图数据库在金融风控中的应用

- 图模式、时序图与多模态的融合之路

乔云从 （Fabarta高级技术专家）

InfoQ 极客传媒

已结束 北京站 2000 人

QCon 全球软件开发大会 暨智能软件开发生态展

架构演进 | AIGC
研发效能 | 向量数据库

6 月 14–15 日 深圳站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

业务架构 | 高可用
AI 架构 | 架构师成长

10 月 18–19 日 上海站 1500 人

QCon 全球软件开发大会 暨智能软件开发生态展

AIGC | AI 原生应用
软件工程 | 金融科技

4月 5月 6月 8月 10月 12月

5 月 17–18 日 北京站 2000 人

AiCon 全球人工智能开发与应用大会 暨大模型应用生态展

通用大模型 | AI Agent
LLMOps | AI Infra

8 月 16–17 日 上海站 1000 人

FCon 全球金融科技大会 暨智能软件开发生态展

金融大模型 | 核心系统
数字化转型 | 数据安全

12 月 20–21 日 北京站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

未来架构 | 高可用高并发
AI 应用 | 架构师思维

极客邦科技 2024 年会议规划

促进软件开发及相关领域知识与创新的传播

访问大会官网

参会咨询

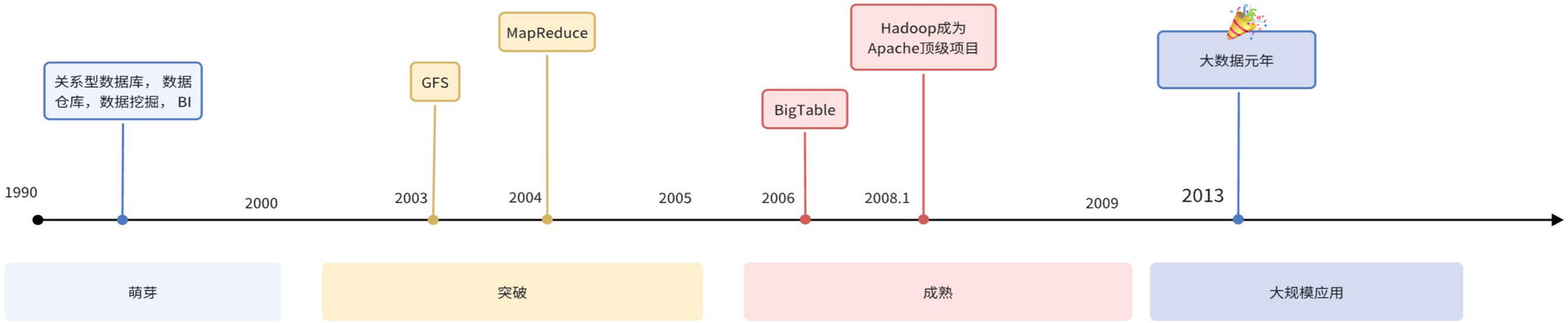
演讲提纲

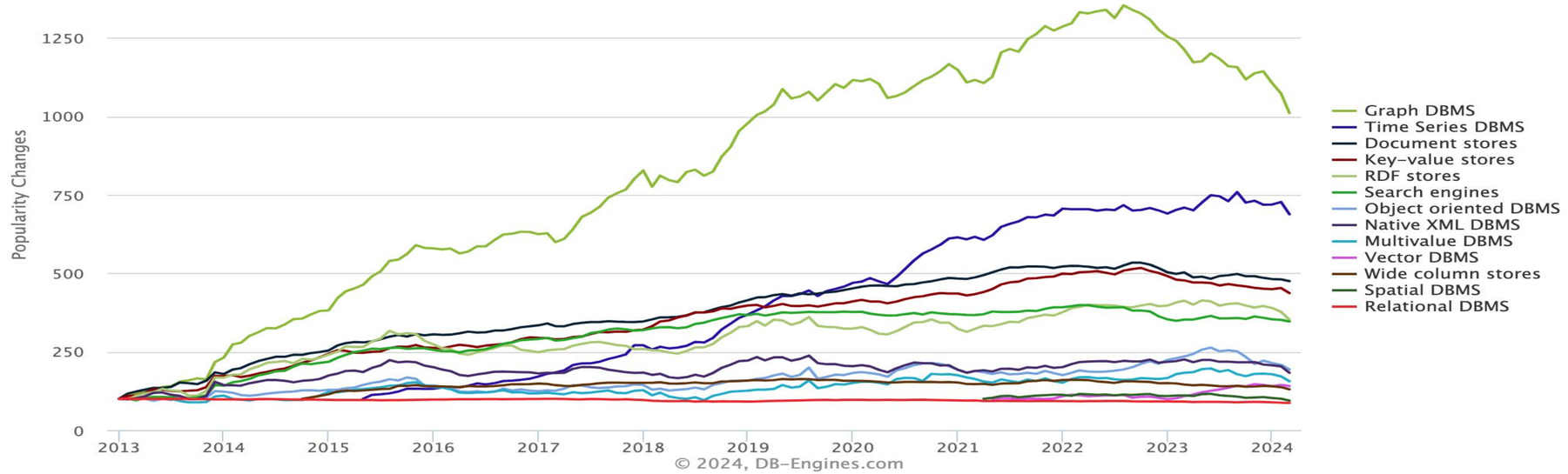
- 背景介绍
  - 大数据时代图数据库的优势及重要性
  - 图数据库在大数据领域的现状
- 风控系统中图数据库的功能及挑战
- 图数据库在金融风控场景中的应用
- 总结与展望

背景介绍

- 大数据的“4V”（Volume，Velocity，Variety，Value）
- 与关系型数据库相比，图数据库在分析深度的关联关系的场景中有明显的优势。
- 随着大数据技术的广泛引用，图数据库的发展也迎来了井喷期。
- 图是描述现实世界最优的模型。

