火山引擎 | DataFun.

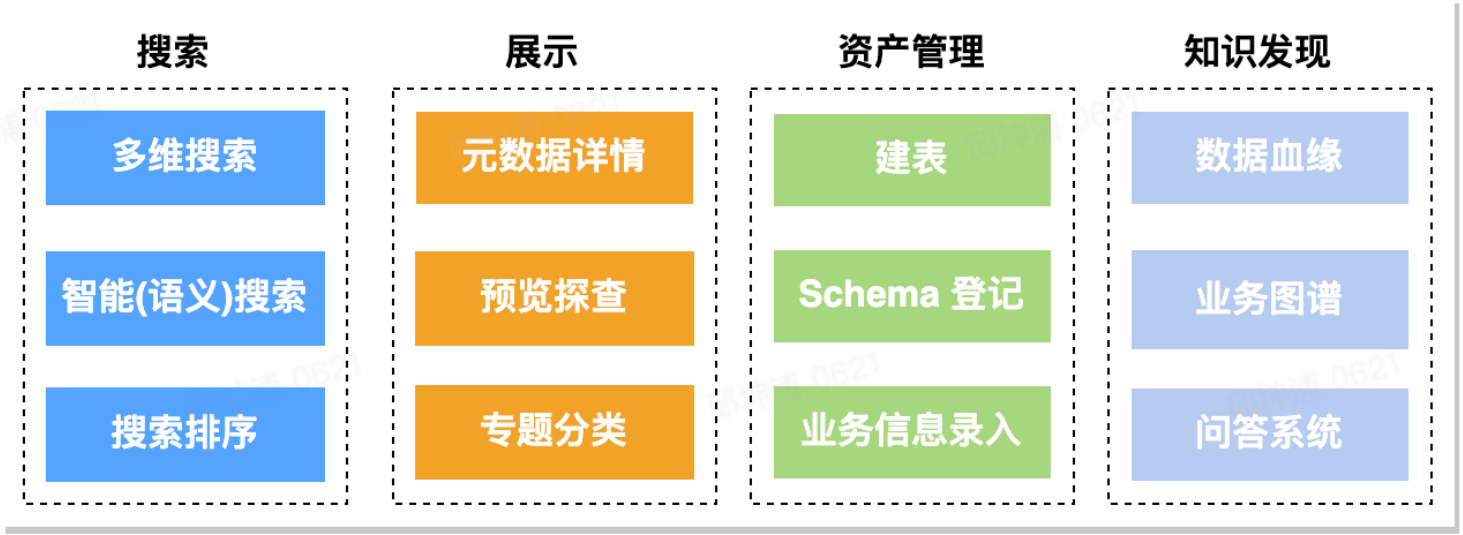
# 字节跳动数据血缘架构的演进之路

罗以亮　火山引擎DataLeap研发工程师

# 目录 CONTENT

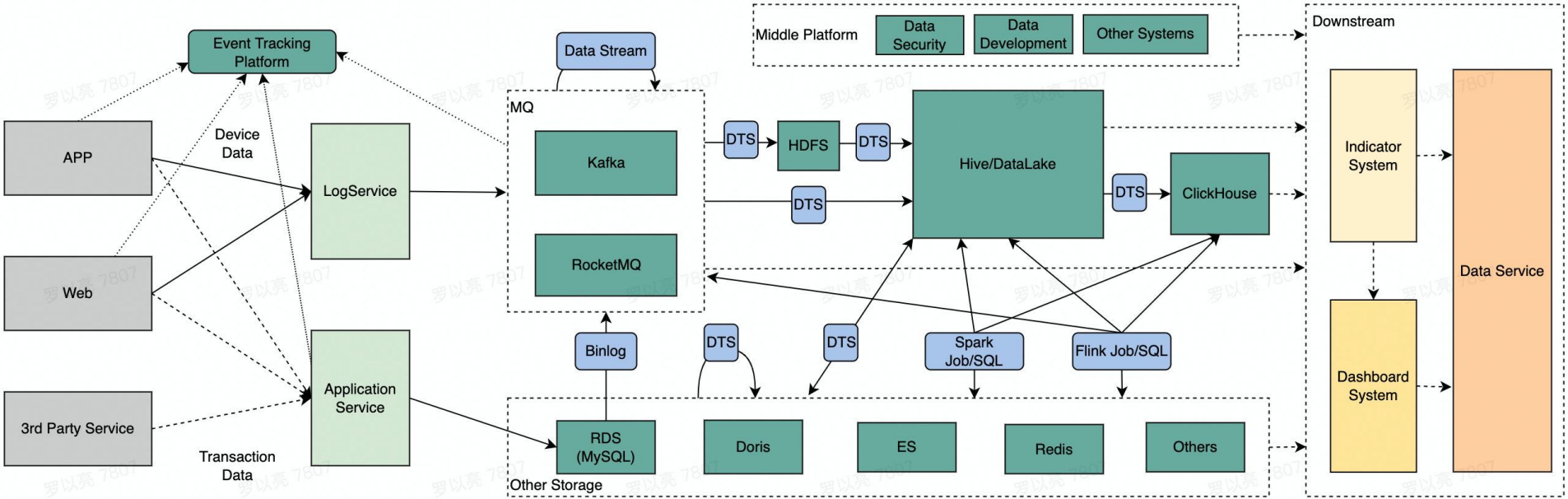
# 01
# 背景介绍

# 数据资产平台介绍




数据血缘是帮助用户找数据、理解数据、以及使数据发挥价值的重要基础能力

