

Hybrid Enterprise Knowledge Graphs

Peter Haase

metaphacts GmbH, Walldorf, Germany
ph@metaphacts.com

Abstract. Knowledge graphs have emerged as a key technology to enable enterprise data management, serving as an integration hub across enterprise data source. In this talk we present Ephedra, a SPARQL federation engine aimed at supporting queries across hybrid knowledge graphs involving both native graph data as well as a range of other data sources, data modalities, and data processing techniques. We will present practical examples and experiences with hybrid enterprise knowledge graphs in real-world applications in various industries.¹

Knowledge graphs established a solid position in the enterprise world, serving as a central element in the organizational data management infrastructure. Knowledge graphs are becoming both the repository for organization-wide master data (ontological schema and static reference knowledge) as well as the integration hub for various legacy data sources: e.g., relational databases or data streams. Typically, a core knowledge graph is used to represent and materialise semantic descriptions of key entities and their relations in RDF, queryable via SPARQL. However, in many practical knowledge graph use cases there is a need to address hybrid information needs. Such needs can be characterized by the following dimensions:

- *Variety of data sources.* There is often a need to integrate data stored in several physical repositories. These repositories can include both native RDF triple stores as well as datasets in other formats presented as RDF (e.g., a relational database exposed using R2RML mappings).
- *Variety of data modalities.* Graph data in RDF often needs to be combined with other data modalities, e.g., textual, temporal or geospatial data. A SPARQL query then needs to support corresponding extensions for full-text, spatial, and other types of search.
- *Variety of data processing techniques.* Retrieved data often has to be further processed using dedicated domain-specific services: e.g., graph analytics (finding the shortest path or interconnected graph cliques), statistical analysis and machine learning (applying a machine learning classifier, finding similar entities using a vector space model), etc.

The main motivation for this work comes from our experience with the metaphactory knowledge graph management platform [1], which is used in a

¹ Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

variety of application domains (e.g., pharmaceuticals, financials, engineering and manufacturing, and IoT). Typical application scenarios often require dealing with a multitude of the above-listed dimensions simultaneously. To support this, the metaphactory platform can connect to various different RDF stores as well as non-RDF sources virtually integrated and exposed via a SPARQL interface. To be able to interact with multiple data sources using virtual data integration, we developed a SPARQL federation engine Ephedra [2]: In Ephedra, we adopt the SPARQL 1.1 federation mechanism using the SERVICE keyword, but broaden its usage to enable custom services to be integrated as data sources. In this way, a complex information request requiring access to several data sources can be expressed using a single query over a virtually integrated knowledge graph.

Ephedra defines a common implementation interface, in which interactions with external services are encapsulated: A custom service can be registered in the repository manager as yet another SPARQL repository and referenced inside SERVICE clauses in SPARQL queries. SPARQL graph patterns specified inside such SERVICE clauses are parsed to extract input parameters for a service call as well as the variables to bind the results returned by the service. The Ephedra SPARQL query execution strategy sends the sub-clauses of a query to corresponding data sources, gathers partial results, combines them using union and join operations, and produces result sets. Processing hybrid queries is transparent and performed in the same way as ordinary SPARQL queries.

In the talk, we will present examples of hybrid knowledge graphs in different industries involving a range of data and compute services virtually integrated:

- We will show how a core knowledge graph of physical devices (such as gas turbines) acts as an integration backbone for heterogeneous data sources including real-time sensor data. To be able to process relevant information in a timely way and achieve intelligent diagnostics, knowledge graph data is integrated with runtime information provided by relevant APIs on demand. The Ephedra federation architecture enables combining the materialized knowledge graph data with the data produced by the API services available.
- In another example, we show how a knowledge graph containing master data about products can be augmented with virtual relations that are not explicitly known or asserted, but virtually integrated from a machine learning algorithm, which provides relations such as similarity and returns as response the classification results.
- We will show examples that illustrate the power of Ephedra to effectively wrap data and compute services accessible via any REST API as a virtual endpoint in a SPARQL federation.

References

1. Haase, P., Kozlov, A., Herzig, D., Nikolov, A., Trame, J.: metaphactory: A platform for knowledge graph management. *Semantic Web Journal* (2019)
2. Nikolov, A., Haase, P., Trame, J., Kozlov, A.: Ephedra: Efficiently combining RDF data and services using SPARQL federation. In: *Knowledge Engineering and Semantic Web, KESW 2017, Szczecin, Poland.* (2017) 246–262