# 图数据库

## 第一章 初 识

### 概念

**图**

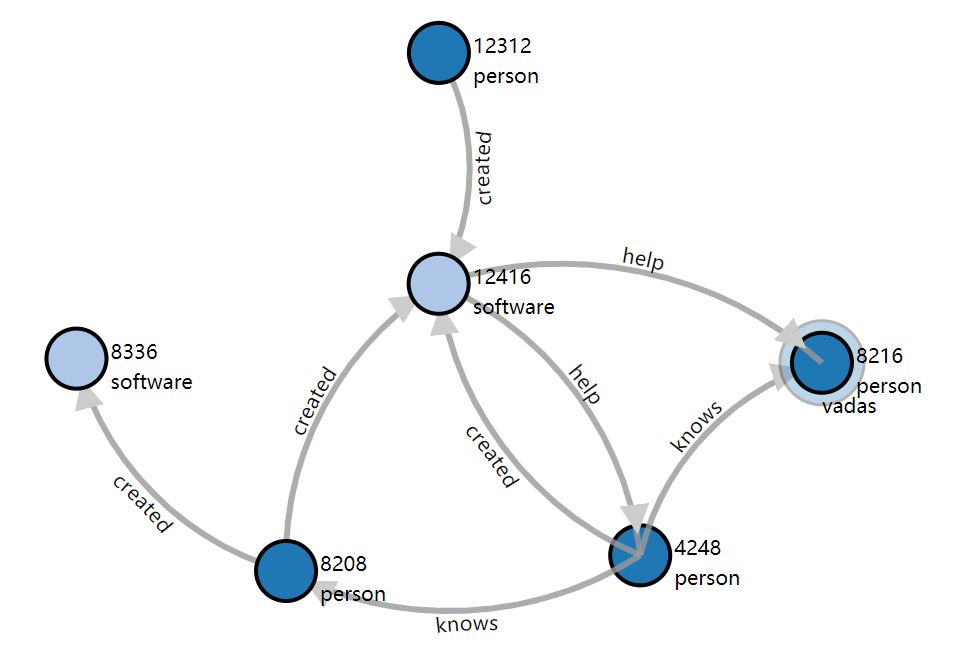
图由两个元素组成：节点、关系/边。每一个节点代表一个实体，每一个关系或边代表两个节点的关联方式。

如果非要和关系数据库映射，可粗略理解如下：

表名--节点标签，表记录--节点，表字段--节点属性，关联表名--关系/边标签，关联表记录--关系/边，关联表字段--关系/边属性。

**图形数据库**

图形数据库是一种非关系型数据库，是NoSQL数据库的一种类型，它应用图形理论存储实体之间的关系信息。



图数据库可视化效果

### **场景**

多对多场景

高价值场景

​ 元素之间的关系和元素本身一样重要，甚至比元素本身还要重要，例如数据治理

大规模低延迟场景

​ 图数据库可以比其他类型的数据库更快地处理大型数据集所表示的关系，尤其是存在在复杂的关系连接

### **领域**

**社交网络**

Facebook, Twitter，Linkedin用它来管理社交关系，实现好友推荐、 社区发现、兴趣用户推荐、垃圾用户甄别、舆论跟踪。

**电商领域**

eBay，沃尔玛使用它实现商品实时推荐，给买家更好的购物体验，例如智能推荐、精准营销、供应链管理、货物推荐、浏览轨迹分析。

**金融领域**

摩根大通，花旗和瑞银等银行在用图数据库做风控处理、银行欺诈、信用卡欺诈、电子商务欺诈、保险欺诈、反洗钱、账号合并、反欺诈、小贷授信、信贷审核、贷后追踪等。

**电信领域**

Verizon, Orange和AT&T等电信公司依靠图数据库来管理网络，例如深度经营分析· 防骚扰· 电信诈骗防范· 运营商经营分析

**政企**

物联网· 智慧城市· 道路规划· 智能交通· 轨迹分析· 钓鱼网站识别

**工业**

沃尔沃，戴姆勒和丰田等顶级汽车制造商依靠图数据库推动创新制造解决方案。其它如电网分析· 供应链管理· 设备管理· 流程管理· 故障诊断

**医疗**

智能诊断· 电子病历· 医保&保险分析

**军工**

情报网络· 研发管理· 人物追踪

### **比较**

目前主流的图数据库有：Neo4j，FlockDB，GraphDB，InfiniteGraph，Titan，JanusGraph，Pregel等。

Neo4j

Neo4j是老牌的图数据库，是一种面向网络的数据库，是一个嵌入式的、基于磁盘的、具备完全的事务特性的Java持久化引擎，它将结构化数据存储在网络上而不是表中。网络是一个灵活的数据结构，可以应用更加敏捷和快速的开发模式。

