neo4j

帮助世界理解数据

从图到知识图谱
踏上收获无限洞察
分析的快捷之旅

Maya Natarajan

产品营销高级总监

引言

企业要想具备竞争优势，当务之急就是充分利用知识的力量。

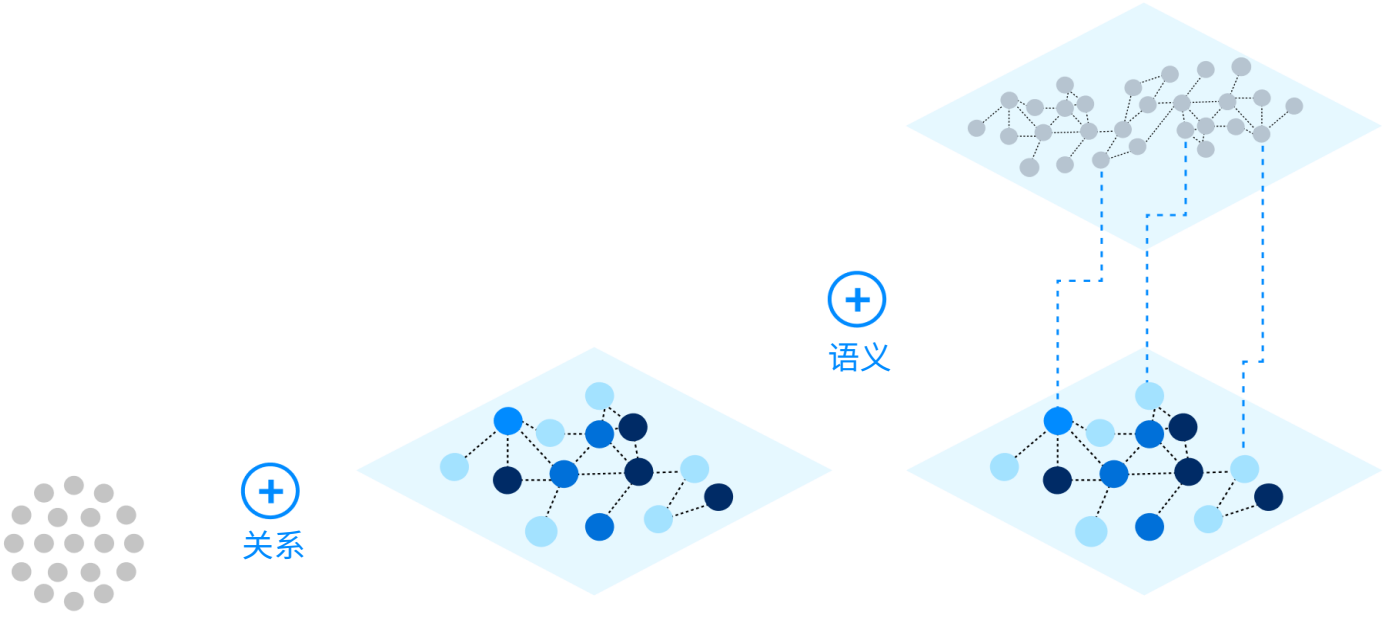
人人都渴望获取知识，以便将自己所掌握的一切信息进行串联。然而，将数据转化为知识在很大程度上仍任重道远，尽管目前已经取得了进展，但大多数数据格局还远未成熟。

我们急需一种方法来关联所有系统、云、备份和数据湖中的数据。一些数据传输管道尽管性能强大，能为仪表板提供实时、干净的数据，并为业务流程提供相关信息，但这还远远不够。我们必须在方方面面都取得此类成效。

现在是关联数据的最佳时机，以便数据易于管理并提高实用性。您组织中需要数据来完成工作的所有员工都应了解数据查找位置、对数据的准确性充满信心，并能够轻松解决自己遇到的问题。

您可以从小处着手，以便收获立竿见影的效果，然后从图开始向外扩展，直到知识图谱，从此开始踏上收获无限洞察分析的快捷之旅。

导致现有数据挑战的原因是什么？



数据趋势和挑战

数据孤岛

首先，有些数据相互孤立。数据存储和应用通常为单个组或部门提供服务。比如，人力资源团队有一个平台，而销售团队可能会使用另一个平台，如 Salesforce。记录系统十分重要，因为这些系统可维护并控制数据，还能建立管控策略，但数据孤岛也会降低分析速度、报告准确性和数据质量。