2013年开始至今，图数据库的流行趋势变化是所有垂直类数据库中最高的，且看趋势至少未来3-5年也依然会保持最高。





风控系统中图数据库的功能及挑战

风控系统的特征

- 数据规模大、数据耦合
- 数据多样性，结构不规则
- 数据增长快，要求实时性
- 关联关系中隐含的价值大

图数据库必备的功能

- HTAP，支持实时图计算
- 图数据存储及优化
- 查询速度快
- 分布式高可用方案
- 时序图

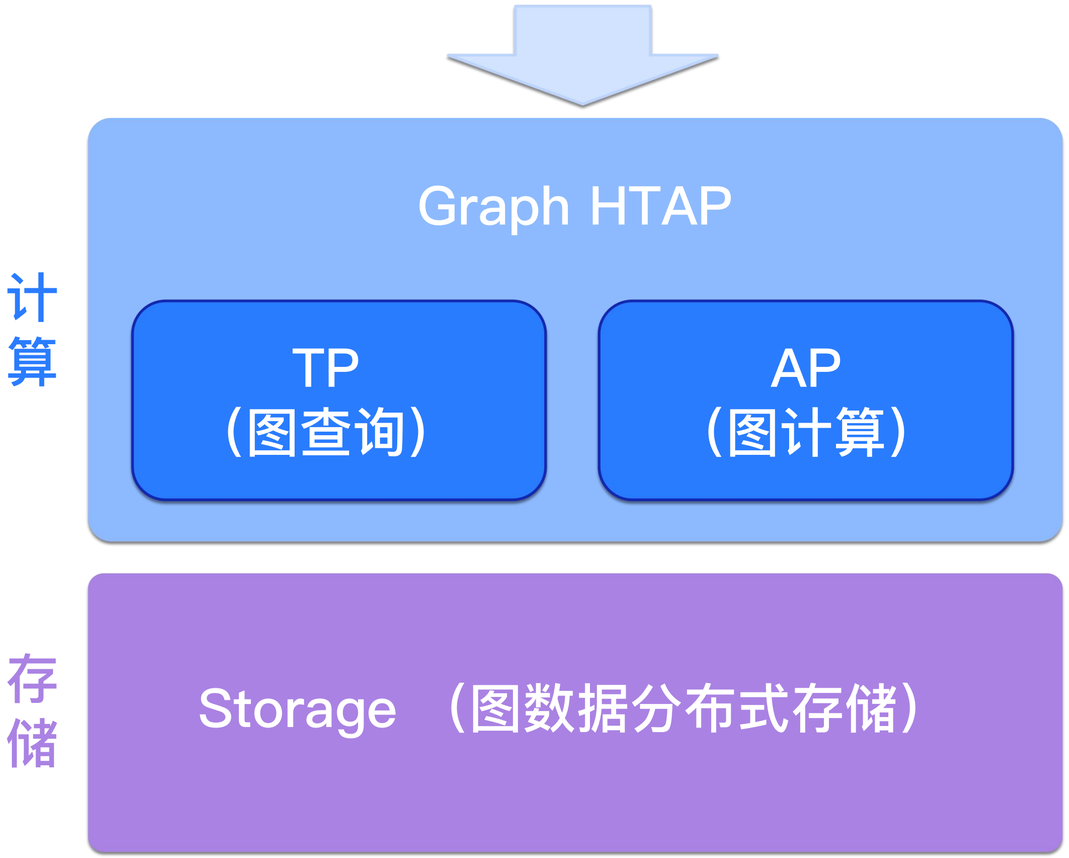
图数据库面临的挑战

- 实时图查询和图计算（HTAP）
- 多模态存储，空间放大
- 查询性能

图数据库的巨大机遇

- 云计算/云存储计算的快速发展，基础设施完善，使用图数据库变得更容易，成本也大大减少
- 随着图数据的价值不断被认可，越来越多的应用场景选择用图模型来表达业务数据
- 信创和国产替代的大势下，图数据库的研发也进入井喷期

HTAP



部署上，All-in-one

单机分布式一体化，一个数据库多种运行模式，方便PoC、测试

资源上，存算分离

存储和计算分离，松耦合，可独立扩展

使用上，Graph HTAP

AP和TP拥有独立计算资源，但共享一份数据，且走TP可直接调用AP算法，使用简单

HTAP -- 具备基本TP能力

- 常见的系统中，图数据库/图计算引擎侧重于OLAP，更偏向于线下、批处理、非实时模式的数据分析；而实时的业务系统通常会根据架构选择其他的数据库来实现。重要一个很重要的问题就出现了：打通两个数据库之间的数据传输（CDC）及传输过程中的数据管理（安全性、正确性等）
- 给图数据库增加OLTP的能力，实时地更新图数据，保持数据的一致性，并提供实时的图计算能力。
- 图数据库实现OLTP的几个重要部分：