# 字节数据链路介绍

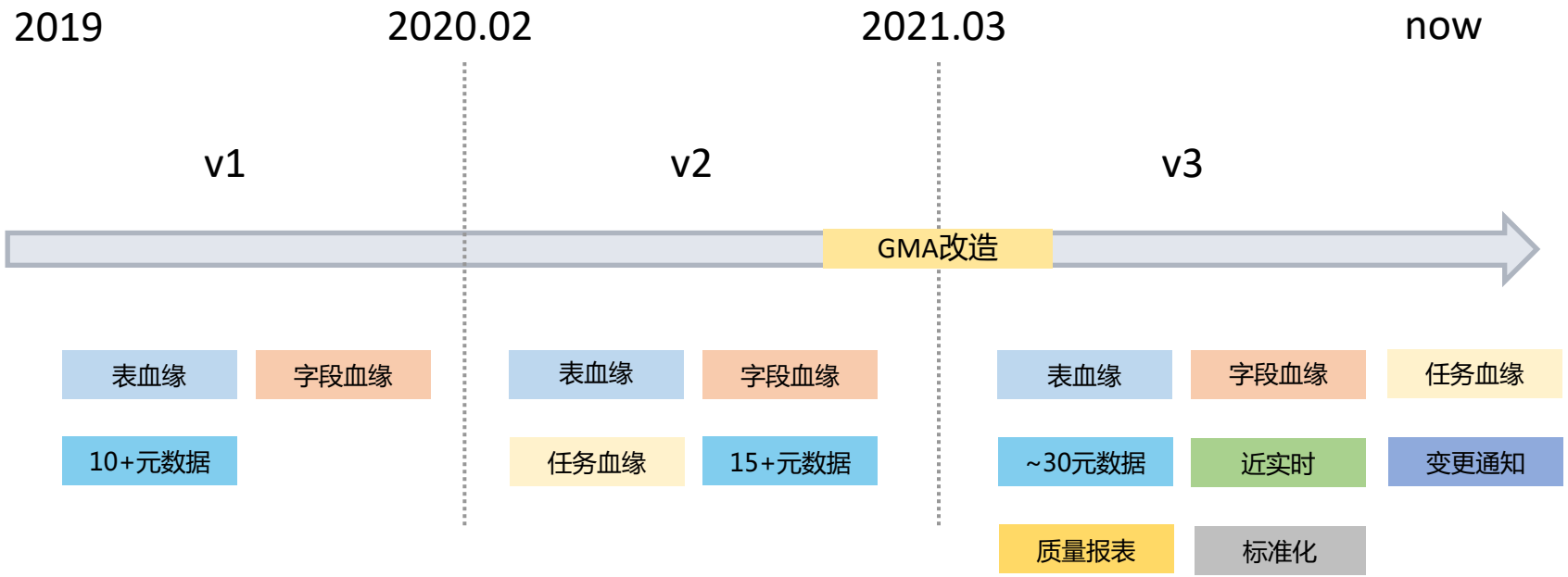



数据来源：埋点数据、业务数据
数据去向：指标系统、报表系统和数据服务
血缘链路：从在线存储和MQ到下游指标、报表系统以及数据服务

# 02
# 血缘发展概况

# 发展概况

GMA: Generalized Metadata Architecture



