DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5459

ISSN No. 0976-5697

Volume 9, No. 1, January-February 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

REVIEW OF MINING TECHNIQUES USED IN THE LOG DATA PROCESSING BASED ON HADOOP AND CLOUD COMPUTING ENVIRONMENT

Mrs.S.Revathi
Assistant Professor/Research Scholor
PG and Research Department of Computer Science
Dr.N.G.P Arts and Science College
Coimbatore-48

Dr.K.Nandhini
Assistant Professor/Research Supervisor
PG and Research Department of Computer Science
Chikkanna Government Arts College
Thiruppur

Abstract: Cloud computing becomes more popular because of their elastic nature and thus makes a greater effects in the day-to-day activities of users. Usage of information by the peoples is increasing tremendously which cannot be handled by users alone. Thus, users intend to utilize the cloud services to manage their large growing data's. Cloud provides a one of the widespread service as Software as a Service (SaaS) which enables users to utilize the latest updated software to accomplish their tasks. The main problem of SaaS is difficult to maintain the log information of user; it is increased directly when the numbers of users are increased. In SaaS environment the log data processing turn into more critical factor, it is more complex to handle for large volume of size. Some of the researches have been conducted towards performing log data processing using data mining techniques for the SaaS applications in the Hadoop environment. The analysis has been carried over on the different researches of log data processing in terms of their working procedure, merits and demerits. This analysis provides the backlogs of various methods with comparison. The performance is analyzed by comparing all methodologies with each other in factor of their merits and demerits.

Keywords: Log data processing, Data mining techniques, summary of usage details, large volume of information

1. INTRODUCTION

SANS Log management survey says that in 2008, 20% of the respondents who were satisfied with their log management system, for log analysis they spent more than one week in every month. Most of those companies were in the Global 2000. The small and medium sized businesses (SMBs) and government organizations are spent more than half-day to five days per month for log analysis. The survey also showed that, most organizations are difficult to setup and integration process and only achieved partial automation of their log management and reporting processes. Many organizations and particularly SMBs are facing that difficulty and wondering if they should turn over log management to an in-cloud provider—one that provides their log management software and log data storage over the Internet.

In January 2008, Stephen Northcutt, president of the SANS Technology Institute, wrote that there are pitfalls with putting log management in the cloud. In positively, he said, "you will almost certainly save money. In addition, real experts on log analysis are hard to find..." Recently, vendors initiated to provide log management in the cloud (otherwise known as Software as a Service or SaaS), as a way to simplify log management because the provider can dedicate the material resources and retain the talented, focused personnel to do a better job for less money. This will be useful for not only for SMBs without the dedicated manpower and also for enterprises whose extend the IT resources they will be try to manage with multiple distributed LANs[1].

While IT managers agree that log management is difficult, they are leery about handing over their log data to a third party application provider because the data might not be available when they need it, not to mention the sensitive nature of some of the data that shows up in log files. Before deploying or overhauling log management systems, organizations are required to consider the benefits and drawbacks of each model in context of their business requirements.

Numerous IT administrators are worried with the security of their log information, and which is well and good: Log information can be risky on the off chance that it falls into the wrong hands. Aggressors can get profitable data from perusing the logs. For instance, they can check whether their assaults work, distinguish inside hosts, and even recognize client names and passwords (which have been known to appear in logs). Log information as normal as Web or email movement regularly contains classified data. Having control of logs can be valuable to aggressors who, now and again, will attempt to clean the log information to expel any hints of their movement.

In this manner, it is vital to take a quick look at the security of log information—whether it is put away on-or off-site. On the off chance that the log information is put away locally, it is frequently continued every individual PC creating the information. Larger associations will have log servers that will store the log information in a concentrated appended stockpiling gadget. Those frameworks are, in a perfect circumstance, secured and hard to break into. In the cloud display, this information stockpiling would be given off to the cloud supplier, which eases the association of the equipment, security and HR loads required with keeping stockpiling in-house. In any case, as they lose control of that information, associations must depend on the cloud administration to handle their information safely. The issue of whether an administration association is capable is hard to decide, and is at last in view of notoriety. Cloud suppliers must make a trust demonstrate as they oversee gathered log information safely and independently in a multi-inhabitant

environment. This makes the requirement for extra layers of security to isolate various occupants from each other on a common server, while likewise shielding the information stores from assailants.

In this analysis work, detailed discussion about the varying techniques that are carried out for efficient log data processing is discussed. The efficient management techniques and also concerns that are taken for ensuring the security of the log data which is going to be stored in the third party servers is also discussed. The merits and demerits of various research methodologies are discussed in detail along with their working procedures and methods.

