

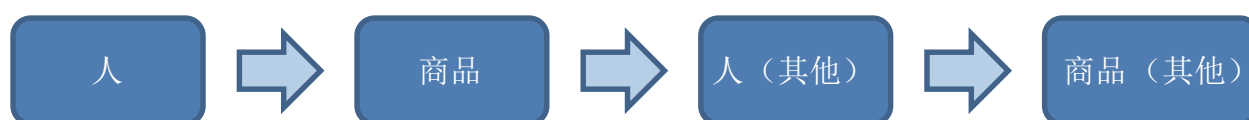


实时深度链接分析： 图分析的下一个里程碑

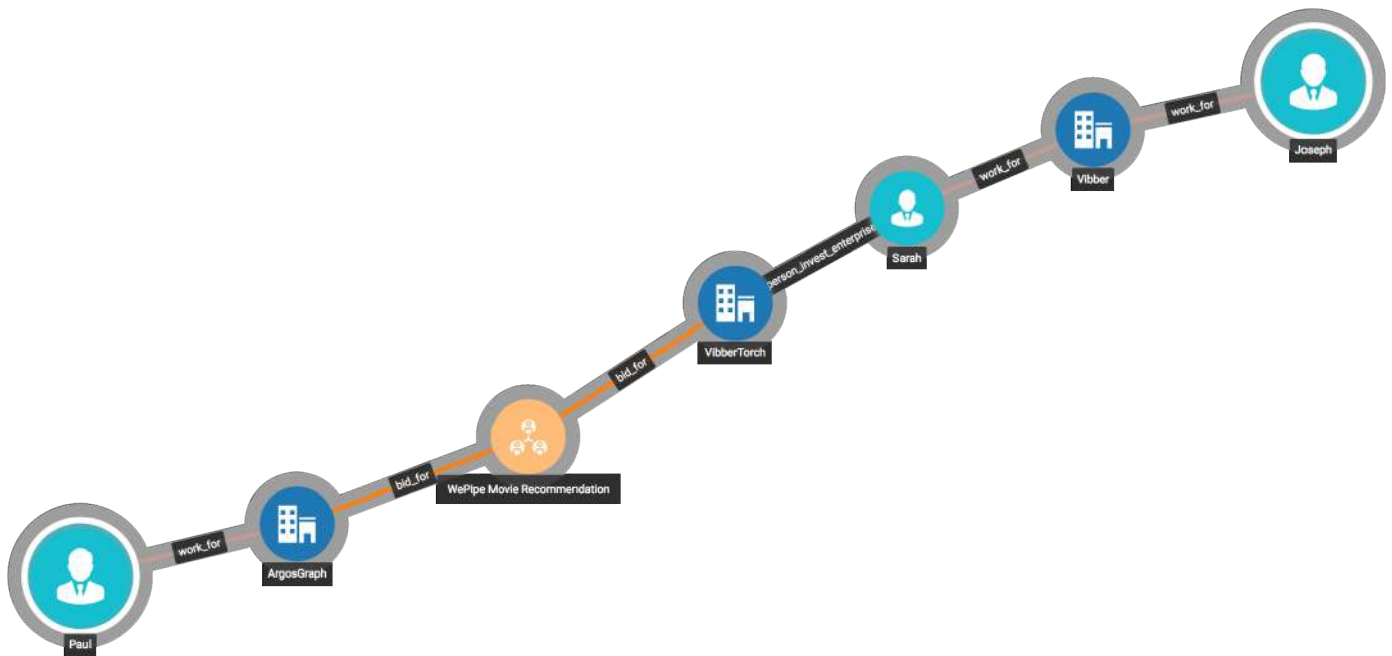
作者：TigerGraph首席执行官 许昱



图数据库的最大优点之一在于它能够分析和回答海量数据中复杂的数据关联问题。与传统的关系数据库应用程序相比，早期图技术在这种关系分析方面已经显现出明显优势。但是，面对当今企业生态系统中普遍日益增长的实时数据时，早期图技术则显得运行缓慢，分析能力有限。这是因为现有的图技术难以加载海量数据，无法提供令人满意的图遍历速度，并且无法导入快速到达的实时数据。因此，当前的图技术在大图中通常限于两步的实时遍历（即最多走两步）。让我们一起来看看这个经典个性化的推荐问题：“那些与你喜好相同的顾客还购买了哪些商品？”。首先，从一个人开始，该查询首先确定你浏览 / 点赞 / 购买过的商品。其次，它确定浏览 / 点赞 / 购买过这些商品的其他顾客。第三，找出这些人购买过的其他商品。



