基于大数据的金融知识图谱建设

李克伟

2017年10月

目录

背景

- 金融行业内部数据和业务系统的现状
  - 海量的历史数据和实时业务数据
  - 结构化和非结构化的数据
  - 数据碎片化，企业内部多个业务系统未完全打通

- 传统的手段无法满足智能风控,智能营销等场景下新的需求

- 金融行业相关的各种违规和异常行为呈现新的特征
  - 方式多元化
  - 手段隐蔽化
  - 行为网络化
  - 实时自动化

金融行业传统防控体系的约束和新的手段

现状	目标	新的手段
• 数据存储于关系型数据库以及数仓, 无法处理非结构化数据，难以适应海量数据的存储和复杂计算 • 业务系统之间没有完全打通，无法真正体现数据的价值 • 技术手段的约束难以发现数据间的“隐性关系”。用户也难以直观查看各种复杂的关联关系	• 建设高可扩展的基于分布式存储和分布式计算框架的平台 • 对各个业务系统进行充分的分析和处理，将散布于各系统间的数据打通并挖掘其价值,形成一张完整的行业知识图谱 • 使用新的技术手段发现数据间的“隐性关系”来帮助构建知识图谱，给用户提供可视化的浏览和操作	• 基于Hadoop的大数据平台作为底层存储和计算框架 • 基于图数据库来构建知识图谱，并提供可视化和交互式的图操作平台 • 使用新的技术手段如文本分析，机器学习，图挖掘等来挖掘数据间的“隐性关系”

为什么需要构建知识图谱？

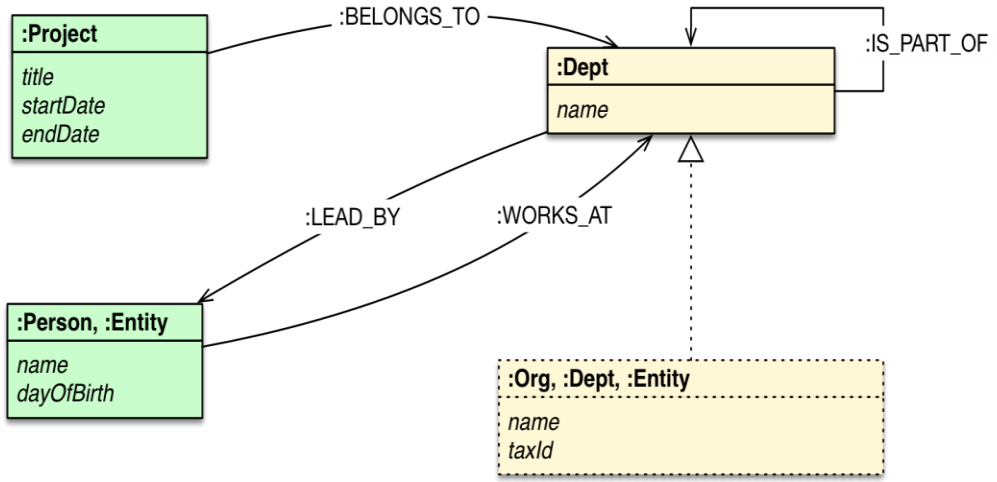
- 知识图谱是最接近真实世界的数据组织结构，通过知识图谱将企业所有的数据连接起来，让数据更加容易被人和机器理解和处理
- 知识图谱的数据组织结构符合人的思维模式，是人工智能的基础环境
- 知识图谱的组织结构能够灵活的应对企业的数据种类变化，新的数据种类可以快速融合并发挥作用
- 知识图谱不仅存储数据，也存储领域知识。包含通用领域和行业领域内的事物的全面属性、精细化分类、可能的概念名称和特定表达等，事物之间的基本规则、行业规则、领域规范等
- 知识图谱涉及的技术领域包括：知识表示、自然语言理解、智能问答、知识抽取、链接数据、图数据库、图挖掘、常识推理等

如何落地企业知识图谱

**交互式服务	批量式服务	可视化服务**	提供各类服务接口，支撑业务场景
数据挖掘与机器学习	业务模型，自然语言处理，全量数据机器学习
知识存储	知识图谱的物理载体，提供各类图相关服务
关联计算	基于规则的关联计算，建立实体之间的关系
数据治理与整合	结构化与非结构化数据智能治理与整合
统一数据视图	建立行业特定的知识图谱所需的数据模型

为什么使用图数据库

- 业务驱动：世界是由关系组成的。
- 技术驱动：关系型数据库处理不好关系，图数据库最适合处理关系。



创建图数据库的目的就是为了存储高度关联的数据，支持灵活多变的数据模型从而能够从关系数据里获取更多信息。

- 图数据库的每个节点（包含实体及属性）直接存储了该节点与其他关联节点的关系记录列表
- 当关系型数据库在做join操作时，图数据库只需要读取关系列表即可找到关联的节点，而不需要进行计算
- 图数据库支持非常灵活及细粒度的数据模型，模型管理更加方便。数据存储与真实世界保持一致，细粒度、范式化、丰富的关联实体，可以支持更多不同的分析场景

