Federated Query Processing over Linked Data

Muhammad Saleem[1], Muhammad Intizar Ali[2],
Ruben Verborgh[3], and Axel-Cyrille Ngonga Ngomo[1]

[1] Universität Leipzig, IFI/AKSW, PO 100920, D-04009 Leipzig
{lastname}@informatik.uni-leipzig.de
[2] Insight Center for Data Analytics, National University of Ireland, Galway
ali.intizar@insight-centre.org
[3] Ghent University – iMinds, Belgium
ruben.verborgh@ugent.be

Abstract. Due to the decentralized and linked architecture of Linked Open Data, answering complex queries often requires accessing and combining information from multiple datasets. Processing such *federated queries* in a virtually integrated fashion is becoming increasingly popular. This tutorial will explore the different approaches used for federated query processing over Linked Data. In particular, we will focus on query federation over SPARQL endpoints, Triple Pattern Fragments, and live Linked Data streams. State-of-the-art techniques will be practically demonstrated with examples along with hands-on experience and exercises to be carried out by the participants. By the end of the tutorial, participants will obtain hands-on knowledge of the federated query processing over Linked Data, understand the main differences between state-of-the-art systems, and be able to position these systems based on their pros and cons.

1 Motivation

The transition from the Web of Documents to the Web of Data has resulted in a large compendium of interlinked datasets from diverse domains. Currently, the Linking Open Data (LOD) Cloud[4] contains over 60 billion triples available from more than 1000 different datasets with large datasets [13] being added frequently. Any complex application which relies on more that one data sources hence needs to be able to execute queries over multiple sources at the same time. We call such queries *federated queries*. Such datasets are not always or only available as SPARQL endpoints. In some cases, they might be available as Triple Pattern Fragments [16], or as data streams originated from various physical and virtual sensors (e.g., smart city infrastructure such as *http://www.odaa.dk/* and *http://iot.ee.surrey.ac.uk:8080/index.html*). Consequently, various engines and applications have been developed to enable the execution of federated queries on these different data infrastructures.

The aim of this tutorial is two-fold. First, we aim to provide the participants with an overview of the state of the art in federated queries. In addition, we aim to provide the participants with practical insights and hands-on experience that will allow them to select the right system for their purposes or even improve upon existing solutions in their future research.

[4] *http://stats.lod2.eu/*

2 Related Events

Hartig and Ozsu [7] presented a tutorial on Linked Data query processing. They focused more on the basics of the Linked Data, SPARQL query, types of source selection, and query optimization strategies. In this tutorial we are aiming to go into further details of the federated query processing with the assumption that the audience already have the basic knowledge of Linked Data and SPARQL queries. Moreover, we will be more focused on the practical demonstrations by using running examples and hand-on exercises. In addition, we will cover query federation in Linked Data Fragments [16] and Linked Data streams.

Tutorials on RDF Stream Processing (RSP) have been presented in ISWC 2013, 2014 and in ESWC 2014. However, RSP is still in its infancy stage and most of the tutorials are focused on the basic presentation of RDF streams and their query semantics. Complimentary to the previous work, in this tutorial we will be focusing on harnessing the true value from RDF streams by showcasing how these streams can be integrated with the background knowledge by executing federated queries over distributed data sources in combination with RDF streams. In order to provide better insights and hands on experience to the audience, in this tutorial we will use real time city data streams (e.g. traffic and temperatures sensors) and integrate them with background knowledge to answer complex user queries.

3 Detailed Description

In this section we describe the contents of tutorial, the aims and learning objectives, presentation style and tutorial format, and the prior knowledge required by the attendees.

3.1 Content Overview

Our tutorial will consist of four sessions:

- **Federated SPARQL query processing:** In this session, we will briefly describe the general steps involved in federated query processing over Linked Data, followed by the description of the state-of-the-art SPARQL query federation engines [14,6,1,10,17] over SPARQL endpoints. In particular, we will explain the source selection, query planning and optimization, and join implementation techniques in these engines. All of the these steps will be explained by using running examples. The aim is this section will be to familiarize the audience with the topic and enable them to understand the main differences between the selected engines. The insights gained in this session will also be central for the hands-on exercises (see last section of the tutorial).
- **Federation with Triple Pattern Fragments:** Since not all datasets are available as SPARQL endpoints all the time, Triple Pattern Fragments [16] were proposed as a light-weight interface to publish queriable Linked Data. Currently, more than 600,000 datasets are available as Triple Pattern Fragments (on *http://lodlaundromat.org/*). Since SPARQL queries are executed on the client side, this interface has