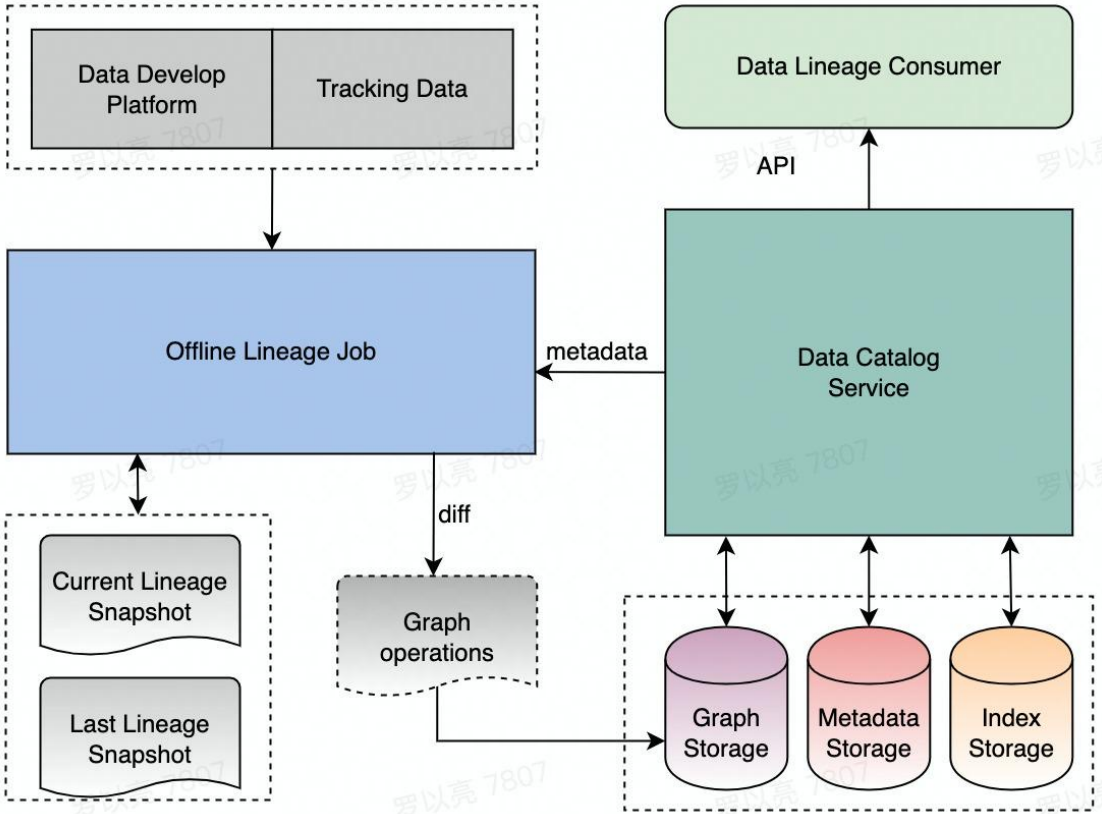
# 03
# 血缘架构演进

# 第一版

血缘基本能力，使用场景初步探索

# 第一版血缘架构

1. 血缘每天全量更新
2. 通过对比血缘快照生成图操作
3. 冗余元数据到图数据库

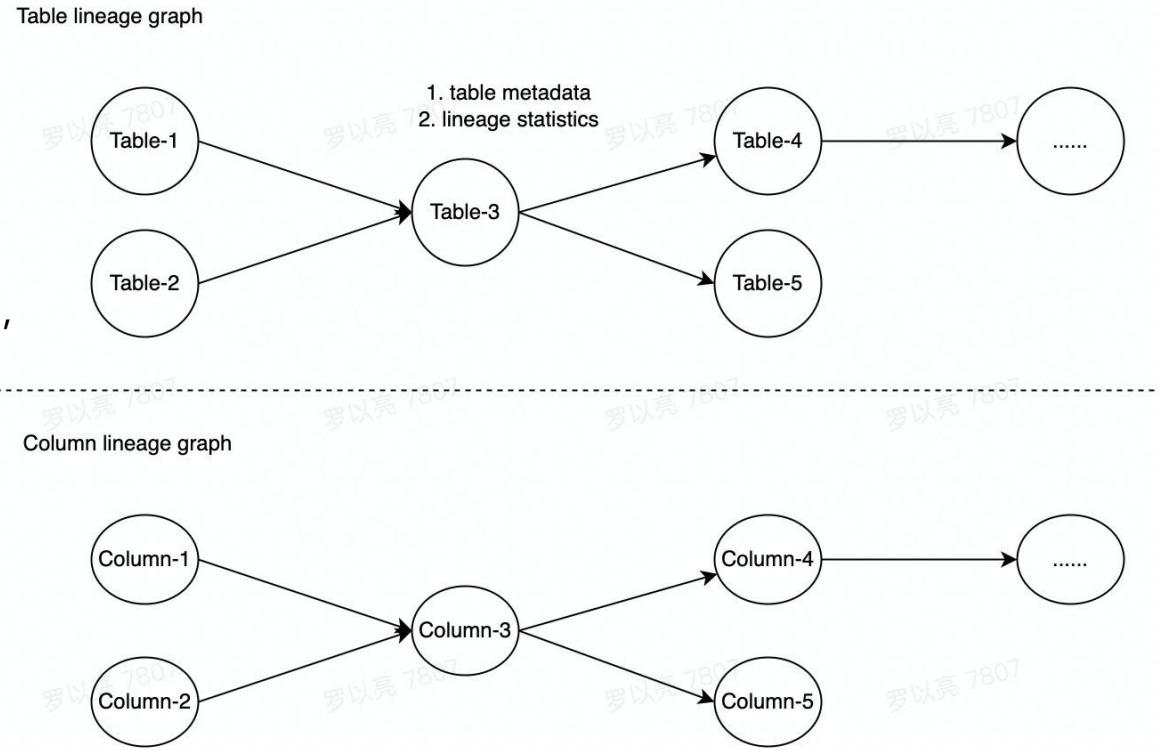


# 第一版存储模型

1. 分离的血缘图谱

