SURVEY ON BIG DATA MINING PLATFORMS, ALGORITHMS AND CHALLENGES

SHERIN A[1], Dr S UMA[2], SARANYA K[3], SARANYA VANI M[4]

[1] PG Scholar, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India
[2]Head of the Department, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India
[3]PG Scholar, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India
[4]PG Scholar, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India

Abstract-- "Big data" is coined to address massive volumes of data sets usually huge, sparse, incomplete, uncertain, complex or dynamic, which are mainly coming from multiple and autonomous sources. The primary sources for big data are from business applications, public web, social media, and sensor data and so on. "Big data mining" involves knowledge discovery from these large data sets. "Big data" is gaining huge significance in the current scenario and consequently, big data mining emerged as an innovative and potential research area. This paper gives an overview of big data along with its type, source and characteristics. A review on various big data mining platforms, algorithms and challenges is also discussed in this paper.

Keywords-- big data, big data mining platforms, big data mining algorithms, big data mining challenges, data mining

I. INTRODUCTION

Today we are living in an era of digital world. With the rapid increase in digitization the amount of structured, semi structured and unstructured data being generated and stored is exploding. Usama Fayyad [1] has presented amazing data numbers about internet usage like "every day 1 billion queries are there in Google, more than 250 million tweets are there in Twitter, more than 800 million updates are there in Face book, and more than 4 billion views are there in You tube". Each day, 2.5 quintillion bytes of data are generated and 90 percent of the data in the world today were created within the past two years [2]. The data produced nowadays is estimated in the order of zeta bytes, and it is growing around 40% every year. International Data Corporation (IDC) terms this as the "Digital Universe" and predicts that this digital universe is set to explode to an unimaginable 8 Zetabytes by the year 2015 [3]. The above examples demonstrate the rise of big data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to manage, capture, and process. The most fundamental challenge for big data applications is to explore the large volumes of data and extract useful information or knowledge for future actions [4]. Thus making big data mining or knowledge discovery of large datasets a difficult process.

II. BIG DATA

Big data technologies defines a new generation of technologies and architectures, designed solely to economically extract useful information's from very large volumes of a wide variety of data, by permitting high velocity capture, discovery, and analysis[5]. O'Reilly [6] defines big data is the data that exceeds the processing capacity of conventional database systems. He also explains that the data is very big, moves very fast, or doesn't fit into traditional database architectures. Further he has extended that to gain value from this data, one has to choose an alternative way to process it .There are different definitions of big data as it is more often used as an all-encompassing term for everything from actual data sets to big data technology and big data analytics. There are mainly 3 types of big data sets- structured, semi structured and unstructured [7]. In structured data, we can group the data to form a relational schema and represent it using rows and columns within a standard database. Based on an organization's parameters and operational needs, structured data responds to simple queries and provides usable information due to its configuration and consistency. Semi structured data [8] does not conform to an explicit and fixed schema. The data is inherently self-describing and contains tags or other markers to enforce hierarchies of records and fields within the data. Certain examples for semi-structured data include weblogs and social media feeds. The formats of unstructured data cannot be easily indexed into relational tables for analysis or querying. Certain examples for unstructured data's are image files, audio files, video files, and health records and so on.

As organizations grow the data concerned with them also expand exponentially. Most of the big organizations have data in multiple applications and in different formats. The data is also spread out in such a way that it is hard to categorize with a single algorithm or logic. Big organizations are in fact facing challenges

to keep all the data on a platform which give them a single consistent view of their data. This unique challenge to identify all the data coming in from multiple sources and to perform knowledge discovery out of it is the revolution big data world is facing. The 3Vs that define big data are Volume, Velocity and Variety [9].

A. Volume

Volume is the most important aspect that comes when dealing with big data. It refers to the huge amounts of data generated every second. In social media channels large amounts of data is seen in the form of images, videos, music's etc. It is very common to have Terabytes, Petabytes and Zetabytes of the storage system for organizations. As the database expands the applications and architecture built to support the data needs to be re-evaluated frequently. Sometimes the same data is re-evaluated with multiple angles and even though the original data is the same the new found intelligence creates explosion of the data. This big volume of data is big data.

