Linked Data

Sherif Sakr • Marcin Wylot • Raghava Mutharaju •
Danh Le Phuoc • Irini Fundulaki

Linked Data

Storing, Querying, and Reasoning

Springer

Sherif Sakr
College of Public Health &
Health Informatics
King Saud bin Abdulaziz University
for Health Sciences
Riyadh, Saudi Arabia

Marcin Wylot
Fakultät IV - Open Distributed Systems
Technische Universität Berlin
Berlin, Germany

Raghava Mutharaju
Knowledge Discovery Lab
GE Global Research
Niskayuna
New York, USA

Danh Le Phuoc
Fakultät IV - Open Distributed Systems
Technische Universität Berlin
Berlin, Germany

Irini Fundulaki
Foundation for Research
and Technology - Hellas (FORTH)
Institute of Computer Science (ICS)
Heraklion, Greece

ISBN 978-3-319-73514-6 ISBN 978-3-319-73515-3 (eBook)
https://doi.org/10.1007/978-3-319-73515-3

Library of Congress Control Number: 2018930261

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my wife, Radwa, my daughter, Jana, and my son, Shehab, for their love, encouragement, and support.

Sherif Sakr

To my wife, Lalitha Sravanthi, my parents, my sister, and her family for their patience and support.

Raghava Mutharaju

To my wife, Nauy Tonnu, my daughter, Kim, and my parents and my sisters for their love, care, and support.

Danh Le Phuoc

To my family and my colleagues.

Irini Fundulaki

Foreword

The Semantic Web has matured over the past years. Since the grand vision of an intelligent agent-based Web of machines that will involve agents acting on our behalf, as outlined in a seminal article by Tim Berners-Lee et al. in 2001, we have seen the development of proven standards, used in Enterprise applications and almost anywhere on the Web. We could—as researchers—well ask ourselves the question "Are we there yet?" and would probably find many reasons to give an affirmative answer. Linked Data and Semantic Web technologies have become mainstream in many Web applications, being used heavily by Web search engines, by big media companies, and very successfully in Healthcare & Life Sciences applications, or e-Science. As such, Linked Data and Sematic Web have become a cornerstone of many "Big Data" applications, providing the "glue" for "Data Lakes" of heterogeneous data or building backbones of large-scale distributed knowledge bases in the form of so-called Knowledge Graphs.

Particularly, on this journey over the past roughly 20 years of research, the focus of the Semantic Web has changed several times, from originally being very much on (1) Knowledge Representation and Reasoning towards (2) more lightweight vocabulary management and Linked Data publishing and—more recently—towards (3) distributed graph databases and scalable data management and processing of Semantic (Web) data, accessible as distributed Linked Data via query interfaces using SPARQL as a standard query language, so-called SPARQL endpoints.

Whereas other textbooks already cover the former two aspects in much detail, now is probably an excellent time for the latter aspect to be presented to a wider audience as it is done in this volume. The success of the Semantic Web has always been driven by enthusiasts who carried the idea forward in education, with preparing and teaching materials for academics and third-level education in general. The availability of good entry-level materials to teach techniques and the state of the art in research is important for both practitioners and researchers approaching the field in order to advance research and our knowledge itself.

This book will provide people with little or no Semantic Web background an entrance, but also covers current hot topics and recent trends about querying and efficient storage of Linked Data at scale. After providing an overview of the relevant standards, the reader will learn how RDF and Linked Data can be stored and queried efficiently, also using recent techniques from NoSQL databases. Next, the processing of and reasoning about dynamic, streaming data is covered as well as distributed techniques for storing and reasoning about Semantic Web data. Last but not least, the book provides information about state-of-the-art benchmarks for systems to store and query such data. A final chapter explains the importance of storing and using provenance data in the context of linked data and how such provenance information can be leveraged in Semantic data management.

Overall, I recommend this book to anyone interested in Semantic Web beyond just data publishing and querying SPARQL endpoints for gaining a deeper insight on what efficient management of Linked Data means under the surface and how it can be implemented and realized using modern NoSQL database systems.

Vienna University of Economics and Business
Vienna, Austria

Axel Polleres

Preface

The World Wide Web is an interlinked collection of documents that are generally devoid of any structure and primarily meant for human consumption. In 2006, Sir Tim Berners-Lee, the inventor of the Web, proposed and described the Semantic Web vision as an extension of the World Wide Web where structure and meaning are provided to the data. This makes the information in the interlinked documents machine understandable. In practice, the nature of the World Wide Web has evolved from a web of linked documents to a web including Linked Data. Traditionally, we were able to publish documents on the Web and create links between them. Those links, however, only allowed the document space to be traversed without understanding the relationships between the documents and without linking to particular pieces of information. In principle, Linked Data allows meaningful links to be created between pieces of data on the Web. The adoption of Linked Data technologies has shifted the Web from a space of connecting documents to a global space where pieces of data from different domains are semantically linked and integrated to create a global Web of Data. Linked Data enables operations to deliver integrated results as new data is added to the global space. This opens new opportunities for applications such as search engines, data browsers, and various domain-specific applications.

