Impact of Data Mining on Big Data Analytics: Challenges and Opportunities

Priyanka Gautam
Assistant Professor of Computer and Information Science ,CPJCHS, NARELA
Prof (Dr.) Y.P Singh
Director of Somany Institute of Technology ,Rewari, haryana

Abstract -*"big data" has become a highlighted buzzword since last year; "big data mining" has almost immediately followed up as an emerging, interrelated research area. Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Data mining techniques are providing great aid in the area of Big Data analytics, since dealing with Big Data are big challenges for the applications. Big Data analytics is the ability of extracting useful information from such huge datasets. This paper presents different approaches used for Big Data Analysis by using Data Mining techniques. Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper provides an overview of big data mining and discusses the related challenges and the new opportunities. We hope our effort will help reshape the subject area of today's data mining technology toward solving tomorrow's bigger challenges emerging in accordance with big data. The basic objective of this paper is to explore the potential impact of big data challenges, open research issues, and various tools associated with it. As a result, this paper provides a platform to explore big data at numerous stages. Additionally, it opens a new horizon for researchers to develop the solution, based on the challenges and open research issues.*

Keywords: Data mining, big data, big data mining, Big Data Analytics

1. Introduction

The era of petabyte has come and almost gone, leaving us to confront the exabytes era now. Technology revolution has been facilitating millions of people by generating tremendous data via ever-increased use of a variety of digital devices and especially remote sensors that generate continuous streams of digital data, resulting in what has been called as "big data". It has been a confirmed phenomenon that enormous amounts of data have been being continually generated at □ miraculous and ever increasing scales. In 2010, Google estimated that every two days at that time the world generated as much data as the sum it generated up to 2003. Regardless of the very recent "Big Data Executive Survey 2013" by New Vantage Partners that states "It's about variety, not volume", many people would still believe the issue with big data is scale or *volume*. Big data sure involves a great *variety* of data forms: text, images, videos, sounds, and whatever that may come into the play, and their arbitrary combinations. Big data frequently comes in the form of streams of a variety of types. Time is an integral dimension of data streams, which implies that the data must be processed in a timely or real-time manner. Besides, the current major consumers of big data, corporate businesses, are especially interested in "a big data environment that can accelerate the time-to-answer critical business questions that demonstrate business values". The time dimension of bid data naturally leads to yet another key characteristic of big data – speed or *velocity*. The era of Big Data has arrived. Every day, 2.5 quintillion bytes of data are created and 90% of the data in the world today were produced within the past two years. Our capability for data generation has never been so powerful and enormous ever since the invention of the Information Technology in the early 19th century.

The theme of this paper is to provide a study on the issue of big data analytics and data mining, its challenges and the opportunities. We also point to a few research topics that are either promising or much needed for solving the big data and big data mining problems. In order to make our discussion logical and smooth, we will start with a review of some essential and relevant concepts, including data mining, big data, big data mining, and the some platforms related to big data and big data mining.

2. Data Mining

Data mining attempts to implement basic processes that facilitate the extraction of meaningful information and knowledge from unstructured data. Data mining extracts patterns, changes, associations and anomalies from large data sets. The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data.

The two "high-level" primary goals of data mining, in practice, are *prediction* and *description*.

1. **Prediction** involves using some variables or fields in the database to predict unknown or future values of other variables of interest.
2. **Description** focuses on finding human-interpretable patterns describing the data.

A) Steps in Data Mining: The following steps are usually followed in data mining. These steps are iterative, with the process moving backward whenever needed.

1. Develop an understanding of the application, relevant prior knowledge, and the end user's goals.
2. Create a target data set to be used for discovery.
3. Clean and pre-process data.
4. Reduce the number of variables and find invariant representations of data if possible.
5. Choose the data mining task (classification, regression, clustering, etc.)
6. Choose the data mining algorithm.
7. Search for patterns of interest.
8. Interpret the pattern mined.
9. Consolidate knowledge discovered and prepare a report.

