



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

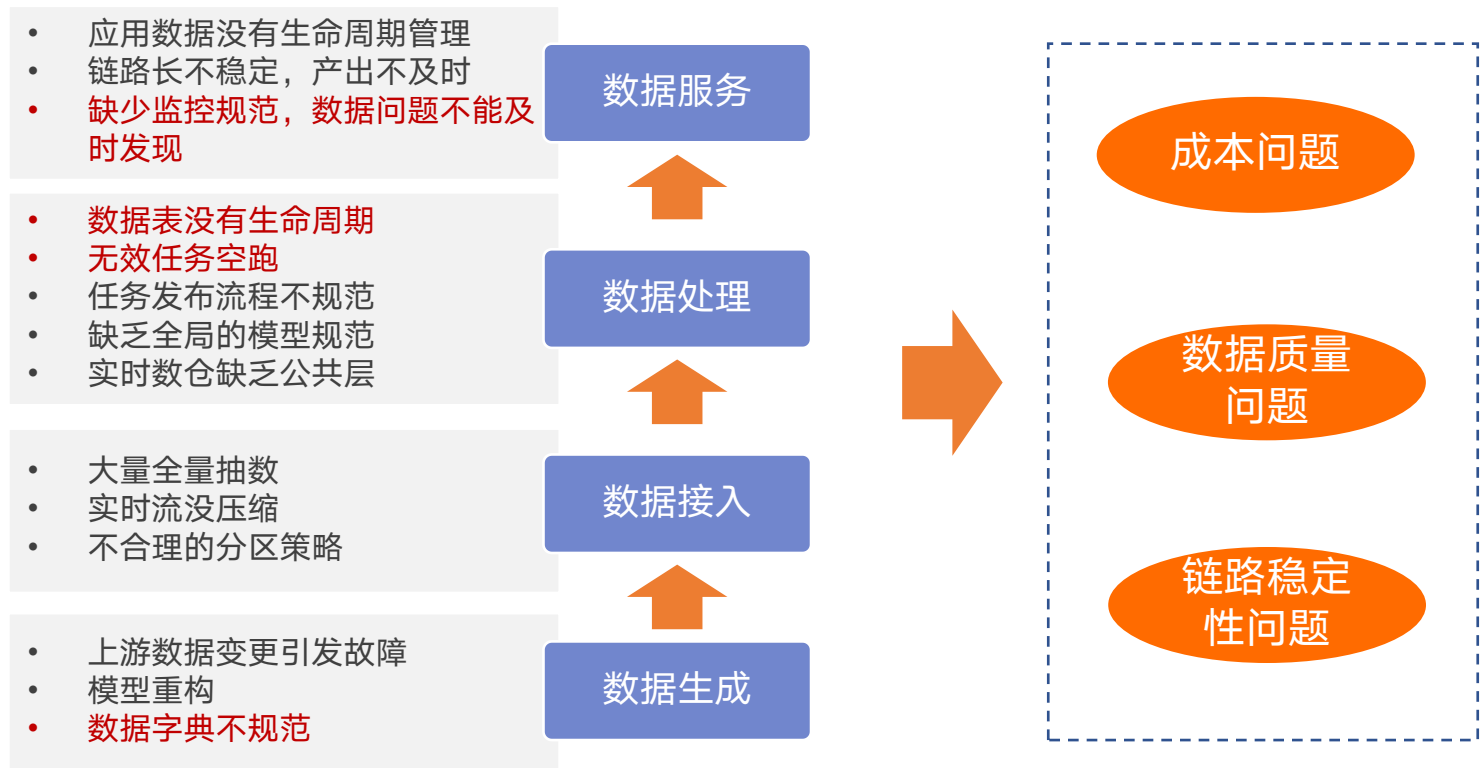
02 数据质量平台建设实践

03 元数据平台建设实践

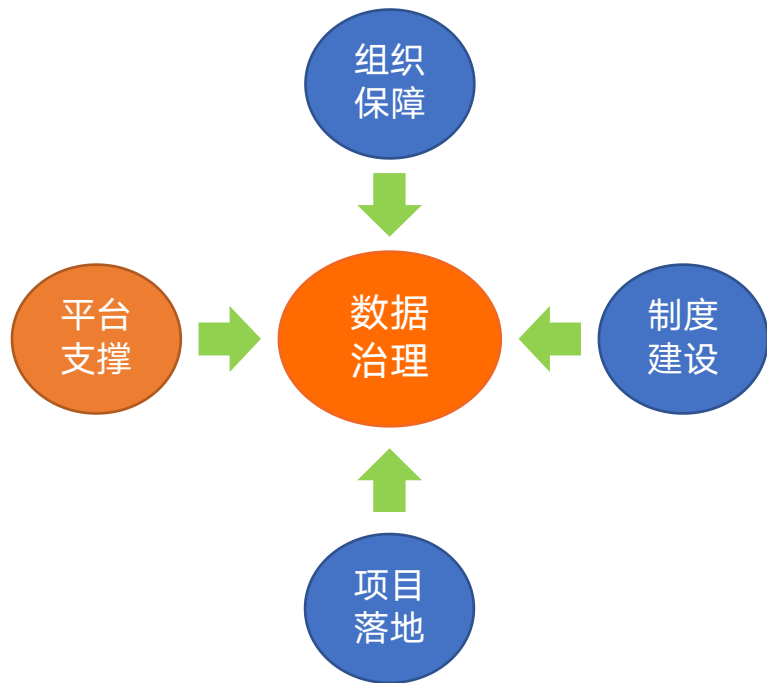
01 货拉拉数据治理体系



背景与现状



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

辅助决策类应用

赋能业务类应用

应用层

经营分析

用户分析

实时报表

鹰眼监控

智能营销

智能广告投放

.....

数据门户

服务层

数据应用支撑服务工具

大数据分析平台

数据工具箱

自助分析

可视化

固定报表

指标库管理

用户画像

数据服务工具

快捷分析

数据上报

快速报表

数据
云服务

数据智能支撑工具

AB Test

特征平台

AI平台

门户
首页

权限
中心

平台层
&数仓

数据研发平台

BQ数据查询

IDP数据集成开发

飞流实时开发

数据治理平台

数据质量管理

质量规则配置

数据质量监控

数据质量报告

质量问题处理

元数据管理

数据地图

数据血缘分析

数据源探查

数据模型管理

数据模型规范

数据库管理

数据建表管理

成本管控

成本度量与展示

辅助治理

成本运营机制

数据资产管理

数据目录管理

数据标准管理

数据资产评估

数据安全治理

数据审计

数据分级分类

数据加密

数据脱敏

库表权限管理

报表权限管理

下载权限管理

数据仓库

集市1

集市2

.....

指标库

DIM

ODS贴源数据层

DWS公共汇总服务层

DWB明细数据整合层

DWD明细数据层

知识库

个人
中心

内容
管理

建议
反馈

接入层

数据接入平台

离线数据接入

实时数据接入

埋点数据接入

数据对账

数据链路监控

基础层

大数据
基础平台

离线计算

资源管理

实时计算

大数据存储

基础元数据 (Hivemeta)

在线数据存储HBase

货拉拉

DataFun.

