货拉拉数据治理平台建设实践

陈元 货拉拉 资深大数据工程师
张放 货拉拉 高级大数据工程师

目录 CONTENT

01 货拉拉数据治理体系

02 数据质量平台建设实践

03 元数据平台建设实践

01 货拉拉数据治理体系

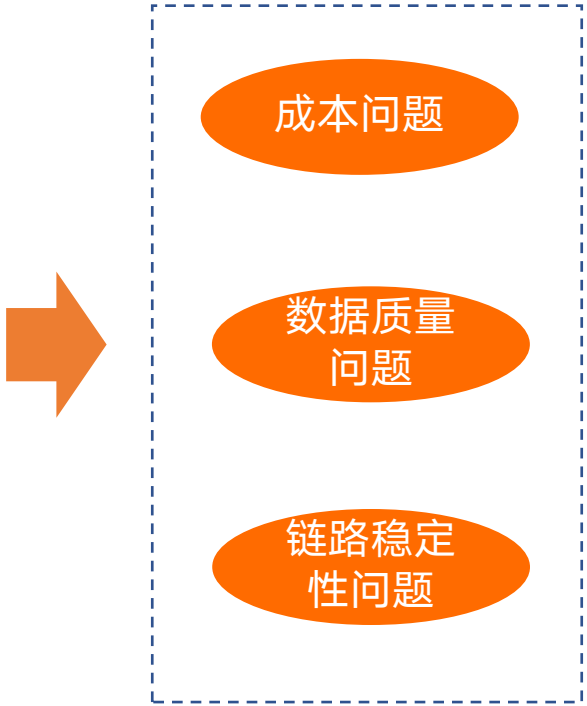
背景与现状

- 应用数据没有生命周期管理
- 链路长不稳定，产出不及时
- 缺少监控规范，数据问题不能及时发现

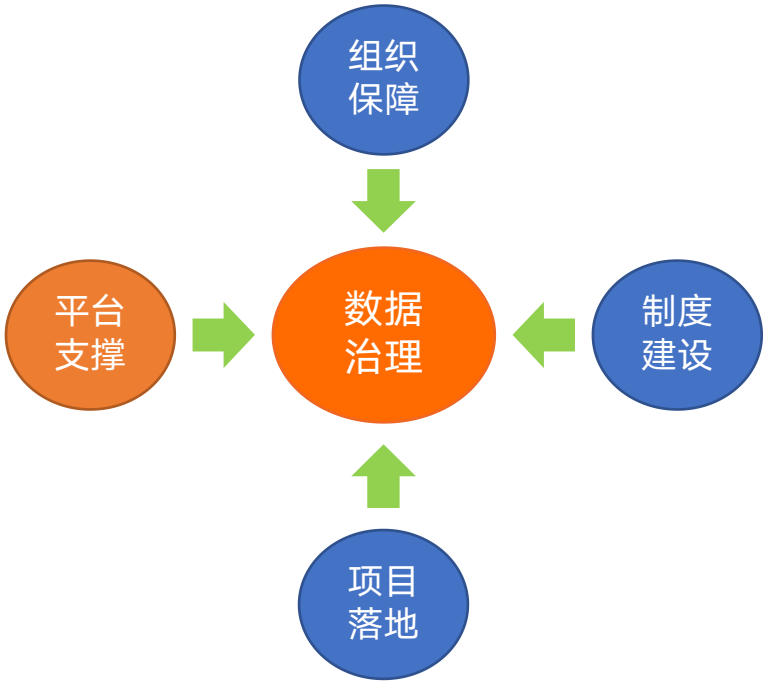
- 数据表没有生命周期
- 无效任务空跑
- 任务发布流程不规范
- 缺乏全局的模型规范
- 实时数仓缺乏公共层

