IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DATA COMPRESSION SECURING DATA IN CLOUD

[1]Sathea sree, [2]Ragunath.R, [3]Samuel.I, [4]Sharan.G
[1]AssistantProfessor, [2,3,4]Students
Department of Computer Science and Engineering,
PERI Institute of Technology, Chennai, India.

Abstract: The rapid development of information technology over the last decade means that data appears in a wide range of sensor data, tweets, photos, raw data and unstructured data formats. With such an overwhelming flood of information, current data management systems cannot scale to this enormous quantity of raw, unstructured data — Big Data, today. We show the basic concepts and designs of big data tools, algorithms and techniques in the present study. We compare the classical data mining algorithms with the Big Data algorithms by using Hadoop / Map Reduce as the core scalable algorithm implementation of Big Data. We implemented the K-means and A-priori algorithms on a 5-node Hadoop cluster with Hadoop / Map Reduce. We use MongoDB as an example to explore NoSQL databases for semi-structured, massive data scaling. Finally, we show the performance of these two algorithms between HDFS (Hadoop Distributed File System) and MongoDB data storage.

I. INTRODUCTION:

1.1. DATA MINING

Data Mining is the advanced process which extracts the potential and effective and comprehensive mode from the vast amounts of Data in accordance with the established business goals. Many people consider data mining as commonly term of knowledge discovery, while others simply put data mining as basic steps in the process of knowledge discovery. It appeared in the late 1980s, which is new areas with a great researching value in the study of the database, and overlapping subject, combined with artificial intelligence, database technology, pattern recognition, machine learning, statistics, data visualization, and other fields of theory and technology. As a kind of technology, the life cycle of data mining is in unclear stage, experts need takes time and energy to research, develop and propel to be mature gradually, eventually been accepted . Data mining is a kind of technology, which combines the traditional data processing methods with different algorithms, to analyze new data types and extract knowledge from huge amounts of data. Found in huge amounts of data, there are two kinds of knowledge, one is on-line analytical processing (OLAP), the other is a data mining (DM). Both are analysis tools based on data warehouse, but on-line analytical processing appears earlier than the data mining, based on a multidimensional view, emphasizing the execution efficiency and quick response for user commands. And data mining is pay attention to the useful model to people hiding in the depths of data, and it is done by automation, without participation of customers.

1.2. ON WHAT DATA MINING

Data mining can examine any type of data and information flow, its difficulty is relative with database type.

1. **Mining for relational database**: A relational database is the set of tables; table is composed of attributes group, depositing large number of tuples. Usually use ER model to represent the connection between the database and the real. Excavating from a relational database, the trends and data model can be obtained. For example, the customer's income, age, education level, and other information can be obtained and by commercial relational database for mining, then making targeted marketing for customer and avoiding fraud, and shape a company's strategy.
2. **Mining for data warehouse**: Data warehouse is a subject oriented, integrated, non-volatile and time-variant collection of data, which contains consistent data used in enterprise decision support. The data warehouse is the data environment that can be used as the single integrated source of data for processing information. Data warehouses deposit aggregated data, which are processed to find hidden patterns and relation to structure analytical model to classify and forecast.
3. **Mining for new database**: The new database includes spatial database, time database and text database and multimedia database. These data include spatial data, text, data, image and audio data and streaming data and web data. The data structure is more complex and dynamic change, more difficult to handle. For example, through the data mining technology can find the evolution of the object characteristics and trends; streaming data are clustered and compared to find interesting patterns.

