Inf Technol Tourism (2015) 15:283–285
DOI 10.1007/s40558-015-0027-y

BOOK REVIEW

Big data related technologies, challenges and future prospects

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

New York: Springer, 2014
ISBN 978-3-319-06244-0 (printed book), 978-3-319-06245-7 (e-book), 89 pp, USD54.99 (printed book), USD39.99 (e-book)

Gang Li[1]

Received: 26 June 2015 / Revised: 27 June 2015 / Accepted: 30 June 2015 /
Published online: 18 July 2015

Big data is the large or complex data that exceed the processing capacity of conventional data processing systems. This book provides a big picture in this broad research area, covering all the phases of its value chains. The authors have attempted to survey most of the relevant technologies in each phrase of big data. The book is recommended for readers interested in advanced research in big data, also for industry practitioners who are interested in building big data applications. If the reader is not with necessary technical background, complementary readings may be needed.

The book contains three parts: Chapter 1–2 introduce the preliminary knowledge on big data. Chapters 3–5 cover four phases of the value chains of big data. Chapters 6–7 present key big data applications and discuss future research and application trend.

Chapter 1 defines the concepts and features of big data, then reviews the history of big data, and discussed the challenges in the development of big data application. By clarifying the difference between 'big' data and "massive data" or "very big data", the authors state that the volume of a dataset is not the only criterion for big data, and big data also features with rapid generation, various data types and huge hidden values.

Chapter 2 gives a preliminary knowledge of fundamental technologies in big data, and examines their relationship with big data. Among these technologies, introduction to the Internet of Things (IoT) describes the data generated from all

✉ Gang Li
gangli@acm.org; gang.li@deakin.edu.au

[1] School of Information Technology, Deakin University, Burwood, VIC 3125, Australia

kinds of devices; discussion on cloud computing and the operation of data center explain how the data resource are managed, stored, and processed over shared resources. Finally, the authors introduce Hadoop, a programming model that provides a systematic solution on different stages, including data storage, system management and data processing.

Chapter 3 focuses on data generation and acquisition. After the data are collected from different operational departments, they go through an integration process. During this process, noise, error and redundancy are removed. The pre-processing strategy needs to be selected carefully regarding the problem and performance requirement. I would add here that, compared to conventional data set, data collection and pre-processing could be the most daunting and time-consuming for big data. For example, with web data, the web crawler often needs to be guided by powerful data mining methods in order to gather most useful data.

Chapter 4 focuses on big data storage, which includes both hardware infrastructure and storage mechanisms. Appropriate Hardware infrastructure can provide reliable storage capability and maximize system scalability. On the other hand, when considering storage mechanisms, you need to select suitable databases and programming models for timely access and analysis. With hardware infrastructure, the authors discussed the storage system for massive data, and distributed storage system. It is not clear what the relationship is between these systems, and what is the guideline in choosing them.

Chapter 5 focuses on big data analysis. Top five data analysis tools are presented based on a 2012 KDNuggets report. I would add here that Python is gaining more popularity in recent years, because of its rich libraries such as Matplotlib, Scikit-learn and pandas, that provide powerful data manipulation, analysis and visualization capability.

Chapter 6 focuses on big data applications, and presents several key big data applications, e.g. application of enterprise data. Chapter 6.2 presents a variety of data analysis approaches that have applied to different data sources. However, I believe that text mining and web mining, should not be considered as different application areas. The complexity of data in these areas actually had led to the invention of new specialized algorithms, and this section may be better in Chapter 5.

Chapter 7 summarizes research issues on big data, and discusses both research and application trends. It is interesting to note that privacy and safety has drawn great attention. The authors then present us the opportunity big data may further bring to us. The fast growth of data volumes has transformed modern programs from algorithm-intensive to data-intensive. The big data will continue to drive the progress of technology, change the social and economic life, and even influence everyone's ways of living.

The book presents an overview on different phrases of big data value chain. The authors also summarized the opportunities and challenges with big data. I would like to recommend this book to researchers and industrial practitioners who work with big data, and it is also useful for researchers in relevant areas such as parallel computing, NoSQL Database, machine learning and data analytics.

Due to the complexity of the topic and the research areas it covers, a reader might prefer the authors to present main ideas in increasing order of complexity. However,

I believe the book could benefit from providing more background on the meaning of big data, as well as concrete examples on big data applications. This will especially benefit readers who are new to this area. In addition, more emphasis is expected to be put on data analytics, and how to incorporate distributed computing to data analysis process. These technologies will play essential roles in real-time big data analytics that help uncovered the hidden insight for timely decision.

Readers who are new to big data may benefit from Mayer-Schönberger's book (Mayer-Schönberger and Cukier 2013), which provides a lighter introduction on big data, and focus on how big data might impact business, government and individuals. If a researcher or practitioner look for technologies that is used in a specific phrase of a big data application, he or she will need to refer to books on specific tools. For example, Richert's book (Richert 2013) walks you through using Python in analysis phrase; whereas White's book (White 2012) teaches how to build scalable distributed big data systems with Hadoop.

References

Mayer-Schönberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, Boston

Richert W (2013) Building machine learning systems with python. Packt Publishing Ltd, Birmingham

White T (2012) Hadoop: the definitive guide. O'Reilly Media Inc, Beijing