- 大量全量抽数
- 实时流没压缩
- 不合理的分区策略

- 上游数据变更引发故障
- 模型重构
- 数据字典不规范

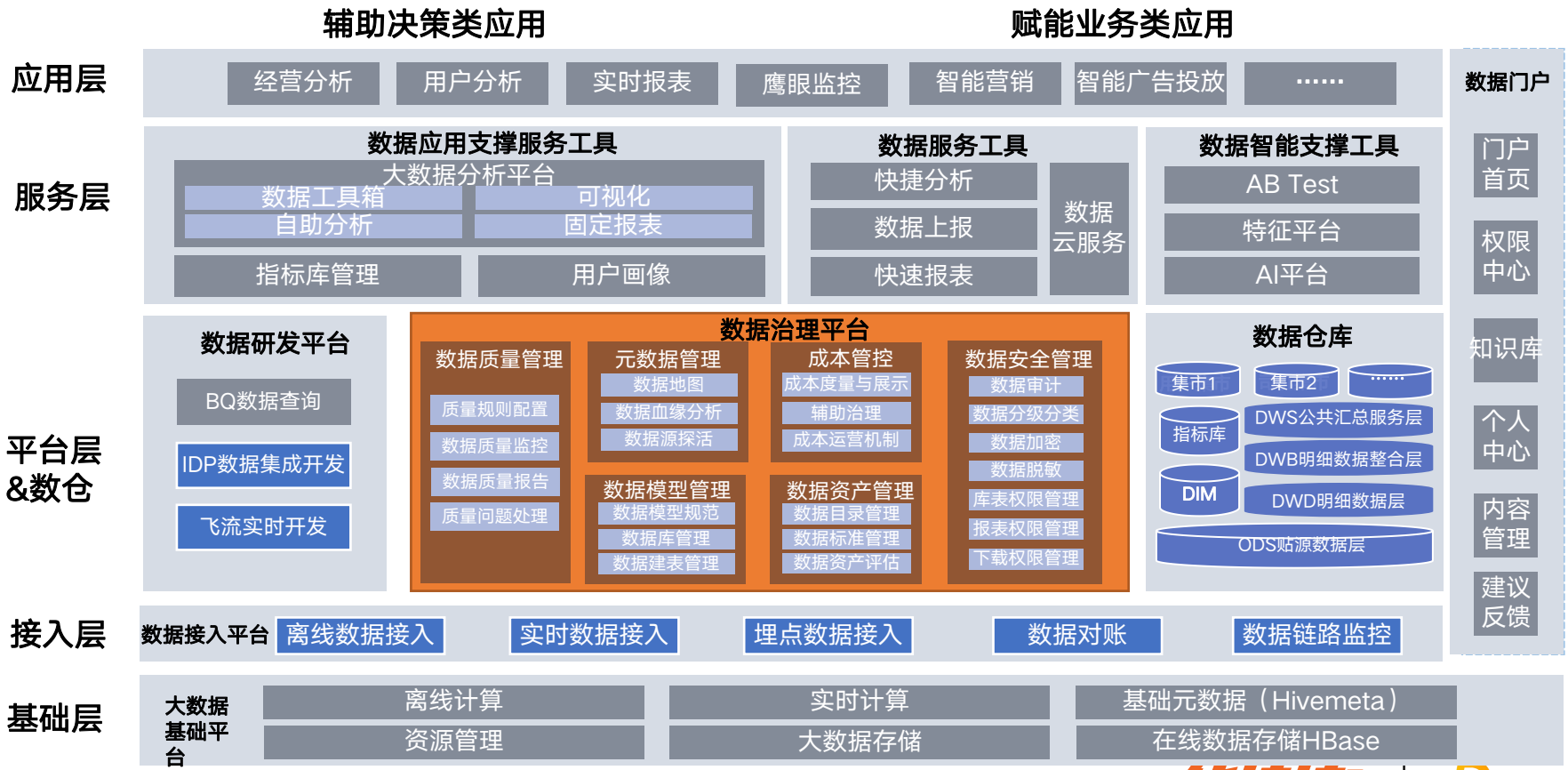


数据治理关键环节



- 组织保障：明确职责分工
- 制度建设：制定标准流程，保障落实执行
- 项目落地：贴合业务，追踪成效
- 平台支撑：研发支撑系统，提质增效

货拉拉数据治理产品体系

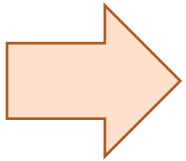


02 数据质量平台建设实践

面临的数据质量问题有哪些？

- 表未按时产出
- 上游表数据错误污染下游
- 埋点数据丢失
- 报表指标数据异常
- ……

影响

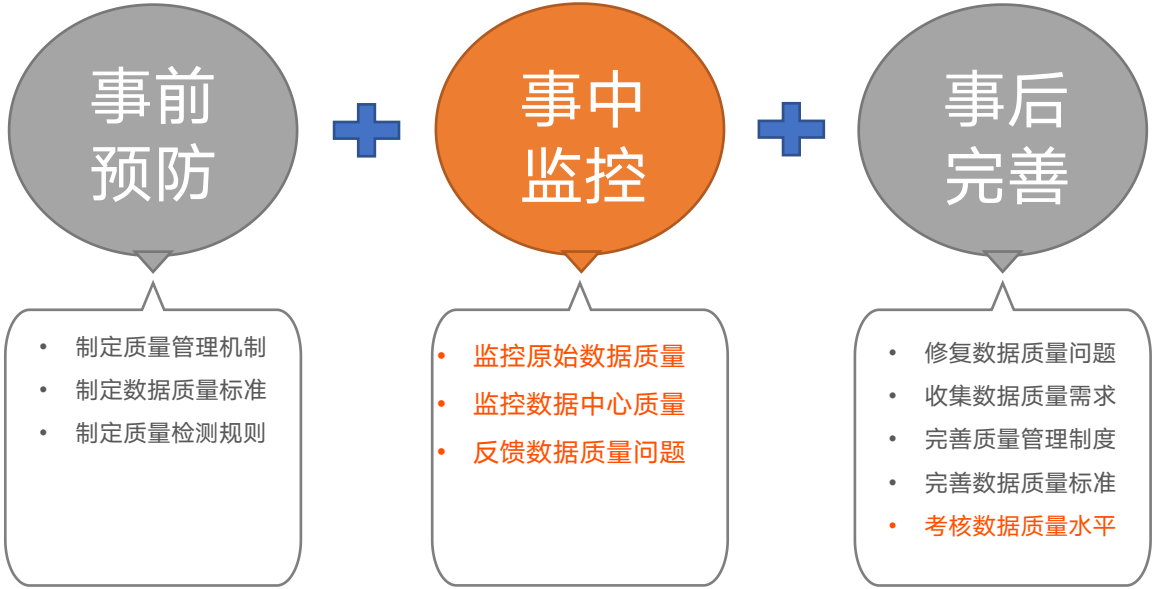


- 报表延迟
- 链路异常
- 数据丢失
- 决策错误

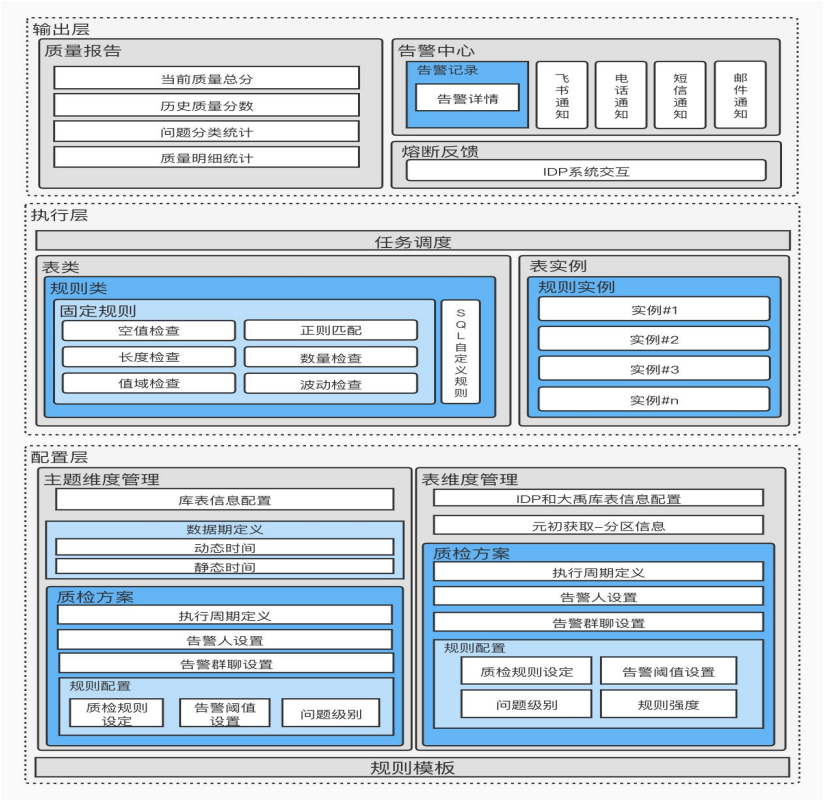
产生数据质量问题的原因

业务端	技术端	基础设施	管理端
• 业务源系统变更 • 业务端数据输入不规范	• 数据开发 SQL bug • 任务流程、参数、配置错误	• 计算资源不足 • 机器硬件问题	• 缺乏质量意识 • 缺乏有效质量问题处理机制

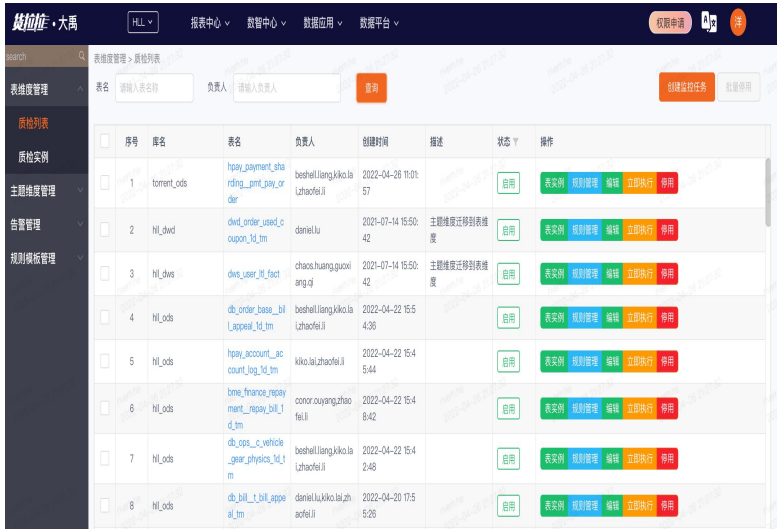
数据质量保障思路


事前
预防
事中
监控
事后
完善
制定质量管理机制
制定数据质量标准
制定质量检测规则
监控原始数据质量
监控数据中心质量
反馈数据质量问题
修复数据质量问题
收集数据质量需求
完善质量管理制度
完善数据质量标准
考核数据质量水平