Neo4是一个高性能的图引擎，该引擎具有成熟和健壮的数据库的所有特性，具备完全的事务特性、企业级的数据库的所有好处。

JanusGraph

JanusGraph是非常成功的开源图数据库TitanDB的Fork版（TitanDB已经不维护了）。janusGraph是一个可扩展的图数据库，可以把包含数千亿个顶点和边的图存储在多机集群上。它支持事务，支持数千用户实时、并发访问存储在其中的图。

HugeGraph

[HugeGraph](https://hugegraph.github.io/hugegraph-doc/" \t "_blank)是百度在2018年中旬开源的一款易用、高效、通用的图数据库系统（Graph Database）， 实现了 [Apache TinkerPop3](https://tinkerpop.apache.org/" \t "https://www.oschina.net/p/_blank) 框架及完全兼容 [Gremlin](https://tinkerpop.apache.org/gremlin.html" \t "https://www.oschina.net/p/_blank) 查询语言， 具备完善的工具链组件，助力用户轻松构建基于图数据库之上的应用和产品。HugeGraph 支持百亿以上的顶点和边快速导入，并提供毫秒级的关联关系查询能力（OLTP）， 并可与 Hadoop、Spark 等大数据平台集成以进行离线分析（OLAP）。

HugeGraph 典型应用场景包括深度关系探索、关联分析、路径搜索、特征抽取、数据聚类、社区检测、 知识图谱等，适用业务领域有如网络安全、电信诈骗、金融风控、广告推荐、社交网络和智能机器人等。

TigerGraph

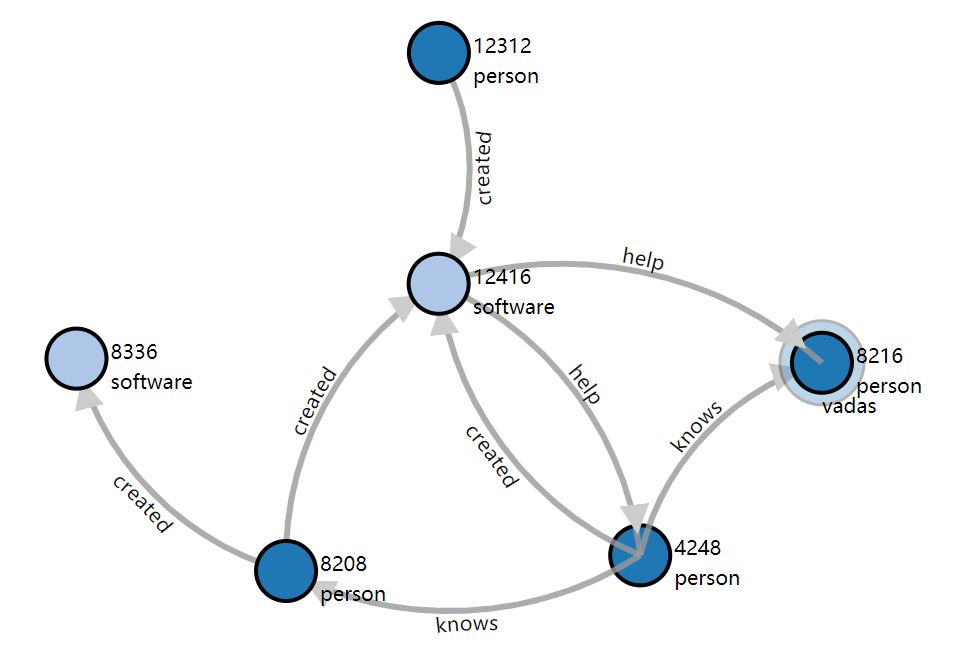
TigerGraph是一款“实时原生并行图数据库”，既可以部署在云端也可以部署在本地，支持垂直扩展和水平扩展，可以对集群中的图数据自动分区，遵循 ACID 标准，并且提供了内置的数据压缩功能。它使用了一种消息传递架构，这种架构具备了可随数据增长而伸缩的并行性，被设计用来执行深层链接分析以及实时在线事务处理（OLTP）和大容量数据加载。

