DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5221

Volume 9, No. 1, January-February 2018

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

MINING BIG DATA FOR EFFICIENT DATA RETRIEVAL: A SURVEY

Sumit Vashishtha
PhD Scholar
MewarUniversity,Rajasthan

Dr. Pradeep Chouksey
Vice-Principal & Associate Professor
Computer Science, TIT, Bhopal,
MadyaPradesh

Dr. Vivek Tiwari
Assistant Professor
Computer Science, IIIT-NR, Raipur, C.G.

Abstract: The vast majority of work related to different data handling tasks is confronting a blast in the information that they need to gather and process so as to direct research. This is valid for both: investigative areas managing trial information, e.g. science, human science, stargazing and so on, additionally logical spaces managing reproduction information, e.g. seismology, material science and so on. To augment the potential result of investigative information examination, individual information administration applications need to satisfy the accompanying coarse undertakings: quick on-interest information handling, and compelling stockpiling and merging of various information accumulations. So handling the big data size along with the processing time is the major concern today. In this paper we have concentrated on big data challenges and analyzed the aspects. For this reason, we have studies and suggest the future possibilities.

Keywords: Big Data, Data Storage, processing time, Data Overhead

1. INTRODUCTION

We are inundated with a surge of information today. In a wide scope of utilization regions, information is being gathered at uncommon scale. Choices that already were in view of mystery, or on
carefully built models of reality, can now be made in view of the information itself [1]. Such Big Data examination now drives about every part of our current society, including portable administrations, retail, fabricating, money related administrations, life sciences, and physical sciences. Exploratory examination has been altered by Big Data [2-9]. In the other sciences, there is currently a well-established custom of storing experimental information into an open storehouse, furthermore of making open databases for utilization by different researchers [10]. Truth be told, there is a whole teach of bioinformatics that is to a great extent dedicated to the duration and investigation of such information [11]. As innovation advances, especially with the coming of Next Generation Sequencing, the size and number of trial information sets accessible is expanding exponentially [12].
Enormous Data can possibly alter research, as well as instruction [13]. Envision a world in which we have admittance to a tremendous database where we gather each definite measure of each understudy's scholarly execution. This information could be utilized to outline the best ways to deal with instruction, beginning from perusing, written work, and math, to propelled, school level, courses [14]. We are a long way from having entry to such information; however there are intense patterns in this bearing. Specifically, there is an in number pattern for gigantic Web organization of instructive exercises, and this will create an inexorably huge measure of definite information about understudies' execution and security [15-18].

Big data is depicted by the following four attributes [19-20]:

- **Volume:** Colossal information sets that are requests of size bigger than information oversaw in conventional stockpiling and investigative arrangements. Think petabytes rather than terabytes.
- **Variety:**Heterogeneous, complex, and variable information, which are created in arrangements as distinctive as email, online networking, feature, pictures, websites, and sensor information and in addition "shadow information, for example, access diaries and Web look histories.
- **Velocity:** Information is created as a consistent stream with constant inquiries for significant data to be served up on interest instead of grouped.
- **Value:** Significant bits of knowledge that convey prescient investigation for future patterns and examples from profound, complex examination in view of machine learning, measurable demonstrating, and diagram calculations. These investigations go past the aftereffects of customary business knowledge questioning and reporting.

We are inundated with a surge of information today. In a wide scope of utilization ranges, information is being gathered at extraordinary scale. Choices that already were in light of mystery, or on meticulously developed models of reality, can now be made taking into account the information itself. Such Big Data investigation now drives about every part of our advanced society, including versatile administrations, retail, producing, money related administrations, life sciences, and physical sciences. Prices related cloud and big data issues are also suggested in [21-24].

