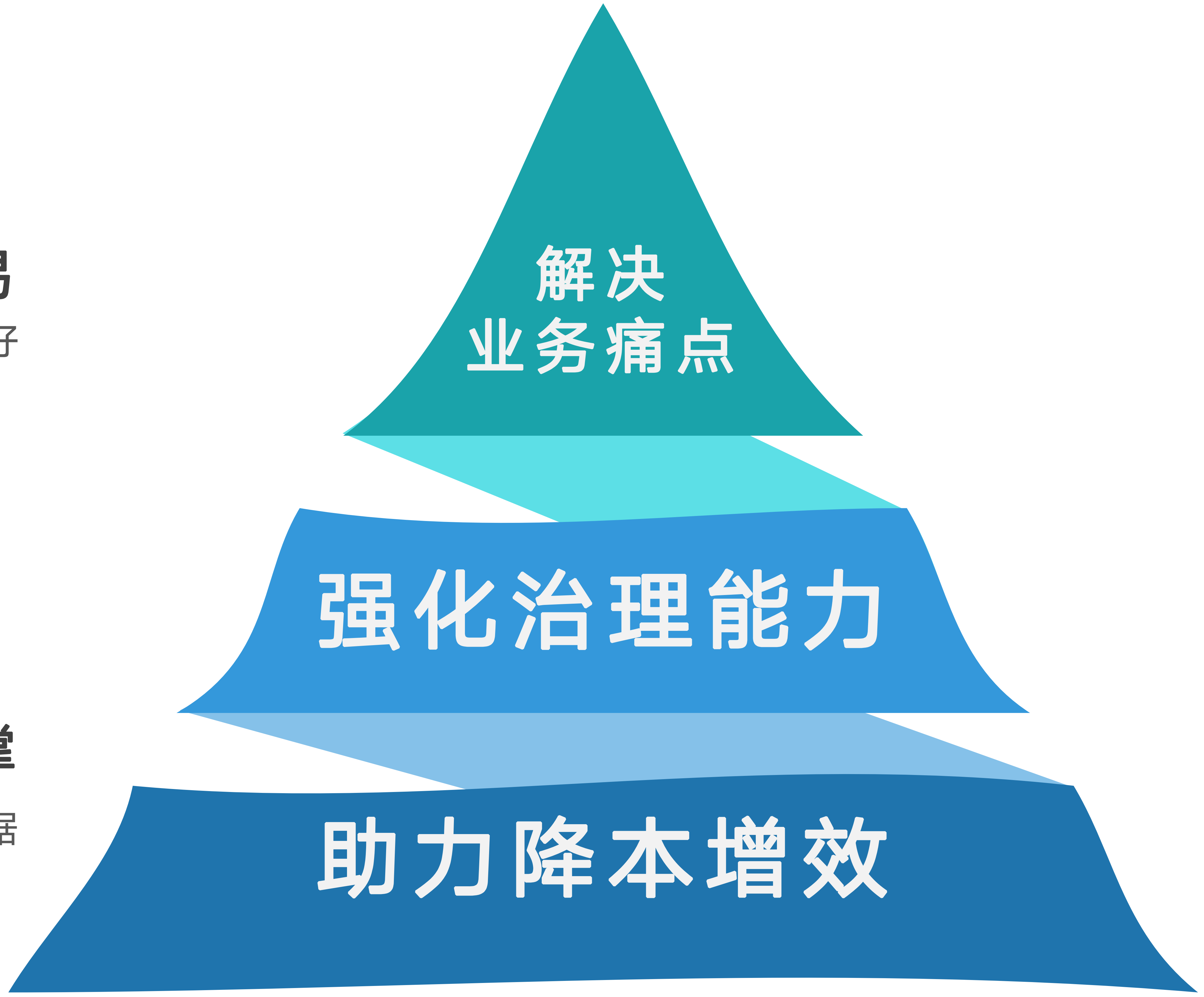
思考：数据治理智能化

操作简易

集成化、结果可度量、效果好

数据支撑

多服务、多引擎、海量数据



算法引擎

规则库、经验分析、自主纠错