图数据库选型

- 是否可直接在**已有大数据平台**上部署
- 是否会引入新的依赖

与大数据平台Hadoop的整合度

功能集易用度

- 是否满足图数据库的常见需求
- 提供的查询接口是否友好
- 安装、部署与运维是否便利

- 在超大图(几亿点，几十亿边)上的查询和修改性能如何
- 能否满足秒级别的响应速度
- 批量操作的性能是否好

性能

可扩展性

- 能否线性扩展
- 服务能否支持多节点部署

- 源码是否清晰易懂
- 团队是否有在源码级别修改和调优的能力

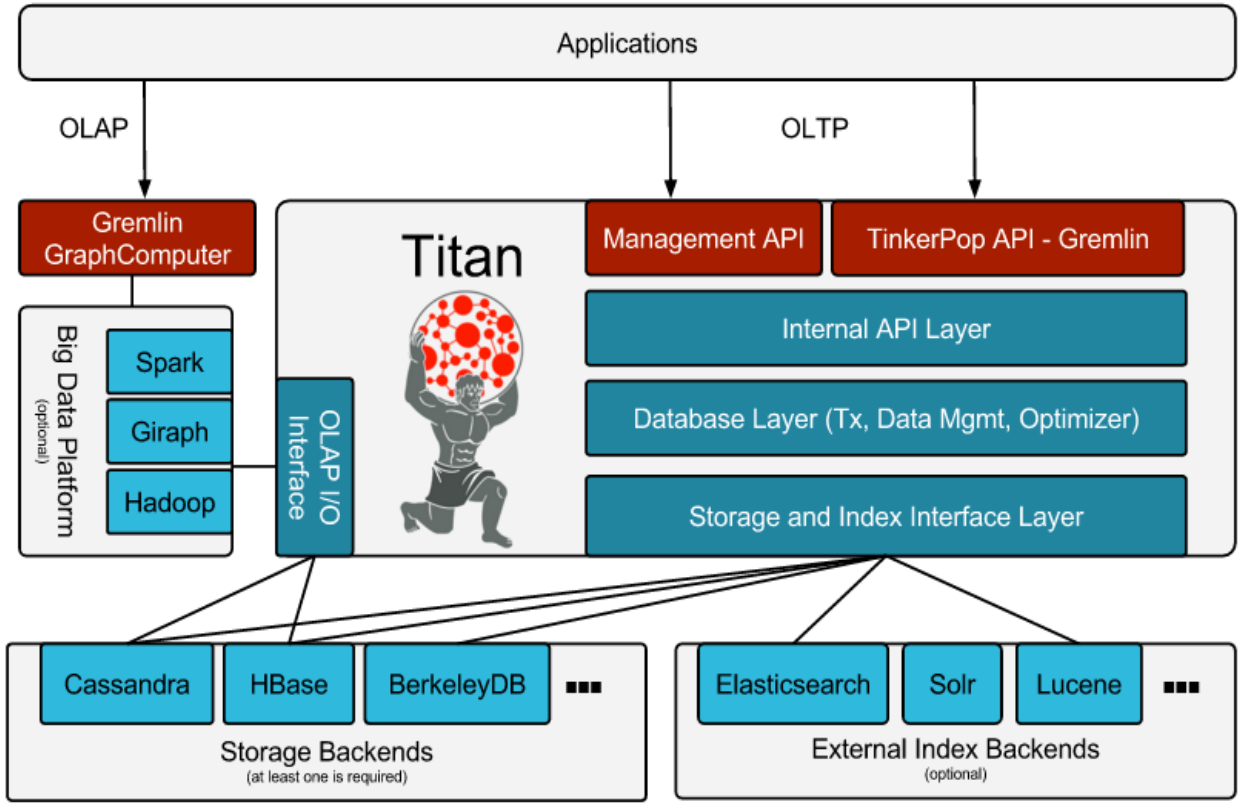
源码级别的掌控程度

License

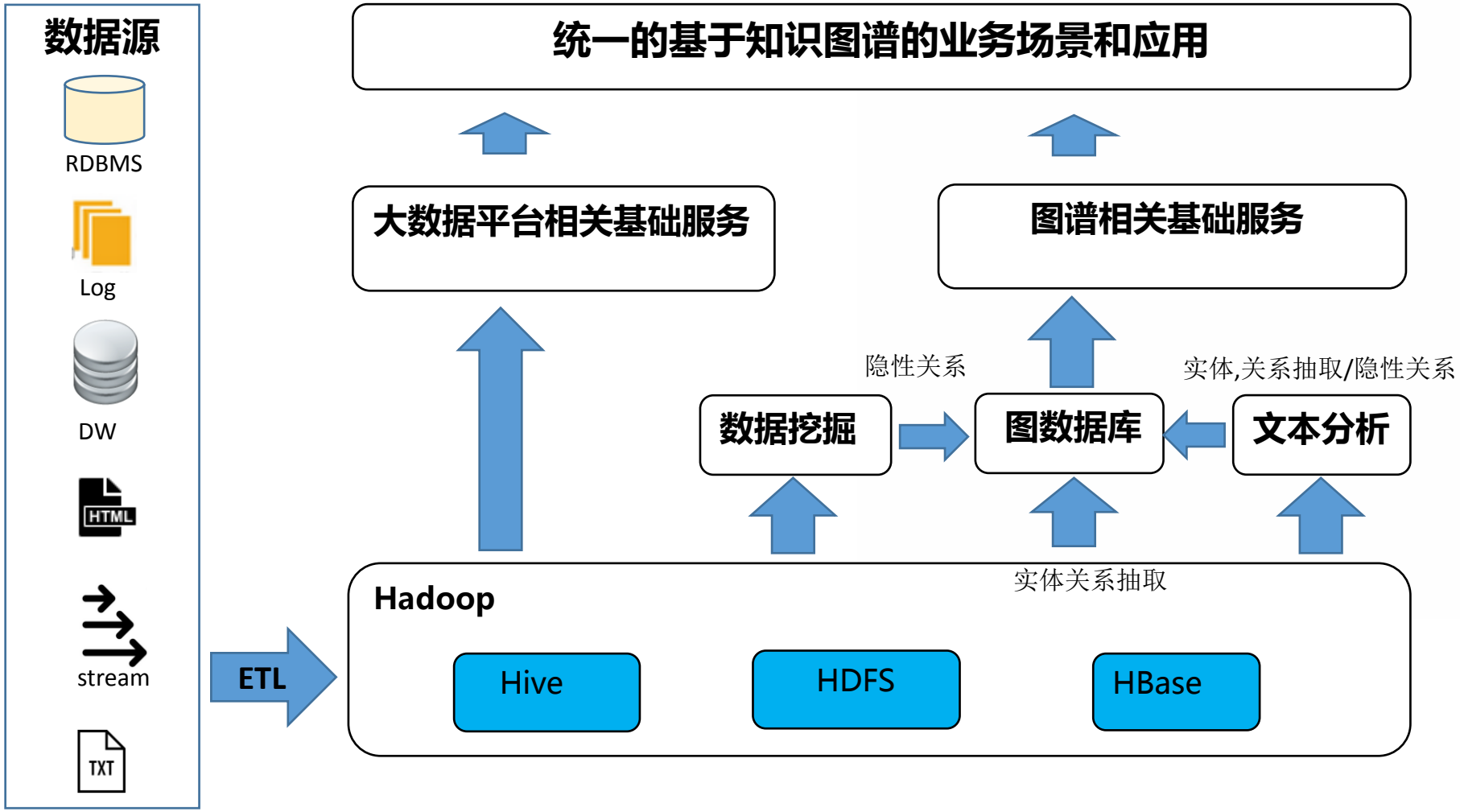
- License是否商业友好
- 最好是Apache License或者类似松散约束的

