A Data Discovery Platform Empowered by Knowledge Graph Technologies: Challenges and Opportunities

Essam Mansour
Concordia University
essam.mansour@concordia.ca

ABSTRACT

In this talk, we present KGLac, a data discovery platform empowered by knowledge graph technologies, and highlights several open research challenges and opportunities.

Reference Format:
Essam Mansour. A Data Discovery Platform Empowered by Knowledge Graph Technologies: Challenges and Opportunities. In the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores (SEA Data 2021).

1 DEVELOPMENT AND OPPORTUNITIES

With the growing importance of data science and open data initiatives, thousands of machine-readable, structured, and semi-structured datasets are collected and made available via data discovery systems in the case of enterprise datasets or via data portals in the case of public datasets. Data portals are maintained, for example, by by governments, e.g., USA, Canada, and EU, organizations, such as WHO and WTO, and ML portals, such as Kaggle and OpenML. Existing portals and systems suffer from limited discovery support and do not track the use of a dataset and insights derived from it. Thus, data integration and enrichment are the primary responsibility of data scientists, who spend most of their time knowing where a relevant dataset exists, understanding its impact on a specific task, finding ways to enrich a dataset, and leverage the derived insights.

Data portals and search engines, such as Google Dataset Search, provide primitive search capabilities to find and download open datasets in different formats, such as CSV, JSON, and XML. Moreover, many organizations are encouraged to build a navigational data structure (data catalogue) to support data discovery [2, 4] or to use tools such as Amundsen. Unfortunately, these systems and tools suffer from limited query support and cannot find data items based on learned representations (embeddings). There is a need for an extensible set of effective discovery operations to find relevant data from their enterprise datasets accessible via data discovery systems or open datasets accessible via data portals.

Several methods were proposed to measure table relatedness [5], support table discovery [1], and find joinable tables [6]. These methods work in isolation from each other and from data portals and discovery systems. Thus, there is a need for data portals and discovery systems with a flexible query language and an extensible set of discovery operations. Moreover, existing data science platforms, such as MLFlow or Cloud AutoML, and tools, such as Jupyter Notebooks or Google Colab, should be able to communicate easily with these portals and systems.

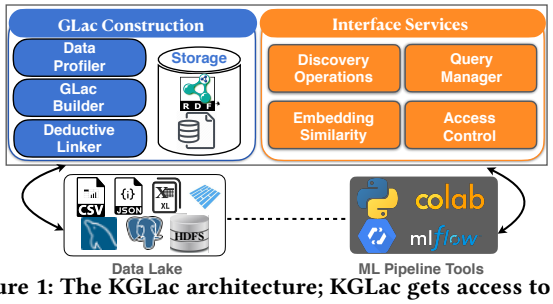

Figure 1: The KGLac architecture; KGLac gets access to a local data lake to construct GLac. Different ML pipeline tools can communicate with KGLac to facilitate data discovery.

The development of KGLac [3], as illustrated in Figure 1, poses research opportunities in various areas spanning data management and AI. These research opportunities cover (i) abstracting and capturing semantics from heterogeneous datasets, (ii) constructing decentralized knowledge graphs (KGs) for datasets, (iii) supporting inference and automatic graph learning to incrementally introduce and enhance the relationships among different nodes in the graph, and (iv) automating several aspects of data science including data preparation, augmentation, and insights analysis.

KGLac is supported by different methods for data profiling and representation learning (embedding) to capture metadata and semantics of datasets to construct a knowledge graph (GLac). KGLac provides an extensible set of data discovery operations implemented using SPARQL queries, and supports ad-hoc queries. KGLac enables automatic graph learning to advance functionalities, such as classification of similar data items, finding unionable and joinable tables, predicting shortest paths between tables, and inferring new relationships. We designed KGLac to be deployed on top of a data owner's data lake to enable efficient and extensible data discovery operations for data scientists who have access to the data lake.

REFERENCES

[1] Christina Christodoulakis, Eric Munson, Moshe Gabel, Angela Demke Brown, and Renée J. Miller. 2020. Pytheas: Pattern-based Table Discovery in CSV Files. *PVLDB* 13, 11.
[2] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. 2018. Aurum: A Data Discovery System. In *ICDE*.
[3] Ahmed Helal, Mossad Helali, Khaled Ammar, and Essam Mansour. 2021. A Demonstration of KGLac: A Data Discovery and Enrichment Platform for Data Science. *PVLDB* 14, 12.
[4] Fatemeh Nargesian, Ken Q. Pu, Erkang Zhu, Bahar Ghadiri Bashardoost, and Renée J. Miller. 2020. Organizing Data Lakes for Navigation. In *SIGMOD*.
[5] Yi Zhang and Zachary G. Ives. 2020. Finding Related Tables in Data Lakes for Interactive Data Science. In *SIGMOD*.
[6] Erkang Zhu, Dong Deng, Fatemeh Nargesian, and Renée J. Miller. 2019. JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes. In *SIGMOD*.

Published in the Proceedings of the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores, co-located with VLDB 2021 (August 16-20, 2021, Copenhagen, Denmark) on CEUR-WS.org.