Chapter 3
Big Data Mining:
A Forecast to the Future

V. Sucharita
K. L. University, India

A. Satya Kalyan
K. L. University, India

P. Venkateswara Rao
ASCET, India

P. Rajarajeswari
K. L. University, India

ABSTRACT
At present in Big Data era mining of Big Data can help us find learning which nobody has possessed the capacity to find some time recently. There is a developing interest for tools and techniques which can prepare and investigate Big Data effectively and proficiently. In this chapter, the accessible information mining tools and techniques which can deal with Big Data have been abridged. This paper additionally concentrates on tools and techniques for mining of data and information streams. Through better analysis of the vast volumes of information that are getting to be accessible, there is the potential for making speedier progress in numerous scientific areas what’s more, making strides the productivity what’s more, victory of numerous organizations. The challenges incorporate not just the self-evident issues of scale, be that as it may too heterogeneity, need of structure, error handling, protection, opportunities at all stages of the analysis from acquisition of data to obtaining to result.

INTRODUCTION
Big Data is a collection of data that represents huge and complex data which is very hard to manage using traditional DBMS and tools of Processing data (Diebold, 2000). Before Big Data emerged, the data was in data warehouses that consisted of structured databases. Slowly the sources of data are diversified and are heterogeneous. The big data analysts analyze the future of the customer data. Though companies are collecting, storing and doing data analysis most of them are struggling with the data of the project which is very big. This chapter explains how the big data tools and techniques gain new insight by doing analysis of various data sets that are very large. The production of the sources of data are associated with various V’s like volume, velocity and variety has more contribution to the growth of big data in the present trend. At present the data is in various unstructured formats like text, sensor reading and videos.

DOI: 10.4018/978-1-5225-2947-7.ch003
and data sets are in petabytes which is very difficult to manage. All the data which is created, duplicated and used per year will reach to Exabyte’s. At present, less data is being analyzed with potential to increase to higher range if the data is properly analyzed. There are wide variety of techniques that have been developed and modified to visualize, analyze, operate and aggregate big data to make this kind of data well structured. The various techniques include Time series analysis, A/B testing, Cluster analysis, Network analysis, Ensemble learning, Data fusion Association rule learning, Machine learning, etc. The A/B testing technique is used to compare various options to determine what treatments will progress a given objective; Cluster analysis is used for classifying objects that splits a diverse group into smaller groups of similar objects. Ensemble learning uses multiple predictive models to obtain better predictive performance. Network analysis analyzes connections between nodes in a network and their strength. Lastly, Machine learning helps in learning and training automatically. There is no meaning of Big Data if unable to analyze the information that is valuable. There are mainly four things in analytics they are collection of data, cleaning of data, modeling and reporting of data. Collecting the data is more relevant is very important. Data cleaning is required when data is collected from various sources. After data cleaning is completed it has to be modeled by using various statistical and various machine learning techniques.

TOOLS ON BIG DATA

In the present trend, huge quantities of data are available for decision makers. Big Data will take a reference of data sets that are very large but maintains velocity and variety. So, it is not possible to manage with ordinary tools and techniques. Therefore, solution must be provided to handle and mine knowledge using such large data so that decision makers could take decision rightly with continuous data changing very rapidly like customer interaction and social networking. This can be done using Big Data Analytics. This Chapter analyzes methods to analyze and also that can be applied to Big Data. Also, various opportunities can be analyzed by analytics in different domains for taking the decisions. Big Data is simply not buzzword it is a reality of business value. Tools are mainly used for identifying trends, mine patterns and getting findings that are very much valuable from the huge quantities available to them. Many relational databases have the capacity to store and manage huge amounts of data may be in petabytes data locally. These relational databases have limited capacity. Semi-structured data is the data stored in a structural form. The data is stored and retrieved using NoSQL. It has not particular scheme, therefore they can handle the data coming from various sources. The distributed storage that is designed for the large databases is BigTable (Chang, 2006). The distributed file system that is designed for providing the access to data using huge clusters of commodity hardware. The programming model for parallel processing is Map-Reduce(Dean, 2004). In the Map reduce the input data is split and distributes across different machines so that parallel processing can be done. Also, it collects those results from different machines as a final result. This can be done in Hadoop. The Hadoop and Bigdata terms are used synonymously. The library of the apache hadoop is like a framework that uses simple programming for distributed computing environment. It is mainly used to come out from failures in the application layer for getting better accuracy. Hadoop uses the distributed file system for the storage of data. To make the efficient processing of Big Data so many Hadoop based tools are available like Apache Sqoop and Apache Flume. Apache Hive is a framework built on top of the Hadoop for making the users to store data in tables. All the tools of Bigdata deals only with volume. So, there is a need for the tools which can also handle data streams. So, Apache storm is used for distributed real time calculations for large streams. All the tools discussed