2. RELATED WORK

In 2012, Xin Luna Dong et al. [25] propose that Big Data time is upon us: information is being produced, gathered and broke down at an uncommon scale, and information driven choice making is clearing through all parts of society. Since the estimation of information blasts when it can be connected and combined with other information, tending to the huge information mix (BDI) test is basic to understanding the guarantee of Big Data. BDI contrasts from customary information reconciliation in numerous measurements: (i) the quantity of information sources, notwithstanding for a solitary space, has become in the several thousands, (ii) a large number of the information sources are exceptionally progressive, as an immense measure of recently gathered information are constantly made accessible, (iii) the information sources are to a great degree heterogeneous in their structure, with extensive mixture notwithstanding for generously comparative substances, and (iv) the information sources are of broadly varying qualities, with noteworthy contrasts in the scope, exactness and convenience of information gave.

In 2012, Aditya B. Patel et al. [26] reports the test chip away at enormous information issue and its ideal arrangement utilizing Hadoop group, Hadoop Distributed File System (HDFS) for capacity and utilizing parallel preparing to process expansive information sets utilizing Map Reduce programming structure. They have done model usage of Hadoop group, HDFS stockpiling and Map Reduce structure for preparing expansive information sets by considering model of huge information application situations. The outcomes acquired from different trials demonstrate great consequences of above way to deal with location enormous information issue.

In 2012, Rini T. Kaushik et al. [27] propose State-of-the-workmanship cooling vitality administration systems depend on warm mindful computational occupation arrangement/relocation and are naturally information position rationalist in nature. It takes a novel, information driven way to deal with diminish cooling vitality costs and to guarantee warm unwavering quality of the servers. T is conscious of the uneven warm profile and contrasts in warm unwavering quality driven burden limits of the servers, and the distinctions in the computational employments landing rate, size, and development life compasses of the Big Data put in the group. Taking into account this learning, and combined with its prescient record models and bits of knowledge, T does proactive, thermalaware document situation, which verifiably brings about thermalaware work position in the Big Data investigation register model. Assessment results with one-month long certifiable Big Data investigation creation follows from Yahoo! appear to 42% decrease in the cooling vitality costs with T politeness of its lower and more uniform warm profile and 9x preferable execution over the best in class information freethinker cooling strategies.

In 2012, EdmonBegoli et al. [28] recommend that Big information wonder alludes to the act of gathering and preparing of expansive information sets and related frameworks and calculations used to examine these gigantic datasets. Architectures for huge information generally extend over different machines and groups, and they normally comprise of various extraordinary reason sub-frameworks. Combined with the information disclosure process, huge information development offers numerous one of kind open doors for associations to advantage (as for new bits of knowledge, business improvements, and so forth.). In any case, because of the trouble of investigating such extensive datasets, huge information presents one of kind frameworks designing and structural difficulties. They introduce three framework plan rule that can educate associations on compelling logical and information gathering procedures, framework association, and information spread practices. The standards introduced get from our own innovative work encounters with enormous information issues from different government organizations, and they delineate every rule with our own particular encounters and suggestions.

In 2012, Gueyoung Jung et al. [29] address the previously stated tradeoff, to decides: (a) what number of and which figuring hubs in united mists ought to be utilized for parallel execution of enormous information examination; (b) entrepreneurial allotting of huge information to these processing hubs in a manner to empower synchronized finishing, best case scenario exertion execution; and (c) arrangement of allocated, diverse sizes of huge information pieces to be registered in every hub so that exchange of a lump is covered however much as could reasonably be expected with the reckoning of the past lump in the hub. They proposed Overlapped Bin-pressing driven Bursting (MOBB) calculation, which enhance the execution by up to 60% against existing methodologies.

In 2013, Antonia Azzini et al. [30] infer that every nearby source is tasked of applying a semantic lifting method for communicating the neighborhood information in term of the normal model. Semantic heterogeneity is then conceivably presented in information. They show a procedure intended to the execution of reliable procedure mining calculations in a 'Major Data' connection. Specifically, we misuse two unique strategies. The first is gone for registering the befuddle among the information sources to be coordinated. The second uses bungle qualities to stretch out information to be handled with a conventional guide diminish calculation.