该查询要求在图上实时走三步，因此它超出了当前图技术对于海量数据集 2 步查询的限制。如果再增加一层关系很容易将查询扩展到 4 步或更多步。



当今的企业因为业务需要，往往要求图技术能够实时做到3到10步以上的浏览和分析，从而发掘和预测错综复杂的关系。每增加一步，即可揭示更多连接和隐含的关系。

这就是实时深度链接分析，它具有以下性能属性：

性能	指标
可连接性高	3到10步以上
图遍历速度快	每个服务器每秒遍历超过1亿个顶点或边
数据更新快	每个服务器每秒更新10万次以上

实时深度链接分析应用案例

实时深度链接分析在当今企业的许多重要的数据分析应用中，成为一把利器。这里我们给出三个案例：实时反欺诈、个性化推荐、和优化供应链。



风险 / 欺诈控制

实时深度链接分析通过辨识高风险交易来打击金融犯罪。例如，从一个新的信用卡交易开始，该交易与其他实体发生关联的方式可以辨识如下：

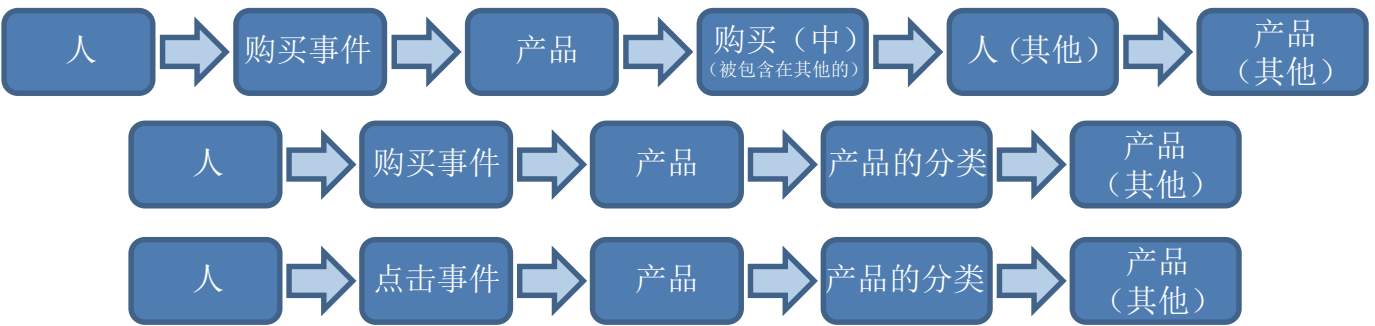


该查询使用4步来寻找与传入交易只有一卡之隔的连接。现今的欺诈者们试图在自己与已知不良行为或不良行为者之间保持迂回连接，藉以掩盖自己的不法活动。任何单条连接路径可能看起来清白无辜，但是如果把从A到B的多条路径联合起来分析，很多欺诈行为被暴露出来的机会就大大增加了。因此，需要更多步才能找到相隔两个或多个交易的链接。这种遍历模式适用于许多其他用例。在这些用例中，你可以简单地使用网页点击事件、电话记录或汇款事件替换交易。通过实时深度链接分析，可以揭示多个隐藏的连接，从而最大程度发掘和及时阻止欺诈行为。



多维个性化推荐

为了找到类似的顾客，或者为产品提供个性化推荐，需要进行实时深度链接分析。请见以下推荐路径示例，其中所有路径均超过 3 步：



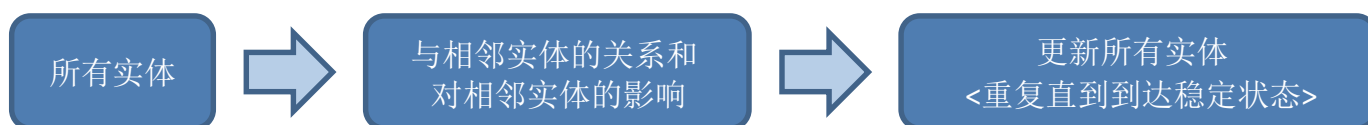
由于当前系统的局限性，今天看到的多数推荐功能都是使用离线计算。这些推荐是在后台提前计算好的，不能利用最新数据执行实时的按需分析，从而丧失很多推荐机会或造成过时的推荐。



