Preface

This book collects high-quality research papers and industrial and practice articles in the areas of big data management, technologies, and applications from academics and industries. It includes research and development results of lasting significance in the theory, design, implementation, analysis, and application of big data, and other critical issues. Seventeen excellent chapters from sixty world-renowned scholars and industry professionals are included in this book, which covers four themes: (1) big data technologies, methods, and algorithms, (2) big data storage, management, and sharing, (3) specific big data, and (4) big data and computer systems and big data benchmarks. This preface presents this book, gives essential big data management, technologies, and applications, introduces each section and chapter of this book, and summarizes the discussions.

INTRODUCTION

The growth of information size is not linear, but exponential. For example, at least one million Web pages are added to the Internet every day, a massive amount of genetic data is created from various genome projects, or vast astronomical data is recorded after studying numerous galaxies. Big data is one of the hottest IT topics these days because many opportunities and great revenue are behind it based on the following reports:

- Big data is a major driver of IT spending these days; for example, $232 billion will be spent on IT including information management and analytics infrastructure from 2012 through 2016 according to Gartner (Beyer, Lovelock, Sommer, & Adrian, 2012).
- IDC (2012) predicted the worldwide market of big data technology and services will grow from $3.2 billion in 2010 to $16.9 billion in 2015, which represents a CAGR (Compound Annual Growth Rate) of 40%. For example, it reported big data storage had the strongest growth rate, growing at 61.4% annually.
- Two observations from Kelly, Floyer, Vellante, and Miniman (2013) are (1) factory revenue generated by the sale of big data-related hardware, software, and services growing by 59% in 2012 over 2011, and (2) the total big data market having an average 31% CAGR over the five-year period from 2012 ($11.4 billion) to 2017 ($47 billion).

Even though the future of big data is bright, traditional IT technologies such as files and relational databases are not able to handle this kind of data anymore because of its vast size, constant changes, and high complication. Other technologies have to be created or used to manage big data, which is complex, unstructured, or semi-structured. Therefore, IT workers and students look forward to books that can help them understand big data and learn effective big data methods. Unfortunately, very few big data books are able to meet the readers' needs at this moment. This is a just-in-time book. It discusses various issues related to big data management, technologies, and applications from a technological perspective. Readers learn fundamental big data knowledge from this book and are able to apply the learned knowledge to their big data problems.

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. The intended audience includes IT industrialists, students, educators, and researchers with big data in mind. It especially benefits the IT personnel of the big corporations, which face a great influx of data. This book will help IT workers smoothly build efficient and effective big data systems based on their traditional IT knowledge. It could be used for a textbook of an advanced IT (or related disciplines) course and could be a reference book for IT professionals and students. Since this book covers the big data subject systematically, it is also for people desiring to learn the big data topics on their own.

This book provides rich topics of big data management, technologies, and applications. This preface is to introduce this book and suggest essential big data management methods, technologies, and applications. The rest of this preface is organized as follows:

Section 2 - Essential Big Data Management, Technologies, and Applications: This section discusses the essential big data management, technologies, and applications. It includes the following steps: (1) big data generation, capturing, and collection, (2) big data storage and preservation, (3) big data analytics, management, visualization, and sharing, and (4) big data applications and other related topics. Each step and its corresponding topics will be introduced in this section.

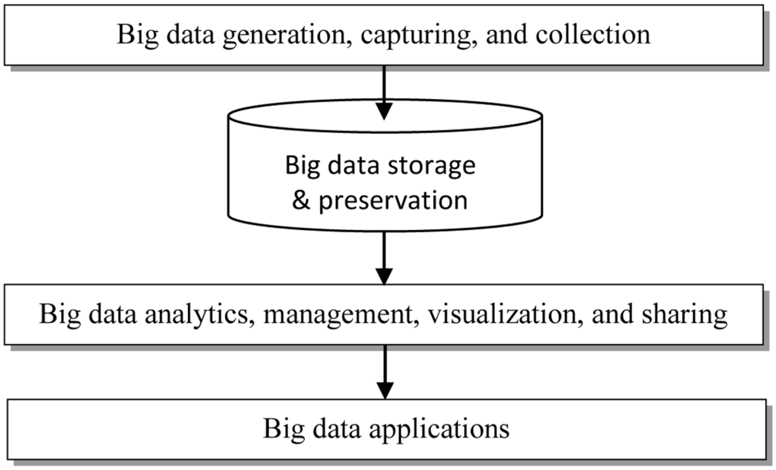
Section 3 - Organization of this Book: This book consists of seventeen chapters divided into four sections: (1) big data technologies, methods, and algorithms, (2) big data storage, management, and sharing, (3) specific big data, and (4) big data and computer systems and big data benchmarks. Each section and chapter will be briefly introduced in this section.

