International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 4 293 - 303

Some Contribution of Statistical Techniques in Big Data: A Review

R. D. Patil[1]
Research Scholar, Department of Statistics,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad – 431004. (M.S) – INDIA
rupali.stat@gmail.com

Omprakash S. Jadhav[2]
Assistant Professor, Department of Statistics,
Dr. Babasaheb Ambedkar Marathwada University
Aurangabad-431004. (M.S)- INDIA
drjadhavop@gmail.com

Abstract- Big Data is a popular topic in research work. Everyone is talking about big data, and it is believed that science, business, industry, government, society etc. will undergo a through change with the impact of big data.Big data is used to refer to very huge data set having large, more complex, hidden pattern, structured and unstructured nature of data with the difficulties to collect, storage, analysing for process or result. So proper advanced techniques to use to gain knowledge about big data. In big data research big challenge is created in storage, process, search, sharing, transfer, analysis and visualizing. To deeply discuss on introduction of big data, issue, management and all used big data techniques. Also in this paper present a review of various advanced statistical techniques to handling the key application of big data have large data set. These advanced techniques handle the structure as well as unstructured big data in different area.

Keywords: *Big Data, Data Mining, Application, Advance Statistical Methods*

*****

I. INTRODUCTION

This paper draws up the basic concepts relating to big data. It attempts to combine the hitherto fragmented discourse on what constitutes big data, what metrics define the size and other characteristics of big data, and what tools and technologies exist to join the possible of big data. From corporate leaders to municipal planners and academics, big data are the subject of attention, and to some extent, fear. The sudden rise of big data has left many unprepared. In the past, new technological developments first appeared in technical and academic publications. The knowledge and combination later seeped into other opportunities of knowledge mobilization, including books. The fast evolution of big data technologies and the ready acceptance of the concept by public and private sectors left little time for the discourse to develop and mature in the academic domain. Authors and practitioners advanced to books and other electronic media for immediate and wide circulation of their work on big data. Thus, one finds several books on big data, including Big Data for Copies, but not enough fundamental discourse in academic publications.

One of the main challenges of the different field is the need to get the rightinformation as fast as possible to the right person in the right time and in the easiest way possible. There is a hugeamount of data produced everywhere. Just Google receives 2 million search queries every minute and Facebookusers post around 700 thousand pieces of content in the same amount of time. There are some analytical tools which help us to orientate in this huge amount of data, to make decisions easierand to speed up the processes. But traditional tools can practically work just with structuralized data and cannotwork with unstructuralized data. But there is also need to cope with huge amount of unstructuralized data, mainlywith the data in the form of text. That means that we have to use and work on more advanced analytical methods toprepare this data, which will help us to better understand not just our own business, customers or suppliers, but alsorelationships between subjects on market and so on.

The concept of 'big data' is just coming to existence and has uncertain origins. The term Big Data appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the tittle of Big Data and NextWave of InfraStress[18]. Big data mining was very relevant from the beginning, as the first book mentioning *Big Data* is a data mining book that appeared also in 1998 by Weiss and Indrukya[79]. However, the first academic paper with words *Big Data* in the tittle appeared a bit later in 2000 in a paper by Diebold[17]. The origin of the term 'Big Data' is due to fact that we are creating a huge amount of data every data.Usama Fayyad (2012) invited the KDD BigMine 12 workshop presented amazing data numbers about internet usage among them follows: each day Google has more than 1 billion queries per day, Twitter has more than 250 million tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in order of zettabytes, and it is growing around 40% every year[75]and the same opinion shared by [43], [25], [16], [27].

The contribution of this paper is to provide an analysis of available literature on big data analytics. Accordingly, some of the various big data tools, methods and technologies which can apply on different application of big data are discussed, and their issue and management provided in several decision domains are portrayed.

The literature was selected based on its novelty and discussion of important topics related to big data, in order to serve the purpose of our research. The publication years range from 2000-2015, with most of the literature focusing on big data ranging from 2010-2015. This due to big data being a recently focused upon topic. Furthermore, our quantity mostly includes research from some of the top journals, conferences, and white papers by leading corporations in the industry. Due to long review process of journal, most of the papers discussing big data analytics, its tools and methods, and its application were found to be conference paper, and white papers. While big data analytics

is being researched in academia, several of industrial advancements and new technologies provided were mostly discussed in industry paper.

A. Definition of Big data

Big data is relatively a new concept and a several definitions have been given to it by researchers, organizations and individuals. As far back as the information technology research company Gartner defines:

Big data are high-volume, high-velocity and high-variety information assets that require new form of processing to enhance decision making, insight discovery and processing optimizing [28].

One traditional and quite popular is definition created by McKinsey Global Institute (2011):

Big Data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse [44].

John Gantz has described Big Data in his article as following [63]:

Big Data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high velocity capture, discovery and analysis.

This definition more describes big data as a global term foe huge and complex sets of technologies, which are aims to manage and analyse mostly unstructured data which are important for the operation and development of a company [49].

According to Wikipedia**:**
Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications[7].

Big data is defined as the representation of the progress, usually include data set with sizes beyond the ability of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time [18].
According to [19], Big Data can be describe as, big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it.

B. Characteristics of Big Data

Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architecture, analytics, and tools in order to enable insights that unlock new sources of business value.Gartner analyst [40] introduced the 3Vs concept in MetaGroup research publication, 3D data management: Controlling data volume, variety and velocity.

VolumeThe volume of the data is its size, and how massive it is. Data volume is primary attribute of big data. Big data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Mover ever, Demchenko, Grosso, de Laat and Membrey stated that volume is the most important and distinctive feature of Big Data, imposing specific requirements to all traditional technologies and tools currently used.

VarietyIt is refers to the structural heterogeneity in a dataset. Technological advances allow firms to use various types of structured, semi-structured and unstructured data. Structured data is an organization data in a predefined format. The formatted data resides in fixed fields within a record or file. Unstructured data is a set of data with a complex structure that might or might not have a repeating pattern like e-mails, text, audio, video, or images files etc. Also in semi-structured data also known as schema-less or self-describing structure. Such type of data does not follow proper structure of data models as in relation database like web data and bibliographic data.

VelocityBig Data is generated in a very rapid speed. It is refer to the time in which Big Data can be processed. The Twitter fire hose (the stream of all tweets, globally), traffic data in mobile communication networks, and streaming video data are prime examples as this data flows at tremendous rates. This type of data at an exponential rate. For example Scrutinize 5 million trade events created each day to identify potential fraud; analyse 500 million daily records in real time to predict customer churn faster.

VeracityIBM coined veracity as the fourth V, which represents the unreliability natural in some sources of data. It is refer to the degree in which a leader trusted information in order to make a decision. For example, customer sentiments in social media are uncertain in nature, since they entail human judgment.

