

Cloud-Based RDF Data Management

Synthesis Lectures on Data Management

Editor

H.V. Jagadish, *University of Michigan*

Founding Editor

M. Tamer Özsu, *University of Waterloo*

Synthesis Lectures on Data Management is edited by H.V. Jagadish of the University of Michigan. The series publishes 80–150 page publications on topics pertaining to data management. Topics include query languages, database system architectures, transaction management, data warehousing, XML and databases, data stream systems, wide scale data distribution, multimedia data management, data mining, and related subjects.

Cloud-Based RDF Data Management

Zoi Kaoudi, Ioana Manolescu, and Stamatis Zampetakis
2020

Community Search over Big Graphs

Xin Huang, Laks V.S. Lakshmanan, and Jianliang Xu
2019

On Transactional Concurrency Control

Goetz Graefe
2019

Data-Intensive Workflow Management: For Clouds and Data-Intensive and Scalable Computing Environments

Daniel C.M. de Oliveira, Ji Liu, and Esther Pacitti
2019

Answering Queries Using Views, Second Edition

Foto Afrati and Rada Chirkova
2019

Transaction Processing on Modern Hardware

Mohammad Sadoghi and Spyros Blanas
2019

Data Management in Machine Learning Systems

Matthias Boehm, Arun Kumar, and Jun Yang
2019

Non-Volatile Memory Database Management Systems

Joy Arulraj and Andrew Pavlo
2019

Scalable Processing of Spatial-Keyword Queries

Ahmed R. Mahmood and Walid G. Aref
2019

Data Exploration Using Example-Based Methods

Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis
2018

Data Profiling

Ziawasch Abedjan, Lukasz Golab, Felix Naumann, and Thorsten Papenbrock
2018

Querying Graphs

Angela Bonifati, George Fletcher, Hannes Voigt, and Nikolay Yakovets
2018

Query Processing over Incomplete Databases

Yunjun Gao and Xiaoye Miao
2018

Natural Language Data Management and Interfaces

Yunyao Li and Davood Rafiei
2018

Human Interaction with Graphs: A Visual Querying Perspective

Sourav S. Bhowmick, Byron Choi, and Chengkai Li
2018

On Uncertain Graphs

Arijit Khan, Yuan Ye, and Lei Chen
2018

Answering Queries Using Views

Foto Afrati and Rada Chirkova
2017

Databases on Modern Hardware: How to Stop Underutilization and Love Multicores

Anatasia Ailamaki, Erieta Liarou, Pinar Tözün, Danica Porobic, and Iraklis Psaroudakis
2017

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, Media Restore, and System Failover, Second Edition

Goetz Graefe, Wey Guy, and Caetano Sauer
2016

Generating Plans from Proofs: The Interpolation-based Approach to Query Reformulation

Michael Benedikt, Julien Leblay, Balder ten Cate, and Efthymia Tsamoura
2016

Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics

Laure Berti-Équille and Javier Borge-Holthoefer
2015

Datalog and Logic Databases

Sergio Greco and Cristina Molinaro
2015

Big Data Integration

Xin Luna Dong and Divesh Srivastava
2015

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, and Media Restore

Goetz Graefe, Wey Guy, and Caetano Sauer
2014

Similarity Joins in Relational Database Systems

Nikolaus Augsten and Michael H. Böhlen
2013

Information and Influence Propagation in Social Networks

Wei Chen, Laks V.S. Lakshmanan, and Carlos Castillo
2013

Data Cleaning: A Practical Perspective

Venkatesh Ganti and Anish Das Sarma
2013

Data Processing on FPGAs

Jens Teubner and Louis Woods
2013

Perspectives on Business Intelligence

Raymond T. Ng, Patricia C. Arocena, Denilson Barbosa, Giuseppe Carenini, Luiz Gomes, Jr., Stephan Jou, Rock Anthony Leung, Evangelos Milios, Renée J. Miller, John Mylopoulos, Rachel A. Pottinger, Frank Tompa, and Eric Yu
2013

