Graph Databases

Transmuting Information to Knowledge with an Enterprise Knowledge Graph

Jans Aasman, *Franz Inc.*

The enterprise knowledge graph for entity 360-views has emerged as one of the most useful graph database technology applications when buttressed by W3C standard semantic technology, modern artificial intelligence, and visual discovery tools.

Every enterprise has a few core entities that it's most interested in. For a hospital, this would be the patient; for a telephone company, a customer; and for an intelligence agency, people and organizations of interest. A *knowledge graph* is a new application of graph technology that collects several layers of knowledge related to an entity of interest (see Table 1). I have been involved in several knowledge graph projects, and here describe this emerging pattern.

Table 1. Knowledge Graph Components

Knowledge Components	*Telecom*	*Hospital*
Core entity	Telephone customer	Patient (this accounts for 99% of the data, but clinicians, drugs, and treatments are entities, too)
Core events	Call detail records, invoices, billing, bill payments, new phones, customer care phone calls	Every interaction with the patient (diagnosis, test, procedure, medication administration, check in, check out, billing)
Domain knowledge	What are the capabilities of a particular phone? What are the available phones and contract plans? What plan is a customer on?	More than 180 ontologies, taxonomies, and terminology systems Databases, including PubMed, adverse reactions, drug databases, and clinical trials
Knowledge inferred by rules	Does the user fit with this plan? Should he or she use another plan? This customer doesn't usually pay his bill on time but he always pays.	This customer has asthma and dermatitis, so we better test her for a peanut allergy. Given that she has diabetes, and already tested positive for A and B, it is now time to do test C.
Machine learning and statistics	There is a high likelihood that this customer is not going to pay his next bill. There is a 70% chance he is going to call about a bill that was too high. This person is the most important	Precision medicine: this customer has this type of lung cancer and most likely drug X will work best for her. Prediction: given her 40 vital signs in the last four hours, there is a 90%

	customer in his social network	chance this patient will go into respiratory failure in the next 48 hours.
Proposed action	Send this customer an email to tell him about the bill that was too high and offer forgiveness if he goes on another plan	Treat this patient for her cancer with drug X. Test her for peanut allergy. Intubate her in the ICU before she goes into respiratory failure. Call an investigator because it seems this patient has both male and female diseases, maybe fraud?

The Enterprise Knowledge Graph

The *enterprise knowledge graph* is one of the more intriguing applications of graph database technology today for two reasons, both of which became immediately obvious at the 8th DBpedia Community Meeting in Sunnyvale, California, in 2016 (wiki.dbpedia.org/meetings/California2016).

The first is that the forerunners in this space represent the very vanguard of IT companies today. With attendees from Google, Yahoo, Microsoft, and LinkedIn all sharing the particulars of their respective knowledge graphs, the worth of such graphs to the enterprise—including small and midsize organizations—became readily apparent. Note, however, that pharmaceutical companies, hospitals, banks, and intelligence agencies are also following this trend and creating their own knowledge graphs on a smaller scale.

The second reason is subtler, yet perhaps even more compelling than the first. As representatives from these organizations spoke about their different knowledge graphs, it became clear that this term—like many in the data ecosystem—has a variety of connotations that vary depending on its use in organizations. Not all knowledge graphs are the same; there seem to be three distinct categories:

- *Internal operations knowledge graph*. Best conceived of as a "company brain," this knowledge graph focuses on integrating an organization's assortment of people, skills, experiences, materials, essential company databases, and projects, which greatly improves its self-knowledge and thereby yields competitive advantage. Compiled from combing through myriad databases, including those for human resources, emails, and manifold other sources, this knowledge graph provides the foundation for a rapid, detailed assessment of what knowledge and skills a company has at its disposal—and their relation to one another.
- *Intermediary products and services knowledge graph*. This graph is designed to create better services and is extremely specific to an organization's industry, line of business, and area of specialization. For example, Google's and Yahoo's search engine endeavors mandate that they collect knowledge about every entity or subject in the world, so they can offer the most relevant, revealing information to their users. LinkedIn's knowledge graph, on the other hand, details people's professions, resumes, and career opportunities.[1] Again, the relationships between these nodes are paramount.
- *External customer knowledge graph*. This graph is best thought of as providing as much detail about customers as the first does about the organization itself. Offering the typical

360-degree view of an organization's clientele,[2] this knowledge graph is dedicated to every point of interaction and piece of information a company can gather about its customers. Typically, the data sources for this graph come from various company silos pertaining to customers; the graph itself highlights the relationships between these pieces of data.