图数据库选型 – 选择Titan

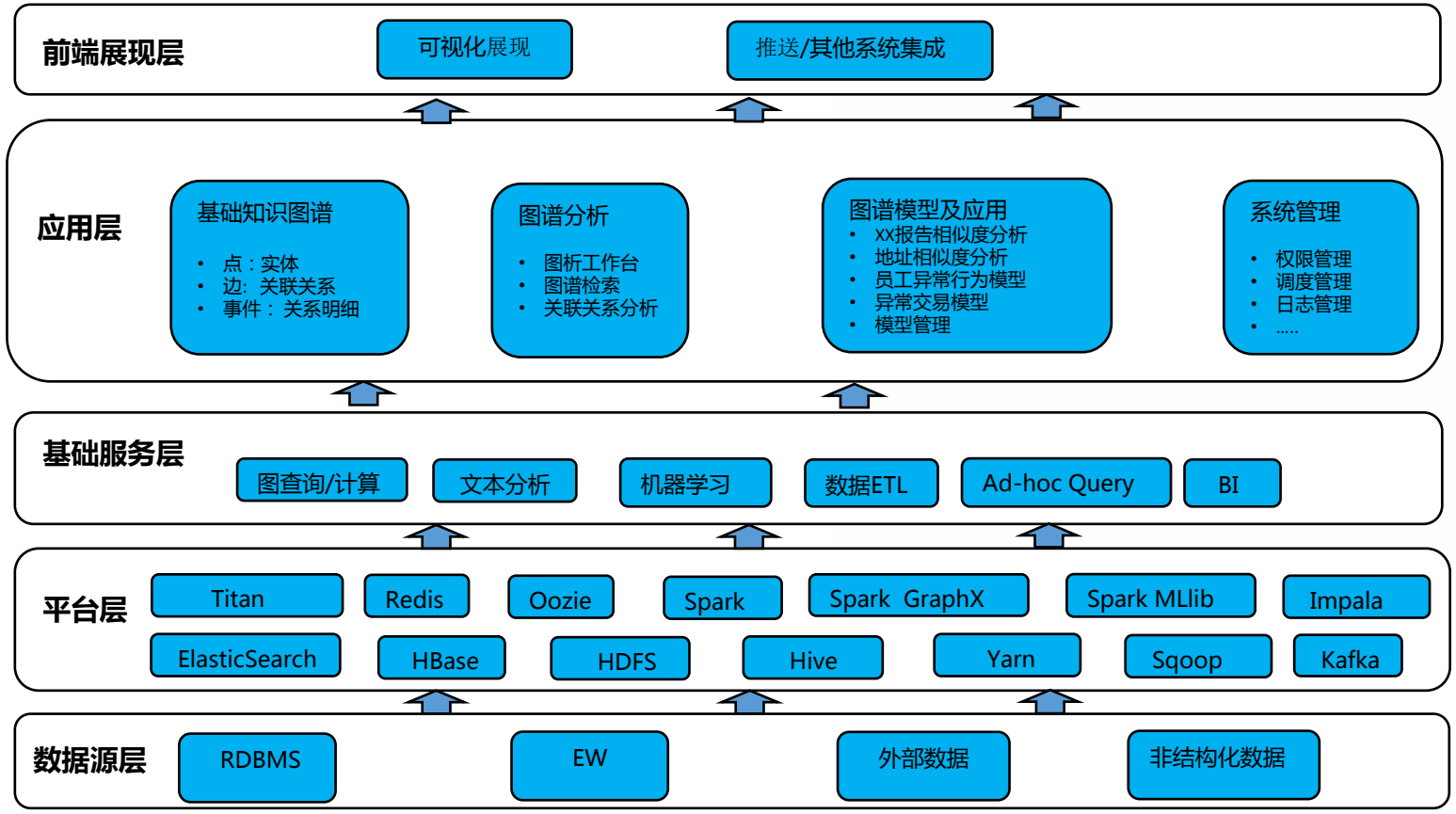
- **分布式架构：**Titan是基于Apache 2 Lience的分布式图数据库，与Hadoop生态圈中的开源组件Hbase、Spark等无缝兼容;
- **开源支持：**全面支持Apache Tinkerpop 图领域的开源软件栈，支持gremlin图查询语言，满足常用的图操作需求;
- **超大规模：**Titan能够存储的数据节点规模更大，支持存储10亿点100亿边及更大量级的数据（实证）；
- **快速响应：**支持丰富的检索规则，性能良好，在超大图上的响应能做到秒级;
- **高可用性：**分离存储层和索引层，基于成熟的大数据框架实现分布式存储和高可用性;
- **方便扩展：**方便横向扩展，可以通过增加计算节点和存储节点提升系统性能；



