MEIZU | IT大咖说 知识共享平台

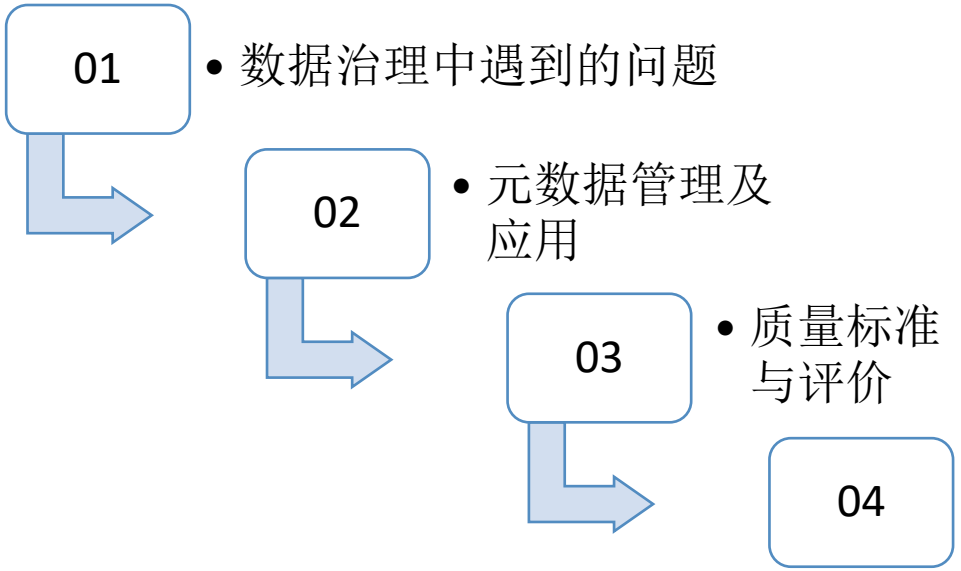
魅族数据治理探索之路

魏战松

MEIZU

魅族技术学院出品

大纲



大纲

01

- 数据治理中遇到的问题

问题背景

业务线	80+	
日新增数据量	＞70TB	每年4~6x增长
技术栈	Hadoop + Spark + Hbase+Hive+TEZ……	
计算平台	调度平台、自助分析平台、机器学习平台、权限……	

面对的问题

这个表有没有其它地方用到?

ETL人员

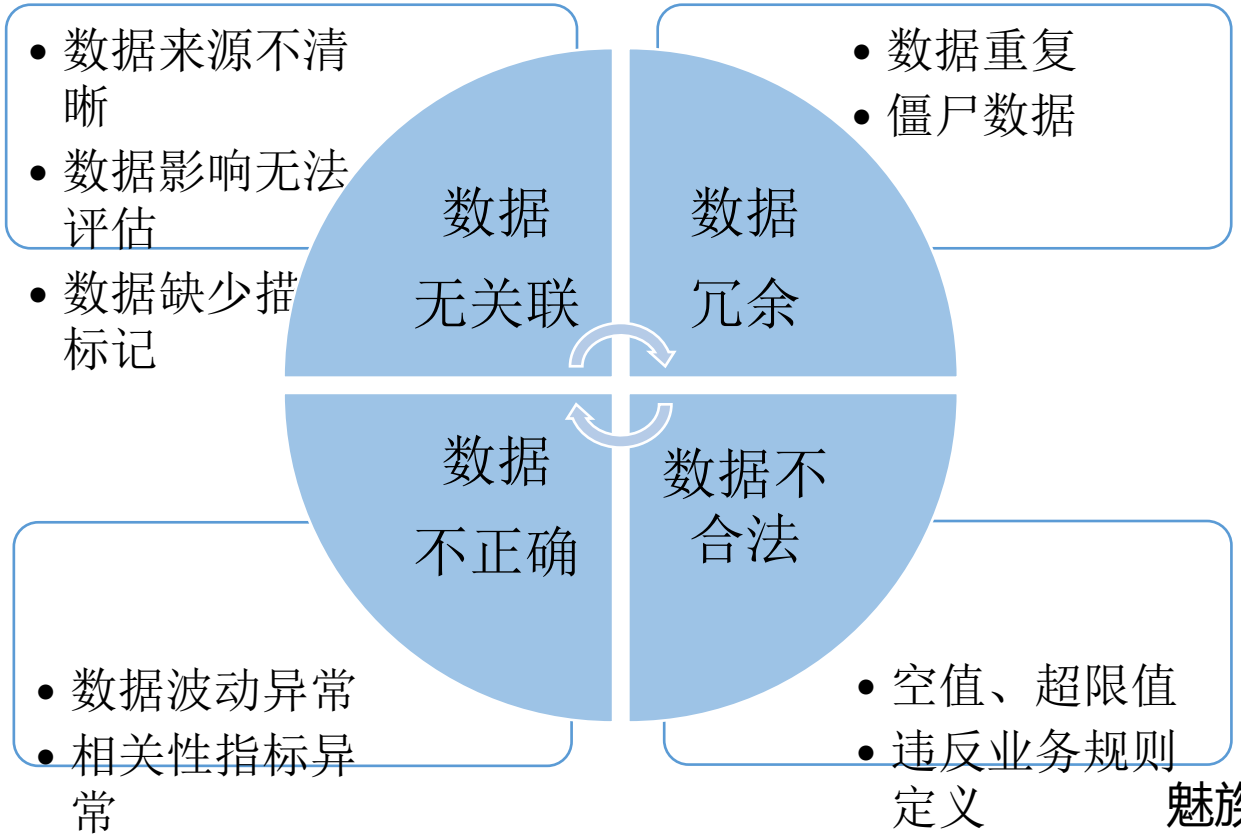
今天的xx指标怎么不正常?数据怀疑?