B. Velocity

Velocity refers to the speed at which new data is generated and the speed at which data is moving around. The data growth and social media explosion have changed the vision of data. In the near future, definitely it is said that data's of yesterday is recent. Today, people rely on social media to update them with the latest happening. On social media perhaps a few seconds old messages (a tweet, status updates etc.) is not of interest for users. They often discard old messages and pay attention to the most recent updates. Nowadays it is said that data movement is done in real time and the update window has reduced to fractions of the seconds. Big data technology allows us to analyze these high speed data while it is being generated, without ever putting it into databases .This high velocity data is big data.

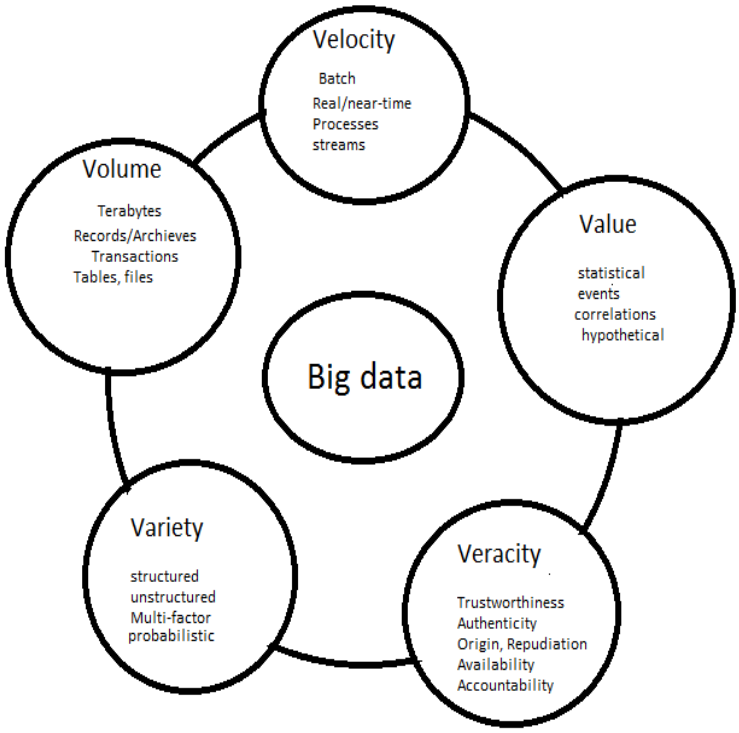

Figure-1 Five V's of big data

C. Variety

Variety refers to the different types of data that can make use of. Big data technology deals with structured, unstructured or semi-structured data's. Traditionally most data that are generated is structured and can be able to store in files or in database. Most of the time data is not in structured format, it may be unstructured in the form of, short message service (SMS), portable document format (pdf) files, social media conversations, and sensor data and so on. This variety of data is big data.

Now two more V's also contributed to big data. They are veracity and value of data [10].

D. Veracity

Veracity refers to cleanliness or trustworthiness of data. Many of the data lacks quality and accuracy. For example twitter posts with hash tags, abbreviations and so on. This veracity of data is big data.

E. Value

Value gives importance to the profit gained by organizations who invest in Big Data technologies

III. BIG DATA CHARACTERISTICS: HACE THEOREM

Big data is voluminous and has large-volume of data coming from heterogeneous, independent sources with dispersed and decentralized control, and attempts to find and explore complex and evolving relationships among them[11].These characteristics make it an extreme challenge for discovering useful knowledge from the big data. The notion of size in determining whether a data is considered as big data or not differs from person to person, since the source of data varies for each person. The big data consists of varying volumes of heterogeneous information aggregated from different resources. According to HACE theorem [11] the most important characteristics of big data are listed below

A. Heterogeneous

Huge heterogeneous diverse data source means data from any number of sources, largely unknown and infinite, and in multiple formats. This heterogeneous huge volume of data comes from various sites like Twitter, Facebook, Orkut and LinkedIn etc

B. *Autonomous*

Autonomous data sources with distributed and decentralized controls are one of the important characteristic of big data applications. Here each data source is able to create and collect information without involving any centralized control. For e.g., In World Wide Web (WWW) setting, each web server gives a certain amount of information and function individually without depending on others

C. *Complex*

As the volume of data increases, complexity of the data also increases. Multi-structured multisource data are highly complex. Example for these types of highly complex data's are bills of commodities, word processing documents, maps, time-series data, images and video files etc.