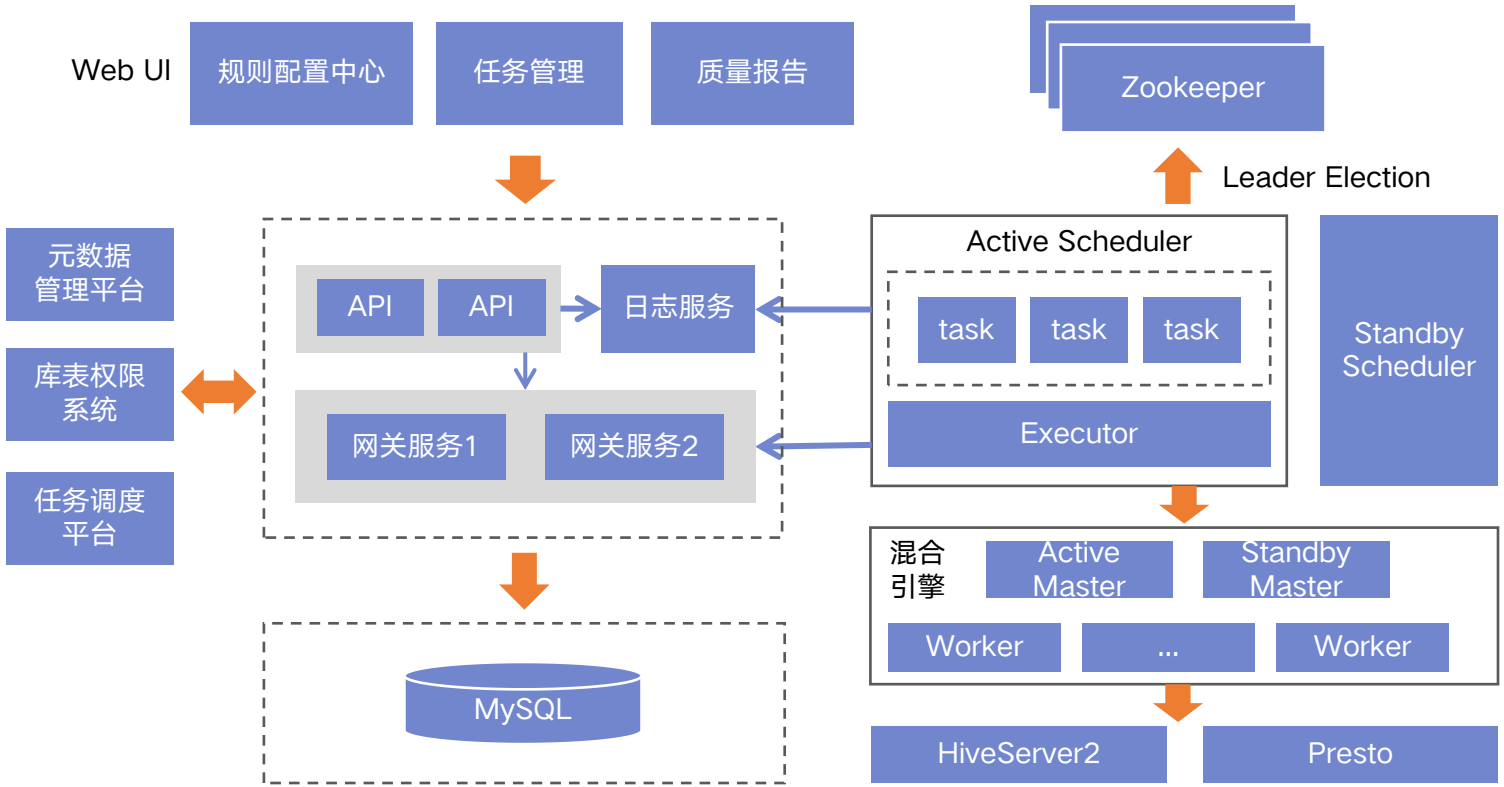
货拉拉数据质量平台



- “零”代码一站式质量检测
- 全链路监控
- 全方位质检报告



货拉拉数据质量平台-系统架构



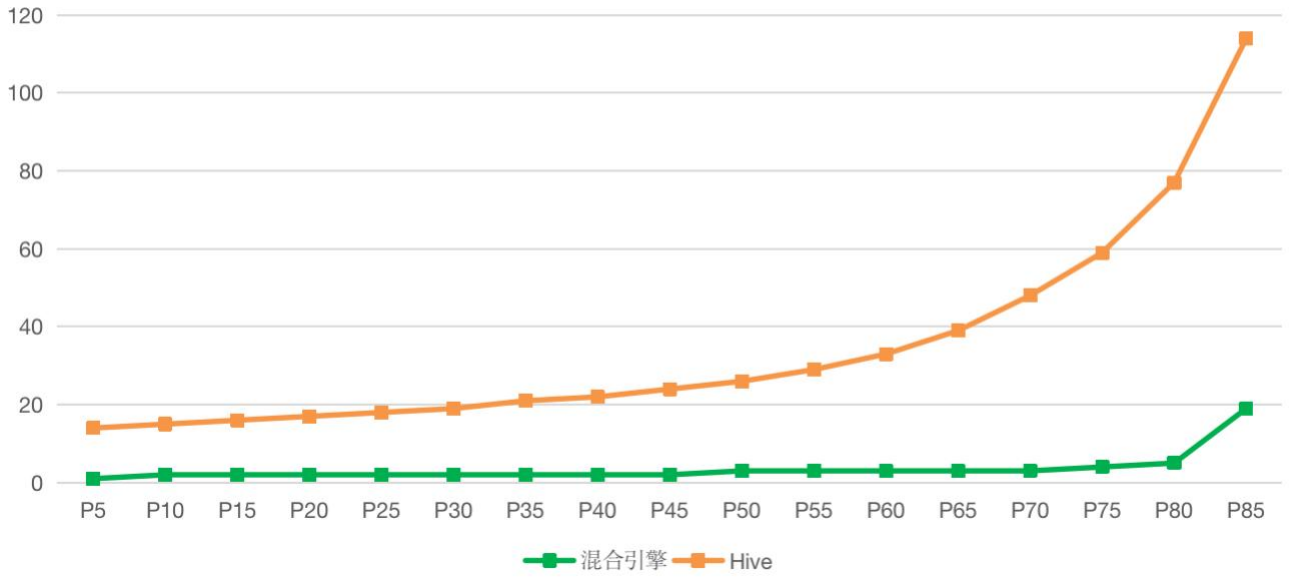
- 高稳定
- 高效率
- 熔断阻塞

数据质量分析效率

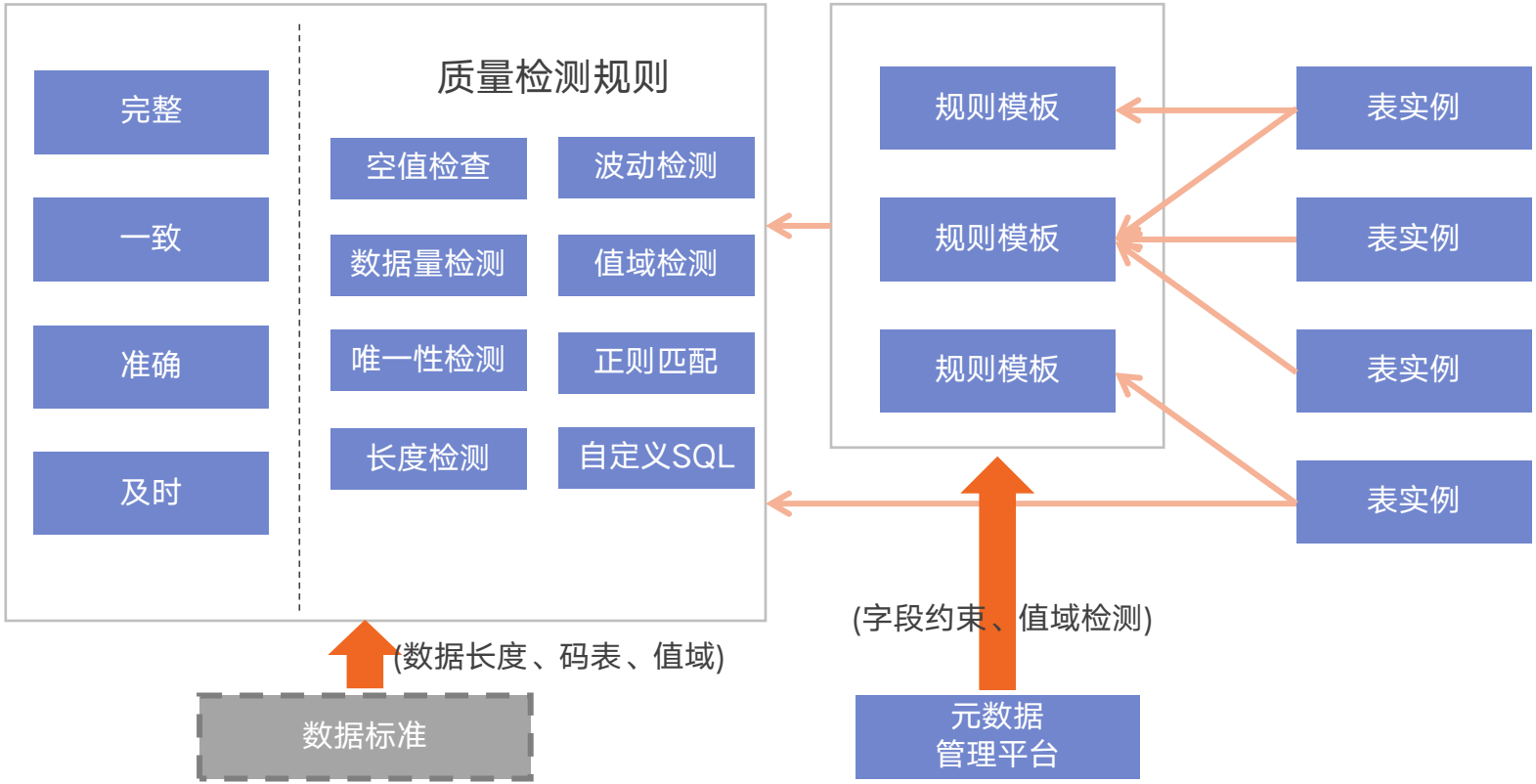
分位	P5	P10	P15	P20	P25	P30	P35	P40	P45	P50	P55	P60	P65	P70	P75	P80	P85	P90	P95	P98
混合引擎	1	2	2	2	2	2	2	2	2	3	3	3	3	3	4	5	19	43	82	192
Hive	14	15	16	17	18	19	21	22	24	26	29	33	39	48	59	77	114	196	396	915
耗时下降	93%	87%	88%	88%	89%	89%	90%	91%	92%	88%	90%	91%	92%	94%	93%	94%	83%	78%	79%	79%