2. 冗余元数据

3. 预计算统计信息保存到节点中，空间换时间

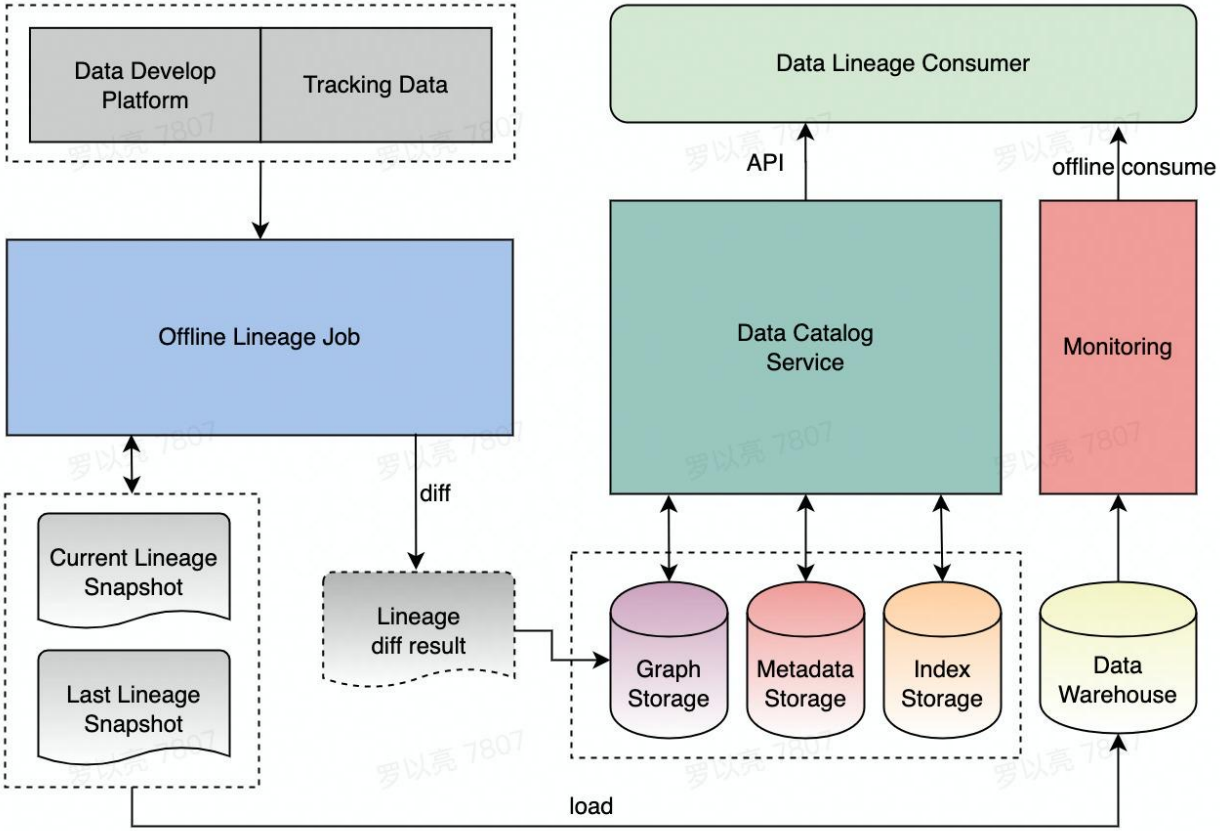


# 第二版

血缘价值逐步体现，使用场景拓宽

# 第二版血缘架构

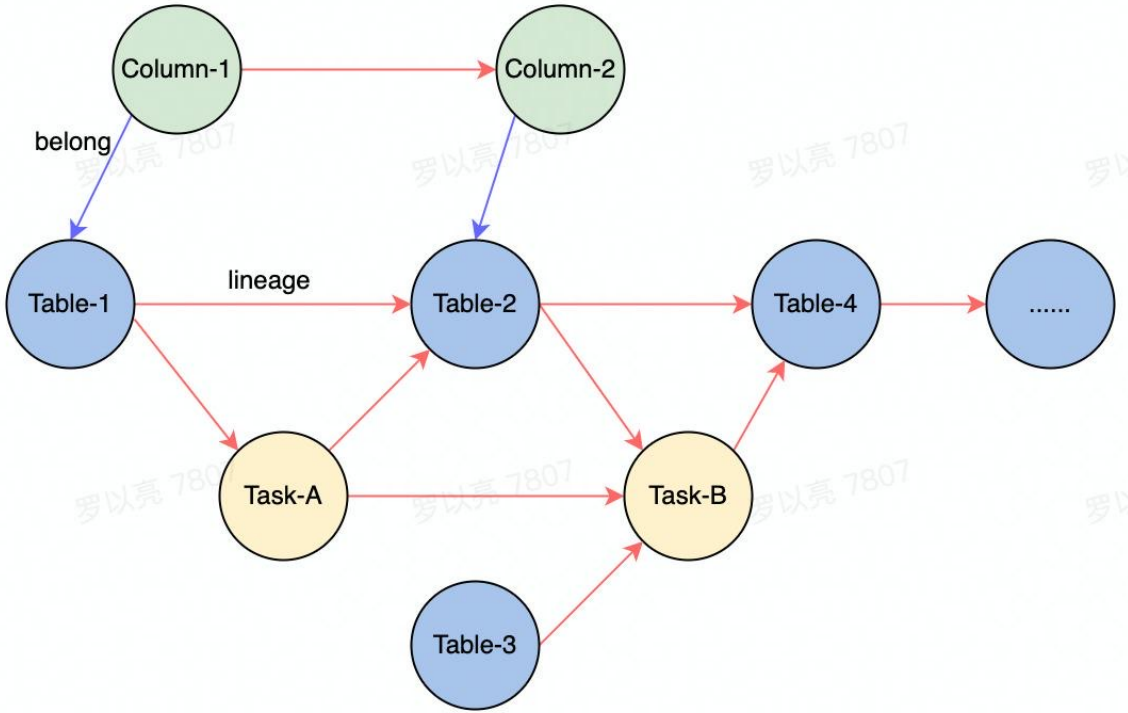
1. 去除元数据的冗余存储
2. 去除血缘统计信息预结算
3. 支持离线消费
4. 全新的存储模型



