2021

DataFunSummit

数据治理与安全

在线峰会

数据治理论坛

2021.08.28，09:00 - 17:30

DataFunSummit

DataFunSummit

2021

小米数据管理与应用实践

勇幸 消息中间件与数据管理负责人

DataFunSummit

引言

数据管理的核心是元数据平台的建设，以元数据支撑数据管理上层应用

目录

01

元数据平台建设

元数据平台的建设现状与架构演进

主要从元数据基础信息、资产信息、衍生信息、作业信息及血缘信息等方面介绍平台的建设情况

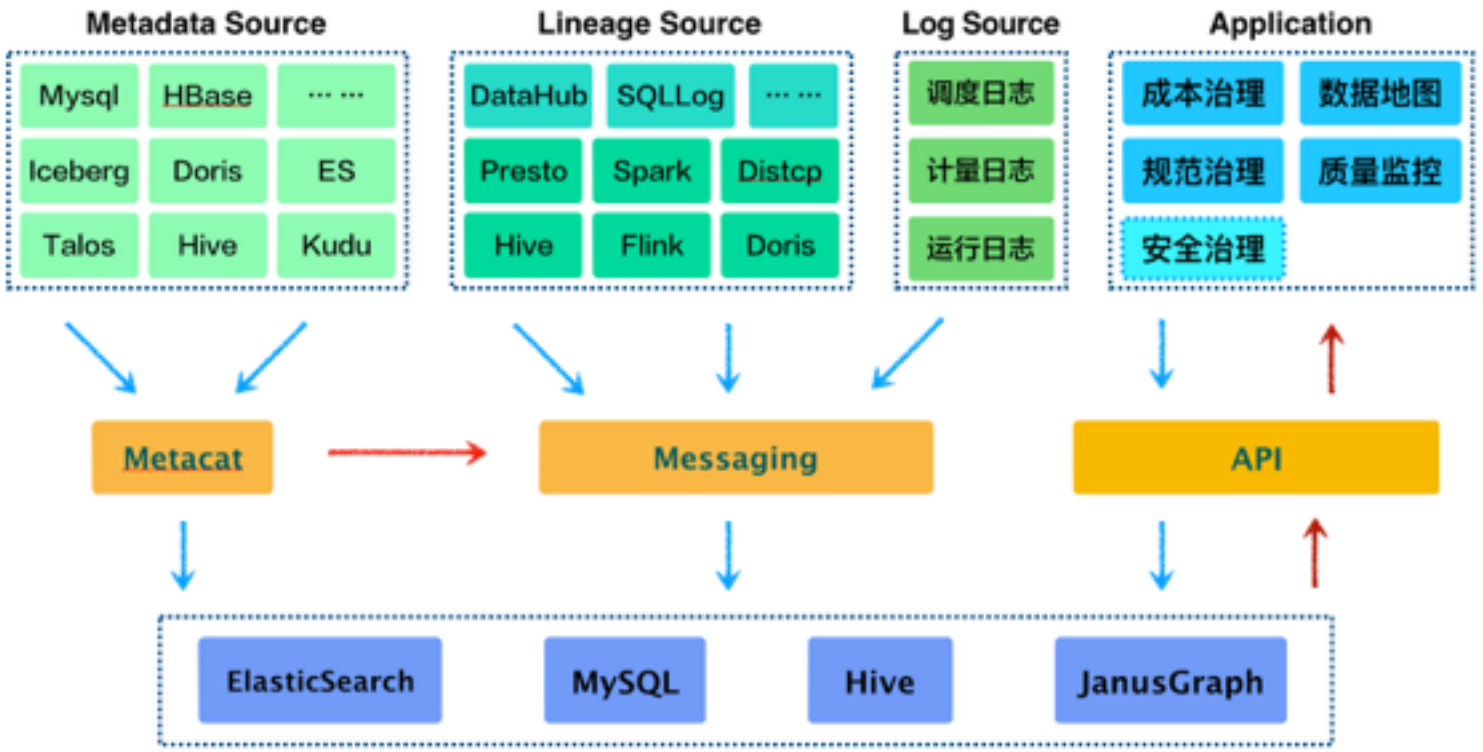
元数据平台 | 元数据

元数据	内容	内容来源	支撑资产管理
技术元数据	表	Hive/Doris/Kudu/MQ/ES/Iceberg	资产地图
	作业	ETL/SQL/Query	
生产元数据	生产	调度系统/Yarn	数据质量 成本治理
业务元数据	数仓分层	建模规范	资产价值 安全治理 规范治理
	数据分类	业务	
	指标关联	指标系统	
	应用信息	BI 看板、报表等	
	隐私分级	业务	
衍生元数据	存储计量	HDFS-Image/Doris/Kudu/MQ/ES	成本治理 资产价值
	访问计量	HDFS-Log/SQL-Log	
