Preface

Rapid advances in digital sensors, networks, storage, and computation, along with their availability at low cost, are leading to the creation of huge collections of data. Initially, the drive for generation and storage of data came from scientists: telescopes and instruments such as the Large Hadron Collider (LHC) generate a huge amount of data that needed to be processed to enable scientific discovery. LHC, for example, was reported as generating as much as 1 TB of data every second. Later, with the popularity of the SMAC (Social, Mobile, Analytics, and Cloud) paradigm, enormous amount of data started to be generated, processed, and stored by enterprises. For instance, Facebook in 2012 reported that they process over 200 TB of data per hour. In fact, SINTEF (The Foundation for Scientific and Industrial Research) from Norway reports that 90% of the world's data generated has been generated in the last 2 years. These were the key motivators towards the Big Data paradigm.

Unlike traditional data warehouses, that rely in highly structured data, this new paradigm unleashes the potential of analyzing any source of data, being it structured and stored in relational databases, semi-structured emerging from sensors, machines, and applications, or unstructured obtained from social media and other human sources.

This data has the potential to enable new insights that can change the way business, science, and governments deliver services to their consumers and can impact society as a whole. Nevertheless, for it to happen, new algorithms, methods, infrastructures and platforms are required that can make sense of all this data and provide the insights while they are still of interest for analysts of diverse domains.

This has led to the emergence of the Big Data Computing paradigm focusing on sensing, collection, storage, management, and analysis of data from variety of sources to enable new value and insights. This paradigm enhanced considerable the capacity of organizations to understand their activities and improve aspects of its business in ways never imagined before; however, at the same time, it raises new concerns of security and privacy whose implications are still not completely understood by society.

To realize the full potential of Big Data, researchers and practitioners need to address several challenges and develop suitable conceptual and technological solutions for tackling them. These include life-cycle management of data, large-scale storage, flexible processing infrastructure, data modeling, scalable machine learning and data analysis algorithms, techniques for sampling and making trade-off between data processing time and accuracy, and dealing with privacy and ethical issues involved in data sensing, storage, processing, and actions.

This book addresses the above issues by presenting a broad view of each of the issues, identifying challenges faced by researchers and opportunities for practitioners embracing the Big Data paradigm.

Organization of the book

This book contains 18 chapters authored by several leading experts in the field of Big Data. The book is presented in a coordinated and integrated manner starting with big data analytics methods, going through the infrastructures and platforms supporting them, aspects of security and privacy, and finally applications.

The content of the book is organised into four parts:

I. Big Data Science
II. Big Data Infrastructures and Platforms
III. Big Data Security and Privacy
IV. Big Data Applications

Part I: Big Data Science

Data Science is a discipline that emerged in the last years, likewise the Big Data concept. Although there are different interpretations of what is Data Science, we adopt the view that Data Science is a discipline that merges concepts from Computer Science (algorithms, programming, machine learning, and data mining), mathematics (statistics and optimization), and domain knowledge (business, applications, and visualization) to extract insights from data and transform it into actions that have an impact in the particular domain of application. Data Science is already challenging when amount of data enables traditional analysis, and thus becomes particularly challenging when traditional methods lose their effectiveness due to large volume and velocity in the data.

Part I presents fundamental concepts and algorithms in the Data Science domain that address the issues rose by Big Data. As a motivation for this part, and in the same direction than what we discussed so far, Chapter 1 discusses how what is now known as Big Data is the result of efforts of two distinct areas, namely Machine Learning and Cloud Computing.

The velocity aspect of Big Data demand analytic algorithms that can operate data in motion, i.e., algorithms that do not assume that all the data is available all the time for decision making, and decisions need to be made “on the go”, probably with summaries of past data. In this direction, Chapter 2 discusses real-time processing systems for Big Data, including stream processing platforms that enable analysis of data in motion and a case study in finance.

The volume aspect of data demands that existing algorithms for different analytics data are adapted to take advantage of distributed systems where memory is not shared, and thus different machines have only part of data to operate. Chapter 3 discusses how it affects natural language processing, text mining, and anomaly detection in the context of social media.

A concept that emerged recently benefiting from Big Data is deep learning. The approach, derived from Artificial Neural Networks (ANN), constructs layered structures that hold different abstractions of the same data and has application in language processing, and image analysis, among others. Chapter 4 discusses algorithms that can leverage modern GPUs to speed up computation of Deep Learning models.

Another concept popularized in the last years is Graph Processing, a programming model where an abstraction of a graph (network) of nodes and vertices represents the computation to be carried out. Likewise the previous chapter, Chapter 5 discusses GPU-based algorithms for graph processing.

Part II: Big Data Infrastructures and Platforms

Although part of the Big Data revolution is enabled by new algorithms and methods to handle large amounts of heterogeneous data in movement and at rest, all of this would be of no value if computing platforms and infrastructures did not evolve to better support Big Data. New platforms providing different abstractions for programmers arose that enable problems to be represented in different ways. Thus, instead of adapting the problem to fit a programming model, developers are now able to select the abstraction that is closer to the problem at hand, enabling faster more correct software solutions to be developed. The same revolution observed in the computing part of the analytics is also observed in the storage part: new methods to persist data that are more flexible than traditional relational databases were developed and adopted in the last years.