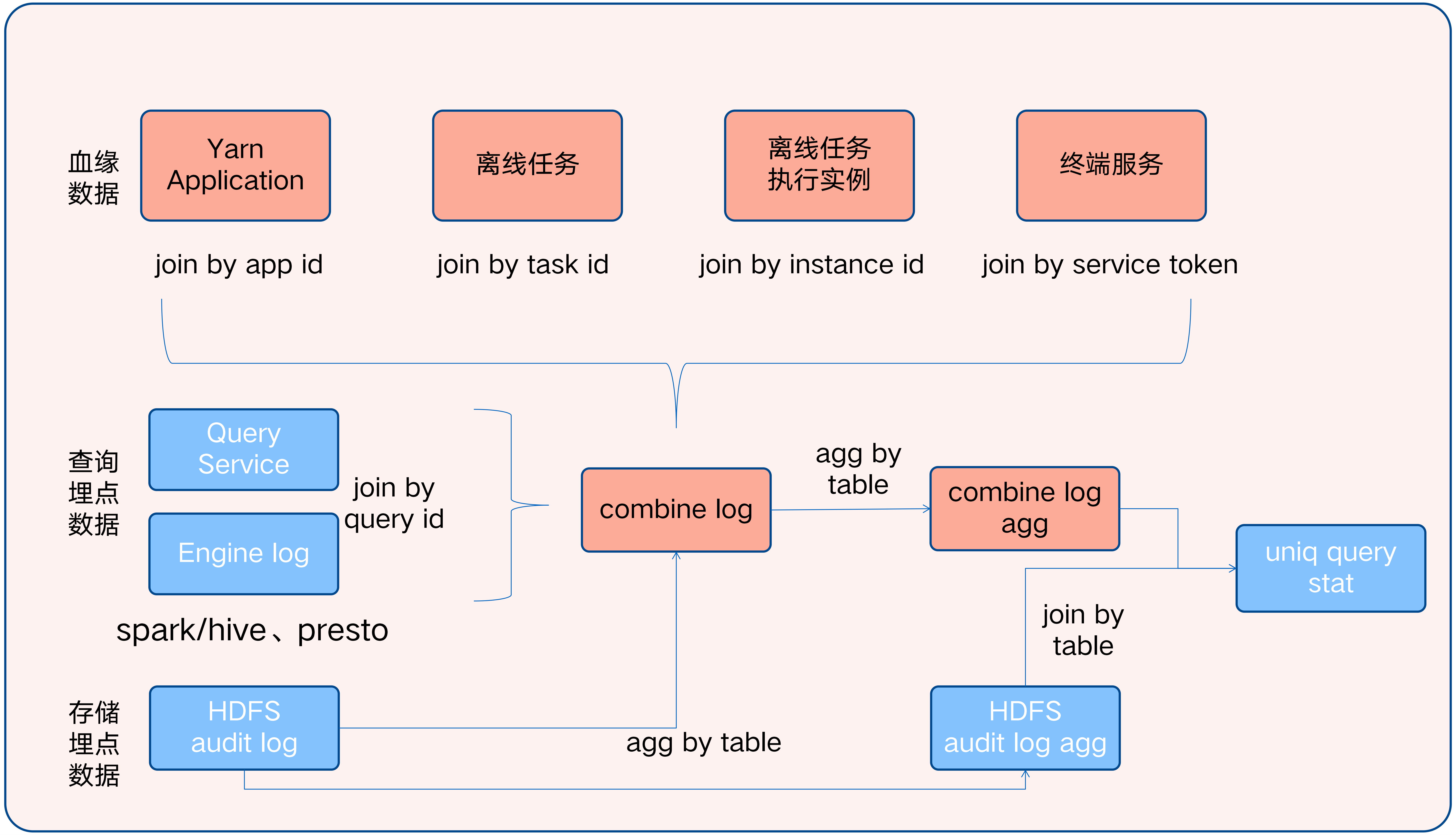
智能化治理实践-基于血缘和热度的推荐

热度数据作为判断数据访问情况的有效输入，其数据的精细化可以为更为细致、激进的治理提供数据支撑。

数据维度广，完善度高，来源可覆盖全公司

处理流程统一，可明确访问次数概念