In 2013, Du Zhang et al. [31] first examine the measurements in huge information and enormous information examination, and afterward center our consideration on the issue of irregularities in huge information and the effect of irregularities in huge information investigation. They offer characterizations of four sorts of irregularities in huge information and call attention to the utility of irregularity instigated adapting as an apparatus for huge information examination.

In 2013, Xin Cheng et al. [32] recommend an information advancement model of Virtual DataSpace (VDS) for dealing with the huge information lifecycle. Firstly, the idea of information development cycle is characterized, and the lifecycle procedure of huge information administration is depicted. In view of these, the information development lifecycle is broke down from the information relationship, the client necessities, and the operation conduct. Also, the characterization and key ideas about the information advancement procedure are depicted in point of interest. As indicated by this, the information development model is built by characterizing the related ideas and breaking down the information relationship in VDS, for the catch and

following of element information in the information advancement cycle. At that point they examine the expense issue about information spread and change. At last, as the application case, the administration procedure of element information in the field of materials science is portrayed and broke down.

In 2013, SerefSagiroglu et al. [33] recommend the procedure of exploration into huge measures of information to uncover shrouded examples and mystery relationships named as large information investigation. These valuable data's for organizations or associations with the assistance of increasing wealthier and more profound bits of knowledge and getting preference over the opposition. Thus, huge information usage should be investigated and executed as precisely as could be expected under the circumstances. They displays a review of huge information's substance, degree, tests, systems, focal points and challenges and talks about security concern on it.

In 2014, Xindong Wu et al. [34] present a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. They analyze the challenging issues in the data-driven model and also in the Big Data revolution.

3. PROBLEM IDENTIFICATION

After studying and analyzing several research papers in this field, we find some major security concerns which are following:

- The amount of raw data is increasing exponentially so the sheer volume and capacity increase the complexity and this situation is calling for new approach.
- Drowning of data but starved for knowledge.
- Need of Trailblazing. The data analysis and based on the analysis decision making is the bigger challenge.
- Because the present innovation empowers us to effectively store and inquiry huge datasets, the attention is currently on procedures that make utilization of the complete information set, rather than inspecting. This has enormous ramifications in territories like machine learning, example acknowledgment and order, to give some examples. Hence, there are various prerequisites for moving past standard information mining systems:

  Strong logical establishment to have the capacity to choose a satisfactory technique or configuration
  Another calculation (and demonstrate its proficiency and versatility, and so forth.)
  An innovation stage and satisfactory improvement aptitudes to have the capacity to actualize it;
  A capacity to see not just the information structure (and the convenience for a given preparing technique), additionally the business esteem. Subsequently, assembling multi-disciplinary groups of "Information researchers" is frequently a crucial method for picking up.

- Due to the increasing mobility of users and devices, context-awareness increases in importance. A suitable and efficient content- and context-aware routing of data is needed in many cases. Facing existing infrastructures and Big Data setups many solutions focus on processing and routing all data at once. For example, in manufacturing existing data has no relation to the context about the user´s history, location, tasks, habits schedule, etc. Concepts for taking the spatial users into account are a major challenge. The goal is to take the context into account for data that is not related to a user or context and present the right data to the right people and devices. Applying contextual awareness can thus be a suitable approach to improve the quality of existing problem solving. In the context of Big Data, contextualization can be an attractive paradigm to combine heterogeneous data streams to improve quality of a mining process or classifier.
- When people devour data, a lot of heterogeneity is serenely endured. Actually, the subtlety and lavishness of common dialect can give important profundity. On the other hand, machine examination calculations expect homogeneous information, and can't comprehend subtlety. In result, information must be painstakingly organized as an initial phase in (or preceding) information investigation.
- The flip side of size is speed. The bigger the information set to be prepared, the more it will take to break down. The configuration of a framework that adequately manages size is likely likewise to result in a framework that can handle a given size of information set speedier. In any case, it is not simply this speed that is generally implied when one discusses Velocity in the setting of Big Data.
- Companies today as of now utilize, and value the estimation of, business insight. Business information is broke down for some reasons: an organization may perform framework log investigation and online networking examination for danger appraisal, client maintenance and brand administration, etc.