使用混合引擎(Presto)提升分析效率：

- P80质量检测任务5秒内完成
- P98质量检测任务提速79%
  （915s -> 192s)

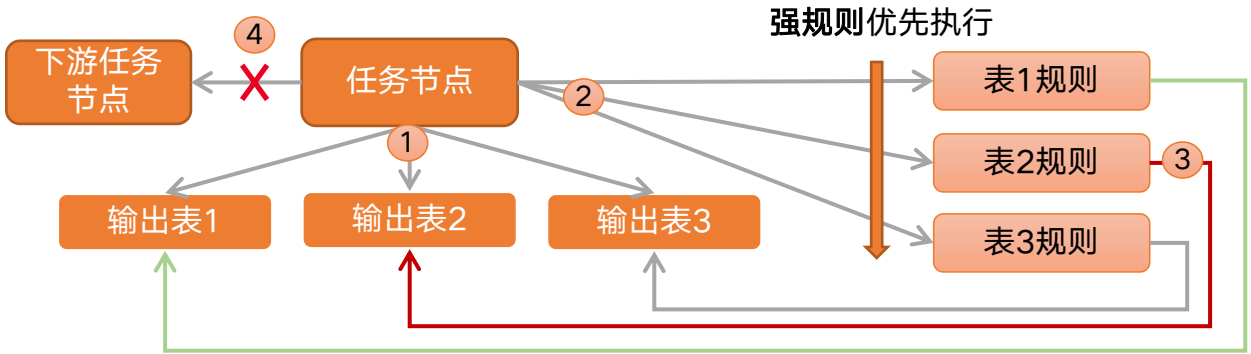


数据质量平台-规则体系



数据质量平台-熔断机制

- 质量问题级别：
  一般、重要、严重
- 强规则：
  强规则不通过，并且是严重质量问题，告警 + 阻塞下游任务节点
- 弱规则：
  弱规则不通过，只警告



1.周期性任务执行 → 2.触发质量规则校验 → 4.强规则不通过 → 5.周期性任务熔断阻塞

数据质量平台-质量报告

全方位质检报告：

- 多维度质量分析报表
- 多角度质量绩效评分
- 支持用户自定义评分依据和权重

数据质量平台-监控告警

监控告警：及时发现数据质量问题

告警级别与告警方式：

- 一般 -> 邮件
- 重要 -> 邮件+飞书
- 严重 -> 邮件 + 飞书 + 电话

数据质量平台

1500+ 张
接入的表数量

100%
核心链路表覆盖

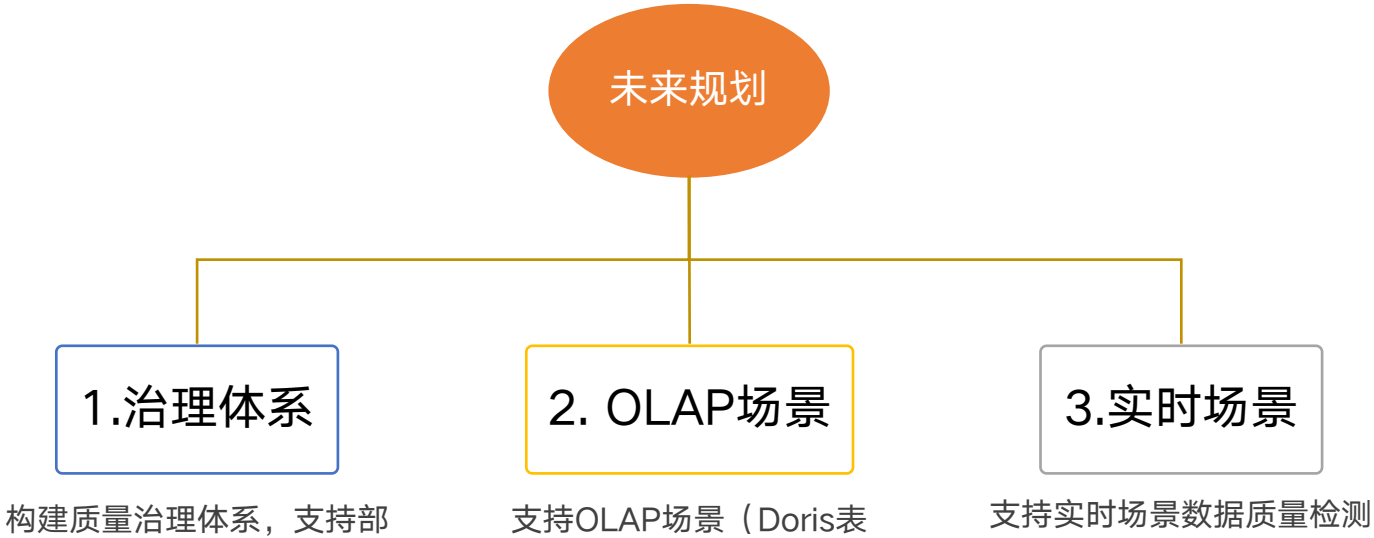
现状

300+ 次/月
检测到的数据质量问题数

14次
2022年累计熔断阻塞

有效保障了数据质量和链路稳定性

数据质量平台未来规划


未来规划
1.治理体系
构建质量治理体系，支持部门/人级别数据质量度量、展示和运营能力
2. OLAP场景
支持OLAP场景（Doris表等）数据质量检测
3.实时场景
支持实时场景数据质量检测

