

SpringerBriefs in Computer Science

Series Editors

Stan Zdonik

Peng Ning

Shashi Shekhar

Jonathan Katz

Xindong Wu

Lakhmi C. Jain

David Padua

Xuemin (Sherman) Shen

Borko Furht

V.S. Subrahmanian

Martial Hebert

Katsushi Ikeuchi

Bruno Siciliano

Sushil Jajodia

For further volumes:

<http://www.springer.com/series/10028>

Min Chen • Shiwen Mao • Yin Zhang
Victor C.M. Leung

Big Data

Related Technologies, Challenges
and Future Prospects

 Springer

Min Chen
School of Computer Science
and Technology
Huazhong University of Science
and Technology
Wuhan, China

Yin Zhang
School of Computer Science
and Technology
Huazhong University of Science
and Technology
Wuhan, China

Shiwen Mao
Auburn University
Auburn, AL, USA

Victor C.M. Leung
Electrical and Computer Engineering
The University of British Columbia
Vancouver, BC
Canada

ISSN 2191-5768

ISBN 978-3-319-06244-0

DOI 10.1007/978-3-319-06245-7

Springer Cham Heidelberg New York Dordrecht London

ISSN 2191-5776 (electronic)

ISBN 978-3-319-06245-7 (eBook)

Library of Congress Control Number: 2014937319

© The Author(s) 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

“How big is big?” Science writer Stephen Strauss asks in his fun book for kids titled *How Big is Big* and explains that “bigness is something no one can consume.”

In this book, we aim to answer this interesting question, but in the context of computer data. In the *big data* era, we are dealing with the explosive increase of global data and enormous datasets. Unlike seemingly similar terms such as “massive data” or “very big data,” *big data* refers to the datasets that could not be perceived, acquired, managed, and processed by traditional Information Technology (IT) and software/hardware tools within a tolerable time. It can be characterized by four Vs, i.e., Volume (great volume), Variety (various modalities), Velocity (rapid generation), and Value (huge value but very low density).

In this book, we provide a comprehensive overview of the background and related technologies, challenges and future prospects of big data. We first introduce the general background of big data and review related technologies, such as cloud computing, Internet of Things (IoT), data centers, and Hadoop. We then focus on the four phases of the value chain of big data, i.e., data generation, data acquisition, data storage, and data analysis. For each phase, we introduce the general background, discuss the technical challenges, and review the latest advances. We next examine the several representative applications of big data, including enterprise management, IoT, online social networks, healthcare and medical applications, collective intelligence, and smart grid. This book is concluded with a discussion of open problems and future directions. We aim to provide the readers a comprehensive overview and big-picture of this exciting area. We hope this monograph could be a useful reference for graduate students and professionals in related fields, and general readers who will benefit from an understanding of the big data field.

We are grateful to Dr. Xuemin (Sherman) Shen, the SpringerBriefs Series Editor on Wireless Communications. This book would not be possible without his kind support during the process. Thanks also to the Springer Editors and Staff, all of whom did their usual excellent job in getting this monograph published.

This work was supported by China National Natural Science Foundation (No. 61300224), the Ministry of Science and Technology (MOST), China, the International Science and Technology Collaboration Program (Project No.:

2014DFT10070), and the Hubei Provincial Key Project (No. 2013CFA051). Shiwen Mao's research is supported in part by the US National Science Foundation (NSF) under Grants CNS-1320664, CNS-1247955, CNS-0953513, and DUE-1044021, and through the NSF Broadband Wireless Access & Applications Center (BWAC) Site at Auburn University (NSF Grant IIP-1266036). The research of Victor Leung is supported by the Canadian Natural Sciences and Engineering Research Council, BC Innovation Council, Qatar Research Foundation, TELUS, and other industrial partners. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the foundation.

