International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(4-1), pp.288-293
DOI: http://dx.doi.org/10.21172/1.841.46
e-ISSN:2278-621X

BIG DATA ANALYTICS: A SURVEY

Gautam[1] & Dr. Chavi Rana[2]

Abstract-It is very difficult tostoring, managing and processing huge amount of data. The term 'Big Data' describes various techniques and technologies to store, distribute, manage and analyze huge amount of data with different structures. Big data consists of structured, unstructured or semi-structured data so there is problems occur regarding incapability of conventional data management methods. To process these huge amounts of data in an inexpensive and efficient way, parallelism is used. Big Data is a data which is in large amount and having complexity in it and this complexity require new architecture, techniques, algorithms, and analytics to manage it and extract knowledge from it.Hadoop is a framework for processing large amount of data and provides better storage capacity for large datasets and performs parallel processing of big data that gives better computational power to all the tasks. It works in batch processing mode and Hadoop is the core platform for structuring Big Data, it also solves the problem of making it useful for analytics purposes. In this paper, we provide a brief overview of Big data management involving hadoopand highlight research efforts and the challenges to big data.
Index Terms: Big Data,Hadoop, Map Reduce, HDFS, Hadoop Component.

1. INTRODUCTION:

1.1. Big Data: Definition

Big data is a term used to describe the exponential growth and availability of data, having structured, unstructured and semi-structured data, whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult or even impossible to be managed and analyzed using conventional software tools and technologies. When the amount of data to be increases than the time to produce results is also increased. Retrieved data from big data is still a complex and time consuming approach. Big dataprovides tremendous opportunities for enterprise information management and decision making. In the recent study big data is not only limited to business needs but also helps in research and scientific issues.
The Big Data problem is characterized by the 3V features:

Volume- a huge amount of data, Volume of big data can be measured in terms or several megabytes, gigabytes, terabytes or petabytes.
Velocity- a high data ingestion rate or the speed with which the data can be analyzed.
Variety- a mix of structured data, semi-structured data, and unstructured data.

These 3V features gives a challenge to data processing systems since these systems cannot either scale to the huge data volume in a cost-effective way or fail to handle data with variety of types. The solutions to the Big Data problem are largely based on the MapReduceframework[9]
andits open source implementation Hadoop. Although Hadoop handles the data volume challenge successfully.Hadoop is the open source software founded by Apache and it is Linux based software. It is used by famous websites like Google, Yahoo, Facebook, Amazon and many more. Hadoop is a framework for processing large amount of data and provides better storage capacity for large datasets and performs parallel processing of big data that gives better computational power to all the tasks. It works in batch processing mode and having two major components HDFS (Hadoop Distributed File System)[12] for huge data storage and MapReduce for processing huge amount of datasets. When the data size is increased it create problems to existing algorithms to manage that so here main problem is to store and process that huge amount of data and this problem is solve by hadoop because it store and process huge amount of data in less time.

1.2. Hadoop:

Hadoop is an open-source software framework used for distributed storage and processing of big data using the MapReduce programming model. Modules present in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework. The core of hadoop consists of two parts the storage part and processing part.

[1] *Research Scholar, UIET, Rohtak*
[2] *Assistant Professor, UIET, Rohtak*

a) Storage part: Storage part of hadoop is HDFS(Hadoop distributed file system) which stores huge amount of data with high degree of throughput and this huge data is stored in form of clusters.
b) Processing part: Processing part of hadoop is Mapreduce which is a software framework which process large amount of data in the form of clusters.
Hadoop distribute clusters to the node so that they process parallely and this approach also takes advantage of data locality This allows the dataset to be processed faster and more efficiently which make it a more conventional supercomputer architecture which work on a parallel file system where computation and data are distributed via high-speed networking