基于Hadoop和图数据库的通用融合架构构建知识图谱



基于Hadoop和图数据库的融合架构来构建知识图谱 - XX项目分层架构

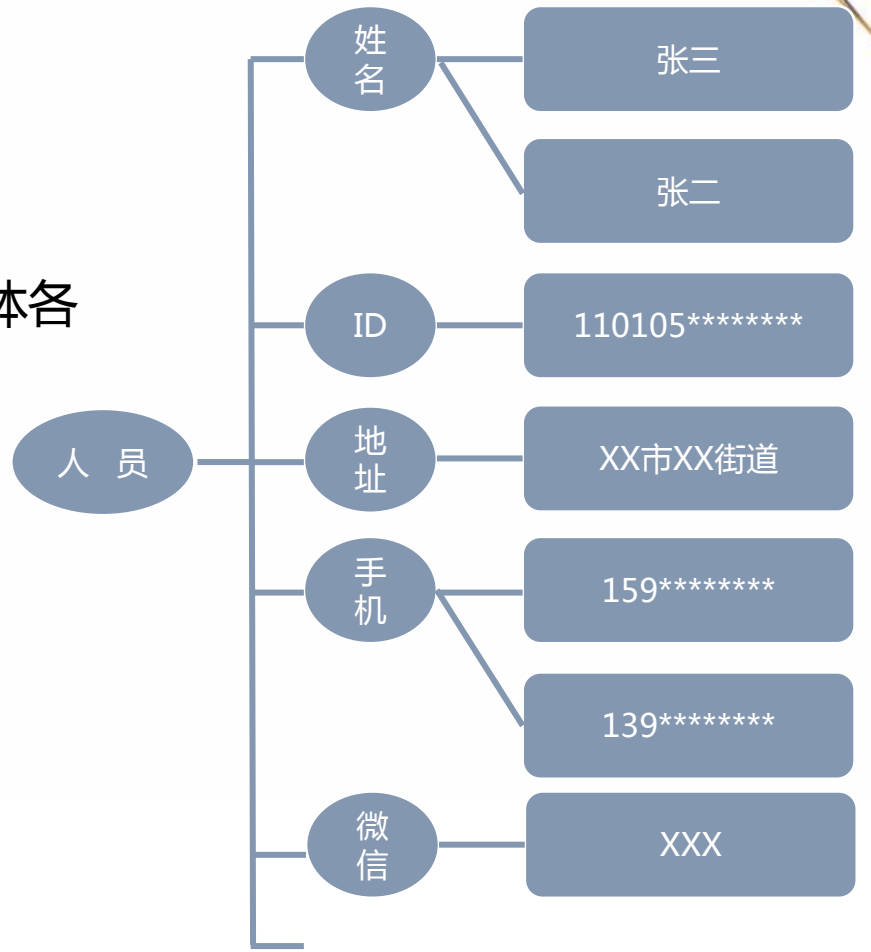


目录

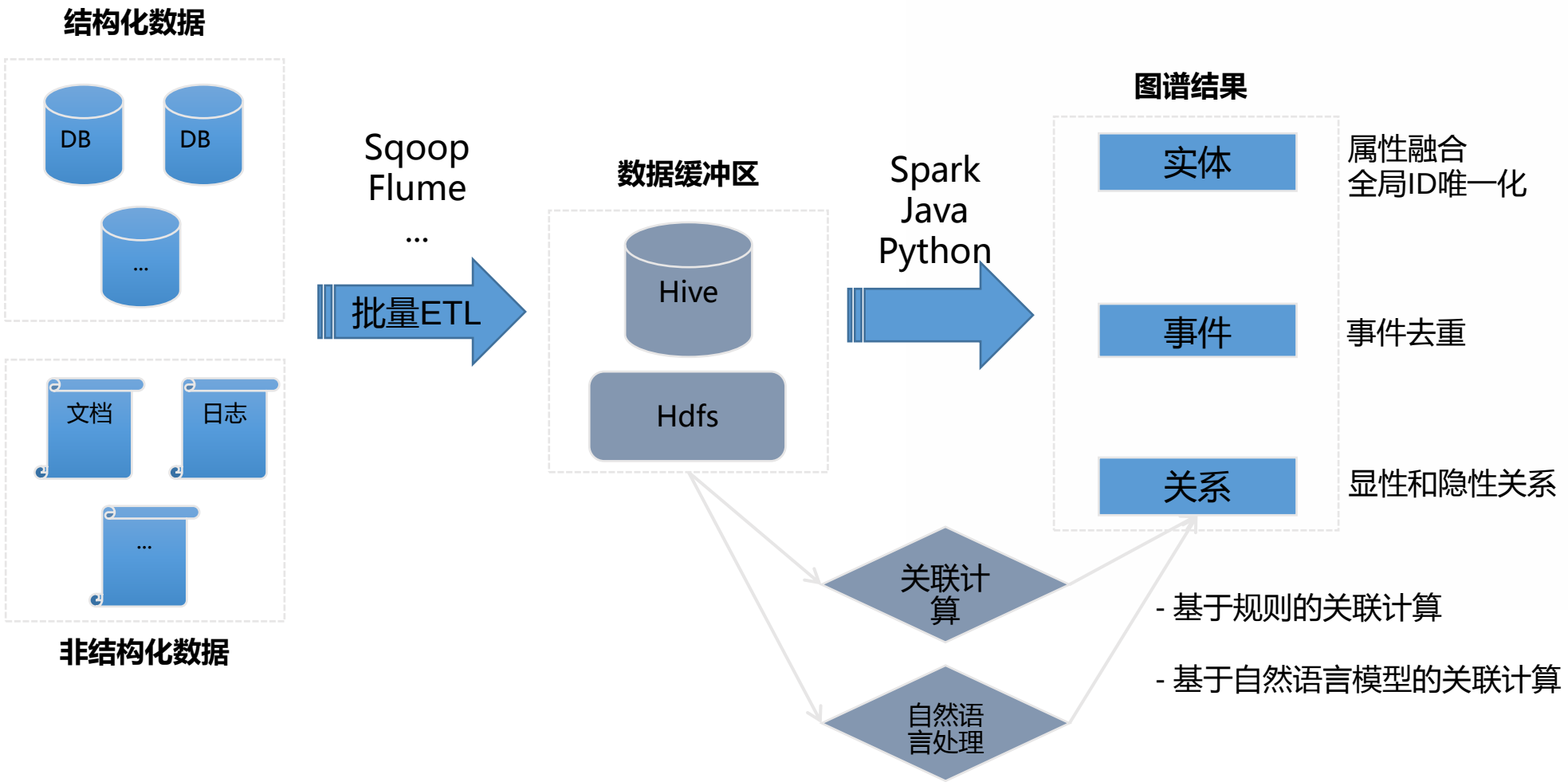
融合数据模型

实体	人 账户 手机 公司 ATM ……
事件	转账事件 抵押事件 担保事件 ……
关系	交易关系 担保关系 地址关系 ……

对于实体，需要融合实体各方面的信息

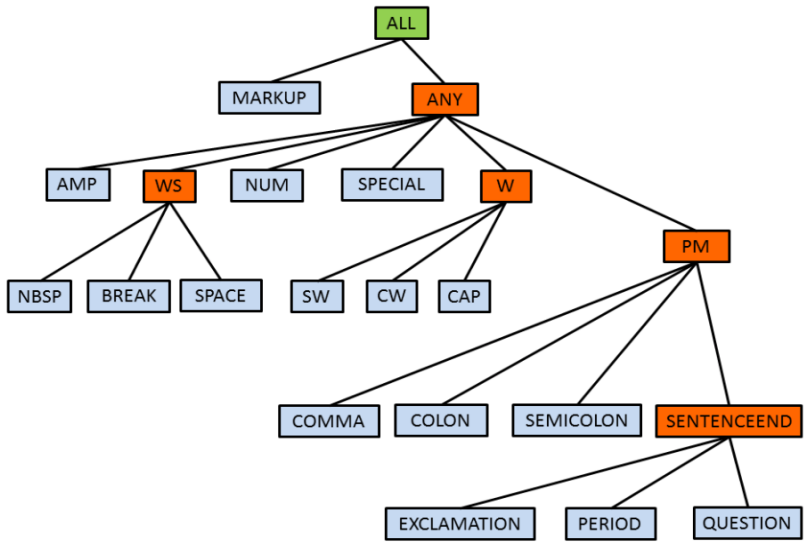


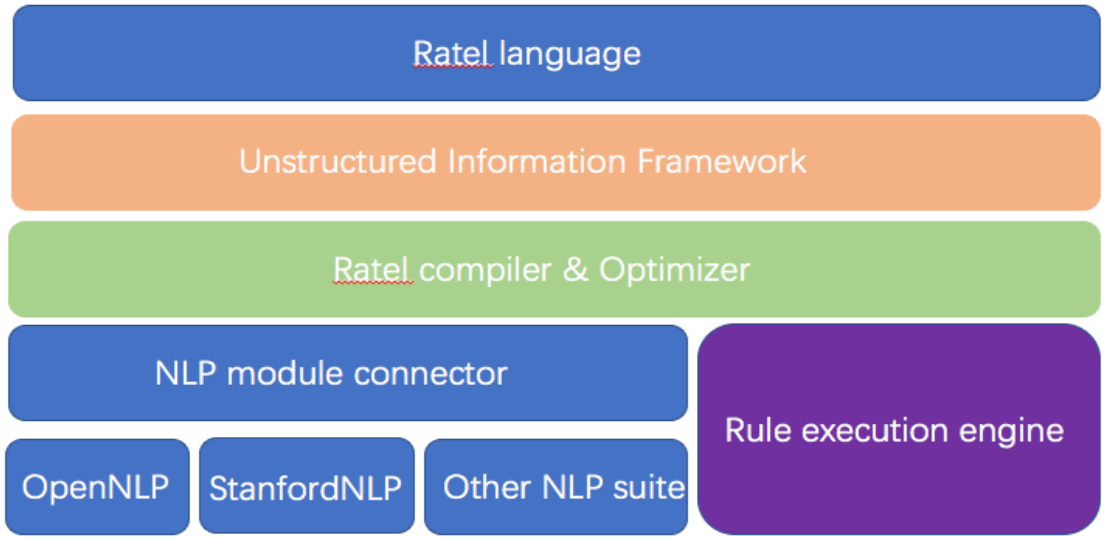
基于知识图谱的数据治理与整合



通用文本挖掘与非结构化关系构建

- 融合机器学习和规则实现的通用文本挖掘工具
- 可用于快速在大量非结构化文档中抽取关系





文档1 文档2 新文档 新文档

case1.txt 3KB

生意缺钱竟用朋友的车骗贷涉嫌诈骗被抓

时间：2016-01-1116:45:19

王某伙同李某、徐某，伪造相关证件，用朋友齐某的本田雅阁轿车向安徽中诚汽车服务公司骗取贷款14万余元，结果无钱还款，该公司将齐某的轿车扣押。王某等三人则因涉嫌诈骗被警方抓获归案。

轿车停楼下莫名失踪

2015年8月26日一大早，家住铜陵市柏庄的齐鹏（男，电话：15902342332）就向警方报案称自己持有的本田雅阁轿车被盗。刑警支队二大队迅速展开调查。齐先生报称头天晚上其回家后，将车子就停在自家楼下，第二天一大早准备开车去上班，发现车子不见了，于是赶紧报警。侦查员得知其轿车上安装了GPS后，立即根据GPS定位系统，确定了轿车的位置。于是侦查员和齐先生一道，根据GPS信号在中诚汽车服务公司的停车场找到了齐先生的轿车。该公司称该车是公司贷款购买，现车主没有按照约定还款，所以对该车进行扣押。听该公司的表述，齐先生一头雾水，此车明明是自己购买，怎么又成了是该公司贷款购买的了。侦查员让双方出具相关材料，进一步展开细致调查。经过调查发现，中诚汽车服务公司出具的相关资料和证件显示，该车车主为王国鹏（男，电话：15592389352，安徽省 铜陵市人）。王国鹏通过该公司办理的是零元购车贷款。齐先生一听王国鹏的名字马上想起来了，此人是自己的朋友，2015年7月份，他曾将车子借给王国鹏使用。通过警方的调查和相关证件的比对，案情很快水落石出。王国鹏伪造证件用齐先生的轿车，从该公司骗取了贷款。王国鹏有重大诈骗嫌疑。