# 第二版存储模型

1. 引入了任务类型节点

2. 表和字段血缘在同一个图中

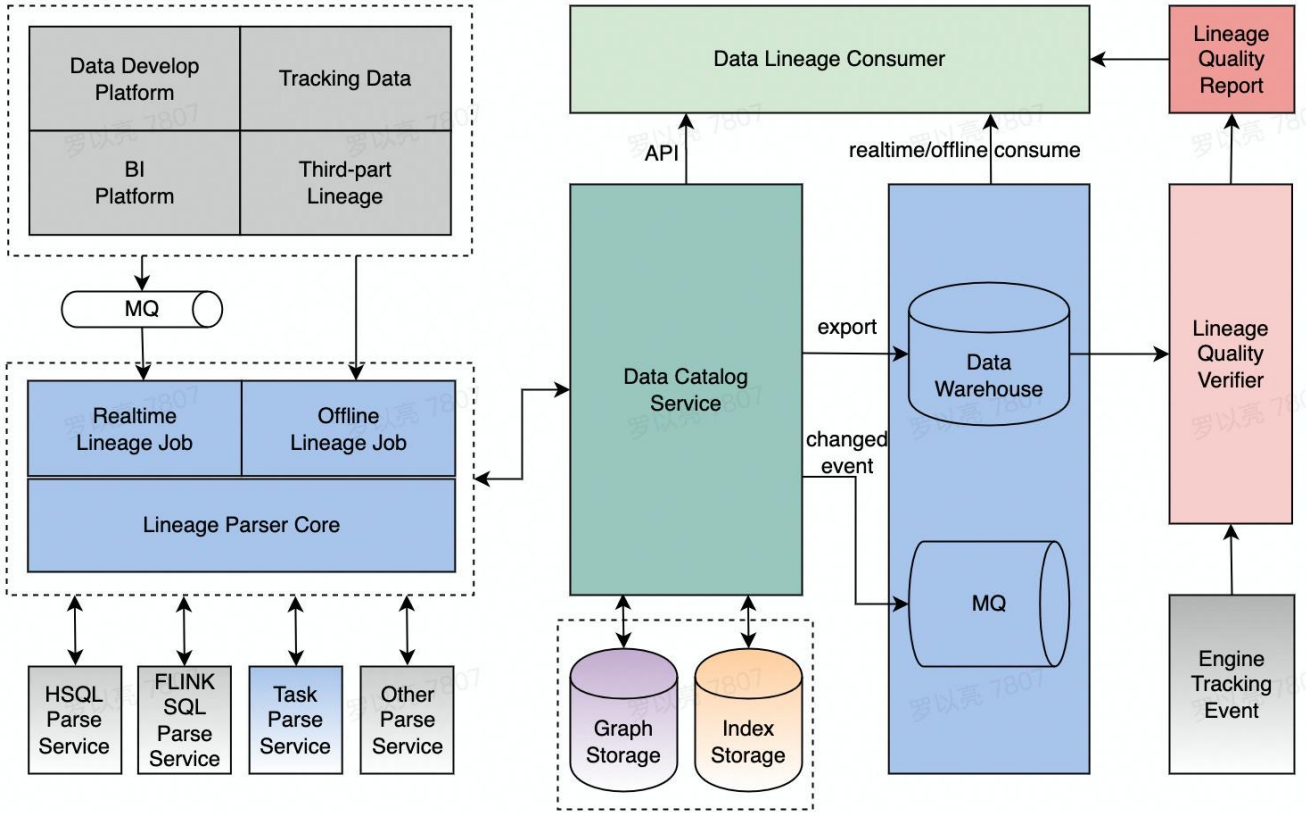


# 第三
# 版

血缘成为数据发挥价值的重要基础能力，对质量要求更高

