MORGAN & CLAYPOOL PUBLISHERS

Big Data Integration

Xin Luna Dong
Divesh Srivastava

SYNTHESIS LECTURES ON DATA MANAGEMENT

Z. Meral Özsoyoğlu, *Series Editor*

Big Data Integration

Xin Luna Dong
Google Inc.

Divesh Srivastava
AT&T Labs-Research

SYNTHESIS LECTURES ON DATA MANAGEMENT #40

MORGAN & CLAYPOOL PUBLISHERS

Big Data Integration

Xin Luna Dong, Divesh Srivastava

www.morganclaypool.com

ISBN: 978-1-62705-223-8 paperback
ISBN: 978-1-62705-224-5 ebook

DOI: 10.2200/S00578ED1V01Y201404DTM040

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON DATA MANAGEMENT
Series ISSN: 2153-5418 print 2153-5426 ebook

Lecture #40
Series Editor: M. Tamer Özsu, *University of Waterloo*

First Edition

10 9 8 7 6 5 4 3 2 1

Synthesis Lectures on Data Management

Synthesis Lectures on Data Management is edited by Meral Özsoyoğlu of Case Western Reserve University. The series publishes 80- to 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.

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, Eric Yu
April 2013

Semantics Empowered Web 3.0: Managing Enterprise, Social, Sensor, and Cloud-Based Data and Services for Advanced Applications
Amit Sheth, Krishnaprasad Thirunarayan
December 2012

Data Management in the Cloud: Challenges and Opportunities
Divyakant Agrawal, Sudipto Das, Amr El Abbadi
December 2012

Query Processing over Uncertain Databases
Lei Chen, Xiang Lian
December 2012

Foundations of Data Quality Management
Wenfei Fan, Floris Geerts
July 2012

Incomplete Data and Data Dependencies in Relational Databases
Sergio Greco, Cristian Molinaro, Francesca Spezzano
July 2012

Business Processes: A Database Perspective
Daniel Deutch, Tova Milo
July 2012

Data Protection from Insider Threats
Elisa Bertino
June 2012

Deep Web Query Interface Understanding and Integration
Eduard C. Dragut, Weiyi Meng, Clement T. Yu
June 2012

P2P Techniques for Decentralized Applications
Esther Pacitti, Reza Akbarinia, Manal El-Dick
April 2012

Query Answer Authentication
HweeHwa Pang, Kian-Lee Tan
February 2012

Declarative Networking
Boon Thau Loo, Wenchao Zhou
January 2012

Full-Text (Substring) Indexes in External Memory
Marina Barsky, Ulrike Stege, Alex Thomo
December 2011

Spatial Data Management
Nikos Mamoulis
November 2011

Database Repairing and Consistent Query Answering
Leopoldo Bertossi
August 2011

Managing Event Information: Modeling, Retrieval, and Applications
Amarnath Gupta, Ramesh Jain
July 2011

Fundamentals of Physical Design and Query Compilation
David Toman, Grant Weddell
July 2011

Methods for Mining and Summarizing Text Conversations
Giuseppe Carenini, Gabriel Murray, Raymond Ng
June 2011

Probabilistic Databases
Dan Suciu, Dan Olteanu, Christopher Ré, Christoph Koch
May 2011

Peer-to-Peer Data Management
Karl Aberer
May 2011

Probabilistic Ranking Techniques in Relational Databases
Ihab F. Ilyas, Mohamed A. Soliman
March 2011

Uncertain Schema Matching
Avigdor Gal
March 2011

Fundamentals of Object Databases: Object-Oriented and Object-Relational Design
Suzanne W. Dietrich, Susan D. Urban
2010

Advanced Metasearch Engine Technology
Weiyi Meng, Clement T. Yu
2010

Web Page Recommendation Models: Theory and Algorithms
Şule Gündüz-Ögüdücü
2010

Multidimensional Databases and Data Warehousing
Christian S. Jensen, Torben Bach Pedersen, Christian Thomsen
2010

Database Replication
Bettina Kemme, Ricardo Jimenez-Peris, Marta Patino-Martinez
2010

Relational and XML Data Exchange
Marcelo Arenas, Pablo Barcelo, Leonid Libkin, Filip Murlak
2010

User-Centered Data Management
Tiziana Catarci, Alan Dix, Stephen Kimani, Giuseppe Santucci
2010