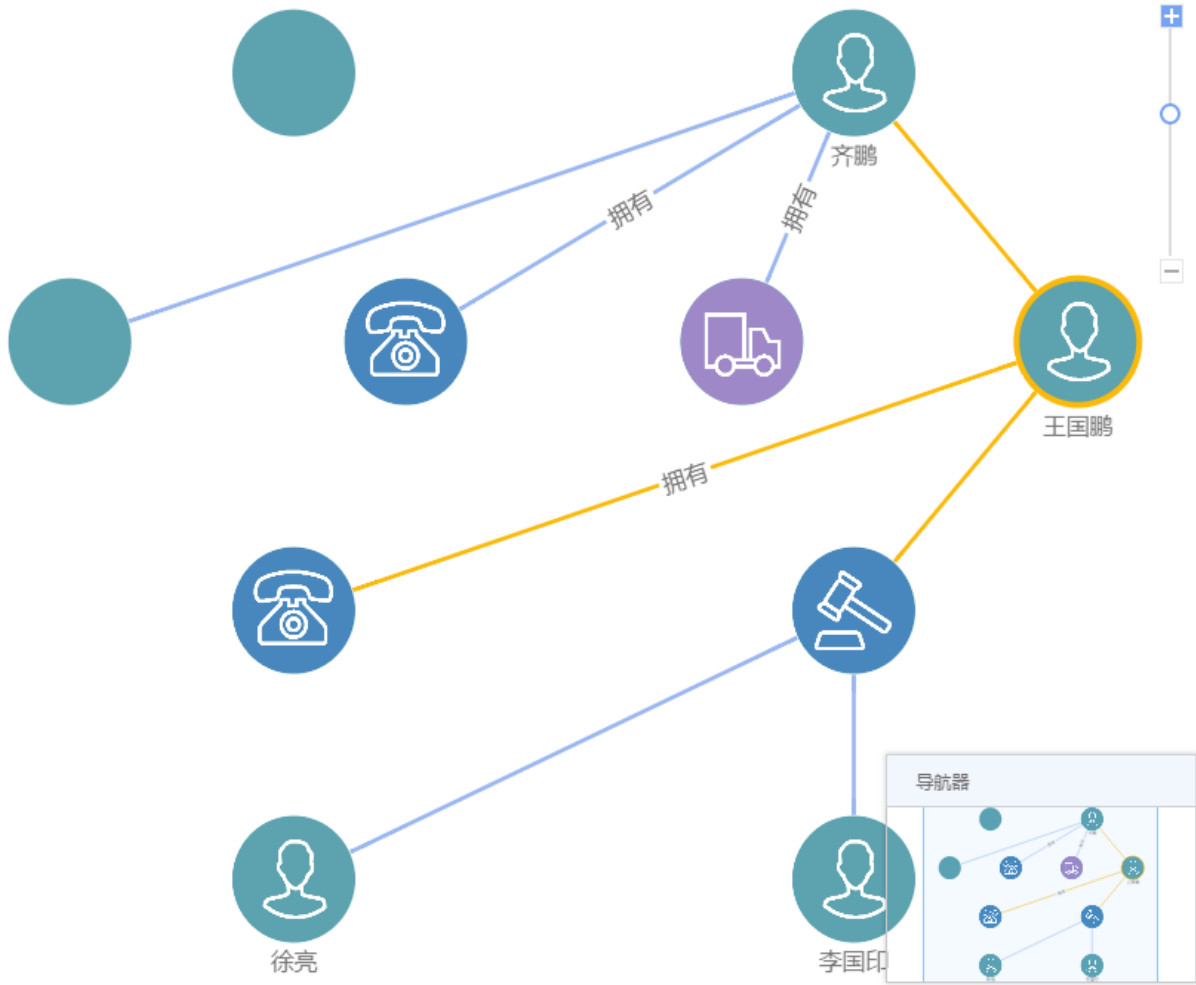
在民警的工作下，中诚汽车服务公司将轿车归还给齐先生。

生意缺钱用朋友车骗贷

警方很快将王国鹏抓获归案。经过审讯，王国鹏交代，因生意缺少资金，2015年6月份，犯罪嫌疑人王国鹏经过李国印介绍，在中诚汽车服务有限公司申请办理零元购车贷款业务，因为王国鹏的银行诚信度较低，无法办理。2015年7月17日，王国鹏伙同李国印、徐亮两人伪造了行驶证、购车发票等相关资料，将其从朋友齐鹏借来的本田雅阁轿车冒充是公司购买的新车骗取了中诚汽车服务公司的第一笔贷款7万元。当天该公司为该辆轿车安装了GPRS定位系统。2015年7月21日，王国鹏又从该公司骗取了第二笔贷款7万余元。共骗取该公司贷款14余万元。骗取的贷款被三人瓜分。贷款后第一个月，王国鹏就没有按约定还款。中诚汽车服务公司根据安装的GPS定位系统找到齐先生停在楼下的轿车，将车开走。案发后，侦查员第一时间搜集了大量的证据材料并对王国鹏采取强制措施，但是犯罪嫌疑人李国印消失在逃。二大队侦查员发扬连续作战的精神，加大对该两名在逃嫌疑人的抓捕力度，终于在大量的事实证据和压力之下，犯罪嫌疑人李国印迫于公安机关强大威慑于2015年12月8日投案自首。李国印的投案，让公安机关掌握了该公司业务经理徐亮的犯罪证据，2015年12月16日，二大队对犯罪嫌疑人徐亮进行依法传唤。证据面前，犯罪嫌疑人李国印、犯罪嫌疑人徐亮如实交代了诈骗事实。

