WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Big Data Analytics: A Survey

[1]K.Saraswathi, and [2]S.Logeshwaran

[1]Assistant Professor, Department of Electronics and Instrumentation Engineering, SCSVMV, Kanchipuram, India
[2]UG Scholar, Department of Computer science and Engineering, PERI Institute of Technology, Chennai, India

ABSTRACT

It's veritably delicate to storing, managing and recycling huge quantum of data. The term' Big Data' describes colorful ways and technologies to store, distribute, manage and dissect huge quantum of data with different structures. Big data consists of structured, unshaped or semi-structured data so there's problems do regarding incapacity of conventional data operation styles. To reuse these huge quantities of data in an affordable and effective way, community is used. Big Data is a data which is in large quantum and having complexity in it and this complexity bear new armature, ways, algorithms, and analytics to manage it and prize knowledge from it. Hadoop is a frame for recycling large quantum of data and provides better storehouse capacity for large datasets and performs resemblant 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 give a brief overview of big data operation involving hadoopand highlight exploration sweats and the challenges to big data.

Indicator 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 vacuity of data, having structured, unshaped and semi-structured data, whose size (volume), complexity (variability), and rate of growth (haste) make them delicate or indeed insolvable to be managed and anatomized using conventional software tools and technologies. When the quantum of data to be increases than the time to produce results is also increased. Recaptured data from big data is still a complex and time consuming approach. Big data provides tremendous openings for enterprise information operation and decision timber. In the recent study big data isn't only limited to business requirements but also helps in exploration and scientific issues.

The Big Data problem is characterized by the 3V features

Volume-a huge quantum of data, Volume of big data can be measured in terms or several megabytes, gigabytes, terabytes or petabytes.

Haste- a high data ingestion rate or the speed with which the data can be anatomized. Variety-a blend of structured data, semi-structured data, and unshaped data.

These 3V features gives a challenge to data processing systems since these systems cannot moreover gauge to the huge data volume in a cost-effective way or fail to handle data with variety of types. The results to the Big Data problem are largely grounded on the Map Reduce framework (9) and its open source perpetration Hadoop. Although Hadoop handles the data volume challenge successfully. Hadoop is the open source software innovated by Apache and its Linux grounded software. It's used by notorious websites like Google, Yahoo, Facebook, Amazon and numerous further. Hadoop is a frame for recycling large quantum of data and provides better storehouse capacity for large datasets and performs resemblant processing of big data that gives better computational power to all the tasks. It works in batch processing mode and having two major factors HDFS (Hadoop Distributed Train System) (12) for huge data storehouse and Map Reduce for recycling huge quantum of datasets. When the data size is increased it produce problems to being algorithms to manage that so then main problem is to store and reuse that huge quantum of data and this problem is break by hadoop because it store and process huge quantum of data in lower time.

Hadoop:

Hadoop is an open-source software framework used for distributed storage and processing of big data using the Map Reduce 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.

a) Storage part Storage part of hadoop is HDFS (Hadoop distributed train system) which stores huge quantum of data with high degree of outturn and this huge data is stored in form of clusters.

b) Processing part of Hadoop is MapReduce which is a software frame which process large quantum of data in the form of clusters.

Hadoop distribute clusters to the knot so that they reuse parallel and this approach also takes advantage of data position This allows the dataset to be reused briskly and more efficiently which make it a more conventional supercomputer armature which work on a resembling train system where calculation and data are distributed via high- speed networking.

