Accelerating Data Discovery with an Ontology-driven Tool for an Enterprise-scale Data Lake Environment

Satyajeet Raje[1], Karina Kervin[2], Nergal Issaie[2], Madhu Channapatna[2]

[1]Global Chief Data Office, IBM, Armonk, NY

[2]Global Chief Data Office, IBM, San Jose, CA

satyajeet.raje1@ibm.com, kkervin@us.ibm.com, nergal.issaie@ibm.com, madhuch@us.ibm.com

Abstract

A large overhead in the analytics process is the time required to find relevant data. We present an ontology-driven data discovery application, implemented over IBM's Cognitive Enterprise Data Platform (CEDP). CEDP contains a large collection of heterogeneous data assets from enterprise-wide data sources. The application accelerates the time required for data consumers to search and find data relevant for their analytics applications.

Introduction

Enterprise Data Lakes are becoming ubiquitous, as more organizations are opting for digitization and realize the role of data and analytics in making effective business decisions. The value delivered by such data lakes and warehouses is directly related to the ease with which data scientists and analysts can find and reuse the relevant data for their applications (Griffiths, 2009). Data repositories may be growing, but many data sets remain unused despite researchers' desire to reuse data (Curty, et al., 2017; Pronk, 2019; Yoon, 2016). User-friendly tools that reduce the time to search and explore data and accelerate reuse are essential to address this problem (Tenopir, et al., 2015). In this submission, we describe our implementation of an ontology-driven tool for data discovery over a large enterprise data lake.

Related Work

Enterprise Data Lakes are collections of data from disparate data sources and warehouses. Data lakes, unlike data warehouses, store data in heterogeneous formats from original sources and do not have a shared database schema. They often rely on external catalog applications that maintain required metadata required for organizing and governing the data in this data lake. These catalogs are mostly suited for data steward and data engineer personas, rather than an end consumer of data like a data scientist or analyst. Thus, finding data effectively within a data lake remains a major challenge. The end-user may not have knowledge of the specific data fields and schema, and, thus, needs the ability to search for data in natural language terms rather using database specifics. This requires a robust metadata conceptual layer that can span across the entire data lake.

Ontologies provide a powerful framework for representing this conceptual metadata layer as they provide the necessary flexibility to define relationships between concepts across the heterogeneous data sources. The search backend can then leverage these relationships that embody the background knowledge built into the ontology (Ramkumar & Poorna, 2014).

Use Case

We have implemented a data discovery application with an intuitive UI that allows users to explore and search data on IBM's internal enterprise data lake, called the IBM Cognitive Enterprise Data Platform (CEDP). CEDP provides a single place to store and find data within IBM with the goal of breaking down data silos to enable company-wide data reuse. According to Feilmayr and Wöß, "the ideal candidate for ontological development is a highly shared, large, and re-usable system," and this description certainly fits the IBM CEDP (Feilmayr & Wöß, 2016). As of Sept 2020, the data lake consists of data from over 300 sources, translating to about 25k data assets (tables) and 150k+ data fields (columns) for relational data alone.

The metadata for these assets is cataloged using IBM Infosphere Governance Catalog (IGC) product (see Figure 1). The catalog also consists of a business glossary with about

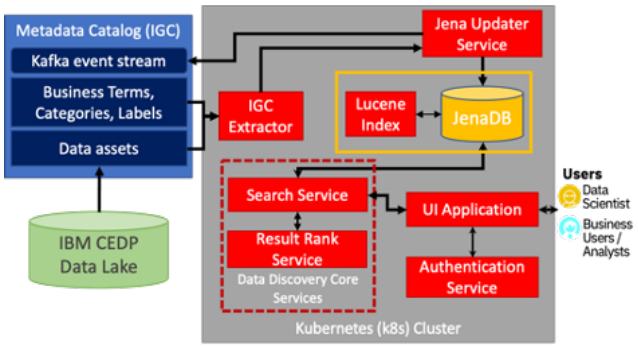

Figure 1. System components and data flow for CEDP Data Discovery

6000 business terms, maintained by data stewards. We linked these business terms together into a unified CEDP ontology. The data fields and assets are annotated with these business terms, thus providing the required conceptual metadata layer. Our Data Discovery application extracts and transforms this metadata layer from the catalog and creates an ontological representation in RDF. This is stored in a triple store (JenaDB) that the UI application interacts with.

