HALF TITLE PAGE

SERIES PAGE

TITLE PAGE

LOC PAGE

Contents

Section I **Big Data Management**

Section II **Big Data Processing**

Section V **Big Data Applications**

Please provide data.

Foreword

Preface

As data sets are being generated at an exponential rate all over the world, *big data* has become an indispensable issue. While organizations are capturing exponentially larger amounts of data than ever these days, they have to rethink and figure out what to digest it. The implicit meaning of data can be interpreted in reality through novel and evolving algorithms, analytics techniques, and innovative and effective use of hardware and software platforms so that organizations can harness the data, discover hidden patterns, and use newly acquired knowledge to act meaningfully for competitive advantages.

This challenging vision has attracted a great deal of attention from the research community, which has reacted with a number of proposals focusing on *fundamental issues*, such as *managing big data*, *querying and mining big data*, *making big data privacy-preserving*, designing and running *sophisticated analytics over big data*, and *critical applications*, which span over a large family of cases, from *biomedical (big) data* to *graph (big) data*, from *social networks* to *sensor* and *spatiotemporal stream networks*, and so forth.

A conceptually relevant point of result that inspired our research is recognizing that classical managing, query, and mining algorithms, even developed with very large data sets, are not suitable to cope with big data, due to both methodological and performance issues. As a consequence, there is an emerging need for devising innovative models, algorithms, and techniques capable of managing and mining big data while dealing with their inherent properties, such as *volume*, *variety*, and *velocity*.

Inspired by this challenging paradigm, this book covers *fundamental and realistic issues about big data*, including efficient algorithmic methods to process data, better analytical strategies to digest data, and representative applications in diverse fields such as medicine, science, and engineering, seeking to bridge the gap between huge amounts of data and appropriate computational methods for scientific and social discovery and to bring technologies for media/data communication, elastic media/data storage, cross-network media/data fusion, SaaS, and others together. It also aims at interesting applications involving big data.

If “SaaS” is an abbreviation, please define at first mention.

According to this methodological vision, this book is organized into five main sections:

- “Big Data Management,” which focuses on research issues related the effective and efficient management of big data, including *indexing* and *scalability* aspects.
- “Big Data Processing,” which moves the attention to the problem of processing big data in a widespread collection of resource-intensive computational settings, for

example, those determined by *MapReduce environments*, *commodity clusters*, and *data-preponderant networks*.

- “Big Data Stream Techniques and Algorithms,” which explores research issues concerning the management and mining of big data in *streaming environments*, a typical scenario where big data show their most problematic drawbacks to deal with—here, the focus in on how to manage big data *on the fly*, with limited resources and approximate computations.
- “Big Data Privacy,” which focuses on models, techniques, and algorithms that aim at making big data *privacy-preserving*, that is, protecting them against *privacy breaches* that may prevent the anonymity of big data in conventional settings (e.g., *cloud environments*).
- “Big Data Applications,” which, finally, addresses a rich collection of practical applications of big data in several domains, ranging from *finance applications* to *multimedia tools*, from *biometrics applications* to *satellite (big) data processing*, and so forth.

In the following, we will provide a description of the chapters contained in the book, according to the previous five sections.

The first section (i.e., “Big Data Management”) is organized into the following chapters.

Chapter 1, “Scalable Indexing for Big Data Processing,” by Hisham Mohamed and Stéphane Marchand-Maillet, focuses on the *K-nearest neighbor* (*K*-NN) search problem, which is the way to find and predict the most closest and similar objects to a given query. It finds many applications for information retrieval and visualization, machine learning, and data mining. The context of big data imposes the finding of approximate solutions. Permutation-based indexing is one of the most recent techniques for approximate similarity search in large-scale domains. Data objects are represented by a list of references (pivots), which are ordered with respect to their distances from the object. In this context, the authors show different distributed algorithms for efficient indexing and searching based on permutation-based indexing, and evaluate them on big high-dimensional data sets.

Chapter 2, “Scalability and Cost Evaluation of Incremental Data Processing using Amazon’s Hadoop Service,” by Xing Wu, Yan Liu, and Ian Gorton, considers the case of *Hadoop* that, based on the *MapReduce* model and *Hadoop Distributed File System* (HDFS), enables the distributed processing of large data sets across clusters with scalability and fault tolerance. Many data-intensive applications involve continuous and incremental updates of data. Understanding the scalability and cost of a Hadoop platform to handle small and independent updates of data sets sheds light on the design of scalable and cost-effective data-intensive applications. With these ideas in mind, the authors introduce a motivating movie recommendation application implemented in the MapReduce model and deployed on *Amazon Elastic MapReduce* (EMR), a Hadoop service provided by Amazon. In particular, the authors present the deployment architecture with implementation details of the Hadoop application. With metrics collected by *Amazon CloudWatch*, they present an empirical scalability and cost evaluation of the *Amazon Hadoop* service on processing continuous and incremental data streams. The evaluation result highlights the potential of auto-scaling for cost reduction on Hadoop services.

Chapter 3, "Singular Value Decomposition, Clustering, and Indexing for Similarity Search for Large Data Sets in High-Dimensional Spaces," by Alexander Thomasian, addresses a popular paradigm, that is, representing objects such as images by their feature vectors and searching for similarity according to the distances of the points representing them in high-dimensional space via *K-nearest neighbors* (*K*-NNs) to a target image. The authors discuss a combination of *singular value decomposition* (SVD), clustering, and indexing to reduce the cost of processing *K*-NN queries for large data sets with high-dimensional data. They first review dimensionality reduction methods with emphasis on SVD and related methods, followed by a survey of clustering and indexing methods for high-dimensional numerical data. After, the authors describe combining SVD and clustering as a framework and the main memory-resident *ordered partition* (*OP*)*-tree index* to speed-up *K*-NN queries. Finally, they discuss techniques to save the OP-tree on disk and specify the *stepwise dimensionality increasing* (SDI) index suited for *K*-NN queries on dimensionally reduced data.