Wuhan, China
Auburn, AL
Wuhan, China
Vancouver, BC, Canada
January 2014

Min Chen
Shiwen Mao
Yin Zhang
Victor C.M. Leung

Contents

1	Introduction	1
1.1	Dawn of the Big Data Era	1
1.2	Definition and Features of Big Data	2
1.3	Big Data Value	5
1.4	The Development of Big Data	6
1.5	Challenges of Big Data	7
	References	9
2	Related Technologies	11
2.1	Cloud Computing	11
2.1.1	Cloud Computing Preliminaries	11
2.1.2	Relationship Between Cloud Computing and Big Data	12
2.2	IoT	13
2.2.1	IoT Preliminaries	13
2.2.2	Relationship Between IoT and Big Data	14
2.3	Data Center	15
2.4	Hadoop	16
2.4.1	Hadoop Preliminaries	16
2.4.2	Relationship between Hadoop and Big Data	17
	References	18
3	Big Data Generation and Acquisition	19
3.1	Big Data Generation	19
3.1.1	Enterprise Data	19
3.1.2	IoT Data	20
3.1.3	Internet Data	21
3.1.4	Bio-medical Data	21
3.1.5	Data Generation from Other Fields	22
3.2	Big Data Acquisition	23
3.2.1	Data Collection	23

3.2.2	Data Transportation	26
3.2.3	Data Pre-processing	27
References	30
4	Big Data Storage	33
4.1	Storage System for Massive Data	33
4.2	Distributed Storage System	35
4.3	Storage Mechanism for Big Data.....	37
4.3.1	Database Technology	38
4.3.2	Design Factors	44
4.3.3	Database Programming Model	45
References	48
5	Big Data Analysis	51
5.1	Traditional Data Analysis.....	51
5.2	Big Data Analytic Methods.....	53
5.3	Architecture for Big Data Analysis	55
5.3.1	Real-Time vs. Offline Analysis	55
5.3.2	Analysis at Different Levels.....	56
5.3.3	Analysis with Different Complexity	57
5.4	Tools for Big Data Mining and Analysis.....	57
References	58
6	Big Data Applications	59
6.1	Application Evolution	59
6.2	Big Data Analysis Fields	61
6.2.1	Structured Data Analysis	61
6.2.2	Text Data Analysis.....	61
6.2.3	Web Data Analysis	63
6.2.4	Multimedia Data Analysis.....	64
6.2.5	Network Data Analysis	65
6.2.6	Mobile Traffic Analysis	67
6.3	Key Applications	69
6.3.1	Application of Big Data in Enterprises	69
6.3.2	Application of IoT Based Big Data	70
6.3.3	Application of Online Social Network-Oriented Big Data	70
6.3.4	Applications of Healthcare and Medical Big Data	73
6.3.5	Collective Intelligence	74
6.3.6	Smart Grid	75
References	76

- 7 Open Issues and Outlook** 81
 - 7.1 Open Issues 81
 - 7.1.1 Theoretical Research 81
 - 7.1.2 Technology Development 83
 - 7.1.3 Practical Implications 84
 - 7.1.4 Data Security 84
 - 7.2 Outlook 86
- References 89

Acronyms

AMI	Advanced Metering Infrastructure
APT	Advanced Persistent Threat
BI	Business Intelligence
BLOB	Binary Large Object or Basic Large Object
BPM	Business Process Management
BSON	Binary JSON
CEO	Chief Executive Officers
CIO	Chief Information Officer
DAS	Direct Attached Storage
DMA	Direct Memory Access
ETL	Extract, Transform and Load
ERCIM	European Research Consortium for Informatics and Mathematics
GUI	Graphic User Interface
HDFS	Hadoop Distributed File System
HGP	Human Genome Project
HQL	HyperTable Query Language
ICT	Information and Communications Technology
IDC	International Data Corporation
IoT	Internet of Things
IT	Information Technology
LHC	Large Hadron Collider
Libpcap	Packet Capture Library
MMF	Multi-Mode Fiber
MPI	Message Passing Interface
MR	MapReduce
MVCC	Muti-Version Concurrency Control
NAS	Network Attached Storage
NER	Named Entity Recognition
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NSF	National Science Foundation

OFDM	Orthogonal Frequency-Division Multiplexing
OLAP	On-Line Analytical Processing
OpenMP	Open Multi-Processing
PB	Petabyte
PMU	Phasor Measurement Unit
PNUTS	Platform for Nimble Universal Table Storage
RAID	Redundant Array of Independent Disks
RDBMS	Relational Database Management System
SAN	Storage Area Network
SDK	Software Development Kit
SDSS	Sloan Digital Sky Survey
SNS	Social Networking Services
SSD	Solid-State Drive
TB	Terabyte
TOMS	Topic-oriented Multimedia Summarization System
TOR	Top Rack Switches
URL	Uniform Resource Locator
WDM	Wavelength Division Multiplexing
ZC	Zero-copy