血缘元数据	表血缘	Spark/Flink/Presto/DataHub/Doris	资产地图 影响分析
	字段血缘	SQL-Log	

元数据：描述数据的数据

- 实体：
  - 表元数据
  - 作业元数据
- 属性：
  - 业务元数据
  - 衍生元数据
- 关系：
  - 血缘元数据

元数据平台 | 技术架构


Metadata Source
Mysql
HBase
··· ···
Iceberg
Doris
ES
Talos
Hive
Kudu
Lineage Source
DataHub
SQLLog
··· ···
Presto
Spark
Distcp
Hive
Flink
Doris
Log Source
调度日志
计量日志
运行日志
Application
成本治理
数据地图
规范治理
质量监控
安全治理
Source & 服务应用
Metacat
Messaging
API
Integration
ElasticSearch
MySQL
Hive
JanusGraph
Storage

元数据平台 | 演化过程：全域元数据

元数据架构演化：全域

域拓展

- Hive

==>

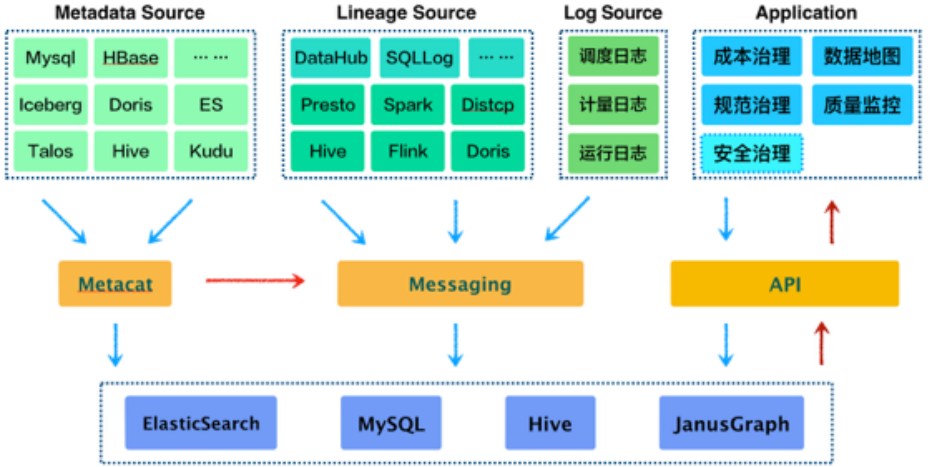
- MySQL/Talos/Hive/Doris/Kudu/ES/Iceberg

