ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Big Data:Data Mining Challenges and Related Research

Mrs. Seema singh solanki
Computer Science & Engineering Deparment,
Everest Educational Society's Group Of Institutios
College Of Engineering & Technology,
Aurangabad.

Syed Abed Ali Hashmi
ME - I (CSE) ,
Computer Science & Engineering Department,
Everest Educational Society's Group Of Institutions
College Of Engineering & Technology, Aurangabad.

Abstract:

Big Data is a new term used to identify the datasets that due to their large size and complexity, we cannot manage them with our current methodologies or data mining software tools. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities. The work for this paper research and seminar is to present details about big data sources, its types, characteristics and data mining challenges with proposed solutions.

It involves introducing and literature survey for at least 3-4 articles from influential scientists and related papers published in this field covering most interesting and state of the art topics for BIG DATA and its analytics challenges.

Introduction:

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences.

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of Infra Stress". The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad in his invited talk at the KDD Big Mine 12Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 million tweets per day, face book 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 the order of zettabytes, and it is growing around 40% every year. Big data has been used to convey all sorts of concepts, including: huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more.

Whatever the label, organizations are starting to understand and explore how to process and analyze a vast array of information in new ways. The purpose of this paper is to provide an in-depth study on data mining challenges in big data with new research projects and areas for perceivable solutions and opportunities.

1. BIG Data:

1.1. BIG Data Types:

There are two types of big data: structured and unstructured.

Structured data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smart phones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data.

Unstructured data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data cannot easily be separated into categories or analyzed numerically. "Unstructured big data is the things that humans are saying," says big data consulting firm vice president Tony Jewitt of Plano, Texas.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

"It uses natural language."Analysis of unstructured data relies on keywords, which allow users to filter the data based on searchable terms. The explosive growth of the Internet in recent years means that the variety and amount of big data continue to grow. Much of that growth comes from unstructured data.

1.2. HACE theorem and Three V's of Big Data:

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. Indeed, the term Big Data literally concerns about data volumes, HACE theorem suggests that the key characteristics of the Big Data are:

A. Huge with heterogeneous and diverse data sources:

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This huge volume of data comes from various sites like Twitter, MySpace, Orkut and LinkedIn etc.

B. Decentralized control:

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function without necessarily relying on other servers.

C. Complex data and knowledge associations:

Multi structure, multisource data is complex data, Examples of complex data types are bills of materials, word processing documents, maps, time-series, images and video. Such combined characteristics suggest that Big Data require a "big mind" to consolidate data for maximum values. Apart from these characteristics, big data is generally explained and understood with help of 3 V's, Doug Laney was the first one talking about 3V"s in Big Data Management.

Volume:

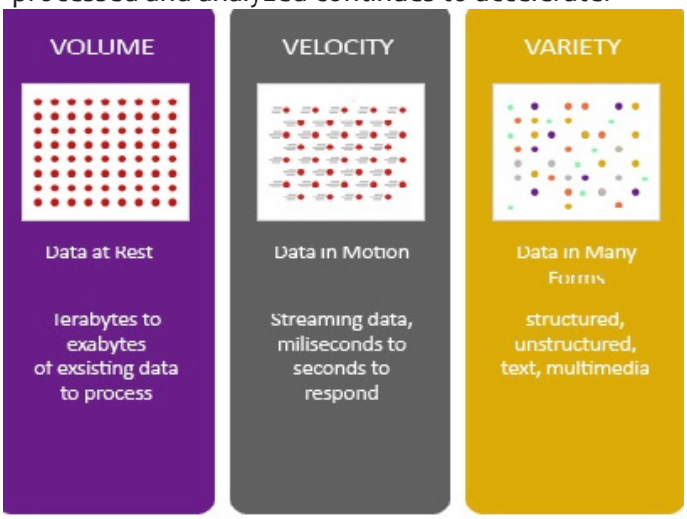
The amount of data. Perhaps the characteristic most associated with big data, volume refers to the mass quantities of data that organizations are trying to harness to improve decision-making across the enterprise. Data volumes continue to increase at an unprecedented rate.

Variety:

Different types of data and data sources. Variety is about managing the complexity of multiple data types, including structured, semi-structured and unstructured data. Organizations need to integrate and analyze data from a complex array of both traditional and non-traditional information sources, from within and outside the enterprise. With the explosion of sensors, smart devices and social collaboration technologies, data is being generated in countless forms, including: text, web data, tweets, audio, video, log files and more.