Semantics Empowered Web 3.0: Managing Enterprise, Social, Sensor, and Cloud-based Data and Services for Advanced Applications

Amit Sheth and Krishnaprasad Thirunarayan
2012

Data Management in the Cloud: Challenges and Opportunities

Divyakant Agrawal, Sudipto Das, and Amr El Abbadi
2012

Query Processing over Uncertain Databases

Lei Chen and Xiang Lian
2012

Foundations of Data Quality Management

Wenfei Fan and Floris Geerts
2012

Incomplete Data and Data Dependencies in Relational Databases

Sergio Greco, Cristian Molinaro, and Francesca Spezzano
2012

Business Processes: A Database Perspective

Daniel Deutch and Tova Milo
2012

Data Protection from Insider Threats

Elisa Bertino
2012

Deep Web Query Interface Understanding and Integration

Eduard C. Dragut, Weiyi Meng, and Clement T. Yu
2012

P2P Techniques for Decentralized Applications

Esther Pacitti, Reza Akbarinia, and Manal El-Dick
2012

Query Answer Authentication

HweeHwa Pang and Kian-Lee Tan
2012

Declarative Networking

Boon Thau Loo and Wenchao Zhou
2012

Full-Text (Substring) Indexes in External Memory

Marina Barsky, Ulrike Stege, and Alex Thomo
2011

Spatial Data Management

Nikos Mamoulis
2011

Database Repairing and Consistent Query Answering

Leopoldo Bertossi
2011

Managing Event Information: Modeling, Retrieval, and Applications

Amarnath Gupta and Ramesh Jain
2011

Fundamentals of Physical Design and Query Compilation

David Toman and Grant Weddell
2011

Methods for Mining and Summarizing Text Conversations

Giuseppe Carenini, Gabriel Murray, and Raymond Ng
2011

Probabilistic Databases

Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch
2011

Peer-to-Peer Data Management

Karl Aberer
2011

Probabilistic Ranking Techniques in Relational Databases

Ihab F. Ilyas and Mohamed A. Soliman
2011

Uncertain Schema Matching

Avigdor Gal
2011

Fundamentals of Object Databases: Object-Oriented and Object-Relational Design

Suzanne W. Dietrich and Susan D. Urban
2010

Advanced Metasearch Engine Technology

Weiyi Meng and Clement T. Yu
2010

Web Page Recommendation Models: Theory and Algorithms

Sule Gündüz-Ögüdücü
2010

Multidimensional Databases and Data Warehousing

Christian S. Jensen, Torben Bach Pedersen, and Christian Thomsen
2010

Database Replication

Bettina Kemme, Ricardo Jimenez-Peris, and Marta Patino-Martinez
2010

Relational and XML Data Exchange

Marcelo Arenas, Pablo Barcelo, Leonid Libkin, and Filip Murlak
2010

User-Centered Data Management

Tiziana Catarci, Alan Dix, Stephen Kimani, and Giuseppe Santucci
2010

Data Stream Management

Lukasz Golab and M. Tamer Özsu
2010

Access Control in Data Management Systems

Elena Ferrari
2010

An Introduction to Duplicate Detection

Felix Naumann and Melanie Herschel
2010

Privacy-Preserving Data Publishing: An Overview

Raymond Chi-Wing Wong and Ada Wai-Chee Fu
2010

Keyword Search in Databases

Jeffrey Xu Yu, Lu Qin, and Lijun Chang
2009

© Springer Nature Switzerland AG 2022
Reprint of original edition © Morgan & Claypool 2020

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Cloud-Based RDF Data Management
Zoi Kaoudi, Ioana Manolescu, and Stamatis Zampetakis

ISBN: 978-3-031-00747-7 paperback
ISBN: 978-3-031-01875-6 ebook
ISBN: 978-3-031-00102-4 hardcover

DOI 10.1007/978-3-031-01875-6

A Publication in the Springer series

SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #62
Series Editor: H.V. Jagadish, *University of Michigan*
Founding Editor: M. Tamer Özsu, *University of Waterloo*
Series ISSN
Print 2153-5418 Electronic 2153-5426