# 第三版血缘架构

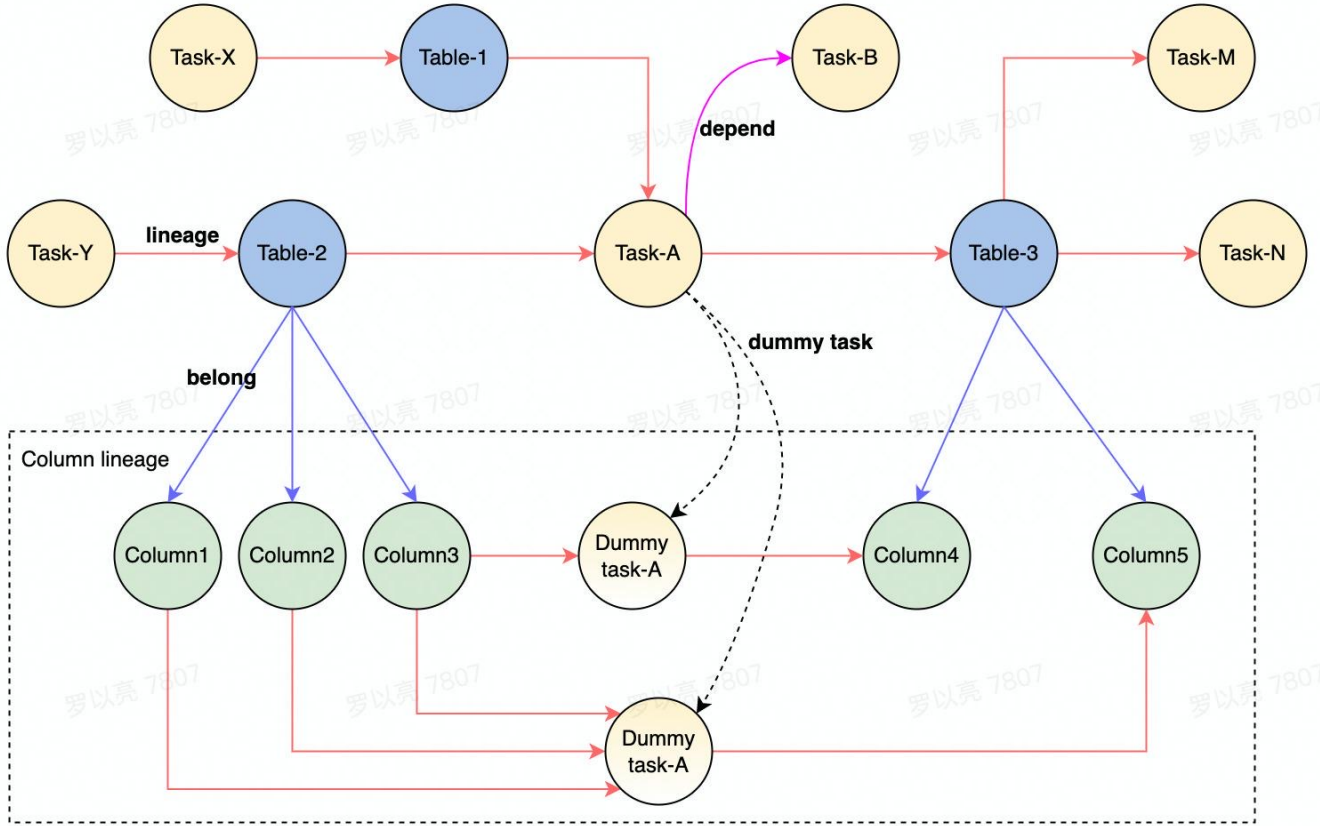
1. 增量和近实时更新
2. 支持血缘标准化接入
3. 插件化的解析服务
4. 更多消费方式
5. 血缘质量报表