The subsequent section discusses about the various research findings regarding log data processing, management and issues. In particular, SaaS log management, security prevention and management of log files. The section three depicts the analysis of log data processing methods and cloud security mechanisms. The fourth section reveals the conclusion of this research.

2. ANALYSIS OF LOG DATA PROCESSING MANAGEMENT

Log information handling is a more vital assignment in the each venture to store their operations in subtle elements. For instance, companies would like to store as much as possible observations regarding their manufacturing details, selling details, accounting details and customer response details. These details would be more useful for enterprises, which can be used in future to take important activities towards enhancing their organization objectives. As the quantities of clients are expanded, log data is additionally expanding which is harder to handle. Consequently, the businesses moving towards distributed computing fashion increases the demand of log information and its handling. However, two noteworthy issues might be happen while performing log information handling on the cloud environment. Those are recorded as takes after: "computational many-sided quality because of developing measure of data and security issue because of putting away organization log data in the outsider server". In the accompanying sub area itemized dialog about the examination systems, which are proposed to handle these issues are given.

3. EFFICIENT LOG DATA PROCESSING MANAGEMENT WITH THE CONCERN OF GROWING AMOUNT OF INFORMATION USING DATA MINING APPROCHES

With the fast advancement of the Web, SaaS applications conveyed as administrations through web turn into an important alternative of traditional software [2, 3]. SaaS applications are conveyed in dynamic data centers [4] with distributed computing innovations overseeing assets to accomplish adaptability and scalability [5, 6]. In this manner, the significant use information of SaaS applications, which can be huge and confused turn into a colossal data space. The utilization data of SaaS is made out of information, which is produced while the applications are gotten to by clients. The getting to time, client source, work utilized, asset devoured are run of the mill use data. While utilizing the SaaS applications to satisfy the business capacities, clients need to know the use data too, in light of the fact that they needs to uncover some Business Insight learning or they simply need to know the amount they have utilized the administration. In this way, the data can be examined rapidly, accurately. Further, it helps to discover much information from data, such as customer details, region, attitude and their motivation to buy the products. Therefore, the log file takes a major role in assessing the process, customers and interruptions.

For instance, through examining the utilization data of administration capacities, we will discover applicable capacity arrangement so that valuable capacity gathering can be intended to enhance the client experience and benefit. To get helpful data, we require information preparing [7] and information mining [8, 9] systems to dissect the utilization information. Given the size of most SaaS applications with a great many clients and huge information, database based continuous investigation techniques can't meet the prerequisites in light of the fact that their restricted execution for taking care of extensive scale issues. As an option arrangement, information log based investigation strategies [10, 11] have been turned out to be successful at explaining vast scale continuous examination issues. These techniques first characterize the arrangement of information log, which contains pieces of execution data required by the examination. SaaS applications create log information as indicated by the characterized organize for every occupant and every client while running in conveyed way. At that point, gigantic information log from various sources is transmitted to handling servers through offbeat transmitted instrument.

As said above, information log should be mined and handled so that valuable data can be uncovered. In any case, traditional data processing preparing and mining strategies are inadequate or not ready to handle enormous measure of information. This induces the requirement of another strategy in data mining to handle the large volume of data in more viable manner. The structure of Hadoop [12] gives an answer for issues of huge information handling, since it runs applications on large cluster built of commodity hardware with failure tolerance. Hadoop executes a computational strategy named as MapReduce, and it offer a distributed file system (HDFS) that stores information on the compute nodes. Along these lines, to get the helpful learning from the huge log information of SaaS use data rapidly and effectively, this work discusses about information log preparing and mining strategies in view of Hadoop to take care of the issue.

On data miming, how to generate mining transaction set reasonably from origin records is an important problem. SaaS usage data has characteristics of the huge amount and long frequent item set. According to these characteristics, current algorithms of association rules mining with item constraints have low efficiency with massive candidate item sets [13]. Besides, how to generate transaction set from records is another problem, reasonable transaction sets are able to improve the speed and quality of data mining, but current methods based on constraints cannot solve this problem either. So, it is important to find an algorithm of association rules mining with item constraints according to the characteristics of SaaS data [14].