D. *Evolving*

The complex data which is non linear and many to many data relationships evolve. Big data often change over time. For instance, customer comments on a website. This type of data must be gathered over significant periods of time inorder to find out patterns and trends.

IV. BIG DATA SOURCES

The major sources of big data are from the following

A. Archives

Archives are mainly maintained by organizations, to show the function of a particular person or organization functions. Accumulation of archives sometimes does not fit into the traditional storage systems and need systems with high processing capabilities. This voluminous archive contributes to big data.

B. Media

Users generate images, videos, audios, live streams, podcasts and so on contributes for big data.

C. Business applications

Huge volumes of data are generated from business applications as part of project management, marketing automation, productivity, customer relation management (CRM), enterprise resource planning (ERP) content management systems, procurement, human resource (HR), storage, talent management, Google Docs, intranets, portals and so on. These data contributes to big data

D. Public web

Many organizations under government sector, weather, competitive, traffic regulatory compliance, health care services, economic, census, public finance, stock, open source intelligence (OSINT), the world bank, electronic data gathering analysis and retrieval (Edgar), Wikipedia and so on uses web services for communication. These data contributes to big data

E. Social Media

Nowadays users rely on social media sites such as twitter, linkedln, facebook, tumblr, blog, slideshare, youtube, google+, instagram, flickr, pinterest, vimeo, wordpress and so on for the creation and exchange of user generated contents. These social networking sites contribute to big data.

F. Data Storage

Data storage in SQL, NoSQL, Hadoop, doc repository, file systems and so on also contributes to big data.

G. Sensor Data

Accumulation of large quantitative datasets from distributed sensors are now becoming widely available online from medical devices, smart electric meters, car sensors, road cameras, satellites, traffic recording devices, processors found within vehicles, video games, cable boxes or household appliances, assembly lines,

cell towers and jet engines, air conditioning units, refrigerators, trucks, farm machinery and so on. This contributes to big data.

V. BIG DATA MINING

Useful data can be retrieved from this large datasets with the aid of big data mining. Here the data which are handled is big data, hence the term big data mining. Usually, data mining is the technique of analyzing data from different prospects and summarizing these data into interesting, understandable and useful models. For better decision making, the large repositories of data collected from different resources require a proper mechanism for extracting knowledge from the databases. Since big data scales far beyond the capacity of single PC, cluster computers, which have high computing powers and rely on parallel programming paradigms, are used. Thus a large attempt to exploit these huge parallel processing architectures was initiated.

Map Reduce Programming model

Map Reduce is the widely used parallel processing programming model and played a significant role in processing big data [12]. Map Reduce is a parallel and distributed programming model developed by Google for processing large datasets [13]. Map Reduce is a batch-oriented parallel computing model. The Map Reduce framework allows users to define two functions, map and reduce, to process large number data entries in parallel. Users, normally specify a map function that processes a key/value pair to create a set of intermediate key/value pairs, and a reduce function that helps to merge all intermediate values associated with the same intermediate key. Advantages of Map Reduce are listed below.

- Simple to use
- High fault tolerance
- Can be used for multiple scenarios
- Scalability
- High throughput

A brief discussion on the various big data mining platforms is given below.

VI. BIG DATA MINING PLATFORMS

A. Google's Map Reduce, Hadoop and Google Big Table

Google's programming model, Map Reduce, and its distributed file system, Google File System (GFS) [14] are the pioneers in the field. Improving the performance of Map Reduce and enhancing the real-time nature of large-scale data processing has received a significant amount of attention, with Map Reduce parallel programming. So with this concept many companies provide big data processing framework that support Map Reduce. After that Yahoo and related companies developed Hadoop [11] uses the Hadoop Distributed File System (HDFS) – an open source version of the Google's GFS. Later in this field to support the Map Reduce computing model strategy, Google developed the BigTable [15] in 2006– a distributed storage system designed for processing structured data with size in the order of petabytes.

B. Dynamo

In 2006 Amazon developed Dynamo [16], which uses a key-value pair storage system. Dynamo is a highly available and scalable distributed data store built for Amazon's platform. Dynamo is used to manage services that need high reliability, availability, consistency, performance and cost effectiveness.