In practice, the Web of Linked Data is rapidly growing from a dozen data collections in 2007 to a space of hundreds of data sources in April 2014. In particular, the number of linked datasets doubled between 2011 and 2014, which shows an accelerating trend of data integration on the Web. The Web of Linked Data contains heterogeneous data coming from multiple sources and various contributors, produced using different methods and degrees of authoritativeness, and gathered automatically from independent and potentially unknown sources. Thus, there is a growing momentum to harness the power of linked data in several application domains. This book is intended to take you in a journey for describing efficient and effective techniques for harnessing the power of Linked Data by tackling the various aspects for managing the growing amounts of Linked Data: storing, querying, reasoning, provenance management, and benchmarking.

Organization of the Book

The World Wide Web is an interlinked collection of documents that are generally devoid of any structure and primarily meant for human consumption. The Semantic Web is an extension of the World Wide Web where structure and meaning are provided to the data. The adoption of Linked Data technologies has shifted the Web from a space of connecting documents to a global space where pieces of data from different domains are semantically linked and integrated to create a global Web of Data. Chapter 1 introduces the main concepts of Semantic Web and Linked Data and provides the book roadmap.

All concepts underpinning Linked Data are standardized by the World Wide Web Consortium. The Consortium publishes recommendations defining and describing in detail the technologies behind Linked Data. Chapter 2 briefly introduces the basic concepts underpinning Linked Data technologies and that are necessary to follow the course of this book. We present a data model, a query language, vocabularies, and a data exchange format. In addition, the chapter provides an overview of emerging big data storage and processing frameworks that are frequently used for RDF data management (e.g., NoSQL databases, Hadoop, Spark).

The wide adoption of the RDF data model has called for efficient and scalable RDF query processing schemes. As a response to this call, a number of centralized RDF query processing systems have been designed to tackle this challenge. In these systems, the storage and query processing of RDF datasets are managed on a single node. Chapter 3 gives an overview of various techniques and systems for centrally querying RDF datasets.

With increasing sizes of RDF datasets, executing complex queries on a single node has turned to be impractical especially when the node's main memory is dwarfed by the volume of the dataset. Therefore, there was a crucial need for distributed systems with a high degree of parallelism that can satisfy the performance demands of complex SPARQL queries. Chapter 4 provides an overview of various techniques and systems for efficiently querying large RDF datasets in distributed environments.

We are witnessing a paradigm shift, where real-time, time-dependent data is becoming ubiquitous. As Linked Data facilitates the data integration process among heterogeneous data sources, RDF Stream Data has the same goal with respect to data streams. It bridges the gap between stream and more static data sources. To support the processing of RDF stream data, there is a need of investigating how to extend RDF to model and represent stream data. Then, from the RDF-based data representation, the query processing models need to be defined to build the stream processing engine that is tailored for streaming data. Chapter 5 provides an overview on how such requirements are addressed in the current state of the art of RDF Stream Data processing.

Large RDF interconnected datasets, especially in the form of open as well as enterprise knowledge graphs, are constructed and consumed in several domains. Reasoning over such large knowledge graphs poses several performance challenges. In practice, although there has been some prior work on scalable approaches to RDF reasoning, the interest in this field started gathering momentum with the rising popularity of modern big data processing systems (e.g., Hadoop, Spark). Chapter 6 covers five main categories of distributed RDF reasoning systems: (1) Peer-to-Peer RDF reasoning systems, (2) NoSQL-based RDF reasoning systems, (3) Hadoop-based RDF reasoning systems, (4) Spark-based RDF reasoning systems, and (5) shared-memory RDF reasoning systems.

Standards and benchmarking have traditionally been used as the main tools to formally define and provably illustrate the level of the adequacy of systems to address the new challenges. Chapter 7 discusses benchmarks for RDF query engines and instance matching systems. In practice, benchmarks are used to inform users of the strengths and weaknesses of competing tools and approaches, but more importantly, they encourage the advancement of technology by providing both academia and industry with clear targets for performance and functionality.

Chapter 8 discusses the provenance management for Linked Data and presents the different provenance models developed for the different fragments of the SPARQL standard language for querying RDF datasets. In addition, we discuss the different models for relational provenance that set the basis for RDF provenance models and proceed with a thorough presentation of the various provenance models for the different fragments of the SPARQL query language.

Chapter 9 briefly summarizes the book journey before providing some insights and highlights on some of the open challenges and research directions for advancing the state of the art of Linked Data towards achieving the ultimate vision of the Semantic Web Domain.

Target Audience

This book is mainly targeting students and academic researchers who are interested in the Linked Data domain. The book provides readers with an updated view of methods, technologies, and systems related to Linked Data.

For Students This book provides an overview of the foundations and underpinning technologies and standards for Linked Data. We comprehensively cover the state of the art and discuss the technical challenges in depth.

For Researchers The material of this book will provide you with a thorough coverage for the emerging and ongoing advancements on Linked Data storing, querying, reasoning, and provenance management systems. You can use this book as a starting point to tackle your next research challenge in the domain of Linked Data management.

