CAMBRIDGE SEMANTICS
THE SMART DATA COMPANY®

ANZO®

Using OLAP Knowledge Graphs for Data Integration and Harmonization

Sean Martin
Cambridge Semantics Inc.

Traditional Enterprise Data Warehouse (EDW) technologies have been used successfully in many settings to integrate and harmonize data so that analysts and users can reliably extract meaning from their large enterprise data-sets. But it is well-established that this approach comes with significant up-front and ongoing costs, enormous risks of failure, and leaves large areas of enterprise data simply unaddressable (due to their complexity). They also lack flexibility, making them too slow to meet the ever-changing demands for data by modern businesses. In this paper, we argue that OLAP Knowledge Graph Technologies (GOLAP) are poised to address these concerns. To explain, we'll briefly review the way EDW technologies are deployed, their limitations, costs, and risks, in order to later contrast how emerging OLAP Graph Technologies have evolved to provide solutions to these issues and are positioned to leapfrog the EDW for modern enterprise OLAP.

On the way, we will also briefly address an inadequate attempt at solving these problems, Hadoop-based Data Lakes and explain how neither of these approaches is fit for the current purpose. Finally, we'll describe GOLAP technology in more detail, showing how it resolves these concerns, and opens up a vista of new data-integration possibilities. OLAP graph technologies and the implementation of broad data integrations, known as knowledge graphs, can link a vast variety of entity types based on many sources of data. We will describe two methods for data integration using GOLAP and an example of a new methodology for data harmonization that is evolving to take great advantage of it.

Introducing Graph Technology

People already know that graph technology is extremely useful because it makes it easy to create, store and query the edges (connections or relationships) that connect vertices (graph nodes or entities – essentially what are presented in RDBMS as records) in a very intuitive manner. The most regularly used examples being the vast graphs representing entities and their interconnections in social networks like Facebook, Twitter, and LinkedIn where the benefit of the connections is self-evident. A key difference between graph and RDBMS databases is that with a graph system it is a simple query that will report all the ways in which two entities are connected (even through multiple vertex hops) whereas with RDBMS it is only possible to formulate SQL queries that test if a specific relationship exists and to repeat that process for each potential relationship to enumerate them.

Another graph technology usage that has entered the mainstream is the "knowledge graph", popularized for example as an improvement to Google's and similar search services, in which often loosely structured but connected information related to the results of a text search are instantly retrieved via a graph database query and displayed in an adjacent supplementary info-box. All of the Wikipedia data has also been turned into a knowledge graph called DBpedia which it is possible to query, and is finding its way into question-and-answer style products. Until recently knowledge graphs have not been widely used for the purposes of OLAP analytics style queries which tend to be complex and usually require reading and acting on far more data than the fast retrieval of a handful of related graph nodes and edges for search result augmentation.

It is also widely, if often vaguely, understood that it is possible for more sophisticated data practitioners to apply relatively "exotic" graph algorithms to extract or calculate all manner of information from data stored in a graph form. Popular examples of these are PageRank (centrality) or calculating how vertices are clustered, or various inference rule techniques where new data is created from existing data, or calculating the shortest path of connections between vertices, etc. Most recently, a vanguard of the wider data science community has begun to discover the predictive value in combining the results of these kinds of algorithms and graph queries as features in machine learning approaches and are reporting positive outcomes. A whole field is growing up around graphs and vertex embeddings, the transformation of graph data to a vector or a set of vectors from which predictive models can be formed.