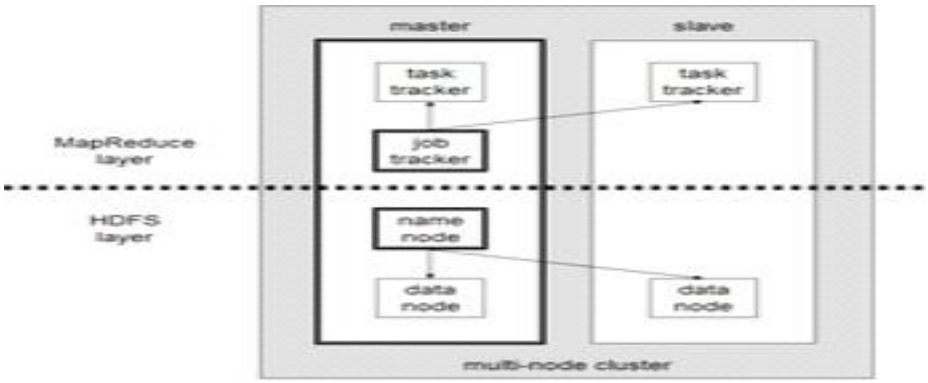

Fig.1.1.Hadop architecture

As depicted in Fig. 1.1, a tiny Hadoop cluster with a single master node and many slave nodes. A Task Tracker, Job Tracker, Name Node, and Data Node make up the master node [14], whilst the slave or worker node serves as both a Data Node and a Task Tracker.

1.2. Hadoop Distributed File System (HDFS): Hadoop Distributed File System (HDFS) is the storage component that stores a large number of structured, unstructured, and seminar-structured data. HDFS is a file system that is based on Java. HDFS is a dependable and easy-to-manage file system. Fast availability, load balancing, security, flexible access, fault tolerance, easy management, and high data throughputs are just a few of the benefits. It allows data to be processed in parallel. The architecture of HDFS is master/slave. [23]

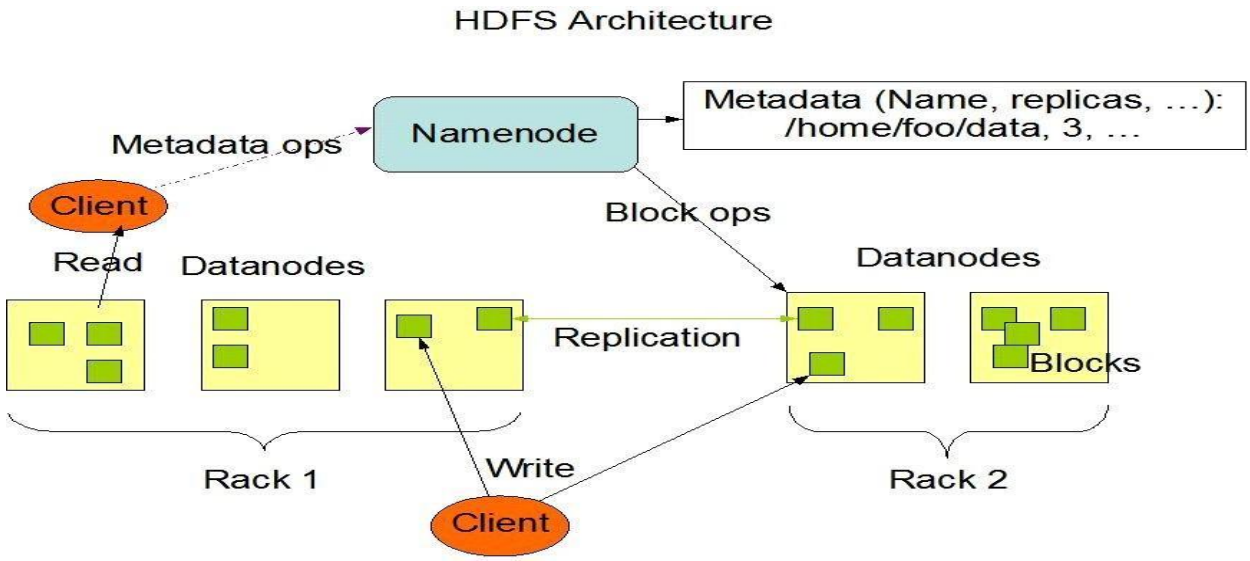

Fig. 1.2. HDFS Architecture

Affair. The reduce job takes the affair of the chart task as the input and combines them to lower set of tuples (reduces the large dataset into a lower one) grounded on the metamorphoses and color full logic. The advantage of Map Reduce is that it's easy to gauge data processing over multiple computing bumps.