Velocity:

Data in motion. The speed at which data is created, processed and analyzed continues to accelerate.



Nowadays there are two more V"s.

Variability:

There are changes in the structure of the data and how users want to interpret that data.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Value: Business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach.

2. Big Data Mining:

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational database. The goals of big data mining techniques go beyond fetching the requested information or even uncovering some hidden relationships and patterns between numeral parameters.

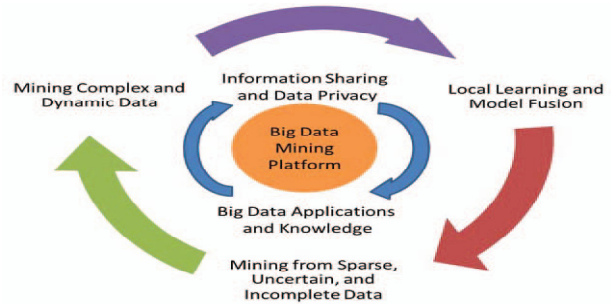
Analyzing fast and massive stream data may lead to new valuable insights and theoretical concepts [2]. Comparing with the results derived from mining the conventional datasets, unveiling the huge volume of interconnected heterogeneous big data has the potential to maximize our knowledge and insights in the target domain.

However, this brings a series of new challenges to the research community. In following sections we will discuss about the challenges and solutions for big data mining. It must deal with heterogeneity, extreme scale, velocity, privacy, accuracy, trust, and interactiveness that existing mining techniques and algorithms are incapable of.

The big data processing in a conceptual view can be described with three dimensions inside out as, data mining platform (access and computing), privacy, semantics and knowledge of the problem space or data space to be mined and lastly the algorithms and techniques involved. These are described in subsequent sections.

3. Challenges in BIG DATA analytics :

Meeting the challenges presented by big data will be difficult. The volume of data is already enormous and increasing every day. The velocity of its generation and growth is increasing, driven in part by the proliferation of internet connected devices.



Big data	Data mining
Big data is a term for large data set	Data mining refers to the activity of going through big data set to look for relevant information
Big data is the asset	Data mining is the handler which provide beneficial
Big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data	Data mining refers to the operation that involve relatively sophisticated search operation

Furthermore, the variety of data being generated is also expanding, and organization's capability to capture and process this data is limited. Current technology, architecture, management and analysis approaches are unable to cope with the flood of data, and organizations will need to change the way they think about, plan, govern, manage, process and report on data to realize the potential of big data. Following are the Challenges and proposed solutions for data mining.

3.1 Data access and Computing Platform:

In typical data mining systems, the mining procedures require computational intensive computing units for data analysis and comparisons. A computing platform is, therefore, needed to have efficient access to, at least, two types of resources: data and computing processors. For small scale data mining tasks, a single desktop computer, which contains hard disk and CPU processors, is sufficient to fulfill the data mining goals.

Indeed, many data mining algorithm are designed for this type of problem settings. For medium scale data mining tasks, data are typically large (and possibly distributed) and cannot be fit into the main memory. Common solutions are to rely on parallel computing [43], [33] or collective mining [12] to sample and aggregate data from different sources and then use parallel computing programming (such as the Message Passing Interface) to carry out the mining process.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

For Big Data mining, because data scale is far beyond the capacity that a single personal computer (PC) can handle, a typical Big Data processing framework will rely on cluster computers with a high-performance computing platform, with a data mining task being deployed by running some parallel programming tools, such as Map-Reduce or Enterprise Control Language (ECL), on a large number of computing nodes (i.e., clusters).

The role of the software component is to make sure that a single data mining task, such as finding the best match of a query from a database with billions of records, is split into many small tasks each of which is running on one or multiple computing nodes.

For example, as of this writing, the world most powerful super computer Titan, which is deployed at Oak Ridge National Laboratory in Tennessee, contains 18,688 nodes each with a16-core CPU. Such a Big Data system, which blends both hardware and software components, is hardly available without key industrial stockholders' support.

In fact, for decades, companies have been making business decisions based on transactional data stored in relational databases. Big Data mining offers opportunities to go beyond traditional relational databases to rely on less structured data: weblogs, social media, e-mail, sensors, and photographs that can be mined for useful information.