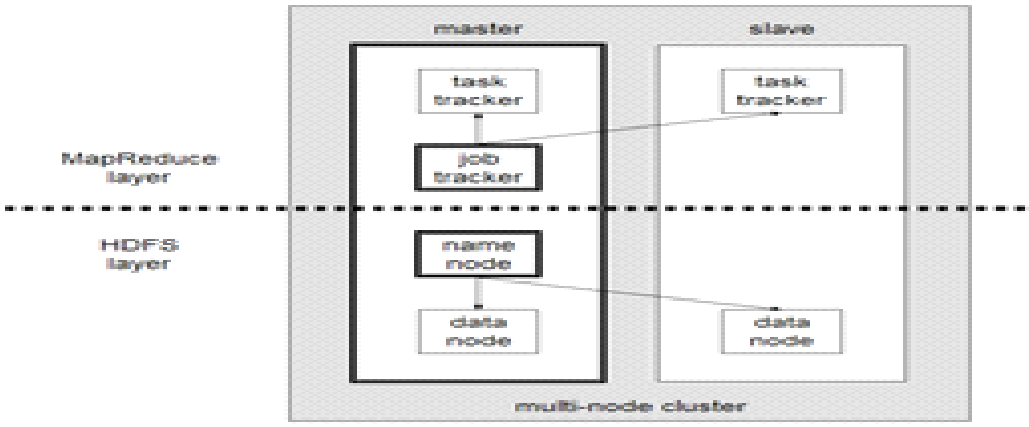

Fig.1.1.Hadoop architecture

A small Hadoop cluster having single master and multiple worker nodes called as slave node as shown in Fig. 1.1.The master node consists of a Task Tracker, Job Tracker, NameNode, and DataNode [14] where as slave or worker node acts as both a DataNode and TaskTracker.

1.3. HDFS:
Hadoop Distributed File System (HDFS) is the storing component in hadoop which store huge amount of structured, unstructured and seminars-structured data.HDFS is java based file system.HDFS is reliable and manageable file system.It has great features such as high availability, load balancing, security, flexible access, fault tolerance,easy management and high data throughputs. It provides parallel processing of data.HDFS has master/ slave architecture.[23]

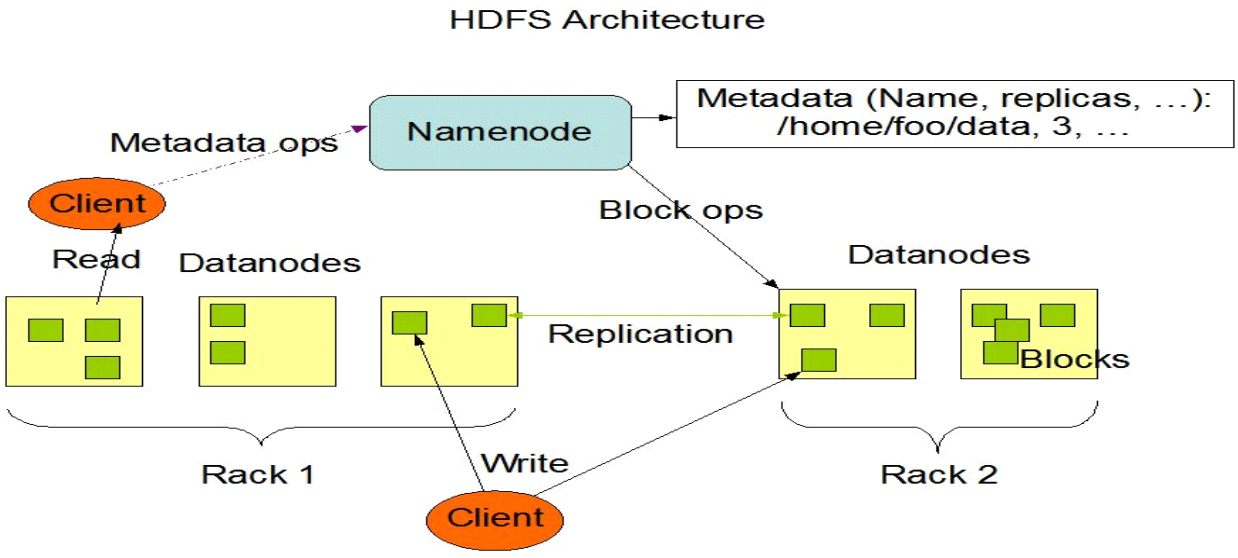

Fig. 1.2. HDFS Architecture

1.4. HadoopMapReduce:

MapReduce is a java based programming paradigm for processing huge amount of data stored in HDFS. MapReduce is the heart of the Hadoop framework that provides scalability across thousands of hadoop cluster. Every MapReduce job performs two tasks - one Map task and theis Reduce task. Map task takes a set of data, processes it at node level and generates the

output. The reduce job takes the output of the map task as the input and combines them to smaller set of tuples (reduces the large dataset into a smaller one) based on the transformations and various logic.The advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes.