03 元数据管理平台实践

张放 货拉拉 高级大数据工程师

目录 CONTENT

01 平台介绍

02 成本治理体系

03 数据血缘

04 未来规划

平台介绍–系统架构

定规范

- ✓ 制定公司级数据模型规范
- ✓ 并逐步推广到所有部门

做治理

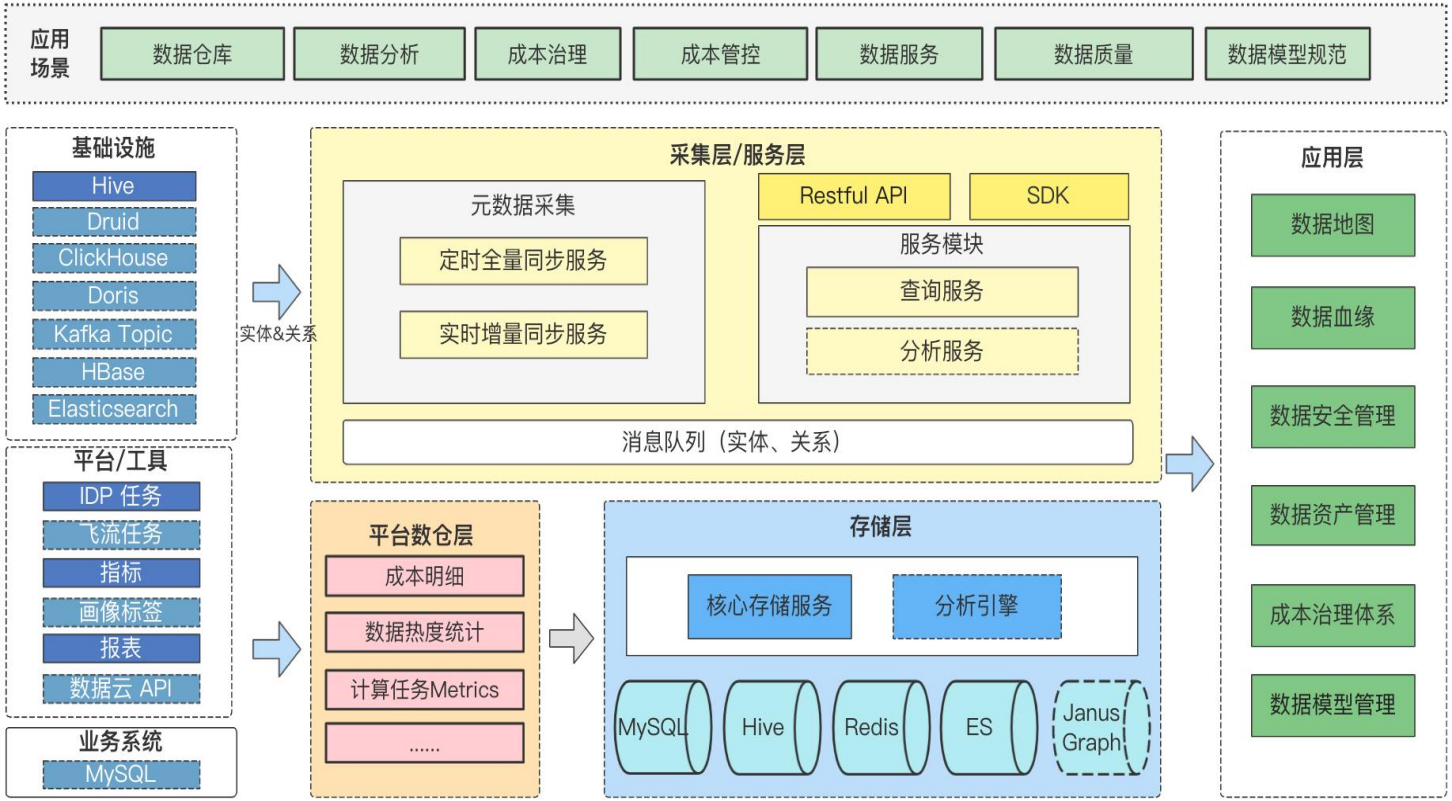
- ✓ 数据生命周期治理
- ✓ 离线存储治理
- ✓ 成本度量

建能力

数据模型管理数据地图、数据血缘、数据资产管理、成本治理体系 . . .

做运营

模型管理宣讲推广、生命周期覆盖、成本运营



平台介绍–对标业界

公司/产品	元数据基建		元数据应用			
	元模型	血缘建设	数据地图	数据管理	数据血缘	其他
快手	• 统一元模型 • 10+元数据类型 • 10W+任务、几十万Hive表	• 实时全链路血缘	• 自定义查询 • 分类检索 • 案例分享	• 生命周期 • 安全等级	• 血缘查询 • 优先级推导 • 下线检测	元数据服务：离线元数据仓库
B站	• 统一元模型 • 10+元数据实体类型 • 10种元数据关系类型 • 6W+ Hive表和11W+ 任务数	• 全链路血缘 • 字段级血缘	• 类SQL查 • 关联查询 • 分类查询 • 热度推荐	• 生命周期 • 安全等级	• 血缘地图 • 影响分析	元数据质量：自动化的采集质量问题解决 数据画像：样例数据和数据探查
网易	6+元数据实体类型	• 字段级血缘	多维度检索数据预览		字段级溯源 血缘的生命周期管理	元数据画像：元数据标签（技术标签、业务标签）
aly DLF&DataWorks	10+表格式 支持API类型 支持数据湖格式	• 字段级血缘	DataWorks(数据地图)	元数据分析和管理、成本分析与优化、冷热分析、有效性分析、安全度分析、性能优化、数据生命周期管理	• 血缘地图 • 影响分析	元数据服务：兼容HMS协议、支持多引擎访问 数据开发提效：元数据驱动的数据建模、驱动ETL
元初	Hive元模型	表级血缘	数据资产目录 元数据检索	生命周期 冷热分层	血缘查询 影响分析	成本治理体系 数据资产管理

