Book Review

Healthc Inform Res. 2014 January;20(1):76-78.
http://dx.doi.org/10.4258/hir.2014.20.1.76
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Big Data Management, Technologies, and Applications

Seewon Ryu, PhD
Department of Health Policy and Healthcare Management, Inje Institute of Advanced Studies, Seoul, Korea
seewon@inje.ac.kr

Author: Wen-Chen Hu and Naima Kaabouch
Year: 2013
Publisher: IGI Global
ISBN: 978-1-666-4699-5

Big data in healthcare is a hot issue as in other fields as well. With the continuous increase of digital data, which is creating in the process of medical services and health management, big data management and analysis is becoming important. Researchers in medical services and health management enclosed within the statistical methods strictly. For example, requirement of randomness and size of sample were limited in the association studies. Big data analysis may liberate researchers from these limitations and introduce new world of analysis and research. Prediction and trend awareness of disease are typical examples of many tentative big data applications in healthcare area.

Although big data has shown us many useful ways of thinking and application cases by innovative methods, there are many issues about collection and analysis for big data. Even though the future of big data is bright, traditional IT technologies are not able to handle this kind of data anymore because of its vast size, constant changes, and high complexity. For these reasons, other technologies have to be created or used to manage big data, which is complex, unstructured, or semi-structured. Therefore, practitioners and researchers in healthcare area are expecting books that can help them understand big data and learn effective big data methods. However, few books are able to meet the readers' needs about big data at this moment.

This book is a timely and urgently needed publication. It discusses various issues related to a big data management, technologies, and applications from a technological perspective. Readers are able to learn fundamental big data knowledge from this book and apply the learned knowledge to their big data problems. It introduces alternative management and processing methods for big data handling from sixty world-renowned scholars and industry professionals. The book includes research and development results of lasting significance in the theory, design, implementation,

analysis, and application of big data, and other critical issues.

Big data exists in a wide variety of data-intensive areas such as atmospheric science, genome research, astronomical studies, and network traffic monitor. This book does not target any specific areas. It is a generic big data book. Therefore, a broader audience could benefit from this book. It especially benefits the IT and data management personnel of the big corporations, such as hospital and health insurance corporations, which face a great influx of data. This book will help them smoothly build efficient and effective big data systems based on their traditional IT knowledge. It could be used for a textbook or a reference book of an advanced healthcare IT course. Since this book covers the big data subject systematically, it is also helpful for people who desire to learn the big data topics on their own.

This book provides rich topics of big data management, technologies, and applications. It is unique among those available big data books because of its great depth and technical approach. This book provides timely, critical management methods, technologies, and applications of big data to IT workers and students. It contains seventeen chapters divided into four sections:

Section 1: big data technologies, methods, and algorithms,
Section 2: big data storage, management, and sharing,
Section 3: specific big data, and
Section 4: big data and computer systems and big data benchmarks.

Section 1 consists of four chapters and discusses some of them including a survey, the k-means algorithm, synchronizing execution, and data reduction. Chapter 1 gives a review and analysis of several key big data technologies including: MapReduce, NOSQL, MPP (Massively Parallel Processing), and in memory databases. Chapter 2 proposes a distributed version of the k-means clustering algorithm for big data mining. It is based on three kinds of software: 1) Apache Hadoop software library, 2) Hadoop Distributed File System (HDFS), and 3) MapReduce. Chapter 3 discusses synchronous parallelization of big data analytics over a distributed environment to optimize performance. Chapter 4 first examines the importance of data reduction techniques for the analysis of big datasets and then presents several basic reduction techniques in detail, stressing on the advantages and disadvantages of each.

Section 2 discusses various issues related to the three themes including sampling, warehouse design, warehousing, and sharing. Chapter 5 first reviews traditional sampling techniques and then suggests adaptations relevant to big data studies of text downloaded from online media, such as email messages, online gaming, blogs, micro-blogs like Twitter, and social networking websites like Facebook. Chapter 6 presents a data warehouse design methodology based on a hybrid approach, which adopts a graph-based multidimensional model. In order to automate the whole design process, the methodology has been implemented using logical programming. Chapter 7 presents an algorithm called Cache Join (CACHEJOIN), which performs asymptotically at least as well as MESHJOIN but performs better in realistic scenarios, particularly if parts of the master data are used with different frequencies. Chapter 8 firstly reviews literature on big data sharing practices using current technology. The second part presents case studies on disciplinary data repositories in terms of their requirements and policies.

Section 3 covers five specific kinds of big data: astronomical telescopes, social networks, digital humanities, geography, and sensor networks. Chapter 9 first discusses the big data challenges in constructing data management systems for astronomical instruments and then suggests open source solutions to them based on software from the Apache Software Foundation including Apache Object-Oriented Data Technology (OODT), Tika, and Solr. Chapter 10 focuses on big data analytics techniques. Developing adequate big data analysis techniques may help to improve the decision-making process and minimize risks by unearthing valuable insights that would otherwise remain hidden. An automated decision-making software can be provided by using big data analytics to automatically fine-tune inventories in response to real-time sales. Chapter 11 introduces Big Data at Scale for Digital Humanities: An Architecture for the HathiTrust Research Center. The HathiTrust Research Center (HTRC) is a cyber infrastructure to support humanities research on big humanities data including the following functions: to make the content easy to find, to make the research tools efficient and effective, to allow researchers to customize their environment, to allow researchers to combine their own data with that of the HTRC, and to allow researchers to contribute tools. The architecture has multiple layers of abstraction providing a secure, scalable, extendable, and generalizable interface for both human and computational users. Chapter 12 proposes the GeoBase, which enables querying over scientific data by improving end-to-end support through two integrated, native components: a linearization-based index to enable rich scientific querying on multidimensional data and a plugin that interfaces key-value stores with array-based binary file formats. Chapter 13 based on several real-world applications, this chapter discusses the challenges involved in large-scale sensor data analysis, and describes practical solutions to address them. Due to the sheer size of the data and the large amount of computation involved, these are clearly

"Big Data" applications.

Section 4 presents big data and computer systems and big data benchmarks and consists of four chapters. The first three chapters in this section are related to computer systems, including graphics processors, hardware selection, and excess entropy. The last chapter presents benchmarking big data workloads. Chapter 14 proposes techniques of accelerating large-scale genome-wide association studies (GWAS) by using graphics processors (GPUs). Large-scale GWAS are a big data application due to the great amount of data to process and high computation intensity and GPUs have been used to accelerate genomic data analytics like minor allele frequency computation. Chapter 15 explores some of the issues at the intersection of cloud computing, metadata, and big data. Chapter 16 applies entropy to large computer systems and shows how entropy, a single concept, can identify problematic groups of servers, strange patterns in load, and changes in composition with minimal human involvement. Chapter 17 looks into various techniques and measures the effectiveness of hardware and software platforms dealing with big data. It reviews system benchmark standards and looks ahead towards an industry standard for benchmarking big data workloads, such as the Transaction Processing Performance Council (TPC) and the Standard Performance Evaluation Corporation (SPEC).

This book provides the most up-to-date, crucial, and practical information for big data management, technologies, and applications. If any researchers and scholars, and workers, who are in healthcare sector with big data in mind, the book will help them understand various big data issues and apply the proposed big data methods to their problems. Collected emerging articles about big data methodologies and technologies also would be beneficial for them.

http://dx.doi.org/10.4258/hir.2014.20.1.76