业务分析人员

存储不够了,哪些数据可以清理?

运维小哥

要治理什么样的数据


数据来源不清晰
数据影响无法评估
数据缺少描标记
数据无关联
数据重复
僵尸数据
数据冗余
数据不正确
数据波动异常
相关性指标异常
数据不合法
空值、超限值
违反业务规则定义

数据治理方案

1. 数据无关联
2. 数据冗余

元数据管理

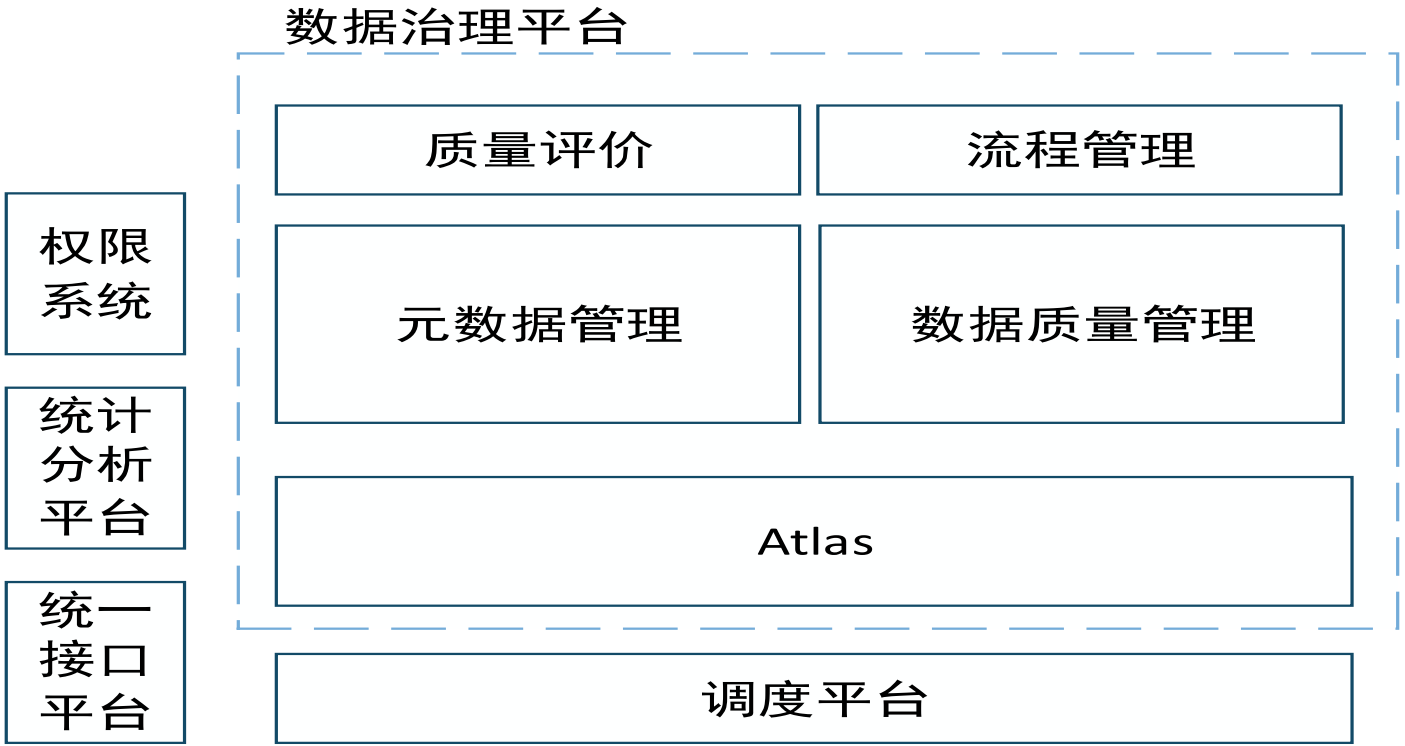
- 指标定义
- 模型管理
- 血缘追踪
- 影响分析

3. 数据不合法
4. 数据不正确

- 数据标准
- 规则校验
- 数据监控
- 质量分析

数据质量管理

数据治理架构图


