火山引擎 | DataFun.

字节跳动血缘技术实现与具体用例

彭洪剑　火山引擎DataLeap研发工程师

目录
CONTENT

火山引擎 | DataFun.

01

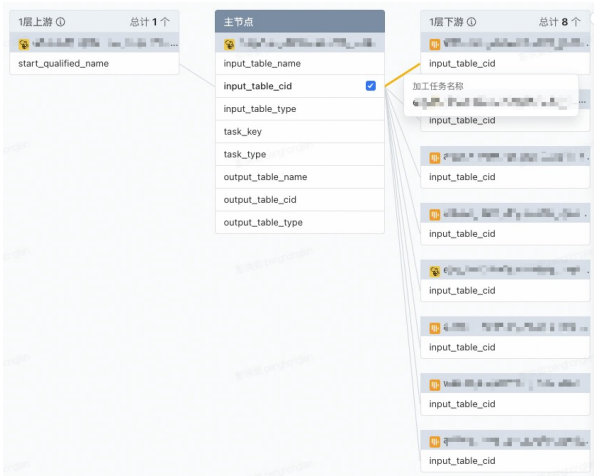
数据血缘模型

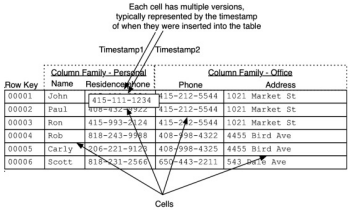
数据血缘模型- 挑战

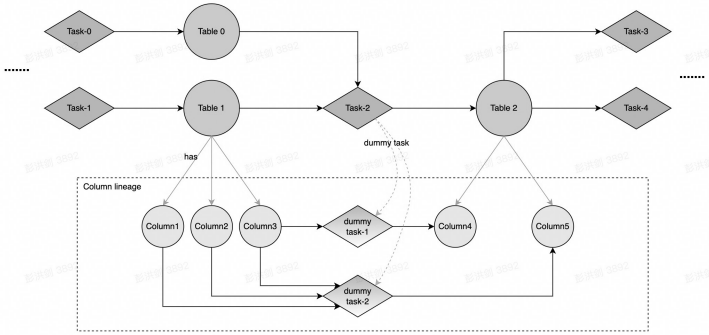
- **如何应对复杂庞大业务、海量存储和任务血缘？**
  - 如何保证模型的扩展性？
  - 如何高效接入血缘？输出血缘？
  - 如何保证血缘的实效性？
  - 赋能**业务**

数据血缘模型 – 展示层

- 以资产为主的血缘视角
- 支持不同粒度资产血缘



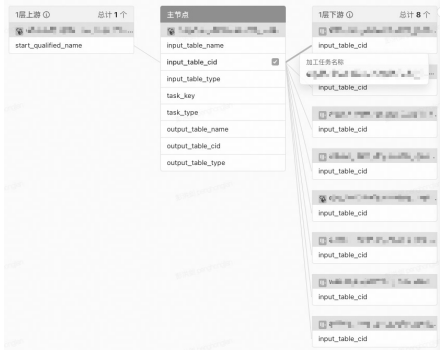


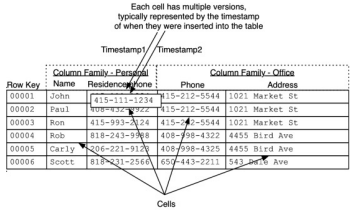


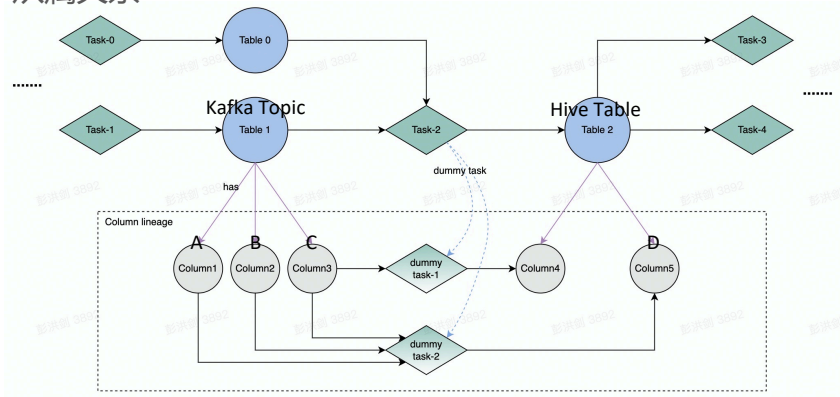
火山引擎 | DataFun.