ValueOracle introduced value as a definition attribute of big data. It is refers to the important feature of the data which is defined by the added-value that the collected data can bring to the intended process, activity or predictive analysis. Big data are often characterized by relatively "low value density". That is, the data received in the original form usually has a low value relative to its volume. However, a high value can be obtained by analysing large volumes of such data.

Variability (and complexity)SAS introduced Variability and complexity as two additional dimensions of big data. Variability refers to the variation in the data flow rates. Data becomes more complex because it comes from multiple sources. The whole characteristics of big data with suitable examples presented [35], [27], [31].

C. Classification of Big Data

The characteristic of big data can be understood better by dividing it into classes. These classes are Data Sources,Content Format, Data Stores, Data Staging and Data Processing [32].

Data Sources: Web & Social, Machine, Sensing, Transactions and IoT
Content Format: Structured, Semi-Structured and Unstructured.

Data Stores: Document-oriented, Column-oriented, Graph based and Key-value

Data Staging: Cleaning, Normalization and Transform
Data Processing: Batch and Real time As the result of the wide variety of data sources, the captured data differ in size with respect to idleness, consistency and noise, etc. [31, 57].

II. ISSUE OF BIG DATA

There are many issues arising in big data. They are management issue, processing issue security issue and storage issues. Each issue has its own task of surviving in big data and mainly focusing on security issues[4].

Management Issue The biggest data management is the collection of large volumes of Structured and unstructured data from the organization, Government sector and Private and Public Administration. The motto of big data management is ensuring a high data quality, data ownership, responsibilities, standardization, documentation and accessibility[60].

Storage Issue The storage is achieved using virtualization in big data where it holds large data set of Sensor information, media, videos, E-business transaction records, and cell phone signal coordinates. The quantity of data is so large that to store the data we required a new storage media that can store hundreds of petabytes of data. Current disk technology limits are about 4 terabytes per disk. So, 1 Exabyte would require 25,000 disks. Overall solution of this issue: First, process the data in place and transmit only the resulting information.

Processing Issue The big data processing analyses the big data size in petabyte, Exabyte or even in Batch processing or Stream processing [9].

Security Issue There are fewer challenges for managing a large data set in secure, manner and inefficient tools, public and private database contain more threats and vulnerabilities, Volunteered and expected leakage of data, and deficiency of public and private policy makes a hackers to collect their resources whenever required. In Distribution programming frameworks, the security issue start working when massive amount of private data stored in a database which is not encrypted or in regular format

.

A. Big Data Management

Basically, data processing is seen as the gathering, Processing, Management of data for producing "new" information for end users. The some key challenges and issue related to big data. Big data can be manage in main three steps: Acquisition, Organization and Analyse. [22], [36], presented deeply big data management with suitable three steps as follow:

Acquisition. Big Data construction has to acquire high speed data from a variety of sources (web, DBMS(OLTP), NoSQL, HDFS) and has to deal with varied access procedures. Itis where a filter could be established to store only data whichcould be helpful or "raw" data with a lower degree of uncertainty. In some applications, the conditions of generationof data are important, thus it could be interesting for furtheranalysis to capture these metadata and store them with the corresponding data.

Organization. At this point the architecture has to arrangementwith various data formats (texts formats, compressed files, variously delimited, etc.) and must be able to explain them andextract the actual information like named entities, relation between them, etc. Also this is the point where datahave to be clean, put in a computable mode, structured orsemi-structured, integrated and stored in the right location(existing data warehouse, data marts, Operational Data Store,Complex Event Processing engine, NoSQL database). Thus a kind of ETL (extract, transform, and load) had to be done.Successfully cleaning in big data construction is not entirely guaranteed in fact "the volume, velocity, variety, and variability of big data may preclude us from tacking the time to cleanse it all thoroughly.

Analyse.Running queries, modeling, andbuilding algorithms to find new insights. Mining requires integrated,cleaned, trustworthy data; at the same time, datamining itself can also be used to help improve the quality andtrustworthiness of the data, understand its semantics, andprovide intelligent querying functions.

Decision. Being able to tack valuable decisions means to be also to efficiently interpret results from analysis. Consequently it is very important for the user to "understand the verify" outputs.

B. Big Data Techniques and Technologies

Since big data is not only large, but also varied and fast growing many technologies and analytical techniques are needed in order to attempt extracting relevant information. For processing large amount of data, the big data requires exceptional technologies. The techniques and technologies have been introduced for manipulating, visualizing and analysing big data[35].[12], discussed current techniques and technologies for exploiting data intensive applications. Big Data techniques involve a number of disciplines, including statistics, data mining, machine learning, neural networks, social network analysis, signal processing, pattern recognition, optimization methods and visualization approaches. Big data requires exceptional technologies to

efficiently process large quantities of data within tolerable elapsed times.

1) Techniques:-There are many types of techniques that could be employed when attacking a big data project. Which ones are used depends on the types of data being analysed, the technology available to you, and research questions you are trying to solve, and some tools can come up frequently:

Optimization Methods have been applied to solve quantitative problems in a lot of fields, such as physics, biology, engineering, and economics. Stochastic optimization, including genetic programming, evolutionary programming, and particle swarm optimization are useful and specific optimization techniques inspired by the process of nature. Real-time optimization is also required in many Big Data application, such as WSNs and ITSs. Data reduction and parallelization are also alternative approaches in optimization problems.

Statistics is the science to collect, organize, and interpret data. Statistical techniques are used to exploit correlationships and causal relationships between different objectives. Numerical descriptions are also provided by statistics. However, standard statistical techniques are usually not well suited to manage Big Data, and many researchers have proposed extensions of classical techniques or completely new methods [23].

Data mining is the process of analysing data from different perspectives and summarizing it into useful information-information that can be used to increase revenue, cuts costs, or both. It including clustering analysis, classification, regression and association rule learning. It including C4.5, k-means, SVM, Apriori, EM, Naive Bayes, and Cart, etc. These ten algorithms cover classification, clustering, regression, statistical learning, association analysis, and linking mining, all of which are the most important problems in data mining research [50, 51, 71].

Text Analytics is used for information retrieval from data. It is refer to techniques that extract information from textual data. Social network feeds, E-mails, blogs, online forums, and news and call centre records are all examples of text data. Text analytics involve machine learning, statistical analysis and computational linguistics. Text analytics enable to extract meaningful summaries from large scale data [27, 57].Information Extraction, Text Summarization, Question Answering and Sentiment Analysis are some of the techniques used in text analytics.

Audio Analytics is used to extract information from unstructured audio data. When applied to human spoken language, audio analytics is also referred to as speech analytics. Call centres and health services are commonly used utilization areas of audio analytics. Audio analytics can be used in numerous fields such as increasing the customer experience, the performance of customer representative and the sales rate; comprehending several tasks such as customer behaviors and the troubles of products [27,57].