目前，犯罪嫌疑人王国鹏已依法移送人民检察院审查起诉，犯罪嫌疑人李国印、嫌疑人徐亮被取保候审，案件正在进一步侦办中。

实体关系图



属性列表

姓名：王国鹏

性别：男

省市县：铜陵市

类型：common

统计

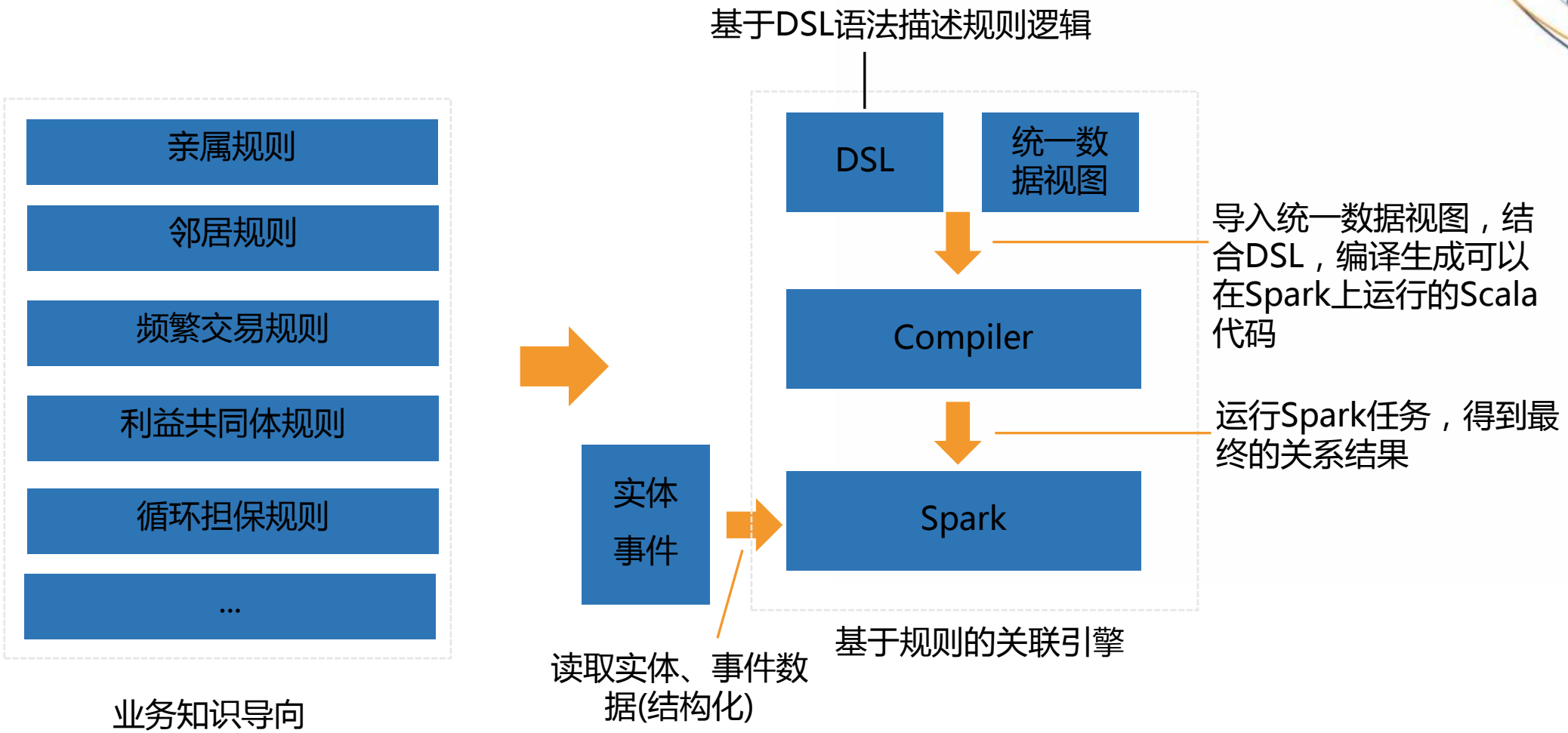
关系类别

拥有 2

拥有 1

事件类别

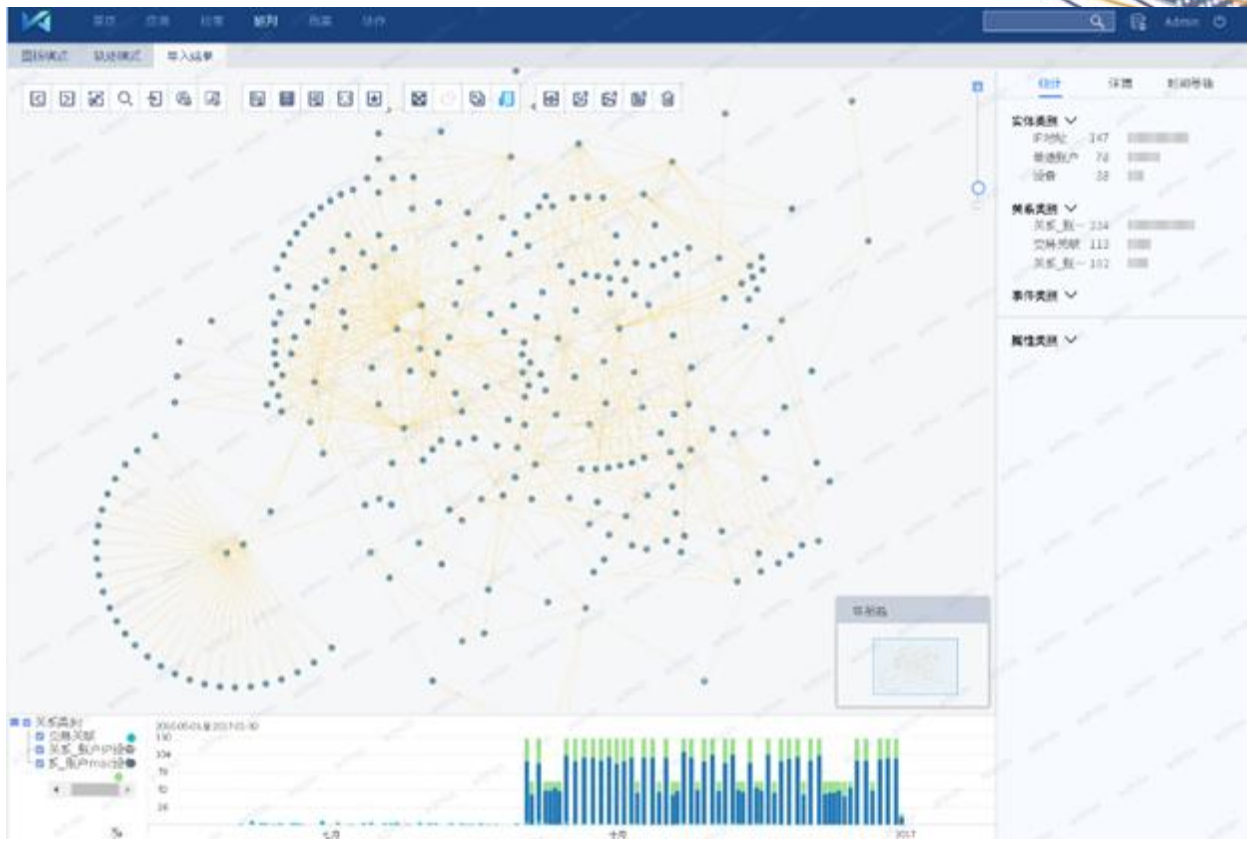
构建隐性关系中的核心技术



某大型证券交易所基于关系挖掘的异常行为发现

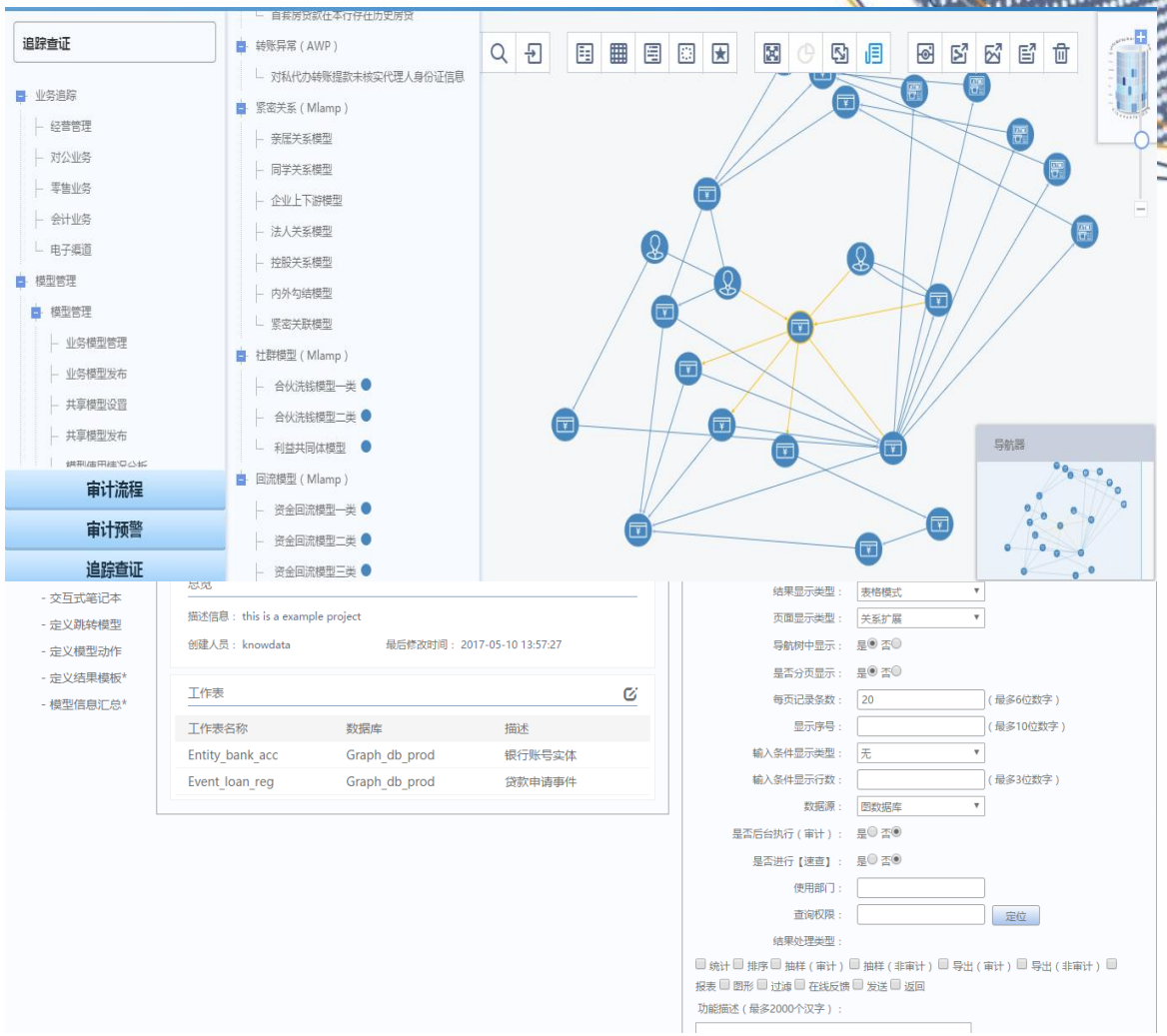
明略数据结合关系挖掘和机器学习发现账号之间的控制关系和异常行为，主要技术手段包括：

- 紧密连通子图
  - 相互有密切交易往来而与其他账户无交易往来
  - 重视提升内聚紧密度
- 标签传播