统一元数据

- Hive Metastore

==>

- 引入 Metacat 统一元数据视角与管理



元数据平台 | 演化过程：实时血缘

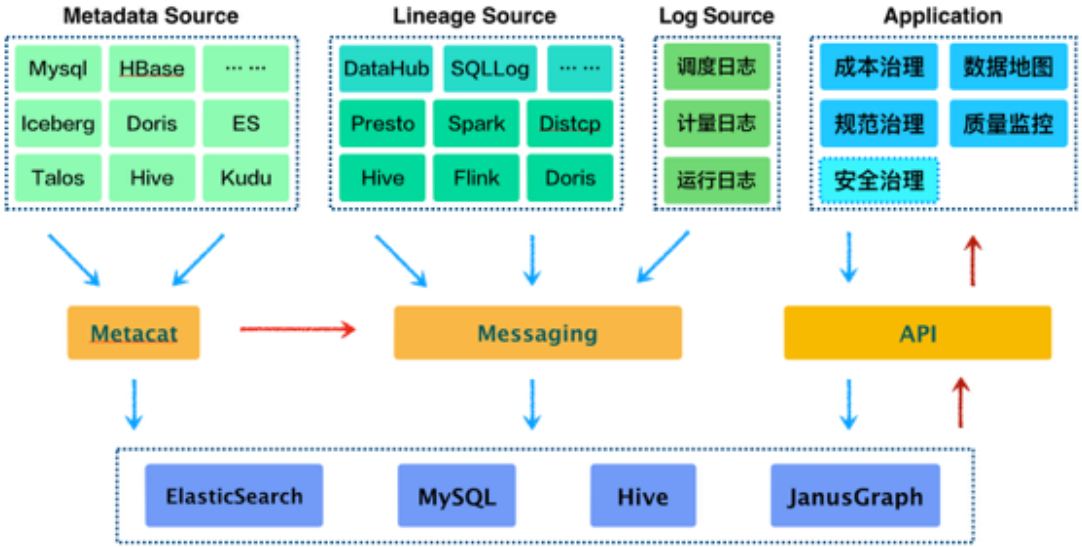
血缘架构演化：实时

原方案

- 解析 HDFS 日志
- T+1
- 不准确

新方案

- 引擎埋点
- 准实时
- 精准解析
- 结合：SQL Proxy Log



元数据平台 | 演化过程：精准计量

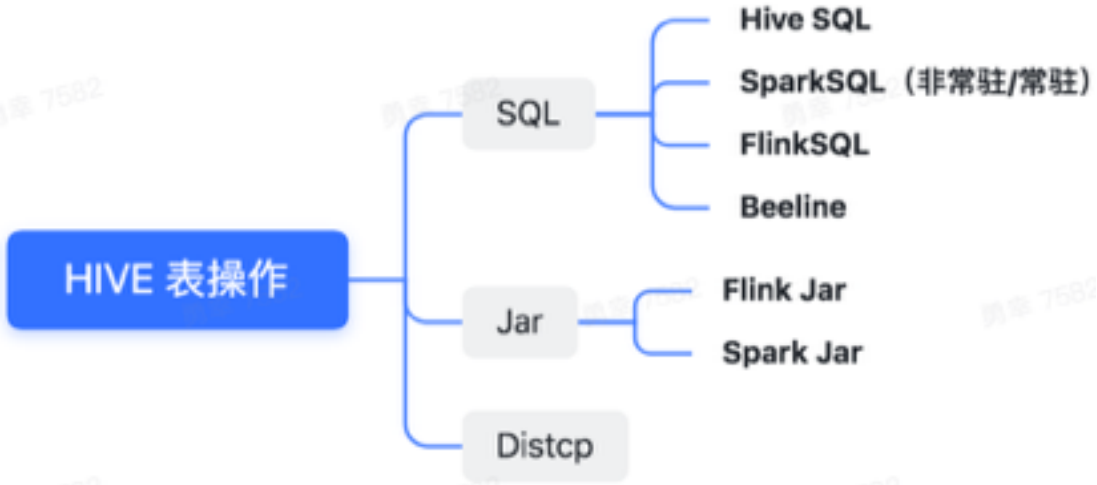
访问计量架构演化：解决 0 与 1

价值

- 数据冷热程度
- 入口不收敛，被访问计成没访问?