Video Analytics is the usage of various techniques to extract meaningful information, track and analyse video stream. The increasing prevalence of closed-circuit television (CCTV) cameras and booming popularity of video sharing websites are the two leading contributors to growth of computerizes video analysis Marketing and operations management is the main application area of video analytics [27].

Machine learning is an important subjection of artificial intelligence which is aimed to design algorithms that allow computers to evolve behaviors based on empirical data. The most obvious characteristic of machine learning is to discovery knowledge and make intelligent decisions automatically[14].

Social Network Analysis (SNA) which has emerged as a key technique in modern sociology, views social relationships in terms of network theory, and it consists of nodes and ties. It has also gained a significant following in anthropology, biology, communication studies, economics, geography, history, information science, organizational studies, social psychology, development studies, and sociolinguistics and is now commonly available as a consumer tool. SNA include social system design, human behavior modeling, social network visualization, social networks evolution analysis, and graph query and mining. Recently, online social networks and Social media analysis have become popular. One of the main obstacles regarding SNA is the vastness of Big Data [27, 57].

Predictive Analytics (PA)is consist of a variety of techniques that predict future outcomes based on historical data and current data. Predictive analysis is used to capture the relationships of data and discover the patterns.PA which is primarily based on statistical methods, is highly applicable on many disciplines [27, 57].

2) Technology:-There are several software products and available technologies to facilitate big data analytics. Some of the most commonly used technology will discuss in this paper. Hadoop is key technology used to handle big data, its analytics and stream computing. It is an open source software project that enable the distributed processing of large data sets across clusters of commodity servers [35].

Hadoop

Hadoop is the most popular and open source implementation of MapReduce programming model. ApacheHadoop is a software framework for reliable, scalable, parallel and distributed computing. It is acostly hardware and expensive systems for processing and storing data, Apache Hadoop empowers parallelprocessing on Big Data on commodity hardware. Apache Hadoop consists of the Hadoop Distributed File System (HDFS), Map Reduce and related projects are Zookeeper, Hbase, and Apache Hive etc. [51].

*HDFS*it is a self-healing, distributed file system that provides reliable, scalable and fault tolerant data storage on commodity hardware. It works closely with MapReduce by distributingstorage and computation across large clusters by

combining storage resources that can scale depending uponrequests and queries while remaining inexpensive and in budget. HDFS accepts data in any format like text,images, videos, etc. regardless of architecture and automatically optimizes for high bandwidth streaming. HDFS consists of nodes:

NameNode: The NameNode is the central location for information about the file system deployed in a Hadoop environment. An environment can have one or two NameNode, configured to provide minimal redundancy between the NameNode. The NameNode is contacted by clients of the Hadoop Distributed File System (HDFS) to locate information within the file system and provide updates for data they have added, moved, manipulated, or deleted.

DataNodes: DataNodes make up the majority of the servers contained in a Hadoop environment. Common Hadoop environments will have more than one DataNodes, and oftentimes they will number in hundreds based on capacity and performance needs. The DataNodes server two functions: it contains a portion of the data in the HDFS and it acts as a compute platform for running jobs, some of which will utilized the local data within the HDFS.

EdgeNode: The EdgeNode is the access point for the external applications, tools, and users that need to utilize the Hadoop environment. The EdgeNode sites between the Hadoop cluster and the corporate network to provide access control, policy enforcement, logging and gateway services to the Hadoop environment. A typical Hadoop environment will have a minimum of on EdgeNode and more based on performance needs.

*MapReduce*it is programming model or a software framework used in Apache Hadoop. HadoopMapReduce is provided for writing applications which process and analyse large data sets in parallel on large multinode clusters of commodity hardware in a scalable, reliable and fault tolerant manner. Data analysis andprocessing uses two different steps namely, Map phase and Reduce phase. MapReduce computation performance as follow:

1. Each Map function is converted to key-value pairs based on input data. The input to map function is tuple or document. The way key-value pairs are produced from the input data is determined by the code written by the user for the Map function.
2. The key-value pairs from each Map task are collected by a master controller and sorted by key. The keys are divides among all the reduce tasks, so all key-value pairs with the same key wind up at the same Reduce task.
3. The Reduce tasks work on one key at a time, and combine all the values associated with that key in some way. The manner of combination of values is determined by the code written by user for the reduce function [76, 31, 68].

According [21], the MapReduce function within Hadoop depends on two different nodes: the Job Tracker and the Task Tracker nodes. The job Tracker nodes are the ones which are responsible for distributing the mapper and reducer functions to the available Task Trackers, as well as monitoring the results. On the other hand, the Task Taker nodes actually run the jobs and communicate results back to the Job Tracker. Also several author discuss these points like [50, 66]. Also other *project in HADOOP*, Other than MapReduce and HDFS, the entire Apache Hadoop is now commonly considered to consist of a number of related project [35]:

HBase: it is scalable and distributed database that supports structured data storage for large tables.
Pig: Apache Pig allows you to write complex MapReduce transportations using a simple scripting language.

Hive: Apache Hive is data warehouse infrastructure built on top of Apache Hadoop for providing data summarization, ad-hoc query, and analysis of large datasets.

Sqoop: Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and Structured data stores such as relational databases.

ZooKeeper: A high-performance coordination service for distributed applications ZooKeeper offers operations services in the Hadoop framework.

Mahout: Apache Mahout is a library of scalable machine-learning algorithms, implemented on top of Apache Hadoop and using the Map Reduce paradigm.

Spark: LA fast and general compute engine for Hadoop data. It is provides a simple and expressive programming model that support a wide range of applications, machine learning, stream process, and huge graph computation.

III. KEY APPLICATION OF BIG DATA

Currently, enormous amounts of data are created every day. With the rapid expansion of data, we are moving from the Terabyte to the Petabyte Age. At the same time, new technologies make it possible to organize and utilize the massive amounts of data currently being generated. However this trend creates a great demand for fresh data storage and analysis methods. Big Data Application Specific emerging application include areas such as healthcare, manufacturing, and traffic management etc. All of these applications rely on huge volumes, velocities and varieties of data to transform the behavior of a market.