Cloud-Based RDF Data Management

Zoi Kaoudi

Technische Universität Berlin

Ioana Manolescu

INRIA

Stamatis Zampetakis

TIBCO Orchestra Networks

SYNTHESIS LECTURES ON DATA MANAGEMENT #62

ABSTRACT

Resource Description Framework (or RDF, in short) is set to deliver many of the original semi-structured data promises: flexible structure, optional schema, and rich, flexible Universal Resource Identifiers as a basis for information sharing. Moreover, RDF is uniquely positioned to benefit from the efforts of scientific communities studying databases, knowledge representation, and Web technologies. As a consequence, the RDF data model is used in a variety of applications today for integrating knowledge and information: in open Web or government data via the Linked Open Data initiative, in scientific domains such as bioinformatics, and more recently in search engines and personal assistants of enterprises in the form of knowledge graphs.

Managing such large volumes of RDF data is challenging due to the sheer size, heterogeneity, and complexity brought by RDF reasoning. To tackle the size challenge, distributed architectures are required. Cloud computing is an emerging paradigm massively adopted in many applications requiring distributed architectures for the scalability, fault tolerance, and elasticity features it provides. At the same time, interest in massively parallel processing has been renewed by the MapReduce model and many follow-up works, which aim at simplifying the deployment of massively parallel data management tasks in a cloud environment.

In this book, we study the state-of-the-art RDF data management in cloud environments and parallel/distributed architectures that were not necessarily intended for the cloud, but can easily be deployed therein. After providing a comprehensive background on RDF and cloud technologies, we explore four aspects that are vital in an RDF data management system: data storage, query processing, query optimization, and reasoning. We conclude the book with a discussion on open problems and future directions.

KEYWORDS

RDF, cloud computing, MapReduce, key-value stores, query optimization, reasoning

Contents

1	Introduction	1
2	Preliminaries	5
2.1	Resource Description Framework (RDF)	5
2.1.1	Data Model	5
2.1.2	The SPARQL Query Language	10
2.2	Distributed Storage and Computing Paradigms	13
2.2.1	Distributed File Systems	13
2.2.2	Distributed Key-Value Stores	14
2.2.3	Distributed Computation Frameworks: MapReduce and Beyond	17
2.3	Summary	20
3	Cloud-Based RDF Storage	21
3.1	Partitioning Strategies	21
3.1.1	Storage Description Grammar	22
3.1.2	Logical Partitioning	24
3.1.3	Graph Partitioning	24
3.2	Storing in Distributed File Systems	25
3.2.1	Triple Model	25
3.2.2	Vertical Partitioning Model	25
3.2.3	Partitioning Based on RDF Entities	28
3.3	Storing in Key-Value Stores	29
3.3.1	Triple-Based	29
3.3.2	Graph-Based	32
3.4	Storing in Multiple Centralized RDF Stores	34
3.5	Storing in Main Memory Stores	38
3.6	Storing in Multiple Back-End Stores	39
3.7	Summary	39
4	Cloud-Based SPARQL Query Processing	43
4.1	Relational-Based Query Processing	43

4.1.1	Data Access Paths	44
4.1.2	Join Evaluation	47
4.2	Graph-Based Query Processing	51
4.2.1	Graph Exploration	51
4.2.2	Partial Evaluation and Assembly Methods	53
4.3	Summary	53
5	SPARQL Query Optimization for the Cloud	57
5.1	Query Plan Search Space	57
5.2	Planning Algorithms	61
5.2.1	Exhaustive Approaches	61
5.2.2	Heuristics-Based Approaches	64
5.2.3	Dynamic Programming	65
5.2.4	Greedy Approaches	65
5.3	Summary	65
6	RDFS Reasoning in the Cloud	67
6.1	Reasoning through RDFS Closure Computation	67
6.2	Reasoning through Query Reformulation	70
6.3	Hybrid Techniques	71
6.4	Summary	72
7	Concluding Remarks	73
	Bibliography	75
	Authors' Biographies	91