AN IMPROVEMENT IN PERFORMANCE BY USING CLUSTERED APPROACH DURING BIG DATA PROCESSING

Ms. Archana, Research Scholar, Department of Computer Science, Singhania University, Rajasthan.
Dr. Gaurav Aggarwal, Research Supervisor, Department of Computer Science, Singhania University, Rajasthan.

Abstract:

The rapid growth of Web in the last decade is growing many folds, with the usage of smart phones and internet, the web is accessible to even remote areas of the world. The rise of e-commerce sites and social media generates lots of information on daily basis. The information can reach up-to terabytes or petabytes. The owners of e-commerce sites, business house, governments and many other agencies or individuals need this information to be analyzed in real time for profit or non-profit purposes. The information generated in web is in all form structured, semi-structured and unstructured formats. Web mining is used to get the useful information from web hyperlinks, page content and usage data. Web mining can be categorized into three types: Web structure mining, Web content mining and Web usage mining.

The research paper is focusing on enhancement of security along with performance issues in big data processing. Proposed work has considered secure clustered approach. It has been observed that proposed research is providing solution for man in middle, brute force and denial of service attack. Research work considered data encryption approach along with clustering to provide better security. On other hand clustering allows better performance by dividing data processing in different clusters. In this way proposed research work is providing secure, high performance and reliable solution.

Keyboard: Big Data, Clustering, Map-Reduce, Security, Performance

Introduction

Data Mining is "the computational cycle of finding designs in enormous data sets" for the objective of "extricating data from a data set and change it's anything but a reasonable construction for additional utilization". It additionally characterized as "the interaction of consequently finding valuable data in enormous data stores". It's anything but a most recent strategy having incredible possibilities to help the applications that investigate the main realities in their data distribution centers. It can likewise separate concealed data from enormous database. The devices in data mining predicts future patterns and practices, making organizations to make proactive, settle on knowledge driven choices and so forth It likewise offered mechanized, forthcoming examinations which move past the investigations of previous occasions given by review devices commonplace choice emotionally supportive networks. It's anything but a cycle to transform crude data into data. Programming can be utilized for designs in enormous bunches of data, to think about clients in the event of business, to foster powerful showcasing procedures, to expand deals and reduction costs. It relies upon compelling data assortment, putting away data in warehousing just as handling. It comprises of calculations and computational ideal models that permit systems to discover designs perform expectation, guaging the future occasions, improves execution by communicating with the data. The KDD cycle in data mining incorporates data determination, cleaning, coding, design acknowledgment, AI techniques, detailing and perception of the created structures.Data mining is the process of posing queries and extracting patterns, often previously unknown from large quantities of data using pattern matching or other reasoning techniques. Cyber security is the area that deals with cyber terrorism. We are hearing that cyber attacks will cause corporations billions of dollars. For example, one could masquerade as a legitimate user and

swindle say a bank of billions of dollars. Data mining may be used to detect and possibly prevent security attacks including cyber attacks. For example, anomaly detection techniques could be used to detect unusual patterns and behaviors. Link analysis may be used to trace the viruses to the perpetrators. Classification may be used to group various cyber attacks and then use the profiles to detect an attack when it occurs. Prediction of attack structure may be used to determine potential future attacks depending in a way on information learnt about terrorists through email and phone conversations. Also, for some threats non real-time data mining may suffice while for certain other threats such as for network intrusions we may need real-time data mining. Many researchers are investigating the use of data mining for intrusion detection. While we need some form of real-time data mining, that is, the results have to be generated in real-time, we also need to build models in real-time. For example, credit card fraud detection is a form of real-time processing. However, here models are built ahead of time. Building models in real-time remains a challenge. Data mining can also be used for analyzing web logs as well as analyzing the audit trails. Based on the results of the data mining tool, one can then determine whether any unauthorized intrusions have occurred and/or whether any unauthorized queries have been posed.

Data Mining Techniques

Association Rule: To discover and represent interesting relation among variable in big database, there is a popular technique in data mining is association rule learning. It is indicated to find out strong rules discovered in databases with the help of different interestingness measures. Introduced association rules for finding or discovering products regularities on large-scale transaction data recorded set by point-of-sale systems in supermarkets which are based on the strong rules concept. An example is found dependent on a connection between things in a similar exchange. Thus this method is otherwise called connection procedure. Eg. Market container analysis.