Chapter 4, "Multiple Sequence Alignment and Clustering with Dot Matrices, Entropy, Genetic Algorithms," by John Tsiligaridis, presents a set of algorithms and their efficiency for *Multiple Sequence Alignment* (MSA) and *clustering problems*, including also solutions in distributive environments with Hadoop. The strength, the adaptability, and the effectiveness of the *genetic algorithms* (GAs) for both problems are pointed out. MSA is among the most important tasks in computational biology. In biological sequence comparison, emphasis is given to the simultaneous alignment of several sequences. GAs are stochastic approaches for efficient and robust search that can play a significant role for MSA and clustering. The *divide-and-conquer* principle ensures undisturbed consistency during vertical sequences' segmentations. Indeed, the *divide-and-conquer method* (DCGA) can provide a solution for MSA utilizing appropriate cut points. As far as clustering is concerned, the aim is to divide the objects into clusters so that the validity inside clusters is minimized. As an internal measure for cluster validity, the *sum of squared error* (SSE) is used. A *clustering genetic algorithm with the SSE criterion* (CGA_SSE), a hybrid approach, using the most popular algorithm, the *K-means* is presented. The CGA_SSE combines local and global search procedures. Comparison of the *K*-means and CGA_SSE is provided in terms of accuracy and quality of solution for clusters of different size and density. The complexity of all proposed algorithms is examined. The Hadoop for the distributed environment provides an alternate solution to the CGA_SSE, following the MapReduce paradigm. Simulation results are provided.

The second section (i.e., "Big Data Processing") is organized into the following chapters.

Chapter 5, "Approaches for High-Performance Big Data Processing: Applications and Challenges," by Ouidad Achahbar, Mohamed Riduan Abid, Mohamed Bakhouya, Chaker El Amrani, Jaafar Gaber, Mohammed Essaaidi, and Tarek A. El Ghazawi, puts emphasis on social media websites, such as *Facebook*, *Twitter*, and *YouTube*, and job posting websites like *LinkedIn* and *CareerBuilder*, which involve a huge amount of data that are very useful to economy assessment and society development. These sites provide sentiments and interests of people connected to web communities and a lot of other information. The big data collected from the web is considered as an unprecedented source to fuel data processing and business intelligence. However, collecting, storing, analyzing, and processing these big data as quickly as possible create new challenges for both scientists and analytics. For example, analyzing big data from social media is now widely accepted by many companies as a way of testing the acceptance of their products and services based on customers' opinions. *Opinion mining* or *sentiment analysis methods* have been recently proposed for extracting positive/negative words from big data. However, highly accurate and timely processing and analysis of the huge amount of data to extract their meaning requires new processing techniques. More precisely, a technology is needed to deal with the massive amounts of unstructured and semi-structured information, in order to understand hidden user behavior. Existing solutions are time consuming given the increase in data volume and complexity. It is possible to use high-performance computing technology to accelerate data processing, through MapReduce ported to cloud computing. This will allow companies to deliver more business value to their end customers in the dynamic and changing business environment. This chapter discusses approaches proposed in literature and their use in the cloud for big data analysis and processing.

Chapter 6, "The Art of Scheduling for Big Data Science," by Florin Pop and Valentin Cristea, moves the attention to applications that generate big data, like social networking and social influence programs, cloud applications, public websites, scientific experiments and simulations, data warehouses, monitoring platforms, and e-government services. Data grow rapidly, since applications produce continuously increasing volumes of both unstructured and structured data. The impact on data processing, transfer, and storage is the need to reevaluate the approaches and solutions to better answer user needs. In this context, scheduling models and algorithms have an important role. A large variety of solutions for specific applications and platforms exist, so a thorough and systematic analysis of existing solutions for *scheduling models*, *methods*, and *algorithms* used in big data processing and storage environments has high importance. This chapter presents the best of existing solutions and creates an overview of current and near-future trends. It highlights, from a research perspective, the performance and limitations of existing solutions and offers an overview of the current situation in the area of scheduling and resource management related to big data processing.

Chapter 7, "Time–Space Scheduling in the MapReduce Framework," by Zhuo Tang, Lingang Jiang, Ling Qi, Kenli Li, and Keqin Li, focuses on *the significance of big data*, that is, analyzing people's behavior, intentions, and preferences in the growing and popular social networks and, in addition to this, processing data with nontraditional structures and exploring their meanings. Big data is often used to describe a company's large amount of unstructured and semi-structured data. Using analysis to create these data in a relational database for downloading will require too much time and money. Big data analysis and cloud computing are often linked together because real-time analysis of large data requires a framework similar to MapReduce to assign work to hundreds or even thousands of computers. After several years of criticism, questioning, discussion, and speculation, big data finally ushered in the era belonging to it. Hadoop presents MapReduce as an analytics engine, and under the hood, it uses a distributed storage layer referred to as the Hadoop Distributed File System (HDFS). As an open-source implementation of MapReduce, Hadoop is, so far, one of the most successful realizations of large-scale data-intensive cloud computing platforms. It has been realized that when and where to start the reduce tasks are the key problems to enhance MapReduce performance. In this so-delineated context, the chapter proposes a framework for supporting *time–space scheduling* in MapReduce. For *time scheduling*, a *self-adaptive reduce task scheduling policy* for reduce tasks' start times in the Hadoop platform is proposed. It can decide the start time point of each reduce task dynamically according to each job context, including the task completion time and the size of the map output. For *space scheduling*, suitable algorithms are released, which synthesize the network locations and sizes of reducers' partitions in their scheduling decisions in order to mitigate network traffic and improve MapReduce performance, thus achieving several ways to avoid scheduling delay, scheduling skew, poor system utilization, and low degree of parallelism.