数据扩张和数据湖

大多数组织都有数据湖、数据仓库、关系型数据库，作为记录系统、客户数据、交易数据、产品数据和订单数据等等的基础。这种数据分散状态将导致数据无序扩张。

数据湖可存储大量结构化、半结构化和非结构化数据，且成本较低，因此备受欢迎。Amazon S3 等对象存储通常用于创建数据湖。

从成本来看，数据湖很有吸引力，可助您存储任何类型的数据，包括应用和服务生成的日志文件。将数据存入数据湖的操作既简单又方便。然而，管理并了解这些数据则挑战重重。

云存储

云计算具有颠覆性意义，但仍面临管理挑战，因为更多数据存储在更多系统中。回忆一下，您的个人云数据是否存储在iCloud、GoogleDrive、Dropbox、Evernote、Gmail 和 Notes 中。

如果是这样，我们需要处理大量数据，这些数据不仅格式不同，且部分数据是多余的，而且大多数数据都没有相互关联。

如果 GDPR 合规性要求您忘记某用户的数据，您知道所有这些数据的存储位置吗？那些您甚至可能没意识到属于用户的 Cookie 呢？还有您的物流系统和您合作伙伴系统中的数据呢？

历史数据惨遭淘汰

历史数据为机器学习预测提供了动力。但新冠疫情爆发后经济动荡，导致历史数据过时。例如，历史数据通常被用来预测购买行为。

但由于疫情期间的封控状态，网购成为市场主导购物方式，购买行为几乎在一夜之间发生了变化。由于消费者行为发生了翻天覆地的变化，历史数据无法对购买行为做出准确的预测。

在数据有限的情况下，数据关联的重要性和价值日益凸显。将数据存入图数据平台可捕获数据关联和数据关系。

您拥有的数据很有价值，但存储数据中和数据之间已存在的关系可以提高您的预测能力，即使在没有相关历史数据的情况下也是如此。这是因为数据关联和关系是数据中最具预测性的元素。

以上所有因素都在推动您这样的企业向关联图数据平台中的数据转型，从而获取知识。下一节具体介绍图如何转化为知识图谱。

> “ 模糊混乱的数据是合规性的噩梦。

什么是知识图谱?

定义如下：

知识图谱是一个语义丰富、相互关联的数据集，以便我们对底层数据进行推理，并大胆将其用于复杂的决策过程。

以下内容将简单解释此定义。

关联数据的价值

知识图谱是一个语义丰富、相互关联的数据集，以便我们对底层数据进行推理，并大胆将其用于复杂的决策过程。

关联数据可带来更多情境信息并改善最终成果。打个比方，医术高超的医生会做足功课。如果医生花时间检查病患的病因，例如医疗条件、病史、做过的所有实验室测试以及生命体征，此次诊疗的效果会更好，这是因为医生掌握病患的情境信息。

领英的做法则是另一个更具代表性的例子，该公司拥有的关联数据是公认的行业标杆（关联数达 500+）。正如所有著名社交网络一样，领英以图数据结构为基础。

一家卫星制造商希望在其高度复杂的产品的整个生命周期中集成所有流程和数据 - 这就是产品 360。

首席数据官表示：“过去，我们需要员工手动确定故障的根本原因，他们要检查一切可能引起零件故障的因素。故障出自工程？采购？还是供应商？是供方问题还是制造缺陷？我们的策略是利用图完成遍历、找到差异并即时报告，而非利用人工花费数周手动完成这项工作。”

语义丰富

知识图谱是一个语义丰富、相互关联的数据集，以便我们对底层数据进行推理，并大胆将其用于复杂的决策过程。

关联数据会使数据本身更有价值，并提供动态情境。增添更多相关信息（语义）将进一步提升关联数据的价值。

RDF：房间里的语义大象

1989 年，Tim Berners-Lee 博士创造了我们现在所知的网络。这项发明创造要归功于瑞士的粒子物理实验室——欧洲核子研究中心 (CERN) 的相关科学家们，这些科学家拥有大量五花八门的计算机，如果他们能够分享相关研究成果，就能更快取得进展。

不到 10 年，也就是 1998 年，Berners-Lee 博士宣布语义网将成为下一重大发明。关于 XML 和资源描述框架 (RDF) 等标准的工作也开展起来。

从那时起，大量的工作投入到了使用基于 RDF 的词汇表的本体上，例如 FIBO 和 SnoMed。维基数据 (WikiData) 这样的大型公开数据集也可以 RDF 的形式提供。

那些对知识图谱转型感兴趣的人如何利用此类数据?