Classification - This strategy depends on AI. It is utilized to arrange everything in a bunch of data into one of predefined set of classes or gatherings. Extortion detection and credit hazard applications are guides to this kind of strategy. It is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

Clustering - It bunches group of articles which have comparable attributes. This procedure characterized the classes and places objects in each class, while in the characterization method; objects are allocated into predefined classes. Clustering or Cluster Analysis as it is widely known is a focused type of data mining technique for large scale analysis of datasets. Cluster Analysis is a pattern discovery procedure, whose goal is to discover patterns in a set of data. It identifies clusters in a set of data and builds a typology of sets using a certain set of data. In the present research analysis clustering technique is applied it is well known unsupervised data mining technique. It is particularly useful where there are many cases and no obvious natural grouping. Here, clustering data mining algorithm can be used to find whatever grouping may exist. A cluster is a collection of data objects that are similar in some sense to one another. A good clustering method produces high quality cluster to insure that the inter-cluster similarity is low and the intra-cluster similarity is high; in other words, members of a cluster are more like to each other than they are like members of different clusters.

Regression - It tends to be adjusted for predication. Relapse analysis can be utilized to demonstrate the connection between at least one autonomous factors and ward factors. Definitely referred to factors are called as Independent factors and what the client need to anticipate is called as reaction factors. Regression is a data mining technique used to predict a range of numeric values (also called *continuous values*), given a particular dataset. For example, regression might be used to predict the cost of a product or service, given other variables.

Prediction - It is utilized to foresee the future result. It likewise finds the connection among reliant and autonomous factors.

Sequential Patterns - It will in general find or recognize comparative examples, standard occasions or patterns in exchange data over a business period.

Decision trees - In this method the base of the choice tree is a straightforward inquiry or condition that has different answers. In light of each answer this strategy prompts set of inquiries or conditions will assist with determining the data so an official choice can be made.

Data that must be kept private and secure must be kept on the network. Customers appreciate a company's reputation highly. Therefore, network provider must install sufficient security measures in order to secure data of its consumers. Customers are also responsible for their own security and should exercise the same care that we did by choosing a strong password and keeping it a secret. In cases when the data is beyond the firewall, the network may be able to help. Security precautions are referred to as "big data security" in the context of analytics and data operations. Data theft, DDoS attacks, ransomware, and other forms of malicious software are some of the threats facing big data platforms. In the event of a robbery, credit card information and other sensitive information may be more vulnerable. If you violate the GDPR's core data security and privacy protection regulations, you risk receiving a fine. Its privacy and security the emergence of big data in the digital age is outpacing conventional methods. It is feasible to hack data that has been de-identified and connected to a specific person, even if the data has been encrypted and protected by access limits and network security IDS. Big data privacy issues, such as inference and aggregation, which enable people to be re-identified even after their personal information has been erased from the dataset, are being addressed via the advocacy of new legislation. Our current concern is the "security triangle," a problem that has troubled us for years. Users have a detrimental impact on the functionality and usefulness of the system, for instance, if a law forbids access to raw data analysis and change. The safety and privacy standards for the whole Big Data ecosystem, from infrastructure and administration to data integrity and quality to confidentiality requirements, need to be reviewed. The area of big data security and privacy needs a lot of work. Security measures must be in place to safeguard Big Data technologies such infrastructure, monitoring and auditing processes, apps, and data sources.

Need of research

Massive amounts of data can only be stored, analysed, and handled utilising a group of "Big Data" technologies. This app was built with the aid of a map as a guide. In comparison to everything else I've tested, it performs a superior job of sorting and filtering. If you ask me, it is capable of doing summative acts. Through the use of big data, businesses gain from the creation of useful data. Organizations utilise big data to enhance their marketing objectives and tactics. Predictive modeling and other applications may be useful for machine learning projects. A vast category of information that cannot be mathematically specified is referred to as "big data." It has been seen that when security is improved, large data processing system speed suffers. When processing massive data, security and speed need to be improved.

