Preface

Many applications generate Big Data, like social networking and social influence programs, Cloud applications, public web sites, scientific experiments and simulations, data warehouse, monitoring platforms, and e-government services. Data grow rapidly since applications produce continuously increasing volumes of both unstructured and structured data. Large-scale interconnected systems aim to aggregate and efficiently exploit the power of widely distributed resources. In this context, major solutions for scalability, mobility, reliability, fault tolerance, and security are required to achieve high performance. The impact on data processing, transfer and storage is the need to re-evaluate the approaches and solutions to better answer the user needs.

Extracting valuable information from raw data is especially difficult considering the velocity of growing data from year to year and the fact that 80 % of data is unstructured. In addition, data sources are heterogeneous (various sensors, users with different profiles, etc.) and are located in different situations or contexts. This is why the Smart City infrastructure runs reliably and permanently to provide the context as a public utility to different services. Context-aware applications exploit the context to adapt accordingly the timing, quality and functionality of their services. The value of these applications and their supporting infrastructure lies in the fact that end users always operate in a context: their role, intentions, locations, and working environment constantly change.

Since the introduction of the Internet, we have witnessed an explosive growth in the volume, velocity, and variety of the data created on a daily basis. This data is originated from numerous sources including mobile devices, sensors, individual archives, the Internet of Things, government data holdings, software logs, public profiles on social networks, commercial datasets, etc. The so-called Big Data problem requires the continuous increase of the processing speeds of the servers and of the whole network infrastructure. In this context, new models for resource management are required. This poses a critically difficult challenge and striking development opportunities to Data-Intensive (DI) and High-Performance Computing (HPC): how to efficiently turn massively large data into valuable

information and meaningful knowledge. Computationally-effective DI and HPC are required in a rapidly increasing number of data-intensive domains.

Successful contributions may range from advanced technologies, applications, and innovative solutions to global optimization problems in scalable large-scale computing systems to development of methods, conceptual and theoretical models related to Big Data applications and massive data storage and processing. Therefore, it is imperative to gather the consent of researchers to muster their efforts in proposing unifying solutions that are practical and applicable in the domain of high-performance computing systems.

The Big Data era poses a critically difficult challenge and striking development opportunities to High-Performance Computing (HPC). The major problem is an efficient transformation of the massive data of various types into valuable information and meaningful knowledge. Computationally effective HPC is required in a rapidly increasing number of data-intensive domains. With its special features of self-service and pay-as-you-use, Cloud computing offers suitable abstractions to manage the complexity of the analysis of large data in various scientific and engineering domains. This book surveys briefly the most recent developments on Cloud computing support for solving the Big Data problems. It presents a comprehensive critical analysis of the existing solutions and shows further possible directions of the research in this domain including new generation multi-datacenter cloud architectures for the storage and management of the huge Big Data streams.

The large volume of data coming from a variety of sources and in various formats, with different storage, transformation, delivery or archiving requirements, complicates the task of context data management. At the same time, fast responses are needed for real-time applications. Despite the potential improvements of the Smart City infrastructure, the number of concurrent applications that need quick data access will remain very high. With the emergence of the recent cloud infrastructures, achieving highly scalable data management in such contexts is a critical challenge, as the overall application performance is highly dependent on the properties of the data management service. The book provides, in this sense, a platform for the dissemination of advanced topics of theory, research efforts and analysis and implementation for Big Data platforms and applications being oriented on Methods, Techniques and Performance Evaluation. The book constitutes a flagship driver toward presenting and supporting advanced research in the area of Big Data platforms and applications.

This book herewith presents novel concepts in the analysis, implementation, and evaluation of the next generation of intelligent techniques for the formulation and solution of complex processing problems in Big Data platforms. Its 23 chapters are structured into four main parts:

1. *Architecture of Big Data Platforms and Applications*: Chapters 1–7 introduce the general concepts of modeling of Big Data oriented architectures, and discusses several important aspects in the design process of Big Data platforms and applications: workflow scheduling and execution, energy efficiency, load balancing methods, and optimization techniques.

2. *Big Data Analysis*: An important aspect of Big Data analysis is how to extract valuable information from large-scale datasets and how to use these data in applications. Chapters 8–12 discuss analysis concepts and techniques for scientific application, information fusion and decision making, scalable and reliable analytics, fault tolerance and security.
3. *Biological and Medical Big Data Applications*: Collectively known as computational resources or simply infrastructure, computing elements, storage, and services represent a crucial component in the formulation of intelligent decisions in large systems. Consequently, Chaps. 13–16 showcase techniques and concepts for big biological data management, DNA sequence analysis, mammographic report classification and life science problems.
4. *Social Media Applications*: Chapters 17–23 address several processing models and use cases for social media applications. This last part of the book presents parallelization techniques for Big Data applications, scalability of multimedia content delivery, large-scale social network graph analysis, predictions for Twitter, crowd-sensing applications and IoT ecosystem, and smart cities.

These subjects represent the main objectives of ICT COST Action IC1406 High-Performance Modelling and Simulation for Big Data Applications (cHiPSet) and the research results presented in these chapters were performed by joint collaboration of members from this action.

Acknowledgments

We strongly believe that this book ought to serve as a reference for students, researchers, and industry practitioners interested or currently working in Big Data domain.

Bucharest, Romania	Florin Pop
Cracow, Poland	Joanna Kołodziej
Naples, Italy	Beniamino Di Martino
July 2016	