4. DISCUSSION AND ANALYSIS

Data Collection Process

Now it's time to look at data collection. Where do you get your data? Any and all sources can be useful to the data mining process. Internal data sources such as web site hits, prospect lists, custom surveys, and old customer data records are all valuable. External data sources such as purchased lists or panel lists are also relevant. Figure 1 shows the statistical changes in data scaling, process time and scale.

Data Preparation

Before raw data can be subjected to statistical analysis, it must be converted into a form suitable for analysis. Data preparation includes editing, coding, data cleaning (consistency checks and missing responses) and statistically adjusting the data (weighting and variable specification). Check your database thoroughly, and be sure to eliminate duplicate records and cases. Figure 2 shows the base line is better.

Data Analysis

Now that the data has been cleaned and formatted, it's time to analyze the data and answer study objectives. A number of statistical techniques can be employed. Statistical techniques can be classified as univariate or multivariate. In some cases, basic statistics may be enough. Frequencies (histograms), means and medians can often tell you a lot. Data reduction, segmentation and modeling techniques may also help. As shown in figure 3 the number of active sources increased steadily but slowly until the end of 2011. But in 2012 and 2013 it is increased in exponential rate.

Table 1: Result Analysis

S.no	Authors	Year	Work	Gap
1	Demchenko et al. [35]	2014	Their work means to give a combined perspective of the Big Data phenomena and related difficulties to present day advancements.	Complex Architecture.
2	Wang et al. [36]	2014	Their suite-BigDataBench covers wide application situations, as well as incorporates various and agent information sets.	Static data Inputs.
3	Pandey et al. [37]	2014	Their investigation of execution components of MapReduce demonstrates that end of their reverse impact by streamlining enhances the execution of Map Reduce.	It can be extended to multi-layer scheme..
4	Hu et al. [38]	2014	They have done the survey on different big data challenges and categorization.	Practical implication is bit complex.
5	Venaik et al. [39]	2015	She has suggested and information pyramid for evaluation and analysis.	Reengineering process in not clear.
6	Sampada et al. [40][41]	2013	They apply Chi-Square test, to test the hypothesis for correctness. The program capability was based on three parameters; first is F-measure (FM), second is odds ratio (OR) and third is power (PO).	Only software metrics are checked
7	Leung et al. [42]	2014	Their approach enormously diminishes the quest space for Big information mining of dubious information, and returns just those examples that are intriguing to the clients for Big information investigation.	Data Storage can be included as the external parameter.