Riyadh, Saudi Arabia	Sherif Sakr
Berlin, Germany	Marcin Wylot
Niskayuna, NY, USA	Raghava Mutharaju
Berlin, Germany	Danh Le Phuoc
Heraklion, Greece	Irini Fundulaki

Acknowledgments

The authors are grateful to many collaborators for their contribution to this book. In particular, we would like to mention Manfred Hauswirth, Martin Grund, Philippe Cudre-Mauroux, Floris Geerts, Vassilis Christophides, Grigoris Karvounarakis, Anastasios Kementsietsidis, Giorgos Flouris, Evangelia Daskalaki, Venelin Kotsev, Peter Boncz, Axel Ngonga-Ngomo, and Paul Groth. Thank you all!

Thanks to Marie Sklodowska-Curie Programme H2020-MSCA-IF-2014 which is supporting one of the authors under his Individual Fellowship, Grant No. 661180.

Thanks to Springer-Verlag for publishing this book. Ralf Gerstner has greatly encouraged and supported us to write this book. Thanks Ralf!

Contents

About the Authors

Sherif Sakr is a professor of computer and information science in the Health Informatics department at King Saud bin Abdulaziz University for Health Sciences. He is also affiliated with the University of New South Wales and DATA61/CSIRO. He received his PhD in Computer and Information Science from the University of Konstanz, Germany, in 2007. He received his BSc and MSc in Computer Science from the Faculty of Computers and Information in Cairo University, Egypt, in 2000 and 2003, respectively. In 2013, Sherif has been awarded the Stanford Innovation and Entrepreneurship Certificate. Sherif's research interests revolve around the areas of efficient and scalable Big Data Management, Processing, and Analytics. He coauthored and coedited seven books covering various fundamental aspects in the field of Data Management. Prof. Sakr is an ACM and IEEE Distinguished Speaker. He is currently serving as the Editor-in-Chief of the Springer Encyclopedia of Big Data Technologies. Homepage: http://www.cse.unsw.edu.au/~ssakr/.

Marcin Wylot is a postdoctoral researcher at TU Berlin in the ODS group. He received his PhD at the University of Fribourg in Switzerland in 2015, with the supervision of Professor Philippe Cudré-Mauroux. He obtained his MSc in Computer Science at the University of Lodz in Poland in 2010, doing part of his studies at the University of Lyon in France. During his studies he was also gaining professional experience working in various industrial companies. His main research interests revolve around database systems for Semantic Web data, provenance in Linked Data, Internet of Things, and Big Data processing. Homepage: http://mwylot.net.

Raghava Mutharaju is a Research Scientist in the AI & Machine Learning Systems division of GE Global Research in Niskayuna, NY, USA. He received his PhD in Computer Science and Engineering from Wright State University, Dayton, OH, USA, in 2016. His dissertation work involved investigating various approaches to distributed reasoning of OWL ontologies. He received his Master of Technology (M.Tech) and Bachelor of Technology (B.Tech) in Computer Science from Motilal

Nehru National Institute of Technology (MNNIT), Allahabad, India, and Jawaharlal Nehru National Institute of Technology (JNTU), Hyderabad, India, respectively. His research interests are in ontology modeling and reasoning, scalable SPARQL query processing, Big Data, and Semantic Web and its applications. He has published at several venues such as ISWC, ESWC, ECAI, and WISE. He cochaired workshops at WebSci 2017 and ISWC 2015. He has co-organized tutorials at IJCAI 2016, AAAI 2015, and ISWC 2014. He has been on the Program Committee of several Semantic Web conferences such as ISWC, ESWC, K-CAP, and SEMANTiCS. Homepage: http://raghavam.github.io/.

Danh Le Phuoc is a Marie Sklodowska-Curie Fellow at the Technical University of Berlin. He received his PhD in Computer Science from the National University of Ireland. He is working on Pervasive Analytics which includes Linked Data/Semantic Web, Pervasive Computing, Future Internet, and Big Data for Internet of Everything. Before joining TUB, he was a Principal Investigator, Research Fellow, and Project Lead of the Insight Centre of Data Analytics or Digital Enterprise Research Institute (DERI) at the National University of Ireland, Galway. Before doing PhD, he spent 8 years working in several industrial positions.

Irini Fundulaki is a Principal Researcher at the Institute of Computer Science of the Foundation for Research and Technology-Hellas. She received her PhD in Computer Science from the Conservatoire National des Arts et Métiers, Paris, France, in 2003. She received her BSc and MSc in Computer Science from the Computer Science Department of the University of Crete, Greece, in 1994 and 1996, respectively. After her PhD she worked as a postdoc and subsequently became a Member of Technical Staff in Bell Laboratories, USA, at the Network Data and Services Research Department. She was also a Research Fellow at the Database Group, Laboratory for Foundations of Computer Science, School of Informatics, University of Edinburgh, UK. Irini's research interests are related to Web Data Management and more specifically the development of benchmarks for RDF engines, instance matching and link discovery systems, the management of provenance for Linked Data, access control for RDF datasets, bias in online information providers, and finally data integration. She has authored a large number of articles and journals and has served as chair of three international workshops.