1.3. THE EVOLUTION PROCESS OF DATA MINING

In 1960, database technology, and information technology has gradually developed from the basic document processing system to more complex and more powerful database system, such as hierarchical and network database are typical representative of this era with little data independence and abstraction. 1970s, relational databases appear, allowing users access to a flexible data access language and interface, OLTP technology make the relational database technology application gained popularity; Mid-1980s, the rise of a powerful database system, and put forward many advanced data models. for example expanding the relational model, object-oriented model, the interpretation of the model, etc. by the end of 80 advanced data model and application oriented database were developed

After 2000, the ability to store large amounts of data is over the capacity of human analysis and understanding; there is no suitable tool to help extracting information and knowledge from the data. The existence of specific patterns and rules can be found by data mining tools in a large amount of data, which can provide the necessary information for commercial activity, scientific exploration and medical research and many other areas. Business intelligence (BI) based on data mining has become the new darling of the IT industry. Currently the data mining has been successfully applied in retail goods basket data analysis, financial risk prediction, product quality, molecular biology, genetic engineering, discovery of Internet site access patterns, the information search and classification and many other fields.

1.4. OBJECTIVES

1. The motivations and objectives of this research are to:
2. Explore Big Data technologies, tools, and concepts.
3. Explore the new database shift paradigm in databases, with what is called NoSQL databases. We focused on document-oriented databases – an example of it being MongoDB database.
4. Implement the common data mining algorithms. Specifically we implemented Apriori algorithm and K-means clustering algorithm by using the Map Reduce model. Then we show the performance between HDFS (Hadoop Distributed File System) data store and MongoDB data store.

II. SYSTEM ANALYSIS

2.1 EXISTING SYSTEM

Relational databases were built based on mathematical foundation introduced by E. F. Codd specifically based on set theory and relational algebra. Relational databases have specific schema design, where data is stored in a table format, and each table may have a relation with another table through some constraints. Data retrieval tasks in RDBMS can be done through SQL (Structured Query Language), which is the standard for storing and retrieving data from a relational database. However, the large growth in data, the different varieties of format, and the need for scalable web applications drove the requirements for a new database development. This new database paradigm is called NoSQL (Not Only SQL). NoSQL is a non-relational databases systems, schema-free, web scalable, BASE (does not support full ACID property) where data is stored in semi-structured or in raw format.

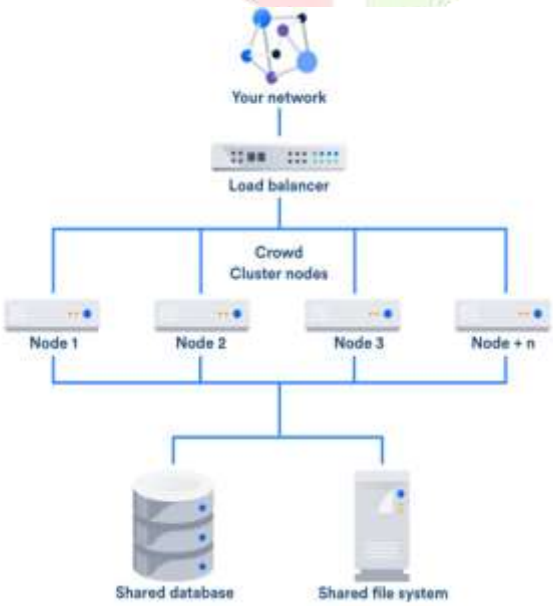
2.2. PROPOSED SYSTEM

As explained earlier, the model is the algorithm that is applied to the data to find similarities, patterns, data summarizations. In this section A-priori algorithm and Kmeans algorithm are covered in detail.

ADVANTAGES

1. **M**DFS store large amount of information
2. MDFS is simple and robust coherency model
3. That is it should store data reliably.
4. MDFS is scalable and fast access to this information and it also possible to serve s large number of clients by simply adding more machines to the cluster.
5. MDFS should integrate well with Matlab MongoDB, allowing data to be read and computed upon locally when possible.
6. MDFS provide streaming read performance.
7. Data will be written to the MDFS once and then read several times.
8. The overhead of cashing is helps the data should simply be re-read from MDFS source.
9. Fault tolerance by detecting faults and applying quick, automatic recovery
10. Processing logic close to the data, rather than the data close to the processing logic

2.3 PROPOSED ARCHITECTURE



III. MODULE SPECIFICATION

According to the system we divided the total system into three modules they are,

3.1. DATA LAYER

As mentioned above, data layer can be database and/or data warehouse systems. This layer is an interface for all data sources. Data mining results are stored in data layer so it can be presented to end-user in form of reports or other kind of visualization.