Chapter 8, "The Graph Engine for Multithread Systems Graph Database System for Commodity Clusters," by Alessandro Morari, Vito Giovanni Caltellana, Oreste Villa, Jesse Weaver, Greg Williams, David Haglin, Antonino Tumeo, and John Feo, considers the specific case of organizing, managing, and analyzing massive amounts of data in several contexts, like social network analysis, financial risk management, threat detection in complex network systems, and medical and biomedical databases. For these areas, there is a problem not only in terms of size but also in terms of performance, because the processing should happen sufficiently fast to be useful. *Graph databases* appear a good candidate to manage these data: They provide an efficient data structure for heterogeneous data or data that are not themselves rigidly structured. However, exploring large-scale graphs on modern high-performance machines is challenging. These systems include processors and networks optimized for regular, floating-point intensive computations and large, batched data transfers. At the opposite, exploring graphs generates fine-grained, unpredictable memory and network accesses, is mostly memory bound, and is synchronization intensive. Furthermore, graphs often are difficult to partition, making their processing prone to load unbalance. Following this evidence, the chapter describes (*Graph Engine for Multithreaded Systems* [GEMS]), a full software stack that implements a graph database on a commodity cluster and enables scaling in data set size while maintaining a constant query throughput when adding more cluster nodes. The GEMS software stack comprises: a SPARQL-to-data parallel C++ compiler; a library of distributed data structures; and a custom, multithreaded, runtime system. Also, an evaluation of GEMS on a typical SPARQL benchmark and on a Resource Description Format (RDF) data set is proposed.

Chapter 9, "KSC-net: Community Detection for Big Data Networks," by Raghvendra Mall and Johan A.K. Suykens, demonstrates the applicability of the *kernel spectral clustering* (KSC) method for community detection in big data networks, also providing a practical exposition of the KSC method on large-scale synthetic and real-world networks with up to 106 nodes and 107 edges. The KSC method uses a primal–dual framework to construct a model on a smaller subset of the big data network. The original large-scale kernel matrix cannot fit in memory. So smaller subgraphs using a *fast and unique representative*

Should "106" and "107" be "10^6" and "10^7"?

subset (FURS) selection technique is selected. These subsets are used for training and validation, respectively, to build the model and obtain the model parameters. It results in a powerful out-of-sample extensions property, which allows inferring of the community affiliation for unseen nodes. The KSC model requires a *kernel function*, which can have kernel parameters and what is needed to identify the number of clusters *k* in the network. A memory-efficient and computationally efficient model selection technique named *balanced angular fitting* (BAF) based on angular similarity in the eigenspace was proposed in the literature. Another parameter-free KSC model was proposed as well. Here, the model selection technique exploits the structure of projections in eigenspace to automatically identify the number of clusters and suggests that a normalized linear kernel is sufficient for networks with millions of nodes. This model selection technique uses the concept of entropy and balanced clusters for identifying the number of clusters *k*. In the scope of this context literature, the chapter describes the software *KSC-net*, which obtains the representative subset by FURS, builds the KSC model, performs one of the two (BAF and parameter-free) model selection techniques, and uses out-of-sample extensions for community affiliation for the big data network.

Chapter 10, "Making Big Data Transparent to the Software Developers' Community," by Yu Wu, Jessica Kropczynski, and John M. Carroll, investigates the *open-source software* (OSS) development community, which has allowed technology to progress at a rapid pace around the globe through shared knowledge, expertise, and collaborations. The broad-reaching open-source movement bases itself on a share-alike principle that allows anybody to use or modify software, and upon completion of a project, its source code is made publically available. Programmers who are a part of this community contribute by voluntarily writing and exchanging code through a collaborative development process in order to produce high-quality software. This method has led to the creation of popular software products including *Mozilla Firefox*, *Linux*, and *Android*. Most OSS development activities are carried out online through formalized platforms (such as *GitHub*), incidentally creating a vast amount of interaction data across an ecosystem of platforms that can be used not only to characterize open-source development work activity more broadly but also to create big data awareness resources for OSS developers. The intention of these awareness resources is to enhance the ability to seek out much-needed information necessary to produce high-quality software in this unique environment that is not conducive to ownership or profits. Currently, it is problematic that interconnected resources are archived across stand-alone websites. Along these research lines, this chapter describes the process through which these resources can be *more conspicuous* through big data, in *three interrelated sections* about the context and issues of the collaborating process in online space and a fourth section on how *big data can be obtained and utilized*.

The third section (i.e., "Big Data Stream Techniques and Algorithms") is organized into the following chapters.

Chapter 11, "Key Technologies for Big Data Stream Computing," by Dawei Sun, Guangyan Zhang, Weimin Zheng, and Keqin Li, focuses on the two main mechanisms for *big data computing*, that is, *big data stream computing* (*BDSC*) and *big data batch computing*. BDSC is a model of straight-through computing, such as *Twitter Storm* and *Yahoo! S4*, which do for stream computing what Hadoop does for batch computing, while big data batch computing is a model of storing and then computing, such as the MapReduce framework, open-sourced by the Hadoop implementation. Essentially, big data batch computing is not sufficient for many real-time application scenarios, where a data stream changes frequently over time and the latest data are the most important and most valuable. For example, when analyzing data from real-time transactions (e.g., financial trades, e-mail messages, user search requests, sensor data tracking), a data stream grows monotonically over time as more transactions take place. Ideally, a real-time application environment can be supported by BDSC. In this specific applicative setting, this chapter introduces data stream graphs and the system architecture for BDSC and key technologies for BDSC systems. Among other contributions, the authors present the system architecture and key technologies of four popular example BDSC systems, that is, *Twitter Storm*, *Yahoo! S4*, *Microsoft TimeStream*, and *Microsoft Naiad*.

Chapter 12, "Streaming Algorithms for Big Data Processing on Multicore Architecture," by Marat Zhanikeev, studies Hadoop and MapReduce in their vest of *de facto standards* in big data processing today. Although they are two separate technologies, they form a single package as far as *big data processing*—not just storage—is concerned. This chapter treats them as one package, according to the depicted vision. Today, Hadoop and/or MapReduce lack popular alternatives. Hadoop solves the practical problem of not being able to store big data on a single machine by distributing the storage over *multiple nodes*. MapReduce is a framework on which one can run *jobs that process the contents of the storage*—also in a distributed manner—and generate statistical summaries. The chapter shows that performance improvements mostly target MapReduce. There are several fundamental problems with MapReduce. First, the *map* and *reduce* operators are restricted to key–value hashes (data type, not hash function), which places a cap on usability. For example, while the *data streaming* is a good alternative for big data processing, MapReduce fails to accommodate the necessary data types. Secondly, MapReduce jobs create heterogeneous environments where jobs compete for the same resource with no guarantee of fairness. Finally, MapReduce jobs lack *time awareness*, while some algorithms might need to process data in their time sequence or using a time window. The core premise of this chapter is to replace MapReduce with a *time-aware storage and processing logic*. Big data is replayed along the timeline, and all the jobs get the time-ordered sequence of data items. The major difference here is that the new method collects all the jobs in one place—the node that replays data—while MapReduce sends jobs to remote nodes so that data can be processed locally. This architecture is chosen for the sole purpose of accommodating a wide range of data streaming algorithms and the data types they create.

