火山引擎 | DataFun.

字节跳动一站式数据治理思考及业务实践

王慧祥　火山引擎DataLeap资深大数据工程师

目录 CONTENT

01 机遇与挑战

02 数据治理思路

03 技术架构演进

04 未来展望

01
机遇与挑战

数据治理挑战：落地难

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，全局策略难落地
- 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

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

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

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、协商工具缺失
- 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

字节特色

字节文化

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

业务第一

规模大

- 业务场景丰富
  - 互娱
  - 资讯
  - 电商
  - 企业服务
  - 商业化
- 海量数据

数据驱动

- 产品闭环
- 业务强依赖数据
  - 商业分析
  - 推荐算法
  - 数据赋能

影响大

- 业务影响
  - 数据延迟
  - 质量问题
  - 数据生命周期

02

数据治理思路

新型数据治理-分布式数据自治

数据治理(Data governance)：治理收益(Profit)、业务影响(Influence)、执行效率(Efficiency)

- 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，何时、何地、如何进行数据治理
  - 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
  - 提供自上而下的规划性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
  - 产品提供完整的开发能力，业务根据自身特性和发展阶段进行接入

集中式 VS 分布式

集中式-传统数据治理

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

权责与管理：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

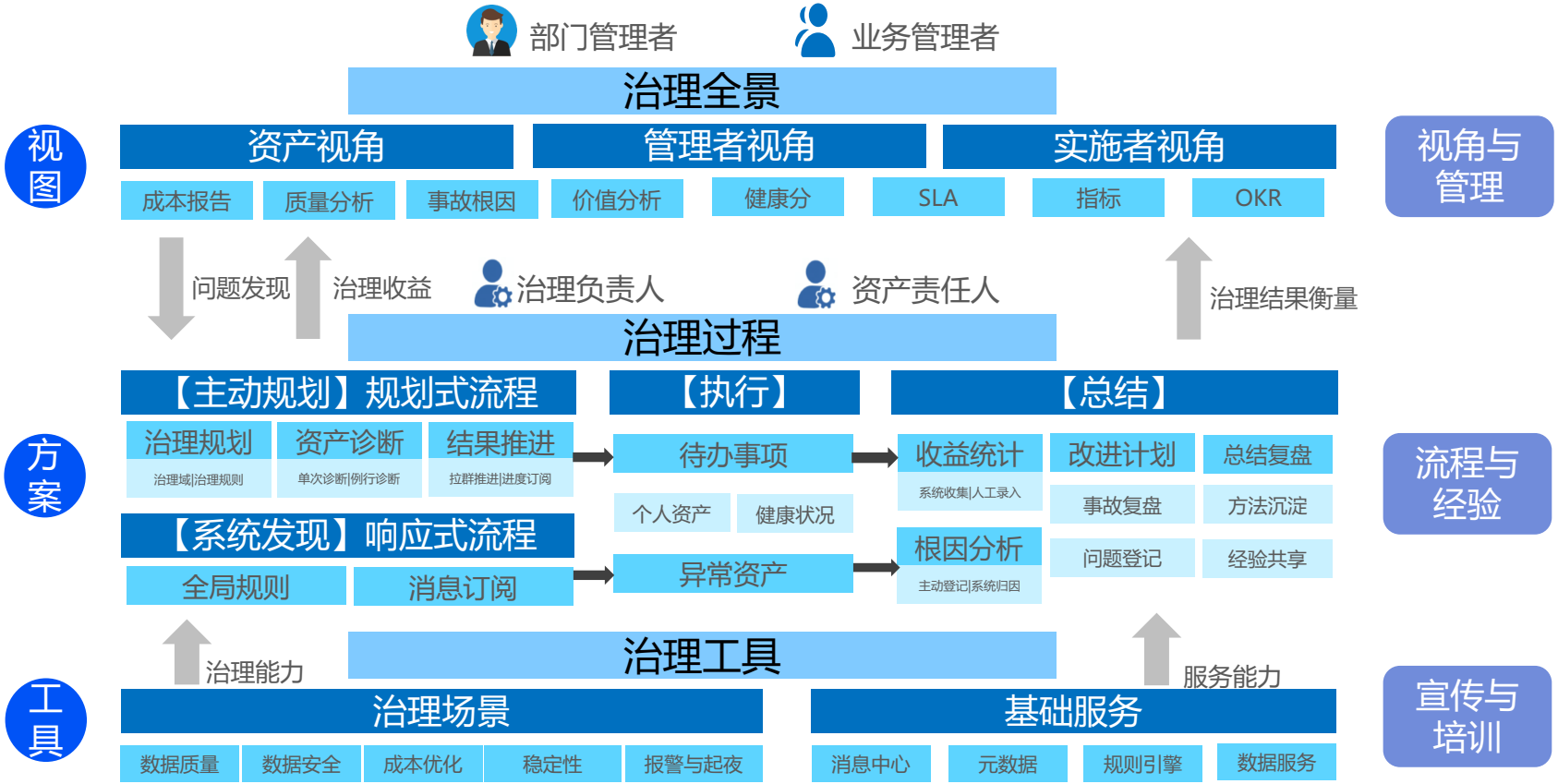
建设周期长、适配能力弱、组织投入多

分布式数据自治

- 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
  - 治理目标对齐、实施、追踪、核算工具化
  - 节省沟通成本，提升协作效率
- 算清账，降成本
  - 治理目标清晰，收益统计自动化