最终结果可衡量，有效提升业务治理效率

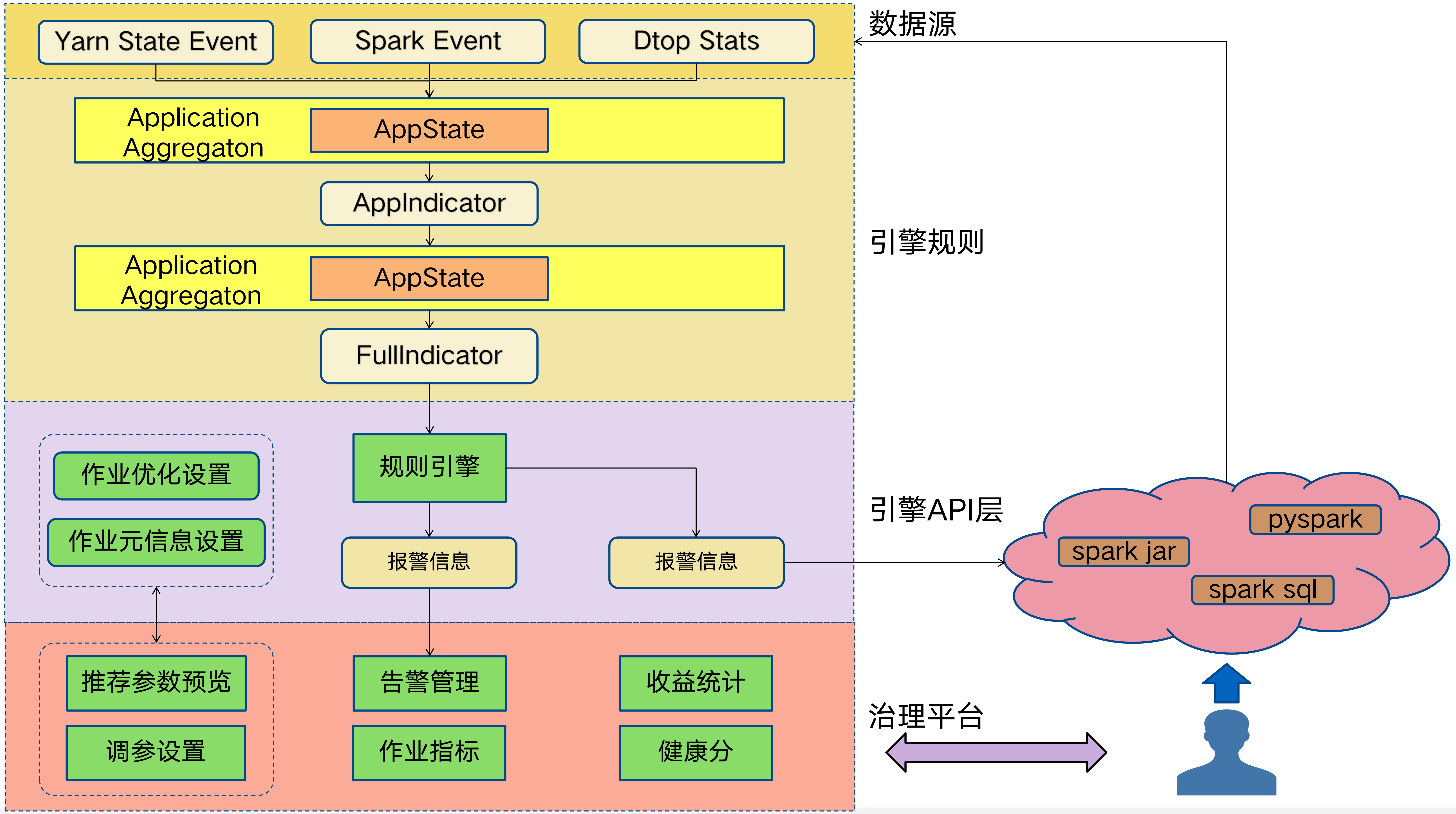


智能化治理实践-任务参数自动优化

根据作业的特点，自动推荐最合适的参数

Spark Engine：
- ◻ shuffle 溢写分裂
- ◻ shuffle 分级限流