Literature Review

Chunhua et al., (2013) proposed a novel induction dependent on improvement for the halfway least squares (PLS) calculation. It shows that just one of either the X-or the Y-grid should be emptied during the sequential cycle of processing idle elements. An improved recursive dramatically weighted PLS relapse calculation was proposed.

Bogomolov et al., (2014) proposed a novel technique to expect crime in a geographic region from more than one data sources, chiefly cell phone and segment records. The guideline commitment of the proposed method lies in the use of amassed and anonymzed human conduct records got from versatile local area interest to handle the crime forecast inconvenience. Indeed, even as past investigations endeavors have utilized both history old knowledge and guilty parties profiling, our discoveries help the hypothesis that amassed human conduct data caught from the cell local area foundation, in total with principal segment records, can be utilized to anticipate crime.

Zhang and Zhang (2014) proposed an altered incomplete least-squares (PLS) relapse displaying technique to assemble a changed relapse model. The proposed model extricated the helpful data in lingering subspace to anticipate the yield factors productively. Thusly more precise quality factors were anticipated. The proposed adjusted PLS strategy and the ordinary PLS technique is likewise applied during the time spent Pencillin aging to show that the proposed strategy was more successful than the regular PLS technique.

Vogelsang and Wagner (2014) proposed a novel assessment strategy dependent on an expanded incomplete total (mix) change of the relapse model. The new assessor was named as Integrated Modified Ordinary Least Squares (IM-OLS). IM-OLS doesn't need assessment of since quite a while ago run fluctuation frameworks and dodges the need to pick tuning boundaries (pieces, transfer speeds, slacks). Deduction wants that a since quite a while ago run fluctuation be scaled out, and propose customary and fixed-b techniques for getting basic qualities for test measurements.

Mukaka et al., (2016) examined about an adaptation that neglects to join, this problem can be tended to most adequately the utilization of both Cheung's OLS approach or via turning into a twofold relapse model with the glm2 bundle in R; every one of these techniques merge and give proficient impartial appraisals of changed possibility varieties. The Cheung's OLS approach offers a measurable protection property over the glm2 strategy. The Poisson variant with character interface and solid favored mistakes fitted the utilization of the glm set of rules in R and the Additive Binomial Regression adaptation equipped the utilization of the blm calculation in R are potential options as they have 100% and practically 100% intermingling rates individually, anyway both produce somewhat one-sided danger differentiation gauges. Also the Poisson form has too high protection while the blm approach produces variable measurable inclusion relying upon adequacy situations.

Mashinchi et al., (2016) proposed a relapse strategy wants to fit the bend on a data set independent of anomalies. This paper adjusts the granular box relapse ways to deal with manage data sets with anomalies. Each approach fuses a three-stage method incorporates granular box setup, anomaly disposal, and straight relapse analysis. The main stage researches two target works each applies distinctive punishment conspires on boxes or cases. The subsequent stage explores two techniques for anomaly end to, then, at that point, plays out the straight relapse in the third stage. The exhibition of the proposed granular box relapses are explored as far as: volume of boxes, obtuseness of boxes to anomalies, passed time for box arrangement, and mistake of relapse. The proposed approach offers a superior straight model, with more modest blunder, on the given data sets containing assortments of anomaly rates.

G.V. Nadiammai, M. Hemalatha has proposed the integrated Data mining system to identify the relevant and hidden data. In this study four issues such as classification of data, high level human interaction, lack of labeled data and effectiveness of distributed Denial of Service Attack by using KDD 99 dataset and various Clustering algorithms performance is measured.

Kirti Jain “An Integrated Approach to Detect Network Attacks Using Data Mining” In this work clustering technique of data mining is used to detect the attack rule structure. In cluster analysis, data objects are clustered together based on the data and relationship among them. Various clustering algorithms such as simple K-Means, Farthest–First, Filtered cluster, Make density algorithms of clustering are used different attack structure are represented by using KDD 1999 dataset. The performance of different algorithms is compared and clustered instances formed by various algorithms are recorded. The architecture of proposed system is to find the solution of above problems, Intrusion detection can allow for the prevention of certainty, attacks severity relative to different type of attack and vulnerability of components under attack the response may be kill the connection, install filtering rules, and disable user account.