Section 4 - Summary: The last section summarizes the management, technologies, and applications of big data discussed in this preface.

ESSENTIAL BIG DATA MANAGEMENT, TECHNOLOGIES, AND APPLICATIONS

Big data covers a wide variety of subjects and methods. This section tries to introduce essential big data management, technologies, and applications by using the following steps: (1) big data generation, capturing, and collection, (2) big data storage and preservation, (3) big data analytics, management, visualization, and sharing, and (4) big data applications and other related topics as shown in Figure 1. Each step is explained next.

Figure 1. A flowchart of generic big data management


Big data generation, capturing, and collection
Big data storage & preservation
Big data analytics, management, visualization, and sharing
Big data applications

Big Data Generation, Capturing, and Collection

This is the first step of big data management including three actions: big data generation, capturing, and collection, which are briefly introduced as follows:

- **Big Data Generation:** During the time when computers were not popular, big data was rare. Since the number of (embedded) computers was greatly increased in the '80s, data is generated explosively. Big data can be generated from many sources; for example, all kinds of sensors, customer purchasing data, astronomical data, and texting messages.
- **Big Data Capturing:** Big data may be generated continuously (like steady satellite image transmission) or abruptly (during the peak hours). Compared to continuously generated big data, capturing abruptly generated big data is a challenge.
- **Big Data Collection:** Not all captured data is worth collection because of limited storage and processing power. For example, the size of videos of traffic monitor could be huge. Instead of saving the whole videos, specific video frames are selected and saved.

Big Data Storage and Preservation

After collecting big data, the next step is to store and preserve it. Because of its high volume, velocity, and variety, it is not a trivial task of storing and preserving big data. Big data storage and preservation are shortly explained below:

- **Big Data Storage:** The size of big data is huge and large scalable storage is required for storing it. Many times datacenters and warehousing are used and data structures are tailored for specific big data.
- **Big Data Preservation:** Big data has three key features: large volume, great variety, and high velocity. The feature of high velocity makes big data preservation volatile and complicated.

Big Data Analytics, Management, Visualization, and Sharing

Before big data can be put into use, it might need to be processed first. Various big data processing methods are available. Four major ones are analytics, management, visualization, and sharing introduced as follows:

- **Big Data Analytics:** It is to examine big data and uncover its hidden information. Examples of using the uncovered information include weather forecasts and economic indicators. Tools for big data analytics include NoSQL databases, Hadoop, and MapReduce.
- **Big Data Management:** After big data is collected, it needs to be well managed and maintained. There are many kinds of big data management methods. Some of them are organizing, searching, processing, mining big data.
- **Big Data Visualization:** Reading big data item by item is not feasible. Visualization tools or functions must be provided so data can be searched, viewed, and managed easily and collectively.
- **Big Data Sharing:** Many issues, like privacy and security, are related to big data sharing. Additionally, big data sharing is considered a hard problem because of its huge size. Cloud computing may relieve this problem.