- ◻ oom 自适应
- ◻ blacklist 功能优化

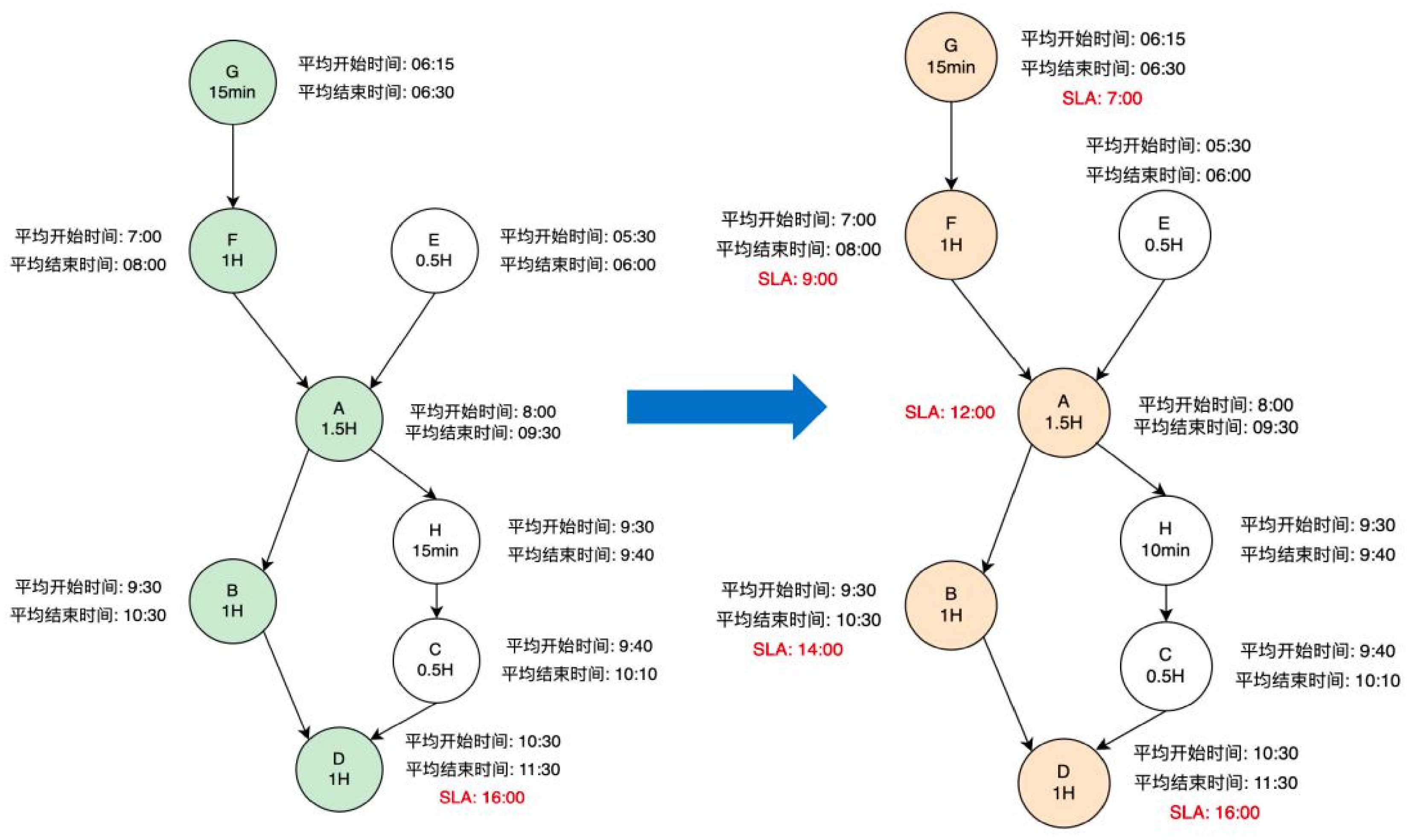
Rule Engine：
- ✓ 资源优化规则
- ✓ Shuffle优化规则
- ✓ 任务读写优化规则