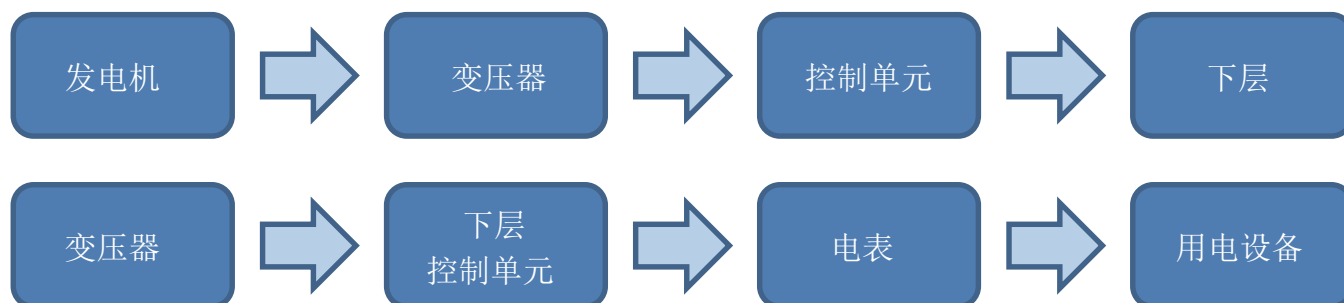
电力流 / 供应链物流 / 道路交通优化



这些应用调整各个实体和/或各个连接，直到建立一个动态平衡或最优的状态。例如，在国家或区域电网中，根据实际使用情况，动态调整每台发电机的电力输出。



为了得到适当的值，需要在图数据上进行迭代计算（类似于 PageRank），直到其中的度量值收敛为止。在这里，每个顶级组件（例如，发电机）是多个从属部件的网络枢纽，从而导致多层关联。在电力分配网络，关联程度通常需要 6 步。



对于这类案例，只有实时深度链接分析才能达到这个级别的计算要求。



实时深度链接分析与企业人工智能

实时深度链接分析是企业人工智能的关键组成部分。由于其技术局限性，早期的数据分析技术人为地打散实体和关系，并将它们作为数据行或键值对，分散存储在多个表 / 行 / 列中。从而导致一些关系丢失，或者增加处理难度。实时深度链接分析可以处理数据实体之间的所有连接，使用户能够更准确、更快速和更深入地浏览、发现和预测关系。这些功能对于个性化推荐、欺诈防范、供应链物流优化、公司知识图谱构建和查询等企业人工智能应用都是至关重要的。

原生并行图平台： 图数据库的下一个里程碑

数据库咨询公司DB Engines调研结果显示，图数据库是所有数据库管理系统中增长最快的一类。它们之所以被很多企业青睐并快速采用的主要原因是基于图数据库多方面的技术优势。首先，图数据库使用顶点、边和属性的图关系格式存储数据，克服了存储海量、复杂和具有互连关系数据的挑战。其次，与传统的关系型数据库管理系统（RDBMS）和近年来新兴的其他大数据平台相比，图数据库的以下优势更为明显：

- 更快的连接相关数据对象
- 更大的数据集扩展性
- 更灵活地应对经常变化的结构

虽然早期的图数据库具有这些优势，但现有的图数据库产品都存在一个难以克服的障碍——对于总量和复杂性不断增长的数据生态系统，不能进行实时分析。这个障碍直到TigerGraph的出现，才得以解决。

介绍 TigerGraph: 全球首个也是唯一的原生并行图平台

作为全球首个和唯一的原生并行图平台，TigerGraph 是一个完整的分布式图数据分析平台，支持互联网规模的实时数据分析。TigerGraph 原生并行图系统建立在本地存储和计算的基础上，它支持实时图谱更新，使用并行计算引擎。这些设计提供了以下独特的优势：

- 快速加载数据来构建图谱——每台计算机每小时可载入 50 到 150GB 的数据
- 快速执行并行图算法——每台计算机每秒可以遍历亿万级别的顶点和边
- 利用 REST 接口实时更新和添加新的数据——在只有 20 台商用服务器的集群上，能够将 20 亿以上的每日事件实时流式传输到具有超过 1000 亿个顶点和超过 6000 亿个边的大数据图谱。
- 把实时分析与大规模离线处理完全统一起来——全球第一个做到这一点的系统