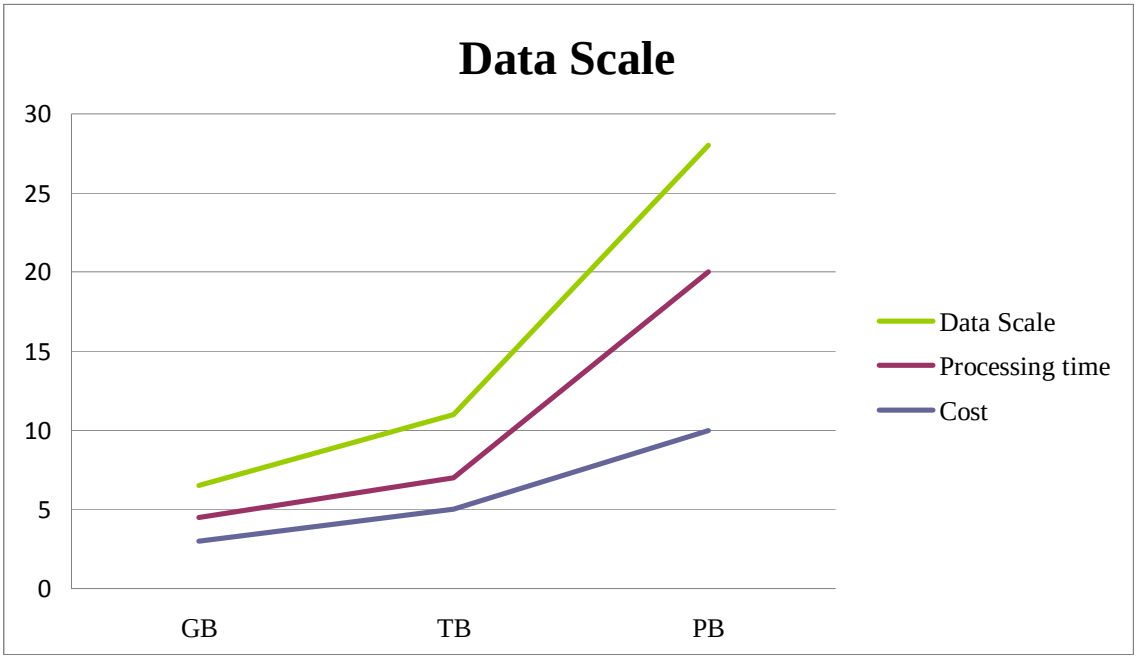

Figure 1: Data Collection Process

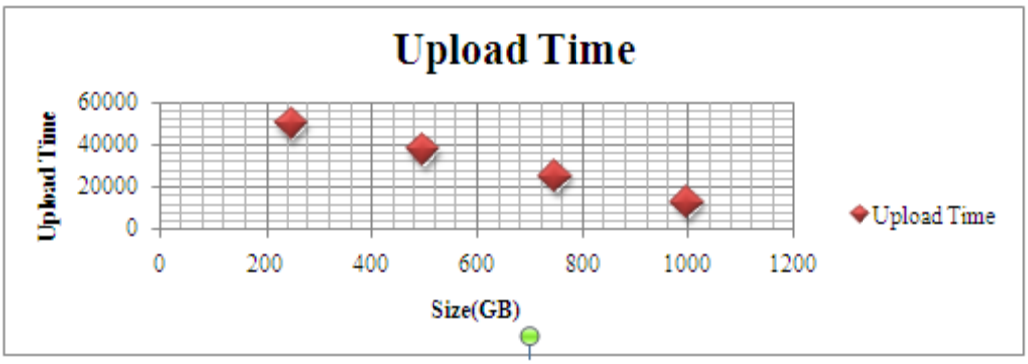

Figure 2: Data Preparation

Figure 3: Data Active Sources [according to Recorded future]

5. CONCLUSIONS AND FUTURE DIRECTIONS

Efficient data handling, data processing and securityare the major research area in the big data.In this paper we have reviewed and analyzed several previous research works and suggest data mining and evolutionary algorithms along with standard encryption techniques to secure the data in the big data handlingenvironment.So our paper guideline mean is to think about and investigates the benefits and discovering the crevice.

The future suggestions are following:

- In today's hypercompetitive business environment, organizations not just have to discover and dissect the applicable information they need, they must discover it rapidly. Representation helps associations perform examinations and settle on choices much all the more quickly, yet the test is going through the sheer volumes of information and getting to the level of point of interest required, all at a rapid.
- Support for standardization and collaboration in software and services technologies.
- Key management, security and key provisioning can be allowed at the end of the big data.
- It takes a considerable measure of comprehension to get information fit as a fiddle with the goal that you can use perception as a feature of information examination.