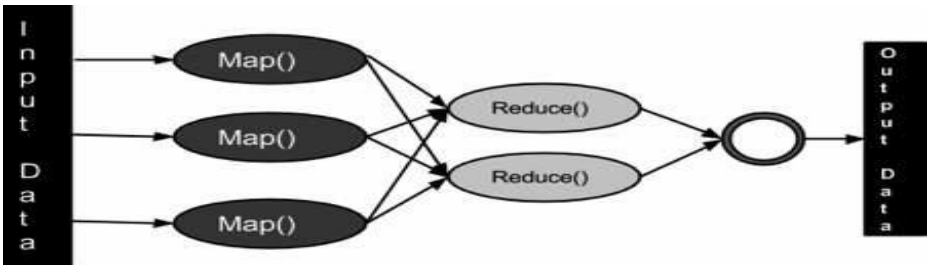

Fig. 1.3.MapReduce Architecture

Map stage: The map stage job is to process the input data as shown in Fig. 1.3. Generally the input data is in the form of file or directory and it is stored in the Hadoop file system (HDFS). The input file is passed to the map function that processes the data and creates several small chunks of data.
Reduce stage: The Reducer's job is to process the data that comes from the map stage. After processing, it produces a new set of output, which will be stored in the Hadoop Distributed File System (HDFS).

2. LITERATURE SURVEY:

This paper provides a detailed review of different approaches used in Big Data in recent years. Table provides the extensive survey of researches; with the name of author, year of publication in descending order of research along with purposed work and approaches used by them as shown below:

Authors	Publication Year	Proposed Work	Technique used
Daniele Apiletti, Elena Baralis, Tania Cerquitelli, Paolo Garza, Fabio Pulvirenti, Luca Venturini. [33]	2017	Reviews Hadoop and Spark based scalable algorithms for mining problem in the Big Data domain having both theoretical and experimental comparative analyses.	Spark algorithms for miningin the Big Data is used.
DineshJ. Prajapati,Sanjay Garg, N.C. Chauhan. [34]	2017	The proposed method initially extracts multilevel association rules including level-crossing for each one using MFFM. From both multilevel consistent and inconsistent rules are evaluated and compared based on different experimental results that lead to the final conclusion.	Use DMFPM for extracts multilevel association rules including level-crossing for each zone.
Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot, Nathalie Ville-Vialaneix. [35]	2017	Proposed a selective review that deal with scaling random forests to Big Data problems and also describe how out of bag error addressed.	Addressing a bag error problem.
M. Bakratsas, P. Basaras, D. Katsaros, L. Tassiulas. [36]	2017	Investigate the relative performance and benefits of SSDs versus hard disk drives (HDDs) when they are used as storage for Hadoop'sMapReduce.	Evaluate SSDs and HDDs by executing algorithm on real social network data.
ZiliangZong, RongGe, QijunGu. [37]	2017	Presented the design of marched system and demonstrate it measurement tools for obtaining power consumption data in different research.	Designed a marched system and its tools.
GuangchenRuan and Hui Zhang. [38]	2017	Proposed framework that integrates information visualization, scalable computing, and user interfaces to explore large-scale multi-model data streams which combine to mediate effective and efficient way to perform cloud-base big data analysis with visualization and scalable computing.	Parallel mining algorithm running on HPC is used.
NavroopKaur, Sandeep K. Sood. [39]	2017	Presented resource management system which solves the problems regarding selecting and allocating appropriate resource to big data and used 4 V's property of big data.	Using Cod and SOM estimate big data characteristics.
Feras A. Batarseh, Eyad Abdel Latif. [40]	2016	Study on healthcare data that is collected from various different sources so that quality and best practices of field is done using big data tools.	Assesses QoS for examines historical health data by analytical infrastructure.
Dawei Jiang, Sai Wu, Gang Chen, Beng Chin Ooi1, Kian-Lee Tan, Jun Xu. [2]	2016	Presents epiC, an extensible system to deliver the Big Data's data variety challenge. They also present the design and implementation of epiC's concurrent programming model and two customized data processing models.	Introduce a new programming model system called epiC.
Marcos D. Assunção, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Netto, RajkumarBuyya. [3]	2015	Discusses environments for carrying out analytics on Clouds for Big Data applications. Through survey they find out possible gaps in technology and provide future directions on Cloud-supported Big Data computing.	Define various method used in data management, model development, visualization and business models.
Sreedhar C.N, Kasiviswanath, P. Chenna Reddy. [1]	2015	The primary purpose of their work is to provide a comprehensive survey on Big data management and to provide an overview on various algorithms related to job scheduling in Hadoop.	Algorithm of Delay and Genetic scheduling is used.
Chao Wang, Xi Li, Peng Chen, Aili Wang, Xuehai Zhou, and Hong Yu. [19]	2015	Proposed a FPGA-based acceleration solution with MapReduce framework. The combination of these two namely hardware acceleration and MapReduce execution flow can enhance the task of aligning short length reads to a known reference genome.	Used FPGA-based acceleration solution with MapReduce framework.
Tao Xu, Dongsheng Wang and Guodong Liu. [20]	2015	Presented an efficient system for managing PB level structured data called Banian, banian overcomes the storage problem.	Used PB level structured data called Banian.
Qinghua Lu, Zheng Li, Maria Kihl, Liming Zhuand Weishan Zhang. [26]	2015	Presented conceptual framework CF4BDA to analyse the existing work done on BDA applications involving the lifecycle of BDA applications and objects involving in BDA applications in the cloud.	Framework CF4BDA to analyze the work on BDA applications.
Claudio A. Ardagna, Ernesto Damiani, FulvioFrati, DavideRebeccani. [25]	2015	Presented score-based benchmark for NoSQL databases, which supports adopters. The proposed benchmark is independent from the specific configurations of the database and deployment environment.	Used score-based benchmark for NoSQL.
Hongbing Wang, Chao Yu, Lei Wan and Qi Yu. [24]	2015	Proposed heterogeneous and trust-based service selection by developing a novel multi-objective optimization approach to make trade-off decision betweenService's trust value and user's QoS preference to rank candidate.	Heterogeneous and trust-based selection by optimization approach.