The following models are also developed to support big data management and processing.

C. HBase

HBase [17] is an open source, non relational, distributed database developed after big table. It works on the top of Hadoop Distributed file system and provides big-table like capabilities for Hadoop

D. Apache Hive

Apache Hive[18] is a data warehouse infrastructure built on top of Hadoop. It provides data summarization, query, and analysis of big data

E. Berkeley Data Analytics Stack(BDAS)

The Berkeley Data Analytics Stack (BDAS) [19] is an open source data analytics stack that integrates software components being built by the UC Berkeley AMPLab for computing and analyzing big data. Many systems in the stack provide higher performance over other big data analytics tools, such as Hadoop. Nowadays, BDAS components are being used in various organizations.

The key open source components of the stack are:

- Spark, a computation engine built on top of the Hadoop which support iterative processing (e.g., Machine Learning algorithms), and interactive queries. Spark gives an easy-to-program interface that is available in Java, Python, and Scala. Spark Streaming, a new component of Spark provides high

scalability, fault-tolerance and stream processing capabilities. Hence Spark provides integrated support for all major computation models such as batch, interactive, and streaming models.

- Shark is a significant data warehouse system .It runs on top of Spark .It allows users to run unmodified Hive queries on existing Hive workhouses because of backward-compatibility with Apache Hive,
- Mesos, a cluster manager that gives an adequate platform for performing resource isolation and sharing efficiently across distributed applications.

F. ASTERIX

ASTERIX [20] is an Open Source System for big data management and analysis. With the help of ASTERIX Semi structured data can be easily ingested, stored, managed, indexed, retrieved and analyzed. Many of the drawbacks of Hadoop and similar platforms such as single system performance, difficulties of future maintenance, inefficiency in extracting data and awareness of record boundaries etc are easily overcome by ASTERIX

G. SciDB

SciDB [21] is an open-source data management and analytics software system (DMAS) that uses a multi-dimensional array data model. SciDB is designed to store petabytes of data distributed over a large number of machines and used in scientific, geospatial, financial, and industrial applications

To tackle, big data mining ,very-large-scale parallel machine learning and data mining algorithms (ML-DM) are developed for e.g Hadoop Map Reduce, NIMBLE[22], Big Cloud-Parallel Data Mining (BC-PDM)[23], Giraph[24], GraphLab[24], Bulk Synchronous Parallel Based Graph Mining (BPGM)[25] etc.

H. Hadoop Map Reduce

These algorithms work on top of Hadoop and make use of Map Reduce programming model.

I. NIMBLE

A open source toolkit for the Implementation of Parallel Data Mining and Machine Learning Algorithms on Map Reduce for large datasets. It allows users to compose parallel ML-DM algorithms using reusable (serial and parallel) building blocks that can be efficiently manipulated using almost all parallel programming models such as Map Reduce. It runs on top of Hadoop.

J. Big Cloud-Parallel Data Mining(BC-PDM)

Big Cloud Parallel Data Mining mainly relies on cloud computing and works on top of Hadoop and mainly used in intelligence data analysis.

K. Graph Mining tools

Graphs are widely used in data mining application domains for identifying relationship patterns, rules, and anomalies. Certain examples for domains include the web graph, social networks etc. The ever-expanding size of graph-structured data for the above applications needs a scalable system that can process large amounts of data efficiently. Giraph, GraphLab, Bulk Synchronous Parallel Based Graph Mining (BPGM) are the examples for the system to process graph structured data.

Many techniques were developed earlier in the analysis of big data. With the advancement in the field of big data, the various analytic techniques such as structural coding, frequencies, co-occurrence, graph theoretic data reduction techniques, hierarchical clustering techniques, multidimensional scaling techniques were developed for large qualitative data sets. It is clearly described that the need for the particular proposal arise with the type of dataset and the way the pattern are to be analyzed [26].

An overview of different classification and clustering algorithms that are mainly defined for handling big data is given below

VII. BIG DATA MINING ALGORITHMS

A. Decision tree induction classification algorithms

In the initial stage different Decision Tree Learning was used to analyze the big data. In decision tree induction algorithms, tree structure has been widely used to represent classification models. Most of these algorithms follow a greedy top down recursive partition strategy for the growth of the tree. Decision tree classifiers break a complex decision into collection of simpler decision. Hall. et al. [27] proposed learning rules for a large set of training data. The work proposed by Hall et al generated a single decision system from a large and independent subset of data. An efficient decision tree algorithm based on rainforest frame work was developed for classifying large data set [28].