Data Stream Management
Lukasz Golab, M. Tamer Özsu
2010

Access Control in Data Management Systems
Elena Ferrari
2010

An Introduction to Duplicate Detection
Felix Naumann, Melanie Herschel
2010

Privacy-Preserving Data Publishing: An Overview
Raymond Chi-Wing Wong, Ada Wai-Chee Fu
2010

Keyword Search in Databases
Jeffrey Xu Yu, Lu Qin, Lijun Chang
2009

ABSTRACT

The big data era is upon us: data are being generated, analyzed, and used at an unprecedented scale, and data-driven decision making is sweeping through all aspects of society. Since the value of data explodes when it can be linked and fused with other data, addressing the *big data integration* (BDI) challenge is critical to realizing the promise of big data.

BDI differs from traditional data integration along the dimensions of *volume*, *velocity*, *variety*, and *veracity*. First, not only can data sources contain a huge volume of data, but also the number of data sources is now in the millions. Second, because of the rate at which newly collected data are made available, many of the data sources are very dynamic, and the number of data sources is also rapidly exploding. Third, data sources are extremely heterogeneous in their structure and content, exhibiting considerable variety even for substantially similar entities. Fourth, the data sources are of widely differing qualities, with significant differences in the coverage, accuracy and timeliness of data provided.

This book explores the progress that has been made by the data integration community on the topics of schema alignment, record linkage and data fusion in addressing these novel challenges faced by big data integration. Each of these topics is covered in a systematic way: first starting with a quick tour of the topic in the context of traditional data integration, followed by a detailed, example-driven exposition of recent innovative techniques that have been proposed to address the BDI challenges of volume, velocity, variety, and veracity. Finally, it presents emerging topics and opportunities that are specific to BDI, identifying promising directions for the data integration community.

KEYWORDS

big data integration, data fusion, record linkage, schema alignment, variety, velocity, veracity, volume

Preface

Big data integration is the confluence of two significant bodies of work: one quite old—data integration—and the other relatively new—big data.

As long as there have been data sets that people have sought to link and fuse to enhance value, data integration has been around. Even before computer scientists started investigating this area, statisticians had already made much progress, given their pressing need to correlate and analyze census data sets collected over time. Data integration is challenging for many reasons, not the least being our ability to represent and misrepresent information about real-world entities in very diverse ways. To effectively address these challenges, considerable progress has been made over the last few decades by the data integration community on the foundational topics of schema alignment, record linkage, and data fusion, especially for well-structured data.

Recent years have seen a dramatic growth in our ability to capture each event and every interaction in the world as digital data. Concomitant with this ability has been our desire to analyze and extract value from this data, ushering in the era of big data. This era has seen an enormous increase in the amount and heterogeneity of data, as well as in the number of data sources, many of which are very dynamic, while being of widely differing qualities. Since the value of data explodes when it can be linked and fused with other data, data integration is critical to realizing the promise of big data of enabling valuable, data-driven decisions to alter all aspects of society.

Data integration for big data is what has come to be known as big data integration. This book explores the progress that has been made by the data integration community in addressing the novel challenges faced by big data integration. It is intended as a starting point for researchers, practitioners and students who would like to learn more about big data integration. We have attempted to cover a diversity of topics and research efforts in this area, fully well realizing that it is impossible to be comprehensive in such a dynamic area. We hope that many of our readers will be inspired by this book to make their own contributions to this important area, to help further the promise of big data.

ACKNOWLEDGMENTS

Several people provided valuable support during the preparation of this book. We warmly thank Tamer Özsu for inviting us to write this book, Diane Cerra for managing the entire publication process, and Paul Anagnostopoulos for producing the book. Without their gentle reminders, periodic nudging, and prompt copyediting, this book may have taken much longer to complete.

Much of this book's material evolved from the tutorials and talks that we presented at ICDE 2013, VLDB 2013, COMAD 2013, University of Zurich (Switzerland), the Ph.D. School of ADC 2014 and BDA 2014. We thank our many colleagues for their constructive feedback during and subsequent to these presentations.

We would also like to acknowledge our many collaborators who have influenced our thoughts and our understanding of this research area over the years.

Finally, we would like to thank our family members, whose constant encouragement and loving support made it all worthwhile.

Xin Luna Dong and Divesh Srivastava
December 2014

Contents

List of Figures

List of Tables