数据治理平台
质量评价
流程管理
元数据管理
数据质量管理
Atlas
权限系统
统计分析平台
统一接口平台
调度平台

大纲

02

- 元数据管理及应用

什么是元数据?

元数据种类

- 技术元数据

基础平台	计算平台	数据产品	线下数据
•HDFS •Hive •Hbase •Spark	•离线任务(CETUS) •流平台 •质量管理平台 •权限系统(SCT) •可视化编码 •机器学习平台	•统计分析平台 •埋点系统 •用户画像&洞察平台 •统一接口平台	•模型元数据 •指标定义 •...

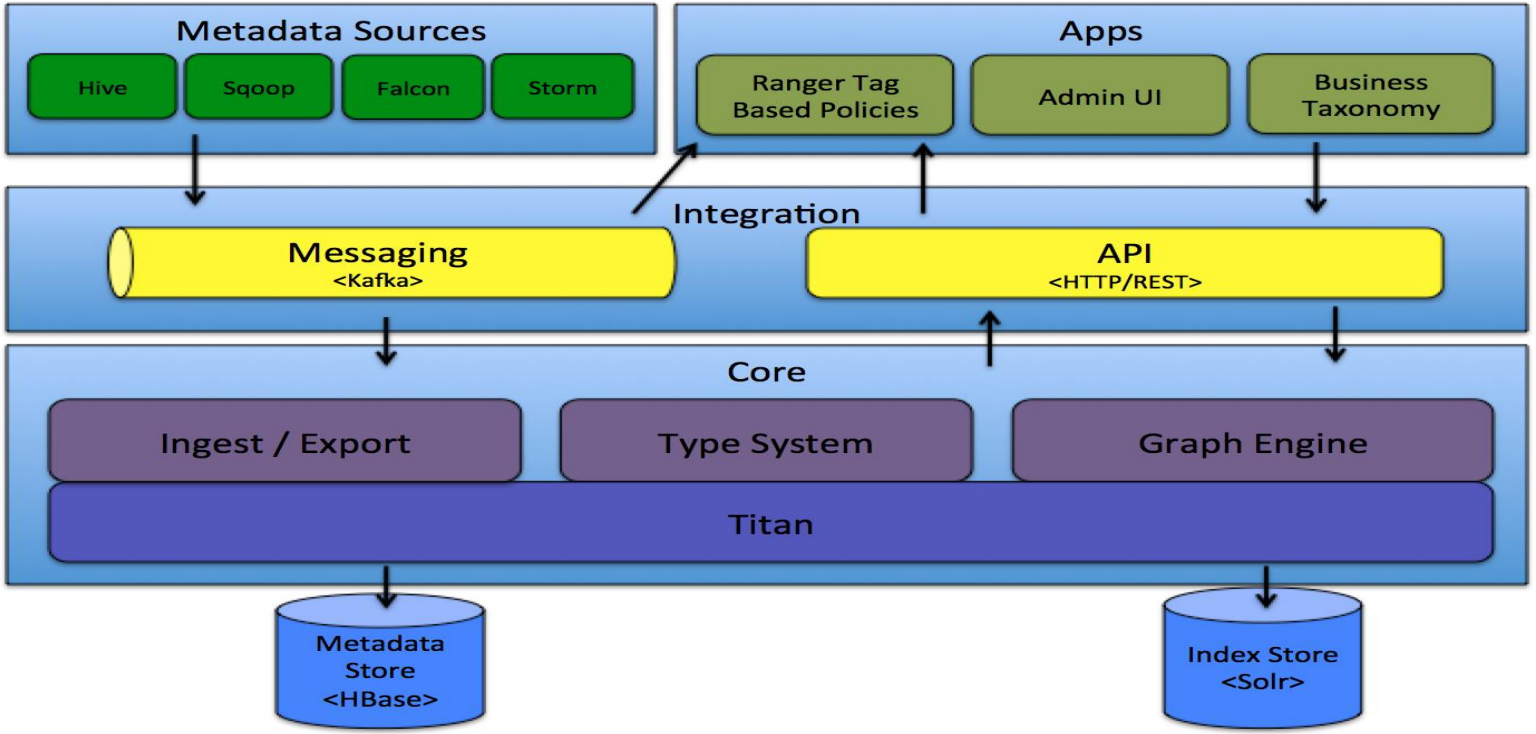
如何管理元数据?

