An Authenticated Survey on Big data Security

Ms. Shivani S. Kaktikar [1], Prof. Sarika V. Bodake [2]

[1] *M. E. Student, Computer Department, PVPIT Pune, Maharashtra, India*
[2] *Head of Department & Guide, Computer Department, PVPIT Pune, Maharashtra, India*

ABSTRACT

The ever-growing interconnectivity of computer networks is to account for the circumstances that permit our expanding worldwide cyber security threat susceptibility. Simultaneously, technological advancements and infrastructure have driven researchers to establish a new challenge: The Big Data issue. As a result, cyber security experts must develop and execute innovative techniques to quickly and successfully neutralize cyber security vulnerabilities in a Big Data context. Numerous information devices have become extremely sophisticated as a result of the recent standardization of information technology. Organizations continue to produce and preserve significant digital data while linked to one another, heralding in the era of big data. However, there's a good chance they'll reveal sensitive information because they share plenty of it via regular conversation. As more electronic systems are added, a system becomes more susceptible. Therefore, this survey paper analyzes researches on this topic to achieve an effective big data security approach which will be elaborated further in the net editions of this research.

Keyword: - *Big Data, MongoDB*

1. INTRODUCTION

Across the world, there has been a considerable increase in the number of users of computing and other electronic devices. This is due to the increased affordability of the devices and the immense technological development that has enabled effective improvement in the manufacturing of these devices. The mass production of the devices is achieved through these improvements which drive down the cost of these devices much further. These situations have been useful and conducive to the entire electronic ecosystem. The devices are being increasingly bought by consumers and the number of these types of devices has been increasing exponentially.

The increase in the number of devices, such as smartphones and other electronic devices, has led to an increase in the number of individuals using the internet platform. The internet platform is a host to a number of services that are provided online, especially the paradigm of social media and useful services such as banking and E-commerce. The internet platform has made the lives of individuals highly convenient and extremely useful. The use of these services and a large number of individual devices add up to a significant amount of data that is generated every day.

This data is constantly being generated and is one of the most useful and highly insightful data that needs to be preserved. This data is massive in nature and keeps on generating constantly. This type of data is referred to as Big Data that has certain characteristics, such as high velocity, high veracity, high volume, etc. These attributes are unique to Big Data and are used to categorize the data effectively and understand the various uses of this type of data. Big Data can be useful in deriving useful statistics and insightful data to achieve improvements in the myriad of processes that achieve the analysis. The Big Data needs to be effectively secured to safeguard the data and also improve the analysis capabilities to ensure the efficient execution of the approach.

However, the widespread use of big data poses significant data security issues, particularly when dealing with sensitive data such as trade secrets, sensitive information, and health records. The majority of the researchers seek to secure sensitive data from possible dangers in order to get the most out of this information. The necessity of data protection in the presence of big data is highlighted, where big data is seen as a focus for attackers looking to steal highly valuable data. Conventional security measures, on the other hand, are incapable of safeguarding large data migration. As a result, protecting big data is a challenge that necessitates the development of new solutions to secure such vast amounts of data. Therefore, a number of related works have been analyzed in this survey article to achieve effective improvement in the security risks associated with big data.

Many studies have extensively emphasized data protection. Most safety precautions, on the other hand, are intended to safeguard fixed data from attacks, which is inadequate for large data and beyond the computing power of conventional databases. The organization's sensitive data is a valuable target for assaults that jeopardize its trust and credit. Big data, for example, may pose a security risk to users' emails by creating phishing sites relying on their email activity and interests. Big data security is regarded as among the most significant threats to cloud computing systems and has a negative impact on a company's productivity.

Security and privacy are the critical requirements for storing, managing, analyzing, and transmitting big data. Any comprehensive big data security solution should meet data confidentiality, integrity, and availability. Big data privacy concerns are considered in building big data environments and protecting big data during either storage or processing, where multiple encryption techniques are implemented to prohibit unofficial users from accessing big data. However, the proposed approaches treat all data with the same priority and their intractable time complexities prove their impracticality, which is why our methodology is being developed and will be elaborated further in the upcoming editions of this research article.

This literature survey paper dedicates section 2 for analysis of past work as a literature survey, and finally, section 3 concludes the paper with traces of future enhancement.