B. Evolutionary based classification algorithms

Evolutionary algorithms use domain independent technique to explore large spaces finding consistently good optimization solutions. There are different types of evolutionary algorithms such as genetic algorithms, genetic

programming, evolution strategies, evolutionary programming and so on. Among these, genetic algorithms were mostly used for mining classification rules in large data sets [29]. Patil et al. [30] proposed a hybrid technique combining both genetic algorithm and decision tree to generate an optimized decision tree thus improving the efficiency and performance of computation. An effective feature and instance selection for supervised classification based on genetic algorithm was developed for high dimensional data [31].

C. Partitioning based clustering algorithms

In partitioning based algorithms, the large data sets are divided into a number of partitions, where each partition represents a cluster. K-means is one such partitioning based method to divide large data sets into number of clusters. Fuzzy- CMeans is a partition based clustering algorithm based on Kmeans to divide big data into several clusters[32]

D. Hierarchical based clustering algorithms

In hierarchical based algorithms large data are organized in a hierarchical manner based on the medium of proximity. The initial or root cluster gradually divides into several clusters. It follows a top down or bottom up strategy to represent the clusters. Birch algorithm is one such algorithm based on hierarchical clustering[33].To handle streaming data in real time, a novel algorithm for extracting semantic content were defined in Hierarchical clustering for concept mining[34].This algorithm was designed to be implemented in hardware, to handle data at very high rates. After that the techniques of self-organizing feature map (SOM) networks and learning vector quantization (LVQ) networks were discussed in Hierarchical Artificial Neural Networks for Recognizing High Similar Large Data Sets [35]. SOM consumes input in an unsupervised manner whereas LVQ in supervised manner. It subdivides large data sets into smaller ones thus improving the overall computation time needed to process the large data set.

E. Density based clustering algorithms

In density based algorithms clusters are formed based on the data objects regions of density, connectivity and boundary. A cluster grows in any direction based on the density growth. DENCLUE is one such algorithm based on density based clustering [36].

F. Grid based clustering algorithms

In grid base algorithms space of data objects are divided into number of grids for fast processing. OptiGrid algorithm is one such algorithm based on optimal grid partitioning [37].

G. Model based clustering algorithms

In model based clustering algorithms clustering is mainly performed by probability distribution. Expectation-Maximization is one such model based algorithm to estimate the maximum likelihood parameters of statistical models [38].

In 2013, a new algorithm called “scalable Visual Assessment of Tendency” (sVAT) [39] algorithm was developed to provide high scalable clustering in big data sets. Afterwards a distributed ensemble classifier algorithm [40] was developed in the field based on the popular Random Forests for big data. This proposed algorithm makes use of Map Reduce for improving the efficiency and stochastic aware random forests for reducing randomness. Later in the field, a mixed visual or numerical clustering algorithm for big data called ClusiVAT [41] was developed to provide fast clustering.

VIII. CHALLENGES OF BIG DATA MINING

A. Evaluating the interestingness of mined patterns

Big data mining or simply mining allows the discovery of knowledge which is potentially useful and unknown. If the knowledge extracted is new, useful or interesting, it is very subjective and mainly depends upon the application and the user who uses the data. It is firm that data mining can create, or discover, a large number of patterns or rules. In some cases the number of patterns is infinite. A meta-mining phase can also be done to mine these infinite data mining results. Additional efforts have to be put in the form of measurements to identify only those patterns that will be of potential importance. Sometimes these patterns may not be interesting and causes the problem of completeness .i.e. the user would want to extract only those that are interesting. The measurement of how interesting an extracted pattern is, often called interestingness, can be relied on quantifiable objective elements such as validity, certainty, and understandability of the patterns. Extracted patterns may also be found interesting if they confirm or validate a hypothesis sought to be confirmed or unexpectedly contradict a general principle. This leads to the issue of describing what is interesting to extract in knowledge discovery or mining. Typically, measurements for interestingness are based on “entry points” set by the user. These “entry points” define the completeness of the rules discovered. Identifying and measuring the interestingness of patterns and rules discovered, or to be discovered is a must for the evaluation of knowledge discovery process. Assessing how interesting a mined pattern is still an important research issue [42].