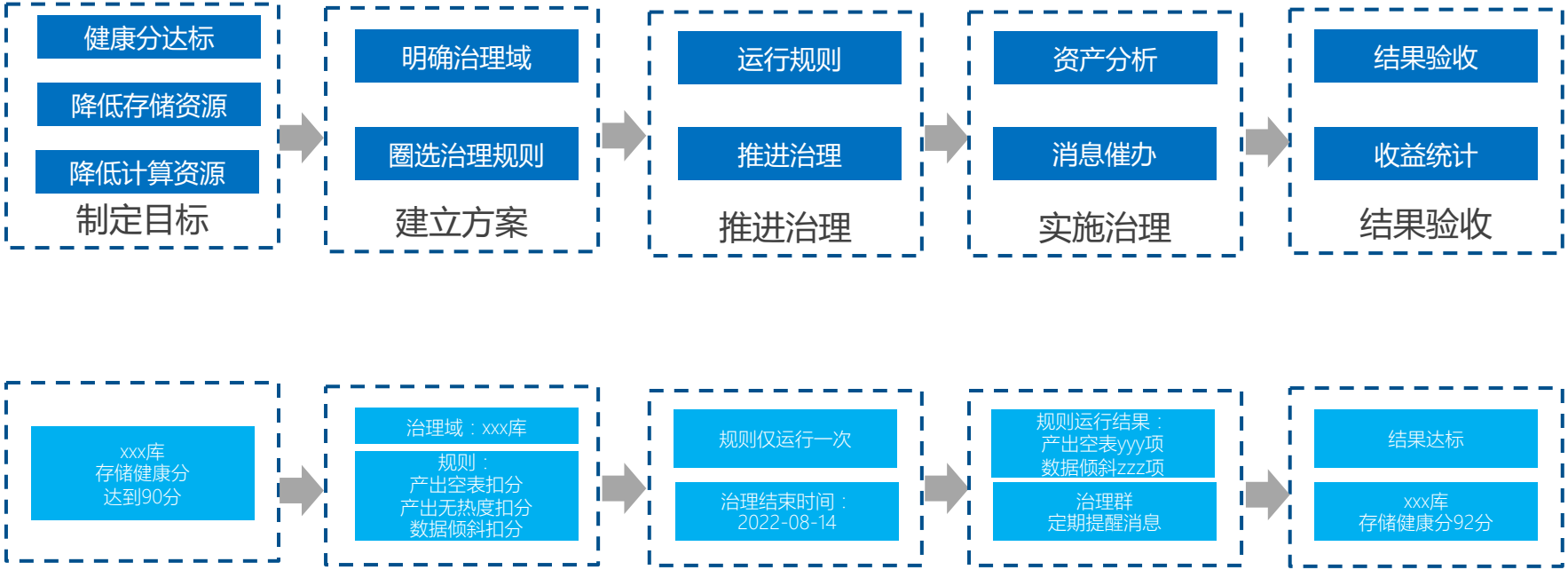
03
技术架构演进

解决方案-一站式



平台建设-治理方案-规划式流程

规划式治理：**资产清晰、规则丰富、动线完整、收益准确**



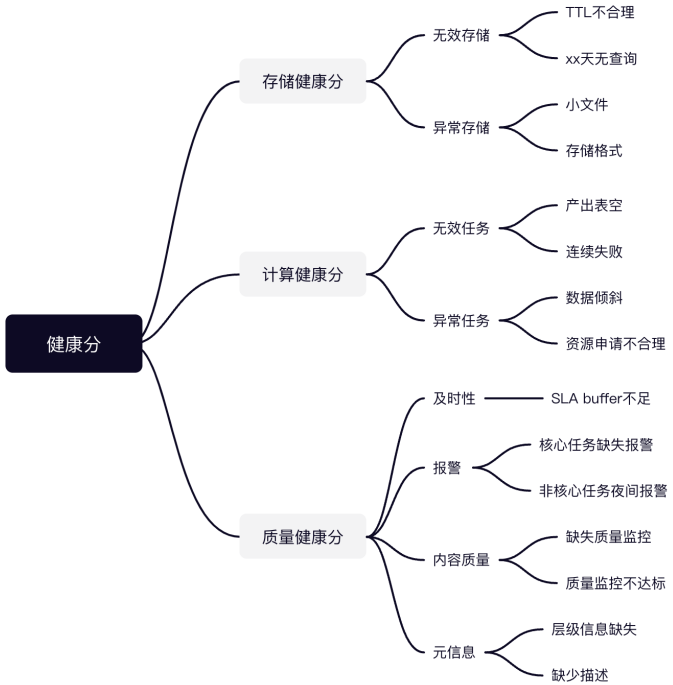
平台建设-治理方案-规划式流程-资产清晰

治理全景

有哪些资产？

健康分

评价体系



平台建设-治理方案-规划式流程-规则丰富

- 完备的治理规则能力
- 存储、计算、质量、报警4大维度（50+）
- 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