3.2. DATA MINING APPLICATION LAYER

This layer is used to retrieve data from database. Some transformation routine can be performed here to transform data into desired format. Then data is processed using various data mining algorithms.

3.3.FRONT-END LAYER

This layer provides intuitive and friendly user interface for end-user to interact with data mining system. Data mining result presented in visualization form to the user in the front-end layer

IV. CONCLUSION

Big Data is new field of study in computer science that applies knowledge from different scientific, technical, and practical applications to seek new answers. Smart phones, digital cameras, smart cars, GPS -- of these devices generate huge amounts of data that have a lot of potential for financial return. For instance, GPS data can be used by insurance companies to track their customers where it helps to predict how likely this driver is to get into an accident and so on.

However, Big Data is defined as having four dimensions or 4Vs. Big Data has volume, which means the data are massively large – TBs, PBs and more. Big Data has velocity, meaning that data need to be processed almost in real time. Big Data also has variety, both semi-structured and unstructured. Finally, the fourth dimension of Big Data is veracity, meaning the truthfulness and the accuracy of an inference model for making decisions.

This thesis research encapsulates common Big Data tools and concepts. In Chapter 3, Hadoop, the core of Big Data, was covered in detail. The Hadoop file system can store up to hundreds of Terabytes of data. More importantly, Hadoop implements the MapReduce computation paradigm, a simple yet powerful computing model. It helps hide the complexity of parallel programming from developers, so they only need to write their Map and reduce function. An implementation of the Hadoop cluster was done through this work – 5 of its nodes were deployed on MET IT VMware's server.

Furthermore, an integration of Hadoop MongoDB was also done as a Big Data technology. Hadoop MongoDB has a high potential, and it is a fully open source. Also, MongoDB can be used to build a real time application on top of it. While the heavy computation can be done offline in a Hadoop cluster, the results can be stored back to MongoDB for presentation. Finally, an implementation of A-priori Algorithm and K-means Algorithm were done on both data stores – MongoDB and MFDS.

4.1. FUTURE WORK

A lot potential for future work is contained in this research, summarized below:

1. Practical implementation of Hadoop MongoDB technology stack such as implementing a recommender system on top of MongoDB, with Hadoop used for offline computational power.
2. Design of software packages for data pre-processing using Hadoop MapReduce.
3. Exploration of data visualization for Big Data.

V. REFERENCES

[1] MongoBD. Available: http://www.mongodb.org/

[2] Wei-ping Zhu, Ming-xin Li, and H. Chen, "Using MongoDB to Implement Textbook Management System instead of MySQL," IEEE, pp. 303-305, 2011.

[3] ACID property Available: http://en.wikipedia.org/wiki/ACID

[4] Chang Fay, Dean Jeffery, Ghemawat Sanjay, Hsieh C. Wilson, and W. A. Deborah, "Bigtable: A Distributed Storage System for Structured Data," OSDI, vol. 7, pp. 1-14, 2006.

[5] Michael Maged, Moreira E. José, Shiloach Doron, and W. W. Robert, "Scale-up x Scale-out: A Case Study using Nutch/Lucene," pp. 1-8, 2007.

[6] (2013-06-1). Design for scalability Available: http://en.wikipedia.org/wiki/Scalability

[7] (2013-06-1). Amdahl's law. Available: http://en.wikipedia.org/wiki/Amdahl%27s_Law

[8] Edx : Online Education Platform. Available: https://www.edx.org/

[9] Bigdata-urban Planning. Available: http://web.mit.edu/newsoffice/2013/deanonymize-cellphone-data-0327.html

[10] Montjoye de Yves-Alexandre, Hidalgo A. Ce´sar, Verleysen Michel, and B. D. Vincent, "Unique in the Crowd: The privacy bounds of human mobility," Scientific Reports, pp. 1-5, 2013.