Neo4j 知识图谱包含 RDF 以及您可能拥有的任何其他类型的数据。Neo4j 是属性图，因此十分简单、直观且可扩展。简单是指可捕获以 RDF 或任何其他数据模型表示的所有信息。直观是指在节点和关系的简单构建块上工作，因此很容易理解和解释。可扩展是指可容纳数十亿个节点和数万亿个关系。

重中之重：充分利用 RDF 的丰富性和 Neo4j 知识图谱的简单、直观和可扩展的特点。

实际上，如果您创建一个图，就会发现自己想要添加更多详细信息，即使是在白板上创建图也是如此。这些新增信息就是语义，即为图赋予意义。

在 NASA，David Meza 试图将数亿份文件、报告、项目数据摘要、经验总结、科学研究、医学分析等信息添加至知识图谱。Meza 通过主题建模算法运行了所有文档的文本，该算法发现了可在内容中应用的标签，从而丰富了图。

根据 Meza 的说法，NASA 的“经验总结”知识图谱加速了火星探索任务，在研发方面节省了至少一年时间和 200 多万美元。

鉴于这些经验，Meza 现在在做什么？他已经进入了下一个前沿领域：在私营部门进行深入研究之际，寻找合适人选来推动 NASA 的发展。

毫无疑问，人才本身就是一个图。NASA 的全新知识图谱如今用来挖掘具有特定知识、技能和能力的人员。但职位和职称并未标准化，所以 Meza 将这些信息从政府数据库（包括美国劳工部的 O*Net、人事管理办公室（OPM）类别和欧盟的 ESCO）映射至现有本体。

虽然 NASA 在人才招聘方面有特殊要求，但招聘问题是大多数组织普遍面临的挑战，无论何种任务，知识图谱都可能有助于解决这一挑战。

对底层数据进行推理

知识图谱是一个语义丰富、相互关联的数据集，以便我们对底层数据进行推理，并大胆将其用于复杂的决策过程。

知识图谱可提供深入的动态情境信息，让人们能够一站式查找所有相关信息以及这些数据之间的所有关系。添加的信息越多，知识图谱也将越有价值。

例如，凭借 NASA 的“经验总结”知识图谱，工程师们可以查看任何特定部件并深入了解其使用方法。工程师可以查询特定部件及相关子系统。他们可能会探索由不同群组管理的所有子系统是如何结合在一起的，并试图从整体上把握这个系统。

欺诈调查人员要处理数以千计的潜在欺诈活动警报。筛选这些警报需要查看几个系统，十分耗时。知识图谱汇集了许多数据源，在这种情况下，汇集了客户属性、信用评分、位置、支付历史等数据，以便调查人员快速锁定欺诈活动，并发现此前注意不到的模式。

放心大胆地使用

知识图谱是一个语义丰富、相互关联的数据集，以便我们对底层数据进行推理，并大胆将其用于复杂的决策过程。

数据冲突。您可能没见过酒吧斗殴，但我们大多数人都见过数据冲突（更文明的说法是数据分歧）。数据冲突由数据不匹配的差异引起。有人会问：“你从哪里得到的数据？我的数据有所不同。”

我们如何才能放心大胆地使用数据？答案是了解数据源、转换次数、是否被清理过、最后一次更新时间，以及更新数据的用户。

借助知识图谱，可轻松将这种类型的元数据添加到图中。知识图谱擅长捕获数据的详细信息，这些元数据将成为知识图谱的一部分。在发生数据冲突（或数据讨论）的情况下，这些类型的详细信息在调查任何差异方面都是无价之宝。