Big Data Applications and Other Related Issues

Results of big data processing can be applied to many areas like businesses and sciences and can be used in many ways like increasing revenue and inventing new drugs. Other related critical big data issues worth mentioning such as privacy and security are given as follows:

- **Big Data Applications:** Most data-intensive areas could be the candidates for big data applications. Some of the examples are (1) data from various sensors for weather forecasts, (2) data from numerous traffic monitors for transportation control, and (3) customer purchasing patterns for revenue discovery.
- **Big Data Privacy and Security:** Without rigorous privacy and security control, big data could not flourish. Strict privacy encourages big data collection and high security assures the safety of big data.
- **Big Data Standards, Policies, and Benchmarks:** Big data is a fairly new research subject. Therefore, its standards, policies, and benchmarks are still developing and investigated.
- **Cloud, Green, and Mobile Computing for Big Data:** Many newest computing paradigms could be used by big data. Among them are (1) cloud computing for sharing big data, (2) green computing for saving time and energy, and (3) mobile computing for accessing big data from anywhere and anytime.

ORGANIZATION OF THE BOOK

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: (1) big data technologies, methods, and algorithms, (2) big data storage, management, and sharing, (3) specific big data, and (4) big data and computer systems and big data benchmarks. Each section and chapter is briefly introduced next.

Big Data Technologies, Methods, and Algorithms

Various technologies, methods, and algorithms are available for big data. This section discusses some of them including a survey, the K-means algorithm, synchronizing execution, and data reduction:

Chapter 1 - Technologies for Big Data: The author, Kapil Bakshi from Cisco Systems, Inc., gives a review and analysis of several key big data technologies including: MapReduce, NOSQL, MPP (Massively Parallel Processing), and in memory databases.

Chapter 2 - Applying the K-Means Algorithm in Big Raw Data Sets with Hadoop and MapReduce: The authors propose 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, a framework for distributed processing of large data sets, (2) Hadoop Distributed File System (HDFS), a distributed file system that provides high-throughput access to data-driven applications, and (3) MapReduce, a software framework for distributed computing of large data sets.

Chapter 3 - Synchronizing Execution of Big Data in Distributed and Parallelized Environments: In order to ensure fast and accurate execution of big data analytics, the computing capability of loosely-coupled distributed infrastructure needs to be maximally leveraged. This chapter discusses synchronous parallelization of big data analytics over a distributed environment to optimize performance.

Chapter 4 - Parallel Data Reduction Techniques for Big Datasets: The major mission of data reduction is to save time and bandwidth in enabling users to deal with larger datasets even in minimal resource environments. This chapter 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.

Big Data Storage, Management, and Sharing

Three themes, storage, management, and sharing, are critical to big data. How to store and share big data is challenging because of its high volume. Furthermore, the potentials of big data cannot be fully discovered if it is not properly managed. This section discusses various issues related to the three themes including sampling, warehouse design, warehousing, and sharing:

Chapter 5 - Techniques for Sampling Online Text-Based Data Sets: The chapter 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 - Big Data Warehouse Automatic Design Methodology: This chapter 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 - Big Data Management in the Context of Real-Time Data Warehousing: In order to make timely and effective decisions, businesses need the latest information from big data warehouse repositories. A well-known algorithm called Mesh Join (MESHJOIN) is to process stream data with disk-based master data, which uses limited memory. This chapter 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 - Big Data Sharing among Academics: The first part of this chapter 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. It describes and compares such requirements and policies at disciplinary repositories in three areas: Dryad for life science, Interuniversity Consortium for Political and Social Research (ICPSR) for social science, and the National Oceanographic Data Center (NODC) for physical science.

Specific Big Data

Big data is prevailing. Each application has its unique big data and requires a particular method to process it. This section covers five specific kinds of big data: astronomical telescopes, social networks, digital humanities, geography, and sensor networks:

Chapter 9 - Scalable Data Mining, Archiving, and Big Data Management for the Next Generation Astronomical Telescopes: This chapter 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 - Efficient Metaheuristic Approaches for Exploration of Online Social Networks: This study 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 - Big Data at Scale for Digital Humanities: An Architecture for the HathiTrust Research Center: The HathiTrust Research Center (HTRC) is a cyberinfrastructure 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 - GeoBase – Indexing NetCDF Files for Large-Scale Data Analysis: The author, Tanu Malik, 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 multi-dimensional data and a plugin that interfaces key-value stores with array-based binary file formats.

Chapter 13 - Large-Scale Sensor Network Analysis – Applications in Structural Health Monitoring: 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.