Is "vest" really what is meant here?

Chapter 13, "Organic Streams: A Unified Framework for Personal Big Data Integration and Organization Towards Social Sharing and Sustainable Individualized Use," by Xiaokang Zhou and Qun Jin, moves the attention to the rapid development of *emerging computing paradigms*, which are conveyed in our continuously experiencing a change in work, life, playing, and learning in the highly developed information society, a kind of seamless integration of the real physical world and cyber digital space. More and more people have been accustomed to sharing their personal contents across the social networks due to the high accessibility of social media along with the increasingly widespread adoption of wireless mobile computing devices. User-generated information has spread more widely and quickly and provided people with opportunities to obtain more knowledge and information than ever before, which leads to an explosive increase of data scale, containing big potential value for an individual, business, domestic, and national economy development. Thus, it has become an increasingly important issue to sustainably manage and utilize *personal big data*, in order to mine useful insight and real value to better support information seeking and knowledge discovery. To deal with this situation in the big data era, a *unified approach to aggregation and integration of personal big data from life logs* in accordance with individual needs is considered essential and effective, which can benefit the sustainable information sharing and utilization process in the social networking environment. Based on this main consideration, this chapter introduces and defines a new concept of *organic stream*, which is designed as a flexibly extensible data carrier, to provide a simple but efficient means to formulate, organize, and represent personal big data. As an abstract data type, organic streams can be regarded as a *logic metaphor*, which aims to meaningfully process the raw stream data into an associatively and methodically organized form, but no concrete implementation for physical data structure and storage is defined. Under the conceptual model of organic streams, a heuristic method is proposed and applied to extract diversified individual needs from the tremendous amount of social stream data through social media. And an integrated mechanism is developed to aggregate and integrate the relevant data together based on individual needs in a meaningful way, in which personal data can be physically stored and distributed in private personal clouds and logically represented and federated by a set of newly introduced metaphors named heuristic stone, associative drop, and associative ripple. The architecture of the system with the foundational modules is described, and the prototype implementation with the experiment's result is presented to demonstrate the usability and effectiveness of the framework and system.

Is "federated" really what is meant here?

Chapter 14, "Managing Big Trajectory Data: Online Processing of Positional Streams," by Kostas Patroumpas and Timos Sellis, considers *location-based services*, which have become all the more important in social networking, mobile applications, advertising, traffic monitoring, and many other domains, following the proliferation of smartphones and GPS-enabled devices. Managing the locations and trajectories of numerous people, vehicles, vessels, commodities, and so forth must be efficient and robust, since this information must be processed online and should provide answers to users' requests in real time. In this *geo-streaming* context, such long-running continuous queries must be repeatedly evaluated against the most recent positions relayed by moving objects, for instance, reporting which people are now moving in a specific area or finding friends closest to the current location of a mobile user. In essence, modern processing engines must cope with huge amounts of streaming, transient, uncertain, and heterogeneous spatiotemporal data, which can be characterized as *big trajectory data*. Inspired by this methodological trend, this chapter examines big data processing techniques over *frequently updated locations* and *trajectories of moving objects*. Indeed, the big data issues regarding *volume*, *velocity*, *variety*, and *veracity* also arise in this case. Thus, authors foster a *close synergy* between the established stream processing paradigm and spatiotemporal properties inherent in motion features. Taking advantage of the spatial locality and temporal timeliness that characterize each trajectory, the authors present methods and heuristics that address such problems.

The fourth section (i.e., “Big Data Privacy”) is organized into the following chapters.

Chapter 15, “Personal Data Protection Aspects of Big Data,” by Paolo Balboni, focuses on *specific personal aspects of managing and processing big data*, by also providing a relevant vision on European privacy and data protection laws. In particular, the analysis considers applicable EU data protection provisions and their impact on both businesses and consumers/data subjects, and it introduces and conceptually assesses a methodology to determine whether (1) data protection law applies and (2) personal big data can be (further) processed (e.g., by way of analytic software programs). Looking into more detail, this chapter deals with diverse aspects of data protection, providing an understanding of big data from the perspective of personal data protection using the *Organization for Economic Co-operation and Development*’s four-step life cycle of personal data along the value chain, paying special attention to the concept of compatible use. Also, the author sheds light on the development of the concept of *personal data* and its relevance in terms of data processing. Further focus is placed on aspects such as *pseudo-anonymization*, *anonymous data*, and *reidentification*. Finally, conclusions and recommendations that focus on the privacy and data implications of big data processing and the importance of data protection compliance management are illustrated.

Chapter 16, “Privacy-Preserving Big Data Management: The Case of OLAP,” by Alfredo Cuzzocrea, highlights the *security and privacy of big data repositories* as among the most challenging topics in *big data research*. As a relevant instance, the author considers the case of *cloud systems*, which are very popular now. Here, cloud nodes are likely to exchange data very often. Therefore, the *privacy breach risk* arises, as distributed data repositories can be accessed from a node to another one, and hence, *sensitive information* can be inferred. Another relevant data management context for big data research is represented by the issue of effectively and efficiently supporting *data warehousing and OLAP over big data*, as multidimensional data analysis paradigms are likely to become an “enabling technology” for *analytics over big data*, a collection of models, algorithms, and techniques oriented to extract useful knowledge from cloud-based big data repositories for decision-making and analysis purposes. At the convergence of the three axioms introduced (i.e., security and privacy of big data, data warehousing and OLAP over big data, analytics over big data), a critical research challenge is represented by the issue of *effectively and efficiently computing privacy-preserving OLAP data cubes over big data*. It is easy to foresee that this problem will become more and more important in future years, as it not only involves relevant theoretical and methodological aspects, not all explored by actual literature, but also regards significant modern scientific applications. Inspired by these clear and evident trends, this chapter moves the attention to privacy-preserving OLAP data cubes over big data and provides two kinds of contributions: (1) a complete survey of privacy-preserving OLAP approaches available in literature, with respect to both *centralized* and *distributed environments*, and (2) an innovative framework that relies on *flexible sampling-based data cube compression techniques for computing privacy-preserving OLAP aggregations on data cubes*.