复杂决策

知识图谱是一个语义丰富、相互关联的数据集，以便我们对底层数据进行推理，并大胆将其用于复杂的决策过程。

NASA 高级数据科学家 David Meza 表示：“一旦开始研究自己拥有的文档类型以及如何将这些文档转化为最终用户可操作的知识时，就意味着已开始改进决策。”

Dooloo 是一家总部位于法国的初创公司，该公司开发了一个基于知识图谱的平台，为忍受慢性疼痛的患者提供帮助。通过将患者的病史与最新研究和治疗方法相结合，患者及其护理人员可迅速获得应对策略，以提高他们的生活质量和病患护理水平。

知识图谱类别

知识图谱就像知识本身一样，涉及方方面面，信息广泛，无穷无尽。一般来说，知识图谱分为两类：**行动型**知识图谱和**决策型**知识图谱。

用于数据管理的行动型知识图谱

数据管理是知识图谱的一个重要用例。Lyft 和爱彼迎等数字公司依靠数据实现蓬勃发展，帮助数据科学家找到最新数据是公司获得成功的关键。

这些公司和其他很多公司都使用知识图谱创建元数据中心，并以此捕获数据沿袭：数据源、转换方式以及清理方式。知识图谱针对复杂的数据传输管道进行建模，以便轻松识别数据的消费者和生产者，并集成新的数据源。

借助数据源的相关强大基础，您就可以针对这些数据采取行动，准确了解数据源、数据生产者和消费者。

除了数据管理用例之外，行动型知识图谱还用于个性化和推荐。行动型知识图谱将客户和产品等所有相关数据汇总到一个 360 度视图中，从而推动采取大量行动，如识别有流失风险的客户以及提供可说服客户留下来的优惠建议。

用于数据分析的决策型知识图谱

知识图谱构成了现代数据和分析的基础。凭借知识图谱捕获的数据，您可以捕获及存储数据中固有的所有关系，无需猜测数据相关性。这样一来，知识图谱就代表了对数据更忠实的表述，并使您能够解锁其预测能力。

借助决策型知识图谱，最终目标是做出更好的决策，无论是人的决策还是算法决策。这些决策可以通过几种方式获得支持。

借助图查询，您可以批量回答有关知识图谱的任何问题。Boston Scientific 使用高级查询分析根本原因，并确定导致缺陷的故障组件组合（这是一种反向推荐）。

图算法可识别数据中的模式，例如两点之间的最短路径或最有影响力的客户。

OrbitMI 使用决策型知识图谱来执行复杂的集装箱船运航线规划。通过寻路算法，他们在不到一秒的时间内就规划出了海上航线。此外，他们的知识图谱还可支持 SaaS 分析产品。知识图谱不仅会产生经济影响，提高复杂路线规划的效率，还减少了 6 万吨碳排放。

图查询和图算法还可以解锁机器学习的预测功能。

阿斯利康在其知识图谱中使用图算法和机器学习来识别患者病历原型和模式。这项研究使该公司能够为早期干预确定强有力的触点，从而改善肾脏等方面疾病的治疗效果。

踏上收获无限洞察分析的快捷之旅

我们叠加组织原则时，图即可转化为知识图谱。因此，知识图谱使数据更加智能。

从图到知识图谱的转变十分微妙自然，就像是制作一张供应链图，再加上您在供应链中与合作伙伴的合作时间、合作伙伴的位置、运输方式、折扣以及任何必须考虑的运输延误等信息，即可转换成知识图谱。

这就是 Neo4j 知识图谱的美妙之处。您可以从小处着手，解决任何类型的实际挑战，并逐步丰富您的知识图谱，以便解决更多用例并为更多利益相关者提供服务。

如果您决定调整知识图谱的结构或添加数据源，Neo4j 可助您轻松修改图谱结构并让数据在该结构中重新流动起来，这样的敏捷性是底层图数据平台的一部分。

这种看似简单的结构可支持高级人工智能和机器学习，因此，每个图数据科学项目都始于知识图谱也就不足为奇了。

Neo4j 为数据管理、数据分析以及机器学习提供了市场上最全面的知识图谱。

访问 neo4j.com/knowledge-graph，详细了解可最快转换至知识图谱的路线。

对 Neo4j 有疑问?

欢迎联系我们:
china@neo4j.com
neo4j.com/contact-us