Major business intelligence companies, such IBM, Oracle, Teradata, and so on, have all featured their own products to help customers acquire and organize these diverse data sources and coordinate with customers' existing data to find new insights and capitalize on hidden relationships.

3.2 Challenges with Semantics and Application Knowledge:

Semantics and application knowledge in Big Data refer to numerous aspects related to the regulations, policies, user knowledge, and domain information. The two most important issues at this tier include 1) data sharing and privacy; and 2) domain and application knowledge.

The former provides answers to resolve concerns on how data are maintained, accessed, and shared; whereas the latter focuses on answering questions like "what are the underlying applications ?" and "what are the knowledge or patterns users intend to discover from the data ?"

3.3 Algorithms:

3.3.1 Local Learning and Model Fusion for Multiple Information Sources:

As Big Data applications are featured with autonomous sources and decentralized controls, aggregating distributed data sources to a centralized site for mining is systematically Prohibitive due to the potential transmission cost and privacy concerns. On the other hand, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each site often leads to biased decisions or models, just like the elephant and blind men case.

Under such a circumstance, a Big Data mining system has to enable an information exchange and fusion mechanism to ensure that all distributed sites (or information sources) can work together to achieve a global optimization goal. Model mining and correlations are the key steps to ensure that models or patterns discovered from multiple information sources can be consolidated to meet the global mining objective.

More specifically, the global mining can be featured with a two-step (local mining and global correlation) process, at data, model, and at knowledge levels. At the data level, each local site can calculate the data statistics based on the local data sources and exchange the statistics between sites to achieve a global data distribution view. At the model or pattern level, each site can carry out local mining activities, with respect to the localized data, to discover local patterns.

By exchanging patterns between multiple sources, new global patterns can be synthetized by aggregating patterns across all sites [50]. At the knowledge level, model correlation analysis investigates the relevance between models generated from different data sources to determine how relevant the data sources are correlated with each other, and how to form accurate decisions based on models built from autonomous sources.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

3.3.2 Mining from Sparse, Uncertain, and Incomplete Data:

Spare, uncertain, and incomplete data are defining features for Big Data applications. Being sparse, the number of data points is too few for drawing reliable conclusions. This is normally a complication of the data dimensionality issues, where data in a high-dimensional space (such as more than 1,000 dimensions) do not show clear trends or distributions.

For most machine learning and data mining algorithms, high-dimensional spare data significantly deteriorate the reliability of the models derived from the data. Common approaches are to employ dimension reduction or feature selection [48] to reduce the data dimensions or to carefully include additional samples to alleviate the data scarcity, such as generic unsupervised learning methods in data mining.

Uncertain data are a special type of data reality where each data field is no longer deterministic but is subject to some random/error distributions. This is mainly linked to domain specific applications with inaccurate data readings and collections. For example, data produced from GPS equipment are inherently uncertain, mainly because the technology barrier of the device limits the precision of the data to certain levels (such as 1 meter). As a result, each recording location is represented by a mean value plus a variance to indicate expected errors. For data privacyrelated applications [36], users may intentionally inject randomness/errors into the data to remain anonymous.

This is similar to the situation that an individual may not feel comfortable to let you know his/her exact income, but will be fine to provide a rough range like [120k, 160k]. For uncertain data, the major challenge is that each data item is represented as sample distributions but not as a single value,so most existing data mining algorithms cannot be directly applied. Common solutions are to take the data distributions into consideration to estimate model parameters. For example, error aware data mining [49] utilizes the mean and the variance values with respect to each single data item to build a Nai¨ve Bayes model for classification. Similar approaches have also been applied for decision trees or database queries.

Incomplete data refer to the missing of data field values for some samples. The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values (e.g., dropping some sensor node readings to save power for transmission).

While most modern data mining algorithms have inbuilt solutions to handle missing values (such as ignoring data fields with missing values), data imputation is an established research field that seeks to impute missing values to produce improved models (compared to the ones built from the original data). Many imputation methods [20] exist for this purpose, and the major approaches are to fill most frequently observed values or to build learning models to predict possible values for each data field, based on the observed values of a given instance.

3.3.3 Mining Complex and Dynamic Data:

The rise of Big Data is driven by the rapid increasing of complex data and their changes in volumes and in nature [6]. Documents posted on WWW servers, Internet backbones, social networks, communication networks, and transportation networks, and so on are all featured with complex data. While complex dependency structures underneath the data raise the difficulty for our learning systems, they also offer exciting opportunities that simple data representations are incapable of achieving.