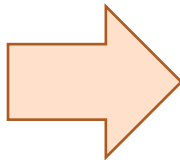
02 数据质量平台建设实践



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



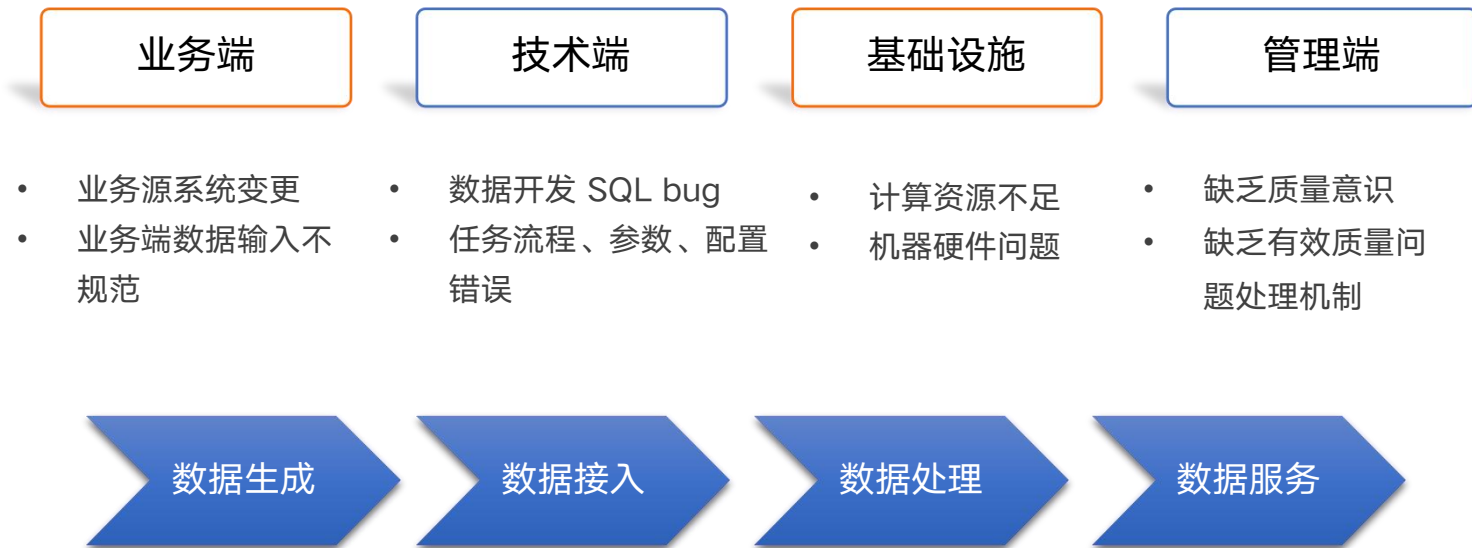
报表延迟

链路异常

数据丢失

决策错误

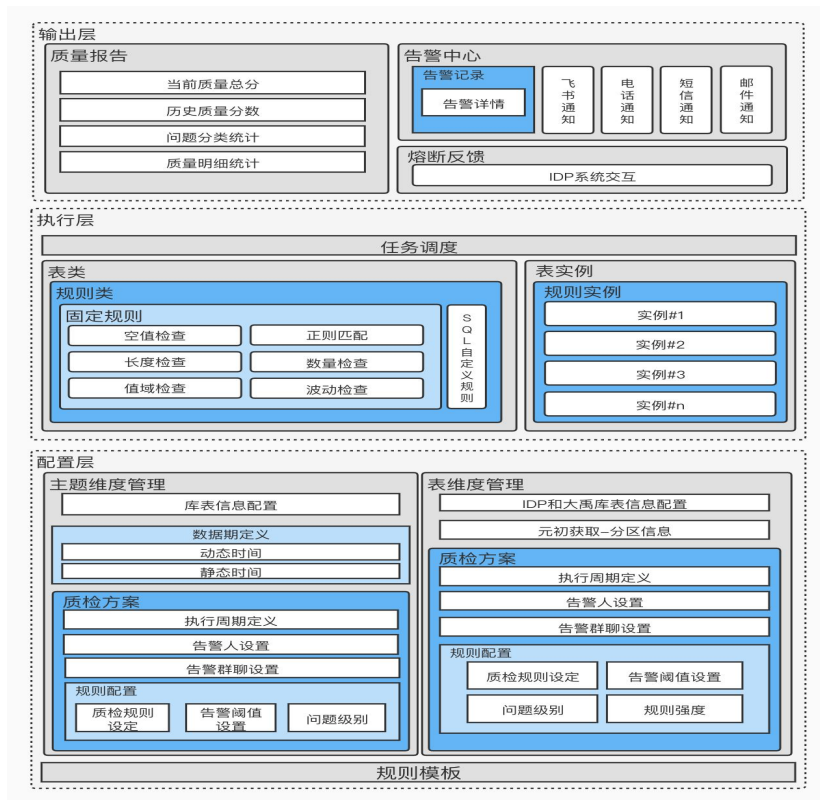
产生数据质量问题的原因



数据质量保障思路



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·大禹

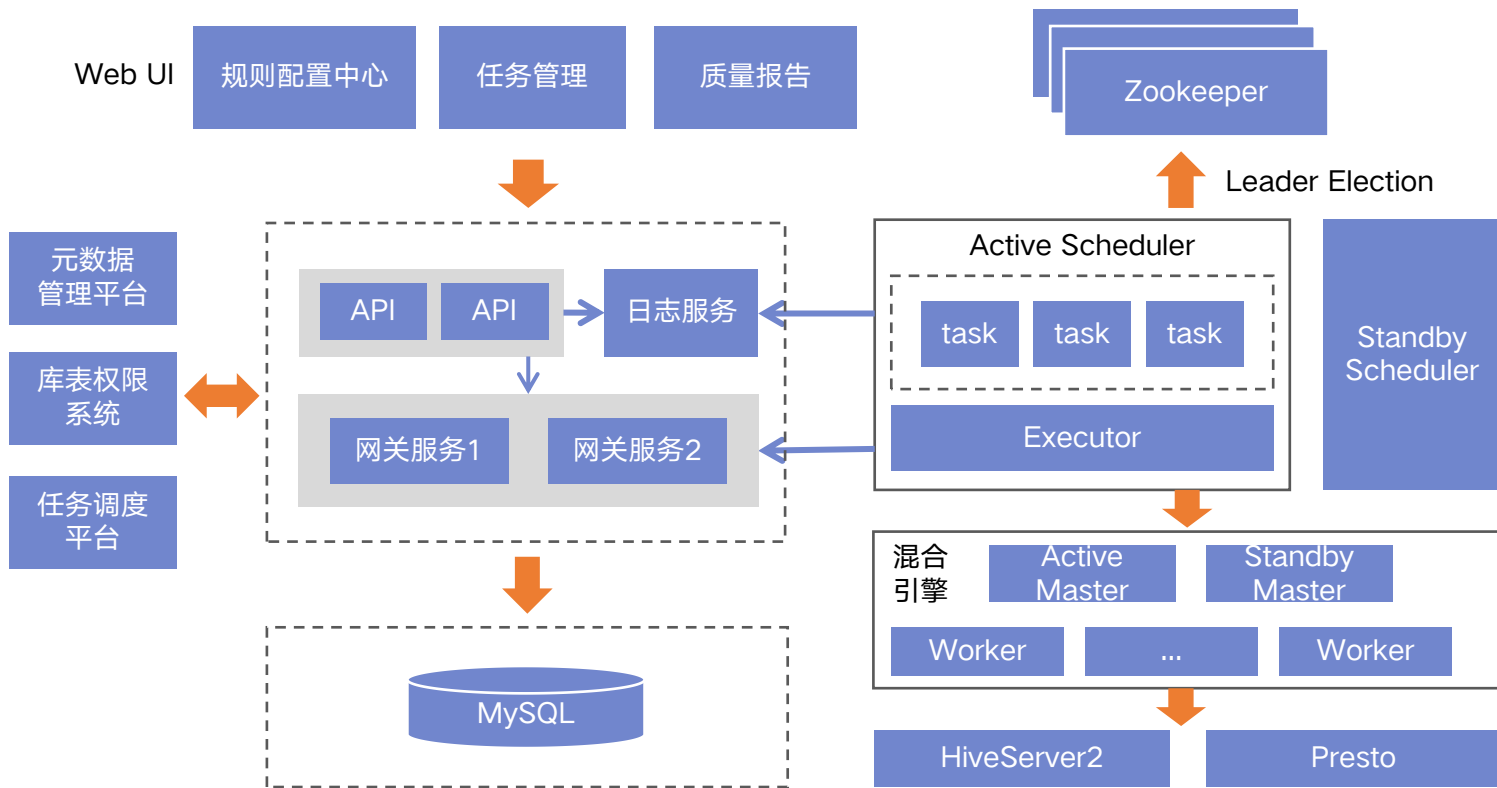
报表中心 | 数据应用 | 数据平台

搜索: 表名 输入表名 负责人 请输入负责人 查询

创建质检任务 设置阈值

序号	库名	表名	负责人	创建时间	描述	状态	操作
1	torrent ods	hpay_payment_guarantee_payment_order	beshell.liang.kiko.la	2022-04-26 11:01:57		启用	表实例 规则管理 编辑 立即执行 停用
2	hl_dwd	dwd_order_coupon_order	daniel.lu	2021-07-14 15:50:42	主题维度迁移到表维度	启用	表实例 规则管理 编辑 立即执行 停用
3	hl_dws	dws_user_fact	chao.huang.guoliang.qi	2021-07-14 15:50:42	主题维度迁移到表维度	启用	表实例 规则管理 编辑 立即执行 停用
4	hl_ods	dp_order_base_bill_append_bill	beshell.liang.kiko.la	2022-04-22 15:54:36		启用	表实例 规则管理 编辑 立即执行 停用