智能化治理实践-其他算法探索

任务SLA签署推荐

- 基于运行时间做权重分配
- 确保下游任务可运行完成
- 关键路径分析计算



$$buffer = suggestSLA - avgCompleteTime_n$$

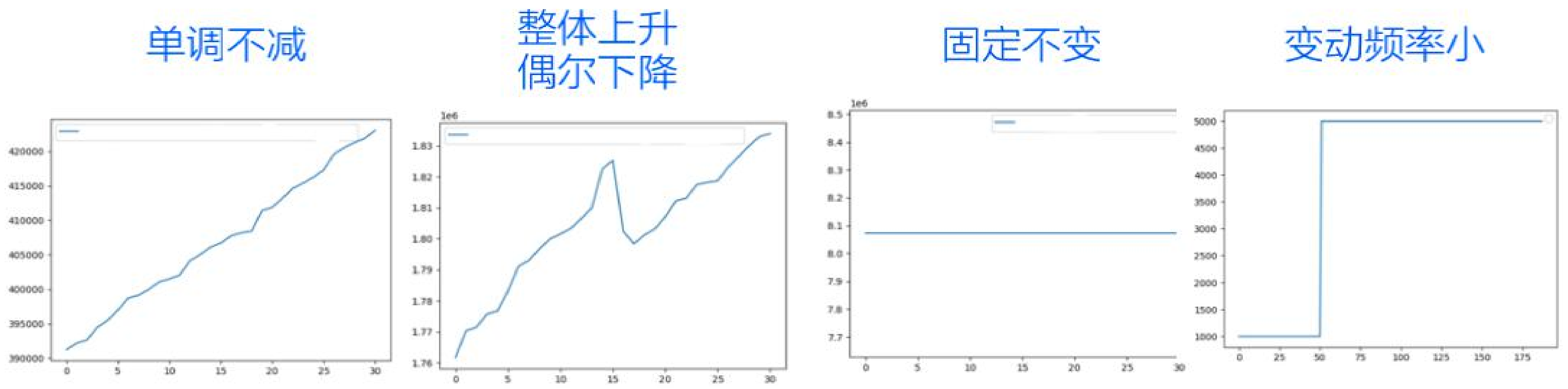
$$totalRunningTime = \sum_{i \in \Lambda} avgRunningTime_i$$

$$buffer_i = buffer * \frac{avgRunningTime_i}{totalRunningTime}$$

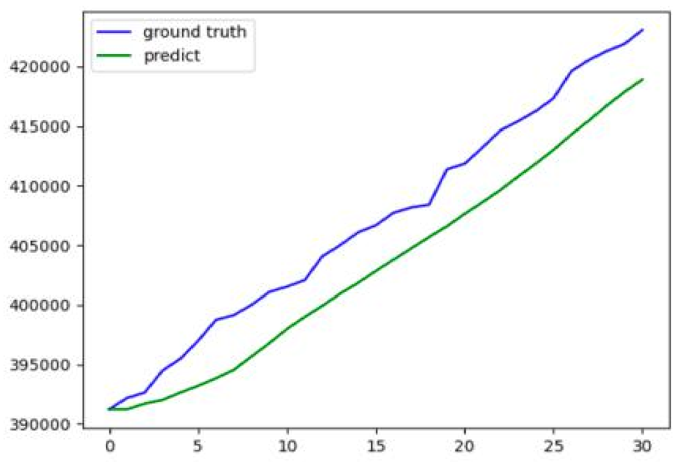
$$ptb = \frac{buffer}{totalRunningTime}$$