If “OLAP” is an abbreviation, please define at first mention.

The fifth section (i.e., "Big Data Applications") is organized into the following chapters.

Chapter 17, "Big Data in Finance," by Taruna Seth and Vipin Chaudhary, addresses big data in the context of the *financial industry*, which has always been driven by data. Today big data is prevalent at various levels of this field, ranging from the financial services sector to capital markets. The availability of big data in this domain has opened up new avenues for innovation and has offered immense opportunities for growth and sustainability. At the same time, it has presented several new challenges that must be overcome to gain the maximum value out of it. Indeed, in recent years, the financial industry has seen an upsurge of interest in big data. This comes as no surprise to finance experts, who understand the potential value of data in this field and are aware that no industry can benefit more from big data than the financial services industry. After all, the industry not only is driven by data but also thrives on data. Today, the data, characterized by the four *V*s, which refer to volume, variety, velocity, and veracity, are prevalent at various levels of this field, ranging from capital markets to the financial services industry. Also, capital markets have gone through an unprecedented change, resulting in the generation of massive amounts of high-velocity and heterogeneous data. For instance, about 70% of the US equity trades today are generated by high-frequency trades (HFTs) and are machine driven. In the so-delineated context, this chapter considers the impact and applications of big data in the financial domain. It examines some of the key advancements and transformations driven by big data in this field. The chapter also highlights important big data challenges that remain to be addressed in the financial domain.

Chapter 18, "Semantic-Based Heterogeneous Multimedia Big Data Retrieval," by Kehua Guo and Jianhua Ma, considers *multimedia retrieval*, as an important technology in many applications such as web-scale multimedia search engines, mobile multimedia search, remote video surveillance, automation creation, and e-government. With the widespread use of multimedia documents, our world will be swamped with multimedia content such as massive images, videos, audios, and other contents. Therefore, traditional multimedia retrieval has been switching into a *big data environment*, and the research into solving some problems according to the features of *multimedia big data retrieval* attracts considerable attention. Having as reference the so-delineated application setting, this chapter proposes a heterogeneous multimedia big data retrieval framework that can achieve good retrieval accuracy and performance. The authors begin by addressing the particularity of heterogeneous multimedia retrieval in a big data environment and introducing the background of the topic. Then, literature related to current multimedia retrieval approaches is briefly reviewed, and the general concept of the proposed framework is introduced briefly. The authors provide in detail a description of this framework, including semantic information extraction, representation, storage, and multimedia big data retrieval. Finally, the proposed framework's performance is experimentally evaluated against several multimedia data sets.

Chapter 19, "Topic Modeling for Large-Scale Multimedia Analysis and Retrieval," by Juan Hu, Yi Fang, Nam Ling, and Li Song, similarly puts emphasis on the *exponential growth of multimedia data* that occurred in recent years, with the arrival of the big data era and thanks to the rapid increase in processor speed, cheaper data storage, prevalence of digital content capture devices, as well as the flooding of social media like *Facebook* and *YouTube*. New data generated each day have reached 2.5 quintillion bytes as of 2012. Particularly, more than 10 h of videos are uploaded onto *YouTube* every minute, and millions of photos are available online every week. The explosion of multimedia data in social media raises a great demand for developing effective and efficient *computational tools* to facilitate producing, analyzing, and retrieving large-scale multimedia content. *Probabilistic topic models* prove to be an effective way to organize large volumes of text documents, while much fewer related models are proposed for other types of unstructured data such as multimedia content, partly due to the high computational cost. With the emergence of cloud computing, *topic models* are expected to become increasingly applicable to multimedia data. Furthermore, the growing demand for a deep understanding of multimedia data on the web drives the development of sophisticated machine learning methods. Thus, it is greatly desirable to develop topic modeling approaches to multimedia applications that are consistently effective, highly efficient, and easily scalable. Following this methodological scheme, this chapter presents a review of topic models for large-scale multimedia analysis and shows the current challenges from various perspectives by presenting a comprehensive overview of related work that addresses these challenges. Finally, the chapter discusses several research directions in the field.

Chapter 20, “Big Data Biometrics Processing: A Case Study of an Iris Matching Algorithm on Intel Xeon Phi,” by Xueyan Li and Chen Liu, investigates the applicative setting of *big data biometrics repositories*. Indeed, the authors recognize that, with the drive towards achieving higher computation capability, the most advanced computing systems have been adopting alternatives from the traditional *general purpose processors* (GPPs) as their main components to better prepare for big data processing. *NVIDIA’s graphic processing units* (GPUs) have powered many of the top-ranked supercomputer systems since 2008. In the latest list published by *Top500.org*, two systems with *Intel Xeon Phi* coprocessors have claimed position 1 and 7. While it is clear that the need to improve efficiency for big data processing will continuously drive changes in hardware, it is important to understand that these new systems have their own advantages as well as limitations. The required effort from the researchers to port their codes onto the new platforms is also of great significance. Unlike other coprocessors and accelerators, the *Intel Xeon Phi* coprocessor does not require learning a new programming language or new parallelization techniques. It presents an opportunity for the researchers to share parallel programming with the GPP. This platform follows the standard parallel programming model, which is familiar to developers who already work with *x86*-based parallel systems. From another perspective, with the rapidly expanded biometric data collected by various sources for identification and verification purposes, how to manage and process such big data draws great concern. On one hand, biometric applications normally involve comparing a huge amount of samples and templates, which has strict requirements on the computational capability of the underlying hardware platform. On the other hand, the number of cores and associated threads that hardware can support has increased greatly; an example is the newly released *Intel Xeon Phi* coprocessor. Hence, big data biometrics processing demands the execution of the applications at a higher parallelism level. Taking an *iris matching algorithm* as a case study, the authors propose an *OpenMP* version of the algorithm to examine its performance on the *Intel Xeon Phi* coprocessor. Their target is to evaluate their parallelization approach and the influence from the optimal number of threads, the impact of thread-to-core affinity, and the built-in vector engine. This does not mean that achieving good performance on this platform is simple. The hardware, while presenting many similarities with other existing multicore systems, has its own characteristics and unique features. In order to port the code in an efficient way, those aspects are fully discussed in the chapter.