方案

- 解析 HDFS 日志
- 结合 SQL 审计做修正



02

元数据应用

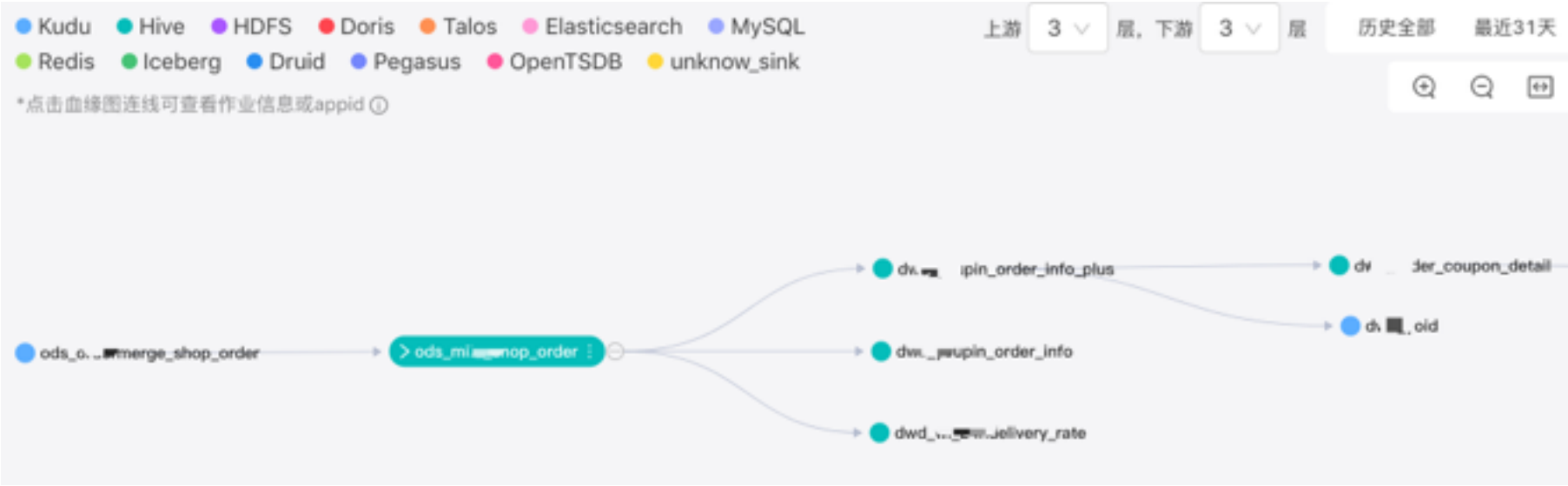
- 数据地图
- 数据规范
- 成本治理
- 质量建设

元数据应用 | 数据地图-搜索

元数据搜索与发现

- 支持表、字段、描述信息、数仓分层、数据分类、标签、部门等信息搜索
- 全域元数据的搜索（完善中）（Talos/Hive/Doris/Kudu/Iceberg/ES/MySQL）
- 支持指标、维度、看板等信息的搜索（未来）

元数据应用 | 数据地图-血缘



数据血缘

- 支持数据处理全链路的血缘展示
- 血缘搜索（完善中）
- 变更通知（完善中）

元数据应用 | 规范治理

规范度

完善度

建模规范度

- 命名：命名是否符合规范
- 分层：超过 70% 的表没有按数仓规范分层
- 打标：数据域分类、标签等，没有打标

建模完善度

- 跨层引用：DWS/ADS 直接访问 ODS
- 查询覆盖：Ad-hoc 查询命中 DWD/DWS/ADS

元数据应用 | 成本治理（存储）

成本分析优化闭环

- 观现状
- 查问题
- 做优化
- 拿反馈

账单逻辑

- 数出一孔
- 天级账单
- 按人归属
- 即时预估

元数据应用 | 成本治理（存储）

排序	一级部门	月总成本	占比	月已优化成本	占比
1					
2					
3					
4					
5					
6					
7					
8					
9					