Problem Statement and Challenges in Data Mining:

Data mining is the analysis of observational dataset to find unsuspected relationship and to summarize large amounts of data in novel ways that are both understandable and useful to data owner in proactive decision making. Data Mining is now possible due

to advances in computer science and machine learning. Data Mining delivers new algorithms that can automatically sift deep into your data at the individual record level to discover patterns, relationships, factors, clusters, associations, profiles, and predictions—that were previously "hidden". Using normal reports, Data mining can produce decisions and create alerts when action is required. Data Mining is being widely used in various fields, such as in business for Customer Relationship Management, Marketing, etc, in medicine for laboratory research, clinical trials, pharmacology, etc, in forecasting of weather, traffic, etc, in aviation for pilot assistance and in research in the areas of astrophysics, medicine, business, security, etc. In order to apply the techniques to information security we needed datasets. We used a commonly applied dataset in information security research.

Since, the network administrator feels difficult to pre-process the data. Due to the overwhelming growth of attacks which makes the task hard, attacks can be identified only after it happens. To overcome this situation, frequent updating of profiles is needed. Reduced workload of administrator increases the detection of attacks. Data mining includes many different algorithms to accomplish the desired tasks. All of these algorithms aims to fit a model to the prescribed data and even analyzes the data and simulate a model which is closest to the data being analyzed.Numerous researches in the area of large data have endorsed clustering. While some researchers advocated utilizing big data platforms to detect malware, others concentrated on data clustering to create intelligent Internet of Things (IoT) solutions. It is essential to provide a more secure clustering approach in order to safeguard Big Data while it is transferred across networks. Both managing and not managing vast volumes of data are options. The results of this study show that research on unmanaged huge data has been limited. Despite the map reduction environment's extensive use in big data research, these strategies were not the best for handling significant volumes of structured data. Additionally, the clustering techniques used in the earlier work are less efficient. Big data management needs to be better safeguarded from cyber attacks, according to a number of studies. Performance is still a problem, too.

Proposed Research Methodology

In this proposed work, research is getting input dataset of queries and keywords. Form this dataset; research gets the queries and keywords. Research is applying Map phase and reduce phase on that Queries and keywords. Now, work gets the frequency count of Keywords and collects the dataset of frequency. Research enhances K-mean clustering and gets the multiple clusters. Last, research is encrypting the clustered data and gets secured the data.

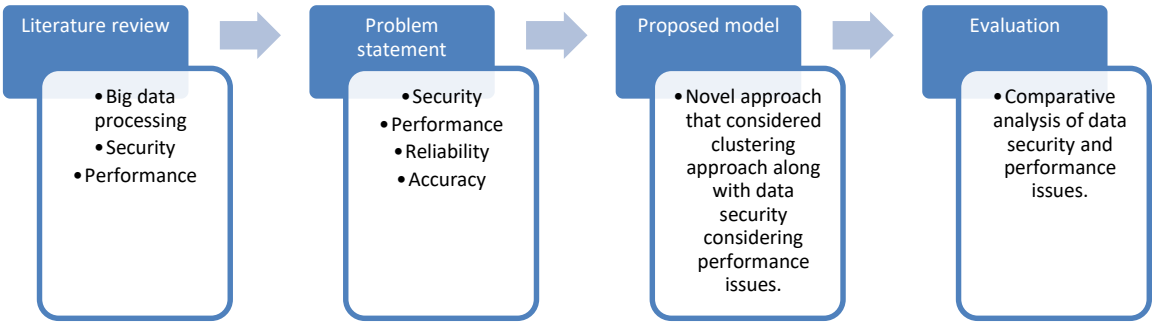

Fig 1 Proposed Research Methodology

Result and discussion

During simulation, the time consumption in the case of previous work and the case of proposed work is noted according to a different number of packets. Simulation work has been performed in a MATLAB environment.

TIME CONSUMPTION

Time taken has been simulated in the case of the proposed system in comparison to previous and proposed research is shown in figure 5.1. Moreover previous researches have not compressed the data before transmission. Thus the time consumption is evidently less as compared to others due to the smaller size of the data packets.