For example, researchers have successfully used Twitter, a well-known social networking site, to detect events such as earthquakes and major social activities, with nearly real time speed and very high accuracy.In addition, by summarizing the queries users submitted to the search engines, which are all over the world, it is now possible to build an early warning system for detecting fast spreading flu outbreaks [23].

Making use of complex data is a major challenge for Big Data applications, because any two parties in a complex network are potentially interested to each other with a social connection. Such a connection is quadratic with respect to the number of nodes in the network, so a million node network may be subject to one trillion connections.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

For a large social network site, like Facebook, the number of active users has already reached 1billion, and analyzing such an enormous network is a big challenge for Big Data mining. If we take daily user actions/interactions into consideration, the scale of difficulty will be even more astonishing.

4.Related work and initiatives for challenges:

4.1 Computing Platforms:

Due to the multisource, massive, heterogeneous, and dynamic characteristics of application data involved in a distributed environment, one of the most important characteristics of Big Data is to carry out computing on the petabyte (PB), even the exabyte (EB)-level data with a complex computing process.

Therefore, utilizing a parallel computing infrastructure, its corresponding programming language support, and software models to efficiently analyze and mine the distributed data are the critical goals for Big Data processing to change from "quantity" to "quality."

Currently, Big Data processing mainly depends on parallel programming models like MapReduce, as well as providing a cloud computing platform of Big Data services for the public. MapReduce is a batch-oriented parallel computing model.

There is still a certain gap in performance with relational databases. Improving the performance of MapReduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with MapReduce parallel programming being applied to many machine learning and data mining algorithms.

Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model parameters. It calls for intensive computing to access the large-scale data frequently.

To improve the efficiency of algorithms, Chu et al. proposed a general-purpose parallel programming method, which is applicable to a large number of machine learning algorithms based on the simple MapReduce programming model on multi core processors.

4.2 Data Privacy Semantics and Application Knowledge:

In privacy protection of massive data, Ye et al. proposed a multilayer rough set model, which can accurately describe the granularity change produced by different levels of generalization and provide a theoretical foundation for measuring the data effectiveness criteria in the anonymization process, and designed a dynamic mechanism for balancing privacy and data utility, to solve the optimal generalization/refinement order for classification.

A recent paper on confidentiality protection in Big Data [4] summarizes a number of methods for protecting public release data, including aggregation (such as K-anonymity, I-diversity, etc.), suppression (i.e., deleting sensitive values), data swapping (i.e., switching values of sensitive data records to prevent users from matching), adding random noise, or simply replacing the whole original data values at a high risk of disclosure with values synthetically generated from simulated distributions.

For applications involving Big Data and tremendous data volumes, it is often the case that data are physically distributed at different locations, which means that users no longer physically possess the storage of their data. To carry out Big Data mining, having an efficient and effective data access mechanism is vital, especially for users who intend tohire a third party (such as data miners or data auditors) to process their data.

Under such a circumstance, users' privacy restrictions may include 1) no local data copies or downloading, 2) all analysis must be deployed based on the existing data storage systems without violating existing privacy settings, and many others. In Wang et al. [48], a privacy-preserving public auditing mechanism for large scale data storage (such as cloud computing systems) has been proposed.

The public key-based mechanism is used to enable third-party auditing (TPA), so users can safely allow a third party to analyze their data without breaching the security settings or compromising the data privacy. For most Big Data applications, privacy concerns focuson excluding the third party (such as data miners) from directly accessing the original data.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Common solutions are to rely on some privacy-preserving approaches or encryption mechanisms to protect the data. A recent effort by Lorch et al. [32] indicates that users' "data access patterns" can also have severe data privacy issues and lead to disclosures of geographically co-located users or users with common interests (e.g., two users searching for the same map locations are likely to be geographically co located). In their system, namely Shround, users' data access patterns from the servers are hidden by using virtual disks. As a result, it can support a variety of Big Data applications, such as micro blog search and social network queries, without compromising the user privacy.

4.3 Data Mining Algorithms:

To adapt to the multisource, massive, dynamic Big Data, researchers have expanded existing data mining methods in many ways, including the efficiency improvement of single-source knowledge discovery methods [11], designing a data mining mechanism from a multisource perspective [50], as well as the study of dynamic data mining methods and the analysis of stream data [18], [12]. The main motivation for discovering knowledge from massive data is improving the efficiency of single-source mining methods. On the basis of gradual improvement of computer hardware functions, researchers continue to explore ways to improve the efficiency of knowledge discovery algorithms to make them better for massive data.