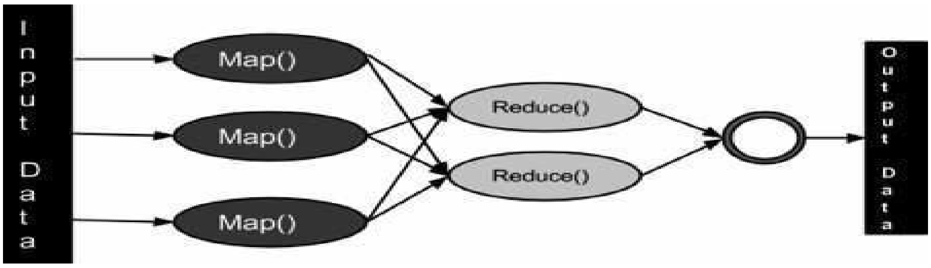

Fig. 1.3.MapReduce Architecture

The map stage's job is to process the input data in the manner indicated in Fig. 1.3. In most cases, the input data is stored in the Hadoop file system as a file or directory (HDFS). The map function receives the input file and processes it, creating numerous little bits of data.

Reducer stage: The Reducer stage's job is to process the data from the map stage. It generates a new set of output after processing, which is saved 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	[illegible]	UseDMFPMinetractsmultilevelassociationrules includinglevel-crossingforeachzone.
Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot, Nathalie Villa-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	Proposedframeworkthatintegratesinformationvisualization,scalablecomputing,anduserinterfacestoexplorelarge-scalemulti-modaldatastreamswhichcombinetorevealaneffectiveandefficientwaytoperformclosed-loopbigdataanalysiswithvisualizationandscalablecomputing.	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 Vs 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 Ooi, Kian-Lee Tan, Jun Xu. [2]	2016	Presents epiC, an extensible system to define the Big Data's data varietychallenge. Theyalso present the design and implementation of epiC's concurrent programming model and twostomized 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 datamanagement,modeldevelopment, visualizationand businessmodels.
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 hardwareacceleration and MapReduce execution flowcan 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 analyzethe existing work done on BDA applicationsinvolving the lifecycle of BDA applications and objects involving in BDA applicationsin 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 adoptes. 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 andtrust-basedservice selectionby developinga novel multi-objective optimizationapproachtomake trade-off decision betweenService's trust value and user's QoS preference towardscandidate.	Heterogeneous and trust-based selection by optimization approach.