5	hl_ods	hpay_account_balance_log_bill	kiko.lu.zhao.ji	2022-04-22 15:45:44		启用	表实例 规则管理 编辑 立即执行 停用
6	hl_ods	time_finance_payment_return_bill	conor.ouyang.zhao.ji	2022-04-22 15:48:42		启用	表实例 规则管理 编辑 立即执行 停用
7	hl_ods	dp_gps_vehicle_gear_physics_bill	beshell.liang.kiko.la	2022-04-22 15:42:48		启用	表实例 规则管理 编辑 立即执行 停用
8	hl_ods	dp_bill_bill_append_bill	daniel.lu.kiko.lu.zhao.ji	2022-04-20 17:56:26		启用	表实例 规则管理 编辑 立即执行 停用

货拉拉数据质量平台-系统架构



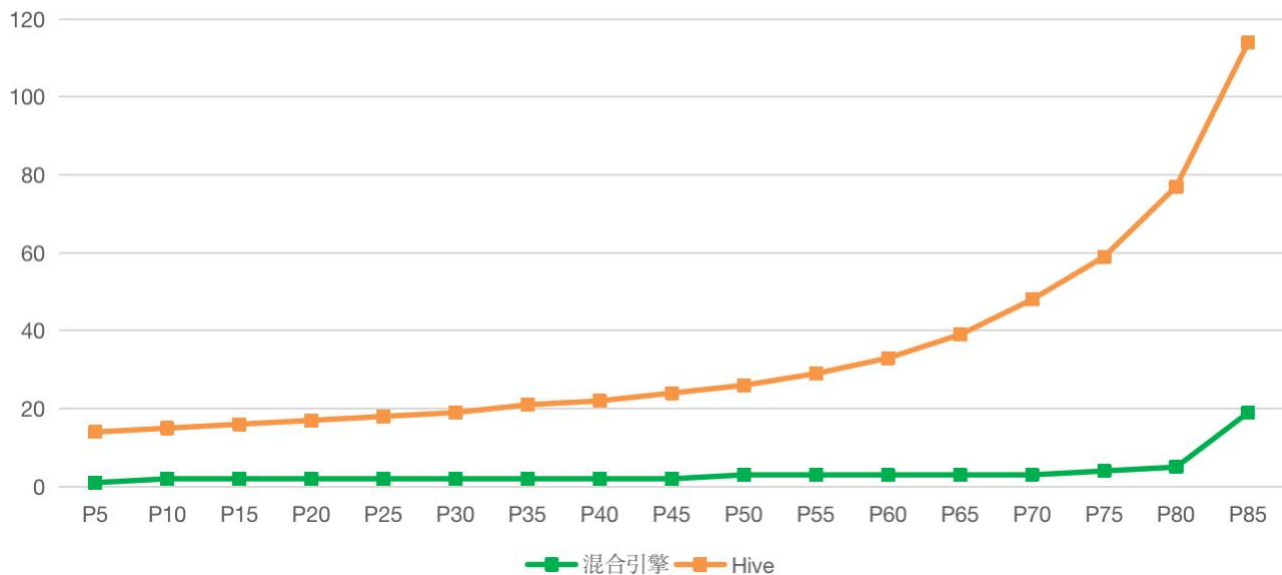
- 高稳定
- 高效率
- 熔断阻塞

数据质量分析效率

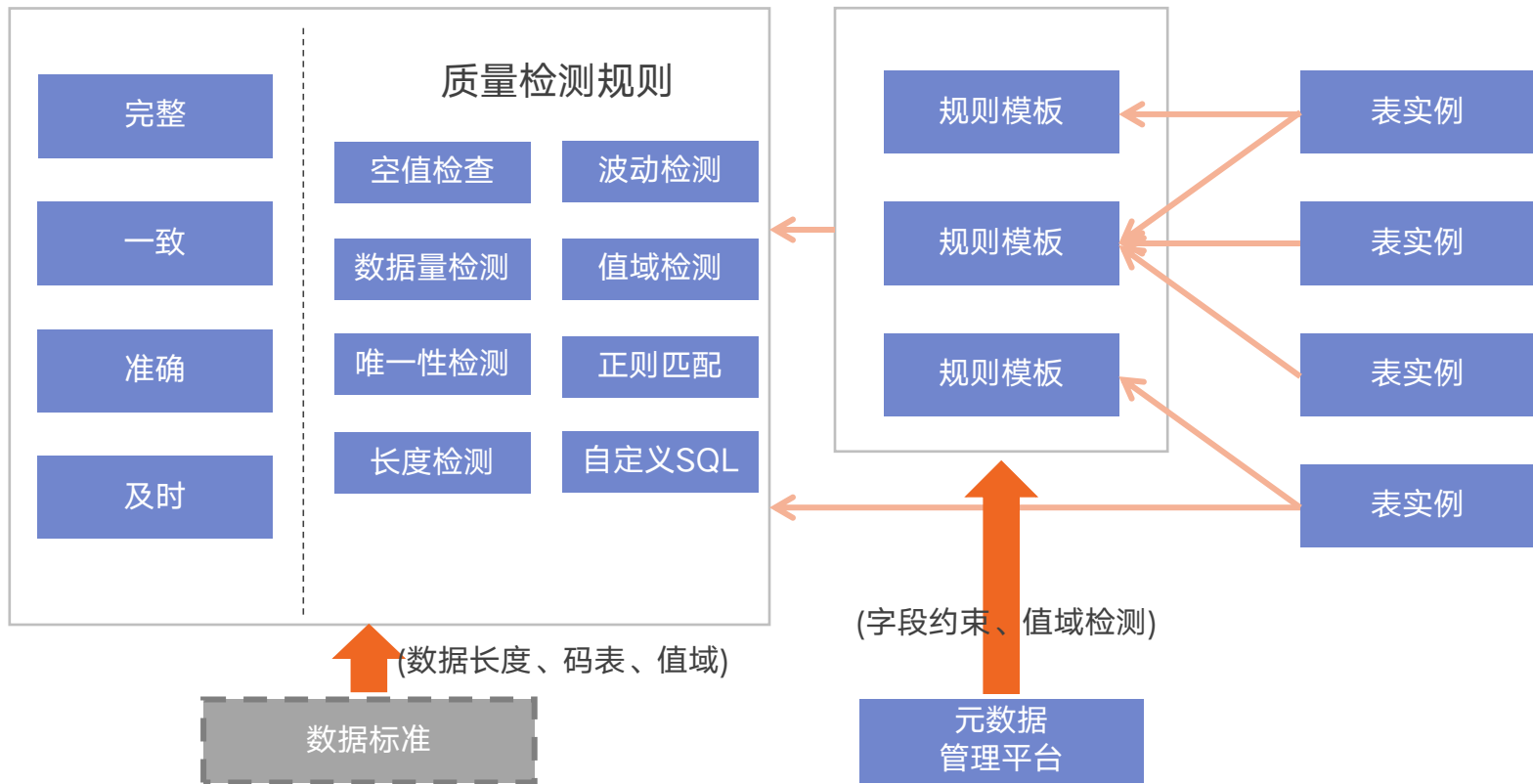
分位	P5	P10	P15	P20	P25	P30	P35	P40	P45	P50	P55	P60	P65	P70	P75	P80	P85	P90	P95	P98
混合引擎	1	2	2	2	2	2	2	2	2	3	3	3	3	3	4	5	19	43	82	192
Hive	14	15	16	17	18	19	21	22	24	26	29	33	39	48	59	77	114	196	396	915
耗时下降	93%	87%	88%	88%	89%	89%	90%	91%	92%	88%	90%	91%	92%	94%	93%	94%	83%	78%	79%	79%