Because massive data are typically collected from different data sources, the knowledge discovery of the massive data must be performed using a multisource mining mechanism. As real-world data often come as a data stream or a characteristic flow, a well-established mechanism is needed to discover knowledge and master the evolution of knowledge in the dynamic data source. Therefore, the massive, heterogeneous and real-time characteristics of multisource data provide essential differences between single-source knowledge discovery and multisource data mining.

Wu et al. [45] proposed and established the theory of local pattern analysis, which has laid a foundation for global knowledge discovery in multisource data mining. This theory provides a solution not only for the problem of full search, but also for finding global models that traditional mining methods cannot find.

Local pattern analysis of data processing can avoid putting different data sources together to carry out centralized computing. Data streams are widely used in financial analysis, online trading, and medical testing, and so on. Static knowledge discovery methods cannot adapt to the characteristics of dynamic data streams, such as continuity, variability, rapidity, and infinity, and can easily lead to the loss of useful information.

Therefore, effective theoretical and technical frameworks are needed to support data stream mining [18]. Knowledge evolution is a common phenomenon in real world systems. For example, the clinician's treatment programs will constantly adjust with the conditions of the patient, such as family economic status, health insurance, the course of treatment, treatment effects, and distribution of cardiovascular and other chronic epidemiological changes with the passage of time.

In the knowledge discovery process, concept drifting aims to analyze the phenomenon of implicit target concept changes or even fundamental changes triggered by dynamics and context in data streams. According to different types of concept drifts, knowledge evolution can take forms of mutation drift, progressive drift, and data distribution drift, based on single features, multiple features, and streaming features.

5. Conclusion:

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully.

The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation.

These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

REFERENCES:

[1] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolutionof Conserved Relational States in Dynamic Networks," Knowledgeand Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

[2] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.

[3] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.

[4] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[5] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

[6] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012.

[7] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[8] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.

[9] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSeyQuarterly, 2010.

[10] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.

[11] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multimedia, (MM '09,) pp. 917-918, 2009.

[12] R. Chen, K. Sivakumar, and H. Kargupta, "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data,"Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.

[13] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Efficient Algorithms for Influence Maximization in Social Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

[14] C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, and K. Olukotun, "Map-Reduce for Machine Learning on Multicore," Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS'06), pp. 281-288, 2006.

[15] G. Cormode and D. Srivastava, "Anonymized Data: Generation, Models, Usage," Proc. ACM SIGMOD Int'l Conf. Management Data, pp. 1015-1018, 2009.

[16] S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J.McPherson, "Ricardo: Integrating R and Hadoop," Proc. ACMSIGMOD Int'l Conf. Management Data (SIGMOD '10), pp. 987-998.2010.

[17] P. Dewdney, P. Hall, R. Schilizzi, and J. Lazio, "The Square Kilometre Array," Proc. IEEE, vol. 97, no. 8, pp. 1482-1496, Aug. 2009.

[18] P. Domingos and G. Hulten, "Mining High-Speed Data Streams," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '00), pp. 71-80, 2000.

[19] G. Duncan, "Privacy by Design," Science, vol. 317, pp. 1178-1179, 2007.

[20] B. Efron, "Missing Data, Imputation, and the Bootstrap," J. Am. Statistical Assoc., vol. 89, no. 426, pp. 463-475, 1994.

[21] A. Ghoting and E. Pednault, "Hadoop-ML: An Infrastructure for the Rapid Implementation of Parallel Reusable Analytics," Proc. Large-Scale Machine Learning: Parallelism and Massive Data Sets Workshop (NIPS '09), 2009.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

[22] D. Gillick, A. Faria, and J. DeNero, MapReduce: Distributed Computing for Machine Learning, Berkley, Dec. 2006.

[23] M. Helft, "Google Uses Searches to Track Flu's Spread," The New York Times, http://www.nytimes.com/2008/11/12/technology/internet/12flu.html. 2008. WU ET AL.: DATA MINING WITH BIG DATA 105.

[24] D. Howe et al., "Big Data: The Future of Biocuration," Nature, vol. 455, pp. 47-50, Sept. 2008.

[25] B. Huberman, "Sociology of Science: Big Data Deserve a Bigger Audience," Nature, vol. 482, p. 308, 2012.