# 第三版存储模型

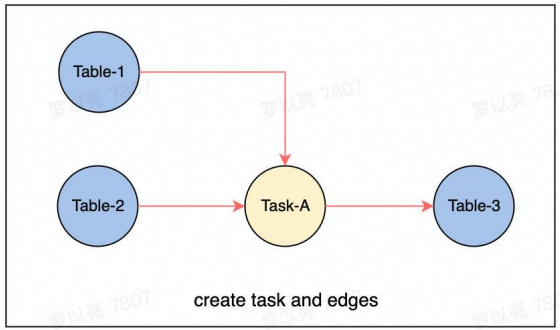
1. 以任务为中心

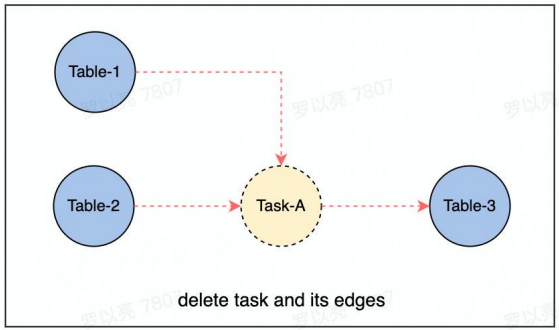
2. 表和字段血缘模型统一

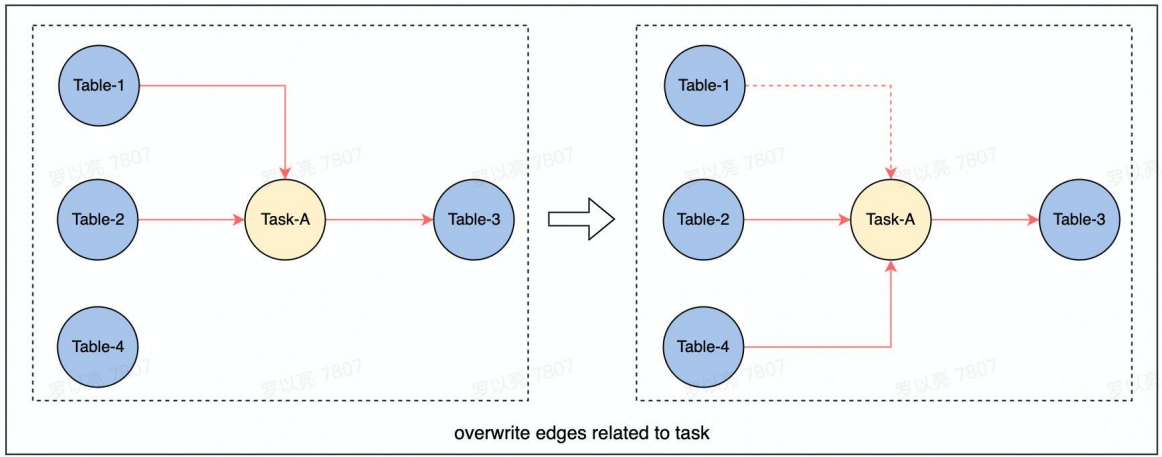


# 增量更新

以任务为中心的存储模式使得血缘增量更新变得简单

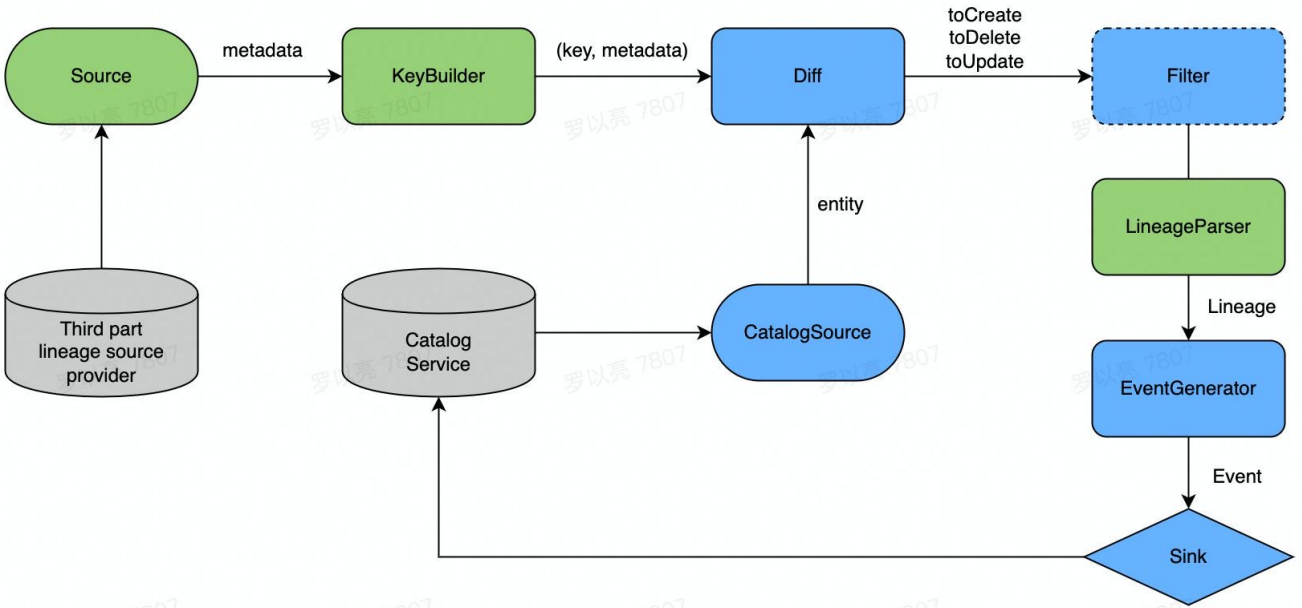


# 血缘标准化

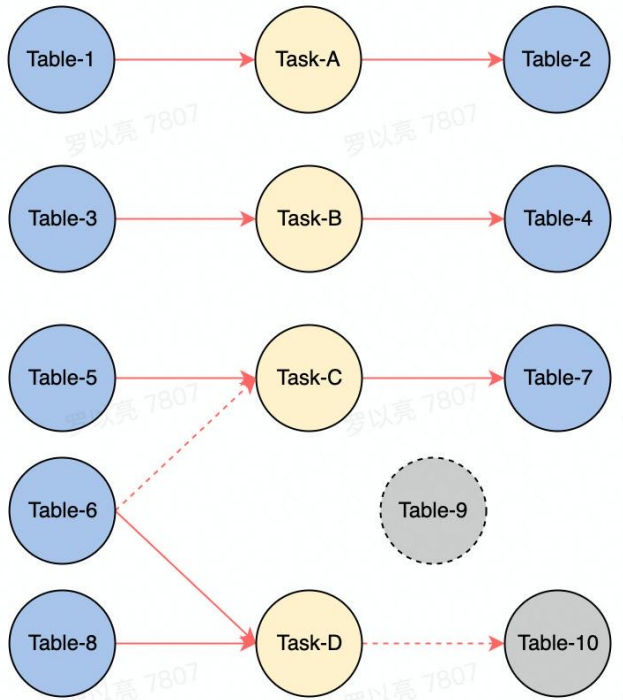
提供通用SDK，复用部分逻辑，高效快速接入血缘



# 血缘数据质量——覆盖率

**覆盖率** = 血缘覆盖的资产数 / 关注的资产数

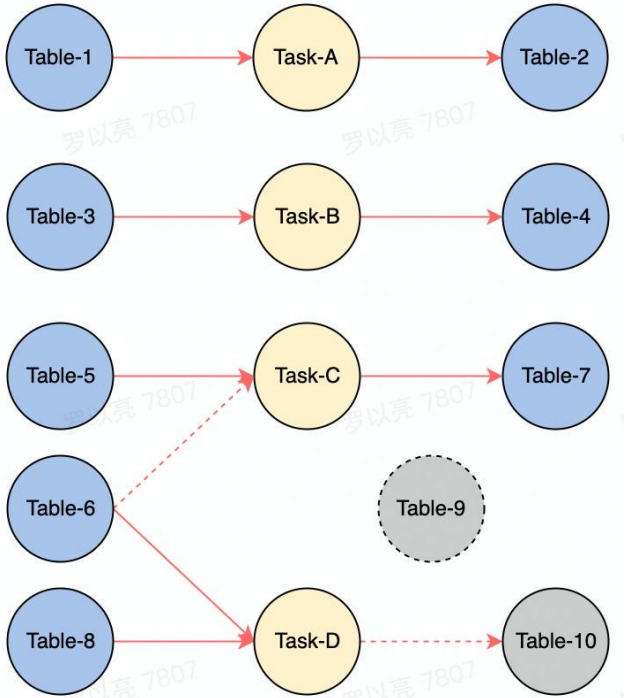
关注的资产: Table1~ 8 + Table10
血缘覆盖的资产: Table1 ~ 8 + Table10
覆盖率: 8/9=88%



# 血缘数据质量——准确率

**准确率** = 血缘准确的任务数 / 同类型的任务数

血缘准确的任务: Task-A and Task-B
血缘不准确的任务: Task-C and Task-D
准确率: 2/4=50%



# 血缘数据质量——字节现状

覆盖率:

Hive 98%

ClickHouse 96%

Kafka 70%

others

准确率:

DTS Task 99%

HSQL Task 97%

FlinkSQL Task 81%

others

# 血缘架构对比

| 内容 | 第一版 | 第二版 | 第三版 |
| --- | --- | --- | --- |
| 血缘消费方式 | API | API、数仓 | API、数仓、MQ |
| 增量更新 | × | × | √ |
| 任务血缘 | × | √ | √ |
| 血缘质量报表 | × | × | √ |
| 与元数据存储统一 | × | × | √ |
| 新血缘接入耗时 | 7~10d | 7~10d | 3~4d |
| | | | |

# 04
# 未来工作

# 未来工作

## 简化

- 接入流程和架构精简
- 插件化能力，支持横向扩展

## 生态化

- 加强对外部生态支持
- 血缘基础能力平台

## 高质量

- 持续提升血缘质量
- 支持异常快速诊断

## 智能化

- 支持智能化场景，如关键链路梳理等

# 欢迎联系我们

# 非常感谢您的观看

火山引擎 | DataFun.