CAMBRIDGE SEMANTICS

Six Essential Requirements for the Scalable Knowledge Graph Platform

Published Date: February 15, 2021
Ben Szekely, Co-Founder and Chief Revenue Officer
Cambridge Semantics Inc.

Introduction

As a concept, framework, and instrument of production, the knowledge graph phenomenon has become more pervasive throughout the data ecosystem than ever before. Its mounting influence is reflected in enterprise-scale deployments of the most ubiquitous technology companies in existence (including Google, Facebook, and LinkedIn) and the enterprise data strategies of the Fortune 1000, aligning to and influencing the foremost trends impacting the analyst community.

The Gartner Group hails knowledge graphs as a reasonable means of preparing data for machine learning and Artificial Intelligence, as well as an expression of AI itself via graph techniques. It recently announced knowledge graph technology had reached the peak of AI expectations. At the same time, Forrester Research tacitly sanctioned this development by introducing a new recurring vendor report exclusively dedicated to graph data platforms.

Consequently, we see an almost endless array of graph vendors, use cases, and variations of knowledge graphs emerging in the marketplace, each claiming to be authentic representations of this valued approach for integrating data and leveraging advanced analytics. With conflicting definitions abounding, it is essential to disambiguate them to understand what is actually required for enterprise-level deployments. To motivate the discussion, consider our working definition of a knowledge graph:

> *A Knowledge Graph is a connected graph of data and associated metadata applied to model, integrate and access an organization's information assets. The knowledge graph represents real-world entities, facts, concepts, and events as well as all the relationships between them yielding a more accurate and more* ***comprehensive representation*** *of an organization's data.*

The *knowledge graph platform* makes the knowledge graph itself the way to access, integrate and analyze the data either directly by end-users or as a feed to operational AI and analytics applications.

Thus, scale represents the primary point of distinction. The success of knowledge graphs has advanced beyond the point in which small applications alone are useful. Today's knowledge graph deployments often underlie strategic initiatives such as Data Fabrics and support advanced analytics use cases both comprising up to and beyond:

- **Hundreds to thousands** of data sources.
- **Hundreds** of use cases.
- **Hundreds of billions** of RDF triples (encompassing nodes and edges) across the graph.
- **Dozens** of users with varying backgrounds building and managing the graph.
- **Thousands** of users analyzing data and accessing data from the graph

Such knowledge graphs require more than just a graph database - the de facto starting place for many first-time innovators in the space. To be practical and successful a complete, scalable knowledge graph platform is needed that addresses the following six requirements:

- **Any Source:** True enterprise knowledge graphs are built and maintained from many sources without compromise and with grace, regardless of the sources' structure variation, format differences, respective data models, or other distinctions at origination.
- **Performant Loading and Efficient Storage:** Enterprise-scale is impossible to achieve without automating—and expediting—loading source data into the graph and providing options for efficient storing of graph data.
- **Flexible Deployments:** Knowledge graphs provide limited utility if enterprises are unable to deploy them wherever they're most advantageous, whether on-premise or in the cloud.
- **Interactive Query at Scale:** Entire knowledge graphs must be traversed interactively to support queries for analytics, data preparation, and data access - leveraging MPP architecture.
- **Easy Interfacing with Open Standards:** Knowledge graphs should use open standards to enable all users and applications (internal and external) to interface with them, regardless of technical capabilities.
- **Granular Security:** Mainstays of security, data governance, and regulatory compliance are essential to scaling knowledge graphs across organizations.

Not all knowledge graph platforms or solutions fulfill these requirements. Some are less expensive than others or take less time to get started. But for addressing the complexity, breadth, and scalability requirements of *leveraging a single investment innumerable times across the enterprise*, knowledge graphs must meet these six requirements for companies to build, manage, and query them for any use case.

Any Source

The first requirement of platforms meeting the scale of enterprise knowledge graphs is the capacity to flexibly integrate any data source, regardless of variation. This necessity involves easily connecting to sources and fluidly changing those connections—and graphs' underlying data models—for unanticipated questions. Subsequently, users can remodel graphs, integrating structured data with the surplus of unstructured and semi-structured data flooding the enterprise. Scalable graphs also accommodate data of any type, format, or data model, whether internal or external to the enterprise. While many platforms require organizations to recalibrate data models, scalable knowledge graph platforms incorporate existing ones within their own. The model naturally evolves from existing sources so organizations can start with what they have and incrementally add to it as they amass more knowledge. In this manner, the scalable knowledge graph platform accommodates the inherent complexity and uncertainty in integrating many data sources together.