2. RELATED WORKS

To meet the demands of data explosion, online information retrieval, and semantic search, the MRSE-HCI architecture is proposed by C. Chen et al. [1]. Simultaneously, a verifiable mechanism is proposed to ensure the accuracy and completeness of search results. The suggested hierarchical approach groups documents based on their lowest relevance thresholds then divides the resulting clusters into sub-clusters until the maximum cluster size is attained. During the search phase, this approach can achieve linear computational complexity despite an exponential increase in document collection size. The authors also look at the search efficiency and security in the context of two prevalent threat models. The search efficiency, accuracy, and rank security are all evaluated using an experimental platform. The experiment results show that the design proposed not only solves a search problem ranked among the multi-keywords, but also improves search efficiency, ranking security, and document relevance.

X. Xiong et al. presented a lurker game model for acquiring huge data that was similar to the public goods game in terms of mechanics. In addition, the lurker game concept now includes individual motivation. A sort of evolutionary public goods game is the lurker game. This model added an individual incentive factor based on his degree, in addition to the features of the public goods game [2]. The authors discovered that the individual strategy to be picked is not related to the user's degree, but rather to a network-wide incentive constant. Individual strategies asymptotically followed three different behaviors, according to the dynamic framework of the individuals and the simulation results. During the evolutionary process, active users emerged as a result of motivation. Without an incentive, active central users would have little impact on their neighbors' states, and to a large number of lurking neighbors, they might even become lurkers. Large amounts of noise reduce the impact of a strong incentive, resulting in network chaos. If the instability persists, active users will lose interest and eventually exit the network. The research contributes to a better knowledge and exploration of social networks' underlying activity mechanisms.

K. Yang et al. presented a data access control mechanism for big data that's both efficient and fine-grained and doesn't leak any personal information. Unlike previous techniques, which can only partially hide attribute values in access policies, their method can hide the complete attribute in access policies [3]. However, legal data consumers

may face significant problems and difficulties in decrypting data as an outcome of this. To solve this problem, the authors created an attribute localization algorithm that determines if an attribute is included in the access policy. The authors showed that their method is selectively secure against specific plaintext assaults. In addition, they developed the ABF using MurmurHash and the access control scheme to demonstrate that their approach can preserve the privacy of any LSSS access policy without imposing a significant burden.

L. Watkins et al. presented a semi-supervised machine learning strategy for dealing with the ever-increasing Big Data problem in cybersecurity. They focused on DNS traffic in their analysis, using a technique to sift through DNS requests and find the smaller amount of questionable network traffic on the network [4]. They accomplished this by using typical clustering methods on a dataset enriched with known harmful domains to filter out the majority of non-malicious network traffic, allowing them to concentrate on a manageable collection of data that most likely contained suspicious or malicious domains. The key to their technique is that they use DNS name-based, TTL value-based, and DNS query answer-based behavior of known problematic domains to motivate clustering algorithms. Then, only the inspired clusters are kept, and these become the reduced dataset of dubious findings. The experimental outcome shows that the strategy can reduce a dataset of 400k query-answer domains to only 3% of the total malicious domains, comprising 99 percent of all malicious domains.

Jin Kim et al. performed research of an artificial intelligence intrusion detection system for successful attack detection utilizing the DNN model, a deep learning technique. For training and testing, it employed the well-known KDD Cup 99 dataset for intrusion detection. Test data were prepared by pre-processing and sample extraction to meet the objectives of the study. The training set consisted of 10% of the corrected data, while the testing set consisted of the complete dataset of about 4.9 million records [5]. The results demonstrate that the accuracy and detection rate are both extremely high, averaging 99 percent. Furthermore, the false alarm rate was 0.08 percent, implying that the chances of incorrectly identifying routine data as assaults are quite unlikely.

H. Xiong et al., proposed a new, efficient ID-based sign-up system, based on a binary tree structure (R-IBSC). Their construction provides, in particular, a shorter ciphertext size and faster sign-coding compared to the sign-then-coding approach. In addition, with the large-scale big-data environment, the key overhead updated at the PKG increases logarithmically with the number of users [6]. The presented scheme is also proven to achieve indistinguishability and existential unforgeability against the chosen-ciphertext attacks adaptively and the chosen message attacks adaptively (shortened as EUF-CMA), assuming the intractability of the Decision Bilinear Diffie-Hellman (DBDH) problem and the computational Diffie-Hellman (CDH) problem. The authors then introduce a concrete construction with the concepts of key rerandomization and ciphertext blinding. There is no evidence to show that any information on the underlying encrypted plaintext cannot be found in the proposed outsourced RIBSC technique and that the receiver may, by using the random oracle model, validate and outsourcing the cloud-server calculation.