[11] (2013-06-04). Obama's campaign Available: http://www.technologyreview.com/featuredstory/508856/obamas-datatechniques-will-rule-future-elections/

[12] IBM Watson. Available: http://www.nytimes.com/2013/02/28/technology/ibmexploring-new-feats-for-watson.html?pagewanted=all&_r=0

[13] Anand Rajaraman and J. D. Ullman, "Mining of Massive Datasets," Cambridge University Press pp. 1-310, 2012.

[14] E. F. CODD, "A Relational Model of Data for Large Shared Data Banks," Communications of the ACM, vol. 13, pp. 377-387, 1970.

[15] NoSQL Databases Available: http://nosql-database.org/

[16] Dean Jeffery and G. Sanjay, "MapReduce: A flexible Data Processing Tool," Communications of the ACM, vol. 53, pp. 72-77, 2010.

[17] H. Karlo, S. Suri, and S. Vassilvitskii. A Model of Computation for MapReduce. Available: theory.stanford.edu/~sergei/papers/soda10-mrc.pdf

[18] A. Thusoo, J. S. Sarma, N. Jain, S. Zheng, P. Chakka, Z. Ning, et al., "Hive - a petabyte scale data warehouse using Hadoop," in IEEE 26th International Conference on Data Engineering (ICDE), 2010, pp. 996-1005.

[19] Document-Oriented Databases. Available: http://en.wikipedia.org/wiki/Document-oriented_database

[20] Apache Cassandra. Available: https://en.wikipedia.org/wiki/Apache_Cassandra

[21] Partner Jonas, VukoticAleksa, and N. Watt, "Neo4J in Action," Manning Publications Co., pp. 1-19, 2013.

[22] B. A. Eric, "Towards robust distributedsystems. (Invited Talk)," Principles of Distributed Computing, Portland, Oregon.

[23] Gilbert Seth and L. Nancy, "Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services," ACM SIGACT News, vol. 33, pp. 51-59.

[24] Robinson Ion, Webber Jim, and E. Emil, Graph Database, First Edition ed.: O'Reilly Media, Inc.,, 2013.

[25] Mahout. Available: http://mahout.apache.org/ [26] Agrawal Rakesh and S. Ramakrishnan, "Fast Algorithms for Mining Association Rules," pp. 1-13, 1994.

[26] J. Lin. Cloud9 using hadoop. Available: https://github.com/lintool

[27] Jiawei Han, Kamber Micheline, and P. Jian, Data Mining Concepts and Techniques: Waltham, MA: Elsevier Inc, 2012.

[28] Qing He, Fuzhen Zhuang, Jincheng Li, and Z. Shi, "Parallel Implementation of Classification Algorithms Based on MapReduce," pp. 655-662, 2010.

[29] Characteristics of Big Data. Available: http://anlenterprises.com/2012/10/30/ibms-4th-v-for-big-data-veracity/

[30] Veracity definition. Available: http://www.thefreedictionary.com/veracity

[31] W. Tom, Hadoop: The Definitive Guide, SECOND EDITION ed.: O'Reilly Media, Inc, 2011.

[32] N. G. Michael. (2013-06-11). A multi-node Hadoop Cluster. Available: http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-singlenode-cluster/

[33] WordCount-example. Available: http://www.confusedcoders.com

[34] Yahoo tutorial-custom data type. Available: http://developer.yahoo.com/hadoop/tutorial/module5.html#types

[35] (2013-06-11). Mongodb-manual. Available: http://docs.mongodb.org/manual/ [37] Shard Cluster Architecture. Available: http://docs.mongodb.org/ecosystem/tutorial/install-mongodb-on-amazon-ec2/

[36] 10gen - the Mongo-DB company. Available: http://www.10gen.com/

[37] A. Boicea, F. Radulescu, and L. I. Agapin, "MongoDB vs Oracle -- Database Comparison," in Third International Conference on Emerging Intelligent Data and Web Technologies (EIDWT), 2012, pp. 330-335.