完善工具治理能力 提升治理执行效率 业务自定义接入

平台建设-治理方案-规划式流程-动线完整

治理全景 | 健康分 | 诊断规划 | 资产大盘

资产诊断

任务治理

- 关闭任务
- 调整参数
- 调整任务
- 修改队列
- 运行监控
- 链路优化

库表规范

- 表管理
- 库管理
- 资产移交
- 属性修改

生命周期

- 设置TTL
- 删除表
- 降副本
- 温存

数据质量

- SLA申报
- 离线数据监控
- 实时数据监控

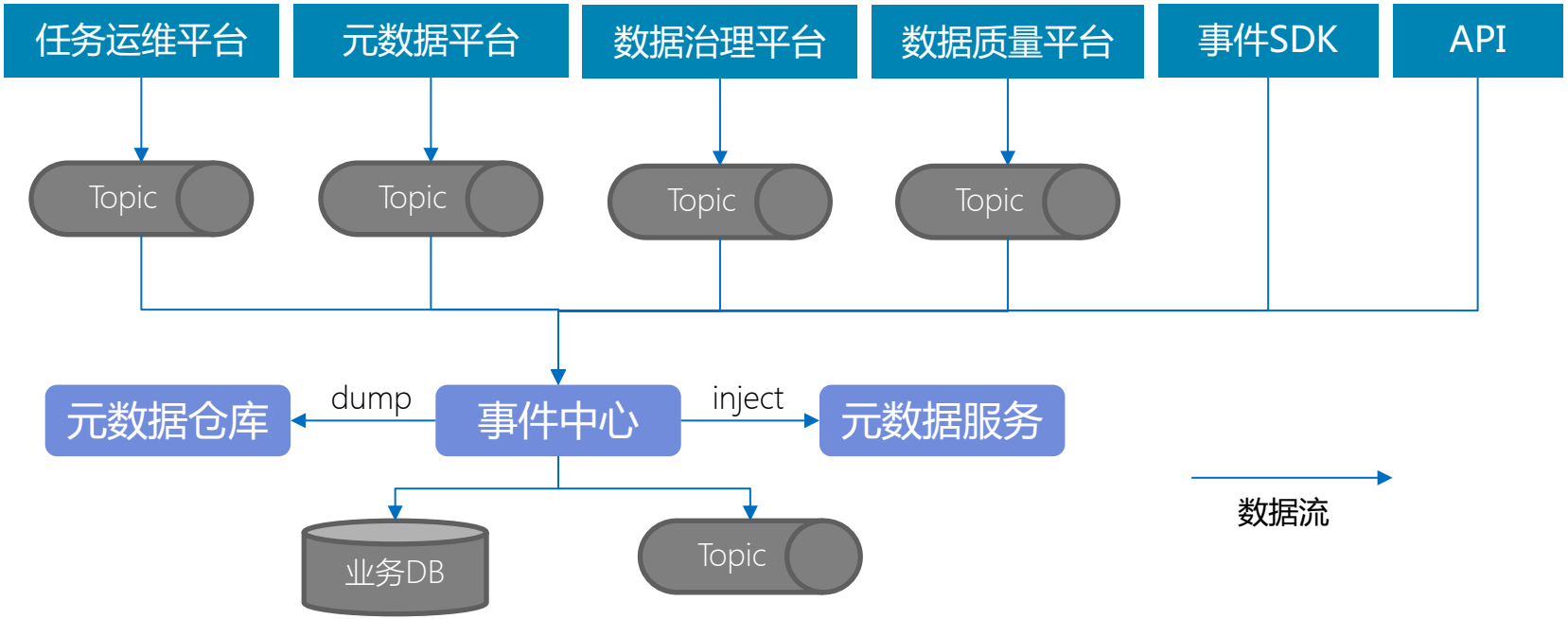