Chapter 21, "Storing, Managing, and Analyzing Big Satellite Data: Experiences and Lessons Learned from a Real-World Application," by Ziliang Zong, realizes how big data has shown great capability in yielding extremely useful information and extraordinary potential in revolutionizing scientific discoveries and traditional commercial models. Indeed, numerous corporations have started to utilize big data to understand their customers' behavior at a fine-grained level, rethink their business process work flow, and increase their productivity and competitiveness. Scientists are using big data to make new discoveries that were not possible before. As the volume, velocity, variety, and veracity of big data keep increasing, significant challenges with respect to innovative big data management, efficient big data analytics, and low-cost big data storage solutions arise. This chapter provides a case study on how the *big satellite data* (at the petabyte level) of the world's largest satellite imagery distribution system is captured, stored, and managed by the *National Aeronautics and Space Administration* (NASA) and the *US Geological Survey* (USGS) and gives a unique example of how a changed policy could significantly affect the traditional ways of big data storage and distribution, which will be quite different from typical commercial cases driven by sales. Also, the chapter discusses how the USGS *Earth Resources Observation and Science* (EROS) center swiftly overcomes the challenges from serving few government users to hundreds of thousands of global users, and how *data visualization* and *data mining* techniques are used to analyze the characteristics of millions of requests and how they can be used to improve the performance, cost, and energy efficiency of the EROS system. Finally, the chapter summarizes the experiences and lessons learned from conducting the target big data project in the past 4 years.

Chapter 22, "Barriers to the Adoption of Big-Data Applications in the Social Sector," by Elena Strange, focuses on *social aspects of dealing with big data*. The author recognizes that effectively working with and leveraging big data has the potential to change the world. Indeed, if there is a ceiling on realizing the benefits of *big data algorithms, applications, and techniques*, we have not yet reached it. The research field is maturing rapidly. No longer are we seeking to understand what "big data" is and whether it is useful. No longer is big data processing the province of niche computer science research. Rather, the concept of big data has been widely accepted as important and inexorable, and the buzzwords "big data" have found their way beyond computer science into the essential tools of business, government, and media. Tools and algorithms to leverage big data have been increasingly democratized over the last 10 years. By 2010, over 100 organizations reported using the distributed file system and framework *Hadoop*. Early adopters leveraged *Hadoop* on in-house *Beowulf clusters* to process tremendous amounts of data. Today, well over 1000 organizations use *Hadoop*. That number is climbing and now includes companies with a range of technical competencies and those with and without access to internal clusters and other tools. Yet, the benefits of big data have not been fully realized by businesses, governments, and particularly the social sector. In this so-delineated background, this chapter describes the impact of this gap on the social sector and the broader implications engendered by the sector in a broader context. Also, the chapter highlights the opportunity gap: the unrealized potential of big data in the social sector, and explores the historical limitations and context that have led up to the current state of big data. Finally, it describes the current perceptions of and reactions to big data algorithms and applications in the social sector and offers some recommendations to accelerate the adoption of big data.

Overall, this book represents a solid research contribution to state-of-the-art studies and practical achievements in algorithms, analytics, and applications on big data, and sets the basis for further efforts in this challenging scientific field that will, more and more, play a leading role in next-generation database, data warehousing, data mining, and cloud computing research. The editors are confident that this book will represent an authoritative milestone in this very challenging scientific road.

Editors

Kuan-Ching Li is currently a professor in the Department of Computer Science and Information Engineering at Providence University, Taiwan. He was department chair in 2009 and special assistant to the university president since 2010 and was appointed as vice dean for the Office of International and Cross-Strait Affairs (OIA) in 2014. He earned his PhD in 2001 from the University of Sao Paulo (USP), Brazil. Dr. Li is a recipient of awards from NVIDIA, the Ministry of Education (MOE)/Taiwan, and the Ministry of Science and Technology (MOST)/Taiwan. He also received guest professorship from universities in China, including Xiamen University (XMU), Huazhong University of Science and Technology (HUST), Lanzhou University (LZU), Shanghai University (SHU), Anhui University of Science and Technology (AUST), and Lanzhou Jiaotong University (LZJTU). He has been involved actively in conferences and workshops as a program/general/steering conference chairman and in numerous conferences and workshops as a program committee member, and he has organized numerous conferences related to high-performance computing and computational science and engineering.

Dr. Li is the editor in chief of the technical publications *International Journal of Computational Science and Engineering (IJCSE)*, *International Journal of Embedded Systems (IJES)*, and *International Journal of High Performance Computing and Networking (IJHPCN)*, all published by Inderscience, also serving on a number of journals' editorial boards and guest editorships. In addition, he has been acting as editor/coeditor of several technical professional books, published by CRC Press and IGI Global. His topics of interest include networked computing, GPU computing, parallel software design, and performance evaluation and benchmarking. Dr. Li is a member of the Taiwan Association of Cloud Computing (TACC), a senior member of the IEEE, and a fellow of the IET.

If "GPU," "IEEE," and "IET" are abbreviations, please define at first mention.