原生并行图系统与早期的 原生图技术有何不同

早期的原生图技术和TigerGraph的原生并行图系统在三个方面有着显著的差异。

- 单机与分布式系统（可扩展性）——早期的原生图技术无法把一个图拆分成多块，分布放到多台机器上存储。现今的图通常大于单机存储能力。TigerGraph 用原生并行图技术很好的解决了这个问题。它能够很自然地把图扩展并分散放到多台服务器上，从而图的大小不再受单机容量的限制。
- 在单个服务器上的性能——由于内置的并行计算功能，原生并行图系统可以在单机上提升查询性能 10 倍至 100 倍。同时，TigerGraph 的原生并行图系统可以在大图中遍历 3 到 10 步以上，而早期的图技术在这种情况下只能超时，无法返回结果。
- 实时更新图的速度——TigerGraph 的原生并行图系统由于其并行数据处理能力，通常可以提供 10 倍以上的加载速度。

TigerGraph 原生并行图系统的独特设计和功能

通过以下核心平台技术，TigerGraph 可以实现令人难以置信的加载速度、快速的查询执行和实时更新功能。



图存储引擎和图计算引擎

TigerGraph 的原生并行图内核采用 C++ 编写，以实现最佳性能。它的技术堆栈很好地集成了存储和计算模块。原生图存储引擎（GSE）与图计算引擎（GPE）同步工作，快速高效地处理数据和算法。图计算引擎（GPE）采用比 Mapreduce 更高效的计算模型，从而提供内置并行性。图存储模块为图遍历和更新做了磁盘和内存的优化设计，从而使系统能够利用硬盘、内存中的数据本地性进行最优调度。



快速数据加载

TigerGraph 凭借其固有的并行性，能够实现每台计算机每小时加载 50 到 150GB 的数据。TigerGraph 凭借批量在线加载、内置加载语言、内置索引等等，提供最完整最方便的加载环境。



高数据压缩比

在数据获取中，TigerGraph 系统执行极其高效的无损失数据压缩和编码。通常有 2 到 10 倍的压缩比（输入数据体积与输出图体积之比）。这种压缩不仅减少内存占用，还提高 CPU 缓存命中率，从而加快整体查询性能。TigerGraph 系统还直接在经过内部压缩的数据上运行，从而带来更多性能优势。



并行计算模型

在 TigerGraph 系统中，图既是存储单位又是计算单位。图的一个顶点或一条边可以存储任何数量的任意信息。此外，每个顶点和边可以与一个计算函数相关联。因此，它们同时作为存储和计算并行单元。这是绝无仅有的，不同于任何现有的图数据库。利用这种方法，图中的顶点和边不仅仅是静态数据单元，而且顶点是主动计算元素，它们通过边进行通信。

TigerGraph 的计算平台将 MapReduce 和并行图处理模型统一起来，它基于大规模同步并行（BSP）的编程模型，使开发人员可以快速而轻松地实现可扩展的并行图算法。类似 SQL 的图查询语言（GSQL）提供大数据的实时查询和交互式分析。利用多核 CPU 计算机和内存计算，TigerGraph 系统在每个顶点 / 边上并行执行计算功能。企业级用户可以实现每台服务器每秒遍历超过 1 亿个顶点或边。



图分割

TigerGraph 系统支持各种图分割算法，以实现最佳性能。在大多数情况下，对输入数据自动执行分割即可获得很好的结果，而无需任何人为优化和调节。TigerGraph 系统十分灵活，用户可以面向具体应用定制分割策略或者采用混合分割策略实现更优的性能。

TigerGraph 系统内置了将多个图引擎作为主动 - 主动网络运行的功能。利用为不同类型的应用查询量身定制的不同分割算法，每个图引擎可以托管相同的图。前端服务器（通常是 REST 服务器）可以根据查询的类型将应用查询路由到不同的图引擎。