治理动线

任务运维平台 | 元数据平台 | 数据治理平台 | 数据质量平台

HDFS | YARN | 调度系统 | 消息中心 | HIVE | SPARK | FLINK

平台建设-治理方案-规划式流程-收益准确

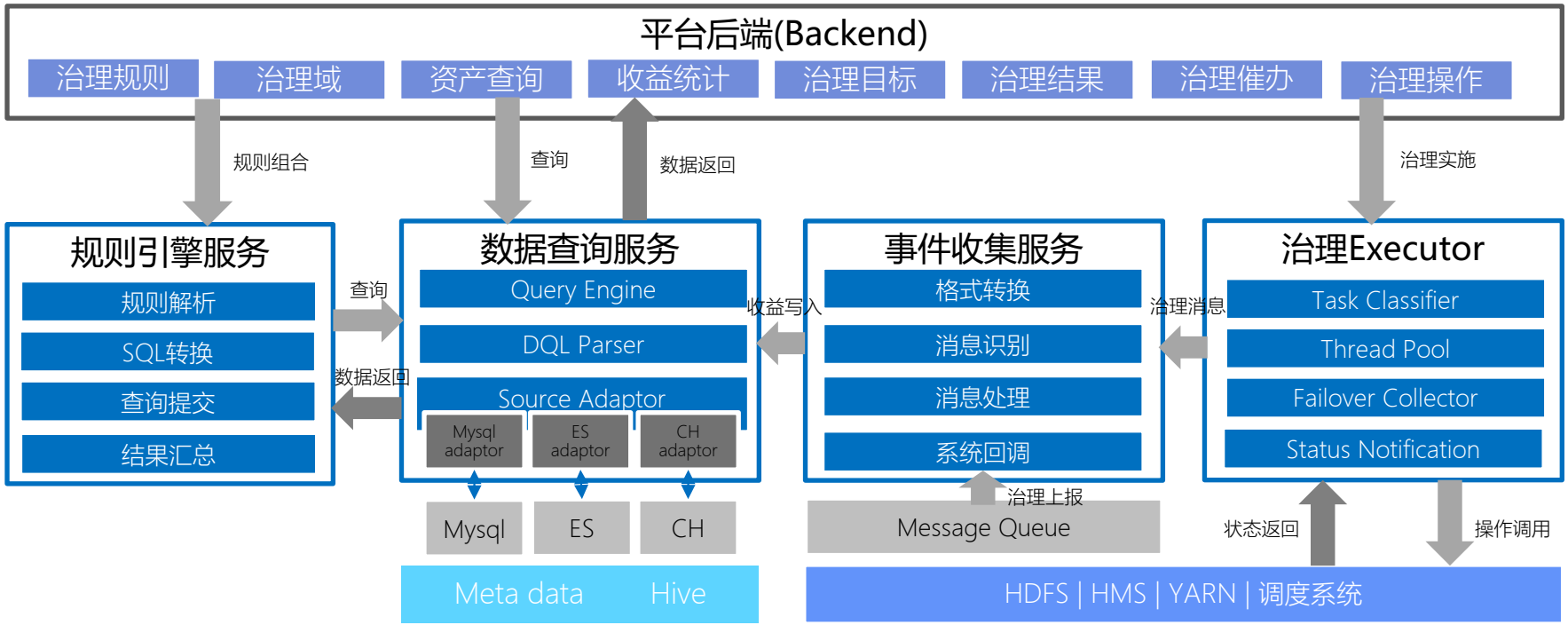
治理目标：存储、计算、健康分



思路：行为埋点、事件上报、关联计算

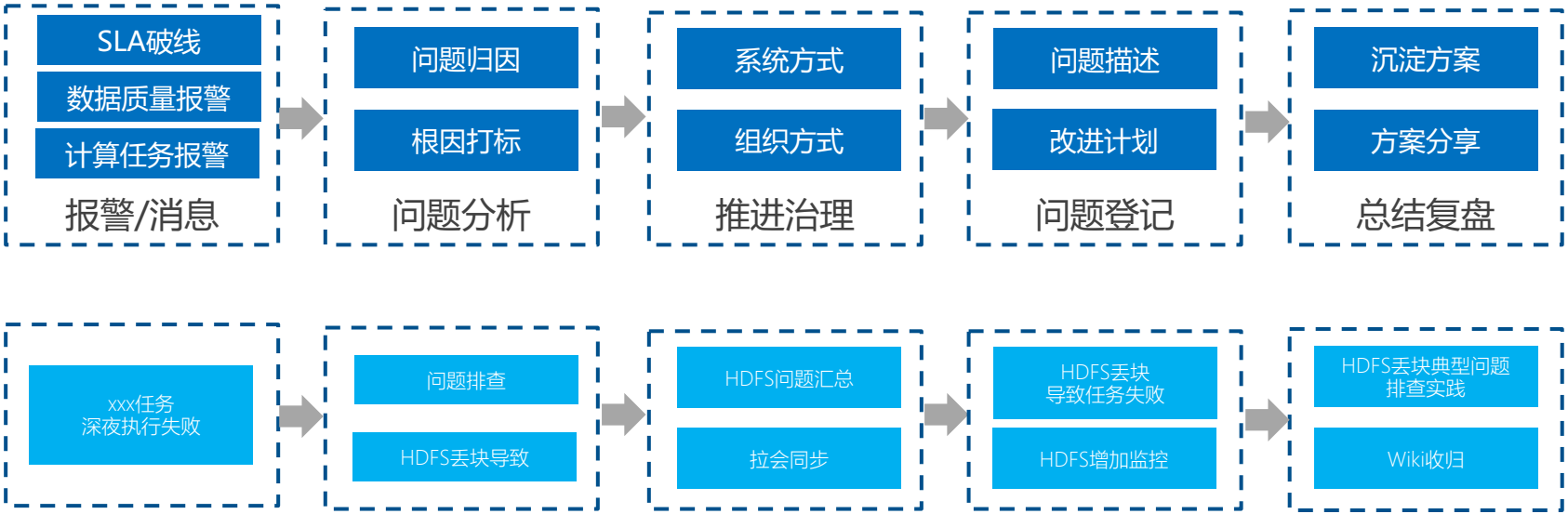