Table 1 Comparative analysis of Time consumption

File	Conventional work	Proposed work
10	1.27	1.76
20	2.22	2.28
30	2.38	2.30
40	3.54	2.95
50	3.14	3.63
60	4.49	4.09

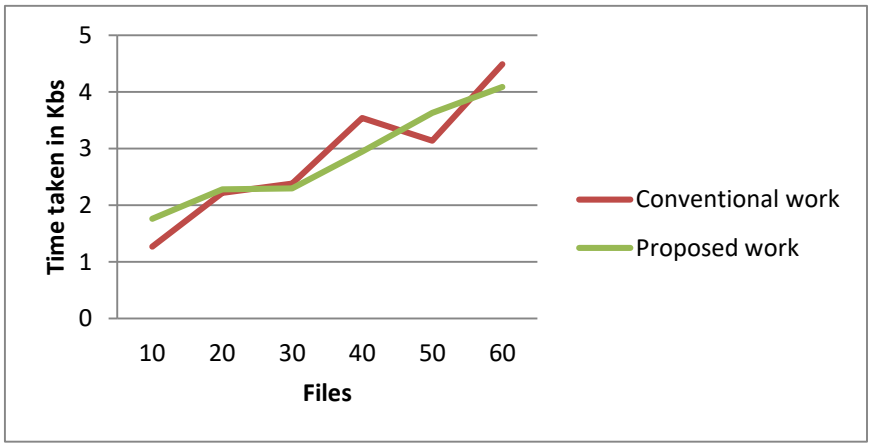

Fig 2 Comparison of time taken

MATLAB SIMULATION FOR COMPARATIVE ANALYSIS OF SECURITY

This section presents the impact of the proposed work on security. In case of the proposed work, the number of files affected is less as the number of attacks increases. From the following figures, it is concluded that the affected files are less in the case of proposed work.

MAN-IN-THE-MIDDLE

Its impact on the files in the case of conventional and proposed work in case of this attack are shown below.

Table 2 Comparative analysis of Man in middle attack

Number of attack	Conventional Work	Proposed work
10	3	2
20	6	6
30	9	8
40	11	10
50	14	14
60	17	16

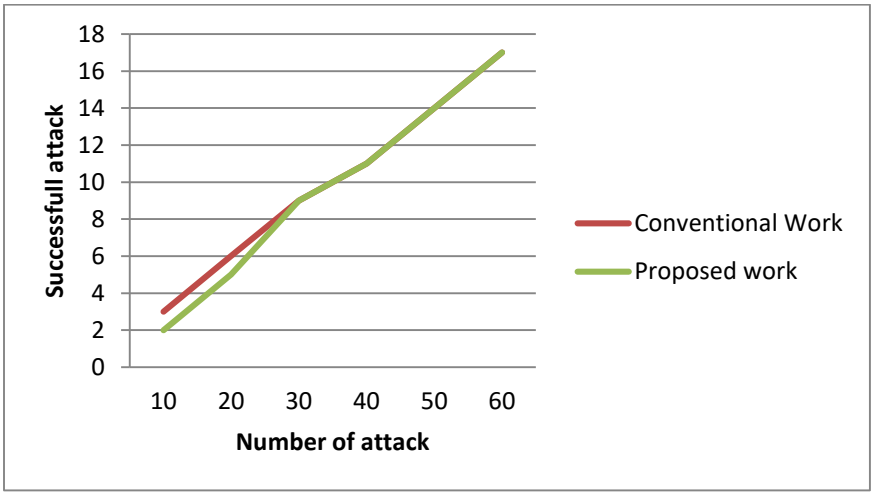

Fig 3 Comparative analysis in case of attack Man-In-The-Middle

BRUTE FORCE ATTACK

A brute force attack involves guessing login information via trial and error. Encryption keys and a hidden web page are also used. Comparative analysis of this attack is shown below.