The cardinal point of commonality between these three graphs becomes magnified when we consider the effect they produce on each other. As with all graphs, the underlying value is in identifying the way each of the data nodes relate to one another. Determining those relationships successfully, sustainably, and in a consistent and well-governed manner that identifies points of similarity that might otherwise remain unnoticed requires a *linked enterprise data* approach fortified by semantic graph technologies. The true value of this proposition is fully realized when we can link these respective knowledge graphs to one another for an interminable number of use cases. In this way, the connections between the first knowledge graph can inform the third by delineating which of the former's processes can best serve particular customer needs or behaviors. Conversely, the latter can illustrate how to tailor operational methods to advantage both customers and the company itself. The possibilities are endless and hinge on the linked data methodology of connecting different data types and attributes.[3]

Corporate Brain Knowledge Graph

The first knowledge graph is the most accessible of the three and a suitable starting point for organizations. Usually, the core entities here are people and projects. The objective is to improve operational efficiency by building a graph using as much detailed information as possible about the personnel resources and internal knowledge the company has. Human resources databases can provide insight into relevant skills and experiences for certain tasks. Email databases—which contain unstructured content but are partly structured in terms of how they interrelate with organizational databases—can identify which employees are proficient at verbal communication and in which particular areas. Project databases contain invaluable information about employees' impact on projects, such as who the most suitable leaders are, what their levels of efficiency and rates of completion are, and who works well together and how. Although product databases are related to the second knowledge graph (which specializes in products and services), they still incorporate a bevy of information about the internal creation of products related to how they were formed, by who, their descriptions, and which employees are most knowledgeable about them.

The practical applications of this internal knowledge graph are integral to effective and efficient operations in the long term. For example, this graph is useful for impact analysis specific to which employees know what about the company and the effect that their leaving or moving might have on how it functions. It also provides solutions for restructuring in a way that is most meaningful for operational efficacy in the event of such circumstances. Still, the most beneficial use of this knowledge graph probably lies in its ability to conduct social network analytics on an organization's employees to see who is most appropriate for dealing with a particular customer. Similarly, it can illustrate which personnel would be most apt at implementing specific strategies—especially recently developed, untested ones. A well-known company with an international presence in IT learned that it could substantially improve its ability to procure and maintain contracts by utilizing personnel who had professional relationships with prospective clients. All this information is detailed in the internal knowledge graph, which denotes employee relationships, skills, and abilities.

Furthermore, this initial knowledge graph effectually provides the foundation for the other two because it is the basis for advantageously facilitating both services and customer relationships. The

corporate brain knowledge graph enables organizations to begin with the data they already possess and expand according to their own specific business domains, eventually culminating in valued customer data. It offers a blueprint for managing operations in a tailored method dependent on an organization's own unique personnel features.

Intermediary Products and Services

Perhaps the most quintessential of the three knowledge graphs, the domain-specific intermediary second variety was widely discussed—and revered—at the Sunnyvale meetup for its salient impact on business value. The major IT players describing their knowledge graphs that afternoon focused on their ability to achieve business objectives by helping them do their jobs better—that is, improve the services they offer. Depending on what those services are, this knowledge graph takes immensely different forms for different organizations.

LinkedIn's knowledge graph, for instance, contains multiple nodes about people's careers—their jobs, skills, educational backgrounds, and both current and previous places of employment. Linking this information in a semantic graph lets LinkedIn users identify additional organizations and employees of other employers with similar backgrounds as needed. This knowledge graph is entirely predicated on improving professional networking via social media. The actual data it contains is distinct from that found in Google's knowledge graph, which pertains to details about any assortment of objects, people, or things so that it's users can attain apropos information regarding a specific search.[4] For each example, the data found in this sort of knowledge graph directly affects its organization's ability to service its customers.

In this regard, this knowledge graph offers an effective segue between operations and customers because it provides the means for obtaining (and retaining) customers. It is akin to a product in that it considerably enhances the mechanisms by which customers interact with the company deploying this semantic technology. Furthermore, it keeps all this relevant information in a single place that is easily traversed on demand. The domain-specific nature of this graph is likely most comprehensible as a digital version of a bill of materials, all of which are linked together and available to users to determine the correlation between different features for products and services. The main difference from the other two knowledge graphs discussed here is that this knowledge graph applies those individual nodes, or materials, to specific services (or examples of services, such as which employees have architectural experience for LinkedIn's graph, for instance) related to a particular customer use case (or search).

Customer Knowledge Graph

This third knowledge graph is best conceived of as a connection of all the data that provides the proverbial 360-degree view of a customer. These data stem from each point of interaction with a customer, from initial marketing efforts to ongoing customer support issues.