Implementation

Figure 1 shows the overall system components and flow. The application uses JenaDB as the central triple store repository. The Jena updater service listens to the Kafka event stream from the IBM Infosphere Governance Catalog instance (IGC) to detect any changes to the glossary assets (business terms, domains, categories, etc.), the data assets (tables, columns, files, etc.) and relations between them. It triggers the extractor service that pulls data from the catalog, transforms it into an RDF representation and updates JenaDB. Users, once authenticated, can browse or search for data. The user queries are sent to the search service, where they are parsed and mapped to business terms in the CEDP ontology. The search service then executes a SPARQL query against JenaDB (and a linked Lucene index). The query returns potential data assets that match the ontology business terms that resulted from the user search. These are ranked using the ranking service before being sent back to the UI.

The Data Discovery tool is built with a microservice architecture. The microservice architecture allows distribution of the functional components into independent but orchestrated services. These services are deployed in a Kubernetes environment, and capacity for specific services can be scaled up or down as per scalability requirements. The architecture allows different services to be written in different languages. Another benefit of this architecture is increased resilience, as individual service failure does not impact the remaining application and can be restarted automatically.

These are key features implemented in the current version:

Translating natural language search tokens to business terms for search: We leverage the built-in Lucene indexing capabilities in JenaDB to index business term names and descriptions. The user can enter multiple search tokens in natural language. These are queried against the Lucene index to match to terms in the business ontology glossary.

Automatic query expansion using ontological relations: The search leverages the ontological relations between business terms (such as "synonym of", "is a", etc.) and OWL reasoner in JenaDB to automatically expand the query results. For example, query for 'email address' automatically returns results for all its descendants, such as 'primary email address', 'organization email', etc.

Result ranking using graph network centrality features: The ranking algorithm leverages the network topology of the graph ontology, considering both query relevance and global centrality of objects to rank search results. The query relevance component is based on the passed search concepts and concepts directly related to the search concepts, such as synonyms of the search concept. This portion of the ranking method prioritizes the search concepts and the concepts closest to the search concepts, particularly synonyms. The second component calculates the network centrality score of all data assets in ontology.

Work Area feature for users to save relevant data assets of interest: The work area feature allows the users to save relevant datasets of interest. The work area is persisted across user sessions allowing users to return to their work.

Conclusion

Based on activity and feedback collected through user surveys and NPS, the Data Discovery tool has provided significant benefits to the user community. We are also continuing to work on additional features, such as integration with the data access tool, fully natural language query interface, and integration with data quality tools. We are also extending the capability of the tool over a federation of catalogs.

The overall framework for the ontology-driven Data Discovery tool described here could be implemented over any archive that supports use of ontologies. It is our hope that the Data Discovery tool can be adopted in repositories in other organizations and domains to make the process of finding data for new research insights and innovations that much easier.

References

Curty, R. G. et al., 2017. Attitudes and norms affecting scientists' data reuse. *PLoS ONE,* 12(12).

Feilmayr, C. & Wöß, W., 2016. An analysis of ontologies and their success factors for application to business. *Data & Knowledge Engineering,* 101(2016), pp. 1-23.

Feilmayr, C. & Wöß, W., 2016. An analysis of ontologies and their success factors for application to business. *Data & Knowledge Engineering,* 101(2016), pp. 1-23.

Griffiths, A., 2009. The Publication of Research Data: Researcher Attitudes and Behaviour. *The International Journal of Digital Curation,* 4(1), pp. 46-56.

Madin, J. S., Bowers, S., Schildhauer, M. P. & Jones, M. B., 2007. Advancing ecological research with ontologies. *Trends in Ecology and Evolution,* 23(3), pp. 159-168.

Pronk, T. E., 2019. The Time Efficiency Gain in Sharing and Reuse of Research Data. *Data Science Journal,* 18(10), pp. 1-8.

Ramkumar, A. S. & Poorna, B., 2014. *Ontology Based Semantic Search: An Introduction and a Survey of Current Approaches.* Coimbatore, India, s.n.

Tenopir, C. et al., 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS ONE,* 10(8).

Wu, H. et al., 2017. Understanding Knowledge Graphs. In: *Exploiting Linked Data and Knowledge Graphs in Large Organizations.* Cham, Switzerland: Springer Nature, pp. 147-180.

Yoon, A., 2016. *Red flags in data: Learning from failed data reuse experiences.* Copenhagen, Denmark, s.n.