平台建设-治理方案-规划式流程-技术架构

数据查询统一、规则组合灵活、治理操作解耦、治理收益准确



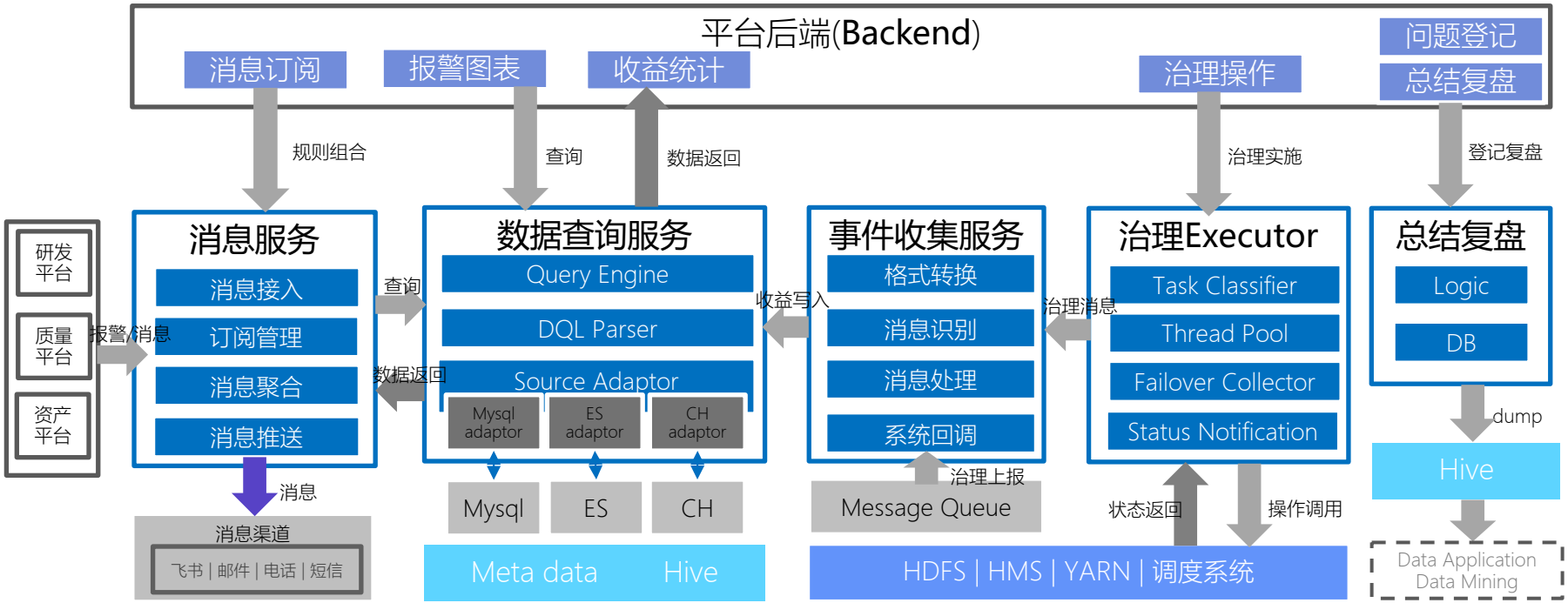
平台建设-治理方案-响应式流程

响应式治理：**事后治理、问题总结、经验沉淀**



平台建设-治理方案-响应式流程-技术架构

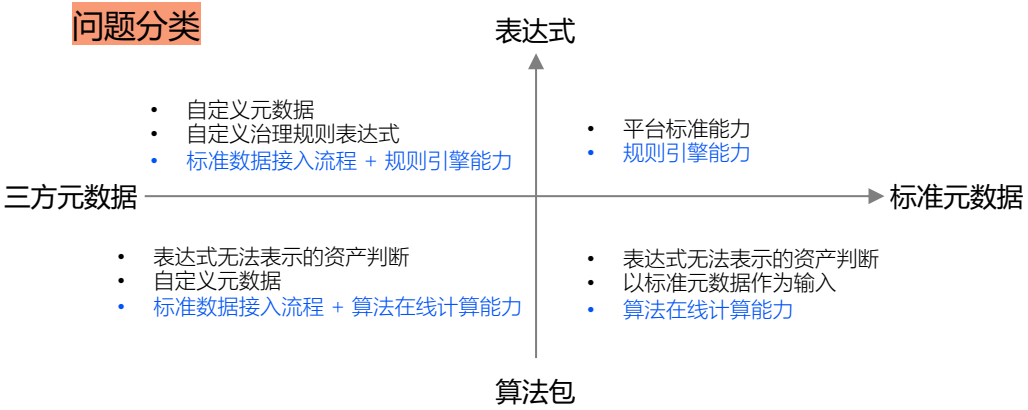
报警、消息统一收归，问题与复盘链路完整

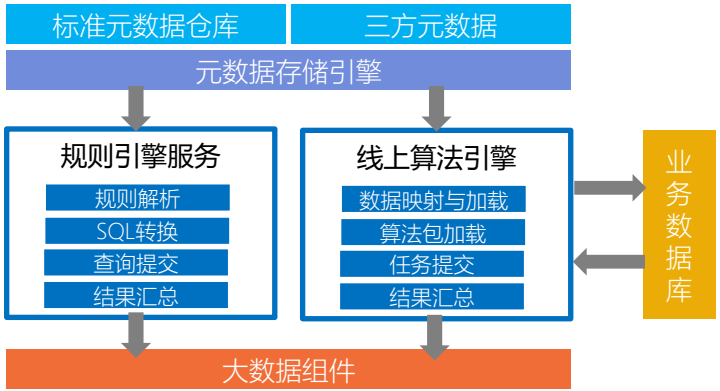