存储治理–面对的问题

无数据生命周期

无冷热分层管理

无成本度量体系

表数量大

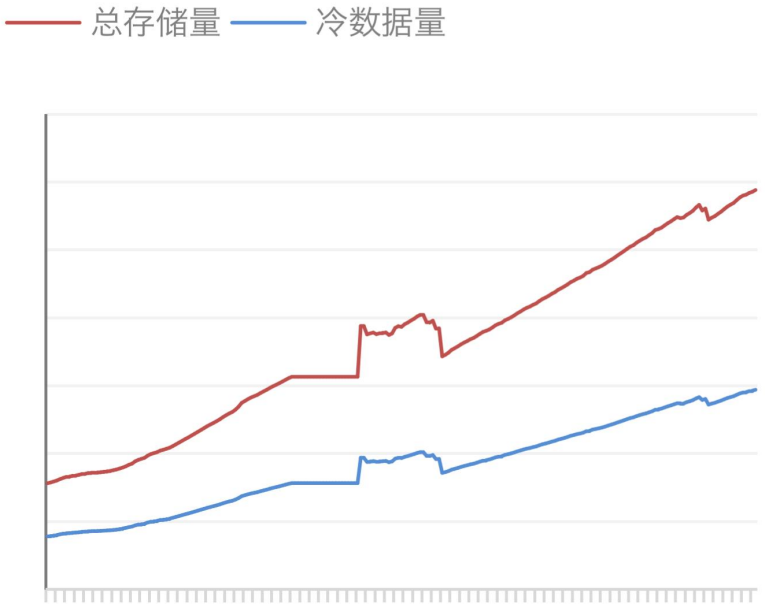
17W+ Hive表

增速快

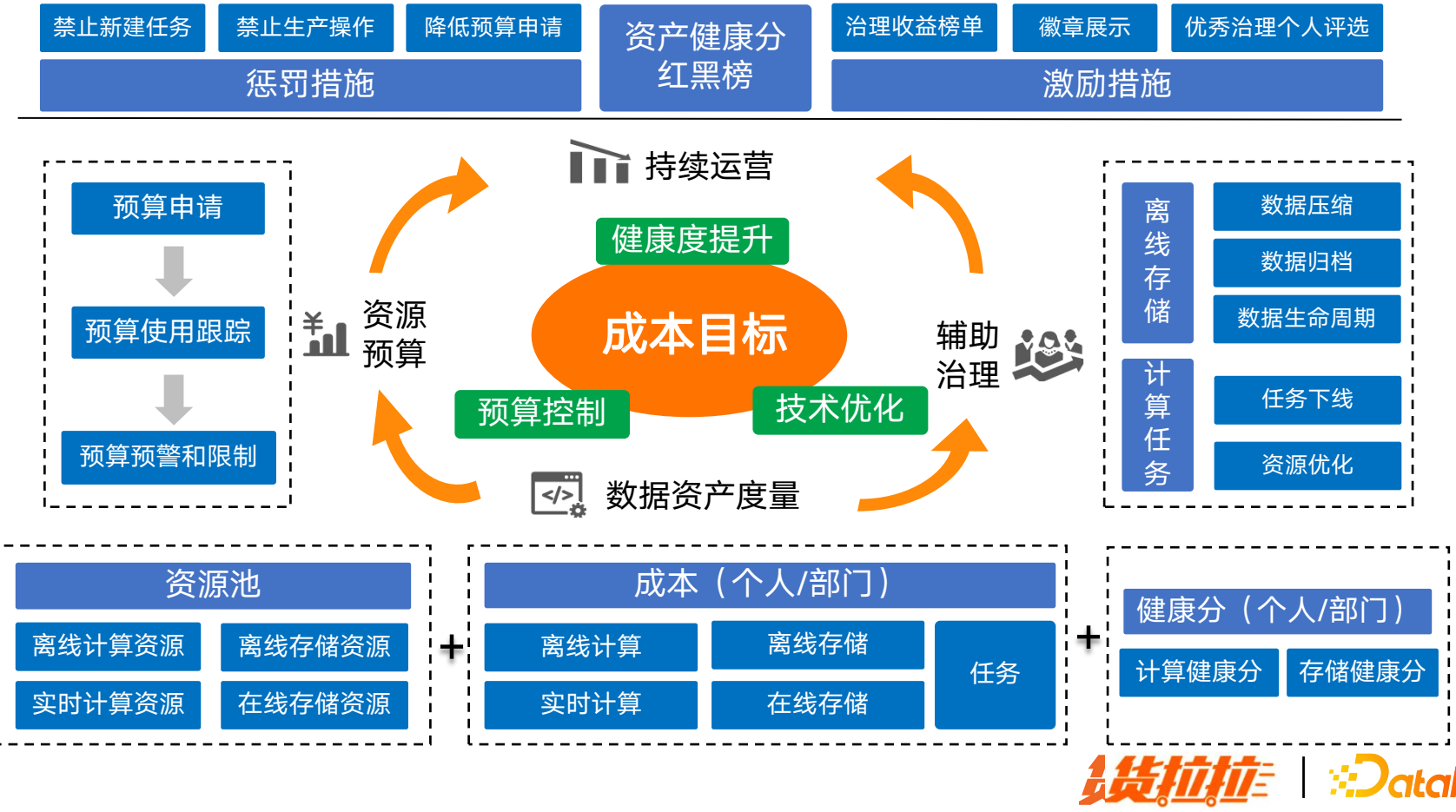
4PB/月自然增长

冷数据量大

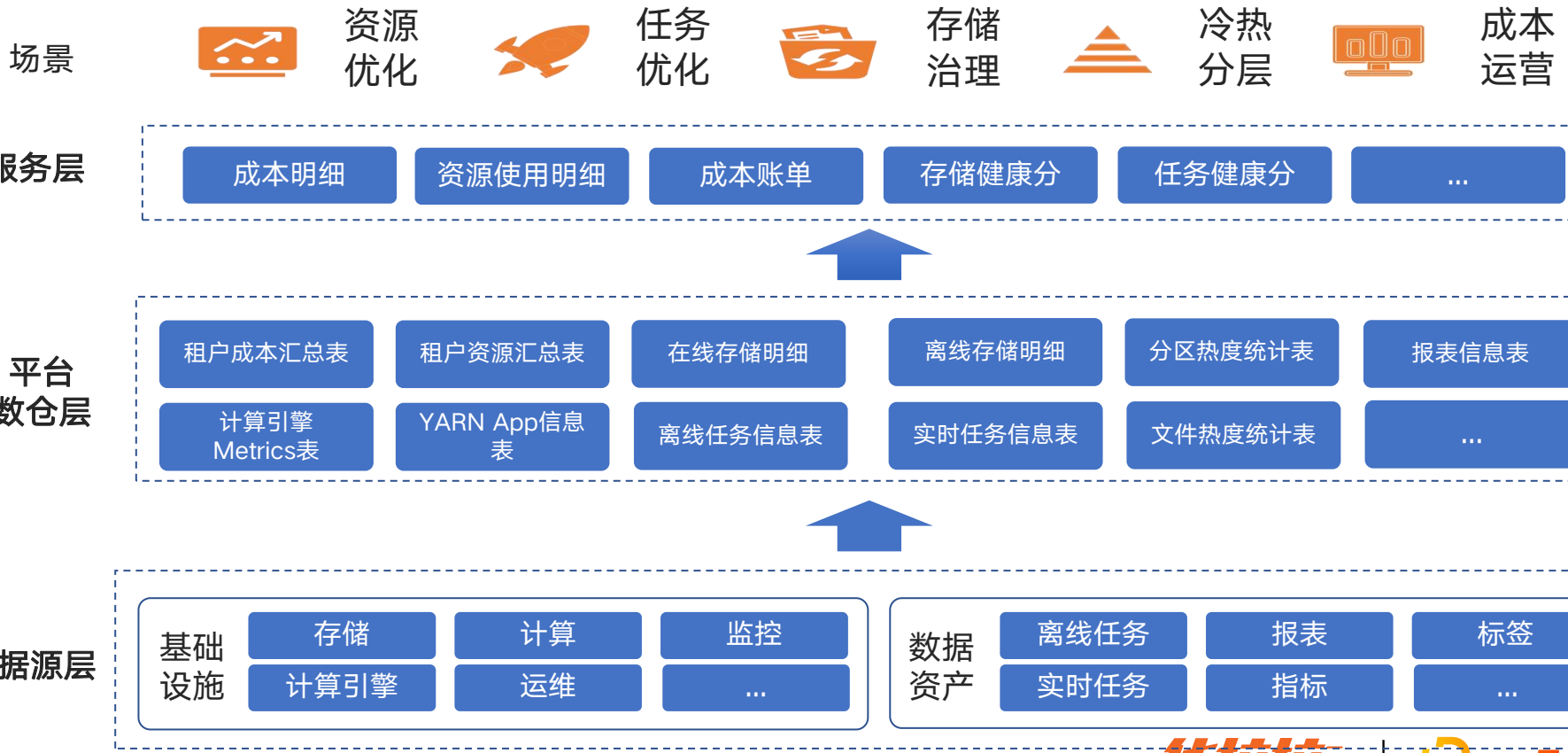
约33%数据90天内无访问



成本治理体系

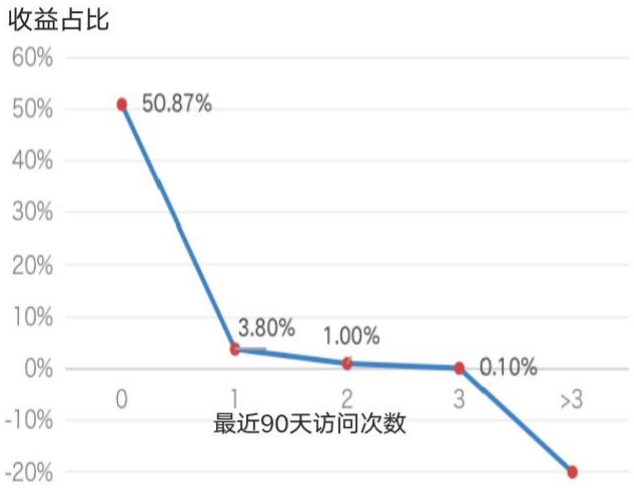


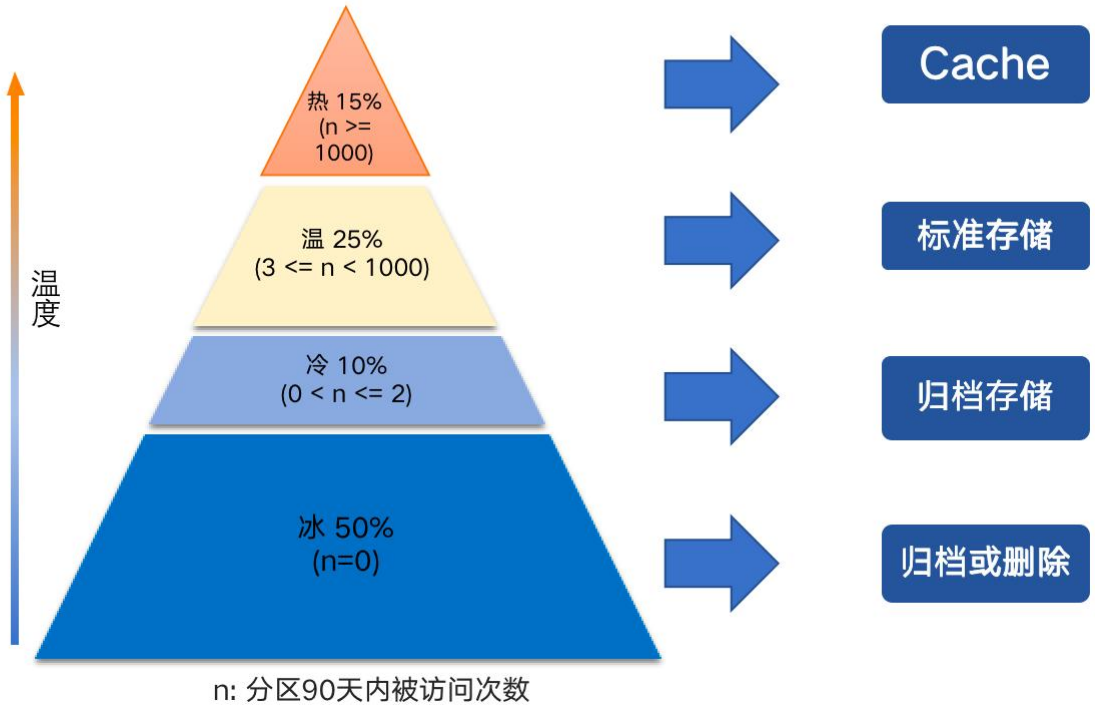
成本治理体系–成本度量和展示