数据血缘模型 – 抽象层

- **表资产节点**：对于存储数据的介质的抽象
- **任务节点：**对于任务（或链路）的抽象
- **子任务节点：**对于处理过程的抽象
- **字段节点:** 存储数据的介质的子结构
- **资产节点与任务节点之间的边:** 生产消费关系
- **同类型节点之间的边**：从属关系



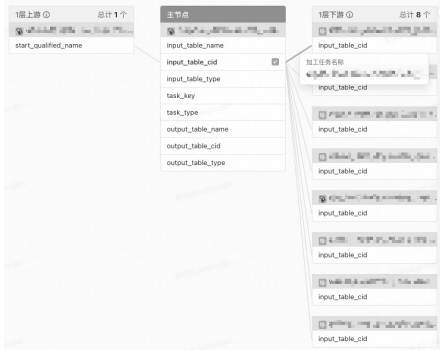


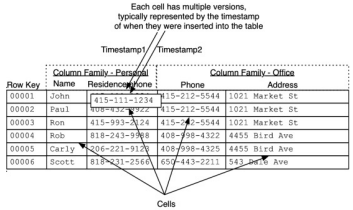


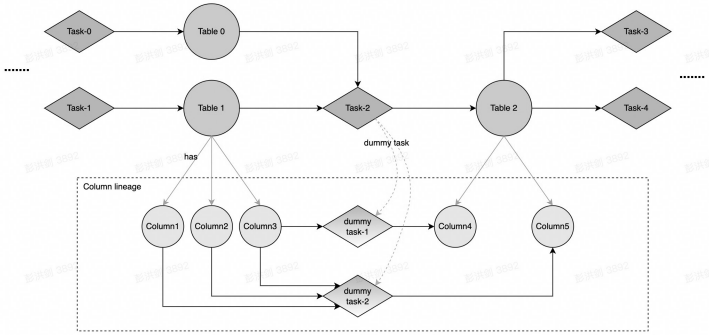
火山引擎 | DataFun.