Application of Healthcare and medical big data a widespread use of big data in the health sector can help doctors make the right choices more quickly, on the basis of information collected by other medical staff. Patients can benefit from more timely and appropriate treatments and be better informed about health care providers. An increased use of data analysis in the health sector can also lead to enormous cost savings through a more precise identification

of unnecessary procedures or duplication of tests. The analysis of large clinical datasets can result in the optimization of the clinical and cost effectiveness of new drugs and treatments. Shelly Gupta et al. (2011), studied the performance analysis of several data mining classification techniques by using three different machine learning tools over the healthcare datasets. In this study, different data mining classification techniques have been tested on four different healthcare datasets. The standards used are percentage of accuracy and error rate of every applied classification technique [30]. David C. Kale et al. (2011), addressed this problem by applied a generalized hashing framework, namely kernelized locality sensitive hashing, to accelerate time series similarity search with a series of represented similarity metrics. Experiments result on three large scale clinical data sets demonstrates the effectiveness of the proposed approach [38].Olegas Nioksu and Olga Kurasovo (2013), showed data mining perception and practical applications in healthcare was a way beyond its steady growth in the academic research field, which raises a hypothesis, that relatively a little percentage of academic research effort results in practical DM applications in healthcare out of which they was conclude that the current interdisciplinary approach in not efficient enough[55].Jianqing Fan and Han Liu (2013), discussed statistical methods for estimating complex correlation structure from large pharmacogenomics datasets. They selectively overviewed several prominent statistical methods for estimating large covariance matrix for understanding correlation structure, inverse covariance matrix for network modelling, large-scale simultaneous tests for selecting significantly differently expressed genes and proteins and genetic markers for complex diseases and high dimensional variable selection for identified important molecules for understanding molecule mechanisms in pharmacogenomics [24].M. Durairaj and V. Ranjani (2013), compared the different data mining application in the healthcare sector for extracting useful information. Used the data mining technique for prediction of diseases and increase the diagnostic accuracy. Also reduced the cost and time constraint in terms of human resources and expertise [20].Alain Yee Loong Chong et al.(2014), extended existing work by integrating unified theory of acceptance and use of technology (UTAUT) and individual differences, namely personality and demographic characteristics to predicted the RFID in healthcare supply chain [1].Tanh Nguyen et al. (2014), presented a combination of wavelet features with fuzzy SAM (Standard additive model) for medical diagnosis. Medical data are often noisy and collected in high dimensional format. Used of fuzzy system helps to handle the noisiness and complexity of medical data. SAM helped to reduce the high dimensional data [53]. OsdenJokonya (2014), proposed a big data integrated framework to assist with prevention and control of HIV/AIDS, TB and silicosis(HATS) in the mining industry. Proposed big dataframework has the potential of addressing the needs of predictive epidemiology which is important in forecasting anddisease control in the mining industry. The paper therefore places a foundation for the use of practical systems model and bigdata to address the challenges of HATS in the mining industry. Also the future work, the framework will be validated using sequential explanatory mixed methods case study approach in mining organizations [56].Changwon Yoo, Luis Ramirez and Juan Liuzzi (2014), introduced modern statistical model learning and bioinformatics approach that have been used in learning statistical relationship from big data in medicine and behavioural science that typically include clinical, genomic and environmental variable. They explained modern statistical method called Bayesian networks i.e. suitable in analysing big data sets that consists with different type of large data form clinical, genomic and environmental data [11]. Ming Yang, Melody Kiang, and Wei Shang (2015), discussed the problem of filtering big data from social media in general and the application to consumer ADR (adverse drug reaction) massage identification. The LAD model reduced the high dimension problem that social media analysis facing [81]. Saravana kumar N. M. et al. (2015), researcher used the predictive analysis algorithm in Hadoop/Map reduces environment to predict the Diabetes types prevalent, complications associated with it and the type of treatment to be provided. Based on the analysis, this system provides an efficient way to cure and care the patients with better outcomes like affordability and availability [65].Jisha Jose Panackal and A. S. Pillai (2015), included performance evaluation of AUA (Adaptive Utility-based Anonymization) model using data sets and proves thatthe data anonymization can be done without compromising the quality of data mining results [58].FilipDabek and Jesus J. Caban (2015), presented an SVM-based model that has been trained with a longitudinal dataset of over 5.3million clinical encounters of 89,840 service members that have sustained a concussion. Themodel has been tested and validated with over 16,045 patients that developed PTSD and it hasshown an accuracy of over 85% (AUC of 86.52%) at predicting the condition within the firstyear following the injury [26].

Application of Transportation big data sector can clearly benefit from big data collected through sensors, GPS data and social media in particular. A smart use of big data supports governments in optimising multimodal transport and managing traffic flows, making our cities smarter. Citizens and companies can save time through the use of route planning support systems.David Levinson and Wei Chen (2006), used multiple regression models to evaluate the long-run performance of four traffic management systems. Finally studied regression analysis was a simple and effective research method for testing the microscopic association between traffic management and traffic system performance [41].Dass P.J.H. et al. (2013), discussed opportunities and challenges associated with using big data and official statistics. Obtained with analyses of large amount of Dutch traffic loop detection records and Dutch social media massages are used to illustrate the topics specific to the statistical analysis of big data. Detected the massage were classified as positive, negative or neutral by using sentiment analysis[15].Chen Zhong, Xianfeng Huang et al. (2014), used data mining and spatial analysis techniques for urban analysis. They introduced a method of combining survey and smart-card data to infer information

about social activities and to generated insights into urban building functions. They proposed a probabilistic model to daily activities from individuals' travel and infer building functions from corresponding daily activities [83]. Georgios Georgiadis et al. (2014), used DEA to evaluate performance of individual bus line, considering them as sub-units of PT system. The DEA exercise is based on the bus PT network of Thessaloniki, Greece and examined the relative efficiency and effectiveness of the city's bus routes. Also used bootstrapping techniques to check robustness of DEA results and the performance assessment showed the more reliable when correcting for bias[29]. Roy Ka-Wei Lee and Tin Seong Kam (2014), studied the explores and discussed the potential use of time-series data mining, a relatively new framework by integrating conventional time-series analysis and data mining techniques, to discover actionable insights and knowledge from the transportation temporal data. Also they studied a case study on the Singapore public train transit was used to demonstrate the time-series data-mining framework and methodology[64]. N. Sari Aslam et al. (2015), provided an insight into the user behaviour of the bicycle hire network and allowed future infrastructure investment decision to be made in an informed manner [3].Qi Shi, and Mohanmed Abdel-Aty (2015), used the DM and Bayesian statistics techniques for identified the leading contribution factors to crashes in real time. First-order Reliability Method (FORM) model was constructed based on the real-time crash prediction model and critical point of system congestion index (CI), volume and speed was calculated. Finally, the proposed framework highlights the association between congestion and safety [70].Li Li et al. (2015), discussed the important traffic prediction problem, Lasso Granger causality models to screen out most irrelevant or redundant data. The corresponding algorithm has good scalability, robustness, real-time responsiveness, and can to a traffic prediction model that strikes a good balance between model complexity and model performance [24].Rashid Mehmood, Gary Graham (2015), aim to make acontribution to big data and city operations theory by exploring how big data can lead to improvements in transport capacity sharing. Researcher explored using Markov models the integration of big data with future city (health-care) transport sharing. Designed a mathematical model to illustrate how sharing transport load (and capacity) in a smart city can improve efficienciesin meeting demand for city services [47].