This versatility naturally extends to how businesses connect to sources. Some, like external news reports for financial trading, are better left in place and accessed as virtual sources. Others, like data warehouses with sales information, warrant tighter integrations with local data and their metadata catalog. It shouldn't matter if sources are on-premise, in the cloud, or in hybrid clouds. For example, knowledge graphs should support the array of unstructured and structured sources impacting hedge fund opportunities. They should readily incorporate new sources like weather data for ad-hoc questions about meteorological trends impacting real estate without lengthy delays in changing the data model, which frequently happens with other platforms. Veritable enterprise-scale graphs future-proof the enterprise, reducing user risk of adopting them.

Performant Loading and Storage

Implicit to the above requirement for source inclusiveness is performant loading and storage of graphs at enterprise scale. Automating loading is vital for expanding graphs across organizations so users can rapidly onboard and connect data for sizable use cases, like clinical trial analytics. In this space, enterprise options automate parallel graph pipelines that create and persist graph data into efficient commodity storage for subsequent in-memory loading or swiftly load desired data directly into memory just-in-time.

With the knowledge graph platform, a metadata-driven catalog is central to managing the lifecycle of the knowledge graph, including transforming any data into a uniform data model. The catalog contains rich metadata descriptions that provide the foundation for metadata-driven graph transformation (ETL or ELT) from any source, like the above sources or other graphs. After transformation, data are harmonized via automated queries specializing in detecting relationships between data for crucial contextualization of any use case. The catalog is fundamental to the toolkit supporting increasing automation of knowledge graph processes and AI and machine learning for data ops, analytics, and model development.

Storage efficiency is pivotal to fulfilling this requirement, too. It is usually a best practice to only store what organizations need; storing entire graphs can increase data ops and storage costs. When assembling huge datasets for clinical trials, for example, the objective is to load fast and only store what is necessary to optimize analytics.

Deployment Versatility

The liberty to deploy knowledge graphs wherever is most beneficial to the business and directly impacts cost. The most affordable operational environments typically do not require additional investments. The scalable knowledge graph platform should give users the flexibility of leveraging commodity VMs on-premises and in any cloud type applying standard deployment mechanisms like Kubernetes. Clouds include on-premise hybrids, public or private options, or public-private hybrids.

Kubernetes' appeal is the ability to dynamically orchestrate the provisioning of massive workloads at scale—like real-time analytics of sales during events like Black Friday or Cyber Monday—then spin them away just as quickly when desired. This approach can also avail organizations of fleeting pricing opportunities in today's multi-cloud reality.

Regardless of their environment, the most cost-effective knowledge graph deployments utilize existing infrastructure. The knowledge graph platform must complement organizations' data platforms, not discard them. For instance, companies operating scalable data lakes in AWS should select a knowledge graph platform that works with or enhances that investment, instead of replacing it.

Interactive Query at Scale

The capacity to seamlessly query an entire graph in real-time represents the crux of the scalability issue distinguishing enterprise-class knowledge graphs from those supporting limited and finite numbers of use cases, datasets, users, and nodes. The latter will never offer enterprise-scale utility if users are unable to expeditiously traverse them for any use case across or within departments. Low latent query speed universally applies to any number of recurring knowledge graph use cases, particularly the business value derived from analytics. For instance, research and development teams can run interdepartmental analytics across sales, marketing, and customer support units to discern which new features for products and services will be most worthwhile to customers in future iterations.

Real-time queries also enrich the ELT or transformation steps required to efficiently and iteratively load and transform data and make them available at enterprise scale. Query rapidity is crucial to accessing data wherever they are in graphs for data discovery and other purposes. Timely querying considerably helps remodeling data efforts and is table stakes for enterprise knowledge graphs since, unless query speed matches or exceeds the growth of knowledge graphs, they'll never successfully scale to rapidly address new use cases and unanticipated questions.

The most effective approach for facilitating these boons is massively parallelizing (MPP) each query across many cores in a cluster and in-memory techniques, which very few knowledge graph engines can achieve. In particular, the query engine must support OLAP-style queries needed for analytics and data integration. Some knowledge graph platforms support fast response times for limited varieties of hand-coded queries, most often with OLTP design-points. To achieve data fabric scale integration and analytics, the graph engine must excel at graph algorithms, traversals, data science primitives, and graph algorithms alike - often combined in the same query.

Easy Interfacing

Ultimately, the merit of any scalable knowledge graph is based on its ease of use and ability to speed and democratize access to business-ready data products across the organization. Non-technical users, developers, internal applications, and external applications should all be able to readily interface with the platform, particularly its functionality and accessibility.