interesting potential for federation. In this part, we will briefly zoom in on the theory behind Triple Pattern Fragments, and then see how to execute custom federated queries. Furthermore, we will discuss how this technology can be integrated in real-world applications.
- **Federation over Streaming Linked Data:** Recent developments in technologies for the internet of things facilitate the availability of dynamic data streams produced by multiple sensors. These streams contain information from multiple domains (e.g., health, traffic, weather). Many RSP engines have been introduced to process semantic RDF streams [8,2,3]. In this session we will demonstrate how to effectively discover relevant data streams (data source selection) and integrate them with the federated data sources containing domain knowledge (federated query processing). We will showcase how applications can effectively discover and select relevant data streams while taking customised constraints and preferences of the user into account [4,5]. We will formulate federated queries to integrate real-time data streams *http://www.odaa.dk/* together with federated static data sources *http://iot.ee.surrey.ac.uk:8080/index.html* to answer complex user queries for smart city applications [15].
- **Hands-on Experience:** In the last session, we will have a hands-on session where the audience will be provided with a set of practical exercises related to each of the above three sessions. In particular, simple federated queries will be executed using some of the SPARQL query federation engines presented in the first session. Moreover, triple-pattern-fragment-based queries will be executed on data from the LOD laundromat. Finally, RDF data streams will be provided to the participants to enable them to execute federated queries using RSP engines.

3.2 Aims and Learning Objectives

Our learning objectives are the following:

- Provide basic knowledge of the federated query processing over Linked Data.
- Elaborate on the main differences between state-of-the-art federated query engines by using examples and hand-on exercises.
- Position these systems based on their pros and cons.
- Present the setup of federated environment for experiments and evaluation in the context of federated query engines over Linked Data.

3.3 Presentation Style and Format

Our presentations will be mostly based on animations, running examples, hands-on exercises and visualization. Basic questions will be asked during the session to keep the audience alert and ensure that the core message is understood. Questions will be allowed throughout the presentations. The first three sessions will last 45 minutes each questions. The hands-on exercise session will be conducted at the end of the tutorial and will last for about 1 hour. If time permits, we will discuss some of the open problems in the federated query processing over Linked Data. Furthermore, a short explanation of the federated evaluation setup and key performance metrics (to be considered for

the performance evaluation of federation engines) will be discussed as well. Given the centrality of federated queries for all complex Linked Data applications, we expect that most of the semantic web community will be interested in our tutorial and are thus expecting between 20 and 30 attendees (conservative estimate).

3.4 Required Prior Knowledge

The audience are required to have basic knowledge of the SPARQL query, Linked Data, and SPARQL endpoints.

4 Length

This will be a half day tutorial.

5 Technical Requirements

We will only need standard projection equipment. Participants should bring their own laptop. VirtualBox 4.3 will need to be installed in order to run the virtual machine image that we will distribute for the hand-on parts.

6 Presenters

Muhammad Saleem (primary contact)
University of Leipzig, Germany,
saleem@informatik.uni-leipzig.de
Homepage: https://sites.google.com/site/saleemsweb/

Muhammad Saleem obtained his Bachelor in Computer Software Engineering from N-W.F.P University of Engineering and Technology and a Master in Computer Science and Engineering from Hanyang University, South Korea. His research interests include federated SPARQL query processing and optimization, Linked Data summaries, personalized query execution, and top-k query processing [9,10,11,12]. He already presented[5] the subject topic as invited speaker at Ghent University – iMinds, Belgium. Previously, he was working as research assistant at Digital Enterprise Research Institute (DERI), National University of Ireland, Ireland. He is now a PhD student at the University of Leipzig (Agile Knowledge Engineering and Semantic Web Group) in Germany.

Dr. Muhammad Intizar Ali
INSIGHT Centre for Data Analytics, NUI Galway,
ali.intizar@insightcentre. org
Homepage: http://www.intizarali.org

Muhammad Intizar Ali is an Adjunct Lecturer & Postdoctoral Researcher at the National University of Ireland, Galway. He also holds an Assistant Professor position at

[5] Presentation slides can be found at: *http://goo.gl/sTI4NO*

COMSATS Institute of Information Technology, Lahore. His research interests include Semantic Web, Data Integration, Internet of Things (IoT), Linked Data, Federated Query Processing, Stream Query Processing and Optimal Query Processing over large scale distributed data sources. He is actively involved in various EU projects as well as industrial collaborative projects aimed at providing IoT enabled adaptive intelligence for smart city applications. He is also a presenter for the tutorial *"Semantic and Analytics for Smart City Application"* at ESWC 2015. Dr. Ali obtained his Ph.D. (with distinction) from Vienna University of Technology, Austria in 2011.

Dr. Ruben Verborgh
Ghent University – iMinds, Belgium
ruben.verborgh@ugent.be
http://ruben.verborgh.org/

Ruben Verborgh is a researcher in semantic hypermedia at Ghent University – iMinds, Belgium, where he obtained his PhD in Computer Science in 2014. He explores the connection between Semantic Web technologies and the Web's architectural properties, with the ultimate goal of building more intelligent clients. Along the way, he became fascinated by Linked Data, REST/hypermedia, Web APIs, and related technologies. He's a co-author of two books on Linked Data, and has written several publications on Web-related topics in international journals.