Application of Banking and Financial big datathe big data revolution happening in and around 21st century has found a significance with financial service firms,considering the valuable data they've been storing since many decades. And even though the collection of this datawas unplanned, since accounting system has always been historical in nature, the potential unlocked by big dataanalytics exceeds any expectation previously expected from this historical record set. This data has now unlockedsecrets of money movements, helped prevent major disasters and thefts and understand consumer behavior. Banksgain the most benefits from big data as they now can extract good information quickly and easily from their data andconvert it into meaningful benefits for themselves and their customers.Edward W. Sun, Yi-Ting Chen, and Min-The Yu (2015), proposed a new information extraction method for big financial data based on wavelet transform- named the generalized optimal wavelet decomposition algorithm (GOWDA) and determined the combination of denoising factors based on multivariate goodness of fit score function i.e. optimally deconstructing the big financial data and removed the market microstructure noise [73].Utkarsh Srivastava, Santosh Gopalkrishnan (2015), the aim of this paper to capture of big data analytics is being successfully used in banking sector, with respect toSpending pattern of customers,Channel usages, Customer Segmentation and Profiling,Product Cross Selling based on the profiling to increase hit rate,Sentiment and feedback analysis,Security and fraud management [72]. Jidong Chen et al. (2015), introduced the Fraud Risk Management at Alibaba under big data. Researchers planed the fraud risk management at Alibaba and a big data prevention product. Also extended the analysis to build a safer and cleaner payment environment [37].XinhuiTian, Rui Han, Lei Wang, Gang Lu, Jianfeng Zhan (2015), analysed the challenges brought by the financial latency critical big datacomputing, then proposed a discussion on how to handle these challenges from a perspective of multi-level system. Also communicated about current researches on low latency in different system levels. Discussion and conclusions of these paper useful to thebanking and financial organizations with the critical latency requirement of big data analytics [80].

Application of Social Media Big data it is refer to the structured and unstructured data from social media channels. Social Media is a broad term is encompassing a variety of online platforms that allow users to create and exchange content.V. U. Parvathi and K. Lyakutti (2012), observed that decision tree model surpasses the neural network model in the prediction of churn and it was also easy to construct. Also explored the application of data mining techniques in predicting likely churners and the impact of attribute selection on identifying the churn [74].William M. Campbell et al. (2013), studied research work on Licoln Laboratory was addressing the problem in constructing networks from unstructured data challenge, analysis the community structure of a network, and inferring information from networks. Graph and content analysis has proven to be valuable tools in solving unstructured data challenges. Used two tools in Laboratory to achieved promising results on real world data. Also sampling techniques presented in this article [9].Zhen-Yu Chen, et al. (2014), developed a hierarchical ensemble learning framework for behaviour aware user response modelling in social media using diverse. In the framework, a general-purpose data transformation and feature extraction strategy is proposed and an improved H-MK-SVM (Hierarchical multiple kernel support vector machine) is developed [82].M. Vasuki, J. Arthi and K. Kayalvizhi (2014), proposed a new algorithm for twitter sentiment analysis and it was based on three way classification framework. This sentiment analysis provided a

solution for obtaining feedback about any product and presenting the overall picture. They used Flume tools; Flume was used to obtained data from any social web site and store in HDFS and produced a dynamic result [77].A. Weichselbraun, S. Gindl, A. Scharl (2014), introduced a novel method to extend sentimentlexicons with concept knowledge, which aims to increase thelexicons' coverage and derive concept information for subsequent opinion mining. Researchers used SenticNet terms and their polarity valuesto generate a baseline sentiment lexicon, identify ambiguous sentimentterms, and extract context information for disambiguatingthese terms in the application phase [78].Mahalakshmi R andSuseela S (2015), proposed SoSA (social sentiment analysis) system was able to collect meaningful information from twitter and effectively perform sentiment analysis. This method was composed of a HDFS system based on the Hadoop ecosystem and Map reduces functions. The map reduced function was used to classify the twitter sentiment polarity and score [45].H. Andrew Schwartz andAndLyle H. Ungar (2015), measured people's thoughts, feelings, and personalities using carefully designed survey questions, which are often given to a relatively small number of volunteers. The proliferation of social media, such as Twitter and Facebook, offers alternative measurement approaches: automatic content coding at unprecedented scales and the statistical power to doopen-vocabulary exploratory analysis [67].Rajni Singh, Rajdeep Kaur (2015), developed a combined dictionary based on social media keywords and online review and also found hidden relationship pattern from this keyword [69].Gema Bello-Orgaza, et al. (2015), presented vision of the new methodologies that is designed to allow for efficient data mining and information fusion from social media and of the new applications and frameworks that are currently appearing under the "umbrella" of the social networks, social media and big data paradigms [6]. Thin Hai Nguyen et al. (2015), the goal of this research paper was to build a mode to predict stock price movement using the sentiment from social media. In this paper presented the novel method to integrate the sentiments in social media for the prediction of stock price movement [54].

Application of Agriculture big data a smart use of big data in agriculture can increase productivity, food security and farmer incomes at the same time. Through an intelligent and widespread use of data coming from sensors the ways we are farming today can be changed entirely for the better. This can lead to a more efficient use of natural resources (including water or sunlight) in our farming practices. With advanced technologies farmers can have access to data in real time on how their farm machinery is working as well as to historic weather patterns, topography and crop performance. Chandrakanth. Biradar and Chatura S Nigudgi (2012),provided a comprehensive review of specific decision tree classifier, discrete wavelet transformation for statistical agricultural data available and graphical user interface method for the purpose of data mining to enhance energy forecasting analysis. The result of this study was developed an effective decision tree classifier used for the prediction of the data and energy in producing and marketing of the seeds [8]. Chuang Ma, Hao Helen Zhang, Xiangfeng Wang (2014), review introduced the basic concepts and procedures of machine-learning could interface with big data technology to facilitate basic research and biotechnology in the plant sciences [13].MengXu and Seung Yon Rhee (2014), researchers highlight the purpose, common problems and general principle in data analysis. Used the RNA sequential analysis to illustrate the rationale behind some of the choices made in statistical data analysis. Finally, provided a list of free online resources that emphasize intuition behind quantitative data analysis [48].

Application of Manufacturing and retailBig Data it means the optimization of operations on a real-time basis for the manufacturing industry. Key benefits of using big data analytics include boosting product quality through improved defect tracking, better manufacturing processes and optimized supply tracking. Big data enables the timely and appropriate delivery of products for consumers and more efficient processes, with cost savings, for business. With big data retail companies better know the needs and interests of costumers and, as a result, offer more personalized products.Tejaswini Abhijit Hilage and R. V. Kulkarni (2011), studied the data mining techniques on large database. Examined the result after applying association rule mining technique, rule induction technique and Apriori algorithm. These techniques are applied to the database of shopping mall. Market basket analysis is performing by using these techniques and some important results are found such as buying behaviour [33].A. Athanasiadis Ioannis and Ioannides Dimitriosb (2015), in this paper researcher considered a large dataset from a big company of wine, with white, rose and red wines (from Greece). The logistic regression givesstronger results in terms of the interpretation, outperforming the linear regression and also researchers expecting that others methods asneural networks and support vector machine improve the last one [34].