Simon Fong, Raymond Wong,and Athanasios V. Vasilakos. [23]	2015	Presentedalgorithmstocollectbigdatawhichispresentinlargedegreeandtestitforperformanceevaluationbyusingacceleratedparticleswarmoptimization(APSO)typeofswarmsearchthatenhancedanalyticalaccuracywithinreasonableprocessingtime.	AccelerateParticleSwarm Optimization(APSO)algorithmstocollectbigdata.
Yanhao Huang and Xiaoxin Zhou. [22]	2015	Proposedthestructure,elements,basiccalculationsandmulti-dimensionalreasoningmethodofthenewknowledgemodel.Researchshowsmorepowerfulandadaptsvariousknowledgerequirementsofelectricpowerbigdata.	The knowledgemodelisestablishedandvariouscalculationisdone.
Marco Viceconti, Peter Hunter, and Rod Hose. [21]	2015	Proposedthatbiddataanalyticscansuccessfullycombinedwith VPHtechnologytogivedesirablemedicalsolutions.	Use VPHtechnologyandcombineditwithbigdataanalytics.
Alun Evans JaviAgenjoJosep Blat. [28]	2015	Presentedaweb-basedapplicationhavinganalyticvisualizationofon-setmediadataandmetadata,whichcombinesresearchfromseveralfieldsofimageprocessingand3Dgraphics.	Use WEBGL3Dinthe webandmeta data visualizationtechniques.
SyedAkhterHossain.[29]	2015	Describedthenascentfieldofbigdataanalyticsineducationwithdiscussiononprospectsandchallengeswayforward.Alsofocusonresearchanddevelopmentissuesforeducationistandpractitionersofbigdataanalytics.	Nascentfieldofbigdataanalyticsineducationanddevelopmentissuesforeducationistofbigdataanalyticsisdefined.
Xue-Wen Chen AND Xiaotong Lin. [30]	2014	Presentedoverviewof deeplearningandalsohighlightcurrentresearcheffortsandthechallengestobigdata,aswellasthefuturetrends.	Unprecedentedchallengestoharnessingdataandinformationispresented.
MatturdiBardi, ZhouXianwei, LI Shuai, LINFuhong. [32]	2014	ReviewedthevariousbenefitsandchallengesofsecurityandprivacyinBigDataandalsopresentedsomepossiblemethodsandtechniquestoensureBigDatasecurityandprivacy.	Bigdatasecurityandprivacytechnique isdefined.
SumanArora, Dr.MadhuGoel. [7]	2014	StudyandanalyzedvarioustechniquesofschedulingwhichenhancetheperformancebyusingHadoop.	Speculative executionandCopycomputesplittingtechnique ofHadoop.
Chang Liu, Jinjun Chen, Chi Yang, Rajiv Ranjan,and RamamohanaraoKotagiri. [16]	2014	Presentedtypesof fine-grainedidataupdatesandscheme that canfullysupportauthorizedauditingandfine-grainedupdaterequests.Also proposeanenhancement that canreducecommunicationoverheads forverifyingsmallupdates.	Describe ascheme forsupportingvariablesizeddata blocks.
ShifengFang,LiDaXu,YunqiangZhu,JiaerhengAhati,HuanPei,JianwuYan,andZhihuiLiu.[17]	2014	IntroducesanovelIISthatcombinesInternetofThings(IoT),CloudComputing,Geoinformatics,geographicalinformationsystem(GIS)ande-Scienceforenvironmentalmonitoringandmanagement,withacasestudyonclimatechangeanditsecologicaleffectsofaparticularregion.	Combine IoT,GISande-science forenvironmentalmonitoringandmanagement.
Daisuke Takaishi, Hiroki Nishiyamai, NeiKatoi and Ryu Miura. [18]	2014	Proposedanewmobilesinkroutinganddatagatheringmethodwiththehelpofnetworkclusteringbasedonmodifiedexpectationmaximizationtechnique.	UseEMalgorithmforclustering.
Andrea Marinoni, Arianna Dagliati, RiccardoBellazzi, Paolo Gamba1. [27]	2013	Providedstudyoftheconnectionbetweenairpollutionandclinicalrecords,thancorrelationsamongblackparticulateconcentration,microandmacro-vasculardiseasecanbedrawnproperly.	microandmacro-vasculardiseasecanbedrawnbycreatingconnectionbetweenvariousapproach.
XiongpaiQin, and Xiaoyun Zhou. [4]	2013	Reviewedlastseveralyearsbigdatabenchmarkworkandtheircharacteristicsareanalyzed.	Use MRBenchforevaluatingthe MapReduceframework.
R a k e s h V a r ma . [6]	2013	ObjectiveoftheresearchistostudyaboutMapReduceandvariousalgorithmsofschedulingwhichenhancetheschedulingperformance.	Formanaging BigData variousschedulingalgorithmsandLATE-Speculative executionisused.
Dan ie l Warn e ke . [1 5]	2011	DiscusstheopportunitiesandchallengesforparalleldataprocessingincloudsandpresentNephele.AndevaluatetheMapReduceprocessandcomparetheresultofframeworkHadoopdataprocessing.	Use Nephele,a newdata processingframework.
JasminAzemovic,Denis Music. [13]	2010	Presentedresearchonusingdifferentdatatypesforstoringunstructureddatawithindatabaseandthisresearchisinspiredwithcurrentsituationofinformationsociety.	Definevariouswayforstoringunstructureddata.
Mengjie Zhou,HaojiHu and Minqi Zhou. [14]	2010	ProposedaSLCA(SmallestLowestCommonAncestor)basedkeywordsearchimplementationforlarge-scaleXMLdatasetsonaMapReducecluster.	SLCAbasedkeywordsearchimplementationforlarge-scale data.
Leonardo Neumeyer,Bruce Robbins,Anish Nair,AnandKesari. [5]	2010	OutlinetheS4architectureanddescribeapplicationsofreal-lifedeployments.Theyinclude	Fordealingwithunboundedstreamofdata,S4architecture isused.
BI Shuoben, Xu Yin, JiaoFeng, LüGuonian, PEI Anping. [12]	2009	Introducesthesingle-dimensionalBooleanassociationruleonApriorialgorithm,andthedataminingalgorithmofthemulti-dimensionalassociationrulebasedonBUCalgorithm.	Single-dimensionalBooleanassociationrule onAprioralgorithmandBUCalgorithm.
Hui Fang, Ming Yang, Ruqing Yang. [11]	2007	Proposedaapproachtolocalizethevehiclepositionwithrespecttoaglobalmap,Itisbasedonthetextureofgroundfromwherethevehiclemoves.	Use dimensioningtechnique forlocalizingglobalmap.
SeemaMetikurke, Vijay K. Vaishnavi. [10]	2006	Describesagrid-enabledapproachforautomaticwebpageclassificationthatappliesthevectorspacemodelinformationretrievalstrategy.	Grid-enabledapproachforautomatic webpage classification.
John. H. Phan, Chang. F. Quo, and May D. Wang.[9]	2004	Reportingtheresultsofthefirstphasedevelopmentofnovelsystem,touseunsupervisedmethodsofclusteringtodiscoverrelationshipofgenesandknowledge-basedsupervisedclassificationisusedtogetaccuratepredictionincancerdiagnosis.	Useunsupervisedmethodsofclusteringtodiscoverrelationshipofgenesandknowledge-basedsupervised.
SushantGoel, HemaSharda, David Tanid. [8]	2003	Distributetheschedulingresponsibilitiestothenodeswheredataisactuallylocatedandalsoproposeanewserializabilitycriterion,ParallelDatabaseQuasi-Serializability.	Anewserializabilitycriterion,ParallelDatabase Quasi-Serializability(PDQS)isused.

