Chapter 7
Knowledge Discovery and Big Data Analytics: Issues, Challenges, and Opportunities

Vinoth Kumar Jambulingam
VIT University, India

V. Santhi
VIT University, India

ABSTRACT

The era of big data has come with the ability to process massive datasets from heterogeneous sources in real-time. But the conventional analytics can't be able to manage such a large amount of varied data. The main issue that is being asked is how to design a high-performance computing platform to effectively carry out analytics on big data and how to develop a right mining scheme to get useful insights from voluminous big data. Hence this chapter elaborates these challenges with a brief introduction on traditional data analytics followed by mining algorithms that are suitable for emerging big data analytics. Subsequently, other issues and future scope are also presented to enhance capabilities of big data.

DOI: 10.4018/978-1-5225-2483-0.ch007

INTRODUCTION

As the era of data communication and expertise reaches across several fields quickly, most of the information has its origin in digital communication in addition internet nowadays. Lyman, P. and Varian, H. (2002) showed in a study that the new knowledge present in digital devices have crossed already over ninety percent all through the 21st millennium, whereas the scale of that new knowledge was additionally over hundreds of petabytes. In fact, the issues of analyzing the massive information did not rise abruptly, however, are there for many years as it has been that the data creation is felt easier than finding hidden knowledge or useful patterns from that information. Albeit personal computers nowadays are loT more quickly than those in the early 1960's, the massive size of information is a pitfall to perform research on the data we've got nowadays. As an answer to the issues of analyzing high volume data, Xu, R. & Wunsch, D (2009) proposed some effective techniques like sampling, density-dependent methods, data condensation, grid-dependent methods, divide and conquer, progressive learning, and distributed computing, are being offered. Obviously, these ways are perpetually accustomed to enhance the efficiency of the mechanisms of data analysis method (Lyman, P. et al., 2002).

The outcomes of those techniques show that with the effective techniques at our disposal, we tend to be able to perform better and larger data analysis in an exceedingly affordable time. Ding, C. & He, X (2004) presented a dimension based technique say PCA could be a classical example that's geared toward minimizing the input file size to speed up the method of knowledge discovery. Kollios, G., Gunopulos, D., Koudas, N., & Berchtold, S. (2003) presented another reduction scheme that minimizes the computations on accumulated data is sampling, which might even be accustomed to accelerate the computation time involved in knowledge discovery process. Even though the improvements in personal computers and web technologies have gone through the phenomenal rise of computing hardware obeying Moore's law since 1970's, the bottlenecks of handling the high-volume information are there though we are getting into the time of big data analytics. Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S (2012) identified that large-scale data refers to the inability of the present information systems to manage and process load them in simpler machines. Also, present data mining algorithms and centralization of analytics won't work in the context of big data directly. Laney D (2001) given a popular definition in addition to the problems of the size of data also known as 3V's to clarify about big data namely volume, variety, and velocity. The terminology of 3Vs shows that the information size is massive, the information is made quickly, and also the information is existed in multiple varieties and taken from heterogeneous sources, correspondingly. Further studies Laney, D. (2001)