Application of Education big data the education has also transformed in the current era. It has turned from offline/Indoors to online. The studentdata size has extended from 100 to millions. The size of data generated in the education domain has grown manyfolds; the education has changed from chalk and talk to mobiles and PDA's. The needs of the online students havechanged w.r.t time, language and scope of the subject. It therefore is of significant concern to handle their queriesand provide tailor made curriculum and suggestion to opt for suitable courses[61].Jin-Tae Park et al. (2013),calculated the students' movement pattern data from the combined patterns of the moving path and behaviour were calculated. By using theGPS collected data, like time, grade, and sex, via smartphone. Determined the specific and detailed student movement patterns. Extended the future work increasing the reliability of the data and gathering of information of the day of the week, weather, blood type, and occupation [59]. ThiagoGraca Ramos et al. (2015), aimed to organize data from different sources, adding in a single database through a

data warehouse so as to make it feasible the analysis in order to understand the assessment of basic education in the perspective of the State Reviewer as a mechanism that generates information regarding positivity andweaknesses of a school or an educational system to provide improvements [62]. Banica Logica, Radulescu Magdalena (2015), discussed aspects regarding the evolution of Big Data technologies, the way of applying them to e-Learning and theirinfluence on the academic environment. Also, designed a three-step system architecture for a consortium of universities,based on actual software solutions, having the purpose to analysed, organize and access huge data sets in the Cloud environment. Researchers focused research on exploring unstructured data using the graphical Gephi tool [5].Marin Fotache, Catalin Strimbei (2015), presented the main coordinates of data processing today and some implicationsfor academic set of courses. It argues that data analysis and business intelligence professionals could benefit if trained to acquire aproper level of SQL and data warehouses knowledge [46].Kanyarat Bussaban, Phanu Waraporn (2015), aimed at showing how to motivate the significanceof mastering data science proficiency as well as depicting examples and resources for lecturers in implementing data science incomputer sciences and mathematics curriculum. Two case studies from Computer Science and Informatics MathematicsPrograms at Faculty of Science and Technology, Suan Sunandha Rajabhat University in Bangkok, Thailand are presented [39].

IV. CONCLUSION

In this paper, we review of whole background of big data. Firstly introduced basic introduction of big data and discuss the issue and management of big data in real life. Also the review of advanced statistical tools and techniques to used big data analysis also software to process huge amount of data like Hadoop, MapReduce etc. Finally review the key application of big data like healthcare, banking, market, education, agriculture and social media. To gain useful knowledge from a structure and unstructured nature of different area of big data in real life. In big data take deep research work using advanced techniques and software in every field having large complex data set.

V. FUTURE SCOPE OF THE STUDY

1. After studying the review on big data, it is found that no satisfactory Statistical study of unstructured big data viz, social media data like email, facebook, twitter, You tube and WhatsApp. Further research carried out by in this direction will certainly help to researchers to aware with different application of big data.

2. However, it is need to prepare Atlas of Statistical studies from different applications of big data. Such an attempt will definitely facilitate, researcher to get aware with different statistical studies of unstructured big data.

3. Future work focuses on the analysis part of the big data sorting by applying a different data mining techniques in it.

ACKNOWLEDGMENT

I acknowledgment my sincere and profound gratitude to my guide, Dr.Omprakash S. Jadhav, for his valuable guidance, dedicated concentration and support throughout this work. I also acknowledge my sincere gratitude to authorities of Dr. Babasaheb Ambedkar Marathwada University, Aurangabad and other teaching staff of Statistics department for their help and support. I am also thankful to my friends for their cooperation.

REFERENCE

[1] Alain Yee-Loong Chong, Martin J. Liu, Jun Luo and Ooi Keng- Boon (2015) Predicting RFID Adoption in Healthcare supply chain from the perspectives of users. Int. J. Production Economics, 159, 66-75.
[2] Ashlesha S. Nagdive, R. M. Tugnayat, Manish P. Tembhurkar (2014) Overview on Performance Testing Approach in Big Data. International Journal of Advanced Research in Computer Science, Vol. 5, No. 8.
[3] Aslam N. Sari, Cheshire J. and Cheng T. (2015) Big Data Analysis of Population Flow between TfL Oyster and Bicycle Hire Networks in London. University College London (UCL), Department of Civil, Environmental and Geomatic Engineering, Gower St, London, UK.
[4] B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M. S. SaleemBasha, P. Dhavachelvan (2015) Big Data and Hadoop-A Study in Security Perspective. 2nd International Symposium on Big data and Cloud Computing (ISBCC'15), Procedia Computer Science 50, 596-601.
[5] Banica Logica, Radulescu Magdalena (2015) Using Big Data in the Academic Environment. 7th International Conference, the Economies of Balkan and Eastern Europe Countries in the Changed World, EBEEC, Procedia Economics and Finance 33, 277-286.
[6] Bello-Orgaza Gema, Jungb Jason J., Camacho David (2015) Social big data Recent Achievements and new Challenges. Information Fusion, 1-15.
[7] Big Data, http://en.wikipedia.org/wiki/Big-Data,2014.
[8] Biradar Chandrakanth., Nigudgi Chatura S (2012) An Statistical Based Agriculture Data Analysis. International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 9.
[9] Campbell William M., Dagli Charlie K., and Weinstein Clifford J. (2013) Social Network Analysis with Content and Graphs. Lincoln Laboratory Journal, Volume 20, Number 1.
[10] Changqing (2012) Big data Processing in Cloud Computing Environments. International Symposium on Pervasive Systems, Algorithms and Networks.
[11] Changwon Yoo, Luis Ramirez, Juan Liuzzi (2014) Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine. International Neurourology Journal.
[12] Chen C.L. Philip, Zhang Chun-Yang (2014) Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences 275, 314–347
[13] Chuang Ma, Hao Helen Zhang, Xiangfeng Wang (2014) Machine learning for big data analytics plants. Trends in plant science, Vol.19, no. 12.
[14] Chun- Wei Tsai, Chin- Feng Lai, Han- Chieh Chao Athanasios V. Vasilakos (2015) Big data analytics: a survey. Tsai et al. Journal of Big Data, 2:21.
[15] Daas P. J. H., Puts M. J., Buelens B. and Van den Hurk P. A. M (2013) Big Data and Official Statistics. Statistics Netherlands, Methodology Sector and Traffic and Transport Sector.
[16] Diebold, F. X. (2012) A personal perspective on the origin(s) and development of Big Data: The phenomenon, the term and the discipline (Scholarly Paper No.ID 2202843).