Table 5.5 Comparative analysis of Brute force attack

Number of attack	Conventional Work	Proposed work
10	3	2
20	7	6
30	10	10
40	13	13
50	17	16
60	20	19

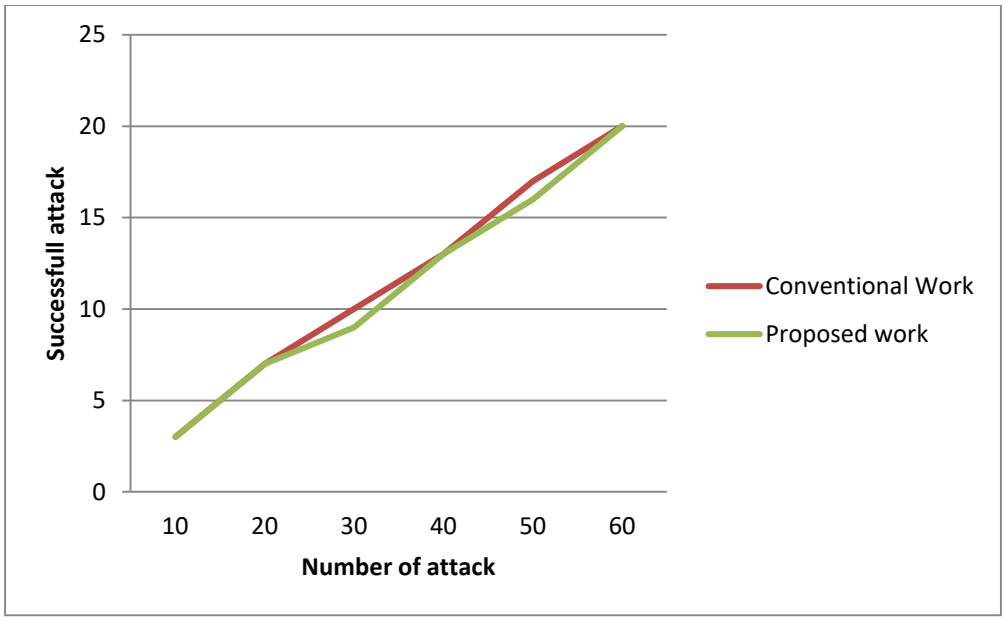

Fig 4 Comparative analysis in case of Brute force attack

Conclusions

Simulation results conclude that proposed work is providing more security along with better performance. During big data processing data security is big challenge. Present work has supported in reducing security threats. Moreover it is observed that proposed work is providing real life flexible solutions. These solutions are scalable.Healthcare, AI, and IOT are three areas where a clustered transmission technique may be helpful. On the other hand, research into unmanaged data may have a significant influence on huge data analysis. This system combines clustering and encryption to improve speed while maintaining data security. There are now comparisons between this approach and other methodologies' performance and security.

Future Scope of work

The scope of this research is to help web usage mining methods. Web usage mining refers to the discovery of user access patterns from web usage logs, which record every click made by each user. The key elements of web usage data preprocessing are data cleaning, page view identification, user identification, and sessionization, path completion and data integration. The objective of the study is to improve and find new model for session identifications using hybrid approaches, which is the important part of data pre-processing, effective session identification provides a reliable platform for web mining models and accurate results can be generated. The use of distributed framework Hadoop for session identification and web log analysis is to be carried in the research work to process massive data and to provide real time response for analysisIn further, more advanced techniques can be implemented to enhance the process of serial crime detection, hotspot moderation and crime estimation. Efficient clustering methods can enhance the precision of dataset classification. Including crime data from sources, other than police stations and social media, for example, print media, detective agencies and more classified sources can improve the accurate crime detection. Incorporating an automatic update system that can update the crime hotspot at different time periods based on the results can be an interesting choice. There are many available opportunities for continued investigation of crimes by using Association rule mining with the use of legal datasets.

