Preface

The purpose of this volume entitled: *Techniques and Environments for Big Data Analysis: Parallel, Cloud, and Grid Computing* is to magnetize and sensitize a wide range of readers and researchers in the area of Big data by presenting the recent advances in the fields of Big data analysis, and also the techniques and tools used to analyze it. Further, it can enlighten them on how the expensive fitness evaluation of evolutionary learning can play a vital role in Big data analysis by adopting parallel, grid, and cloud computing environments.

Rapid growth of computational resources produces huge amounts of data to be used in the field of engineering and technology known as Big data. So it is essential to know computational theories and tools which assist humans in extracting knowledge from Big data. The amount of data collected across different areas exceeds human ability to reduce and analyze without the help of the automated machines. There is much knowledge accumulated in the voluminous data. On the other hand, these computational resources can also be used to better understand the data, by performing large-scale evaluations, parameter sweeps, etc. We refer to the overall use of massive on-demand computation (cloud or GPUs) for machine learning as Big Learning. Evolutionary machine learning techniques are perfect candidates for big learning tasks as they have flexibility in knowledge representations, learning paradigms, and their innate parallelism.

To achieve the objectives of Big data, this volume includes ten chapters contributed by promising authors. In Chap. 1, Mishra and Pattanaik have presented an introduction and architecture of Big data. They also present a brief description about Big-table, MapReduce, and Hadoop. Mishra and Sagnika discuss different parallel environments and their architectures in Chap. 2. They have presented a descriptive vision on how to work in different parallel environments.

Dev and Patgiri present an overview on Hadoop distributed file system (HDFS) in Chap. 3. They also evaluate the performance of the Hadoop in different environments by considering several factors. How files less than the block size affects Hadoop's R/W performance and how the time of execution of a job depends on

block size and number of reducers are illustrated. They also enumerate some of the challenges of Hadoop.

Mustafi in Chap. 4 draws focus on how Big data challenges can be handled from the data science perspective. The data available for analysis are in different forms in terms of volume, velocity, variety, and veracity. The objective is to resolve some of these real-world problems using natural language processing, where the unstructured data can be transformed into meaningful structured information; and machine learning to get more insights out of the information available or derived. This chapter fairly covers important methodologies where, what, and when to apply. Some open research problems are also shared for the budding data scientists.

In Chap. 5, Panigrahi, Tiwari, Pati, and Das present the development of cyber foraging systems by introducing the concept of cloudlets and the role of cloudlets in cyber foraging systems as well as discuss the working and limitations of cloudlets. This chapter also explores the new architectures where the cloudlets can be helpful in providing Big data solutions in areas with less Internet connectivity and where the user device disruption is high. The chapter then deals with different applications of cloudlets for Big data and focuses on the details of the existing work done with cyber foraging systems to manage different characteristics of Big data.

As the scope of computation is extending across domains where large and complex databases are needed to be dealt with, it has become a very useful approach to subdivide the tasks and to perform them in parallel, which leads to a significant reduction in the processing time. On the other hand, evolutionary algorithms are rapidly gaining popularity to solve intractable problems. However, they are also suffering with an intrinsic problem of expensive fitness evaluation; hence parallelization of evolutionary algorithms proves to be beneficial in solving intensive tasks within a feasible execution time. Therefore to address the aforesaid issues, in Chap. 6 Mishra, Sagnika, and Dehuri present different parallel genetic algorithm models and uses of different Big data mechanisms like MapReduce over parallel GA models.

In Chap. 7, Ghosh and Desarkar present the limitations of general search optimization algorithm for Big data. They also discuss how evolutionary algorithms like GA can be suitable tool for Big data analysis. In Chap. 8, Meena and Ibrahim present how, by using MapReduce programming model, feature selection can be done, when documents are represented as a bag of syntactic phrases. They also explain how ACO can be parallelized using a MapReduce programming model.

In Chap. 9, Mishra and Patel present the key challenges, issues, and applications of grid technologies in the management of Big data.

Finally, we hope that the readers enjoy reading this book, and most importantly, that they learn all new computing paradigms enumerated in this book for analyzing Big data.

Bhabani Shankar Prasad Mishra
Satchidananda Dehuri
Euiwhan Kim
Gi-Nam Wang

http://www.springer.com/978-3-319-27518-5

Techniques and Environments for Big Data Analysis
Parallel, Cloud, and Grid Computing
Prasad Mishra, B.S.; Dehuri, S.; Kim, E.; Wang, G.-N.
(Eds.)
2016, XI, 191 p. 103 illus., 76 illus. in color., Hardcover
ISBN: 978-3-319-27518-5