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

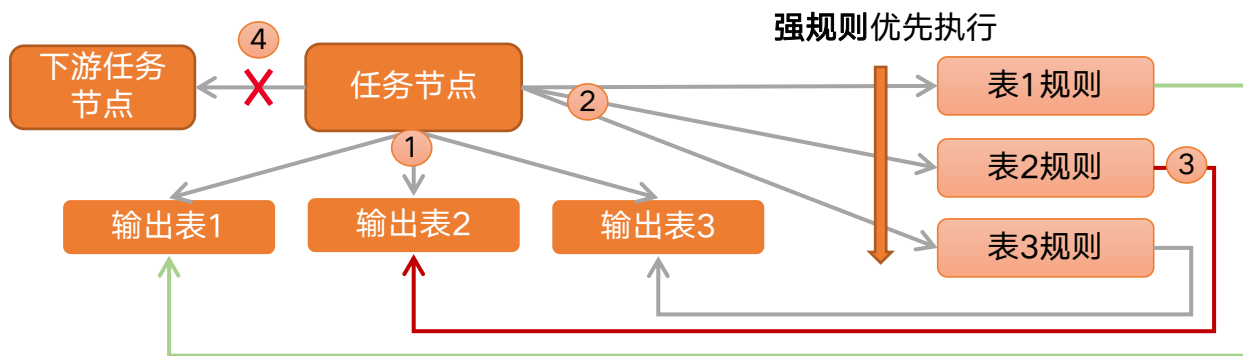
任务详情--64211 不是任务负责人无法编辑

依赖关系: 全周期依赖

本任务输出表:

请输入库名.表名(例: ods.tb_task)搜索 添加目标表

输出表名	表注释	负责人	添加方式	大禹检测	目标类型	操作
hl_dwd.dwd_order_driver_eda_1d_tm	司机抢接单eda数据	daniel.lu,akin.jiang	代码解析	<input type="checkbox"/>	表	删除



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

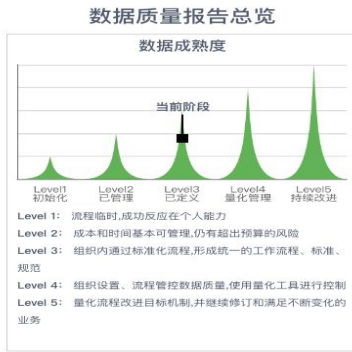
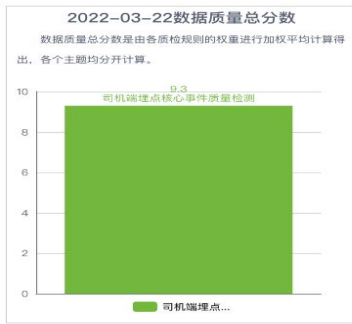
货运司机核心埋点质检报告

HLL DW Data Quality Reports

第2022-03-22期

机密文件禁止泄漏

数据质量报告 | 2022-03-22
HUOLALA Data Quality Report



统计数据

质检规则	数值
质检规则总数	342
通过规则数	158
失败规则数	184

质检属性项	数值
司机端埋点核心事...	12,496,342,930



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

货拉拉·大禹									
HLL 报表中心 数智中心 数据应用 数据平台 权限申请									
search	告警管理 > 告警中心								
表维度管理	表维度告警列表 主题维度告警列表								
主题维度管理	规则名称 表名称 告警时间 2022-04-24 22:52:35 - 2022-04-26 22:52:3 查询								
告警管理									
告警中心									
规则模板管理									
规则类型	表名	告警生成时间	告警级别	告警状态	告警通知	通知执行结果	告警处理状态	操作	
SQL自定义	hll_dm.dm_monitor_order_fact_ld_in	2022-04-26 04:33:15	重要	飞书: 成功 邮件: 成功	通知联系人: daniel.lu 通知群聊: 暂未设置	成功	待处理	告警详情 告警备注	
SQL自定义	hll_dm.dm_monitor_order_fact_ld_in	2022-04-26 04:33:15	重要	飞书: 成功 邮件: 失败	通知联系人: daniel.lu 通知群聊: 暂未设置	成功	待处理	告警详情 告警备注	
SQL自定义	hll_dm.dm_monitor_order_fact_ld_in	2022-04-26 04:33:15	重要	飞书: 成功 邮件: 失败	通知联系人: daniel.lu 通知群聊: 暂未设置	成功	待处理	告警详情 告警备注	
波动检查	risk_dwd.dwd_iot_device_car_td_tm	2022-04-26 04:07:50	重要	飞书: 成功 邮件: 成功	通知联系人: qianfan.yang 通知群聊: 暂未设置	成功	待处理	告警详情 告警备注	
波动检查	risk_dwd.dwd_iot_device_car_td_tm	2022-04-26 03:12:43	重要	飞书: 成功 邮件: 成功	通知联系人: qianfan.yang 通知群聊: 暂未设置	成功	待处理	告警详情 告警备注	
SQL自定义	torrent_ods.lts_hbase_match_rate	2022-04-26 02:50:28	重要	飞书: 成功 邮件: 成功	通知联系人: guoxiang.qi 通知群聊: 暂未设置	成功	待处理	告警详情 告警备注	
波动检查	hll_dwb.dwb_user_questionnaire_nps_touch_1d_tm	2022-04-26 02:34:03	重要	飞书: 成功	通知联系人: guoxiang.qi chaos.huang	成功	待处理	告警详情 告警备注	