Simon Fong, Raymond Wong, and Athanasios V. Vasilakos. [23]	2	0	1	5	Presented algorithms to collect big data which is presented in large degree authest for performance evaluation by using accelerated particle swarm optimization (APSO) type of swarm search that enhanced analytical accuracy within reasonable processing time.	Accelerated Particle Swarm Optimization (APSO) algorithms to collect big data.
Yanhao Huang and Xiaoxin Zhou. [22]	2	0	1	5	Proposed the structure, elements, basic calculations and multi-dimensional reasoning method of the new knowledge model. Research shows more powerful and adopts various knowledge requirements of electric power big data.	The knowledge model is established and various calculations is done.
Marco Viceconti, Peter Hunter, and Rod Hose. [21]	2	0	1	5	Proposed that bid data analytics can successfully combined with VPH technology to give desirable medical solutions.	Use VPH technology and combined it with big data analytics.
Alun Evans JaviAgenjoJosep Blat. [28]	2	0	1	5	Presented a web-based application having analytic visualization of on-set media data and metadata, which combines research from several fields of image processing and 3D graphics.	Use WEBGL 3D on the web and meta data visualization techniques.
SyedAkhterHossain. [29]	2	0	1	5	Described the nascent field of big data analytics in education with discussion on prospects and challenges way forward. Also focus on research and development issues for educationists and practitioners of big data analytics.	Nascent field of big data analytics in education and development issues for educationist of big data analytics defined.
Xue-Wen Chen AND Xiaotong Lin. [30]	2	0	1	4	Presented overview of deep learning, and also highlight current research efforts and the challenges to big data, as well as the future trends.	Unprecedented challenges to harnessing data and information is presented.
MatturdiBardi, ZhouXianwei, LI Shuai, LINFuhong. [32]	2	0	1	4	Reviewed the various benefits and challenges of security and privacy in Big Data and also presented some possible methods and techniques to ensure Big Data security and privacy.	Big data security and privacy technique is defined.
SumanArora, Dr.MadhuGoel. [7]	2	0	1	4	Study and analyzed various techniques of scheduling which enhance the performance by using Hadoop.	Speculative execution and Copy compute splitting technique of Hadoop.
Chang Liu, Jinjun Chen, Chi Yang, Rajiv Ranjan, and RamamohanaraoKotagiri. [16]	2	0	1	4	Presented types of fine-grained data updates and scheme that can fully support authorized auditing and fine-grained update requests. Also propose an enhancement that can reduce communication overheads for verifying small updates.	Describe a scheme for supporting variable sized data blocks.
Shifeng Fang, Li DaXu, Yunqiang Zhu, JianhengAhati, Huan Pei, Jianwu Yan, and Zhihui Liu. [17]	2	0	1	4	Introduces a novel IOs (Internet of Things (IoT), Cloud Computing, Geoinformatics, geographical information system (GIS) push-it-learn for environmental monitoring and management, within case study on climate change and its ecological effects of a particular region	Combine IoT, GIS and e-science for environmental monitoring and management.
Daisuke Takaishi, Hiroki Nishiyamai, NeiKatoi and Ryu Miura. [18]	2	0	1	4	Proposed a new mobile sink routing and data gathering method with the help of network clustering based on modified expectation maximization technique.	Use EM algorithm for clustering.
Andrea Marinoni, Arianna Degiati, RiccardoBellazzi, Paolo Gamba1. [27]	2	0	1	3	Provided study of the connection between air pollution and clinical records, than correlations among black particulate concentration, micro and macro-vascular disease can be drawn properly.	micro and macro-vascular disease can be drawn by creating connection between various approach.
XiongpaiQin, and Xiaoyun Zhou. [4]	2	0	1	3	Reviewed last several years big data benchmark work and their characteristics are analyzed.	Use MRBench for evaluating the MapReduceframework.
RakeshVarma. [6]	2	0	1	3	Objective of the research is to study about MapReduce and various algorithms of scheduling which enhance the scheduling performance.	For managing Big Data various scheduling algorithms and LATE-Speculative execution is used.
Daniel Warneke. [15]	2	0	1	1	Discuss the opportunities and challenges for parallel data processing inclouds and present Nephele. And evaluate the MapReduce process and compare the result of framework Hadoop data processing.	Use Nephele, a new data processing framework.
JasminAzemovic,Denis Music. [13]	2	0	1	0	Presented research on using different data types for storing unstructured data within database and this research is inspired with current situation of information society.	Define various way for storing unstructured data.
Mengjie Zhou,HaojiHu and Minqi Zhou. [14]	2	0	1	0	Proposed a SLCA (Smallest Lowest Common Ancestor) based keyword search implementation for large-scale XML data sets on a MapReduce cluster.	SLCA based keyword search implementation for large-scale data.
Leonardo Neumeyer,Bruce Robbins,Anish Nair,AnandKesari. [5]	2	0	1	0	Outline the S4 architecture and describe applications of real-life deployments. They include	For dealing with unbounded stream of data S4 architecture is used.
BI Shuoben, Xu Yin, JiaoFeng, LiGuonian, PEI Anping. [12]	2	0	0	9	Introduces the single-dimensionalBoolean association rule on Apriori algorithm, and the data mining algorithm of the multi-dimensional association rule based on BUC algorithm.	Single-dimensionalBoolean association rule on Apriorialgorithm and BUC algorithm.
Hui Fang, Ming Yang, Ruqing Yang. [11]	2	0	0	7	Proposed a approach to localize the vehicle position with respect to a global map. It is based on the texture of ground from where the vehicle moves.	Use dimensioning technique for localizing global map.
SeemaMetikurke, Vijay K. Vaishnavi. [10]	2	0	0	6	Describes a grid-enabled approach for automatic web page classification that applies the vector space model information retrieval strategy.	Grid-enabled approach for automatic web page classification.
John. H. Phan, Chang. F. Quo, and May D. Wang.[9]	2	0	0	4	Reporting the results of the first phase development of novel system, in use unsupervised methods of clustering to discover relationship of genes and knowledge-based supervised classification is used to get accurate prediction in cancer diagnosis.	Use unsupervised methods of clustering to discover relationship of genes and knowledge-based supervised.
SushantGoel, HemaSharda, David Tanid. [8]	2	0	0	3	Distribute the scheduling responsibilities to the nodes where data is actually located and also propose a new serializability criterion, Parallel Database Quasi-Serializability.	A new serializability criterion, Parallel Database Quasi-Serializability (PDQS) is used.