Part II of this book is dedicated to such infrastructure and platforms supporting Big Data. Starting with databases support, Chapter 6 discusses the different models of NOSQL database models and systems that are available for storage of large amounts of structured, semi-structured and structured data, including key-value column-based, graph-based, and document-based stores.

As the infrastructures of choice for running Big Data analytics are shared (think of clusters and clouds), new methods were necessary to rationalize the use of resources so all applications get their fair share of resources and can progress to a result in a reasonable amount of time. In this direction, Chapter 7 discusses the general problem of resource management techniques for Big Data frameworks and a new efficient technique for resource management implemented in Apache YARN. Chapter 8 presents a novel technique for increasing resource usage and performance of Big Data platforms by applying a “resource shaping” technique, whereas Chapter 9 contains a survey on various techniques for optimization of many aspects of the Hadoop framework, including the job scheduler, HDFS, and Hbase.

Whereas the previous three chapters focused on distributed platforms for Big Data analytics, parallel platforms, which rely on many computing cores sharing memory, are also viable platforms for Big Data analytics. In this direction, Chapter 10 discusses an alternative solution that is optimized to take advantage of the large amount of memory and large number of cores available in current servers.

Part III: Big Data Security and Privacy

For economic reasons, physical infrastructures supporting Big Data are shared. This helps in rationalizing the huge costs involved in building such large-scale cloud infrastructures. Thus, whether the infrastructure is a public cloud or a private cloud, multi-tenancy is a certainty that raises security and privacy concerns. Moreover, the sources of data can reveal many things about its source: although many times sources will be applications, and the data generated is in public domain, tit is also possible that data generated by devices and actions of humans (for example, via posts in social networks) can be analyzed in a way that individuals can be identified and/or localized, an issue that also raises privacy issues. Part III of this book is dedicated to such security and privacy issues of Big Data.

Chapter 11 addresses the issue of spatial privacy of users of social networks and the threats to it enabled by Big Data analytics. Chapter 12 addresses the issue of the use of shared resources for Big Data computing, and ways to protect queries and prevent loss of privacy on correlated data.

Chapter 13 is dedicated to methods to perform consumer analytics when shopping. It introduces methods to infer the location of mobile devices and to estimate human behavior in shopping activities.

Part IV: Big Data Applications

All the advances in methods and platforms would be of no value if the capabilities offered them did not generate value (whatever definition of value we take into consideration). Thankfully, this is not the case, and a range of applications in the most diverse areas were developed that fulfill the goal of delivering value via Big Data analytics. These days, finance institutions, governments, education institutions, and researchers, to name a few, are applying Big Data analytics on a daily basis as part of their business as usual tasks. Part IV of this book is dedicated to such applications, bringing interesting use cases of the application of Big Data analytics.

Social media arose in the last 10 years, initially as a means to connect people. Now it has emerged as a platform for businesses purposes, advertisements, delivery of news of public interest, and for people to express their opinions and emotions. Chapter 14 introduces an application in this context, namely a Big Data framework for mining opinion from social media in Thailand. In the same direction, Chapter 15 presents an interesting case study of application of Big Data Analytics to mine social media to e valuate the effect of the weather in people’s emotions.

The entertainment industry can also benefit from Big Data, as demonstrated in Chapter 16 with an application of Big Data analytics for optimization of delivery of video on demand via the Internet.

Big Data analytics is also disrupting core traditional sectors. As an example, Chapter 17 presents a case study on application of Big Data Analytics in the energy sector: the chapter shows how data generated by smart distribution lines (smart grids) can be analyzed to enable identification of faults in the transmission line.

E-Science is one of the first applications driving Big Data paradigm in which scientific discovery enabled by large scale computing infrastructures. As clusters and grids became popular among research institutions, it became clear that new discoveries could be made if these infrastructures were put at work to crunch massive volumes of data collected from many scientific instruments. Acknowledging the importance of e-Science as a motivator for a substantial amount of innovation in the field leading to the establishment of Big Data, Chapter 18 concludes with various e-Science applications and key elements of their deployment in a cloud environment.

Acknowledgments

We thank all the contributing authors for their time, effort, and dedication during the preparation of the book.

Raj would like to thank his family members, especially his wife, Smrithi and daughters, Soumya and Radha Buyya, for their love, understanding, and support during the preparation of the book. Rodrigo would like to thank his wife, Kimie, his son Roger, and his daughter Laura. Amir would like to thank his wife Elly and daughter Diana.

Finally, we would like to thank the staff at Morgan Kauffman, particularly, Amy Invernizzi, Brian Romer, and Todd Green for managing the publication in record time.

Rajkumar Buyya
The University of Melbourne and Manjrasoft Pty Ltd, Australia

Rodrigo N. Calheiros
The University of Melbourne, Australia

Amir Vahid Dastjerdi
The University of Melbourne, Australia