B. Building a global unifying theory of big data mining

Many techniques are designed for performing classification or clustering individually, but there is no theoretical framework that unifies different tasks such as classification, clustering and association rules and so on. Therefore building a global unifying theory for mining big data is an active research area [43].

C. Scaling up to meet the growing needs of large data sets

In order to meet the growing demands of data, we need to scale up both in terms of capacity and performance measures effortlessly. Big data needs more capacity, scalability, and efficient processing capabilities without increasing the resource demands. In traditional systems, storage architectures were designed in such a way to scale up with the growing needs of data. But it really affects the performance capacity of the storage systems. So organizations dealing with big data should design an optimal storage architecture which offers the features such as scalability, high performance, high efficiency, operational simplicity, interoperability and so on to manage growth.

D. Building efficient big data mining platform

To handle big data and its characteristics, an efficient big data Processing and computing framework is needed. Traditional data mining algorithms only needed all the data to be loaded into the main memory and perform the operation of data mining. In medium scale data processing, parallel computing is used with limited number of processors. As big data applications are characterized by autonomous sources and decentralized controls, consolidating distributed data sources to a centralized node for mining is systematically discouraging due to the potential transmission cost and privacy issues. So building an efficient platform to mine big data is essential.

E. Building efficient mining algorithms/models for big data

With the exponential growth of data, traditional data mining algorithms have been unable to meet large data processing needs. In order to deal with big data, an efficient model that deals with cost effective computation of huge, heterogeneous, sparse, incomplete, complex data are needed. The main drawbacks of big data mining algorithms are lack of user-friendly interaction support**,** quality and performance. Data mining algorithms usually needs scanning of entire data for obtaining perfect statistics and there may require intensive computing. Therefore it is essential to improve the efficiency and performance of data mining algorithms to handle big data.

F. Maintaining security, trust and data integrity

Security is a major concern with big data. In order to ensure security, organizations need to establish policies, which are self-configurable. Another major issue in the field is trust of data sources which are not well – known and not at all verifiable. Data Integrity should be maintained by adopting the best practices in the industry.

G. Data privacy issues

Data privacy has been always a serious issue right from the beginning of Data mining applications. The concern has become extremely vigorous with big data mining that often needs personal information in order to give relevant and accurate outputs. Also, the massive volume of big data such as in social media sites that contains tremendous amount of highly interconnected personal information can be easily mined out and when all pieces of the information about a person are mined out and put together, any privacy about that individual rapidly disappears.

IX. CONCLUSION

We are living in a digital world of big data where massive amounts of heterogeneous, autonomous, complex and evolving data sets are constantly generated at unprecedented scale. In this paper, an overview of big data along with it types, sources, characteristics and challenges are discussed. This paper reviews about the various big data mining platforms and algorithms. To support big data mining, high-performance computing platforms are required. It is understood that interestingness of discovered patterns, developing a global unifying theory, building efficient mining platforms or algorithms, privacy preserving, security, trust and data integrity are the major challenging issues in the current big data mining scenario. It is known that big data mining is an emerging trend in all science and engineering domains and also a promising research area. In spite of the limited work done on big data mining so far, it is believed that much work is required to overcome its challenges related to the above mentioned issues.

REFERENCES

[1] U. Fayyad Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling. http:// big-data-mining.org/keynotes/#fayyad 2012
[2] "IBM What Is Big Data: Bring Big Data to the Enterprise," http://www-01.ibm.com/software/data/bigdata/ 2012
[3] Big data spectrum Infosys. http://www.infosys.com/cloud/resource-center/Documents/big-data-spectrum.pdf
[4] A. Rajaraman and J. Ullman, Mining of Massive Data Sets.Cambridge Univ. Press.2011
[5] IDC, Extracting Value from Chaos: http://idcdocserv.com/1142, june 2011
[6] O'Reilly Radar, What is bigdata? http://radar.oreilly.com/2012/01/what-is-big-data.html. January 11,2012
[7] IDC, 2012
[8] Peter Buneman, Semistructured Data http://homepages.inf.ed.ac.uk/opb/papers/PODS1997a.pdf, 1997,
[9] Doug Laney of META Group ,"3-D Data Management: Controlling Data Volume, Velocity and Variety, 2001