3. CHALLENGES:

Big data is very huge amount of data so set of challenges occur because difficulties regarding management, storing, scheduling security and processing occur. First, Data preparation, efficiently distributed storage and search is required for effective online analysis which requires effective techniques for data mining.Efficient handling of big data stream is big challenge which uses various programming models. Second, Scheduling, scheduling approach should be smart enough to make real-time responses t o a changing environment. Third, Data Integration, new protocols and interfaces are require which are able to manage structur ed, semi- structured and unstructured data. Fourth, Visualisation and userinteraction. There are many research challenges prese nt in big data visualisation so more efficient techniques are required in real time visualization.

In addition, Security and Privacy is also a big issue in big data. Security is crucial phase in any organization so strong mechani sms for the privacy of data should be needed.

4. CONCLUSION:

A survey of different big data approaches is presented of recent years. It is found that solutions to Big Data problem are largely based on the MapReduce framework and its open source implementation Hadoop,Hadoop handles the data volume challenge successfully. Big data management includes different tools, techniques and various algorithms for job scheduling in hadoop. This paper helps to a novice who wants to pursue his/her career in the field of big data.

5. FUTURE DIRECTION:

This work can be extended by developing a new job scheduling algorithm which consider all the parameters which can produc e better performance. Second, the user profile (similar users) and usage profile (invoked services) should taken and some relat ed collaborative filtering techniques can be considered to integrate with ourservice selection approach.

6. REFERENCES:

[1] Sreedhar C.N, Kasiviswanath, P. Chenna Reddy,“A Survey on Big Data Management and Job Scheduling"International Journal of Computer Applications (0975 – 8887) Volume 130 – No.13, November 2015.

[2] Dawei Jiang, Sai Wu, Gang Chen, Beng Chin Ooi1and Kian-Lee Tan, Jun Xu, "epiC: an extensible and scalable system for processing Big Data"The VLDB Journal (2016) 25:3–26 DOI 10.1007/s00778-015-0393-2.

[3] Marcos D. Assunçãoa, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Nettoc, RajkumarBuyya, "Big Data computing and clouds: Trends and future directions"J. Parallel Distrib. Comput. 79–80 (2015) 3–15.

[4] Xiongpai Qin, and Xiaoyun Zhou,"A Survey on Benchmarks for Big Data and Some More Considerations"H. Yin et al. (Eds.): IDEAL 2013, LNCS 8206, pp. 619–627, 2013. Springer-VerlagBerlin Heidelberg 2013.

[5] Leonardo Neumeyer,BruceRobbins,AnishNair,AnandKesari,"S4: Distributed Stream Computing Platform" 2010 IEEE International Conference on Data Mining Workshops.

[6] RakeshVarma,"Survey on MapReduce and Scheduling Algorithms in Hadoop" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438.

[7] Suman Storage and Dr.MadhuGoel,"Survey Paper on Scheduling in Hadoop" Volume 4, Issue 5, May 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[8] SushantGoel, HemaShardaand DavidTanid,"Distributed Scheduler for High Performance Data-Centric Systems"b7803-76CI-XIO1lB17.00 0 2003 IEEE.

[9] John. H. Phan, Chang. F. Quo, and May D. Wang,"Comparative Study of Microarray Data for Cancer Research" proceedings of the 26th Annual International Conference of IEEE EMBS San Francisco, CA, USA * September 1-5, 2004.

[10] SeemaMetikurke and Vijay K. Vaishnavi, "Grid-Enabled Automatic Web Page Classification" 2006 IEEE International Conference on Fuzzy Systems Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006.