  - 与同类节点的往来较其他节点更密切
  - 重视提升模块性（Modularity ）
- 谱聚类
  - 对相似矩阵做特征分解
  - 对特征向量进行聚类
- 置信传播
  - 认为邻居节点的角色更符合资金往来模式



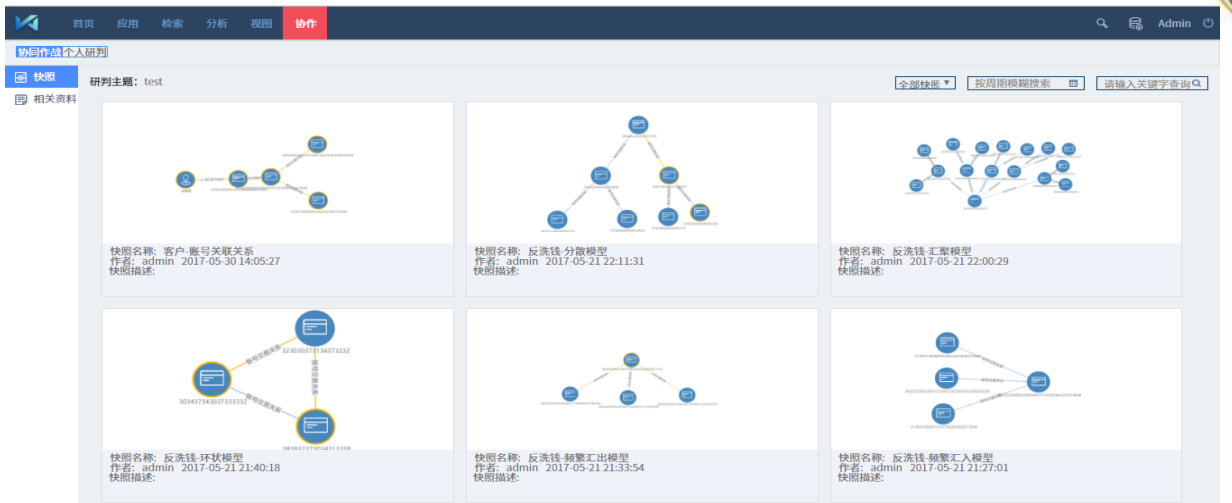
某股份制商业银行内审内控项目

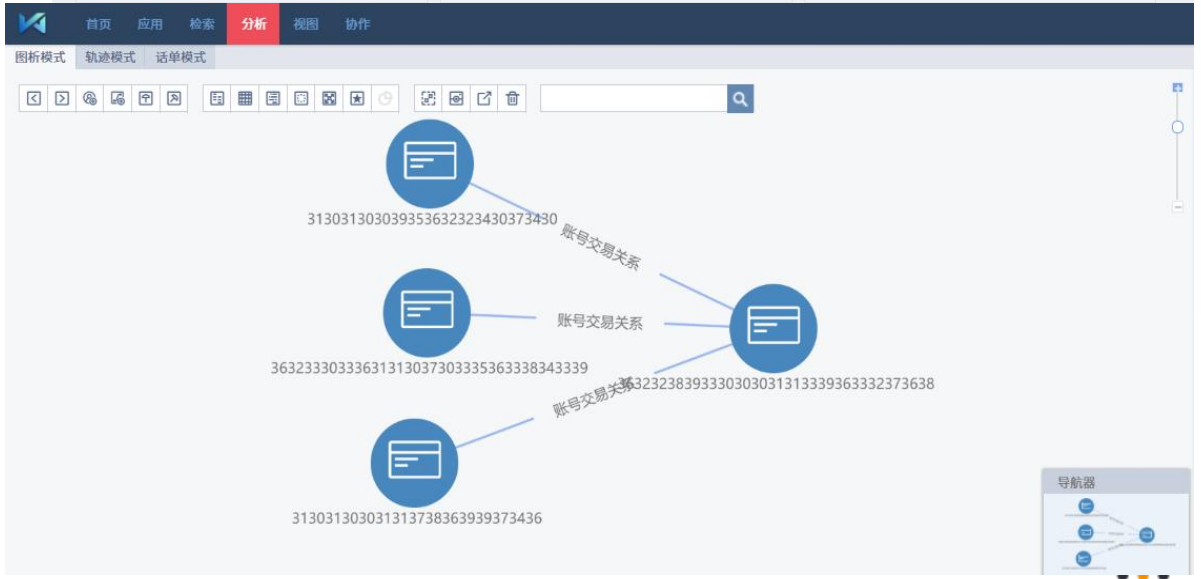
- **数据维度**：整合全行多业务条线数据，构建全行大数据关联关系分析平台，重构银行数据架构，建立银行内部的知识图谱数据库
- **技术维度**：运用机器学习和知识图谱技术，实现复杂业务条线下的数据架构整合。运用自然语言处理技术，解析半结构化文本数据，拓展数据维度，提升数据处理能力
- **业务维度**：全面提该银风险内控水平，辅助业务人员提高效率，对审计的业务有非常大的革新和价值。