自研?

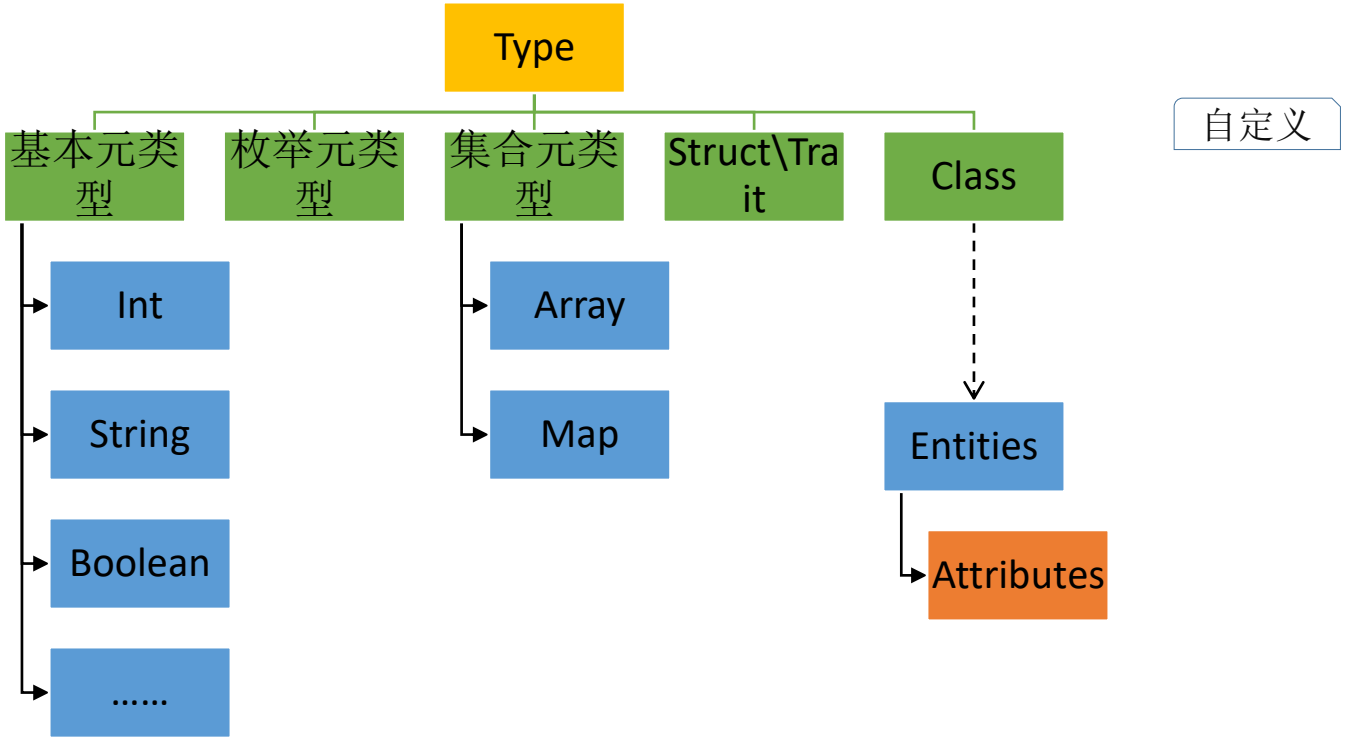
OR

寻求数据治理的开源解决方案?