B) Data Mining Process: Data Mining is an iterative process that uses a variety of data analysis tools to discover patterns and relationships in data. Data mining is an interactive and iterative process involving data pre-processing], search for patterns, knowledge evaluation, and possible refinement of the process based on input from domain experts or feedback from one of the steps. The pre-processing of the data is a time-consuming, but critical, first steps in the data mining process. It is often domain and application dependent; however, several techniques developed in the context of one application or domain can be applied to other applications and domains as well.

3. Big Data

We are living in an interesting era – the era of big data, full of challenges and opportunities. Organizations have already started to deal with petabyte-scale collections of data; and they are about to face the exabyte scale of big data and the accompanying benefits and challenges. Big data is playing a crucial role in the future in all things of our lives and our societies. For example, governments have now started mining the data of social media networks and blogs, and online-transactions and other sources of information to recognize the need for government facilities, the suspicious organizational groups, and to predict future events. Even, service providers start to track their customers' purchases made through online, instore, and on-phone, and customers' behaviors through recorded streams of online clicks, as well as product reviews and ranking, for improving their marketing efforts, predicting new growth points of profits, and increasing customer satisfaction. The mismatch between the demands of the big data management and the capabilities that current DBMSs can provide has reached the historically high peak.

The three Vs (volume, variety, and velocity) of big data each implies one distinct aspect of critical deficiencies of today's DBMSs. Gigantic volume requires equally great scalability and massive parallelism that are beyond the capability of today's DBMSs; the great variety of data types of big data particularly unfits the restriction of the closed processing architecture of current database systems and the speed/velocity request of big data processing asks for commensurate real-time efficiency which again is far beyond where current DBMSs could reach. The limited availability of current DBMSs defeats the velocity request of big data from yet another angle.

To overcome this scalability challenge of big data, several attempts have been made on exploiting massive parallel processing architectures. The first such attempt was made by Google. Google created a programming model named MapReduce that was coupled with the GFS (Google File System) , a distributed file system where the data can be easily partitioned over thousands of nodes in a cluster. Later, Yahoo and other big companies created an Apache open-source version of Google's MapReduce framework, called Hadoop MapReduce.

- *Variety* makes big data really big. Big data comes from a great variety of sources and generally has in three types: structured, semi structured and unstructured. Structured data inserts a data warehouse already tagged and easily sorted but unstructured data is random and difficult to analyze. Semi-structured data does not conform to fixed fields but contains tags to separate data elements.

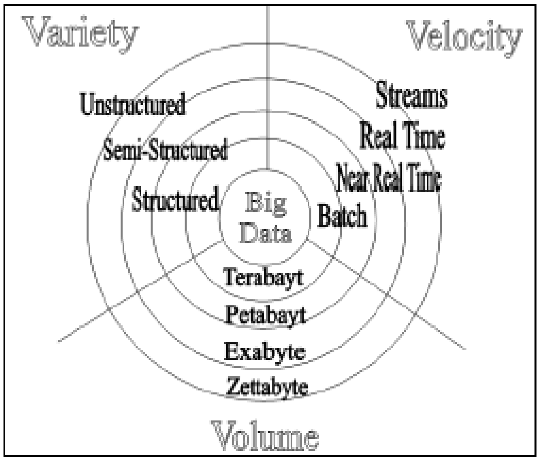

Figure 1.The three Vs of big data

- *Volume* or the size of data now is larger than terabytes and petabytes. The grand scale and rise of data outstrips traditional store and analysis techniques.
- *Velocity* is required not only for big data, but also all processes. For time limited processes, big data should be used as it streams into the organization in order to maximize its value.

During in the intensity of this information, another component is the verification of data flow. It is difficult to control large data so data security must be provided. In addition, after producing and processing of big data, it should create a plus value for the organization and Industry.