Effective data mining is able to dig out meaningful and valuable information from raw SaaS data. Usage behavior of certain users of specific region, group or tenant can be found and used in Business Intelligence forecasts. The ideal algorithm needs to be scalable, which means it is able to mine large amount of data in acceptable time span [15]. On data mining, the Apriori algorithm is a classic algorithm for getting association rules. Apriori is designed to operate on data transaction sets for finding association rules. Given a set of item sets, the algorithm attempts to find frequent item sets which are common (occur more times than a certain lower bound). Apriori uses a "bottom up" approach, where frequent itemsets from the each transaction (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found [16].

4. SECURITY PREVENTION METHOD WHILE STORING LOG DATA IN THE CLOUD

In this Survey relative mechanisms and the methods which are employed earlier to attain a security and privacy are discussed. And also the advantages and disadvantages of each technique are discussed. According to the survey of the earlier mechanism, it finds that the current system implemented has more advantages.

C.Wang et al [17] concentrates on accomplishing secured and dependable storage in the cloud environment where the information is outsourced in distributed manner. The powerful method for accomplishing secured cloud storage is accomplished. To rectify this issue developed a method as integrity auditing. The system of this work comprises of three entities, to be named as client, cloud server (CS), Third party auditor (TPA). Cloud provider is accountable for offering storage space to the cloud user for storing their data in provided space. TPA will handle the process of providing trusted environment for to access the data. This work assess the verified the data in dynamically in order to provide secure dependable data storage.

N.Cao, [18] solves a reliability issues with optimal solution for providing cloud service as secure storage. In order to maintain the data integrity, the data owner needs to keep up his state in the online itself where it will build the burden of data owners. This issue is controlled by a exact repair solution where none of the data should be produced during the process. The LT-Codes based secure and dependable cloud storage service (LTCS) approach utilized as a part of this work is executed to give an efficient data storage for data owner and data user. This technique comprises of a few phases. Those are Setup, Data outsourcing, Data retrieval, Integrity check and Data repair. This method solves the problem show in previous strategy called erasure based coding technique. This technique is utilized to accomplish less storage cost, and quick data retrieval. Moreover, it concerns about only managing of data integrity in all duplicates copies of data and guaranteeing the data security efficiently.

C.Wang et al [19] assessed a privacy-concerned technique with a public auditing system which guarantees zero-knowledge leakage by utilizing the cloud data. Cryptographic measures are not sufficiently just to give security over the cloud server where massive data is shared publicly. Since the information put away in the cloud are exceedingly alert and that are outsourced publicly with the other users to share their knowledge. To conquer these issues in this paper TPA idea is processed. Homomorphic authenticator and random masking guarantees that TPA couldn't increase any learning during the process of auditing. It empowers TPA to get to the information from the distributed storage to share the confidential information of clients. The vast majority of the association begins to store their information in a cloud situation to share it among their staff individuals successfully. It will give a economic feasibility and adaptability, access within group members. In previous works the privacy and security issues are examined just in individual user will sharing the data though it can't be connected to the multiple users. In the group sharing of information there might be the issues founded like initiate new members and revoking the existed used.

X.Liu et al [20] presents a technique for sharing information in a multi owner way. This approach is utilized to accomplish a security preservation of information and identity of data owner information. This strategy forces an idea of imparting information among the multiple clouds with the different characteristics by any user in the group. Sharing information in a various cloud by any client rather than information proprietor will prompts to a security danger in the untrusted cloud. This issue is overcome by utilizing the approach, to be specific Secure Multi Owner Data Sharing (MONA). The objective of this work is to accomplish access control, data confidentiality, anonymity, traceability and efficiency. The dynamic data sharing among the dynamic group is accomplished by combining the group signature concept and dynamic broadcast encryption method. In this approach, the group manager is permitted to register the revocation parameter and move them to the public cloud for sharing the information. The computation overhead happened when the user computing the revocation parameter independently can decrease it extensively.

H.Wang, [21] focus on enabling user to control the security over remote data. Remote data possession is the most complex issue to be accomplished by the clients when the client is in the outside ranges from rremote area like being in prison, being in the battlefield and so on. In this case proxy provable data possession has to be provided. To rectify this issue, the Proxy Provable Data Possession (PPDP) protocol is assessed in this work. This PPDP protocol comprises of a six stages. Those are SetUp, TagGen, SignVerify, CheckTag, GenProof, and CheckProof. In PPDP plan, CheckTag is added to every customer those who are accesing to the data in order to prevent from the malicious client. The performance of this protocol is evaluated by utilizing the two parameters, namely communication overhead and computation overhead. This PPDP protocol is ended up being a secured one to offering an efficient data possession checking of remote information by clients. This approach utilizes the public verifiability to demonstrate that information in not changes by the unauthorized persons. By utilizing this approach anybody can utilize this strategy to demonstrate their accuracy.