While multiple graph standards are emerging, the most accessible knowledge graphs employ semantics-based standards describing data and relationships in business terms understood across departments. Understanding these graphs' content reduces IT involvement and the need for esoteric coding skills to interact—unlike using other types of platforms. Led by RDF and OWL, these semantic standards are essential for systemic interoperability between graphs, use cases, vocabularies, and even entire organizations (such as subsidiaries, partners, or supply chain networks).

Semantic graphs are designed for collaborating and exchanging data. Throughout the enterprise, users can assemble and employ the same knowledge graph—from just one investment—for individual needs, especially when governance and security concerns addressed in the next requirement are met. Semantic graph’s ease of use is enhanced by knowledge graph platforms that support automatic query generation, enabling ad-hoc data exploration for even laypeople. Powered by intuitive, browser-based experiences similar to those of self-service Business Intelligence tools, such experiences are delivered through simple, visual means accessible to any user.

The very same query generation capability should offer REST and ODBC/JDBC endpoints for external applications so users can access the knowledge graph from popular analytics tools (Qlik, Power BI, etc.) not yet part of the graph-aware ecosystem. These APIs afford easy access for three main knowledge graph consumption patterns: analytic insights, operational analytics, and custom applications built on the knowledge graph.

Many knowledge graphs lack this ease-of-use requirement for scalability. These platforms may not employ standardized models in business defined terms, creating an undue reliance on IT and highly specialized developers to understand them and their data. Regardless of how advanced such engines might be for analyzing this data, the surfeit of coding required to load, query, and work with this data circumscribes its use to technology sophisticates—preventing it from scaling throughout the enterprise.

Enterprise Security

True enterprise scalability pertains to more than just querying speed and democratization of use, although these concerns are critical. A large part of readying knowledge graphs for the enterprise pertains to fundamentals of security, data governance, and regulatory compliance. Platforms that fail to account for these crucial mainstays will never support organization-wide use cases. Fine-grained access restrictions are required to reinforce these aspects of data management. Such mechanisms specify which users can access, query, and update which parts of the graph and how they can use them. The previously discussed knowledge graph management catalog facilitates this access via metadata-driven security against the underlying graph engine. When querying graphs for diabetes patients, for example, users will only get results from the sub-graphs they have access to.

This granular security has significant data governance implications. It reinforces role-based access for data privacy, while its applicability to PII is perfect for regulatory compliance. Organizations can specify access based on whatever compliance factor is relevant. Maintaining these security and data governance requisites is paramount to implementing scalable knowledge graph applications. For example, they allow a parent company to grant access to its knowledge graph to subsidiaries for interoperability while still preserving data privacy and data integrity.

Conclusion: Better Safe than Sorry

A knowledge graph platform will only scale to the extent it fulfills these six requirements. Leveraging any source and enabling performant loading and storage is necessary for quickly building comprehensive graphs. Real-time queries and easy interfacing maximize end-user value. Flexible deployments and granular security access are required to manage graphs successfully.

Meeting these requirements enables organizations to build knowledge graphs fast and sustainable enough to support any enterprise use case. Selecting a solution based on cost or availability that fails to meet these requirements will only impede the success and adoption of the knowledge graph, potentially derailing the entire initiative.

Author

Ben Szekely is co-founder and Chief Revenue Officer of Cambridge Semantics, creator of the industry-leading Anzo scalable knowledge graph platform. He has spent more than a decade working with global organizations to implement enterprise-scale knowledge graphs. Before founding Cambridge Semantics, Ben was a development and technology leader in IBM's Advanced Technology Office, focusing on next-generation data modeling technologies for analytic, big data, and digital transformation initiatives. Ben holds an MS in Computer Science from Harvard University and a BA from Cornell University.π

Cambridge Semantics Inc.

Cambridge Semantics Inc. is a modern data management and enterprise analytics software company. Our solutions transform siloed data into enterprise-scale knowledge graphs, revealing previously hidden insights, fueling pervasive analytics, and making previously unanswerable questions answerable.

Cambridge Semantics solution Anzo® is a scalable knowledge graph platform for modern data integration and analytics. Anzo dramatically simplifies and accelerates the integration, modeling, and blending of siloed data into insight-rich knowledge graphs at enterprise scale. Anzo is built on AnzoGraph®, the fastest & most scalable knowledge graph engine supporting data integration, graph algorithms, data warehouse-style analytics, feature engineering for Machine Learning, and more. The company delivers solutions that enable IT departments and business users across Life Sciences, Financial Services, Government, Manufacturing, and other industries to accelerate data delivery and provide meaningful insights across the organization at hyper-speed and scale.

Learn more at www.cambridgesemantics.com or contact us at info@cambridgesemantics.com