[26] "IBM What Is Big Data: Bring Big Data to the Enterprise," http://www-01.ibm.com/software/data/bigdata/, IBM, 2012.

[27] A. Jacobs, "The Pathologies of Big Data," Comm. ACM, vol. 52, no. 8, pp. 36-44, 2009.

[28] I. Kopanas, N. Avouris, and S. Daskalaki, "The Role of Domain Knowledge in a Large Scale Data Mining Project," Proc. Second Hellenic Conf. AI: Methods and Applications of Artificial Intelligence, I.P. Vlahavas, C.D. Spyropoulos, eds., pp. 288-299, 2002.

[29] A. Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033, 2012.

[30] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.

[31] W. Liu and T. Wang, "Online Active Multi-Field Learning for Efficient Email Spam Filtering," Knowledge and Information Systems, vol. 33, no. 1, pp. 117-136, Oct. 2012.

[32] J. Lorch, B. Parno, J. Mickens, M. Raykova, and J. Schiffman, "Shoroud: Ensuring Private Access to Large-Scale Data in the Data Center," Proc. 11th USENIX Conf. File and Storage Technologies (FAST '13), 2013.

[33] D. Luo, C. Ding, and H. Huang, "Parallelization with Multiplicative Algorithms for Big Data Mining," Proc. IEEE 12th Int'l Conf. Data Mining, pp. 489-498, 2012.

[34] J. Mervis, "U.S. Science Policy: Agencies Rally to Tackle Big Data," Science, vol. 336, no. 6077, p. 22, 2012.

[35] F. Michel, "How Many Photos Are Uploaded to Flickr Every Day and Month?" http://www.flickr.com/photos/franckmichel/ 6855169886/, 2012.

[36] T. Mitchell, "Mining our Reality," Science, vol. 326, pp. 1644-1645, 2009.

[37] Nature Editorial, "Community Cleverness Required," Nature, vol. 455, no. 7209, p. 1, Sept. 2008.

[38] S. Papadimitriou and J. Sun, "Disco: Distributed Co-Clustering with Map-Reduce: A Case Study Towards Petabyte-Scale End-to- End Mining," Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08),pp. 512-521, 2008.

[39] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "Evaluating MapReduce for Multi-Core and Multiprocessor Systems," Proc. IEEE 13th Int'l Symp. High Performance Computer Architecture (HPCA '07), pp. 13-24, 2007.

[40] A. Rajaraman and J. Ullman, Mining of Massive Data Sets. Cambridge Univ. Press, 2011.

[41] C. Reed, D. Thompson, W. Majid, and K. Wagstaff, "Real Time Machine Learning to Find Fast Transient Radio Anomalies: A Semi-Supervised Approach Combining Detection and RFI Excision,"Proc. Int'l Astronomical Union Symp. Time Domain Astronomy, Sept. 2011.

[42] E. Schadt, "The Changing Privacy Landscape in the Era of Big Data," Molecular Systems, vol. 8, article 612, 2012.

[43] J. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining," Proc. 22nd VLDB Conf., 1996.

[44] A. da Silva, R. Chiky, and G. He´brail, "A Clustering Approach for Sampling Data Streams in Sensor Networks," Knowledge and Information Systems, vol. 32, no. 1, pp. 1-23, July 2012.

[45] K. Su, H. Huang, X. Wu, and S. Zhang, "A Logical Framework for Identifying Quality Knowledge from Different Data Sources," Decision Support Systems, vol. 42, no. 3, pp. 1673-1683, 2006.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

[46] "Twitter Blog, Dispatch from the Denver Debate," http://blog.twitter.com/2012/10/dispatch-from-denver-debate.html,Oct. 2012.

[47] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters," Proc. Int'l Conf. Data Mining Workshops (ICDMW '09), pp. 296-301, 2009.

[48] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy- Preserving Public Auditing for Secure Cloud Storage" IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[49] X. Wu and X. Zhu, "Mining with Noise Knowledge: Error-Aware Data Mining," IEEE Trans. Systems, Man and Cybernetics, Part A, vol. 38, no. 4, pp. 917-932, July 2008.

[50] X. Wu and S. Zhang, "Synthesizing High-Frequency Rules from Different Data Sources," IEEE Trans. Knowledge and Data Eng.,vol. 15, no. 2, pp. 353-367, Mar./Apr. 2003.