4. Data Mining & Big Data

Big data and data mining are two different things. Both of them relate to the use of large data sets to handle the collection or reporting of data that serves businesses or other recipients. However, the two terms are used for two different elements of this kind of operation. Big data is a term for a large data set. Big data sets are those that outgrow the simple kind of database and data handling architectures that were used in earlier times, when big data was more expensive and less feasible. For example, sets of data that are too large to be easily handled in a Microsoft Excel spreadsheet could be referred to as big data sets. Data mining refers to the activity of going through big data sets to look for relevant or pertinent information. This type of activity is really a good example of the old axiom "looking for a needle in a haystack." The idea is that businesses collect massive sets of data that may be homogeneous or automatically collected. Decision-makers need access to smaller, more specific pieces of data from those large sets. They use data mining to uncover the pieces of information that will inform leadership and help chart the course for a business.

5. Big Data Analytics via Data Mining

Data analytics refers to the Business Intelligence & Analytics technologies that are grounded mostly in data mining and statistical analysis. Most of these techniques rely on the mature commercial technologies of relational DBMS, data warehousing, ETL, OLAP, and BPM .Since the late 1980s, various data mining algorithms have been developed by researchers from the artificial intelligence, algorithm, and database communities. In the IEEE 2006 International Conference on Data Mining (ICDM), the 10 most influential data mining algorithms were identified based on expert nominations, citation counts, and a community survey. In ranked order, they are C4.5, k-means, Support Vector Machine, Apriori, EM (expectation maximization), PageRank, AdaBoost, Naïve Bayes, and CART Algorithms. These algorithms cover classification, clustering, regression, association analysis, and network analysis. Most of these popular data mining algorithms have been incorporated in commercial and open source data mining systems. Other advances such as neural networks for classification/prediction and clustering and genetic algorithms for optimization and machine learning have all contributed to the success of data mining in different applications.

Two other data analytics approaches commonly taught in business school, that are use in statistical theories and models, multivariate statistical analysis covers analytical techniques such as regression, factor analysis, clustering, and discriminant analysis that have been used successfully in various business applications. Developed in the management science community, optimization techniques and heuristic search are also suitable Most of these techniques can be found in business school curricula.

Due to the success achieved collectively by the data mining and statistical analysis community, data analytics continues to be an active area of research. Statistical machine learning often based on well-grounded mathematical models and powerful algorithms, techniques such as Bayesian networks, Hidden Markov models, support vector machine, reinforcement learning, and ensemble models, have been applied to data, text, and web analytics applications. Other new data analytics techniques explore and leverage unique data characteristics, from sequential/temporal mining and spatial mining, to data mining for high-speed data streams and sensor data. Increased privacy concerns in various e-commerce, e-government, and healthcare applications have caused privacy preserving data mining to become an emerging area of research. Many of these methods are data-driven, relying on various anonymization techniques, while others are process driven, defining how data can be accessed and used.

In addition to active academic research on data analytics, industry research and development has also generated much excitement, especially with respect to big data analytics for semi-structured content. Unlike the structured data that can be handled repeatedly through a RDBMS, semi-structured data may call for *ad hoc* and one-time extraction, parsing, processing, indexing, and analytics in a scalable and distributed MapReduce or Hadoop environment. MapReduce has been hailed as a revolutionary new platform for large scale, massively parallel data access. Inspired in part by MapReduce, Hadoop provides a Java based software framework for distributed processing of data intensive transformation and analytics. The top three commercial database suppliers—Oracle, IBM, and Microsoft— have all adopted Hadoop.

Big Data Analytics is the application that enables organizations to analyze large sets of data to discover patters and other useful information with the help of data mining tools. Due to the significant contribution of Big Data Analytics, the amount of data was exponentially increased within the past decade i. e 2005-15.

The technological advances in storage, processing, and analysis of Big Data include:

- The rapidly decreasing cost of storage and CPU power in recent years.
- The flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage.
- The development of new frameworks such as Hadoop, which allow users to take advantage of these, distributed computing systems storing large quantities of data through flexible parallel processing.