成本分析：大盘 & 下钻到人

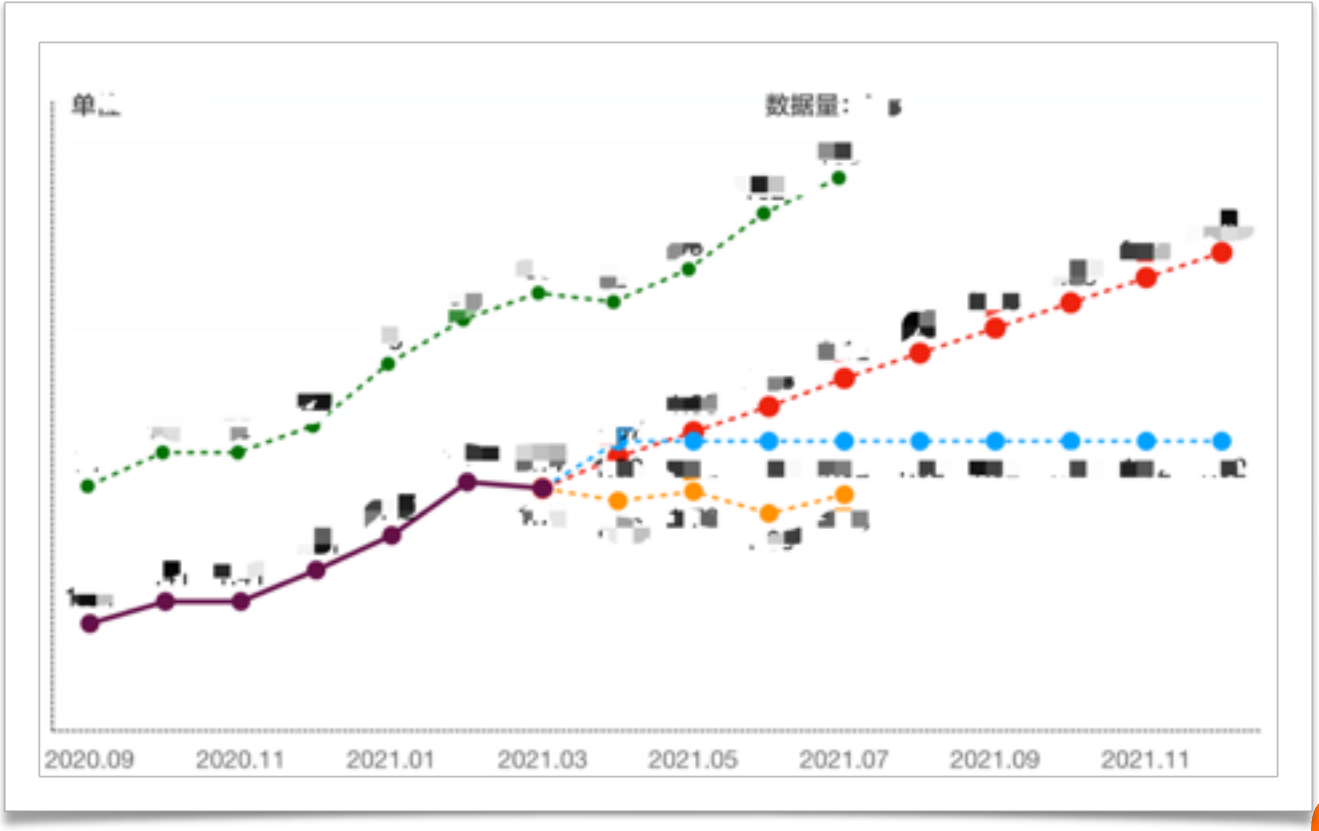
- 公司看部门
- 部门看子部门
- 小组看个人
- 个人看名下的表

元数据应用 | 成本治理（存储）

成本优化

- 冷备（低频访问）
- 删除（冷数据）
- 生命周期管理

元数据应用 | 成本治理（存储）



存储成本优化效果
（模拟数字）

- 数据量增长趋势线
- 成本历史线
- 成本趋势线（业务正常增长）
- 成本停滞线（业务不增长）
- 成本实际走势曲线

元数据应用 | 质量建设

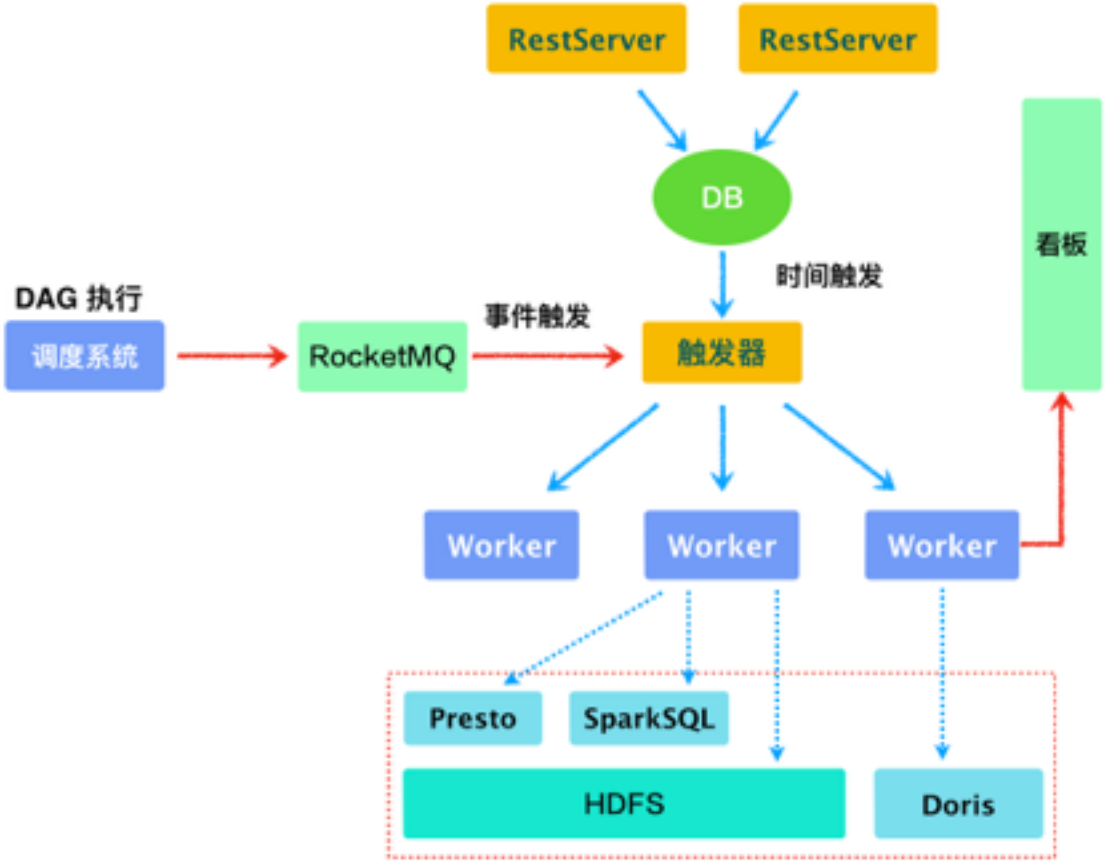
id	name	type	scope	catalog	database_name	table_name	target	cal_expr
1	xx表主键唯一	唯一性	private				Hive	[RULE_1]=TRUE
2	xx表主键非空	准确性	private				Hive	[RULE_2]=TRUE
3	xx表数据量符合预期	完整性	private				Hive	([RULE_3] - [RULE_4]) / [RULE_4] < 0.2
4	xx表空值率符合预期	完整性	private				Hive	([RULE_5] - [RULE_6]) / [RULE_6] < 0.2
5	xx表取值范围符合预期-数值型	正确性	private				Hive	[RULE_7]=TRUE
6	xx表取值范围符合预期	正确性	private				Hive	[RULE_8]=TRUE
7	xx表取值关联符合预期	正确性	private				Hive	[RULE_9]=TRUE
8	xx表和yy表取值关联符合预期	一致性	private				Hive	[RULE_10]=TRUE
9	xx表zz字段分布符合预期	正确性	private				Hive	[RULE_11]=TRUE
10	xx表yy字段格式统一-日期	统一性	private				Hive	[RULE_12]=TRUE
11	xx表yy字段格式统一-数值	统一性	private				Hive	[RULE_13]=TRUE