数据质量平台

1500+ 张
接入的表数量

100%
核心链路表覆盖

现状

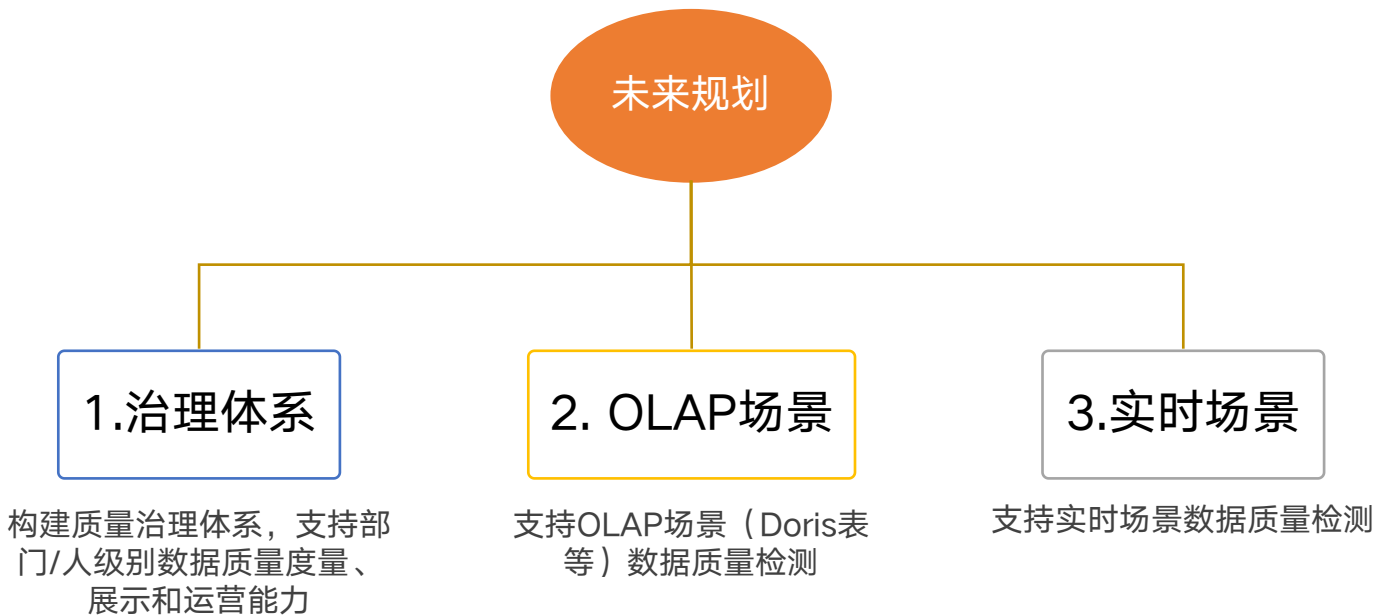
现状

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



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

03 数据血缘

02 成本治理体系

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

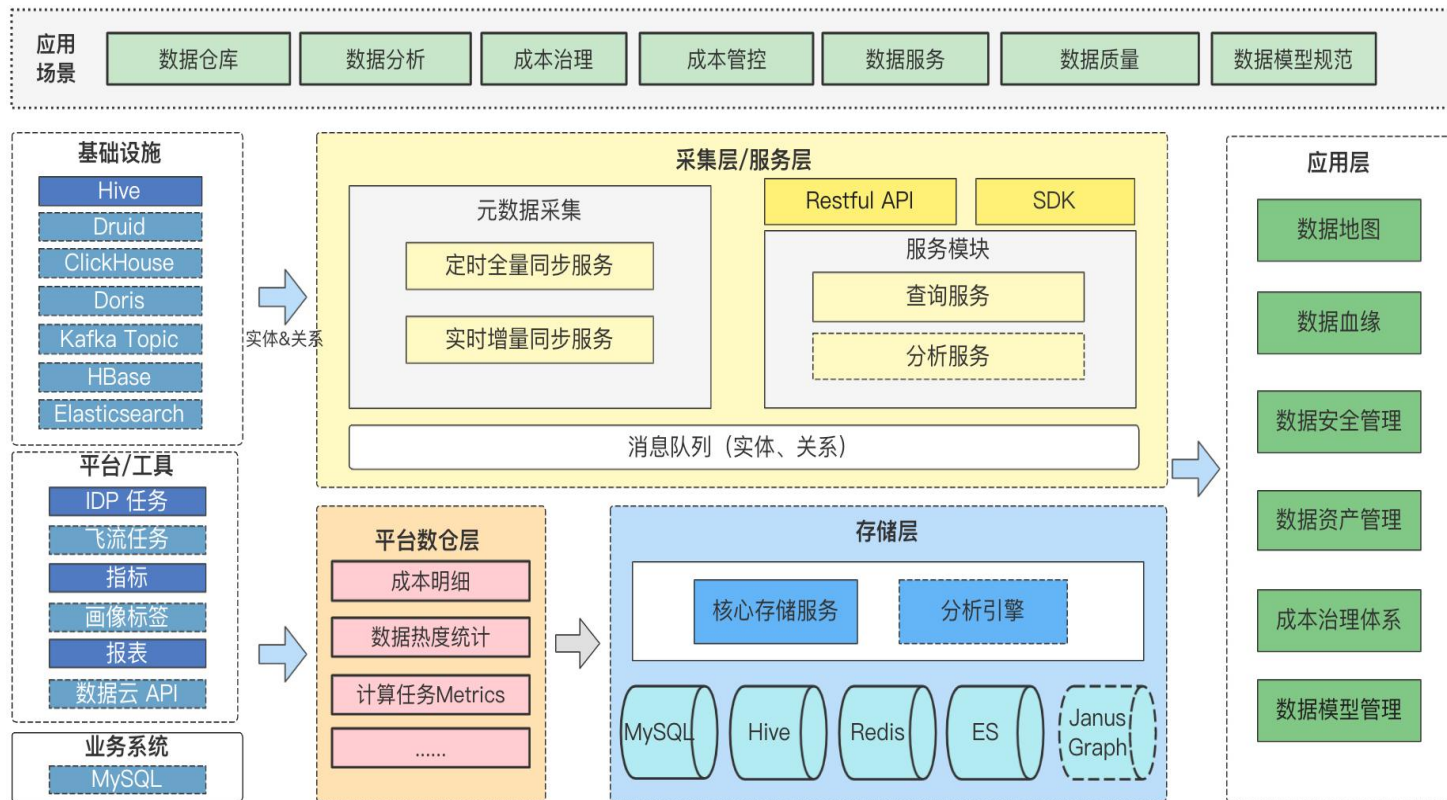
- ✓ 数据生命周期治理
- ✓ 离线存储治理
- ✓ 成本度量

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	<ul style="list-style-type: none">• 统一元模型• 10+元数据类型• 10W+任务、几十万Hive表	<ul style="list-style-type: none">• 实时全链路血缘	<ul style="list-style-type: none">• 自定义查询• 分类检索• 案例分享	<ul style="list-style-type: none">• 生命周期• 安全等级	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 全链路血缘• 字段级血缘	<ul style="list-style-type: none">• 类SQL查• 关联查询• 分类查询• 热度推荐	<ul style="list-style-type: none">• 生命周期• 安全等级	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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&D ataWor ks	10+表格式 支持API类型 支持数据湖格式	<ul style="list-style-type: none">• 字段级血缘	DataWorks(数据地图)	元数据分析和管理的、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	<ul style="list-style-type: none">• 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