成本治理体系–冷热分层和归档

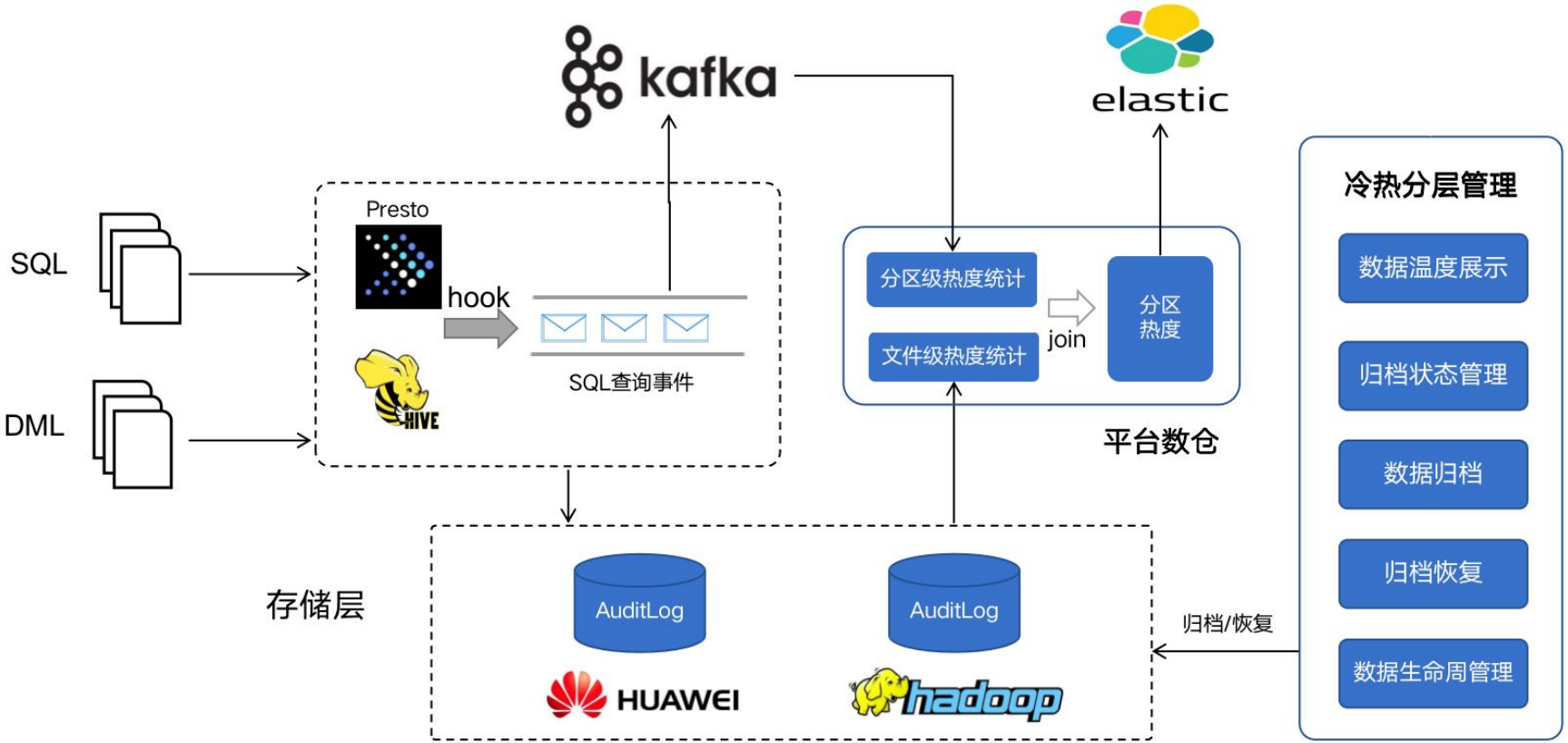
归档收益曲线 → 冷热分层 → 存储策略



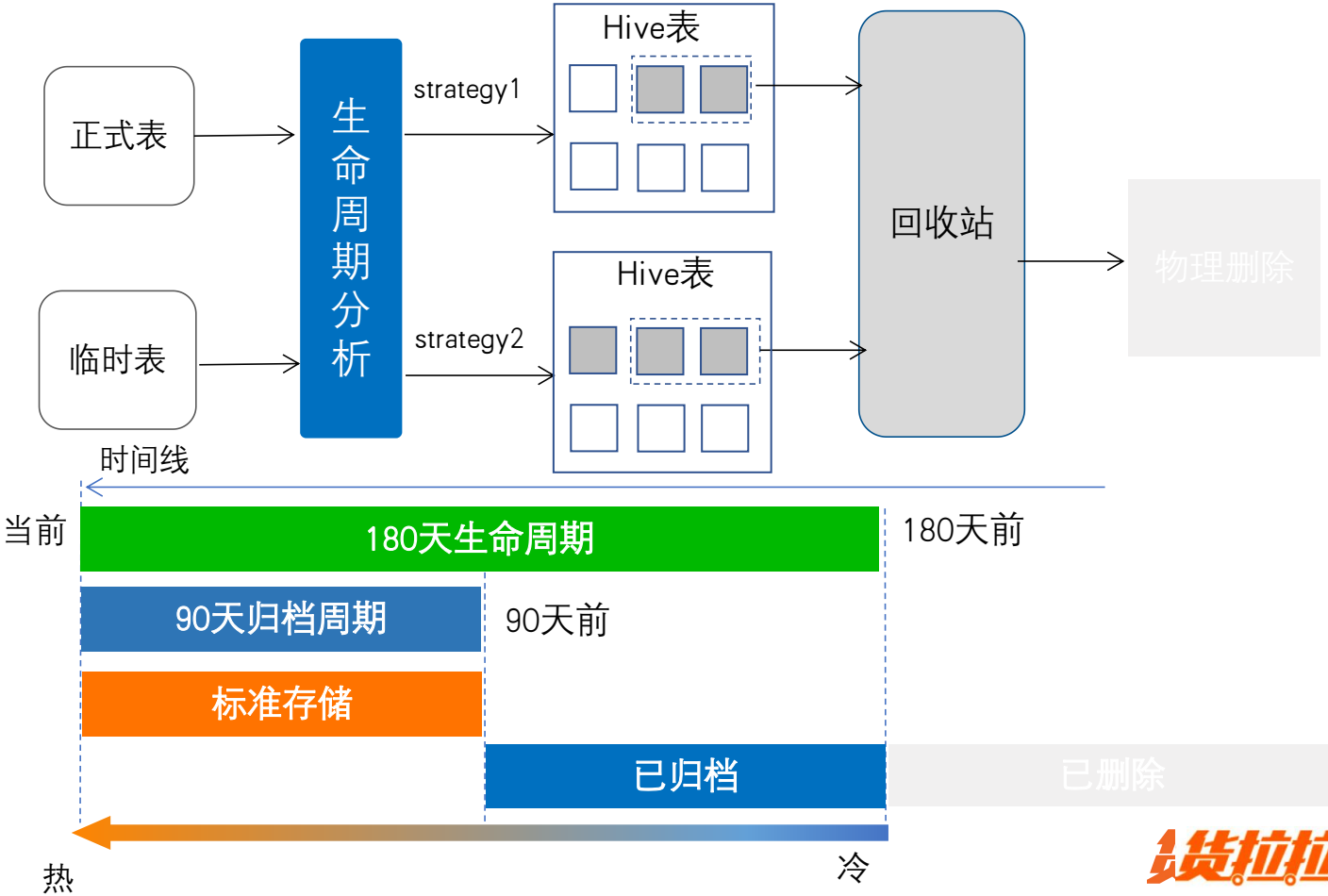


n: 分区90天内被访问次数

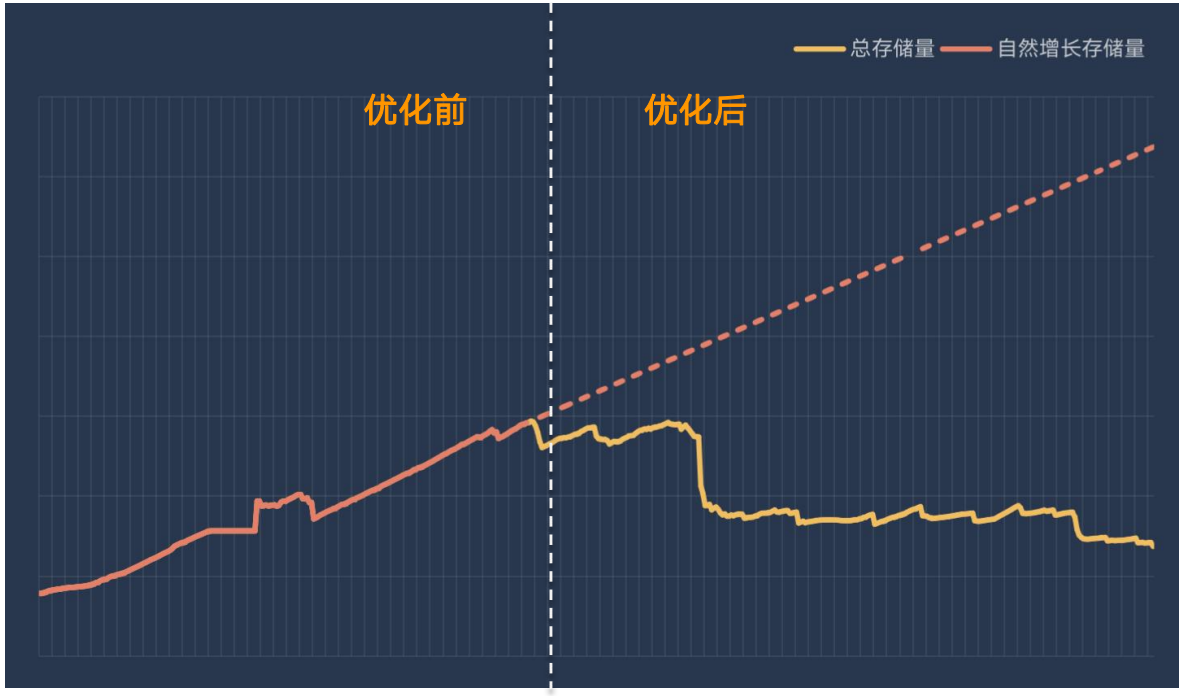
成本治理体系–冷热分层和归档



成本治理体系–数据生命周期管理



成本治理体系–存储治理收益



- 存储优化明显：
  - 优化前：存储线性快速增长
  - 优化后：存储8个月**零增长并持续下降**
- 累计节省了54%的存储成本

数据血缘–应用场景

01 数据资产