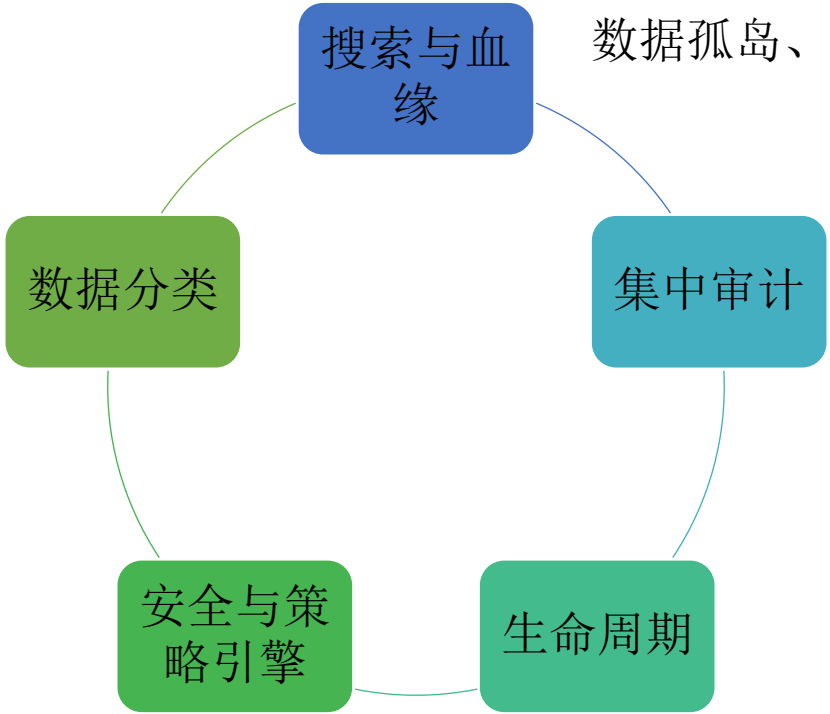
Apache Atlas



Apache Atlas 类型



魅族元数据管理需求



数据孤岛、血缘追踪、数据地图

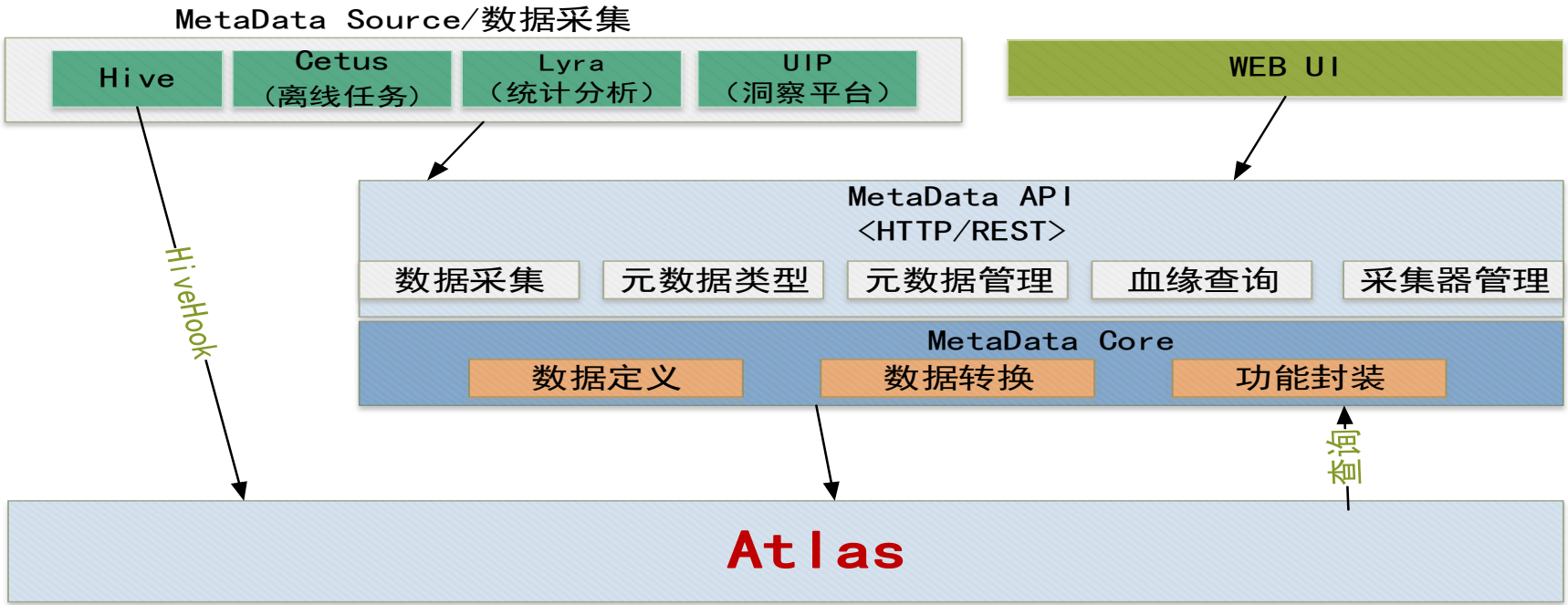
监测控制

僵尸数据

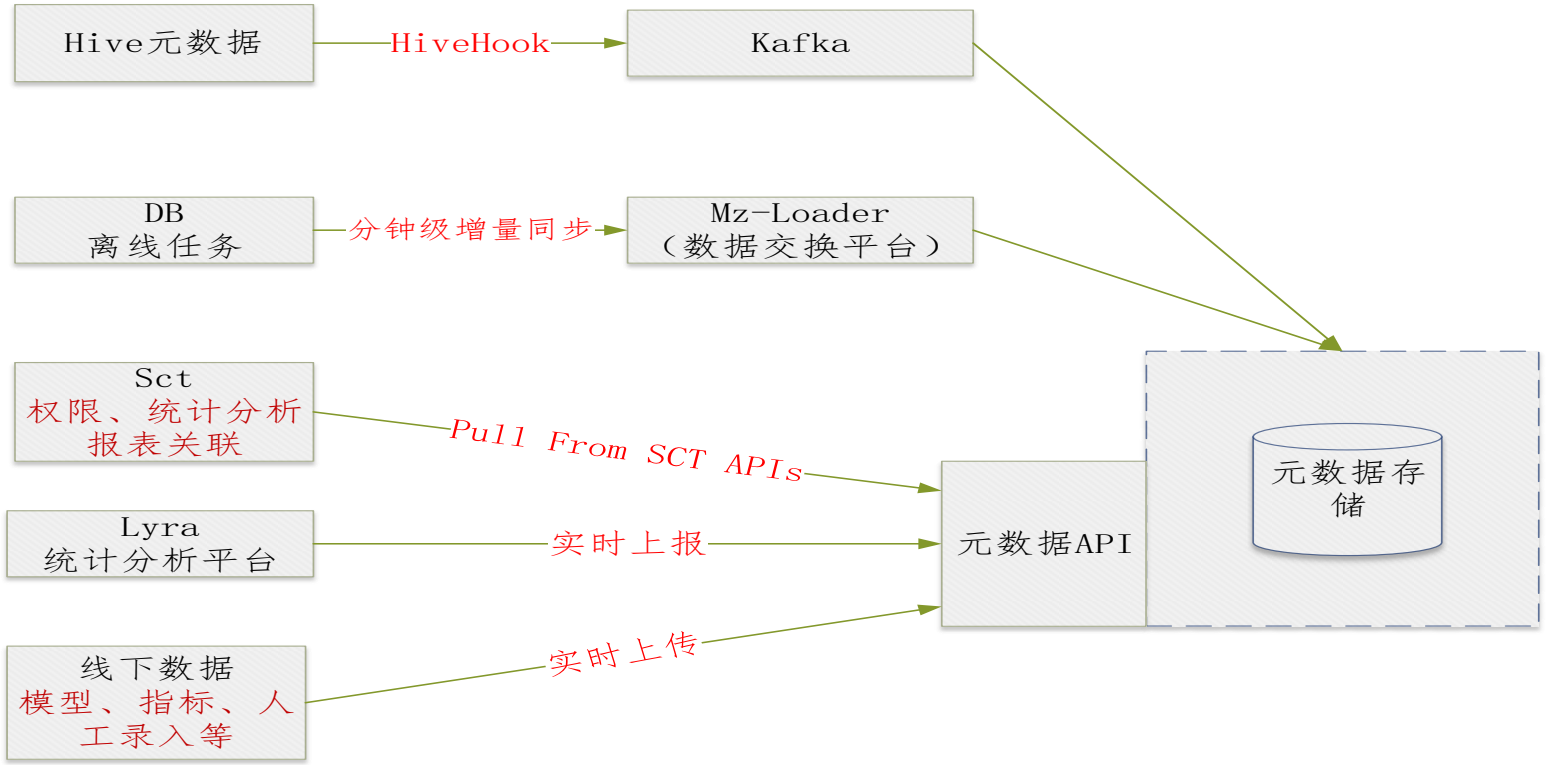
敏感授权、集成Ranger

查询、标签

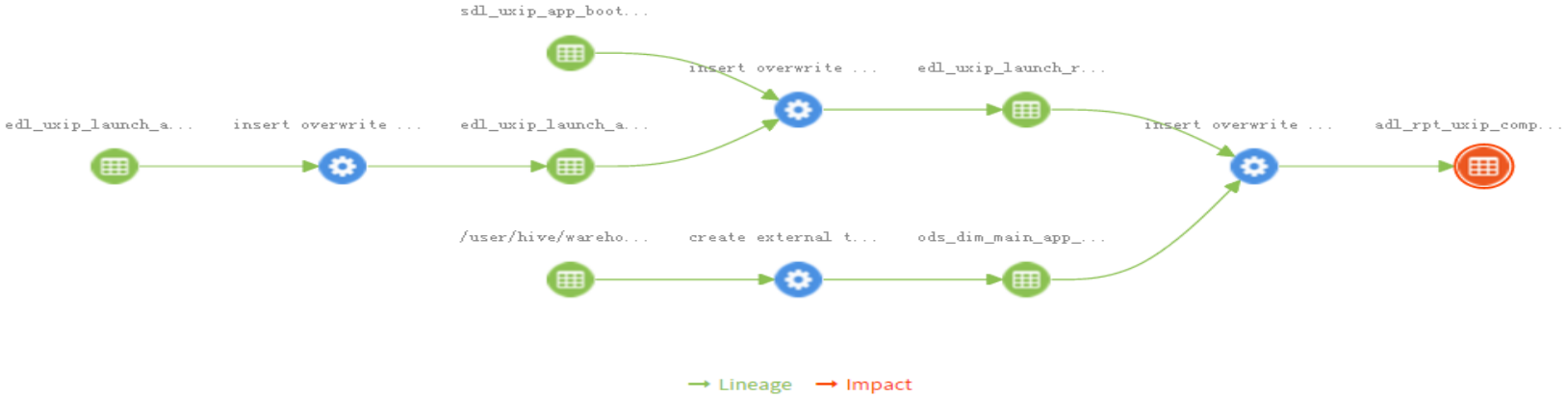
元数据平台架构


MetaData Source/数据采集
Hive
Cetus
（离线任务）
Lyra
（统计分析）
UIP
（洞察平台）
WEB UI
MetaData API
<HTTP/REST>
数据采集
元数据类型
元数据管理
血缘查询
采集器管理
MetaData Core
数据定义
