



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

Copyright © 2015 by Morgan & Claypool Publishers

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.

Big Data Integration

Xin Luna Dong, Divesh Srivastava

[www.morganclaypool.com](http://www.morganclaypool.com)

ISBN: 978-1-62705-223-8    paperback

ISBN: 978-1-62705-224-5    ebook

DOI: [10.2200/S00578ED1V01Y201404DTM040](https://doi.org/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

Editor

**Z. Meral Özsoyoğlu**, *Case Western Reserve University*

Founding Editor

**M. Tamer Özsu**, *University of Waterloo*

**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.

**Big Data Integration**

Xin Luna Dong, Divesh Srivastava

March 2015

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

Goetz Graefe, Wey Guy, Caetano Sauer

December 2014

**Similarity Joins in Relational Database Systems**

Nikolaus Augsten, Michael H. Böhlen

November 2013

**Information and Influence Propagation in Social Networks**

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

October 2013

**Data Cleaning: A Practical Perspective**

Venkatesh Ganti, Anish Das Sarma

September 2013

**Data Processing on FPGAs**

Jens Teubner, Louis Woods

June 2013

### Perspectives on Business Intelligence

Raymond T. Ng, Patricia C. Arocena, Denilson Barbosa, Giuseppe Carenini, Luiz Gomes, Jr., Stephan Jou, Rock Anthony Leung, Evangelos Miliotis, 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 Suci, 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 . . . . .	xv
List of Tables . . . . .	xvii
Preface . . . . .	xix
Acknowledgments . . . . .	xix
<b>1. Motivation: Challenges and Opportunities for BDI . . . . .</b>	<b>1</b>
1.1 Traditional Data Integration . . . . .	2
1.1.1 The Flights Example: Data Sources . . . . .	2
1.1.2 The Flights Example: Data Integration . . . . .	6
1.1.3 Data Integration: Architecture & Three Major Steps . . . . .	9
1.2 BDI: Challenges . . . . .	11
1.2.1 The “V” Dimensions . . . . .	11
1.2.2 Case Study: Quantity of Deep Web Data . . . . .	13
1.2.3 Case Study: Extracted Domain-Specific Data . . . . .	15
1.2.4 Case Study: Quality of Deep Web Data . . . . .	20
1.2.5 Case Study: Surface Web Structured Data . . . . .	23
1.2.6 Case Study: Extracted Knowledge Triples . . . . .	26
1.3 BDI: Opportunities . . . . .	27
1.3.1 Data Redundancy . . . . .	27
1.3.2 Long Data . . . . .	28
1.3.3 Big Data Platforms . . . . .	29
1.4 Outline of Book . . . . .	29
<b>2. Schema Alignment . . . . .</b>	<b>31</b>
2.1 Traditional Schema Alignment: A Quick Tour . . . . .	32
2.1.1 Mediated Schema . . . . .	32
2.1.2 Attribute Matching . . . . .	32
2.1.3 Schema Mapping . . . . .	33
2.1.4 Query Answering . . . . .	34
2.2 Addressing the Variety and Velocity Challenges . . . . .	35
2.2.1 Probabilistic Schema Alignment . . . . .	36
2.2.2 Pay-As-You-Go User Feedback . . . . .	47