实时深度链接分析

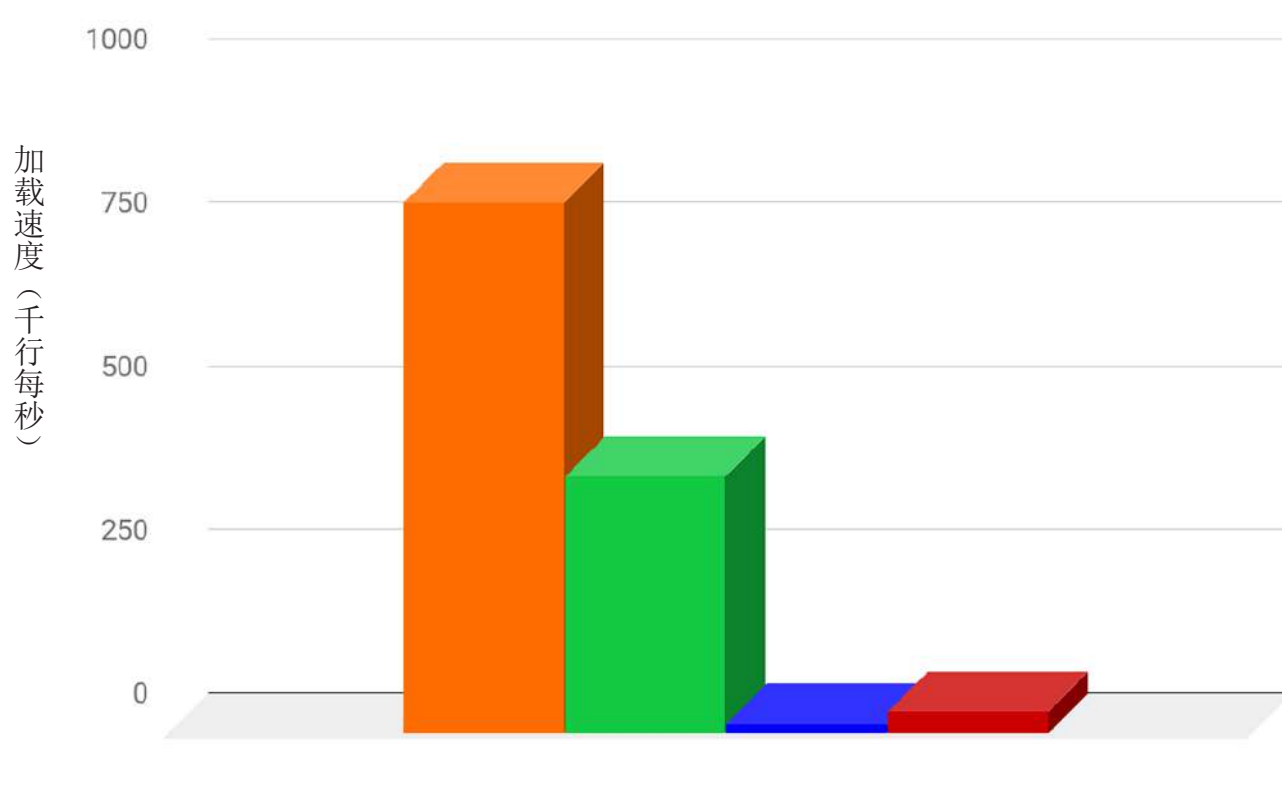
当前图技术通常限于大图内的两步的遍历（即两度分离），而 TigerGraph 的原生并行图系统能够实现 3 到 10 步以上的遍历。每增加一步，即可揭示更多连接和隐含信息。这种深度链接分析能够达到每台服务器每秒 1 亿以上个顶点 / 边遍历。相应请求同时可以每台服务器每秒 10 万次以上的快速数据更新。



超大图的速度和可扩展性

平均加载速度，针对具有 14 亿条边的数据表

■ TigerGraph online ■ Neo4j-offline ■ Neo4j-Cypher ■ Titan



TigerGraph 对超大图支持并行算法。即使在单个服务器中也采用并行化，以最大限度地提高速度，并确保性能随着计算机的增加而提升。这是非常重要的，因为图总是会越来越大。图存储引擎对图大小没有上限；它会简单地增加地址空间，并将图分割到更多的计算机上。原生并行图系统既适用于接触图的一小部分到数百万个顶点和边的快速查询，也适用于必须涉及十亿量级图中每个顶点的更复杂的分析。

扫码关注微信公众号下载试用版



关于 TigerGraph

TigerGraph 是基于原生并行图 (NPG) 技术的全球首个实时图分析平台。TigerGraph 通过为具有复杂和海量数据的企业提供实时深度链接分析支持, 实现图平台的真正承诺和好处。TigerGraph 的成熟技术已经被蚂蚁金服、VISA、软银、中国国家电网公司、Wish、Elementum 等客户所采用。TigerGraph 由许昱博士于 2012 年创立, 并获得启明创投、百度、蚂蚁金融、AME 云创投、莫拉多风险投资公司、佐德·纳齐姆、丹华资本和 DCVC 风投基金公司的投资。TigerGraph 的总部位于加利福尼亚州红木市。如欲了解更多信息, 请访问 www.tigergraph.com.cn。