- ✓ 热度计算
- ✓ 理解数据上下文

02 数据开发

- ✓ 影响分析
- ✓ 问题数据溯源

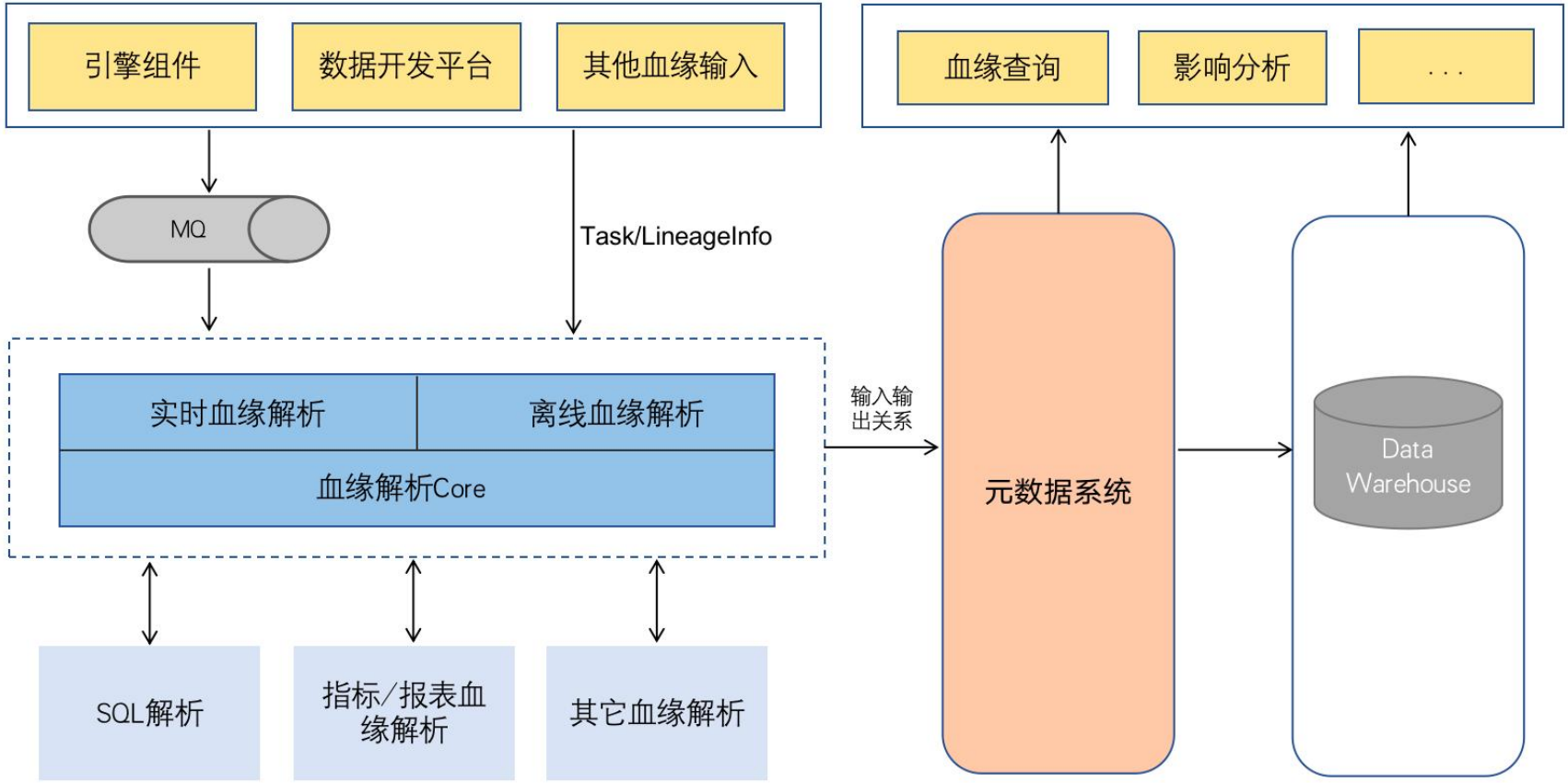
03 数据治理

- ✓ 链路状态追踪
- ✓ 数仓治理

04 数据安全

- ✓ 安全合规检查
- ✓ 字段安全级别传播

数据血缘–架构图



未来规划

更高效的找数服务

构建全域数据资产地图

找数

成本

增强成本治理能力

加强成本度量手段
提升辅助治理分析能力
推进成本运营机制

更细粒度血缘

构建广义全链路字段级血缘

血缘

规范

统一模型和数据标准

落地公司级模型规范
制定统一数据标准

非常感谢您的观看

货拉拉 | DataFun.