Y.Zhu et al [22] proposed a novel dynamic audit service for the untrusted and outsourced data from the cloud. The main goal of this approach is to provide a data integrity check when the data are shared to the untrusted cloud. This

work tries to achieve the security metric given in the following list to check the performance of this approach. Those security metrics are Public auditability, Dynamic operation, Timely detection, Effective forensic, Light weight.The virtualized nature of cloud computing will cause many of the security attacks in the cloud. The data will be gathered in the one place of cloud for effective management where there are lots of possible DDoS attacks like Html and Xml are available. It will create the threat to the cloud environment which will also affect the cloud service consumers.

L.Ferretti et al [23] proposed a novel approach to provide a secured data access over a distributed concurrent database. Cloud environment is the virtualized environment where the data are shared publicly. Security threat occurs in the cloud environment when data owners' placing a sensitive data on the cloud providers which may cause the collision of data. In this work secure DBaas framework is designed to allow multiple clients who act independently to connect with the untrusted Dbaas without any intermediate servers.

All the works discussed above clearly show the different methodologies used to provide a privacy and security prevention for the cloud data users as well as for cloud data owners. All of the above discussed technologies are meant to be solved various types of security threats and also possible ways to provide privacy.

ANALYSIS OF RESEARCH METHODOLOGIES

S. No.	AUTHOR	METHOD	ADVANTAGES	DISADVANTAGES
DATA MINING TECHNIQUES ON LOG DATA PROCESSING USING HADOOP				
1	Liang Zhong et al (2010) [5]	Virtualization-based SaaS	Can support the user requirements efficiently by providing their requirements in individual server More user satisfaction level	Log data maintenance would be more difficult task due to arrival of more service requests
2	Yuan Yuan et al (2011) [6]	Resource management process in cloud	User can be provided with their required resources based on their request Log data management would be more flexible	QoS requirements isn't concentrated Log management is complex in case of arrival of large volume of tasks
3	Maruster L et al (2004) [7]	Process mining	Log identification is done efficiently Log extraction is done accurately with the help of process flow	Large volume logs which is increasing dynamically cannot be supported well
4	Weijters T et al (2001) [8]	Event based Process mining	Log identification is done efficiently	Large volume logs which is increasing dynamically cannot be supported well It only considers active events for the log data gathering which might reduce the processing quality
CLOUD SECURITY MECHANISM				
1	Cong Wang et al [17]	Distributed storage integrity auditing mechanism	Dynamic data verification Resilient against Byzantine failure and malicious data modification attack	High redundant copies are present which may cause high memory occupation Data Integrity is not achieved
2	Ning Cao et al [18]	LT codes-based cloud storage service	Efficient and fast data retrieval Less storage cost	Need to retrieve entire data to check data integrity TPA is not trustable
3	Cong Wang et al [19]	Privacy-preserving public Auditing mechanism	Assures zero knowledge leakage Better privacy preservation	Group access of data cannot be secured
4	Xuefeng Liu et al [20]	Secure multi owner data sharing scheme	Better security over group of users User revocation is handled effectively	Remote data integrity is not considered

5	Huaqun Wang [21]	Proxy provable data possession	Efficient user controlled data management	Public verifiability may causes intruders collision on data
6	Yan Zhu et al [22]	Dynamic audit services	Less communication overhead Less memory storage	Highly causes from security attacks
7	Luca Ferretti et al [23]	Novel architecture that integrates cloud database services with data confidentiality	Guaranteed data confidentiality	High computational cost

5. CONCLUSION

In real world environment the more essential part is Log data processing where the services are offered to the users based on their requirements. Log data processing required to be done efficiently to improve their performance by knowing the previous log data. In this research, analysis of different research methodologies that attempts to maintain their log data and processed in the Hadoop environment is done with the concern of handling massive amount of data. This paper discussed various security preventing approaches which attempts to ensure the security while storing log data in the third party servers. The analysis has been carried out by comparing the merits and demerits of the varying related research works is done which shown in the tabular format to identify the processing issues that occur during log data processing.

REFERENCES

[1] Varma, Anil, and Nicholas Edward Roddy. "Method and system for processing repair data and fault log data to facilitate diagnostics." U.S. Patent 6,415,395, issued July 2, 2002.

[2] N. Gold, C. Knight, A. Mohan, and M. Munro, "Understanding Service-Oriented Software." IEEE Software, vol. 21(2), pp. 71-77, 2004.