[17] Diebold Francis X (2000) Big Data Dynamic Factor Models for macroeconomic measurement andforecasting. Advance in Economics and Econometrics, Eight World Congress of the Econometric Society.
[18] Douglas and Laney (2008) The importantance of big data: A definition.
[19] DumhillE. (2012) What is big data?http://strata.oreilly.com/2012/01/what-is-big-data- html.
[20] Durairaj M., Ranjani V. (2013) Data Mining Applications in Healthcare Sector: A Study. International Journal of Scientific & Technology Research Volume 2, Issue 10, ISSN 2277-8616.
[21] EMC: Data Science and Big Data Analytics (2012). In: EMC Education Services, PP. 1-508.
[22] Emani Kacfah Cheikh, Nadine Cullot, Christophe Nicolle (2015) Understandable Big Data: A Survey. Computer Science Review.
[23] Fan Jianqing, Han Fang and Liu Han (2014) Challenges of Big Data Analysis. National Science Review, 1:293-314.
[24] Fan Jianqing and Liu Han (2013) Statistical Analysis of Big Data on Pharmacogenomics. Preprint Submitted to Advanced Drug Delivery Reviews.
[25] Fan Wei and Bifet Albert (2013) Mining Big Data: Current Status, and Forecast to the Future ACM SIGKDD Explorations Newsletter, Vol. 14, Issue 2, pp. 1-5.
[26] FilipDabek, Jesus J. Caban (2015) Leveraging Big Data to Model the Likelihood of Developing Psychological Conditions after a Concussion. Procedia Computer Science, INNS Conference on Big Data, Vol.53, Pages 265-273.
[27] Gandomi Amir and Haidar Murtaza (2015) Beyond the hype: Big data concepts, methods,Analytics. International Journal of Information Management, 35, 137-144.
[28] Gartner IT Glossary (n.d). Retrieved from http://www.gartner.com/it-glossary/big-data/
[29] Georgiadis Georgios, Politis Loannis, Papaioannou Panagiotis (2014) Measuring and Improving the Efficiency and effectiveness of bus public transport system. Research in Transportation Economics 48, 84-91.
[30] Gupta Shelly, Kumar Dharminder and Sharma Anand (2011) Performance Analysis of Various Data Mining Classification Techniques on Healthcare Data. International Journal of Computer Science & Information Technology (IJCSIT) Vol. 3, No 4.
[31] Hadi Hiba Jasim, Shnain Ammar Hameed, Hadishaheed Sarah, Ahmad Azizahbt Haji (2014) Big Data and Five V's Characteristics. Proceedings of IRF International Conference, Tirupati, India, ISBN: 978-93-84209-61-2.
[32] Hashem Ibrahim Abaker Targio, Yaqoob Ibrar, Anuar Nor Badrul, Mokhtar Salimah, Gani Abdullah, Khan Samee Ullah (2015) The rise of "Big Data" on cloud computing: Review and open research issues. Information System 47, 98-115.
[33] Hilage Tejaswini Abhijit &Kulkarni R. V. (2011) Application of data mining Techniques to a selected business organization with special reference to buying behaviour. International Journal of Database Management Systems (IJDMS) Vol.3, No.4.
[34] Ioannis Athanasiadis, Dimitrios Ioannides (2015) A Statistical Analysis of Big Web Market Data Structure Using a Big Dataset of Wines. 7th International Conference, the Economies of Balkan and Eastern Europe Countries in the Changed World, EBEEC, Procedia Economics and Finance 33, 256 – 268.
[35] Ishwarappa, Anuradha J. (2015) A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. International Conference on Intelligent Computing, Communication and Convergence (ICCC-2015), Procedia Computer Science 48, 319-324.
[36] H.V. Jagadish, D. Agrawal, P. Bernstein, E. e. a. Bertino (2015) Challenges and Opportunities with Big Data, The Community Research Association.
[37] Jidong Chen, Ye Tao, Haoran Wang, Tao Chen (2015) Big Data based fraud risk management at Alibaba. The Journal of Finanace and Data Science 1, 1-10, http://www.keaipublishing.com/en/journal/jfds/
[38] Kale David C., Gong Dian, Wetzel Zhengping Che Randall, Ross Patrick (2011) An Examination Multivariate Time Series Hashing with Applications to Health Care. American Hospital Association Hospital Statistics survey conducted in 2009 and published in 2011 by Health Forum, LLC, and the American Hospital Association.
[39] Kanyarat Bussaban, Phanu Waraporn (2015) Preparing undergraduate students majoring in Computer Science and Mathematics with Data Science perspectives and awareness in the age of big data. 7th World Conference on Educational Sciences, (WCES-2015), Novotel Athens Convention Centre, Athens, Greece, Procedia- Social and Behavioral Sciences 197, 1443-1446.
[40] Laney. D. (2011) 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note.
[41] Levinson David, Chen Wei (2006) Traffic Management System Performance Using Regression Analysis. California Path Program Institute of Transportation Studies University of California, Berkeley, UCB-ITS-PWP-3.
[42] Li Li, Xiaonan Su, Yanwei Wang, Yuetong Lin, Zhiheng Li and Yuebiao Li (2015) Robust causal dependence mining in big data network and its application to traffic flow predictions. Transportation Research Paper C 58, 292-307.
[43] Mannem Rama Devi, Nukala Vijay Kumar (2015) Information Mining Thorugh Big Data. International Journal of Research in Computer and Communication Technology, Vol. 4, Issue 8.
[44] McKinsey Global Institute (2011) Big data: The next frontier for innovation, competition and productivity. Retrieved February 14, 2015 from http://bigdatawg.nist.gov/MGI_big_data_full_report.pdf
[45] Mahalakshimi R., Sussela S. (2015) Big-Sosa: Social Sentiment Analysis and Data Visualization on Big Data. International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 4.
[46] Marin Fotache, Catalin Strimbei (2015) SQL and data analysis some implications for data analysis and higher education. 7th International Conference on Globalization and Higher Education in Economics and Business Administration, GEBA 2013, Procedia Economics and Finance 20, 243-251.
[47] Mehmood Rashid, Graham Gary (2015) Big data logistics: a health-care transport capacity sharing model. Conference on Enterprise Information Systems / International Conference on Project Management / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist, Procedia Computer Science 64, 1107 – 1114.
[48] MengXu and Seung Yon Rhee (2014) Becoming data-savvy in a big data world. Trend in plant Science, Vol. 19, No. 10.
[49] Milan Kubina, Michal Varmus, Irena Kubinova (2015) Use of big data for competitive advantage of company. 4th World Conference on Business, Economics and Management, WCBEM, Procedia Economics and Finances 26, 561-565.
[50] Min Chen, Shiwen Mao, Yunhao Liu (2014) Big Data: A Survey. Mobile Network Application, 19:171-209.
[51] Mohd Rehan Ghazi, Durgaprasad Gangodkar (2015) Hadoop, MapReduce and HDFS: A Developers Perspective. International Conference on Intelligent Computing, Communication and Convergence (ICCC-2015), Procedia Computer Science, 48, 45-50.

