Challenges in Big Data Analysis

5

M. Govindarajan
Annamalai University, India

INTRODUCTION

Recent years big data has been accumulated in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents, and internet search indexing. Social computing includes social network analysis, online communities, recommender systems, reputation systems, and prediction markets where as internet search indexing includes ISI, IEEE Xplorer, Scopus, Thomson Reuters etc. Considering this advantages of big data it provides a new opportunities in the knowledge processing tasks for the upcoming researchers. However opportunities always follow some challenges.

To handle the challenges, various computational complexities, information security, and computational method are needed to be known to analyze big data. For example, many statistical methods that perform well for small data size do not scale to voluminous data. Similarly, many computational techniques that perform well for small data face significant challenges in analyzing big data. Big Data problems such as Heterogeneity, Scalability, Storage and Processing Issues, Noise Accumulation, Spurious Correlation, Incidental Endogeneity, Data Access and Sharing of Information, Quality of Data, Fault Tolerance, Speed/Velocity, Accuracy, Privacy and Security, Garbage Mining, Timeliness, Human Collaboration are to be addressed to design effective statistical procedures for exploring and predicting big data. These issues are discussed briefly in the following subsections.

BACKGROUND

David Lazer et al., (2009) discusses an emerging field that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors. Stadler et al., (2010) developed an efficient EM algorithm for numerical optimization with provable convergence properties. High dimensionality also gives rise to incidental endogeneity, a phenomenon that many unrelated covariates may incidentally be correlated with the residual noises. The endogeneity creates statistical biases and causes model selection inconsistency that lead to wrong scientific discoveries (Liao and Jiang, 2011; Fan and Liao, 2012). Jianqing Fan et al., (2013) gives overviews on the salient features of Big Data and how these features impact on paradigm change on statistical and computational methods as well as computing architectures. Nawsher Khan et al., (2014) comprehensively surveys and classifies the various attributes of Big Data, including its nature, definitions, rapid growth rate, volume, management, analysis, and security. This study also proposes a data life cycle that uses the technologies and terminologies of Big Data. Future research directions in this field are determined based on opportunities and several open issues in Big Data domination. These research directions facilitate the exploration of the domain and the development of optimal techniques to address Big Data. Lenka Venkata Satyanarayana

DOI: 10.4018/978-1-7998-3479-3.ch041

(2015) provides an in-depth analysis of different platforms available for performing big data analytics. This paper surveys different hardware platforms available for big data analytics and assesses the advantages and drawbacks of Big Data. D. P. Acharjya et al., (2016) explore the potential impact of big data challenges, open research issues, and various tools associated with it. Akhil et al., (2017) analyzed the potential effect of big data challenges, open research issues, and different tools related with it. Ripon Patgiri (2018) presents a study report on numerous research issues and challenges of Big Data which is employed in very large dataset. Reihaneh H. Hariri et al., (2019) reviews previous work in big data analytics and presents a discussion of open challenges and future directions for recognizing and mitigating uncertainty in this domain.

FOCUS OF THE ARTICLE

The purpose of this chapter is to highlight the Big Data challenges and also provide a brief description of each challenge.

Heterogeneity

Big data are often created via aggregating many data sources corresponding to different sub-populations. Each sub-population might exhibit some unique features not shared by others. In classical settings where the sample size is small or moderate, data points from small sub-populations are generally categorized as "outliers" and it is hard to systematically model them due to insufficient observations. However, in the Big data era, the large sample size enables us to better understand heterogeneity, shedding lights toward studies such as exploring the association between certain covariates (e.g., genes or SNPs) and rare outcomes (e.g., rare diseases or diseases in small populations) and understanding why certain treatments (e.g. chemotherapy) benefit a subpopulation and harm another subpopulation. In short, the main advantage brought by Big Data is to understand heterogeneity of sub-populations such as the benefits of certain personalized treatments, which are infeasible when sample size is small or moderate.

Scalability

The processor technology has changed in recent years. The clock speeds have largely stalled and processors are being built with more number of cores instead. Previously data processing systems had to worry about parallelism across nodes in a cluster but now the concern has shifted to parallelism within a single node. In past the techniques which were used to do parallel data processing across data nodes aren't capable of handling intra-node parallelism. This is because of the fact that many more hardware resources such as cache and processor memory channels are shared across a core in a single node.

The scalability issue of Big data has lead towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters. This requires a high level of sharing of resources which is expensive and also brings with it various challenges like how to run and execute various jobs so that we can meet the goal of each workload cost effectively. It also requires dealing with the system failures in an efficient manner which occurs more frequently if operating on large clusters. These factors combined put the concern on how to express the programs, even complex machine learning tasks. There has been a huge shift in the technologies being used. Hard Disk Drives (HDD) are being replaced by the solid state Drives and Phase Change technology which are not having