数据内容质量检查

- 及时性（数据生产保障，建设中）

- 唯一性
- 准确性
- 完整性
- 一致性

元数据应用 | 质量建设



技术架构

- 时间触发、事件触发
- 可扩展无状态 Worker
- 多数据源设计（Hive/HDFS/Doris）
- 便捷的规则模板与产品化

03

未来规划

- 数据管理长期路线

未来规划 | 生产保障联动资源调度

数据生产时效保障：基线 -> 作业 -> 调度 -> Yarn 全链路打通

- 基线管理：基线级别、产出时间
- 生产执行：联动 Yarn Job 优先级支持
- 监控预警：执行进度、破线预警

计算资源治理：

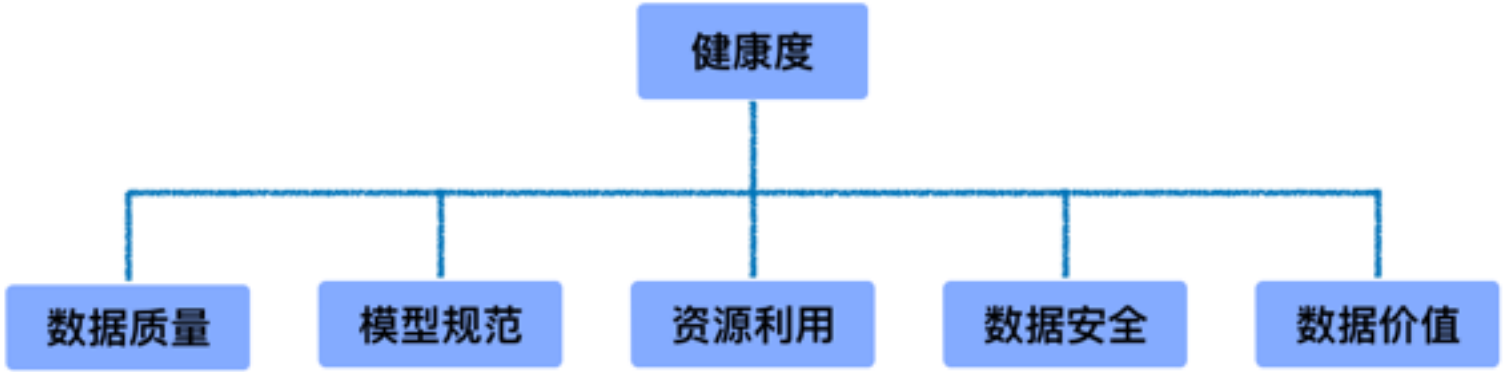
- 僵尸作业
- 暴力扫描
- 数据倾斜
- ……

未来规划 | 长期路线

元数据建设长期路线：回答好 2 个问题，

- **数据健康定义**：什么样的数据是一份好的数据？数据的健康程度如何？
- **数据健康治理**：一份不健康或不够好的数据，应该如何治理？治理后有什么收益？

- 从数据治理、模型规范、资源使用、数据安全、数据价值等方面总和定义数据健康度



未来规划 | 业务赋能

（讨论）如何让业务愿意把数据接入到中台，从业务痛点出发

质量：

- 重保数据能够保障产出
- 数据产出后的质量检查

效率：

- 规范建模、查询优化让出数加快
- 找数加快
- 问题追溯

成本：

THANKS!