数据血缘模型 – 实现层

- 目前基于Apache Atlas实现
- 扩展任务/资产的类型定义

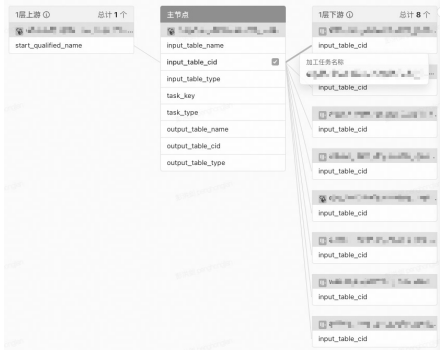


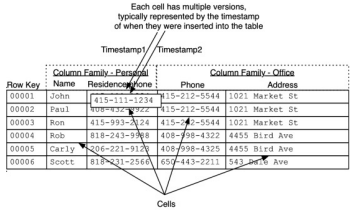


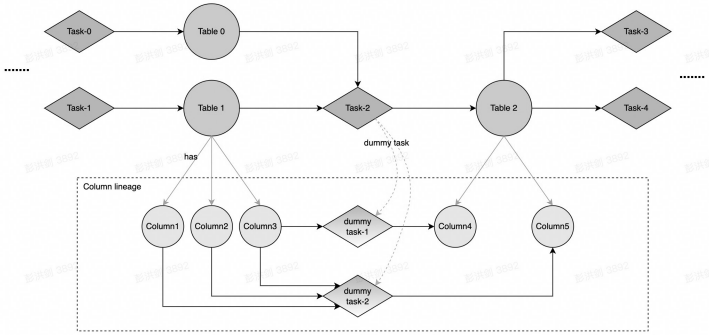


数据血缘模型 – 存储层

- 数据库扩展
- 支持自研图数据库，OLTP数据库

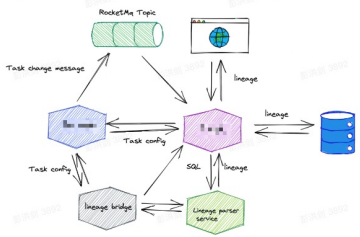






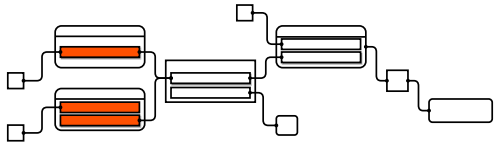
02
数据血缘优化

数据血缘优化



实时数据血缘

- 实时数据血缘更新链路



血缘查询优化

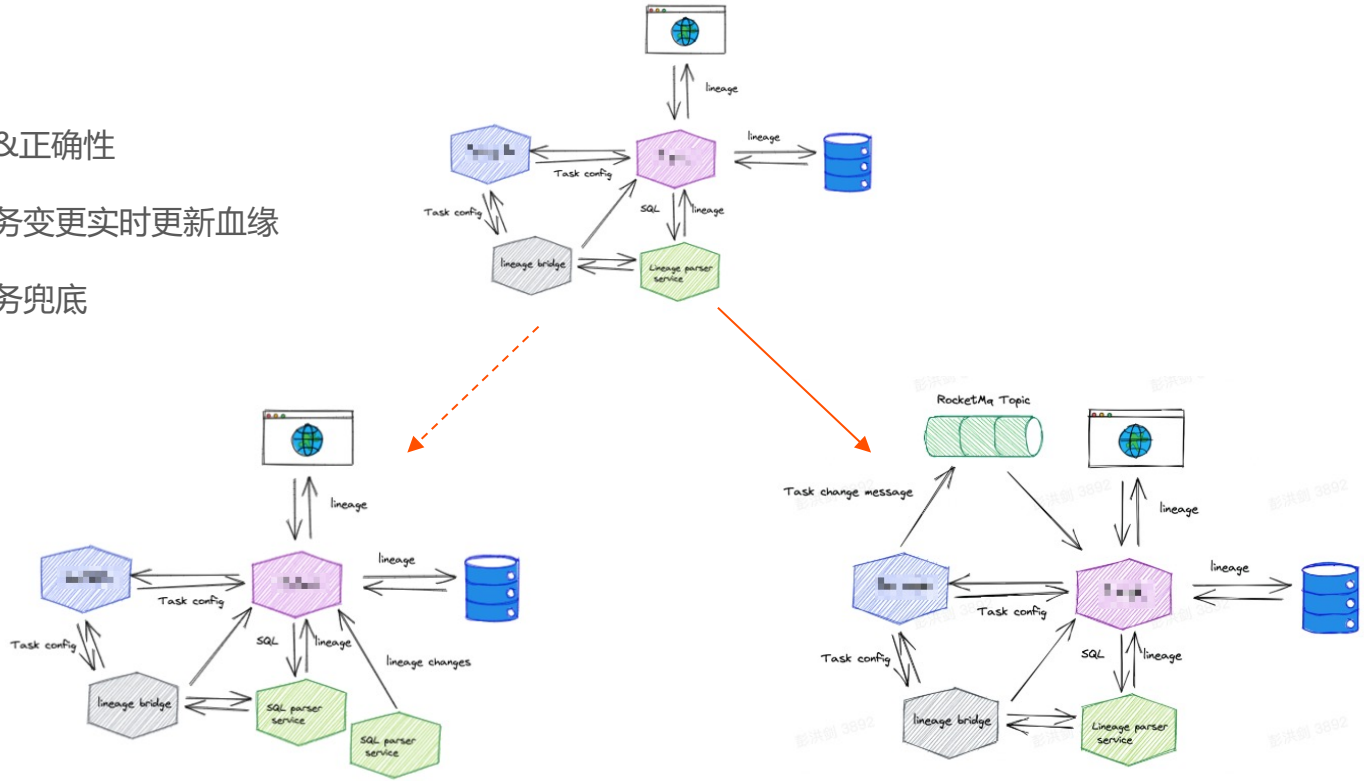
- 批量顶点并发查询

血缘数据开放式导出

- Catalog系统/数仓/API/Topic订阅