In this section, we focus on the data mining Process model for Big Data analytics. The overall model is divided into 2 Sub-Processes: Data Management and Analytics, which further broken down into 5 stages.

Big Data Phases and Processes	
Data Management	**Analytics**
A. Requirements & recording B. Extraction, Cleaning & Annotation C. Integration, Aggregation& Presentation	D. Modeling & Analysis E. Interpretation

Table1:**Process for Extracting Information from Big Data Set**

In short, big data is the asset and data mining is the "handler" of that is used to provide beneficial results.

Therefore, Data mining correlate with discovering useful models in massive data sets by itself, machine learning combine with data mining and statistical methods enabling machines to understand datasets. With the help of the different data mining techniques and tools, we can easily extract required information from the big data set. Data mining can involve the use of different kinds of software packages such as analytics tools. It can be automated, or it can be largely labor-intensive, where individual workers send specific queries for information to an archive or database. Generally, data mining refers to operations that involve relatively sophisticated search operations that return targeted and specific results. For example, a data mining tool may look through dozens of years of accounting information to find a specific column of expenses or accounts receivable for a specific operating year. So with the help of data mining techniques, it can make an easy process to extract the useful and relative knowledge and information from a big data set. Conclusionly, we can also says that the analysis of big data become an easier process with the help of the different tools and techniques of data mining.

6. Research Challenges & Opportunities in Data mining and Big data Analysis:

Recent year's big data has been acquired in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents, and internet search indexing. Social computing includes social network analysis, online communities, recommender systems, reputation systems, and prediction markets where as internet search indexing includes ISI, IEEE Xplorer, Scopus, etc. Considering this advantages of big data it provides a new opportunities in the knowledge processing tasks for the upcoming researchers. However opportunities always follow some challenges.

To handle the challenges we need to know various computational complexities, information security, and computational method, to analyze big data. For example, many statistical methods that perform well for small data size do not scale to voluminous data. Similarly, many computational techniques that perform well for small data face significant challenges in analyzing big data. Various challenges that the health sector face was being researched by many researchers.

The challenges of big data analytics and data mining are classified into the different broad categories namely

6.1 Mining data streams in extremely large database One important problem is mining data streams in extremely large databases. Satellite and computer network data [3] can easily be of this scale. However, today's data mining technology is still too slow to handle data of this scale. In addition, data mining should be a continuous, online process, rather than an occasional one-shot process. Organizations that can do this will have a decisive advantage over ones that do not. Data streams present a new challenge for data mining researchers.

6.2 Mining complex knowledge from complex data One important type of complex knowledge is in the form of graphs. Recent research has touched on the topic of discovering graphs and structured patterns from large data, but clearly, more needs to be done. Another form of complexity is from data that are non independent and identically distributed. This problem can occur when mining data from multiple relations. In most domains, the objects of interest are not independent of each other, and are not of a single type. We need data mining systems that can soundly mine the rich structure of relations among objects.

6.3 Privacy preserving data mining Privacy preserving data management is an important emerging research area that emerged in response to two important needs: data analysis and ensuring the privacy of the data owners. Privacy preserving data publishing emphasizes the importance of need for privacy threats in data sharing. A new approach seeks to protect data without focusing on the infrastructure level, but at element or aggregate data type. This type of pervasive security can be achieved by classifying data and enforcing access control.

6.4 Hetrogeinty and Incompleteness In the past, data mining techniques have been used to discover unknown patterns and relationships of interest from structured, homogeneous, and small datasets. Variety, as one of the essential characteristics of big data, is resulted from the phenomenon that there exists nearly unlimited different sources that generate or contribute to big data. This phenomenon naturally leads to the great variety or heterogeneity of big data. The data from different sources inherently possesses a great many different types and representation forms, and is greatly interconnected, interrelated, and delicately and inconsistently represented. Mining such a dataset, the great challenge is perceivable and the degree of complexity is not even imaginable before we deeply get there. Heterogeneity in big data also means that it is an obligation (rather than an option) to accept and deal with structured, semi-structured, and even entirely unstructured data simultaneously. This is especially so in data-intensive, scientific computation areas .Nevertheless, though bringing up greater technical challenges, the heterogeneity feature of big data means a new opportunity of unveiling, previously impossible, hidden patterns or knowledge dwelt at the intersections within heterogeneous big data.