Initially, it requires gathering data from the various points of interaction in a customer's experience with a specific company. That data is then connected via the linked data approach in a semantic graph. The pivotal element of learning, however, is facilitated in two chief ways. The first of these is based on the data themselves and is rule-based. This learning requires simple reasoning and involves determining basic patterns gleaned from customer data about, for instance, whether a customer is likely to pay his or her bill, churn, or possibly seek a supplementary product or service. It involves establishing and adhering to rules regarding customer behavior validated by data.

The second form of learning via predictive analytics mechanisms for artificial intelligence (AI) and machine learning is much more statistically oriented. Regardless, with the former method, AI

algorithms find patterns in previous data and utilize them as the basis for predicting future data (results) about customer behavior. The combination of these two forms of learning (both plain data and AI inferences) issues layers of learning applicable to customer behavior. The result is a much more nuanced, contextualized awareness of an organization's customers, their behaviors, and the best way to effectively service them. For example, it would behoove a consultant company to compile data about all its different clientele involving everything it ever did for them and the things its clients wanted. The contextual codifications of such a graph are highly specific and should include budgetary information, the current state of the company and its state at the time it was consulted, which individuals have authority to sign off on projects and get them approved, and so on.

Connecting this Knowledge

The granular customer data from the customer knowledge graph supplies an interesting data source rife for exploitation when analyzed—and acted on—in conjunction with data from the other two graphs. This fact is especially convincing when linking the external customer data with the internal operations data. To continue the consultant use case, the relationship between the internal party of the consultancy and its effect on its customers as demarcated in the external graphs delivers all sorts of insight into which tactics and methods were most beneficial, between whom, and how to build on such a paradigm so that the entire organization (meaning other consultants) improves its performance from this knowledge.

More pragmatically, linking these graphs identifies the most appropriate candidate to interact with potential and current clients, while evincing which approaches have demonstrated effectiveness in doing so. Therefore, organizations can ascertain how their operations can best influence customers by actually learning from the outcome of those operations on real clients. This knowledge can form the basis for developing features for existing products or services and indicate which new products or services can sate customer needs based on verifiable, factual data. It can dictate strategies and more time-sensitive tactics while offering an overall direction for the company to achieve business objectives. The knowledge gained from linking the different graphs can also pinpoint relevancy between competitors and competitors' solutions when such data is included. This insight is also instrumental in planning employee relationships with others—either additional customers or particular companies. All these use cases deliver tangible business value by creating a conduit through which to monetize data by influencing future actions.

Being able to perceive these connections among knowledge graphs to obtain these boons depends on several factors. Organizations must be able to understand the correlation between data of various types and structures, as well as how they relate among the differing graphs. Moreover, they must be able to do so in time to seize on opportunities at the current pace of business. They need to ensure that the preparation work for curating and making such differential data integrate with one another is an ongoing evolution as opposed to a manual process that is reconstructed every time requirements or data sources fluctuate. In short, organizations need a linked data approach—what is referred to as linked enterprise data—to effect these advantages at scale in a way that meaningfully impacts their investments in data-driven technologies.

Linked Enterprise Data

Linked enterprise data is the nucleus of the aforementioned knowledge graphs.[5] This technology borrows its fundamental concepts from the notion of *linked open data*. The chief distinction

between the two is that the latter involves external, publicly available sources, whereas the former typically revolves around more internal and proprietorial data. Thus, linked open data functions as a precedent of sorts for linked enterprise data. Linked data (whether for the enterprise or the general public) is swiftly gaining traction throughout the public and private sectors as preferable to *master data management* and the conventional silo approaches of relational options. The semantic technologies upholding linked data are specifically designed to integrate data regardless of type, schema, or any other traditional concern so that **users** can exchange data between sources or repositories to ultimately connect them. It heralds the end of silo culture and the end of the typical schema limitations that laboriously underpinned it.

Knowledge graphs can exploit this approach by making all enterprise data exchangeable throughout an organization. Initially, users would simply need to mirror all their important databases—specifically, those that appertain to the three types of knowledge graphs in a Resource Description Framework (RDF) graph, which is commonly known as a *semantic graph*.[6] Ideally, all enterprise data should be linked in RDF, but organizations starting out can simply replicate relevant relational data into this format. From an implementation perspective, the various database administrators who oversee these individual relational databases can be tasked with maintaining the replicated RDF versions as well. Doing so enables organizations to transition from containing data in numerous silos to *linked data repositories*. As this moniker implies, such repositories are connected and render obsolete individual silos and their time-consuming, arduous schema requirements, which are provincial as opposed to enterprise-wide.