[11] Hui Fang, Ming Yang and Ruqing Yang, "Ground Texture Matching based Global Localization for Intelligent Vehicles in Urban Environment" Proceedings of the 2007 IEEE Intelligent Vehicles Symposium Istanbul, Turkey, June 13-15, 2007.

[12] BI Shuoben, XU Yin, JIAO Feng, LüGuonian, PEI Anping, "Study on Data Mining in First Period of Jiangzhai Site Based on the Association Algorithms"2009 International Conference on Artificial Intelligence and Computational Intelligence, 2009 IEEE DOI 10.1109/AICI.2009.

[13] JasminAzemovic,Denis Music,"Comparative analysis of efficient methods for storing unstructured data into database with accent on performance"201O,IEEE 2nd International Conference on Education Technology and Computer (ICETC).

[14] MengjieZhou,HaojiHu and Minqi Zhou, "Searching XML Data by SLCA on a MapReduce Cluster” 2010 IEEE.

[15] Daniel Warneke,"Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud"IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 22, NO. 6, JUNE 2011.

[16] Chang Liu, Jinjun Chen, Chi Yang, Rajiv Ranjan and RamamohanaraoKotagiri,"Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine-Grained Updates"IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 9, SEPTEMBER 2014.

[17] Shifeng Fang, Li DaXu, Yunqiang Zhu, JiaerhengAhati, Huan Pei, Jianwu Yan, and Zhihui Liu,"An Integrated System for Regional Environmental Monitoring and Management Based on Internet of Things"IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 10, NO. 2, MAY 2014.

[18] Daisuke Takaishi, HirokiNishiyamai, NeiKatoi and Ryu Miura,"Toward Energy Efficient Big Data Gathering in Densely Distributed Sensor Networks"2014 IEEE.

[19] Chao Wang, Xi Li, Peng Chen, Aili Wang, Xuehai Zhou and Hong Yu,"Heterogeneous Cloud Frameworkfor Big Data Genome Sequencing"IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 12, NO. 1, JANUARY/FEBRUARY 2015.

[20] Tao Xu, Dongsheng Wang and Guodong Liu,"Banian: A Cross-Platform Interactive Query Systemfor Structured Big Data"TSINGHUA SCIENCE AND TECHNOLOGY ISSN 1007-021 07/11 p p 6 2- 7 1 Volume 20, Number 1, February 2015.

[21] Marco Viceconti, Peter Hunter and Rod Hose,"Big Data, Big Knowledge: Big Data for Personalized Healthcare"IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 4, JULY 2015.

[22] Yanhao Huang and Xiaoxin Zhou,"Knowledge Model for Electric Power Big Data Based on Ontology and Semantic Web"CSEE JOURNAL OF POWER AND ENERGY SYSTEMS, VOL. I, NO. I, MARCH 2015.

[23] Simon Fong, Raymond Wong, and Athanasios V. Vasilakos,"Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data"IEEE TRANSACTIONS ON JOURNAL NAME, 2015.

[24] Hongbing Wang, Chao Yu, Lei Wan and Qi Yu,"Effective BigData-Space Service Selection overTrust and Heterogeneous QoS Preferences"IEEE, 2015.

[25] Claudio A. Ardagna, Ernesto Damiani, FulvioFrati, DavideRebeccani,"A Configuration-Independent Score-Based Benchmark for Distributed Databases"DOI 10.1109/TSC.2015.2485985, IEEE Transactions on Services Computing.