As a classic data mining example, we consider a simple grocery transaction dataset that records only one type of data, i.e., goods items. Examples insights that might be mined from this dataset may include, e.g., the famous association of "beer and diapers" showing a strong linkage between the two items, and popular items like milk that are almost always purchased by customers, showing strong linkage of milk to all other items. In contrast to that, big data mining must deal with semi-structured and heterogeneous data. Now we generalize the aforementioned simple example by extending the scenario to an online market such as eBay. The dataset now is a richer network consisting of at least three different types of objects: items, buyers, and sellers. Interrelation may broadly exist, e.g., between commodity items in the form of "bought with", between sellers and items in the form of "sell" and "sold by", between buyers and items in the form of "buy" or "bought by", and between buyers and sellers in the form of "buy from" and "sold to". This data network has different types of objects and relationships. We speculate that existing data mining techniques would not maximally uncover the hidden associations and insights in this data network.

For a heterogeneous set of big data, trying to construct a single model would most likely not result in good-enough mining results; thus constructing specialized, more complex, multi-model systems is expected . An interesting algorithm following this spirit is proposed in that first determines whether the given dataset is truly heterogeneous, and if so, it then partitions the set into homogeneous subsets and constructs a specialized model for each homogeneous subset. Partitioning, as an intuitive approach, would speed up the process of knowledge discovery from heterogeneous big data. However, potential patterns and knowledge may miss the opportunity of being discovered after partitioning if important relationships crossing distinct homogeneous regions are not adequately retained. The social community mining problem has recently received a lot attention from the researchers. This problem desires "multi-network, user-dependent, and query based analysis". It conveys that the intersections between multiple networks bear potential knowledge and insights that may not be discovered if a homogenous model is to be enforced.

However, the degree of the heterogeneity captured does not reflect the real degree of the inherent heterogeneity existing in the big data. Mining hidden patterns from heterogeneous multimedia streams of diverse sources represents another frontier of data mining research. The output of this research has broad applicability such as detection of spreading dangerous diseases and prediction of traffic patterns and other critical social events.

6.5 Scalability The unprecedented volume/scale of big data requires commensurately high scalability of its data management and mining tools. Instead of being timid, we shall proclaim the extreme scale of big data because more data bears more potential insights and knowledge that we have no chance to discover from conventional data. We are optimistic with the following approaches that, if exploited properly, may lead to remarkable scalability required for future data and mining systems to manage and mine the big data:

(1) Cloud computing that has already demonstrated admirable elasticity, which, combined with massively parallel computing architectures, bears the hope of realizing the needed scalability for dealing with the volume challenge of big data.

(2) Advanced user interaction support (either GUI- or language-based) that facilitates prompt and effective system-user interaction. Big data mining straightforwardly implies extremely time-consuming navigation in a gigantic search and prompt feedback/interference/guidance from users must be beneficially exploited to help make early decisions, adjust search/mining strategies on the fly, and narrow down to smaller but promising sub-spaces.

6.6 Speed/Velocity For big data, speed/velocity really matters. The capability of fast accessing and mining big data is not just a subjective desire, it is an obligation especially for data streams– we must finish a processing/mining task within a certain period of time, otherwise, the processing/mining results becomes less valuable or even worthless. Exemplary applications with real-time requests include earthquake prediction, stock market prediction and agent-based autonomous exchange (buying/selling) systems. Speed is also relevant to scalability – conquering or partially solving anyone helps the other one.