Hai Jiang is an associate professor in the Department of Computer Science at Arkansas State University, United States. He earned his BS degree from Beijing University of Posts and Telecommunications, China, and MA and PhD degrees from Wayne State University, Detroit, Michigan, United States. His current research interests include parallel and distributed systems, computer and network security, high-performance computing and communication, big data, and modeling and simulation. He has published one book and research papers in major international journals and conference proceedings. He has served as a US National Science Foundation proposal review panelists and a US Department of Energy (DoE) Smart Grid Investment Grant (SGIG) reviewer multiple times. He serves as an editor for the *International Journal of High Performance Computing and Networking* (*IJHPCN*); a regional editor for the *International Journal of Computational Science and Engineering* (*IJCSE*) as well as the *International Journal of Embedded Systems* (*IJES*); an editorial board member for the *International Journal of Big Data Intelligence* (*IJBDI*), the *Scientific World Journal* (*TSWJ*), the *Open Journal of Internet of Things* (*OJIOT*), and the *GSTF Journal on Social Computing* (*JSC*); a guest editor for the *IEEE Systems Journal*, *International Journal of Ad Hoc and Ubiquitous Computing*, *Cluster Computing*, and *The Scientific World Journal* for multiple special issues. He has also served as a general chair or program chair for some major conferences/workshops (CSE, HPCC, ISPA, GPC, ScalCom, ESCAPE, GPU-Cloud, FutureTech, GPUTA, FC, SGC). He has been involved in 90 conferences and workshops as a session chair or as a program committee member, including major conferences such as AINA, ICPP, IUCC, ICPADS, TrustCom, HPCC, GPC, EUC, ICIS, SNPD, TSP, PDSEC, SECRUPT, and ScalCom. He has reviewed six cloud computing–related books (*Distributed and Cloud Computing*, *Virtual Machines*, *Cloud Computing: Theory and Practice*, *Virtualized Infrastructure and Cloud Services Management*, *Cloud Computing: Technologies and Applications Programming*, *The Basics of Cloud Computing*) for major publishers such as Morgan Kaufmann, Elsevier, and Wiley. He serves as a review board member for a large number of international journals (*TC*, *TPDS*, *TNSM*, *TASE*, *JPDC*, *Supercomputing*, *CCPE*, *FGCS*, *CJ*, and *IJPP*). He is a professional member of ACM and the IEEE Computer Society. Locally, he serves as US NSF XSEDE (Extreme Science and Engineering Discovery Environment) Campus Champion for Arkansas State University.

Please define abbreviations at first mention.

Dr. Laurence T. Yang is a professor in the Department of Computer Science of St. Francis Xavier University, Canada. His current research includes parallel and distributed computing, embedded and ubiquitous/pervasive computing, cyber–physical–social systems, and big data.

He has published around 200+ refereed international journal papers in the above areas; around one-third are on IEEE/ACM transactions/journals, and the rest mostly are on Elsevier, Springer, and Wiley journals). He has been involved actively in conferences and workshops as a program/general/steering conference chair and as a program committee member. He served as the vice chair of the IEEE Technical Committee of Supercomputing Applications (2001–2004), the chair of the IEEE Technical Committee of Scalable Computing (2008–2011), and the chair of the IEEE Task Force on Ubiquitous Computing and Intelligence (2009–present). He was in the steering committee of the IEEE/ACM Supercomputing Conference series (2008–2011) and in the National Resource Allocation Committee (NRAC) of Compute Canada (2009–2013).

If "IEEE/ACM" is an abbreviation, please define at first mention.

In addition, he is the editor in chief of several international journals. He is serving as an editor for many international journals. He has been acting as an author/coauthor or an editor/coeditor of more than 25 books from well-known publishers. The book *Mobile Intelligence* from Wiley (2010) received an Honorable Mention by the American Publishers Awards for Professional and Scholarly Excellence (the PROSE Awards). He has won several best paper awards (including IEEE best and outstanding conference awards, such as at the IEEE 20th International Conference on Advanced Information Networking and Applications [IEEE AINA-06]); one best paper nomination; Distinguished Achievement Award (2005, 2011); and the Canada Foundation for Innovation Award (2003). He has been invited to give around 30 keynote talks at various international conferences and symposia.

Alfredo Cuzzocrea is currently a senior researcher at the Institute of High Performance Computing and Networking of the Italian National Research Council, Italy, and an adjunct professor at the University of Calabria, Italy. He is habilitated as an associate professor in computer science engineering by the Italian National Scientific Habilitation of the Italian Ministry of Education, University and Research (MIUR). He also obtained the habilitation as an associate professor in computer science by the Aalborg University, Denmark, and the habilitation as an associate professor in computer science by the University of Rome Tre, Italy. He is an adjunct professor at the University of Catanzaro "Magna Graecia," Italy, the University of Messina, Italy, and the University of Naples "Federico II," Italy. Previously, he was an adjunct professor at the University of Naples "Parthenope," Italy. He holds 35 visiting professor positions worldwide (Europe, United States, Asia, and Australia). He serves as a Springer Fellow Editor and as an Elsevier Ambassador. He holds several roles in international scientific societies, steering committees for international conferences, and international panels, some of them having directional responsibility. He served as a panel leader and moderator in international conferences. He was an invited speaker in several international conferences worldwide (Europe, United States, and Asia). He is a member of scientific boards of several PhD programs worldwide (Europe and Australia). He serves as an editor for the Springer series *Communications in Computer and Information Science*. He covers a large number of roles in international journals, such as editor-in-chief, associate editor, and special issue editor (including *JCSS, IS, KAIS, FGCS, DKE, INS,* and *Big Data Research*). He edited more than 30 international books and conference proceedings. He is a member of the editorial advisory boards of several international books. He covers a large number of roles in international conferences, such as general chair, program chair, workshop chair, local chair, liaison chair, and publicity chair (including ODBASE, DaWaK, DOLAP, ICA3PP, ICEIS, APWeb, SSTDM, IDEAS, and IDEAL). He served as the session chair in a large number of international conferences (including EDBT, CIKM, DaWaK, DOLAP, and ADBIS). He serves as a review board member in a large number of international journals (including *TODS, TKDE, TKDD, TSC, TIST, TSMC, THMS, JCSS, IS, KAIS, FGCS, DKE,* and *INS*). He also serves as a review board member in a large number of international books and as a program committee member in a very large number of international conferences (including VLDB, ICDE, EDBT, CIKM, IJCAI, KDD, ICDM, PKDD, and SDM). His current research interests include multidimensional data modeling and querying, data stream modeling and querying, data warehousing and OLAP, OLAM, XML data management, web information systems modeling and engineering, knowledge representation and management models and techniques, Grid and P2P computing, privacy and security of very large databases and OLAP data cubes, models and algorithms for managing uncertain and imprecise information and knowledge, models and algorithms for managing complex data on the web, and models and algorithms for high-performance distributed computing and architectures. He is the author or a coauthor of more than 330 papers in international conferences (including EDBT, CIKM, SSDBM, MDM, DaWaK, and DOLAP), international journals (including *JCSS, IS, KAIS, DKE,* and *INS*), and international books (mostly edited by Springer). He is also involved in several national and international research projects, where he also covers responsibility roles.