However, the crux of the linked data approach and the means by which it either succeeds or fails is using the same words for the same things, which requires a careful synthesis of semantic technologies in the form of *classifications*, *ontologies*, and *terminologies*. This synthesis begins with the underlying terminology: using the same words for the same things necessitates creating and enforcing enterprise-wide definitions for them. Taxonomies and classifications are instrumental in this regard, creating consistent hierarchical definitions and layers of meaning for terms throughout the enterprise.[7] Ontologies replace traditional schema in a naturally evolving way predicated on uniform standards created in part by the W3C. These ontologies evolve to incorporate new data types and business requirements by simply expanding those standards to include whatever novelties are required, so that all data can interrelate regardless of their location. As previously mentioned, this data is linked in a semantic, RDF graph once these standards are erected and maintained by uniform terminology, taxonomies, and ontologies.

Universally Identifiable URLs

The fundamental difference between the linked data approach and the more commonly found relational methodology becomes manifest with a cursory comparison between the two. In the latter, people, customers, products, and any other business object are assigned random numeric values that are parochial and meaningful only to the particular silo in which they exist. Attempting to link these silos requires a significantly time-consuming reworking of not only schematic concerns but also those of the identifiers themselves. When we consider all the respective databases necessary for any of the three knowledge graphs (which might include silos for email, personnel skills, salaries, and other appropriate variations of these necessities), doing upfront preparation work becomes all but impossible to effect competitive advantage in a truly time-sensitive manner.

The linked data methodology readily ameliorates these concerns by giving every object in a semantic graph a universally identifiable URL. This URL is the same wherever data is located throughout the enterprise—whether dumped in a data lake or located in a particular repository for

a certain purpose. In fact, the rendering of objects with company-wide, uniform URLs is one basic difference between semantic graph databases and nonsemantic, NoSQL graph databases. Every object in an RDF graph (products, people, customers, and so on) is a URL, which is why they can be connected; linking these URLs is the fundamental concept behind the Semantic Web (World Wide Web).

The other core difference between RDF graphs and NoSQL property graphs is that the former focus on the edges or connections between objects,[8] whereas the latter are dedicated to objects' names or properties. The semantic graphs inherently understand relationships and connections between even disparate types of data much more intuitively, as well as quickly and with less data preparation work. Furthermore, the URL-centered approach of semantic graphs renders them innately machine readable, a fact that is taking on renewed importance with advances in artificial intelligence, the Internet of Things, and some of their more prominent applications in autonomous vehicles or personal digital assistants.

Supported by the aforementioned Semantic Web techniques, the linked data methodology of knowledge graphs is an excellent means of implementing measures for data quality—which takes on additional importance when attempting to connect all the data in an enterprise. Issues of duplication, ambiguity, updates, and others[9] are greatly simplified by the fact that all the data in a knowledge graph has a unique, enterprise-wide identifier that is easily read and understood by machines.[10] These data quality concerns are part of the reason it is so difficult to connect silos with other approaches. However, because the knowledge graph method uses the same words for the same things (which means that it relies on URLs that consistently have the same meaning wherever they're found or deployed), those terms—those URLs—are inherently queryable for improved analytic insight and celerity of use.

Deriving Knowledge from Insight

The queryable nature of URLs, which is the crux of the linked enterprise data approach of knowledge graphs, is directly attributable to their ability to transmute insight into established knowledge. The principal way such graphs transform data into proven facts is by taking analytics a step beyond its conventional utility. Most organizations predominantly view the output of analytics as information or "answers" to questions, which are ends in themselves. The highly queryable nature of knowledge graphs, however, surpasses this utility by enabling users to input the results of analytics back into their stacks. In this case, those analytic results are the foundation of concrete knowledge, which is then further used to pinpoint accuracy for additional analytics.[11] Each subsequent wave of analytics only furthers this knowledge, which contributes to the expedient, targeted results provided, for example, by search engines or other applications.

A return to the consultancy use case clarifies this principle while hinting at the depths of understanding that knowledge graphs purvey. The project databases for these types of organizations contain detailed data about the involvement of different employees for respective customers, as well as the input of the company as a whole. Such data, when linked together in a knowledge graph, can deliver insight into the individual effectiveness of these employers. Moreover, these computations can also produce this effect for any number of combinations of employees who might be working together on assigned projects.