[38] J. Changqing, L. Yu, Q. Wenming, U. Awada, and L. Keqiu, "Big Data Processing in Cloud Computing Environments," in Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on, 2012, pp. 17-23.

[39] JatanaNishtha, Puri Sahil, Ahuja Mehak, Kathuria Ishita, and G. Dishant, "A Survey and Comparison of Relational and Non-Relational Database," International Journal of Engineering Research & Technology (IJERT), vol. 1, pp. 2-5, 2012.

[40] H. Jing, E. Haihong, L. Guan, and D. Jian, "Survey on NoSQL database," in 6th International Conference on Pervasive Computing and Applications (ICPCA), 2011, pp. 363-366.

[41] C. Ming-Syan, H. Jiawei, and P. S. Yu, "Data mining: an overview from a database perspective," IEEE Transactions on Knowledge and Data Engineering, vol. 8, pp. 866-883, 1996.

[42] XindongWu,Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J, McLachlan, Angus Ng, Bing Liu, Philip S, Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg: Top 10 algorithms in data mining", Springer-Verlag, 4 December 2007.

[43] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, "Business Intelligence And Analytics: From Big Data To Big Impact, Big Data Analytics An Oracle White Paper", MIS Quarterly vol. 36 no. 4, pp. 1165-1188/December 2012.

[44] San M. Negnevitsky, N. Hatziargyriou, "Applications of Data Mining and Analysis Techniques in Wind Power Systems", 42440178X/06/$20.00 ©2006 IEEE.

[45] Anushree A. Wasu, HarshadaM .Kariya, Shreyas S. Tote, "Evaluating renewable energy using data mining techniques in developing India", Journal of IJSER, IJSER (International Journal of Scientific & Engineering Research), vol. 4, Issue 12, December 2013.

[46] Lionel Fugon, J´er´emieJuban and George Kariniotakis, "Data mining for wind power forecasting", European Wind Energy Conference - Brussels, Belgium, April 2008.

[47] Muhammad Shaheen, Muhammad Shahbaz, Khalid Afsar Khan Jadoon, "Data Mining For Wind Energy Site Selection", Proceedings of the World Congress on Engineering and Computer Science 2012 vol I WCECS 2012, October 24-26, 2012, San Francisco, USA.

[48] Youssef, M., Gamal Attiya, and El-Sayed Ayman. "New Framework For Improving Big Data Analysis Using Mobile Agent."

[49] Krioukov, Andrew, "Integrating Renewable Energy Using Data Analytics Systems: Challenges and Opportunities." IEEE Data Eng. Bull. 34.1 (2011): 3-11.

[50] Niu, Kun, Fang Zhao, and Shubo Zhang. "A fast classification algorithm for big data based on KNN." Journal of Applied Sciences 13, no. 12, pp.2208.

[51] ArintoMurdopo, "Distributed Decision Tree Learning for Mining Big Data Streams", July 2013.

[52] A Min Tjoa, Iman Paryudi, Ahmad Ashari, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool", Journal of IJACSA, IJACSA (International Journal of Advanced Computer Science and Applications), vol. 4, no. 11, 2013.

[53] Abu-Taha, Rimal. "Multi-criteria applications in renewable energy analysis: A literature review." Proceedings of PICMET (Technology Management in the Energy Smart World), 11: IEEE, 2011.

[54] Riondato, Matteo, and Eli Upfal. "Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees." Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2012. 25-41.

[55] Machová, Kristína, Frantisek Barcak, and Peter Bednár. "A bagging method using decision trees in the role of base classifiers." Acta Polytechnica Hungarica3.2 (2006): 121-132.