[26] QINGHUA LU, ZHENG LI, MARIA KIHL, LIMING ZHU AND WEISHAN ZHANG1,"CF4BDA: A Conceptual Framework for Big Data Analytics Applications in the Cloud" IEEE October 27, 2015.
[27] Andrea Marinoni, Arianna Dagliati, RiccardoBellazzi, Paolo Gamba1,"INFERRING AIR QUALITY MAPS FROM REMOTELY SENSED DATA TO EXPLOIT EOREFERENCED CLINICAL ONSETS: THE PAVIA 2013 CASE" IEEE, 2015.
[28] Alun Evans JaviAgenjoJosep Blat,"COMBINED 2D AND 3D WEB-BASED VISUALISATION OF ON-SET BIG MEDIA DATA"978-1-4799-8339-1/15 2015 IEEE.
[29] SyedAkhterHossain,"Big Data Analytics in Education: Prospects and Challenges"978-1-4673-7231-2/15/ 2015 IEEE.
[30] XUE-WEN CHEN1, AND XIAOTONG LIN,"Big Data Deep Learning: Challenges and Perspectives"May 16, 2014, IEEE.
[31] Zhi-HuaZhou,Nitesh V. Chawla,YaochuJin,Graham J. Williams,"Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives"IEEE Computational intelligence magazine | November 2014.
[32] MATTURDI Bardi, ZHOU Xianwei, LI Shuai, LINFuhong,"Big Data security and privacy: A review" China Communications Supplement No.2 2014.
[33] Daniele Apiletti, Elena Baralis, Tania Cerquitelli, Paolo Garza, Fabio Pulvirenti, Luca Venturini , "Frequent itemsetsmining for big data: A Comparative Analysis"IEEE ,Aug 2017.
[34] Dinesh J. Prajapati,SanjayGarg, N.C. Chauhan,"MapReduce Based Multilevel Consistent and Inconsistent Association Rule Detection from Big Data Using Interestingness Measures" vol-9 September 2017,IEEE.
[35] Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot, Nathalie Villa-Vialaneix,"Random Forests for Big Data"Vol-23,IEEE 2017.
[36] M. Bakratsas, P. Basaras, D. Katsaros , L. Tassiulas,"HadoopMapReduce Performance on SSDs for Analyzing Social Networks " IEEE 2017.
[37] ZiliangZong, RongGe, QijunGu,"Marcher: A Heterogeneous System Supporting Energy-Aware High Performance Computing and Big Data Analytics"Volume 8, July 2017.
[38] GuangchenRuanand HuiZhang,"Closed-loop Big Data Analysis with Visualization and Scalable Computing ". Volume 8, July 2017.
[39] NavroopKaur ,Sandeep K. Sood,"Efficient Resource Management System Based on 4Vs of Big Data Streams " Volume 13,April 2017.
[40] Feras A. Batarseh, Eyad Abdel Latif ,"Assessing the Quality of Service Using Big Data Analytics: With Application to Healthcare" Volume4, June 2016.
[41] Gantz J, Reinsel D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East [J]. IDC iView: IDC Analyze the Future, 2012.
[42] Weiss R, Zgorski L. Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments [J]. Office of Science and Technology Policy, Washington, DC, 2012.Data P. The Emergence of a New Asset Class[C]//World Economic Forum Report. 2011.
[43] Anderson C. The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine 16.07[J]. 2008.
[44] Mayer-Schönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think [M]. Houghton Mifflin Harcourt,2013.
[45] Ardagna C A, Damiani E. Business Intelligence meets Big Data: An Overview on Security and Privacy [J].
[46] Manyika J, Chui M, Brown B, et al. Big data:The next frontier for innovation, competition, and productivity [J]. 2011.
[47] Laney D. 3-D Data Management: Controlling Data Volume [J]. Velocity and Variety, META Group Original Research Note, 2001.
[48] Beyer M. Gartner says solving big data challenge involves more than just managing volumes of data. Gartner [J]. 2011.
[49] Beyer M A, Laney D. The importance of 'big data': a definition [J]. Stamford, CT: Gartner, 2012.
[50] Lefevre C. LHC: the guide (English version) [R]. 2009.[14] Brumfiel G. Down the petabyte highway[J]. Nature, 2011, 469(20): 282-283.
[51] Mangelsdorf J. Supercomputing the climate: Nasa's big data mission[J]. Accessed online, 2013: 11-27.
[52] Kalil T. Big data is a big deal[J]. The White House, 2012.
[53] Sheet F. Big Data Across the Federal Government[J]. 2012 03-29)[2013-03-06]. http://www. whitehouse, gov/sites/default/ files/microsites/ostp/big_ data fact sheet final. pdf, 2012.
[54] Lampitt A. 'The real story of how Big Data analytics helped Obama win'[J].Info World, 2013, 14.