By gauging whether projects were completed, if they were done so on time, whether they were done to the customer's satisfaction, and other relevant facets of project management, a consultant company can determine numeric ratings for its employees (or their combinations). Going forward,

these ratings can be input back into the knowledge graph and its repository to determine future analytics regarding consultancy work. Additional queries could grant insight into how much money was generated from those with the top ratings; conversely, they can denote how much money was lost due to the interactions of those with the lowest ratings. Each result of a different query can be input back into the graph and form the basis of further queries rooted in the knowledge of what could be a perpetual analytics process.

The very nature of the queries facilitated by knowledge graphs is tremendously affected by the visual aspect of graph analytics, which is based on an understanding of the connections between data elements. Competitive visualization mechanisms for semantic graphs, for instance, enable users to intuitively point and click, drag and drop, and highlight data they can see to illustrate those connections the semantic technology gleans. Frequently, these visualization mechanisms only require users to indicate which data elements they wish to obtain information from (as well as in what way) prior to actually writing the formal query automatically. In many instances, code is not required, enabling users to parse through their data much more intuitively to determine relationships and answers to questions they wouldn't otherwise think to ask. The ensuing contextualization gained from this approach improves the value obtained from queries, especially when users can see the ingoing and outgoing links for any concept of interest to a particular business function.

Overall, enterprise knowledge graphs and their requisite linked enterprise data methodology are swiftly gaining credence throughout the data landscape. The utilitarian nature of these technologies, however, is far too pervasive to be reserved exclusively for the largest, most well-funded organizations in the IT space (such as those identified at the Sunnyvale conference). Small and midsized organizations can significantly enhance their ROI on data-driven technologies by leveraging these graphs to improve their operations, products and services, and overall cognizance of customers. This reality signifies the true utility of the underlying linked data approach at the core of these graphs—they are applicable to any organization and create the same value regardless of the scope or focus of the company deploying them.

References

1. Q. He, "Machine Learning in LinkedIn Knowledge Graph," LinkedIn, 26 Aug. 2016; bit.ly/2fwUQKF.
2. R. Bibb and E. Gehm, "The 360-Degree View," DestinationCRM, June 2001; bit.ly/2ylE0G4.
3. "What is Linked Data," W3C, 2015; www.w3.org/standards/semanticweb/data.
4. O. McCorry, "When Google Knowledge Graph Meets Healthcare," TechCrunch, 11 Mar. 2015; tcrn.ch/2fvzVYb.
5. M. Galkin, S. Auer, and S. Scerri, "Enterprise Knowledge Graphs: A Backbone of Linked Enterprise Data," ResearchGate, May 2016; bit.ly/2xi5s9l.
6. *Resource Description Framework (RDF)*, W3C recommendation, 2014; www.w3.org/2001/sw/wiki/RDF.
7. "LinkedIn Knowledge Graph Enriches Data Value," InsideBigData, 31 Mar. 2017; bit.ly/2hf6UzU.
8. M. Allen, "What Are the Differences Between a Graph Database and a Triple Store?" *Quora*, 1 May 2014; bit.ly/2wqEWeZ.
9. S. Judah, "Data Quality Improvement," Gartner blog, 17 Oct. 2014; gtnr.it/1zhUahv.
10. G. Ross, "An Introduction to Tim Berners-Lee's Semantic Web," TechRepublic, 31 Jan. 2005; tek.io/2xdwfC2.
11. Q. He, "Building the LinkedIn Knowledge Graph," LinkedIn, 6 Oct. 2016; bit.ly/2dB4JFS.

Jans Aasman is the CEO of Franz Inc., the leading supplier of commercial, persistent, and scalable graph database products. His research interests are in telecommunications, specializing in intelligent user interfaces and applied artificial intelligence projects. Aasman started his career as an experimental and cognitive psychologist, and was a part-time professor in the industrial design department at the Technical University of Delft, the Netherlands. He received a PhD in cognitive science from The University of Groningen, with a detailed model of car driver behavior using Lisp and Soar. Contact him at ja@franz.com.

The enterprise knowledge graph has emerged as one of the most useful applications of graph database technology today. Manifest in three different varieties for internal operations, products and services, and customer 360-views, respectively, these graphs hinge on a linked enterprise data approach to determine relationships between graphs and individual nodes of data. Buttressed by semantic technology standards for ontologies, RDF graphs, and taxonomies, enterprise knowledge graphs provide a range of opportunities for contemporary enterprises when used alongside additional modern developments in artificial intelligence and visual querying.

knowledge graph, graph database, artificial intelligence, semantic technology, machine learning, SPARQL, NoSQL, RDF, linked data, ontology, taxonomy