Big Data and Computer Systems and Big Data Benchmarks

Big data is complicated and its size is huge. It requires the services from high performance computer systems. The first three chapters in this section are related to computer systems, including graphics processors, hardware selection, and excess entropy. The last chapter, benchmarking big data workloads, does not belong to any topics of the four sections and is put here. The four chapters are briefly described as follows:

Chapter 14 - Accelerating Large-Scale Genome-Wide Association Studies with Graphics Processors: Large-scale Genome-Wide Association Studies (GWAS) are a big data application due to the great amount of data to process and high computation intensity and Graphics Processors (GPUs) have been used to accelerate genomic data analytics like Minor Allele Frequency (MAF) computation. This chapter proposes techniques of accelerating MAF computation by using GPUs.

Chapter 15 - The Need to Consider Hardware Selection When Designing Big Data Applications Supported by Metadata: The selection of hardware to support big data systems is complex. The trend of cloud computing has emerged as an effective method of leasing compute time and metadata enables many applications and users to access datasets and effectively use them without relying on extensive knowledge from humans about the data. This chapter explores some of the issues at the intersection of cloud computing, metadata, and big data.

Chapter 16 - Excess Entropy in Computer Systems: Entropy is well researched in physics and used in economics. This research applies it to large computer systems. The author, Charles Loboz, 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 - A Review of System Benchmark Standards and a Look Ahead Towards an Industry Standard for Benchmarking Big Data Workloads: Transaction Processing Performance Council (TPC) and the Standard Performance Evaluation Corporation (SPEC) have developed several industry standards for performance benchmarking for various applications, including big data applications. This chapter looks into various techniques and measures the effectiveness of hardware and software platforms dealing with big data.

SUMMARY

Big data has existed for a long time, but it did not catch great attention until recently because it did not prevail. However, owing to the high popularity of computers and great IT advancements, big data is everywhere. A tremendous amount of data is generated every day, everywhere, from fields such as businesses, research, and sciences. The high volume, velocity, and variety of data cause a great headache to people because the traditional methods are no longer working for this kind of data. A book covering big data from a technological perspective is in need. Unfortunately, not many books about big data are available at this moment (2013). This book, *Big Data Management, Technologies, and Applications*, is unique among those big-data books because of its great depth and technical approach. It consists of four themes: (1) big data technologies, methods, and algorithms, (2) big data storage, management, and sharing, (3) specific big data, and (4) big data and computer systems and big data benchmarks. It is a timely and urgently needed publication, and it provides the most up-to-date, crucial, and practical information for big data management, technologies, and applications. It is a must-read book for IT students, researchers, scholars, and workers with big data in mind.

This book covers various important issues of big data from fundamental knowledge to advanced algorithms. There are many advantages provided by this unique, much needed big data book. Some of the advantages are (1) covering various critical issues related to big data management, technologies, and applications, (2) helping IT students and workers gain essential knowledge of big data, (3) assisting researchers and professionals to master big data technologies, and (4) providing a just-in-time textbook for a course of big data. Additionally, it could be used as a reference book for data-intensive workers and students, who are interested in big data. Overall, this book will help readers better understand big data and apply the methods learned from this book to real world problems. Hope you enjoy reading it.

Wen-Chen Hu
University of North Dakota, USA

Naima Kaabouch
University of North Dakota, USA

REFERENCES

Beyer, M. A., Lovelock, J.-D., Sommer, D., & Adrian, M. (2012, October 12). *Big data drives rapid changes in infrastructure and $232 billion in IT spending through 2016*. Retrieved June 12, 2013, from http://www.gartner.com/id=2195915

IDC. (2012, March 7). *IDC releases first worldwide big data technology and services market forecast, shows big data as the next essential capability and a foundation for the intelligent economy*. Retrieved May 4, 2013, from http://www.idc.com/getdoc.jsp?containerId=prUS23355112

Kelly, J., Floyer, D., Vellante, D., & Miniman, S. (2013, April 17). *Big data vendor revenue and market forecast 2012-2017*. Retrieved May 22, 2013, from http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017