[10] Anuradha, G., , Suggested techniques for clustering and mining of data streams, Published in:Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on Date of Conference:4-5 April 2014
[11] Xindong Wu , Gong-Quing Wu and Wei Ding " Data Mining with Big data ", IEEE Transactions on Knoweledge and Data Engineering Vol 26 No1 Jan 2014
[12] Pandey, Shweta;Tokekar, Vrinda, Prominence of MapReduce in Big Data Processing Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on April 2014
[13] Dean, J., Ghemawat, S MapReduce: Simplified Data Processing on Large Clusters. In: 6th Symposium on Operating System Design and Implementation (OSDI), pp. 137-150 , 2004.
[14] Ghemawat, S. Gobioff, H., Leung, S.T, The Google File System. In: 19th ACM Symposium on Operating Systems Principles, pp. 29-43. Bolton Landing, New York ,2003
[15] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber, Bigtable: A Distributed Storage System for Structured Data, 2006
[16] DeCandia, G., Hastorun, D., Jampani, et al : Dynamo: Amazon's Highly Available Key-Value Store. In: 21st ACM SIGOPS Symposium on Operating Systems Principles, pp.14-17. Stevenson, Washington, USA, 2007
[17] Lizhi Cai,Shidong Huang;Leilei Chen;Yang Zheng,Performance analysis and testing of HBase based on its architecture, Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on June 2013
[18] Taoying Liu, Jing Liu ; Hong Liu ; Wei Li , A performance evaluation of Hive for scientific data management. Big Data, IEEE International Conference on October 2013
[19] Mike FranklinUC Berkeley, USA, The Berkeley Data Analytics Stack: Present and future Big Data, 2013 IEEE International Conference 9 Oct. 2013
[20] Sattam Alsubaiee, Yasser Altowim, Hotham Altwaijry, Alexander Behm, Vinayak R. Borkar, Yingyi Bu, Michael J. Carey, Raman Grover, Zachary Heilbron, Young-Seok Kim, Chen Li, Nicola Onose, Pouria Pirzadeh, Rares Vernica, Jian Wen. ASTERIX: An Open Source System for "Big Data" Management and Analysis, 2012
[21] Stonebraker, M.;Brown, P.;Donghui Zhang;Becla, J., SciDB: A Database Management System for Applications with Complex. Computing in Science & Engineering (Volume:15 , Issue: 3),july 2013
[22] Amol Ghoting, Prabhanjan Kambadur, Edwin Pednault, and Ramakrishnan Kannan, NIMBLE: a toolkit for the implementation of parallel data mining and machine learning algorithms on mapreduce,2011.
[23] Chao Deng, Ling Qian, Meng Xu, Federated Cloud-based Big Data Platform in Telecommunications, Workshop on Cloud Services, Federation, and the 8th Open Cirrus Summit, , September 21 2012, San Jose, CA, USA
[24] Sherif Sakr ,Senior Research Scientist , Processing large-scale graph data: A guide to current technology Learn about Apache Giraph, GraphLab, and other open source systems for processing graph-structured big data National ICT Australia 10 June 2013
[25] Yang Liu, Bin Wu, Hongxu Wang, and Pengjiang, BPGM: A Big Graph Mining Tool, Tsinghua Science and Technology, February 2014
[26] Emily Namey, Greg Guest, Lucy Thairu, Laura Johnson,"Data Reduction Techniques for Large Qualitative Data Sets",2007
[27] Lawrence 0. Hall, Nitesh Chawla , Kevin W. Bowyer, "Decision Tree Learning on Very Large Data Sets", IEEE, Oct 1998
[28] Thangaparvathi, B., Anandhavalli, D An improved algorithm of decision tree for classifying large data set based on rainforest framework, Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference on Oct. 2010 Page(s):800 – 805
[29] D. L. A Araujo., H. S. Lopes, A. A. Freitas, "A parallel genetic algorithm for rule discovery in large databases" , Proc. IEEE Systems, Man and Cybernetics Conference, Volume 3, Tokyo, 940-945, 1999.
[30] Mr. D. V. Patil, Prof. Dr. R. S. Bichkar,"A Hybrid Evolutionary Approach To Construct Optimal Decision Trees with Large Data Sets", IEEE, 2006
[31] Ros, F., Harba, R. ; Pintore, M. Fast dual selection using genetic algorithms for large data sets, ,Intelligent Systems Design and Applications (ISDA), 12th International Conference on Date of Conference:27-29 Nov. 2012 Page(s):815 – 820, 2012.
[32] J. C. Bezdek, R. Ehrlich, and W. Full. Fcm: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2):191–203,1984.
[33] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. ACM SIGMOD Record, volume 25, pp. 103–114, 1996
[34] Moshe Looks, Andrew Levine, G. Adam Covington, Ronald P. Loui, John W. Lockwood, Young H. Cho, 2007,"Streaming Hierarchical Clustering for Concept Mining", IEEE, 2007
[35] Yen-ling Lu, chin-shyurng fahn, 2007, "Hierarchical Artificial Neural Networks For Recognizing High Similar Large Data Sets. ", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, August 2007
[36] A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. Proc. of the ACM SIGKDD Conference on Knowledge Discovery ad Data Mining (KDD), pp. 58–65, 1998.
[37] A. Hinneburg, D. A. Keim, et al. Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. Proc. Very Large Data Bases (VLDB), pp. 506–517, 1999.
[38] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), pp. 1–38,1977.
[39] Havens, T.C. ; Electr. & Comput. Eng. Dept., Michigan Technol. Univ., Houghton, MI, USA ; Bezdek, J.C. ; Palaniswami, M ,2013,Scalable single linkage hierarchical clustering for big data, Intelligent Sensors, Sensor Networks and Information Processing, IEEE Eighth International Conference on April 2013
[40] Assuncao, J. ; Comput. Sci. Dept., PUCRS Univ., Porto Alegre, Brazil ; Fernandes, P. ; Lopes, L. ; Normey, S, Distributed Stochastic Aware Random Forests -- Efficient Data Mining for Big Data, Big Data (Big Data Congress), 2013 IEEE International Congress on June -July 2 2013
[41] Kumar, D. ; EEE, U. of Melbourne, Melbourne, VIC, Australia ; Palaniswami, M. ; Rajasegarar, S. ; Leckie, C,2013, clusiVAT: A mixed visual/numerical clustering algorithm for big data. Big Data, 2013 IEEE International Conference on Oct. 2013
[42] Vashishtha, J. GJUST, Kumar, D. ; Ratnoo, S., Revisiting Interestingness Measures for Knowledge Discovery in Databases, Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on Jan. 2012 Page(s):72 – 78
[43] QIANG YANG, XINDONG WU, 10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH, International Journal of Information Technology & Decision Making,Vol. 5, No. 4 (2006) 597–604, World Scientific Publishing Company,