某大型清算机构反洗钱项目

- **数据来源：**
  - 客户账户数据
  - 交易流水数据
  - 重点监控名单
- **反洗钱模型：**
  - 频繁汇入模型
  - 频繁汇出模型
  - 强连通模型
  - 分散模型
  - 汇聚模型
- **业务支持功能**
  - 关联账户/客户
  - 可疑交易管理
  - 黑名单管理和实时监控
  - 模型管理
  - …





目录

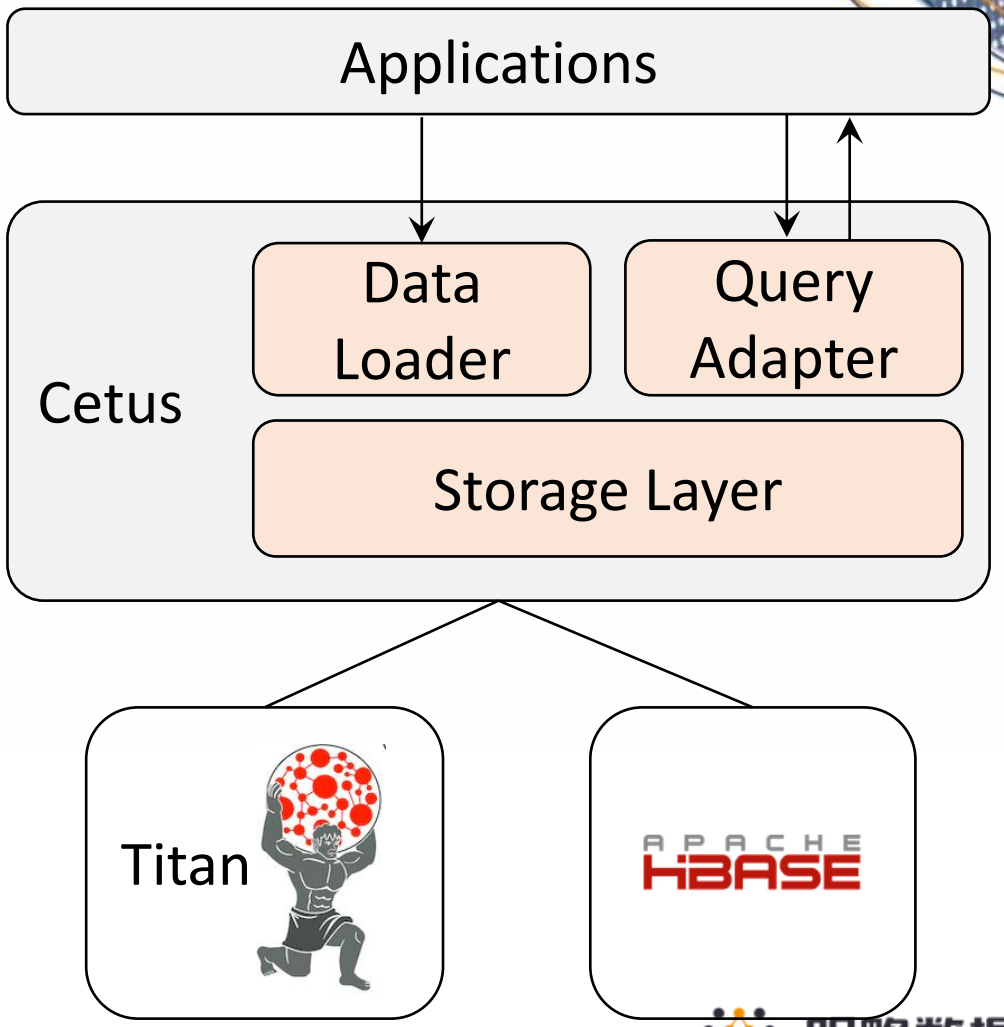
图数据库原生Titan的可优化之处

- 关系爆炸问题
  - 合并同类边的混合存储架构
- Super Node问题
  - 合并同类边的混合存储架构
- 多点碰撞查询效率
  - 基于场景建立合理的索引
- 批量导入数据性能
  - 基于子图划分的并行导入方案
- 灵活的索引管理
  - 多类型的索引统一管理

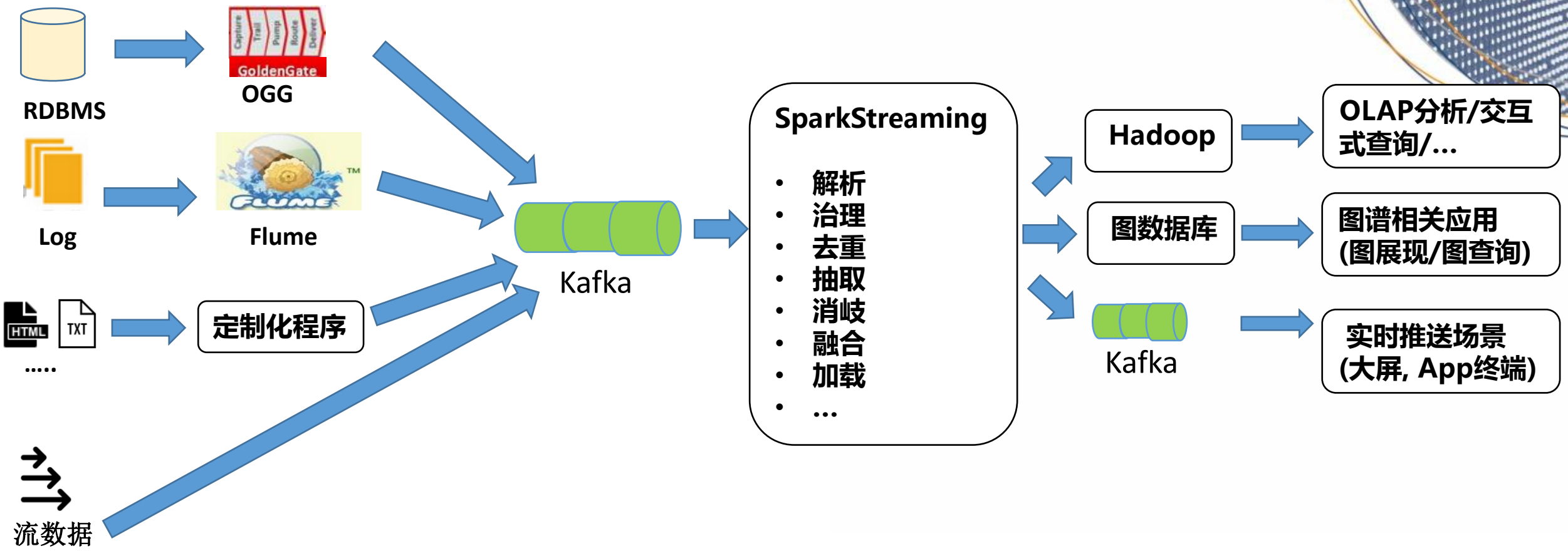
Cetus: 明略研发的面向图的混合存储架构

Cetus: 一种能高效处理大量重边的图数据库架构

- 基于Titan和HBase实现不同类型数据的存储
- Data Loader和Query Adapter为应用程序提供图管理接口



流式数据和图数据库的结合



谢 谢

明略数据 MININGLAMP