2.3	Addressing the Variety and Volume Challenges . . . . .	49
2.3.1	Integrating Deep Web Data . . . . .	49
2.3.2	Integrating Web Tables . . . . .	54
<b>3.</b>	<b>Record Linkage . . . . .</b>	<b>63</b>
3.1	Traditional Record Linkage: A Quick Tour . . . . .	64
3.1.1	Pairwise Matching . . . . .	65
3.1.2	Clustering . . . . .	67
3.1.3	Blocking . . . . .	68
3.2	Addressing the Volume Challenge . . . . .	71
3.2.1	Using MapReduce to Parallelize Blocking . . . . .	71
3.2.2	Meta-blocking: Pruning Pairwise Matchings . . . . .	77
3.3	Addressing the Velocity Challenge . . . . .	82
3.3.1	Incremental Record Linkage . . . . .	82
3.4	Addressing the Variety Challenge . . . . .	88
3.4.1	Linking Text Snippets to Structured Data . . . . .	89
3.5	Addressing the Veracity Challenge . . . . .	94
3.5.1	Temporal Record Linkage . . . . .	94
3.5.2	Record Linkage with Uniqueness Constraints . . . . .	100
<b>4.</b>	<b>BDI: Data Fusion . . . . .</b>	<b>107</b>
4.1	Traditional Data Fusion: A Quick Tour . . . . .	108
4.2	Addressing the Veracity Challenge . . . . .	109
4.2.1	Accuracy of a Source . . . . .	111
4.2.2	Probability of a Value Being True . . . . .	111
4.2.3	Copying Between Sources . . . . .	114
4.2.4	The End-to-End Solution . . . . .	120
4.2.5	Extensions and Alternatives . . . . .	123
4.3	Addressing the Volume Challenge . . . . .	126
4.3.1	A MapReduce-Based Framework for Offline Fusion . . . . .	126
4.3.2	Online Data Fusion . . . . .	127
4.4	Addressing the Velocity Challenge . . . . .	133
4.5	Addressing the Variety Challenge . . . . .	136
<b>5.</b>	<b>BDI: Emerging Topics . . . . .</b>	<b>139</b>
5.1	Role of Crowdsourcing . . . . .	139
5.1.1	Leveraging Transitive Relations . . . . .	140
5.1.2	Crowdsourcing the End-to-End Workflow . . . . .	144



5.1.3	Future Work . . . . .	146
5.2	Source Selection . . . . .	146
5.2.1	Static Sources . . . . .	148
5.2.2	Dynamic Sources . . . . .	150
5.2.3	Future Work . . . . .	153
5.3	Source Profiling . . . . .	153
5.3.1	The Bellman System . . . . .	155
5.3.2	Summarizing Sources . . . . .	157
5.3.3	Future Work . . . . .	160
<b>6.</b>	<b>Conclusions . . . . .</b>	<b>163</b>
	Bibliography . . . . .	165
	Authors' Biographies . . . . .	175
	Index . . . . .	177



## List of Figures

1.1	Traditional data integration: architecture. . . . .	9
1.2	K-coverage (the fraction of entities in the database that are present in at least $k$ different sources) for phone numbers in the restaurant domain [Dalvi et al. 2012]. . . . .	18
1.3	Connectivity (between entities and sources) for the nine domains studied by Dalvi et al. [2012]. . . . .	19
1.4	Consistency of data items in the Stock and Flight domains [Li et al. 2012]. . . . .	22
1.5	High-quality table on the web. . . . .	23
1.6	Contributions and overlaps between different types of web contents [Dong et al. 2014b].	27
2.1	Traditional schema alignment: three steps. . . . .	32
2.2	Attribute matching from <code>Airline1.Flight</code> to <code>Mediate.Flight</code> . . . . .	33
2.3	Query answering in a traditional data-integration system. . . . .	34
2.4	Example web form for searching flights at <code>Orbitz.com</code> (accessed on April 1, 2014). . . . .	50
2.5	Example web table (Airlines) with some major airlines of the world (accessed on April 1, 2014). . . . .	54
2.6	Two web tables ( <code>CapitalCity</code> ) describing major cities in Asia and in Africa from <code>nationsonline.org</code> (accessed on April 1, 2014). . . . .	58
2.7	Graphical model for annotating a 3x3 web table [Limaye et al. 2010]. . . . .	61
3.1	Traditional record linkage: three steps. . . . .	65
3.2	Pairwise matching graph. . . . .	67
3.3	Use of a single blocking function. . . . .	69
3.4	Use of multiple blocking functions. . . . .	70
3.5	Using MapReduce: a basic approach. . . . .	72
3.6	Using MapReduce: <code>BLOCKSPPLIT</code> . . . . .	74
3.7	Using schema agnostic blocking on multiple values. . . . .	79
3.8	Using meta-blocking with schema agnostic blocking. . . . .	81
3.9	Record linkage results on $\overline{\text{Flights}}_0$ . . . . .	84
3.10	Record linkage results on $\overline{\text{Flights}}_0 + \Delta \overline{\text{Flights}}_1$ . . . . .	85