REFERENCES

[1] B. Ratner, Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data, CRC Press, 2003.
[2] T. Craig and M.E. Ludlof, Privacy and Big Data: The Players, Regulators, and Stakeholders, O'Reilly Media, 2011.
[3] R. Clarke, "Human Identification in Information Systems: Management Challenges and Public Policy Issues," Information Technology & People, vol. 7, no. 4, 1994, pp. 6-37.
[4] M.R. Wigan, "Owning Identity-One or Many-Do We Have a Choice?" IEEE Technology and Society, vol. 29,no. 2, 2010, pp. 33-38.
[5] E.W.T. Ngai, L. Xiu, and D.C.K. Chau, "Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification," Expert Systems with Applications, vol. 36, no. 2, 2009, pp. 2592-2602.
[6] Ruchita Gupta, C.S.Satsangi, " An Efficient Range Partitioning Method for Finding Frequent Patterns from Huge Database " , International Journal of Advanced

Computer Research (IJACR), Volume-2, Issue-4, June-2012 ,pp.62-69.
[7] Z. Bellahsene, A. Bonifati, and E. Rahm, editors. Schema Matching and Mapping. Springer, 2011.
[8] J. Bleiholder and F. Naumann. Data fusion. ACM Computing Surveys, 41(1):1–41, 2008.
[9] M. J. Cafarella and A. Y. Halevy. Web data management. In Sigmod, pages 1199–1200, 2011.
[10] M. J. Cafarella, A. Y. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web. In PVLDB, pages 538–549, 2008.
[11] Debopam De, Deblina Banerjee, Sneha Mukherjee and JayatiGhoshDastidar, " A Simplistic Mechanism for Query Cost Optimization " , International Journal of Advanced Computer Research (IJACR), Volume-5, Issue-19, June-2015 ,pp.205-211.
[12] K. C.-C. Chang, B. He, and Z. Zhang. Toward large scale integration: Building a metaquerier over databases on the web. In CIDR, pages 44–55, 2005.
[13] N. N. Dalvi, A. Machanavajjhala, and B. Pang. An analysis of structured data on the web. PVLDB, 5(7):680–691, 2012.
[14] U.Chandrasekhar, Sandeep Kumar. K, Yakkala Uma Mahesh, " A Survey of latest Algorithms for Frequent Itemset Mining in Data Stream" , International Journal of Advanced Computer Research (IJACR), Volume-3, Issue-9, March-2013 ,pp.60-65.
[15] X. Dong and A. Y. Halevy. Indexing dataspaces. In C. Y. Chan, B. C. Ooi, and A. Zhou, editors, SIGMOD Conference, pages 43–54. ACM, 2007.
[16] X. Dong, A. Y. Halevy, and C. Yu. Data integration with uncertainties. In VLDB, 2007.
[17] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. PVLDB, 2(1), 2009.
[18] Ramesh R and Divya G, “Dynamic Security Architecture among E-Commerce Websites”, International Journal of Advanced Computer Research (IJACR), Volume-5, Issue-19, June-2015, pp.184-191.
[19] X. L. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. PVLDB, 2(1), 2009.
[20] X. L. Dong and F. Naumann. Data fusion–resolving data conflicts for integration. PVLDB, 2009.
[21] Soni A, Hasan M. Pricing schemes in cloud computing: a review. International Journal of Advanced Computer Research. 2017 Mar 1;7(29):60.
[22] Lasheng Y, Xu W, Yu Y. Research on visualization methods of online education data based on IDL and hadoop. International Journal of Advanced Computer Research. 2017 Jul 1;7(31):136.
[23] NaikMR, Sathyanarayana SV. Key management infrastructure in cloud computing environment-a survey. ACCENTS Transactions on Information Security. 2017; 2(7): 52-61.
[24] Shobha K, Nickolas S. Time domain attribute based encryption for big data access control in cloud environment. Transactions on Information Security. 2017; 2(7):73-77.
[25] Dong, X.L.; Srivastava, D., "Big data integration," Data Engineering (ICDE), 2013 IEEE 29th International Conference on , vol., no., pp.1245,1248, 8-12 April 2013.