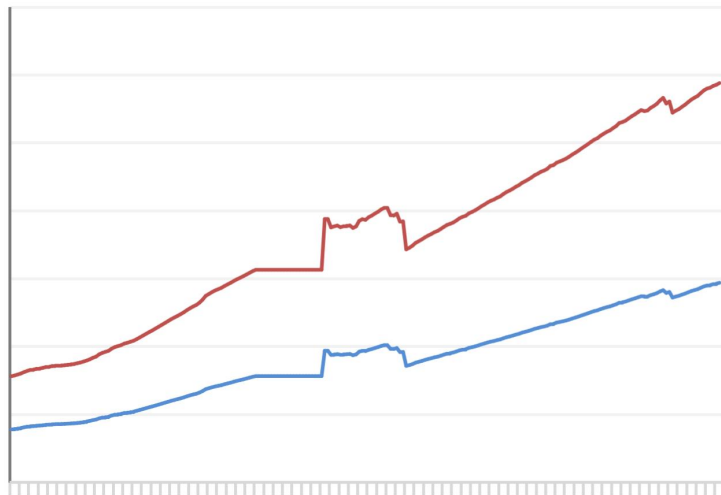


表数量大
17W+ Hive表

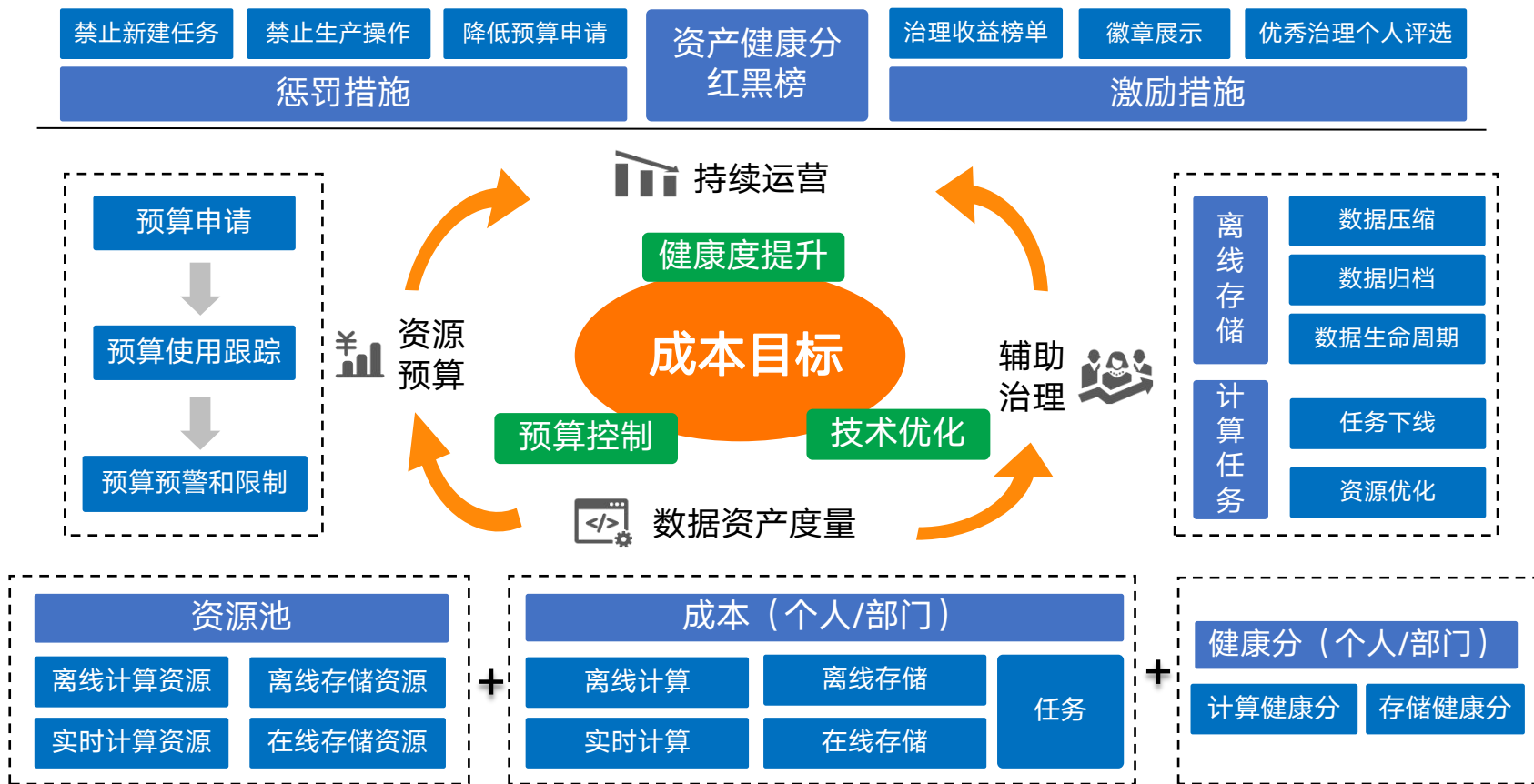
增速快
4PB/月自然增长

冷数据量大
约33%数据90天内无访问

— 总存储量 — 冷数据量



成本治理体系



成本治理体系-成本度量和展示



服务层

成本明细

资源使用明细

成本账单

存储健康分

任务健康分

...

平台
数仓层

租户成本汇总表

租户资源汇总表

在线存储明细

离线存储明细

分区热度统计表

报表信息表

计算引擎
Metrics表

YARN App信息
表

离线任务信息表

实时任务信息表

文件热度统计表

...

数据源层

基础
设施

存储

计算

监控

计算引擎

运维

...

数据
资产

离线任务

报表

标签

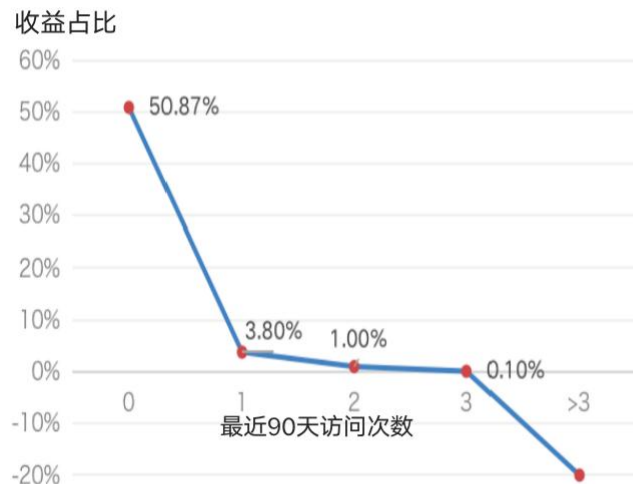
实时任务

指标

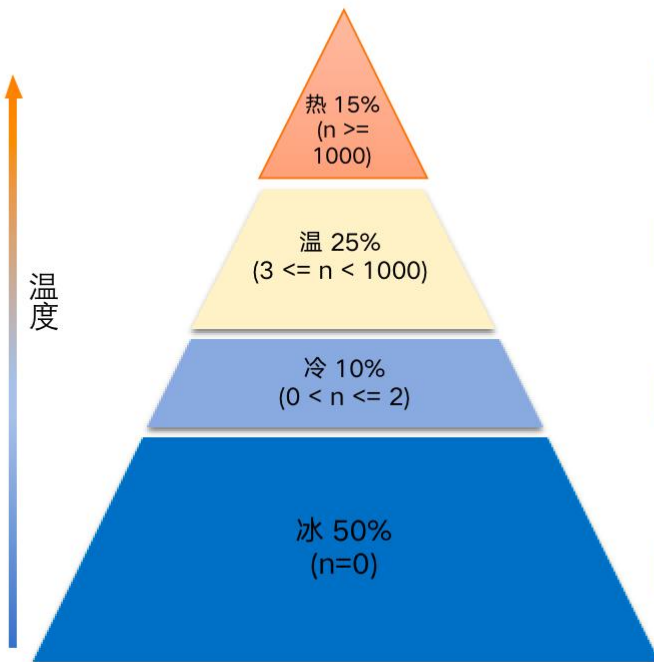
...