平台建设-开放接入

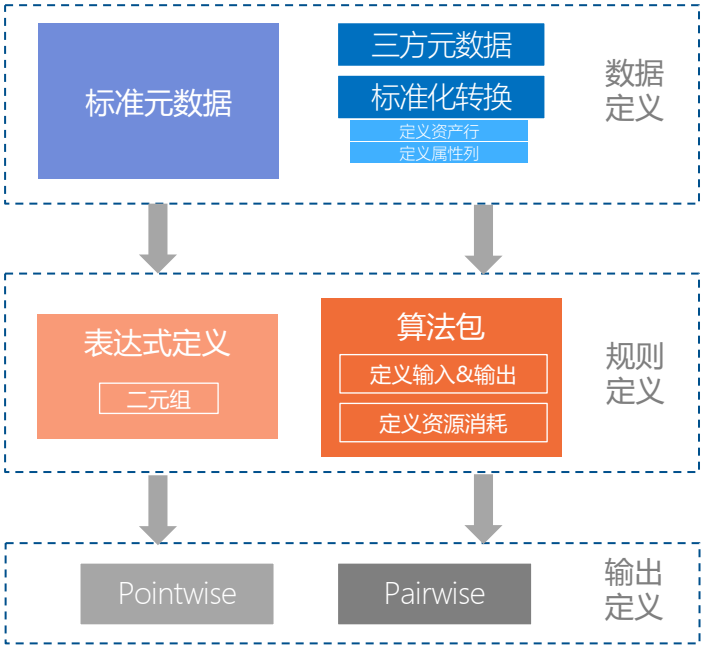
构建治理生态、建设开放平台，不同业务快速、灵活接入

问题分类





接入流程

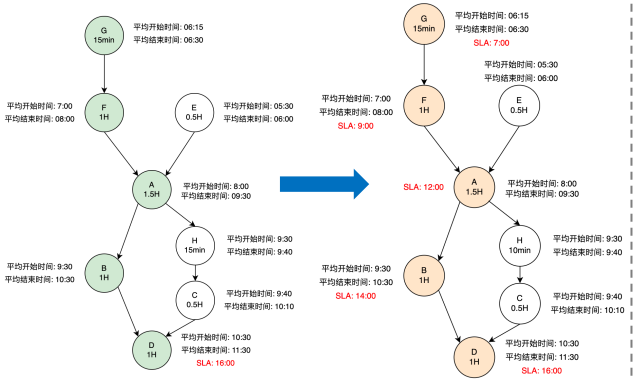


平台建设-智能化能力

挖掘数据价值，提效数据治理

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



$$buffer = suggestSLA - avgCompleteTime_n$$