[52] Murale Narayanan, Aswani Kumar Cherukuri (2016) A study and analysis of recommendation systems for location-based social network (LBSN) with big data. IIMB Management Review, XX, 1-6.
[53] Neguyen Thanh, Khosravi, Creighton Douglas and Nahavandi Saeid (2015) Classification of Healthcare data using genetic fuzzy logic system and waveletes. Expert System with Application 42, 2184-2197.
[54] Nguyen Thien Hai, Shirai Kiyoaki and Velcin Julien (2015) Sentiment analysis on Social media for stock movement prediction. Expert System with Applications, 1-9.
[55] Niaksu Olegas and Kurasova Olga (2013) Data Mining Applications in Healthcare: Research vs. Practice. Vilnius University, Institute of Mathematics and Informatics, Akademijos str.4, LT-08663, Vilnius, Lithuania.
[56] OsdenJokonya (2014) Towards a Big Data Framework for the prevention and control of HIV/AIDS, TB and Silicosis in the mining industry. CENTERIS 2014-Conference on ENTERprise Information System/ ProjMAN 2014-International Conference on Project MANagement/HCIST 2014- International Conference on Health and Social Care Information Systems and Technologies. Procedia Technology 16, 1533-1541.
[57] Ozkose Hakan, Ar Emin Sertac, Gencer Cevriye (2015) Yesterday, Today and Tomorrow of Big Data. World Conference on Technology, Innovation and Entrepreneurship, Procedia-Social and Behavioral Sciences 195, 1042-1050.
[58] Panackal Jisha Jose, Pillaib Anitha S (2015) Adaptive Utility-based Anonymization Model: Performance Evaluation on Big Data Sets. 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15),Procedia Computer Science 50, 347 – 352
[59] Park Jin-Tae, Kang Hee-Soo, Yun Jun-Soo, Moon IL-Young (2013) Analysis of Students' Movement Patterns through Big Data. Advanced Science and Technology Letters, Vol. 44, pp.1-4.
[60] Philip Russom (2013) Managing Big Data. TDWI research, Fourth Quarter.
[61] Rajiv Pandey, Manoj Dhoundiyal (2015) Quantitative Evaluation of Big Data Categorical Variables through R. International Conference on Information and Communication Technologies (ICICT 2014), Procedia Computer Science 46, 582-588.
[62] Ramos Thiago Graca, Machado Jean Cristian Ferreira, Cordeiro Bruna Principe Vieira (2015) Primary Education Evaluation in Brazil using Big Data and Cluster Analysis. Information Technology and Quantitative Management (ITQM 2015), Procedia Computer Science 55, 1031 – 1039.
[63] Reinsel D., Gantz J. (2011) Extracting value from Chaos. Retrieved April 02, 2015 from http://www.emc.com/collateral/analyst-report/idcextracting-value-from-chaos-ar.pdf.
[64] Roy Ka-Wei Lee and Tin Seong Kam (2014) Time-Series Data Mining in transportation: A Case Study on Singapore Public Train Commuter Travel Patterns. IACSIT International Journal of Engineering and Technology, Vol. 6, No. 5.
[65] Saravana Kumar N.M., Eswari T. Sampath P. and Lavanya S. (2015) Predictive Methodology for Diabetic Data Analysis in Big Data. 2nd International Symposium on Big Data and Cloud Computing (ISBCC 15), Procedia Computer Science 50, 203-208.
[66] SeemaMaitrey, C. K. Jha (2015) MapReduce: Simplified Data Analysis of Big Data, Procedia Computer Science 57, 563-571.
[67] Schwartz H. Andrew and Ungar Lyle H. (2015) Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods. ANNALS, AAPSS, 659.
[68] Singh Puneet Duggal, Paul Sanchita (2013) Big Data Analysis: Challenges and Solutions. International Conference on Cloud, Big Data and Trust, 13-15, RGPV.
[69] Singh Rajni and Kaur Rajdeep (2015) Sentiment Analysis on Social Media and Online Review. International Journal of Computer Applications (0975 – 8887) Volume 121 – No.20.
[70] Shi Qi and Abdel-Aty Mohamed (2015) Big data application in real-time traffic operation and safety Monitoring and improvement on urban expressways. Transportation Research Part C 58, 380-394.
[71] Shyam R., Bharathi Ganesh HB, Sachin Kumar S., Prabaharan Poornachandran, Soman K. P. (2015) Apache Spark a Big Data Analysis Platform for Smart Grid. SMART GRID Technologies, Procedia Technology 21,171-178.
[72] Srivastava Utkarsh, Gopalkrishnan Santosh (2015) Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks. 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Procedia Computer Science 50, 643 – 652.
[73] Sun Edward W., Chen Yu-Ting, Min-The Yu (2015) Generalized Optimal Wavelet Decomposing Algorithm for Big Financial Data. Int. J. Production Economics 165, 194-214.
[74] Umayaparvathi V. and Lyakutti K. (2012) Application of data mining techniques in telecom churn application. International Journal of computer applications (0975-8887), Vol. 42, No.-20.
[75] Usama Fayyad (2012) Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling. http://big-data-mining.org/keynotes/#fayyad.
[76] Varma Yashika, Smit Hooda (2015) A Review Paper on Big Data and Hadoop. International Journal for Scientific Research & Development| Vol. 3, Issue 02.
[77] Vasuki M., Arthi J., Kayalvizhi K. (2014) Decision Making Using Sentiment Analysis from Twitter International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 12.
[78] Weichselbraun A., Gindl S., Scharl A. (2014) Enriching semantic knowledge bases for opinion mining in big data applications. Knowledge-Based Systems 69, 78–85.
[79] Weiss S.H and Indukhya N. (1998) Predictive Data mining: A Practical Guide, Morgan Kaufmann Publishers, and San Francisco, CA, 1998.
[80] XinhuiTian, Rui Han, Lei Wang, Gang Lu, Jianfeng Zhan (2015) Latency critical big data computing in finance. The Journal of Finanace and Data Science 1, 33-41.
[81] Yang Ming, Kiang Melody, Shang Wei (2015) Filtering big data from social media – Building an early warning system for adverse drug reactions. Journal of Biomedical Informatics 54, 230–240.
[82] Zhen-Yu Chen, Zhi-Ping Fan and Minghe Sun (2015) Behaviour- aware user response modelling in social media: Learning from diverse heterogeneous data. European Journal of Operation Research 241, 422-434.
[83] Zhong Chen, Huang Xianfeng, Arisona Stefan Muller, Schmitt Gerhard, Batty Michal (2014) Inferring building functions from a probabilistic model using public transportation data. Computers, Environment and Urban Systems 48,124-137.