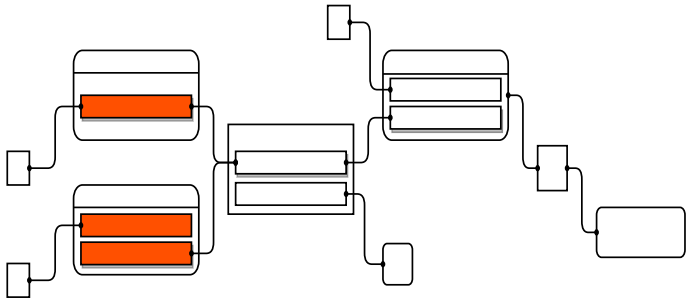
数据血缘优化

- **实时数据血缘**
  - 时效性&正确性
  - 监听任务变更实时更新血缘
  - 离线任务兜底



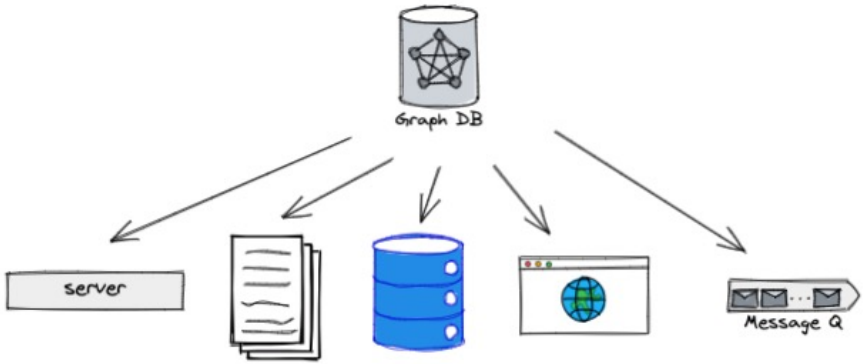
血缘查询优化

- **批量顶点并发查询**
  - 血缘服务批量查询图节点
  - 图节点异步批量转化为实体



数据血缘优化

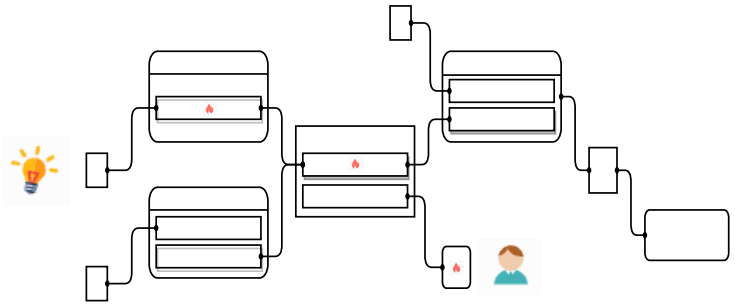
- **血缘数据开放式导出**
  - 元数据系统
  - 数仓
  - API
  - Topic订阅



03
数据血缘用例

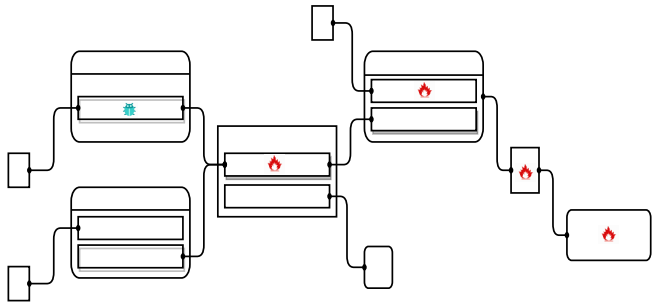
数据血缘用例 – 资产领域

- **资产热度计算**
  - 热度和资产的权威性正相关
  - 资产下游热度反馈上游
- **理解数据**
  - 我是谁？我从哪来？我要到哪里去？
  - 根据上下游判断资产是否满足需求