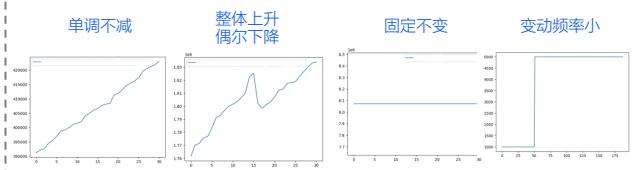
$$totalRunningTime = \sum_{i \in \Lambda} avgRunningTime_i$$

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

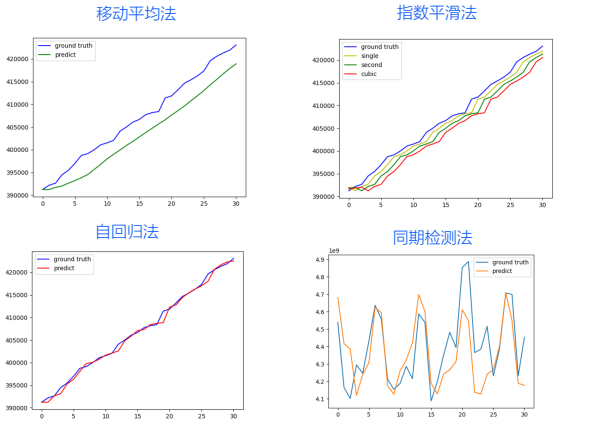
$$ptb = \frac{buffer}{totalRunningTime}$$