  ACID, MVCC, 多模态, 存算分离, 分布式

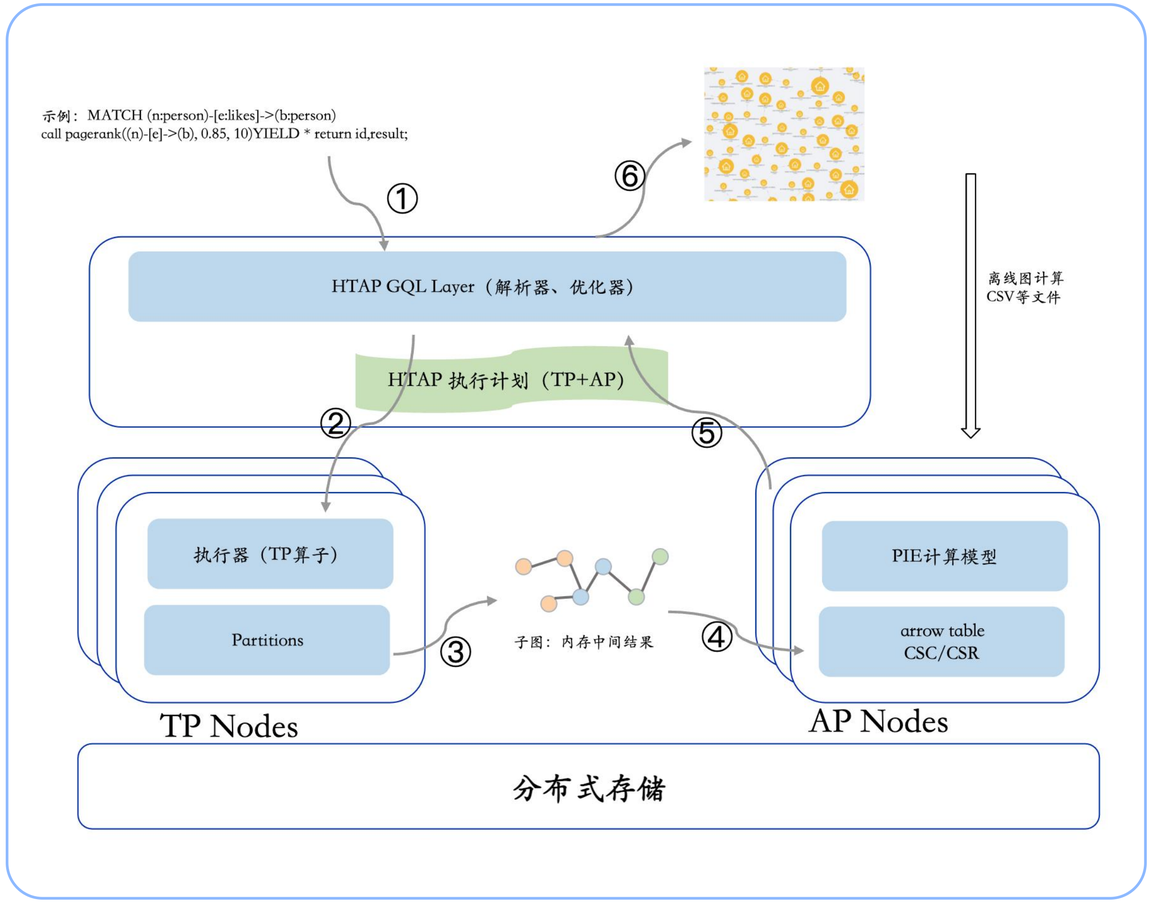
- 增加架构的灵活性和可扩展性

  即可以self-serviced，也可以跟其他数据库/计算引擎组合

HTAP -- 图算法集成（1）

- 大数据技术中的一些算法，在图中可以更好更快的支持，例如，中心性算法，社群检测算法，寻路算法等。
- 在图数据库中支持图算法
  - 可以共享图数据库的图数据和索引，减少资源开销和数据传输成本
  - 统一的查询接口
  - 架构上TP引擎和AP引擎更紧密的集成

HTAP -- 图算法集成（2）



- 一个类Cypher语句，完成子图查询和图算法调用（比如：对最近一个月的交易进行环路算法查找）
- TP主要应对1-3跳高并发
- 可使用AP，加速多跳查询的性能
- AP实时从TP中读取最新数据，避免数据延迟
- 技术：
  - HTAP 统一执行计划
  - TP&AP 共享存储
  - TP^&AP 独立扩展

图数据存储及优化（1）

数据的存储与图的查询紧密相关

图的集中存储方式

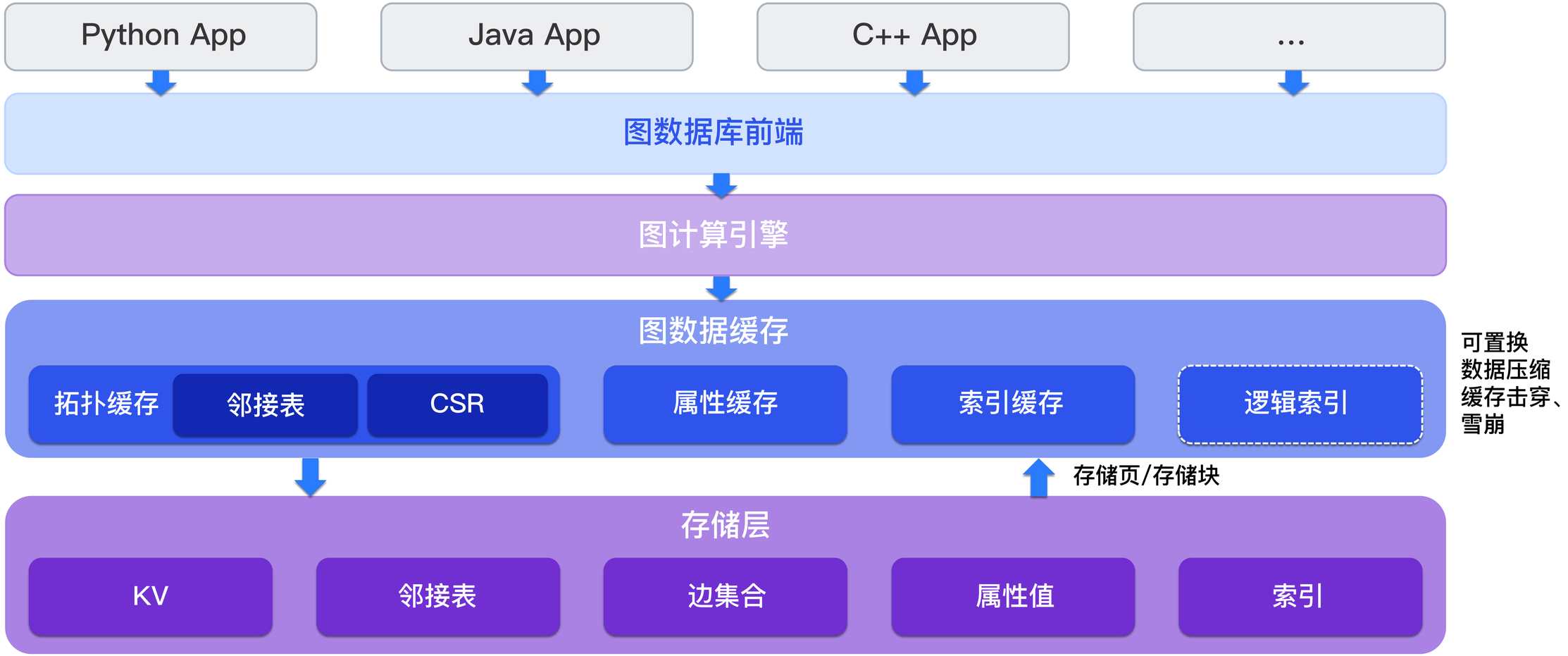
- KV，邻接表、邻接矩阵

多模态

- 时序边、JSON、向量、文档等

内存缓存

图数据存储及优化（2）


Python App
Java App
C++ App
...
图数据库前端
图计算引擎
图数据缓存
拓扑缓存
邻接表
CSR
属性缓存
索引缓存
逻辑索引
可置换
数据压缩
缓存击穿、
雪崩
存储页/存储块
存储层
KV
邻接表
边集合
属性值
索引