Brief Author biography

Sherin A received Bachelor of Technology degree in Computer Science from Government College of Engineering, Sreekrishnapuram, Palakkad, Kerala, India affiliated to Calicut University. Now she is pursuing Masters of Engineering in Computer Science & Engineering from Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India affiliated to Anna University.

Dr S.Uma is Professor and Head of PG Department of Computer Science and Engineering at Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India. She received her B.E., degree in Computer Science and Engineering in First Class with Distinction from PSG College of technology in 1991 and the M.S., degree from Anna University, Chennai, Tamilnadu, India. She received her Ph.D., in Computer Science and Engineering Anna University, Chennai, Tamilnadu, India with High Commendation. She has nearly 24 years of academic experience. She has organized many National Level events like seminars, workshops and conferences. She has published many research papers in National and International Conferences and Journals. She is a potential reviewer of International Journals and life member of ISTE professional body. Her research interests are pattern recognition and analysis of non linear time series data.

Saranya K received Bachelor of Technology degree in Information Technology from Sri Ramakrishna Institute of Technology, Coimbatore, India. Now she is pursuing Masters of Engineering in Computer Science & Engineering from Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India affiliated to Anna University.

Saranya Vani M received Bachelor of Technology degree in Information Ttechnology from Amrita Vishwa Vidyapeetham, Coimbatore, India. Now she is pursuing Masters of Engineering in Computer Science & Engineering from Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India affiliated to Anna University.