TigerGraph 使用了一种新的查询语言 GSQL。GSQL 将 SQL 风格的查询语法与 Cypher 风格的图导航语法结合在一起，并加入了过程编程和用户自定义函数，支持并行化，且可以将 Cypher 转换为 GSQL，便于从 Neo4j 数据库转到 TigerGraph 的开发者使用。TigerGraph 使用了几种流行的开源组件：用于处理 Web 流量的 Nginx、消息队列 Apache Kafka、用于管理 Kafka 集群的 Apache Zookeeper。

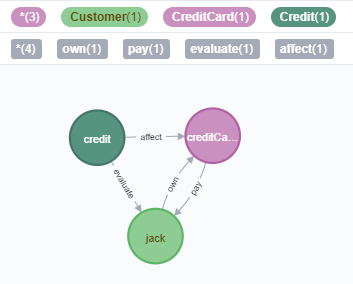
neo4j、janusgraph和hugegraph对比：（本人不想进一步整理，爱看不看）

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Neo4j** | **JanusGraph** | **HugeGraph** |
| 开源 | 社区版开源，企业版收费 | 完全开源，Apache协议，开放性最好，兼容tinkerpop生态 | 百度开源，基于JanusGraph，兼容tinkerpop生态 |
| 技术特点 | 一站式服务、工具齐全 | 自行组合，技术门坎高 |  |
| 查询语言 | Cypher | Gremlin | Gremlin |
| 开发语言 | Java,scala | Java | java |
| 集群 | 企业版支持，社区版不支持 | 支持 | 值 |
| 量级 | 轻量级 | 重量级 |  |
| 额外组件依赖 | 官方提供组件，可以与ES、MongoDB、Cassandra等NoSqlDb进行交互 | 数据存储服务可以选用Cassandra、HBase或Berkeley DB服务。 数据索引可以选用Es、Solr或Lucene服务。 | 数据存储可以选用mysql oracle pgsql sqlserver |
| 事务 |  | 事务数据库，支持数千并发实时复杂查询，支持ACID和最终一致性 | 使用关系数据库存储支持事务 |
| 数据备份恢复 | 企业版支持，社区版手工停服务备份，支持热备 | 手动写程序 | 支持在线远程备份，支持在线远程恢复 |
| 访问控制 | 为数据库管理员提供指定的用户角色功能 | 支持用户认证 | 支持用户认证，基于用户角色的权限控制 |
| 数据库监控 | 提供监控数据库运行状况的整套指标 | 支持metrics监控 | 支持监控接口 |
| 运营管理工具 | 管理员命令行界面 |  |  |
| 查询能力 | 3度，快 | 3度，较快，性能抖动较严重 | 3度，快，较稳定 |
| 使用情况 | 使用最广最流行 | 很广 | 国内知名度高 |
| 存储 | 非分布式存储 | 分布式 |  |
| 数据规模 | 社区支持十亿级 | 百亿级 | 千亿级 |
| 大规模写入性能 | 脱机导入较快 | 较慢 | 在线导入快，支持覆盖写 |
| 功能完善程度 | 最完善 | 完善 | 完善 |
| 数据导入 | 支持CSV在线导入，速度1万/秒内；支持脱机导入，速度在10万/秒内，只能用于初始化导入 | 未提供支持 | 支持在线导入，速度在10万/秒，支持csv txt json，支持HDFS导入，支持从mysql oracle pgsql sqlserver导入，支持消息队列导入 |
| 增量备份 | 企业版支持，且支持备份数据加密 | 不支持 | 不支持 |
| API和客户端 | 支持http api，支持java c# js语言版本客户端 | 支持http api或web socket，支持java python c# js版本客户端 | 支持http restful api，支持java客户端，支持gremlin api |
| 可视化 | 支持，功能丰富 | 不支持，第三方支持 | 支持，功能丰富，支持可视化数据建模、导入、分析等 |
| 社区活跃度 | 遥遥领先 | 文档一般 | 部分国内文档，github活跃度很高 |

## **第二章 查询语言**



Gremlin



Cypher

Gsql