查询优化

一般查询计划的优化：
逻辑重写，CBO，HBO

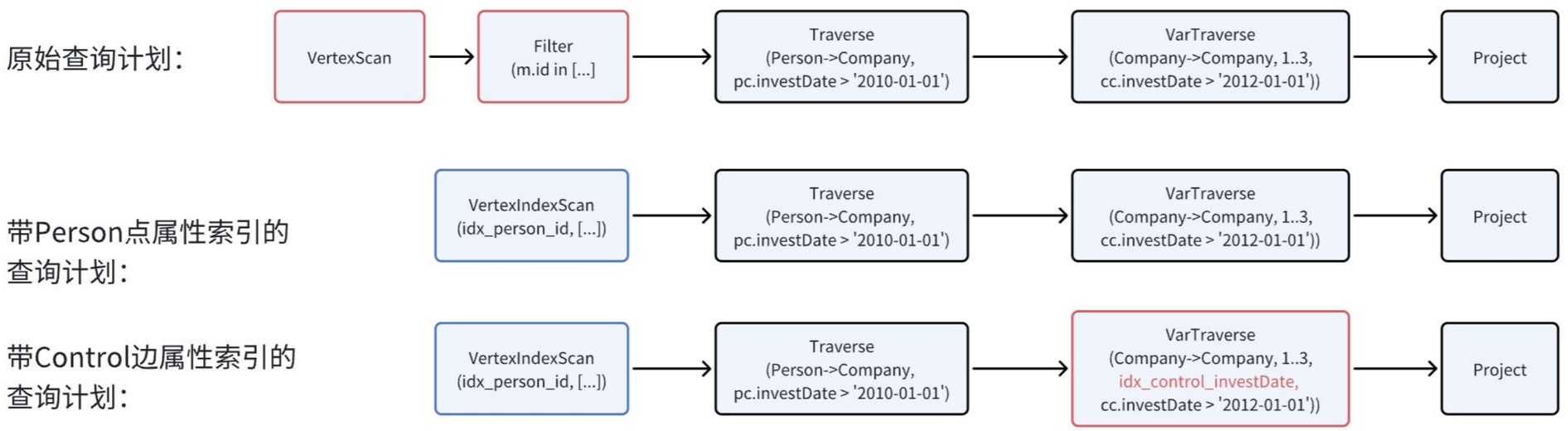
物理优化（空间换时间）：
属性索引，物理视图等

图特有的优化：
矩阵、hop 索引等

其他：
调整数据模型（图结构）

示例：扩层中属性索引的使用

```
MATCH (m:Person)-[pc:Invest]->(c:Company)-[cc:Control * 1..3]-> (c2:Company)
WHERE m.id in [ ‘p300129’, ‘p300130’, ‘p300140’ ]
                AND pc.investDate > ‘2012-01-01’
            AND cc.investDate > ‘2012-01-01’
RETURN m, c, c2;
```



图： ~3000万点，
~1.7亿边，
返回35万条结果

没有边的索引： ~235s

带边的索引： ~160s

挑战1 -- K跳可达

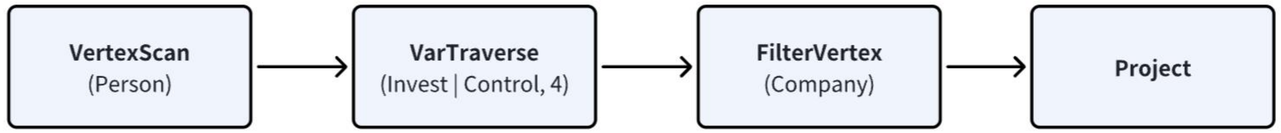
问题	挑战	优化方法
K 跳可达问题	• 查询速度慢，中间结果占用大量内存，重复路径 • 邻接表和矩阵，能快速地找到1跳邻居问题，对于k跳邻居，只能逐层查询，在稠密图上通常效率不高。	• hop索引： 2-hop，3-hop，path-hop等 • 带过滤条件的查询优化

示例： hop-index

K跳查询示例:

```
MATCH (p:Person {id: 'P3000213')-[inv:Invest|Control * 4]->[c:Company {id:
'C1000321' ]
WHERE inv.
RETURN p, c, LENGTH(inv) AS link_length;
```

普通查询计划:



使用2跳索引的执行计划:

挑战2 -- 超级节点问题

问题

超级节点

挑战

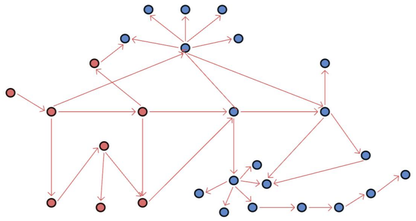
- 查询中遇到超级节点，内容占用大，速度慢，查不出来

优化方法

一般的查询计划：

从起点出发，逐步扩层，再过滤出点

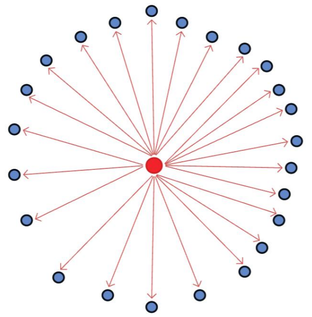
或者分别从起点和终点出发，做join操作



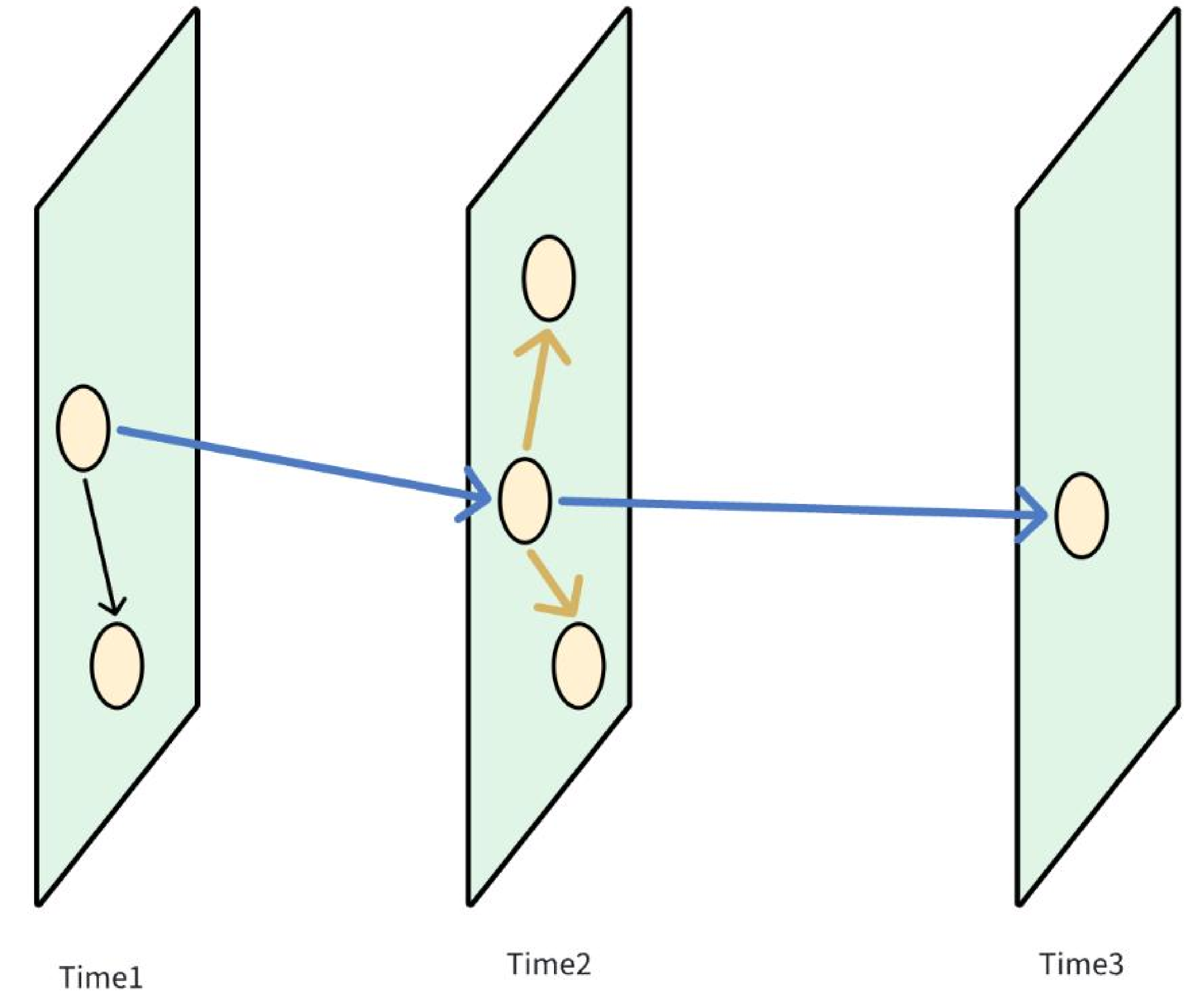
超级顶点的处理：

服务端：设置阀值，只访问阀值内的边；随机算法；优化存储结构

应用端：**调整数据模型（重构）**，拆分顶点类型，拆分边类型，合并边数据

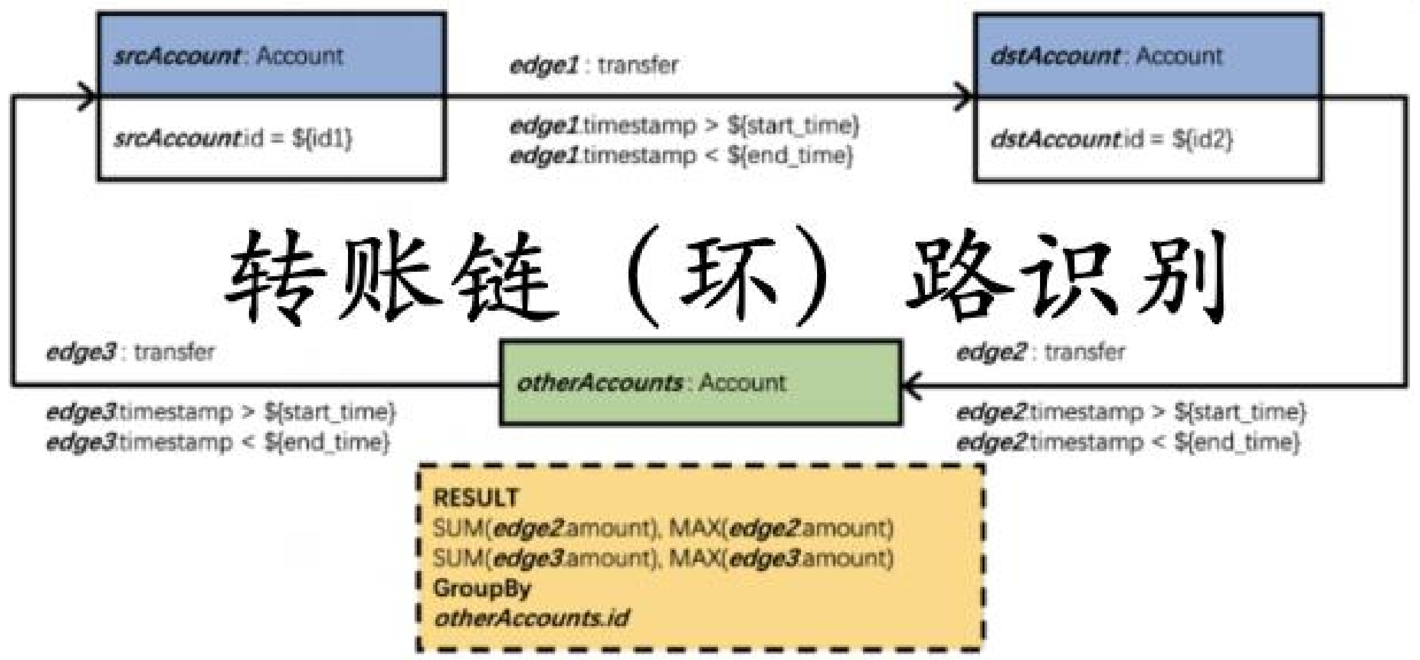


特色能力：时态图



按时间发生的顺序在图中查找

Time2 时刻的关系链



- 查询最新转账信息（时态数据）：

```
>> match (m)-[e: transfer]->(n) return e;
start executing statement: match (m)-[e: transfer]->(n) return e;
+------------------------------------------------------------------------------+
| e                                                                            |
+==============================================================================+
| eid: [src: [_oid: 108976354583642112, label: person], dest: [_oid: 108976354583642113, label: person], label: transfer, _rank: 1678838400000], properties: [2000] |
+------------------------------------------------------------------------------+
total rows 1 spends: 26ms
SystemTime { tv_sec: 1709708084, tv_nsec: 686805000 }
```

- 查询历史某时间点转账信息（时态数据）：

```
>> match (m)-[e: transfer]->(n) return e temporal 1673740800000;
start executing statement: match (m)-[e: transfer]->(n) return e temporal 1673740800000;
+------------------------------------------------------------------------------+
| e                                                                            |
+==============================================================================+
| eid: [src: [_oid: 108976354583642112, label: person], dest: [_oid: 108976354583642113, label: person], label: transfer, _rank: 1673740800000], properties: [1000] |
+------------------------------------------------------------------------------+
```