3.11	Record linkage results on $\overline{\text{Flights}}_0 + \Delta\overline{\text{Flights}}_1 + \Delta\overline{\text{Flights}}_2$ . . . . .	85
3.12	Tagging of text snippet. . . . .	91
3.13	Plausible parses of text snippet. . . . .	92
3.14	Ground truth due to entity evolution. . . . .	95
3.15	Linkage with high value consistency. . . . .	96
3.16	Linkage with only name similarity. . . . .	97
3.17	$K$ -partite graph encoding. . . . .	103
3.18	Linkage with hard constraints. . . . .	104
3.19	Linkage with soft constraints. . . . .	104
4.1	Architecture of data fusion [Dong et al. 2009a]. . . . .	110
4.2	Probabilities of copying computed by AccuCOPY on the motivating example [Dong et al. 2009a]. An arrow from source $S$ to $S'$ indicates that $S$ copies from $S'$ . Copyings are shown only when the sum of the probabilities in both directions is over 0.1. . . . .	121
4.3	MapReduce-based implementation for truth discovery and trustworthiness evaluation [Dong et al. 2014b]. . . . .	126
4.4	Nine sources that provide the estimated arrival time for <i>Flight 49</i> . For each source, the answer it provides is shown in parenthesis and its accuracy is shown in a circle. An arrow from $S$ to $S'$ means that $S$ copies some data from $S'$ . . . . .	128
4.5	Architecture of online data fusion [Liu et al. 2011]. . . . .	129
4.6	Input for data fusion is two-dimensional, whereas input for extended data fusion is three-dimensional [Dong et al. 2014b]. . . . .	136
4.7	Fixing #provenances, (data item, value) pairs from more extractors are more likely to be true [Dong et al. 2014b]. . . . .	138
5.1	Example to illustrate labeling by crowd for transitive relations [Wang et al. 2013]. . . . .	141
5.2	Fusion result recall for the Stock domain [Li et al. 2012]. . . . .	147
5.3	Freshness versus update frequency for business listing sources [Rekatsinas et al. 2014]. . . . .	151
5.4	Evolution of coverage of the integration result for two subsets of the business listing sources [Rekatsinas et al. 2014]. . . . .	152
5.5	TPCE schema graph [Yang et al. 2009]. . . . .	158

## List of Tables

1.1	Sample data for Airline1.Schedule . . . . .	3
1.2	Sample data for Airline1.Flight . . . . .	3
1.3	Sample data for Airline2.Flight . . . . .	4
1.4	Sample data for Airport3.Departures . . . . .	4
1.5	Sample data for Airport3.Arrivals . . . . .	5
1.6	Sample data for Airfare4.Flight . . . . .	6
1.7	Sample data for Airfare4.Fares . . . . .	6
1.8	Sample data for Airinfo5.AirportCodes, Airinfo5.AirlineCodes . . . . .	6
1.9	Abbreviated attribute names . . . . .	7
1.10	Domain category distribution of web databases [He et al. 2007] . . . . .	16
1.11	Row statistics on high-quality relational tables on the web [Cafarella et al. 2008b] . . . . .	25
2.1	Selected text-derived features used in search rankers. The most important features are in italic [Cafarella et al. 2008a] . . . . .	56
3.1	Sample Flights records . . . . .	65
3.2	Virtual global enumeration in PAIRRANGE . . . . .	76
3.3	Sample Flights records with schematic heterogeneity . . . . .	78
3.4	Flights records and updates . . . . .	83
3.5	Sample Flights records from Table 3.1 . . . . .	89
3.6	Traveller flight profiles . . . . .	95
3.7	Airline business listings . . . . .	101
4.1	Five data sources provide information on the scheduled departure time of five flights. False values are in italics. Only S1 provides all true values. . . . .	109
4.2	Accuracy of data sources computed by AccuCopy on the motivating example . . . . .	122
4.3	Vote count computed for the scheduled departure time for <i>Flight 4</i> and <i>Flight 5</i> in the motivating example . . . . .	122
4.4	Output at each time point in Example 4.8. The time is made up for illustration purposes . . . . .	128

4.5	Three data sources updating information on the scheduled departure time of five flights. False values are in italic. . . . .	133
4.6	CEF-measures for the data sources in Table 4.5 . . . . .	135