The speed of data mining depends on two major factors: data access time and the efficiency of the mining algorithms themselves. Exploitation of advanced indexing schemes is the key to the speed issue. Multidimensional index structures are especially useful for big data. For example, a combination of R-Tree and KD-tree and the more recently proposed Fast Bit shall be considered for big data. Besides, design of new and more efficient indexing schemes is much desired, but remains one of the greatest challenges to the research community.

An additional approach to boost the speed of big data access and mining is through maximally identifying and exploiting the potential parallelism in the access and mining algorithms. The elasticity and parallelism support of cloud computing are the most promising facilities for boosting the performance and scalability of big data mining systems.

6.7 Timeliness As the size of the data sets to be processed increases, it will take more time to analyse. In some situations results of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed by preventing the transaction from taking place at all. Obviously a full analysis of a user's purchase history is not likely to be feasible in real time. So we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination. Given a large data set, it is often necessary to find elements in it that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. In such cases Index structures are created in advance to permit finding qualifying elements quickly. The problem is that each index structure is designed to support only some classes of criteria.

6.8 Security And privacy Challenges for Big Data Analysis Big data refers to collections of data sets with sizes outside the ability of commonly used software tools such as database management tools or traditional data processing applications to capture, manage, and analyze within an acceptable elapsed time. Big data sizes are constantly increasing, ranging from a few dozen terabytes in 2012 to today many petabytes of data in a single data set. Big data creates tremendous opportunity for the world economy both in the field of national security and also in areas ranging from marketing and credit risk analysis to medical research and urban planning. The extraordinary benefits of big data are lessened by concerns over privacy and data protection.

As big data expands the sources of data it can use, the trust worthiness of each data source needs
to be verified and techniques should be explored in order to identify maliciously inserted data.

Information security is becoming a big data analytics problem where massive amount of data will be correlated, analyzed and mined for meaningful patterns. Any security control used for big data must meet the following requirements:

- It must not compromise the basic functionality of the cluster.
- It should scale in the same manner as the cluster.
- It should not compromise essential big data characteristics.
- It should address a security threat to big data environments or data stored within the cluster.

Unauthorized release of information, unauthorized modification of information and denial of
resources are the three categories of security violation.
The following are some of the security
threats:

- An unauthorized user may access files and could execute arbitrary code or carry out further attacks.
- An unauthorized user may eavesdrop/sniff to data packets being sent to client.
- An unauthorized client may read/write a data block of a file.
- An unauthorized client may gain access privileges and may submit a job to a queue or delete or change priority of the job.

The following are some of the methods used for protecting big data:

Using authentication methods: Authentication is the process verifying user or system identity before accessing the system. Authentication methods such as Kerberos can be employed for this.

Use file encryption: Encryption ensures confidentiality and privacy of user information, and it Secures the sensitive data. Encryption protects data if malicious users or administrators gain access to data and directly inspect files, and renders stolen files or copied disk images unreadable.File layer encryption provides consistent protection across different platforms regardless of OS/platform type. Encryption meets our requirements for big data security. Open source products are available for most Linux systems; commercial products additionally offer external key management, and full support. This is a cost effective way to deal with several data security threats.

Implementing access controls: Authorization is a process of specifying access control privileges for user or system to enhance security.

Use key management: File layer encryption is not effective if an attacker can access encryption keys. Many big data cluster administrators store keys on local disk drives because it's quick and
easy, but it's also insecure as keys can be collected by the platform administrator.Use key management service to distribute keys and certificates and manage different keys for each group, application, and user.

Logging: To detect attacks, diagnose failures, or investigate unusual behavior, we need a record of activity. Unlike less scalable data management platforms, big data is a natural fit for collecting and managing event data. Many web companies start with big data particularly to manage log files. It gives us a place to look when something fails, or if someone thinks you might have been hacked. So to meet the security requirements, we need to audit the entire system on a periodic basis.