- 查询历史某时间范围转账信息（时态数据）：

```
>> match (m)-[e: transfer]->(n) return e temporal (0, now());
start executing statement: match (m)-[e: transfer]->(n) return e temporal (0, now());
+------------------------------------------------------------------------------+
| e                                                                            |
+==============================================================================+
| eid: [src: [_oid: 108976354583642112, label: person], dest: [_oid: 108976354583642113, label: person], label: transfer, _rank: 1678838400000], properties: [2000] |
+------------------------------------------------------------------------------+
| eid: [src: [_oid: 108976354583642112, label: person], dest: [_oid: 108976354583642113, label: person], label: transfer, _rank: 1673740800000], properties: [1000] |
+------------------------------------------------------------------------------+
```

图数据库的典型金融应用场景

反欺诈场景

- 资金环路、回流检测
- 欺诈交易发现
- 信贷申请欺诈团伙发现
- 对公信贷申请欺诈发现

反洗钱场景

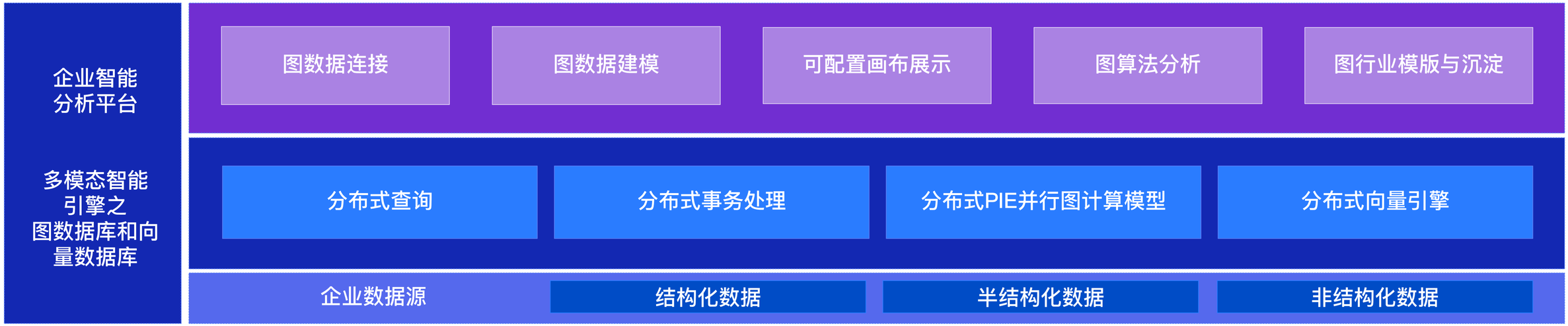
- 构建面向反洗钱的图谱
- 分析建模发掘可疑个体
- 关联数据进行团伙发现
- 模型固化后自动化预警