数据转换
功能封装
HiveHook
查询
Atlas

元数据采集


Hive元数据
HiveHook
Kafka
DB
离线任务
分钟级增量同步
Mz-Loader
（数据交换平台）
Sct
权限、统计分析
报表关联
Pull From SCT APIs
Lyra
统计分析平台
实时上报
线下数据
模型、指标、人
工录入等
实时上传
元数据API
元数据存
储

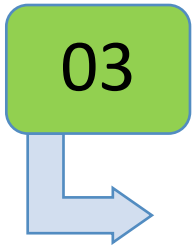
元数据应用-血缘分析



hive_process:

```
insert overwrite table edl_uxip_launch_retention_rate partition(stat_date=20170924)
SELECT *
FROM
     (SELECT *
     FROM sdl_uxip_app_boot_stat_d s
     WHERE stat_date = 20170924
              AND event_name = 'boot_app'
     GROUP BY  umid,packagename) s
RIGHT JOIN
     (SELECT *
     FROM edl_uxip_launch_app_user t
     WHERE (stat_date >= 20170917
              AND stat_date <= 20170923)
              OR (stat_date IN (20170910, 20170825))) t
     ON s.umid=t.umid
          AND s.packagename=t.packagename
GROUP BY  t.packagename ,t.stat_date) s
```