Contributors

Mohamed Riduan Abid
Al Akhawayn University
Ifrane, Morocco

Ouidad Achahbar
Al Akhawayn University
Ifrane, Morocco

Mohamed Bakhouya
International University of Rabat
Sala el Jadida, Morocco

Please provide locations for Paolo Balboni's affiliations.

Paolo Balboni
ICT Legal Consulting
and
European Privacy Association

Vito Giovanni Caltellana
Pacific Northwest National Laboratory
Richland, Washington

Please provide city of affiliation of John M. Carroll.

John M. Carroll
College of Information Sciences and Technology
Penn State University
Pennsylvania

Vipin Chaudhary
Department of Computer Science and Engineering
University at Buffalo (SUNY)
Buffalo, New York

Valentin Cristea
Computer Science Department
Faculty of Automatic Control and Computers
University Politehnica of Bucharest
Bucharest, Romania

Alfredo Cuzzocrea
ICAR-CNR
and
University of Calabria
Italy

If "ICAR-CNR" is an abbreviation, please expand. Also, please provide the complete location (city, state/country) for the affiliations of Alfredo Cuzzocrea.

Chaker El Amrani
Université Abdelmalek Essaadi-Tanger
Morocco

Please provide the city for the affiliation of Chaker El Amrani.

Tarek A. El Ghazawi
George Washington University
Washington, District of Columbia

Mohammed Essaaidi
Ecole Nationale Supérieure d'Informatique et d'Analyse des Systemes
Agdal Rabat, Morocco

Yi Fang
Department of Computer Engineering
Santa Clara University
Santa Clara, California

John Feo
Pacific Northwest National Laboratory
Richland, Washington

Jaafar Gaber
Universite de Technologie de
Belfort-Montneliard
Belfort, France

Please provide city for the affiliation of Ian Gorton.

Ian Gorton
Software Engineering Institute
Carnegie Mellon University
Pennsylvania

Kehua Guo
School of Information Science and
Engineering
Central South University
Changsha, China

David Haglin
Pacific Northwest National Laboratory
Richland, Washington

Juan Hu
Department of Computer Engineering
Santa Clara University
Santa Clara, California

Lingang Jiang
College of Information Science and
Engineering
Hunan University
Changsha, China

Please provide city of affiliation for Qun Jin.

Qun Jin
Waseda University
Japan

Please provide city of affiliation of Jessica Kropczynski.

Jessica Kropczynski
College of Information Sciences and
Technology
Penn State University
Pennsylvania, United States

Kenli Li
College of Information Science and
Engineering
Hunan University
Changsha, China

Keqin Li
College of Information Science and
Engineering
Hunan University
Changsha, China

Note: Author Keqin Li wrote more than one chapter, and different affiliations were stated in different chapters. Both were captured here.

and

Department of Computer Science and
Technology
Tsinghua University
Beijing, China

Xueyan Li
Department of Electrical and Computer
Engineering
Clarkson University
Potsdam, New York

Nam Ling
Department of Computer Engineering
Santa Clara University
Santa Clara, California

Chen Liu
Department of Electrical and Computer
Engineering
Clarkson University
Potsdam, New York

Yan Liu
Faculty of Computer Science and
Engineering
Concordia University
Montreal, Canada

Jianhua Ma
Faculty of Computer and Information Sciences
Hosei University
Tokyo, Japan

Please check if the affiliation for Raghvendra Mall is correct.

Raghvendra Mall
KU Leuven–ESAT/STADIUS
Leuven, Belgium

Please ensure that the affiliation information for Stéphane Marchand-Maillet and Hisham Mohamed is complete (Computer Science Department of what institution?).

Stéphane Marchand-Maillet
Viper Group, Computer Science Department
Geneva, Switzerland

Hisham Mohamed
Viper Group, Computer Science Department
Geneva, Switzerland

Alessandro Morari
Pacific Northwest National Laboratory
Richland, Washington

Kostas Patroumpas
National Technical University of Athens
Athens, Greece

Florin Pop
Computer Science Department
Faculty of Automatic Control and Computers
University Politehnica of Bucharest
Bucharest, Romania

Ling Qi
College of Information Science and Engineering
Hunan University
Changsha, China

Please provide city of affiliation for Timos Sellis.

Timos Sellis
RMIT University
Australia

Taruna Seth
Department of Computer Science and Engineering
University at Buffalo (SUNY)
Buffalo, New York

Li Song
Institute of Image Communication and Information Processing
Shanghai Jiao Tong University
Minhang, Shanghai, China

Elena Strange
University of Memphis
Memphis, Tennessee

Dawei Sun
Department of Computer Science and Technology
Tsinghua University
Beijing, China

Please check if the affiliation information for Johan A.K. Suykens is correct.

Johan A.K. Suykens
KU Leuven–ESAT/STADIUS
Leuven, Belgium

Zhuo Tang
College of Information Science and Engineering
Hunan University
Changsha, China

Alexander Thomasian
Thomasian & Associates
Pleasantville, New York

John Tsiligaridis
Math and Computer Science Department
Heritage University
Toppenish, Washington

Antonino Tumeo
Pacific Northwest National Laboratory
Richland, Washington

Oreste Villa
NVIDIA Research
Santa Clara, California

Jesse Weaver
Pacific Northwest National Laboratory
Richland, Washington

Greg Williams
Rensselaer Polytechnic Institute
Troy, New York

Xing Wu
Faculty of Computer Science and Engineering
Concordia University
Montreal, Canada

Please provide city of affiliation of Yu Wu.

Yu Wu
College of Information Sciences and Technology
Penn State University
Pennsylvania

Guangyan Zhang
Department of Computer Science and Technology
Tsinghua University
Beijing, China

Marat Zhanikeev
Department of Artificial Intelligence
Computer Science and Systems Engineering
Kyushu Institute of Technology
Iizuka, Fukuoka Prefecture, Japan

Weimin Zheng
Department of Computer Science and Technology
Tsinghua University
Beijing, China

Xiaokang Zhou
Waseda University
Japan

Please provide city of affiliation for Xiaokang Zhou.

Ziliang Zong
Texas State University
Texas

Please provide city of affiliation for Ziliang Zong.