3. CHALLENGES:

Because big data is such a large volume of data, it poses a number of obstacles in terms of management, storage, scheduling, security, and processing. For starters, effective online analysis necessitates data preparation, efficiently distributed storage, and search, all of which necessitate excellent data mining tools. The efficient processing of a large data stream is a major challenge that requires a variety of programming models. Second, the scheduling technique should be intelligent enough to respond in real time to a changing environment. Third, data integration necessitates the development of new protocols and interfaces capable of managing structured, semi-structured, and unstructured data. The fourth point is user interaction and visualization. There are numerous research problems in big data visualization, therefore more efficient real-time visualization techniques are required.

Furthermore, security and privacy are important considerations.

CONCLUSION:

In recent years, a survey of various big data techniques has been presented. It has been discovered that Big Data solutions are mostly built on the Map Reduce architecture and its open source implementation Hadoop. Hadoop successfully solves the data volume difficulty. For task scheduling in Hadoop, big data management encompasses a variety of tools, strategies, and algorithms. This paper is beneficial to a newcomer who want to pursue a profession in the field of big data.

FUTURE DIRECTION:

This work could be expanded by creating a new job scheduling algorithm that takes into account all of the variables that influence performance. Second, the user profile (similar people) and usage profile (invoked services) should be taken into consideration, as well as some related collaborative filtering strategies to combine with our service selection methodology.

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 DataMining 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, Hema Shardaand David Tanid,"Distributed Scheduler for High Performance Data-Centric Systems"b7803-76CI-XIO1lB17.00 0 2003IEEE.

[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] 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.

[15] Chang Liu, Jinjun Chen, Chi Yang, Rajiv Ranjan and Ramamohanarao Kotagiri,"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.

[16] 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.