[56] "Big data & green energy opportunities", Copyright IBM Corporation 2010, Copyright IBM Corporation 2010

[57] Shahrokni, van der Heijde, Lazarevic, Brandt, "Big Data GIS Analytics Towards Efficient Waste Management in Stockholm", 2nd International Conference on ICT for Sustainability (ICT4S 2014).

[58] Chinmay Bhawe, "Big Data Classification Using Decision Trees On The Cloud", Master's Projects. Paper 317.

[59] Chanchal Yadav, Shuliang Wang, Manoj Kumar, "Algorithm and approaches to handle large Data-A Survey", Journal of IJCSN, IJCSN (International Journal of Computer Science and Network), vol. 2, no. 3, 2013.

[60] ErdiÖlmezoğulları, Ismail Ari, Online Association Rule Mining over Fast Data, 2013 IEEE International Congress on Big Data.

[61] Suthaharan, Shan, "Big data classification: problems and challenges in network intrusion prediction with machine learning." ACM SIGMETRICS Performance Evaluation Review 41.4 (2014): 70-73.

[62] Evans, Michael R., "Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities." CyberGIS: Fostering a New Wave of Geospatial Innovation and Discovery, Springer Book, 2013.

[63] Mark J. Embrechts, "bigDAARE: Big Data Analytics for Renewable Energy", CFES 2012-2013 Annual Conference January 25, 2013.

[64] Anoop Verma, Andrew Kusiak, "Prediction of Status Patterns of Wind Turbines: A Data-Mining Approach", Journal of JSEE, JSEE (Journal of Solar Energy Engineering), February 2011.

[65] Kuncheva, Ludmila I., and Juan J. Rodríguez. "An experimental study on rotation forest ensembles." Multiple Classifier Systems. Springer Berlin Heidelberg, 2007. 459-468.

[66] Kale Suvarna Vilas, "Big Data Mining", Journal of CSMR, CSMR (International Journal of Computer Science and Management Research eETECME), October 2013.

[67] Mrs. Deepali KishorJadhav, "The New Challenges in Data Mining", Journal of IJIRCST, IJIRCST (International Journal of Innovative Research in Computer Science & Technology), September 2013.

[68] Rong Liu, Qicheng Li, Feng Li, Lijun Mei, Juhnyoung Lee, Big Data Architecture for IT Incident Management, 2014 IEEE.

[69] Han, Jiawei, Micheline Kamber, and Jian Pei, " Data mining: concepts and techniques: concepts and techniques." Elsevier, 2011.

[70] Minaei-Bidgoli, Behrouz, and William F. Punch. "Using genetic algorithms for data mining optimization in an educational web-based system." Genetic and Evolutionary Computation—GECCO 2003. Springer Berlin Heidelberg, 2003.

[71] Slimani, Thabet. "Application of rough set theory in data mining." arXiv preprint arXiv: 1311.4121 (2013).

[72] Zdzisław Pawlak, Roughsets And Data Mining, Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, ul. Baltycka 5, 44 100 Gliwice, Poland.

[73] Hegland, Markus. "Data mining techniques." Acta Numerica 2001 10 (2001): 313-355.

[74] Mohammed J. Zaki, Limsoon Wong, Data Mining Techniques, August 9, 2003 WSPC/Lecture Notes.

[75] Freitas, Alex A, "A survey of evolutionary algorithms for data mining and knowledge discovery." Advances in evolutionary computing. Springer Berlin Heidelberg, 2003. 819-845.

[76] Ozer, Patrick, "Data Mining Algorithms for Classification." Radboud University Nijmegen, January 2008.

[77] Berkhin, Pavel, "A survey of clustering data mining techniques." Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.

[78] Aloisioa, G., "Scientific big data analytics challenges at large scale" Proceedings of Big Data and Extreme-scale Computing (BDEC) (2013).