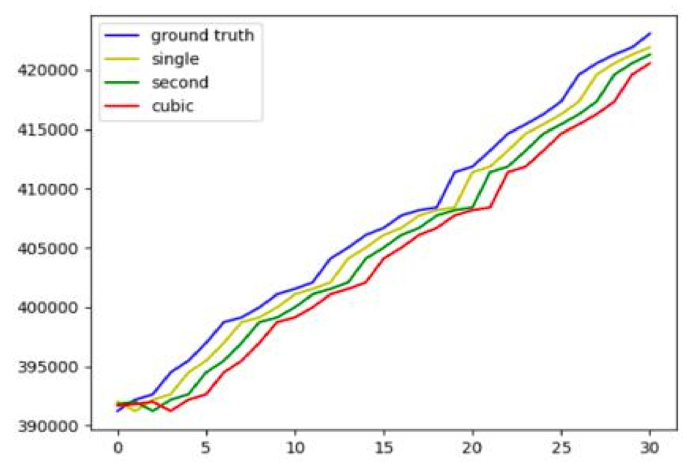
动态阈值监控

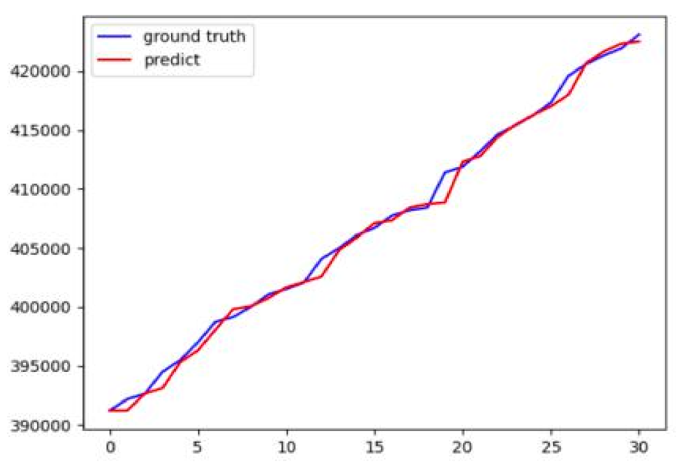
- 报警阈值 = 预测表行数 * 倍数
- 数据分布

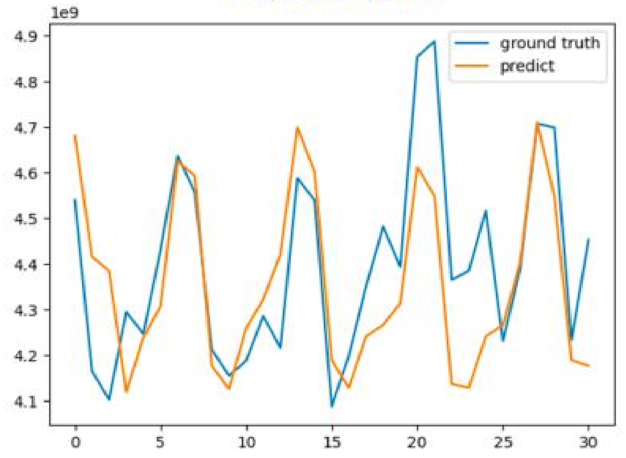


- 预测方法



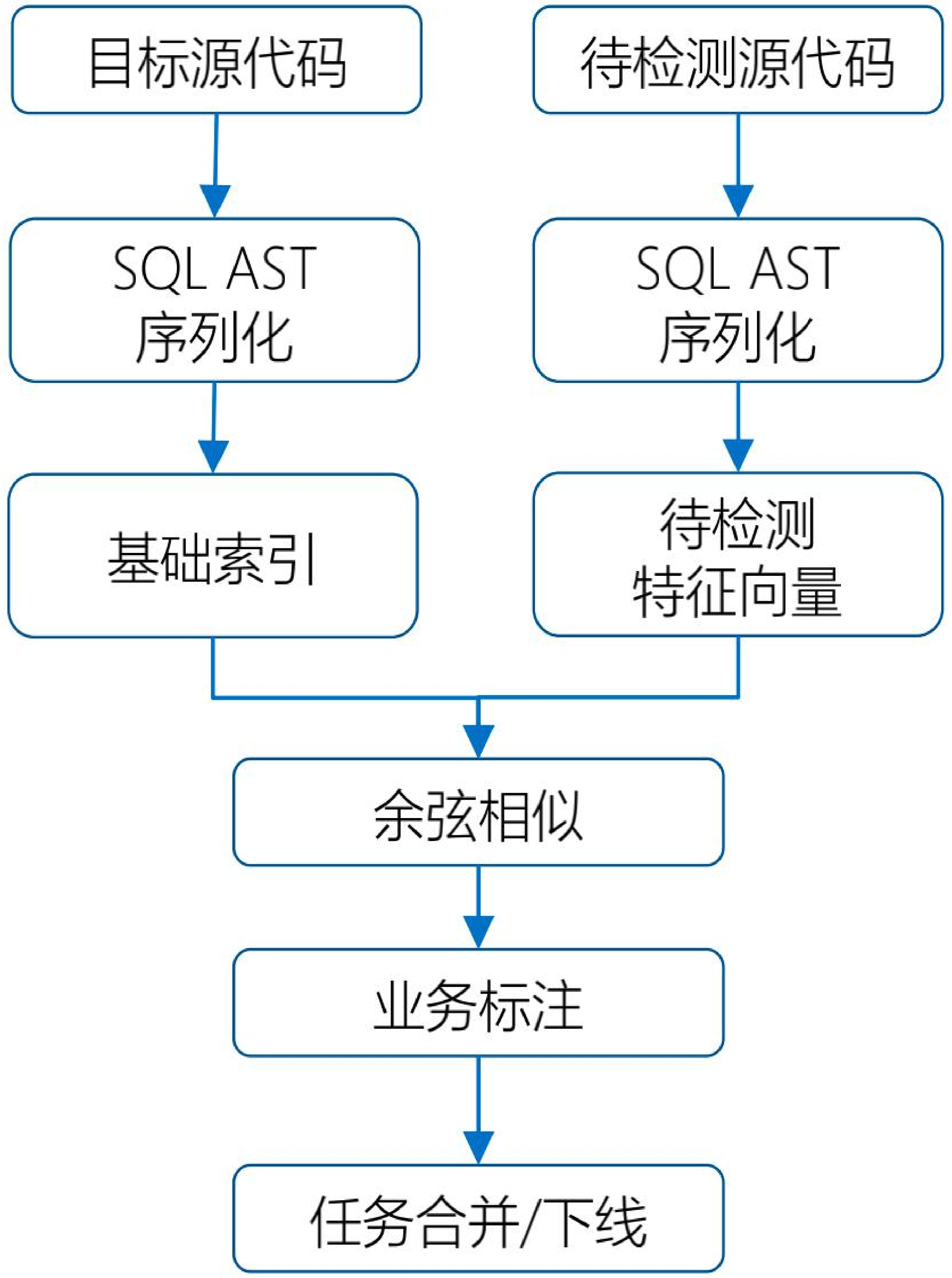






相似任务识别

- 业务合作典范



06 总结

总结

分布式治理

低业务影响 | 高度灵活治理 | 高效治理

闭环链路：资产清晰 → 规则丰富 → 动线完成 → 收益明确

平台功能：

- 治理分析：治理全景、健康分、治理大盘
- 治理目标：绝对值、分数值、目标推荐
- 治理方案：规划诊断、治理周期
- 治理实施：通知催办、进度跟进、操作动作
- 治理收益：数据收集、收益归属、收益计算