Ruben currently leads the Linked Data Fragments [16] effort, and is therefore highly knowledgeable to present about federation using Fragments. Furthermore, he has an extensive experience presenting hands-on workshops in the context of the Free Your Metadata initiative (*http://freeyourmetadata.org*), and is teaching practical exercises to Web technology students at Ghent University.

Dr. Axel-Cyrille Ngonga Ngomo
University of Leipzig, Germany,
ngonga@informatik.uni-leipzig.de
Homepage: https://aksw.org/AxelNgonga

Axel Ngonga co-leads the Agile Knowledge Engineering and Semantic Web research group at the University of Leipzig. His research interests revolve around Semantic Web technologies, especially link discovery, federated queries, machine learning and natural-language processing. Axel (co-)authored more than 100 reviewed publications, has developed/led the development of several widely used frameworks such as LIMES and FOX. In addition, he has received manifold awards including best (student) research paper awards at CiCLING 2008, ISWC 2011 and 2014 as well as ESWC 2013 and 2014. He has also won several challenges such as the I'Challenge 2013 and the Big Data Challenge at ISWC 2013. Axel has given university lectures on Semantic Web and Information Retrieval/Text Mining.

References

1. M. Acosta, M.-E. Vidal, T. Lampo, J. Castillo, and E. Ruckhaus. ANAPSID: an adaptive query processing engine for SPARQL endpoints. In *ISWC*, 2011.

2. D. F. Barbieri, D. Braga, S. Ceri, E. Della Valle, and M. Grossniklaus. C-sparql: Sparql for continuous querying. In *In Proc. of WWW*, pages 1061–1062, 2009.
3. A. Bolles, M. Grawunder, and J. Jacobi. Streaming sparql extending sparql to process data streams. In *In Proc. of ESWC*, pages 448–462, 2008.
4. F. Gao, M. I. Ali, and A. Mileo. Semantic discovery and integration of urban data streams. In *SSC at ISWC*, 2014.
5. F. Gao, E. Curry, M. Ali, S. Bhiri, and A. Mileo. Qos-aware complex event service composition and optimization using genetic algorithms. In *In Proc. of ICSOC,*, pages 386–393. 2014.
6. O. Görlitz and S. Staab. Splendid: Sparql endpoint federation exploiting void descriptions. In *COLD at ISWC*, 2011.
7. O. Hartig and M. T. Ozsu. Linked data query processing. In *2014 IEEE 30th International Conference on Data Engineering (ICDE)*, pages 1286–1289. IEEE, 2014.
8. D. Le-Phuoc, M. Dao-Tran, J. X. Parreira, and M. Hauswirth. A native and adaptive approach for unified processing of linked streams and linked data. In *In Proc. of ISWC*, pages 370–388. 2011.
9. M. Saleem, Y. Khan, A. Hasnain, I. Ermilov, and A.-C. N. Ngomo. A fine-grained evaluation of sparql endpoint federation systems. *SWJ*, 2014.
10. M. Saleem and A.-C. Ngonga Ngomo. HiBISCuS: Hypergraph-based source selection for sparql endpoint federation. In *ESWC*, 2014.
11. M. Saleem, A.-C. Ngonga Ngomo, J. X. Parreira, H. F. Deus, and M. Hauswirth. Daw: Duplicate-aware federated query processing over the web of data. In *ISWC*, 2013.
12. M. Saleem, S. S. Padmanabhuni, A.-C. Ngonga Ngomo, A. Iqbal, J. S. Almeida, S. Decker, and H. F. Deus. TopFed: TCGA tailored federated query processing and linking to LOD. *JBMS*, 2014.
13. M. Saleem, S. Shanmukha, A.-C. Ngonga, J. S. Almeida, S. Decker, and H. F. Deus. Linked cancer genome atlas database. In *I-Semantics 2013*, 2013.
14. A. Schwarte, P. Haase, K. Hose, R. Schenkel, and M. Schmidt. Fedx: Optimization techniques for federated query processing on linked data. In *ISWC*, 2011.
15. R. Tönjes, P. Barnaghi, M. I. Ali, A. Mileo, M. Hauswirth, F. Ganz, S. Ganea, B. Kjærgaard, D. Kuemper, S. Nechifor, D. Puiu, A. Sheth, V. Tsiatsis, and L. Vestergaard. Real time iot stream processing and large-scale data analytics for smart city applications. In *Presented at European Conference on Networks and Communications (EUCNC)*. 2014.
16. R. Verborgh, O. Hartig, B. De Meester, G. Haesendonck, L. De Vocht, M. Vander Sande, R. Cyganiak, P. Colpaert, E. Mannens, and R. Van de Walle. Querying datasets on the Web with high availability. In *Proceedings of the 13th International Semantic Web Conference*. Springer, Oct. 2014.
17. X. Wang, T. Tiropanis, and H. C. Davis. Lhd: Optimising linked data query processing using parallelisation. In *LDOW at WWW*, 2013.