[3] P. Laplante, J. Zhang, and J. Voas, "What's in a name? distinguishing between SaaS and SOA." IT Professional, vol. 10(3), pp. 46–50, 2008.

[4] Jing Bi, Zhiliang Zhu, Ruixiong Tian, and Qingbo Wang, 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD), Miami, Florida, pp. 370–377, 2010.

[5] Liang Zhong, Tianyu Wo, Jianxin Li, Bo Li. "A Virtualization-based SaaS Enabling Architecture for Cloud Computing" , 2010 Sixth International Conference on Autonomic and Autonomous Systems (ICAS), Cancun Mexico, pp. 144-149, 2010.

[6] Yuan Yuan, and Wen-Cai Liu, 2011 International Conference on System Science, Engineering Design and Manufacturing Informatization (ICSEM), "Efficient resource management for cloud computing", pp. 233-236, 2011.

[7] Maruster L, Weijters A, van der Aalst W M P, et al. "Process mining: discovering direct successors in process logs", Computers in Industry, vol. 53(3), pp. 231–244, 2004.

[8] Weijters T, van der Aalst W M P. "Process mining: discovering workflow models from event-based data", 13th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC2001) Amsterdam, Netherlands pp. 283–290, 2001.

[9] Nayak R, Tong C. "Applications of data mining in Web services", 5th International Conferences on Web Information Systems, Brisbane, pp. 199–205, 2004.

[10] Van der Aalst W M P, Weijters T, Marudter L. "Workflow mining: discovering process models from event logs", IEEE Transactions on Knowledge and Data Engineering, pp. 101–132, 2002.

[11] Van der Aalst W M P, Van Dongen B F. "Discovering Workflow Performance Models from Timed Logs", International Conference on Engineering and Deployment of Cooperative Information Systems(EDCIS 2002), pp. 45–63, 2002.

[12] Hadoop, http://hadoop.apache.org/, 2011

[13] Xin, D., Han, J., Yan, X., & Cheng, H. (2005, August). Mining compressed frequent-pattern sets. In Proceedings of the 31st international conference on Very large data bases (pp. 709-720). VLDB Endowment.

[14] Peng, Y., Kou, G., Shi, Y., & Chen, Z. (2008). A descriptive framework for the field of data mining and knowledge discovery. International Journal of Information Technology & Decision Making, 7(04), 639-682.

[15] Vaidya, J., & Clifton, C. (2002, July). Privacy preserving association rule mining in vertically partitioned data. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 639-644). ACM.

[16] Ye, Y., & Chiang, C. C. (2006, August). A parallel apriori algorithm for frequent itemsets mining. In Fourth International Conference on Software Engineering Research, Management and Applications (SERA'06) (pp. 87-94). IEEE.

[17] Cong Wang, Qian Wang, KuiRen, Ning Cao and Wenjing Lou, "Towards Secure and Dependable Storage Services in Cloud Computing", IEEE Transactions on Cloud Computing Volume: 5 , Issue: 2, April-June 2012, ISSN: 1939-1374

[18] Ning Cao, Shucheng Yu,Zhenyu Yang, Wenjing Lou and Y. Thomas Hou, "LT Codes-based Secure and ReliableCloud Storage Service", Proceedings of IEEE Infocom, 2012, ISSN: 0743-166X

[19] Cong Wang, Sherman S.M. Chow, Qian Wang, KuRen and Wenjing Lou, "Privacy-Preserving Public Auditing for Secure Cloud Storage", IEEE transactions on computers, vol. 62, no. 2, February 2013, ISSN: 0018-9340

[20] Xuefeng Liu, Yuqing Zhang, Boyang Wang, and Jingbo Yan, "Mona: Secure Multi-Owner Data Sharing for Dynamic Groups in the Cloud", IEEE transactions on parallel and distributed systems, vol. 24, no. 6, June 2013, ISSN: 1045-9219

[21] Huaqun Wang, "Proxy Provable Data Possession in Public Clouds", IEEE Transactions on Services Computing, Vol. 6, No. 4, October-December 2013, ISSN: 1939-1374

[22] Yan Zhu, Gail-JoonAhn, Hongxin Hu, Stephen S. Yau, Ho G. An, and Chang-Jun Hu, "Dynamic Audit Services for Outsourced Storages in Clouds", IEEE Transactions On Services Computing, Vol. 6, No. 2, April-June 2013, ISSN: 1939-1374

[23] Luca Ferretti, Michele Colajanni, and MircoMarchetti, "Distributed, Concurrent, and Independent Access to Encrypted Cloud Databases", IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 2, February 2014, ISSN: 1045-9219