智能风控场景

- 贷款流向异常分析
- 贷款还款异常分析
- 企业状态异常分析
- 企业违约风险分析

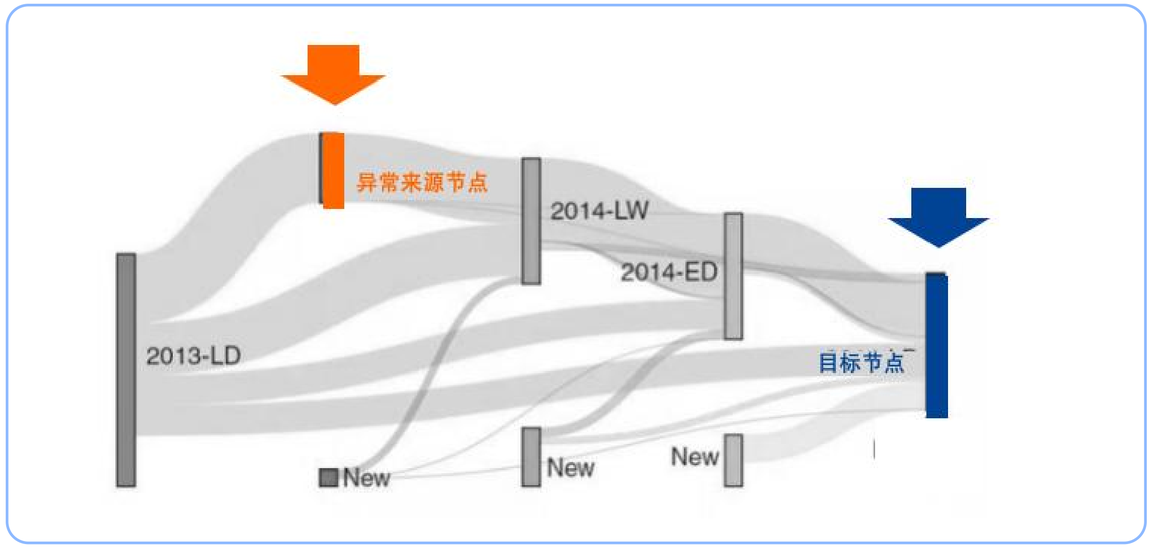
智能营销场景

- 私人银行潜在高净值客户挖掘
- 私人银行客户生态圈分析
- 对公客户生态圈分析
- 信用卡高价值成长路径分析



深度数据的链路分析：

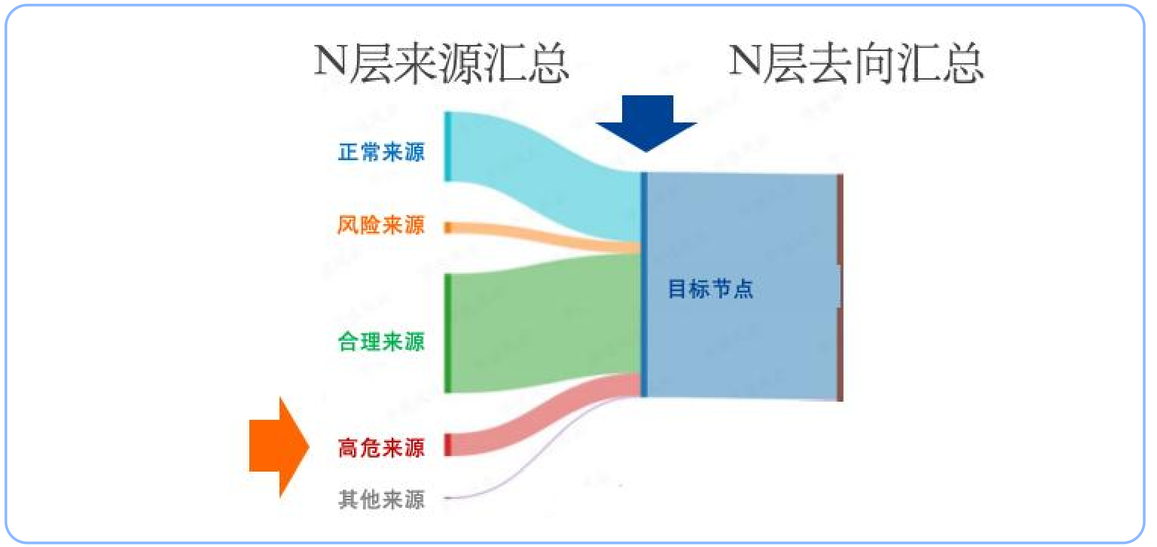
资金来源去向的深度分析



交易明细链路的异常发现

资金来源、去向穿透

链路中异常节点识别



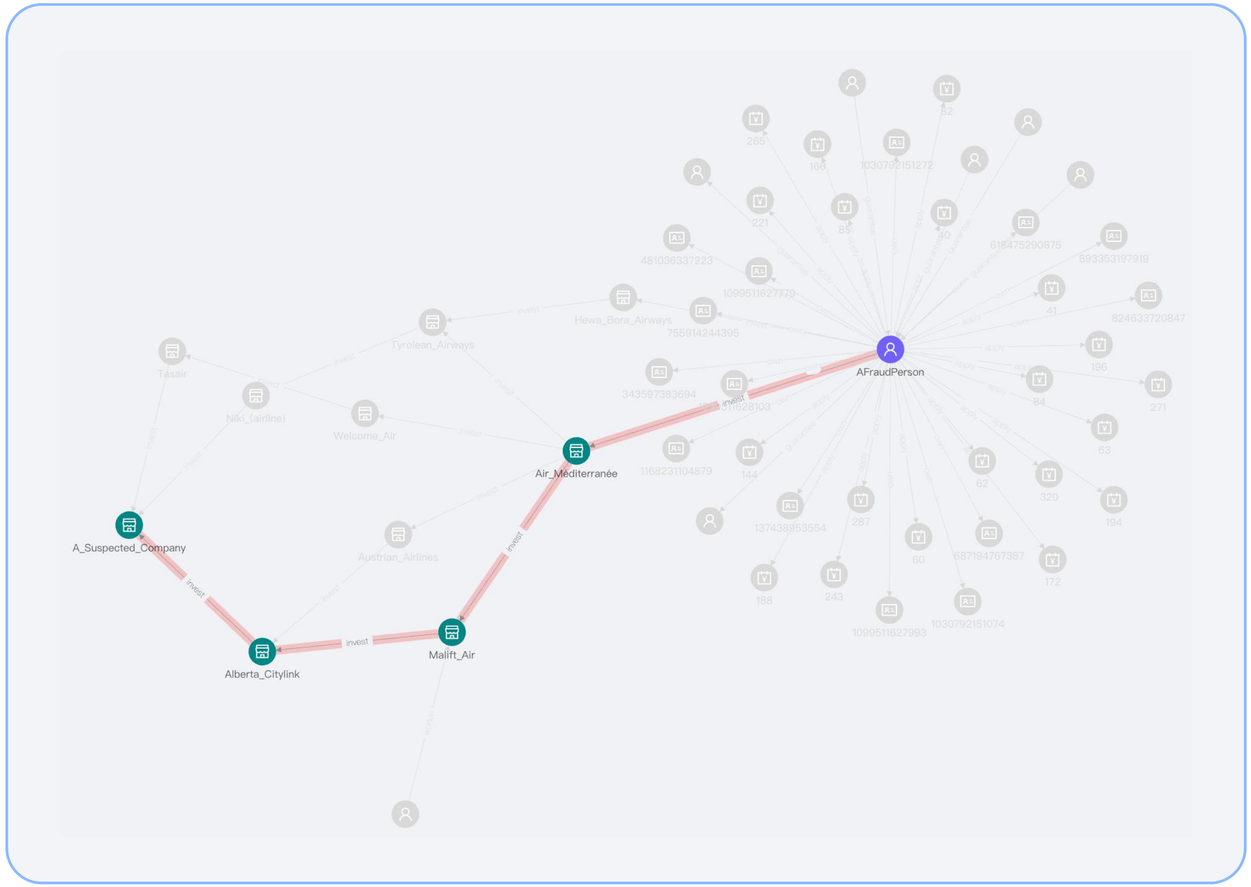
N层穿透汇总

来源、去向汇总分类

最终流向、来源定位

深度数据的链路分析：

基于深度交易数据的图计算

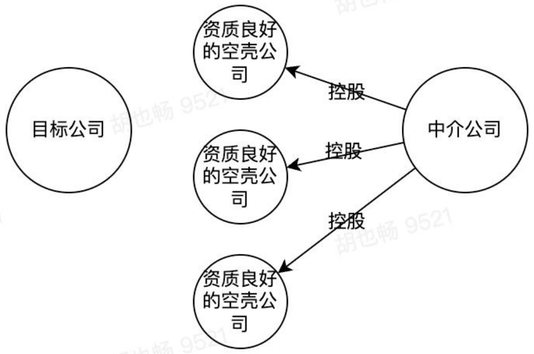


链路传导

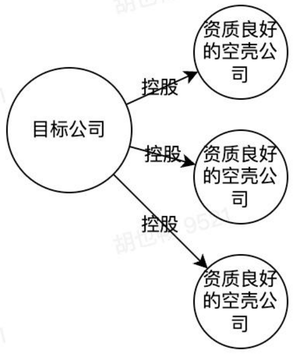
根据交易、投资、担保等关系从已定位的数据中发掘潜在的风险传导路径

深度数据的链路分析：

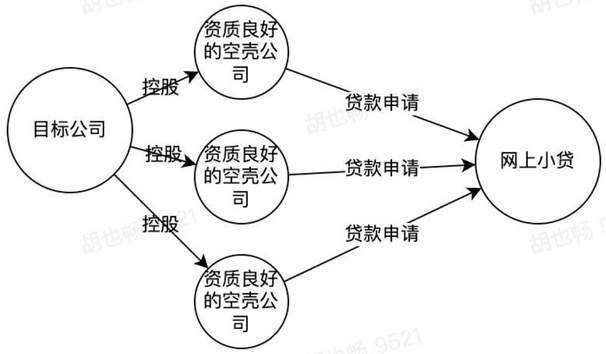
基于时序演进的图模式



t1
中介公司持有一系列空壳公司



t2
目标公司“购买”一系列的空壳公司



t3
目标公司控制空壳公司向银行申请在线小贷“金额相对较少”