成本治理体系-冷热分层和归档

归档收益曲线



冷热分层



n: 分区90天内被访问次数

存储策略

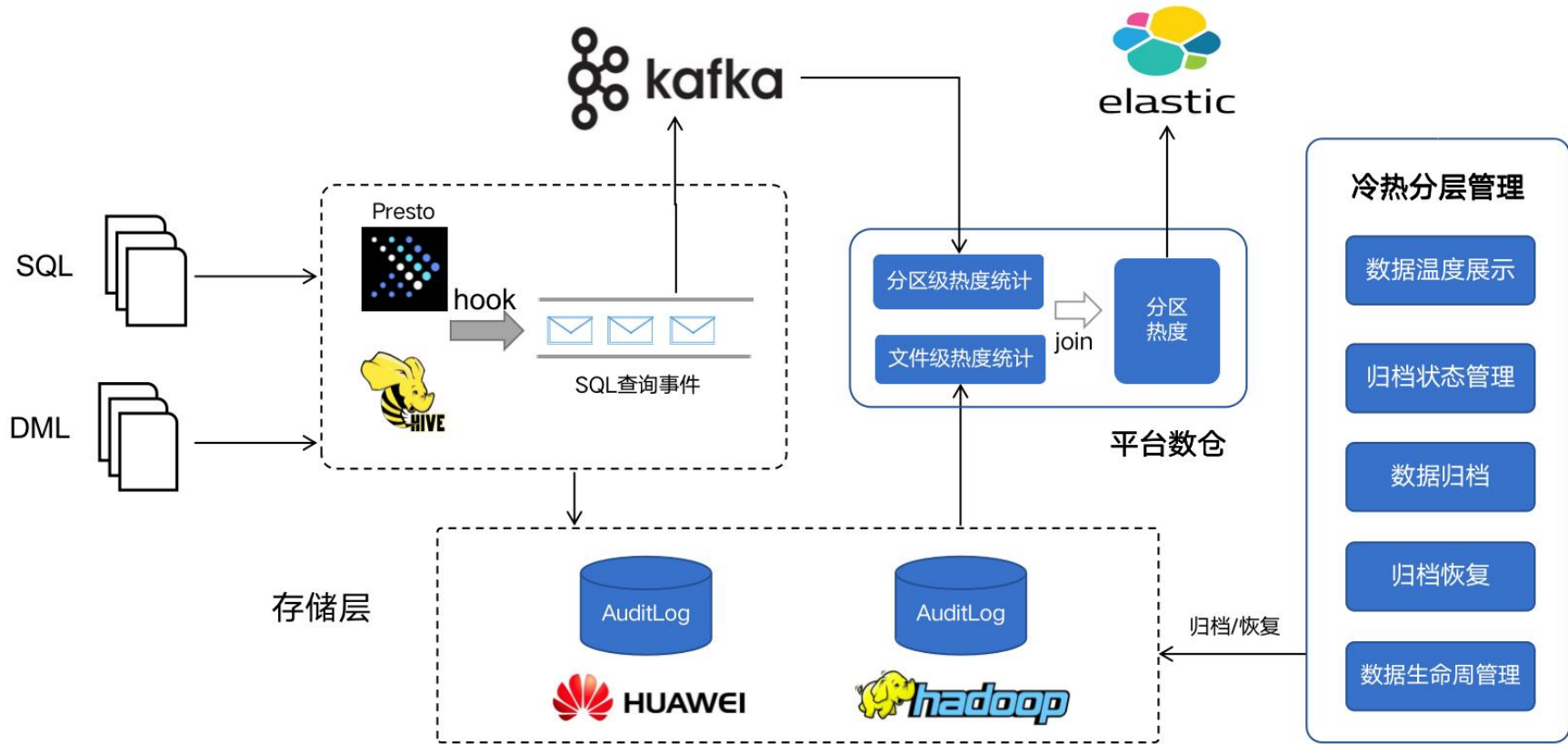
Cache

标准存储

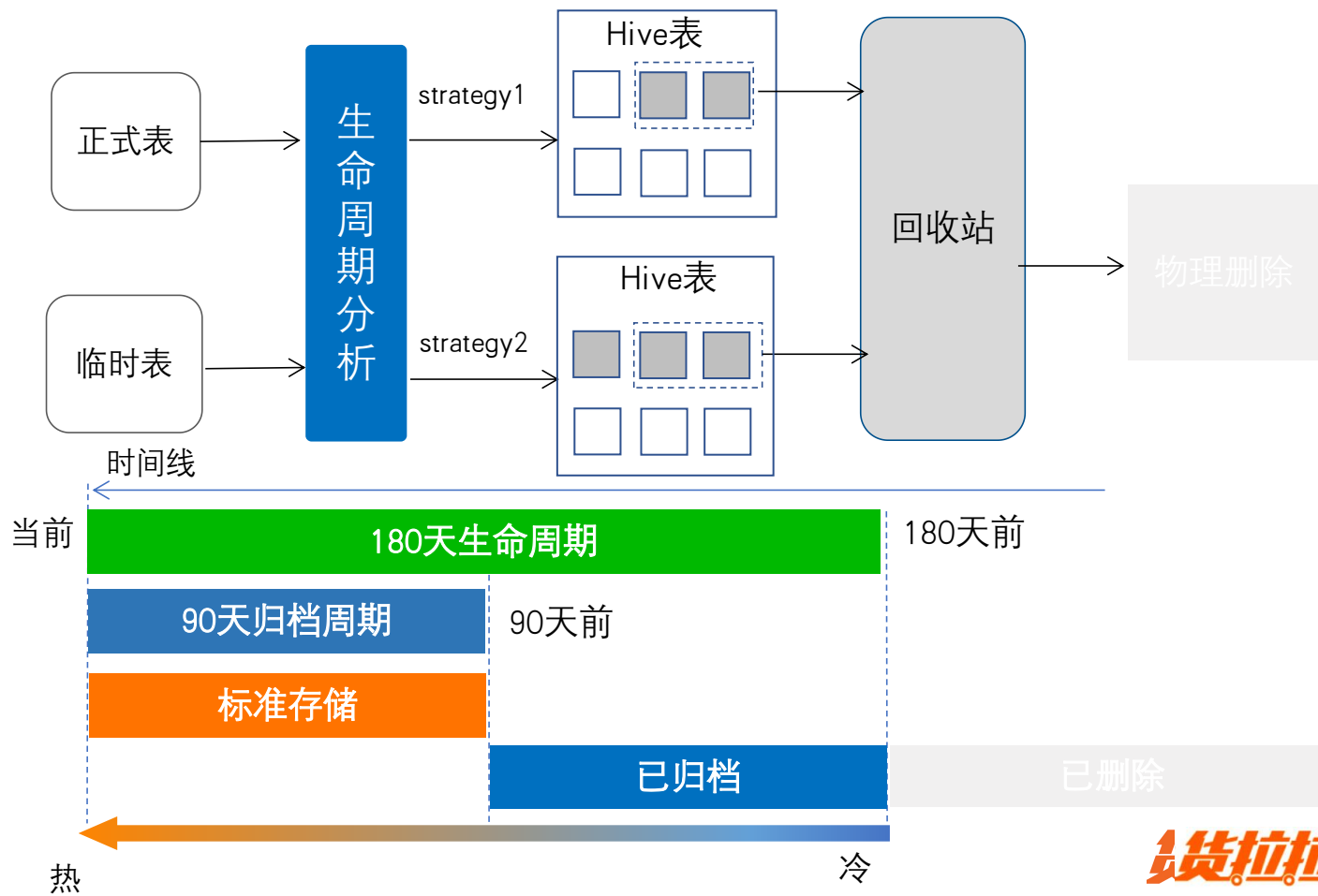
归档存储

归档或删除

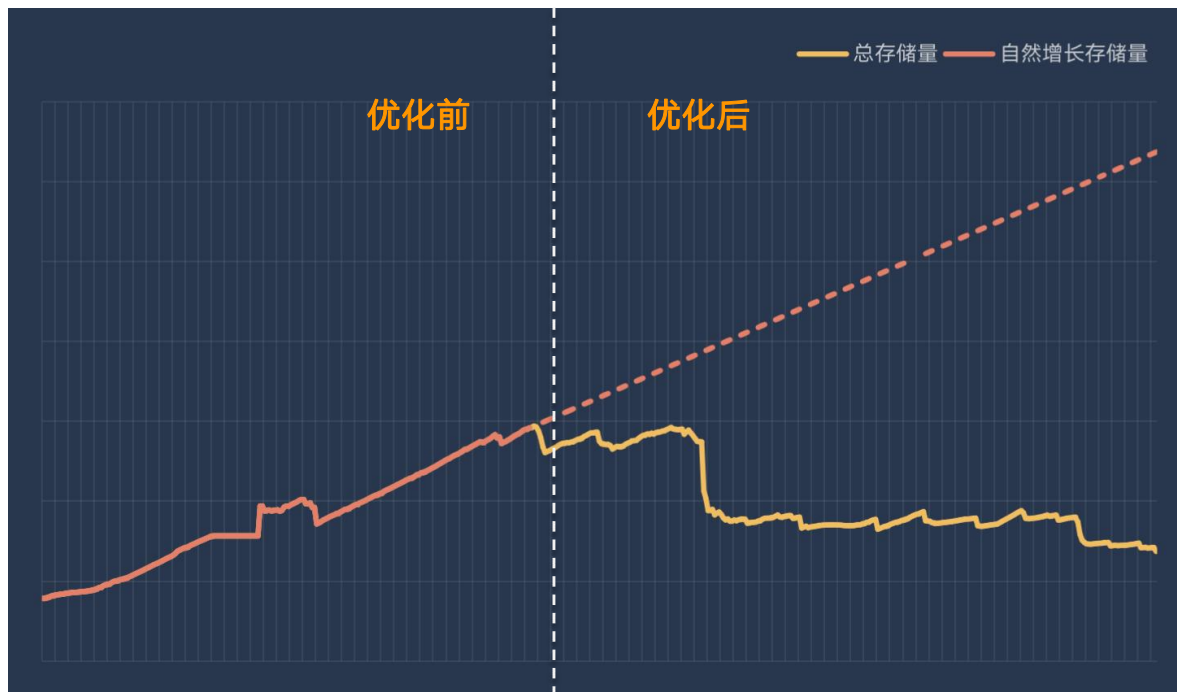
