Taxonomy Outline of Big Data Analytics Literature

Sapna Sinha
Amity School of Engineering and Technology, India

Vishal Bhatnagar
Ambedkar Institute of Advance Communication Technologies & Research, India

Abhay Bansal
Amity School of Engineering and Technology, India

INTRODUCTION

Advancement of information technology made its use in every domain and in every nook and corners of the world. Mobile technology has a major contribution in this regards. The use of sensors, smart meters, credit/debit cards, internet and mobile are contributing in this data deluge. Due to availability of low cost storage, has made storage of data generated by these devices feasible. The terminology used by industry to refer this huge data is Big Data.

The term Big Data was first coined in 2005 by Roger Magolus from O'Really. Big Data is the term for data sizing zeta bytes or more, structured/semi structured/quasi structured/unstructured in nature, need processing in batches or real time. Organization across the globe, realized that Big data can be used to draw best decision based on facts which will give a competitive edge from their competitors.

All this made traditional tools and technologies obsolete. New tools and technologies are needed to be developed to handle the data deluge. According to (IDC, 2012), big data with its latest improved architecture and techniques are able to extract values at low cost by trapping this high velocity data and analysis. Big Data are the datasets that grow very large and very fast which are difficult to handle by using traditional tools and techniques. Predictive nature, volume and near real time result production have put Big Data into new domain requiring research in all facets (storage, network, operating system, analytical software etc.).

Big Data is an interesting area of research due to its novelty and hype. This paper presents review and classification of the literatures of Big Data research conducted between year 1997 to 2013. The classification scheme presented is as per our perception, indicating salient features of the domain providing ramps to both academics and industries involved. The growth in the publishing of the research paper and its current state will draw the interest of many.

This paper is organized as follows: Segment II presents the steps of Big Data Analytics, Segment III discusses research methodology adopted in the paper, Segment IV outlines the classification framework specifying dimensions and various approaches. Subsection V discusses the result of classification of articles. Subsequent subsection VI highlight research implications and discussions. Next Segment VII shows the limitation of the study and the last Subsection VIII concludes by presenting some direction for future research and the conclusion of the paper.

STEPS OF BIG DATA ANALYTICS

The research in the domain of Big Data Analytics is exhaustive and spread across various journals and the white paper published by corporate involve in research of Big Data Analytics. The roadmap

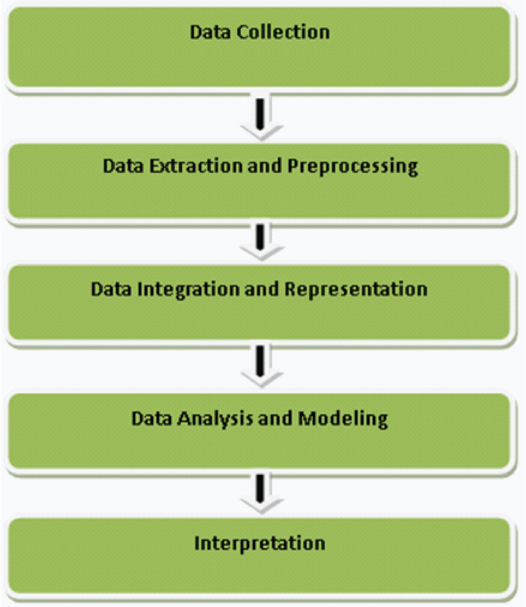
DOI: 10.4018/978-1-4666-5202-6.ch219

of Big Data Analytics is depicted in the form of steps as shown in Figure 1.

The factors to be considered in the each step of Big Data Analytics are discussed below:

1. **In data acquisition and recording**: Due to technology advances, different type of devices came into existence. In (Gantz, J., 2010), gave 14 classes of device & application that are having major contribution in generation of Big Data, few among them are in the field of auto intelligence system, automation systems like household application etc. and other applications like scientific applications. These devices and application are contributing in data deluge, information growth is exceeding the Moore's Law. Data generated by the devices have various forms like structured, semi-structured, quasi-structured and unstructured. The main challenge of this step is, what to store and what to discard and how to store the relevant information.

Figure 1. Roadmap of steps involved in big data analytics


Data Collection
Data Extraction and Preprocessing
Data Integration and Representation
Data Analysis and Modeling
Interpretation

2. **For Data Extraction and Preprocessing:** In this step, data collected from the previous step is cleaned and preprocessed to remove ambiguity, duplicate and records having null value are treated. Data is collected from different sources and in different form, it is the major challenge to integrate structured and unstructured data. Adding structure to unstructured data has become another major challenge now.
3. **For Data Integration and representation:** Preprocessed data are stored in the data stores residing on the distributed heterogeneous storage units. Hadoop and NoSQL are the two important components of Big Data processing. Hadoop provides storage capabilities through distributed, shared nothing file system and analysis capability and NoSQL has capability to capture read and update real time unstructured data. Linking data with the previously collected data is the major challenge of this step.
4. **For Data Analysis and Modeling:** Big Data Analytics tools and techniques are used, in-memory databases, real time reports and dashboards, text mining, advanced analytics tools, predictive analysis etc is done. Complex analytical queries are off loaded to analytical sandbox which enables integration with other relevant data and run complicated and difficult queries. MapReduce that runs on the Hadoop platform can be used to simplify creation of fault tolerant application. Hive is used for data summarization, impromptu queries and analysis of large data sets SQL like interface. High level procedural language Pig is used for the database access. Zoo Keeper, highly available system is used for coordinating distributed processes and application. But lack of scalable and underlying algorithm and complexity of data is the major challenge of this step.
5. **For Data Interpretation:** Advance Visualization technique is used for representing the result of the analysis in the appropriate