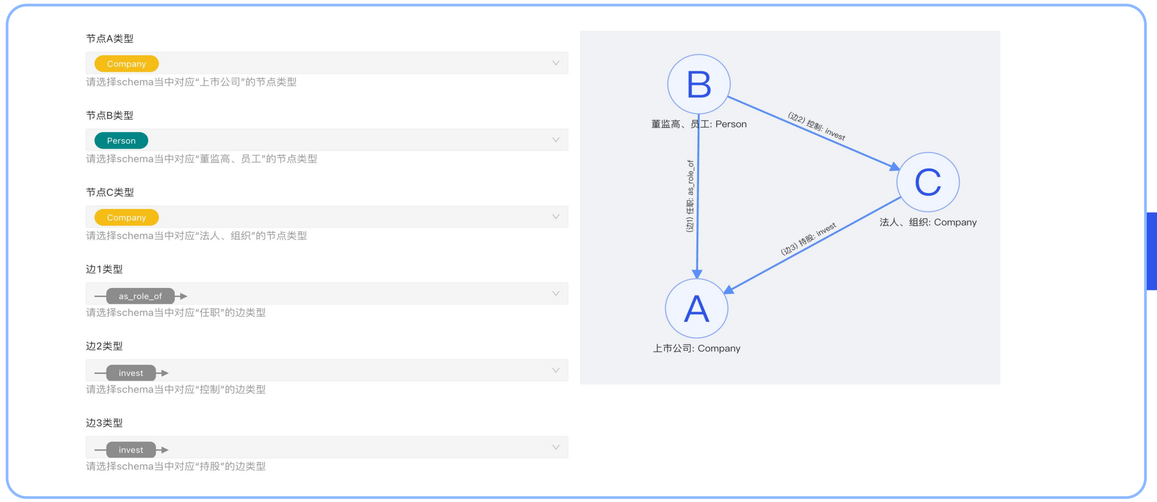
从时序发展中，定位动态的变化的图模式

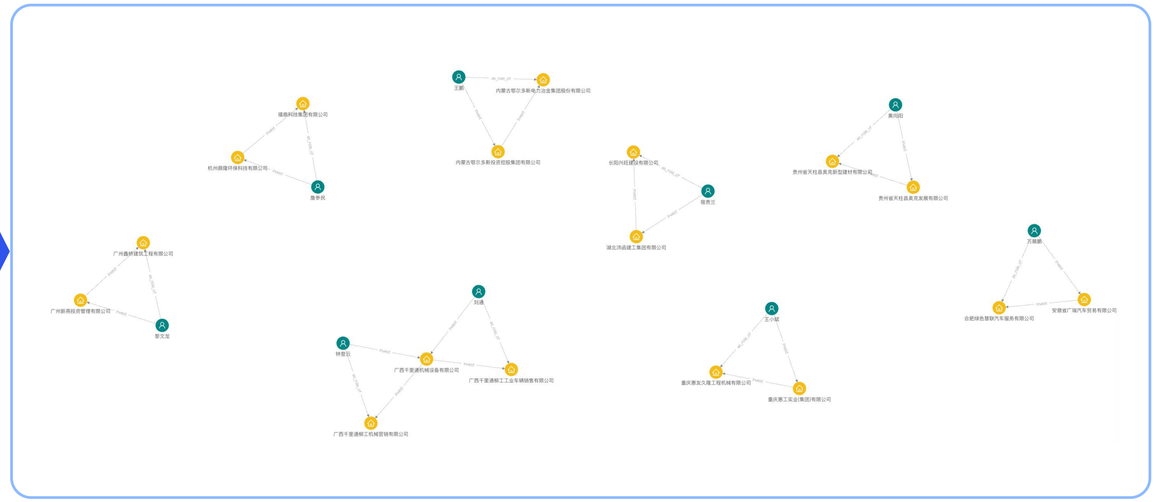
社团、模式演进

具备典型时间窗口动作行为的群体发现

深度数据的链路分析：

静态图模式发现





定义业务上典型的图模式

- 发现图数据中具备该模式的数据
- 环路交易
- 分散转入、集中转出
- 快进快出

使用图数据库的优势

	传统关系型数据库	图技术
数据加工效率	【慢】多个数据源的多次拼接操作 某行实践：14次SQL处理，6张临时表	【快】提供connector快速进行数据导入 将客户主体关联的担保、合同、交易数据入图后，通过图查询快速探索数据，响应业务需求 某行实践：从已有数据库中将交易流水数据直接导入
数据深度	只能进行【1跳】转账记录的处理查询 由于结构化数据库的限制，无法应对探索深度关联数据 某行实践：在处理涉及交易数据的还款来源异常探查中，仅处理一次交易链路的探查就需要～50分钟（包含数据预加工）	可以扩展至【N跳】转账记录的处理查询 图数据天然支持深度的关联数据探查 某行实践：在关联数据入图后，包含多次交易链路的探查可以做到分钟级返回
数据复用度	【低】需要逐个场景从数据源开始开发	【高】入图后基于图模型可直接用于后续进行分析，图模型易于扩展的特性可以便于积累更完整的业务模型，利于后续复用

总结与展望

图数据库实践的收获和经验

- 隐含关系挖掘场景下加速显著
- 分布式计算和实时性越来越重要
- HTAP

对未来图数据库在大数据领域的发展的看法和建议

- 统一的图数据库的查询语言标准（GQL）
- 图数据库与图AI技术的结合
- 在这个数据互联时代，图将是从海量数据中挖掘出隐式数据关联的最优模型，
- 图是“未来”。

智能时代的多模态图数据库的新趋势

- 更多异构数据源
- 数据入图成为必选项
- 大图时代

InfoQ 极客传媒

已结束 北京站 2000 人

QCon 全球软件开发大会 暨智能软件开发生态展

架构演进 | AIGC
研发效能 | 向量数据库

6 月 14–15 日 深圳站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

业务架构 | 高可用
AI 架构 | 架构师成长

10 月 18–19 日 上海站 1500 人

QCon 全球软件开发大会 暨智能软件开发生态展

AIGC | AI 原生应用
软件工程 | 金融科技

4月 5月 6月 8月 10月 12月

5 月 17–18 日 北京站 2000 人

AiCon 全球人工智能开发与应用大会 暨大模型应用生态展

通用大模型 | AI Agent
LLMOps | AI Infra

8 月 16–17 日 上海站 1000 人

FCon 全球金融科技大会 暨智能软件开发生态展

金融大模型 | 核心系统
数字化转型 | 数据安全

12 月 20–21 日 北京站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

未来架构 | 高可用高并发
AI 应用 | 架构师思维

极客邦科技 2024 年会议规划

促进软件开发及相关领域知识与创新的传播

访问大会官网

参会咨询

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software