数据赋能：

- 数据资产体系建设、治理评估体系建设、治理规则体系建设、智能化策略沉淀
- 元数据采集、模型建设、数据挖掘、特征构建

治理推动者

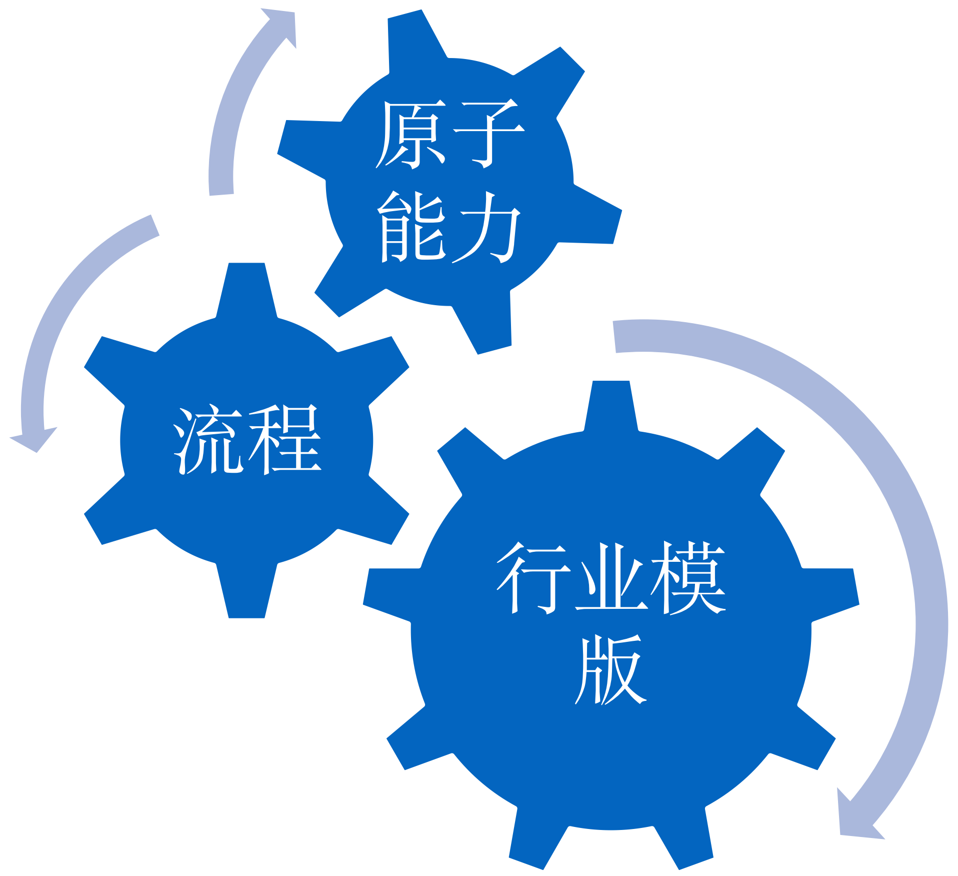
治理实施者

07 未来展望

未来展望

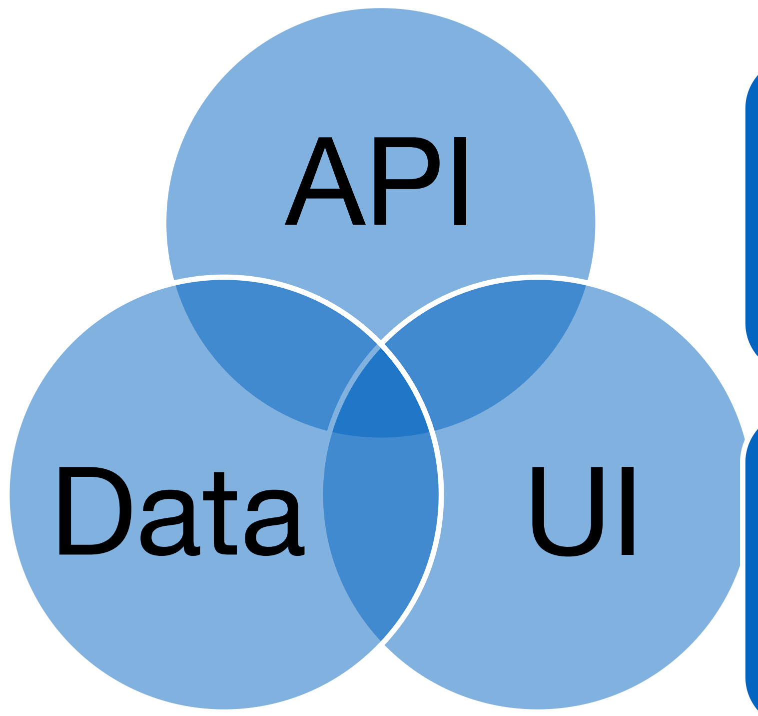
行业模版沉淀

- 行业模版
  电商、互娱治理模版
- 治理流程
  治理驾驶舱、治理运营、治理策略
- 治理能力原子化
  行业治理规则、治理操作



开放生态打造

- 接入
  - 元数据
  - 规则
  - 收益
- 配置
  - 数据团队
  - 资产范围
  - 运营流程
- 接出
  - 治理方案
  - 治理收益
  - 治理操作



大模型能力赋能

元数据沉淀
- 丰富
- 准确

模型能力建设
- 总结&推断
- SFT

关于我们

进入火山引擎DataLeap官网

了解更多产品信息

进入官方交流群

获取更多技术干货、活动信息

想一想，我该如何把这些技术应用在工作实践中？

THANKS