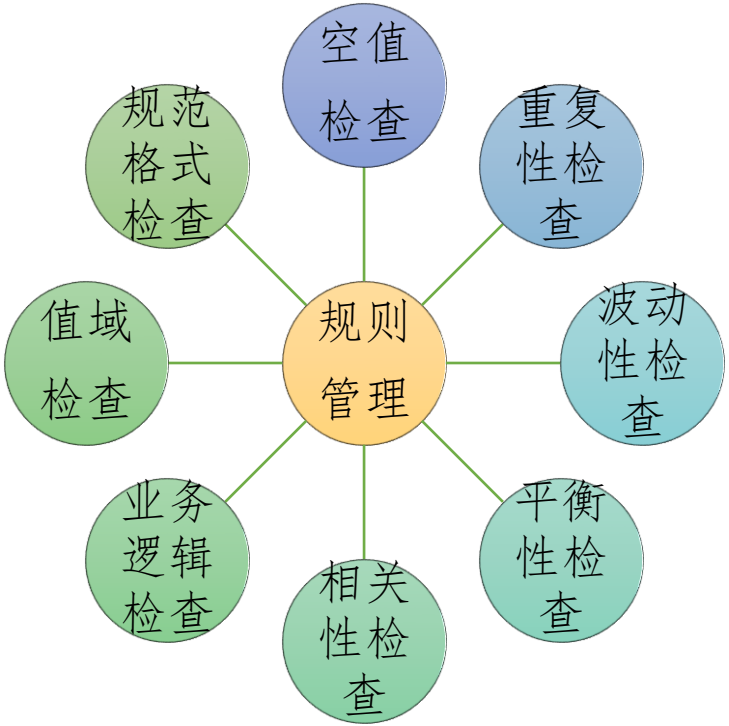
大纲



- 数据质量管理及评价体系

如何评价数据质量？

度量规则



质量管理系统

故障管理平台

告警平台

流程管理平台（CVN）

度量规则管理

- 基础数据度量规则
- 指标度量规则

质量问题发现

- 数据核检结果
- 质量问题提交
- 质量问题告警

质量问题分析

- 数据质量分析报告
- 数据质量对比分析

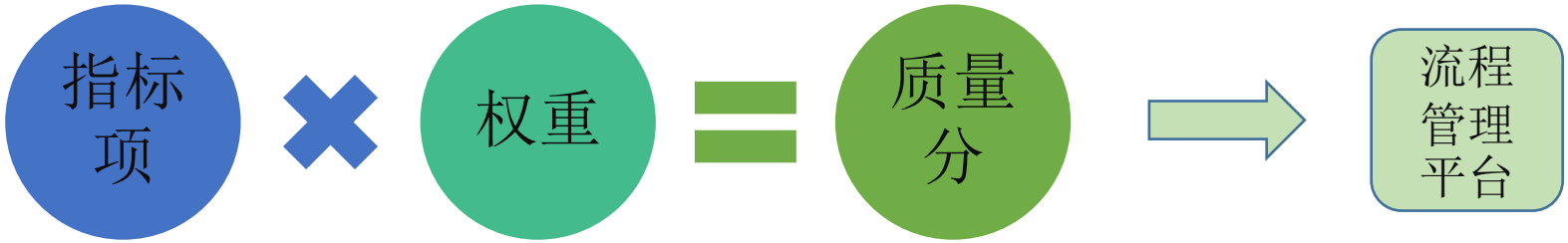
查询功能

- 检核规则查询
- 检核结果查询
- 规则执行情况查询(执行时间、时长、资源)

检核规则调度

- HQL任务完成后自动执行规则调度
- 独立任务调度

质量评价体系



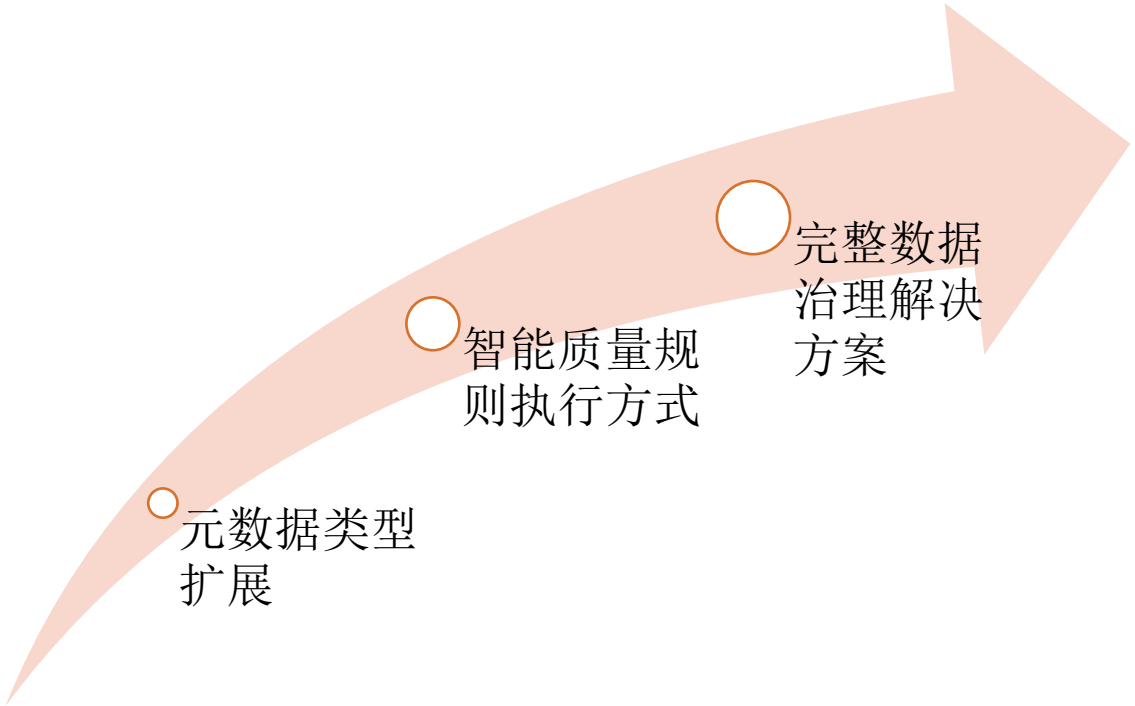
指标定义	1.离线任务失败个数
	2.离线任务整体延迟分钟数
	3.九点之前未完成的任务数
	4.流平台告警次数
	5.规则执行失败次数

大纲

04

- 未来规划

未来规划



The End