Chapter 2

LITERATURE REVIEW

- **Aditya B. Patelet al (2012)**, reports the experimental work on the Big data problems. It describes the optimal solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and Map Reduce programming framework for parallel processing to process large data sets.

- **Mukherjee, A. et al (2012)** stated that Big data analytics classify the analysis of large amount of data to get the useful information and reveal the hidden patterns. Big data analytics refers to the Map-Reduce Framework which is developed by the Google. Apache Hadoop is the open source platform which is used for the purpose of realization of Google’s Map-Reduce Model. In this the performance of SF-CFS is compared with the HDFS using the SWIM by the face book job traces. SWIM contains the workloads of thousands of jobs with complex data arrival and computation patterns.

- **Garlasu D et al (2013)** Authors stated that Grid Computing offered the advantage about the storage capabilities and the processing power and the Hadoop technology is used for the implementation purpose. Grid Computing provides the concept of distributed computing. The benefit of Grid computing center is the high storage capability and the high processing power. Grid Computing makes the big contributions among the scientific research, help the scientists to analyze and store the large and complex data.

- **Sagioglu Set al (2013)** describe the big data content, its scope, methods, samples, advantages and challenges of Data. The serious issue about the Big data is the privacy and security. Big data samples explain the review about the atmosphere, biological science and research. Life sciences etc. By this paper, we can wrap up that any organization in any industry having big data can catch the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the Big
data easy to get the information from the complex data sets. The overall Evaluation
describe that the data is increasing and becoming multifaceted. The challenge is not
only to collect and run the data also how to dig out the useful information from that
collected data. According to the Intel IT Center, there are many challenges related to
Big Data which are data growth, data infrastructure, data range, data visualization,
data velocity.

- **Ling Liu et al (2013)** explained that the data scientists, we live in interesting times.
  Data has been the No. 1 fast growing incident on the Internet for the last decade. Big
data analytics have the prospective to expose deep insights concealed by big data that
exceeds the dispensation capacity of existing systems, such as gaze influence among
customers, revealed by analyzing shoppers' dealings, social and geographical data. In
the past 40 years, data was primarily used to trace and account business activities and
scientific events, and in the next 40 years data will be used also to get new insights, to
manipulate business decisions and to hasten scientific discovery. The key challenge is
to give the right platforms and tools to make calculation of big data easy and simple.
In this key note talk, I will discover reuse opportunities and challenges from multiple
dimensions towards delivering big data analytics as a service.

- **Keith C.C. Chan (2013)** explained that the Big Data refers to data sets that are so
  large and compound that conventional data processing tools and technologies cannot
cope with. The process of probing such data to uncover buried patterns in them is
referred to as Big Data Analytics. Drug discovery is associated to big data analytics as
the process may require the assortment, dispensation and analysis of extremely large
volume of structured and unstructured biomedical data stemming from a wide range of
experiments and surveys collected by hospitals, laboratories, pharmaceutical
companies or even social media. These data may comprise sequencing and gene
expression data, drug data including molecular data, protein and drug interaction data,
clinical trial and electronic patient record data, patient behavior and self-reporting data
in social media, regulatory monitoring data, and literatures where trends and drug
repurposing and protein-protein interface data may be found.
• **Roger Schell (2013)** explained that the Big data implies performing calculation and database operations for huge amounts of data, remotely from the data owner’s enterprise. Since a key value proposition of big data is access to data from multiple and varied domains, security and privacy will play a very important role in big data research and technology. Making efficient use of big data requires access from any domain to data in that domain, or any other domain it is certified to access.

• **Shunmei Meng; WanchunDou; Xuyun Zhang; Jinjun Chen; (Dec. 2013)“Big Data Applications”** explained that Service recommender systems have been shown as valuable tools for providing suitable recommendations to users. In the last decade, the amount of customers, services and online information has grown rapidly, yielding the big data analysis problem for service recommender systems. Consequently, traditional service recommender systems often suffer from scalability and inefficiency problems when processing or analyzing such large-scale data. Moreover, most of presented service recommender systems present the same ratings and rankings of services to different users without considering diverse users' preferences, and therefore fails to meet users' personalized requirements.

• **Xingdong Wu; Xingquan Zhu; Gong-Qing Wu; Wei (VOL.26, January 2014), “Data Mining with Big Data”** explained that Big Data concern large-volume, complex, growing data sets with multiple, independent sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly increasing in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolt, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and study, user interest modeling, security and privacy considerations. We study the challenging issues in the data-driven model and also in the Big Data revolution. While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and
knowledge associations. Such combined characteristics suggest that Big Data need a “big mind” to consolidate data for maximum values.

- **Shan Suthaharan** explained that the specific problem of Big Data classification of network imposition traffic. It discusses the system challenges offered by the Big Data problems associated with network intrusion forecast. The prediction of a possible invasion attack in a network requires constant collection of traffic data and learning of their characteristics on the fly. The continuous collection of traffic data by the network leads to Big Data problems that are caused by the volume, diversity and velocity properties of Big Data. The learning of the network characteristics requires machine learning techniques that capture global knowledge of the traffic patterns. The Big Data properties will lead to significant system challenges to implement machine learning frameworks.

- **Ahmed E. Yourself et al (2014)** explained that in this paper we initiate a framework for Healthcare Information Systems (HISs) based on big data analytics in mobile cloud computing environments. This framework provides a high level of mixing, interoperability, availability and sharing of healthcare data among healthcare providers, patients, and practitioners. Electronic Medical Records (EMRs) of patients dispersed among different Care Delivery Organizations (CDOs) are integrated and stored in the Cloud storage area; this creates an Electronic Health Records (EHRs) for each patient. Mobile Cloud allows fast Internet access and condition of EHRs from anywhere and at any time via different platforms.

- **Ashwin Belleet et al (2015)** explained that the rapidly expanding field of big data analytics has ongoing to play a pivotal role in the evolution of healthcare practices and research. It has provided tools to gather, direct, analyze and incorporate large volumes of dissimilar, structured, and unstructured data produced by current healthcare systems. Big data analytics has been recently applied towards aiding the process of care delivery and disease exploration.

- **Ho Ting Wong et al (2015)** explained that the big data is a hot topic in the academic sector, and healthcare researchers are definitely not an exemption. This article aims to provide a showcase in emergency medicine research to reveal the advantages of
conducting such research using big data. Big data is a novel and cost-effective research approach, and emergency medicine researchers could take advantage by using this approach and by doing so produces high-quality research at a faster pace. Big data, by contrast, provides flexibility in analyzing a data set from different perspectives. Moreover, because governmental organizations already maintain a significant amount of big data, research using big data can be conducted at a very low cost.