动态阈值监控

- 报警阈值 = 预测表行数 * 倍数
- 数据分布

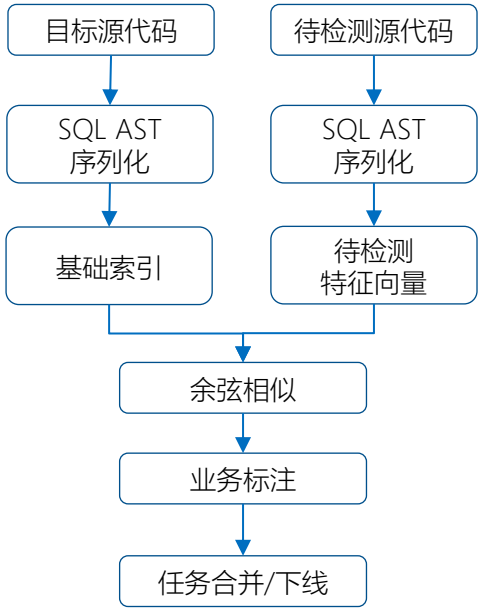


- 预测方法

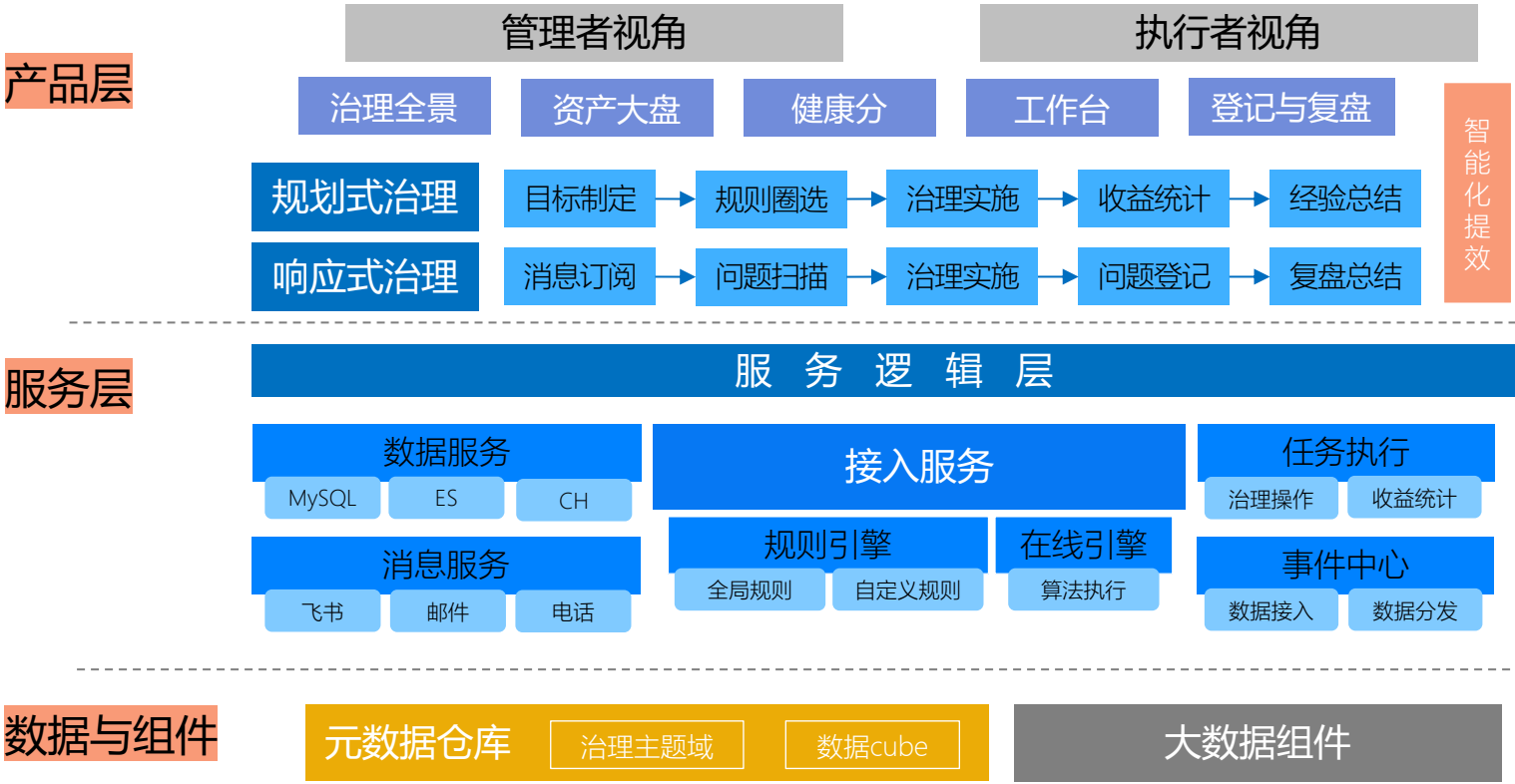


相似任务识别

- 业务合作典范



平台建设-架构总结



04
未来展望

未来展望

体验打磨

资产清晰 → 规则丰富 → 动线完整 → 收益准确

问题登记 → 根因归因 → 归纳总结 → 经验沉淀

开放能力

1. 自定义指标：自定义健康分、自定义组织（团队）
2. 自定义方案：自定义治理规则、灵活维度组合条件
3. 业务打通：以业务的视角看治理问题，治理增加业务属性

增强型数据治理

Data Statistics
- 事后处理
- 统计类规则

→

Data Mining
- 事中处理
- 推荐治理

→

Artificial Intelligence
- 事前避免
- 预测与预处理

欢迎联系我们

扫码关注

“字节跳动数据平台”微信公众号

扫码添加小助手

进入“字节跳动数据平台”官方交流群

非常感谢您的观看

火山引擎 | DataFun.