成本治理体系-冷热分层和归档



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

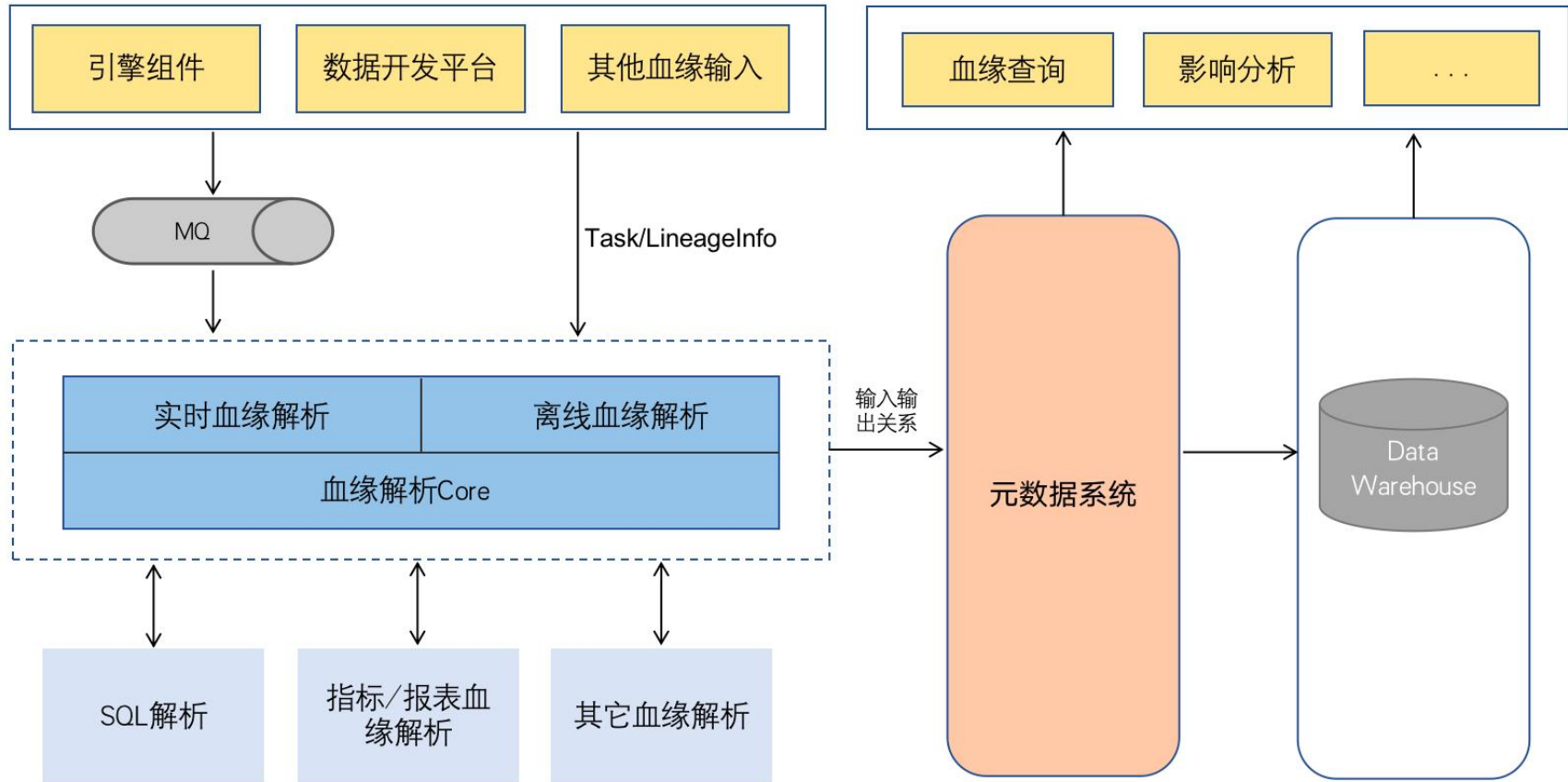
- ✓ 链路状态追踪
- ✓ 数仓治理

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.