[79] Ularu, Elena Geanina, "Perspectives on Big Data and Big Data Analytics", Journal of DBSJ, DBSJ (Database Systems Journal) pp.3-14.

[80] Labrinidis, Alexandros, and H. V. Jagadish. "Challenges and opportunities with big data." Proceedings of the VLDB Endowment 5.12 (2012): 2032-2033.

[81] Ms. Ashwini Mandale, and Prof. ShriniwasGadage, "Big Data Analytics: Challenges, Tools", Journal of IJIRCST, IJIRCST (nternational Journal of Innovative Research in Computer Science & Technology), vol.3, no.3, May 2015.

[82] Yadav, Chanchal, Shuliang Wang, and Manoj Kumar, "Algorithm and approaches to handle large Data-A Survey."arXiv preprint arXiv: 1307.5437(2013).

[83] Wu, Xindong, "Data mining with big data." Knowledge and Data Engineering, IEEE Transactions on 26.1 (2014): 97-107.

[84] Li, Deren, and Shuliang Wang. "Concepts, principles and applications of spatial data mining and knowledge discovery." Proceedings of the International Symposium on Spatio-Temporal Modeling, (STM'05), Beijing, China. 2005.

[85] Gupta, Richa, "Journey from Data Mining to Web Mining to Big Data." arXiv preprint arXiv: 1404.4140 (2014).

[86] Fan, Wei, and Albert Bifet, "Mining big data: current status, and forecast to the future." ACM SIGKDD Explorations Newsletter 14.2 (2013): 1-5.

[87] Davenport, Thomas H., and Jill Dyché, "Big data in big companies." May 2013(2013).

[88] Zaki, Mohammed J., and Wagner Meira Jr, "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, 2014.

[89] [48] Shunxiang, Xu, and Chen Dezhi. "2013 Third International Conference on Intelligent System Design and Engineering Applications ISDEA 2013."

[90] Han, Jiawei, Micheline Kamber, and Jian Pei, "Data mining, southeast Asia edition: Concepts and techniques", 2006.

[91] Sastry, Kumara, David Goldberg, and Graham Kendall. "Genetic algorithms." Search methodologies. Springer US, 2005. 97-125.

[92] Washio, Takashi, and Hiroshi Motoda, "State of the art of graph-based data mining." ACM SIGKDD Explorations Newsletter 5.1 (2003): 59-68.

[93] Tamhane, Deepak S., and Sultana N. Sayyad, "Big Data Analysis Using Hace Theorem", Journal of IJARCET, IJARCET (International Journal of Advanced Research in Computer Engineering & Technology), vol.4, 2015.

[94] Shafaque, Uzma, and Parag D. Thakare, "Algorithm and Approaches to Handle Big Data." IJCA Proceedings on National Level Technical Conference X-PLORE 2014.no. 1. Foundation of Computer Science (FCS), 2014.

[95] Ularu, Elena Geanina, "Perspectives on Big Data and Big Data Analytics." Journal of DBSJ, DBSJ (Database Systems Journal) 2012.

[96] De Francisci Morales, Gianmarco, "SAMOA: A platform for mining big data streams. "Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013.

[97] Cai, Xiao, FeipingNie, and Heng Huang, "Multi-view k-means clustering on big data." Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, 2013.

[98] Lim, A., L. Breiman, and A. Cutler, "BIGRF: Big Random Forests: Classification and Regression Forests for Large Data Sets, 2014.

[99] Kleiner, Ariel, "The big data bootstrap." [59] Hand, David J., "Statistics and data mining: intersecting disciplines." ACM SIGKDD Explorations Newsletter 1.1 (1999): 16-19.

[100] Ceci, Michelangelo, "Big Data Techniques For Renewable Energy Market.

[101] Buck, Samuel F, "A method of estimation of missing values in multivariate data suitable for use with an electronic computer." Journal of the Royal Statistical Society, 1960.