[26] Patel, A.B.; Birla, M.; Nair, U., "Addressing big data problem using Hadoop and Map Reduce," Engineering (NUiCONE), 2012 Nirma University International Conference on , pp.1,5, 6-8 Dec. 2012.
[27] Kaushik, R.T.; Nahrstedt, K., "T*: A data-centric cooling energy costs reduction approach for Big Data analytics cloud," High Performance Computing, Networking, Storage and Analysis (SC), 2012 International Conference for , app.1,11, 10-16 Nov. 2012.
[28] Begoli, E.; Horey, J., "Design Principles for Effective Knowledge Discovery from Big Data," Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on , pp.215,218, 20-24 Aug. 2012.
[29] Gueyoung Jung; Gnanasambandam, N.; Mukherjee, T., "Synchronous Parallel Processing of Big-Data Analytics Services to Optimize Performance in Federated Clouds," Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on, pp.811,818, 24-29 June 2012.
[30] Azzini, A.; Ceravolo, P., "Consistent Process Mining over Big Data Triple Stores," Big Data (BigData Congress), 2013 IEEE International Congress on, pp.54, 61, June 27 2013-July 2 2013.
[31] Du Zhang, "Inconsistencies in big data," Cognitive Informatics & Cognitive Computing (ICCI*CC), 2013 12th IEEE International Conference on, pp.61, 67, 16-18 July 2013.
[32] Xin Cheng; Chungjin Hu; Yang Li; Wei Lin; HaoleiZuo, "Data Evolution Analysis of Virtual DataSpace for Managing the Big Data Lifecycle," Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2013 IEEE 27th International, pp.2054,2063, 20-24 May 2013.
[33] Sagiroglu, S.; Sinanc, D., "Big data: A review," Collaboration Technologies and Systems (CTS), 2013 International Conference on, pp.42,47, 20-24 May 2013.
[34] Xindong Wu; Xingquan Zhu; Gong-Qing Wu; Wei Ding, "Data mining with big data," Knowledge and Data Engineering, IEEE Transactions on , vol.26, no.1, pp.97,107, Jan. 2014.
[35] Demchenko, Y.; De Laat, C.; Membrey, P., "Defining architecture components of the Big Data Ecosystem," Collaboration Technologies and Systems (CTS), 2014 International Conference on, pp.104, 112, 19-23 May 2014.
[36] Lei Wang; Jianfeng Zhan; ChunjieLuo; Yuqing Zhu; Qiang Yang; Yongqiang He; WanlingGao; Zhen Jia; Yingjie Shi; Shujie Zhang; Chen Zheng; Gang Lu; Zhan, K.; Xiaona Li; BizhuQiu, "BigDataBench: A big data benchmark suite from internet services," High Performance Computer Architecture (HPCA), 2014 IEEE 20th International Symposium on , pp.488,499, 15-19 Feb. 2014.
[37] Pandey, S.; Tokekar, V., "Prominence of MapReduce in Big Data Processing," Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on , vol., no., pp.555,560, 7-9 April 2014.
[38] Han Hu; Yonggang Wen; Tat-Seng Chua; Xuelong Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," Access, IEEE, vol.2, pp.652, 687, 2014.
[39] Anita Venaik , " Qualitative risk level estimation of Business Process Re-engineering efforts and effects (With special reference to IT-sector) " , International Journal of Advanced Computer Research (IJACR), Volume-5, Issue-18, March-2015 ,pp.11-18.
[40] SampadaKembhavi, Ravindra Gupta, Gajendra Singh, " An Efficient Algorithm for Auto Upload and Chi-Square Test on Application Software " , International Journal of Advanced Computer Research (IJACR), Volume-3, Issue-10, June-2013 ,pp.121-125.
[41] SampadaKembhavi and Gajendra Singh, " Auto Upload and Chi-Square Test on Application Software as a Service for Cloud Computing Environment " ,

International Journal of Advanced Technology and Engineering Exploration (IJATEE), Volume-1, Issue-1, December-2014 ,pp.26-31.

[42] Leung, C.K.-S.; MacKinnon, R.K.; Fan Jiang, "Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data," Big Data (BigData Congress), 2014 IEEE International Congress on , vol., no., pp.315,322, June 27 2014-July