References

1. D. Asir Antony Gnana Singh, A. Escalin Fernando, E. Jebamalar Leavline (2016) Performance analysis on Clustering Approaches for Gene Expression Data.
2. Kritwara Rattanaopas, Sureerat Kaewkeeree(2017) Improving Hadoop MapReduce Performance with Data Compression: A Study using Wordcount Job.
3. Yongxin Huang, ManXie, JiataoZhang(2017) A novel border-rich Prussian blue synthetized by inhibitor control as cathode for sodium ion batteries,
4. Sivarajah U, Kamal MM, Irani Z, Weerakkody V. (2017, January) Critical analysis of Big Data challenges and analytical methods. J Bus Res;70:263-286.
5. Llewelyn H. (2017) Sensitivity and specificity are not appropriate for diagnostic reasoning. BMJ;358:j4071.
6. Ahmed Oussous , Fatima-Zahra Benjelloun (2017) Big Data technologies: A survey.
7. Tabary MY, Memariani A, Ebadati E. (2019) Chapter 3 - Developing a decision support system for big data analysis and cost allocation in national healthcare. In: Dey N, Ashour AS, Bhat C, Fong SJ, editors. Healthcare data analytics and management. Cambridge, MA: Academic Press:89-109.
8. Muhammad UmerSarwar, Muhammad KashifHanify, RamzanTalibz, AwaisMobeenx, Muhammad Aslam (2017) "A Survey of Big Data Analytics in Healthcare", proceedings of International Journal of Advanced Computer Science & Applications, , vol. 8 (6), pp. 355-359.
9. J. Kaur, K. Sachdeva, and G. Singh (2017) "Image processing on multinode hadoop cluster," 2017 Int. Conf. Electr. Electron. Commun. Comput. Optim. Tech., pp. 21–26.
10. ArunPushpan, Ali Akbar N (2017) "Data Mining Applications in Healthcare", proceedings ofIOSR Journal of Computer Engineering, pp. 04-07.
11. K. Chitra1, Dr. D. Maheswari(2017) "Comparative Study of Various Clustering Algorithms in Data Mining", proceedings of International Journal of Computer Science and Mobile Computing, 2017, vol. 6 (8), pp.109 – 115.
12. K. Rattanaopas and S. Kaewkeeree (2017) "Improving Hadoop MapReduce performance with data compression: A study using wordcount job," ECTI-CON 2017 - 2017 14th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol., pp. 564–567.
13. Chen M. (2017) "Soft clustering for very large data sets", proceedings of International Journal of Computer Science and Network Security, vol.17 (1), pp. 102-108.
14. P. R. Merla and Y. Liang (2018) "Data analysis using hadoop MapReduce environment," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol., pp. 4783–4785, 2018.

15. S. Kumar and Z. Raza (2018) "A K-Means Clustering Based Message Forwarding Model for Internet of Things (IoT)," 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, pp. 604-609.
16. Ristevski B, Chen M. (2018) Big data analytics in medicine and healthcare. J Integr Bioinform 10;15(3):20170030
17. C.P. Patidar, NehaVerma (2018). Comparison of Visual Content for Different Browsers. International Journal of Computer Science and Engineering, vol. 6, no. 4, pp177.
18. Sitalakshmi Venkatraman andMamounAlazab(2108) "Use of Data Visualisation for Zero-Day Malware Detection", Security and Communication Networks, Article ID 1728303, 13 pages.
19. Deepak Gupta and Rinkle Rani(2018). Big Data Framework for Zero-Day Malware Detection", Cybernetics and Systems, DOI: 10.1080/01969722.2018.1429835.
20. Dhilip Kumar V, Vinoth Kumar V, Kandar D (2018), "Data Transmission Between Dedicated Short-Range Communication and WiMAX for Efficient Vehicular Communication" Journal of Computational and Theoretical Nanoscience, Vol.15, No.8, pp.2649-2654.
21. Jayasuruthi L,Shalini A,Vinoth Kumar V.,(2018) " Application of rough set theory in data mining market analysis using rough sets data explorer" Journal of Computational and Theoretical Nanoscience, 15(6-7), pp. 2126-213
22. J. Xiong et al. (2019, april). Enhancing Privacy and Availability for Data Clustering in Intelligent Electrical Service of IoT," in IEEE Internet of Things Journal, vol. 6, no. 2, pp. 1530-1540.
23. T. Farnham(2019). Indoor Localisation of IoT Devices by Dynamic Radio Environment Mapping," 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), Limerick, Ireland, pp. 340-345.
24. J. Jung, K. Kim and J. Park(2019), Framework of Big data Analysis about IoT-Home-device for supporting a decision making an effective strategy about new product design, International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Okinawa, Japan, 2019, pp. 582-584
25. Lee S, Mohr NM, Street WN, Nadkarni P. (2019, march). Machine learning in relation to emergency medicine clinical and operational scenarios: an overview. West J Emerg Med;20(2):219-227