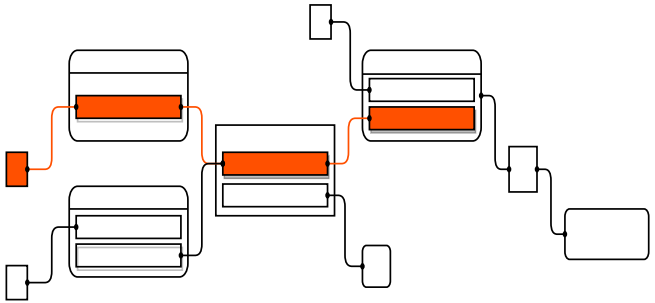
数据血缘用例 – 开发领域

- **影响分析**
  - 判断变更影响范围
  - 通知下游变更
- **归因分析**
  - 透过现象看本质



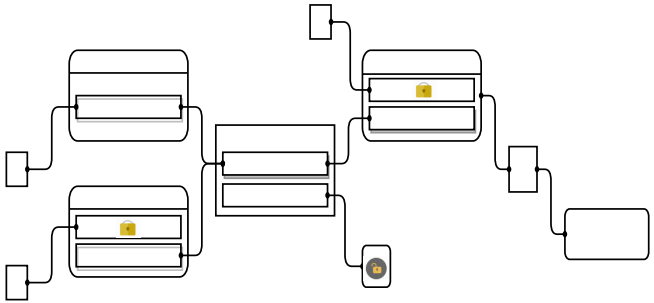
数据血缘用例 – 治理领域

- **链路状态追踪**
  - 重保核心链路
  - 签署链路SLA保障
- **数仓治理**
  - 数仓规范化治理



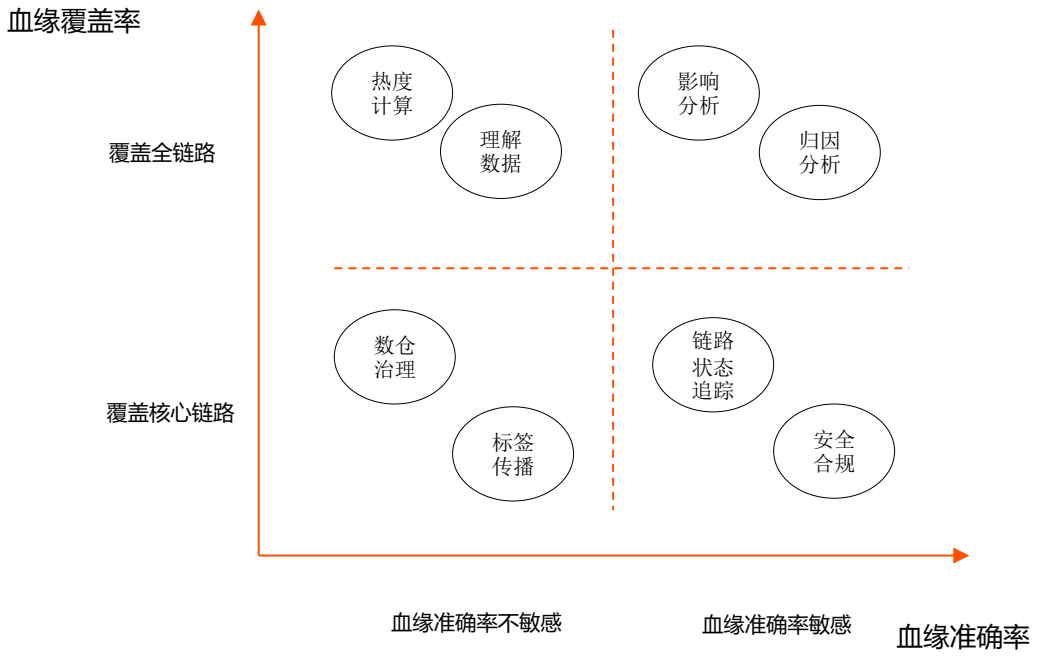
数据血缘用例 – 安全领域

- **安全合规检查**
  - 安全等级排查
- **标签传播**
  - 安全标签向下传播



数据血缘用例

- 血缘覆盖率需求
- 血缘准确率需求
- 血缘实效性需求



04
未来展望

未来展望

- **数据血缘技术趋势**
  - 通用的血缘解析能力
  - 非侵入式的非SQL类型血缘采集
  - 时序血缘
- **数据血缘应用趋势**
  - 标准化
  - 端到端的血缘打通

未来展望

云上的全链路血缘能力

- 全数据源类型覆盖，自定义接入新类型血缘
- 标准化的血缘应用
- 开放的血缘生态

欢迎联系我们

非常感谢您的观看

火山引擎 | DataFun.