Use secure communication: Implement secure communication between nodes and between nodes and applications. This requires an SSL/TLS implementation that actually protects all network communications rather than just a subset. Thus the privacy of data is a huge concern in the context of Big Data.

7. Conclusion: We are living in the big data era where enormous amounts of heterogeneous, semistructured and unstructured data are continually generated at unprecedented scale. Big data discloses the limitations of existing data mining techniques, resulted in a series of new challenges related to big data mining. Big data mining is a promising research area. In spite of the limited work done on big data mining so far, we believe that much work is required to overcome its challenges related to heterogeneity, scalability, speed, accuracy, trust, privacy. Big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently occurring patterns and hidden rules. Big data analysis helps companies to take better decisions, to predict and identify changes and to identify new opportunities. In this paper we discussed about the opportunities and challenges related to big data mining and also Big Data analysis tools like Map Reduce over Hadoop which helps organizations to better understand their customers and the marketplace and to take better decisions and also helps researchers and scientists to extract useful knowledge out of Big data. In addition to that we introduce some big data mining tools and how to extract a significant knowledge from the Big Data set, which will help the research scholars to choose the best mining tool for their research work and for big data analytics. Our future work would primarily focuses on the Big Data analytics approach discussed above using various data mining techniques.

REFERENCES

[1] Puneet Singh Duggal and Sanchita Paul, Big Data Analysis: Challenges and Solutions.

[2] Han Hu, Yongyang Nen, Tat Seng Chua, Xuelong Li, Towards Scalable System for Big Data Analytics: A Technology Tutorial, IEEE Access, Volume 2, Page No 653, June 2014.

[3] Wei Fan and Albert Bifet, Mining Big Data: Current Status, and Forecast to the Future, SIGKDD Explorations, Volume 14, Issue 2, 2012.

[4] S.Vikram Phaneendra and E.Madhusudhan Reddy, Big Data- solutions for RDBMS problems- A survey, IEEE/IFIP Network Operations & Management Symposium (NOMS 2010),Osaka Japan, Apr 19-23 2013.

[5] Hardeep Kaur, A Review of Applications of Data Mining in the Field of Education, IJARCCE, Vol. 4, Issue 4, April 2015.

[6] Kishor, D., Big Data: The New Challenges in Data Mining, IJIRCST, 1(2), pp. 39-42, 2013.

[7] Dheeraj Agarwal, A comprehensive study of data mining and applications, IJARCET, Vol , issue 1, January 2013.

[8] Sagiroglu, S. and Sinanc, D., Big Data: A Review, International Conference on Collaboration Technologies and Systems (CTS), pp.42-47, 20-24, May 2013.

[9] Richa Gupta, Sunny Gupta and Anuradha Singhal, Big Data: Overview, IJCTT, Vol 9, Number 5, March 2014.

[10] Xindong Wu, Gong-Quing Wu and Wei Ding, Data Mining with Big Data, Jan 2014.

[11] Bharti Thakur and Manish Mann, Data Mining for Big Data: A Review, IJARCSSE, Volume 4, Issue 5, May 2014.

[12] Gantz, J., & Reinsel, D. (2011). The 2011 Digital Universe Study: Extracting Value from Chaos.

[13] Sabia and Sheetal Kalra, Application of Big Data: Current Status and Future Scope, IJACTE, Vol 3, Issue 5, 2014.

[14] James Manyika, Michael Chui, Brad Brown, Jacques Bhuhin, Richard Dobbs, Charles Roxburgh and Angela Hungh Byers, Big Data: The next frontier for innovation, competition and productivity, June 2011.

[15] Sagiroglu, S and Sinanc D., Big Data: A Review, International Conference on Collaboration Technologies and Systems (CTS), pp.42-47, 20-24 May 2013

[16] Ankita S. Tiwarkhede and Vinit Kakde, A Review Paper on Big Data Analytics, IJSR, Volume 4 Issue 4, April 2015 Article ID 431047, March 2015.