Y. Wang et al. proposed MtMR to ensure the completeness of MapReduce calculations as the tree-based verification framework for Merkle. To ensure a high degree of integrity in the outcome of jobs, MtMR uses a hybrid cloud architecture and uses pre-education (i.e., map and shuffle) verifications based on Merkle-Tree to reduce the MapReduce jobs [7]. Their qualitative analysis has shown that a semi-honest employee cannot perform safe fraud under the MtMR. The authors showed that MtMR can produce high result integrity and moderate overhead performance.

J. Wang et al. [8], theoretically analyzed the effect on parallel computing of imbalanced distribution of the data resulting from the clustering of the data within HDFS blocks. The authors proposed DataNet to support distributed computing for sub-datasets. To store the sub dataset distributions, DataNet uses an elastic structure, known as ElasticMap. In addition, the construction of ElasticMap is supported by a dominant sub-dataset separation algorithm. DataNet allows the analysis of sub-datasets to easily balance their workload between computer nodes. The authors can easily find the most efficient blocks of data from the entire dataset as input for the sub dataset sampling. They use DataNet to conduct extensive experiments for various sub-dataset applications, and the findings suggest that DataNet has a promising performance.

FASTEN, an FPGA-based system is proposed by B. Hong et al., as a new option for providing the most secure alternative for big data processing. The motivation of FASTEN was that software-based Big Data processing systems contain unavoidable security vulnerabilities. In FASTEN, the data processing kernel code is ported on the

FPGA tissue to the host computer's hardware [9]. The execution takes place inside the computer. As data processing and crypto-working take place on the hardware, it is not possible to expose plaintext data practically via practically possible paths. It also makes use of modern FPGAs' tamper-resistant features to deal with side-channel attacks such as differential power analysis (DPA). These security features are combined with the reconfigurable nature of FPGAs to be used in the cloud. Using Xilinx Zynq-7000 devices, the authors built a 24-node Linux system. Experiment results with MapReduce applications show that FAS-TEN provides not only enhanced security but also a performance advantage over the most recent Hadoop release for security, at the expense of additional hardware.

H. Cui et al. presented a novel approach for implementing an attribute-based storage system that supports secure deduplication. Their storage system is designed using a hybrid cloud architecture, in which a private cloud handles computation and a public cloud handles storage [10]. The private cloud has a trapdoor key with the respective ciphertext that enables the text of a single access policy to be transferred from other access policies in ciphertexts of the same plaintext without knowing the plaintext. The private cloud validates the uploaded item first with the attached evidence when a storage request is received. If necessary, the ciphertext is regenerated into the same plaintext ciphertext via an Access Policy which is the union set of the two Access Policy(s). There are two major advantages to the proposed storage system. Firstly, it is possible to share data with other users in confidence, rather than to share the decryption key, by specifying an access policy. Secondly, it achieves the standard concept of semantic security, while only a weaker security concept applies to existing deduplication schemes.

W. Liao et. al. developed efficient translation and permutation systems for vectors and matrics based solely on linear algebra to protect the customers' private data. The authors showed that transformed data values and positions are computationally indistinguishable from random vectors and matrices under a CPA or CPA secure plaintext attack [11]. The customer can therefore trustfully share the transformed data with the cloud. The proposed secure algorithm for linear approximations enables the cloud server to find the solution efficiently while protecting the privacy of the client. Furthermore, to prevent malicious behavior of the cloud the correctness of the returned results can be verified efficiently by the customer. The theoretical analysis of data security has shown that customer data privacy is well maintained. The suggested algorithm effectively handles large-scale CSPs with significant customer time savings, according to Amazon Elastic Compute Cloud (EC2) trial findings. The proposed system requires the client to assist the cloud in solving the communication issue.

J. H. Abawajy et al., introduced Hybrid Consensus Pruning (HCP), a new ensemble pruning strategy that modifies the number of ensemble classifiers to make them suited for big data processing on the cloud. HCP shrinks the size of large ensemble classifiers while preserving or improving their performance. The HCP has three novel features: (i) it ranks all instances of the base classifiers in the ensembles and selects a percentage of the higher-ranked classifiers for further analysis; (ii) it uses a fast consensus function to obtain a stable consensus clustering that partitions the selected base classifiers; and (iii) it uses an optimization technique to choose base classifiers for the final ensemble within each of the created clusters [12]. The authors conduct an experimental case study to determine the efficacy of the HCP technique for the problem of malware detection, as this is one of the most critical issues in cloud security.

Big data recovery from the network is susceptible to malicious requests and data poisoning attacks. R. Li et al. proposed DCAuth to prevent such attacks, which provides data-centered authentication with a fusion of CA-based trust and neighborly trust. It allows users, IPEs, copyholders, and publishers, independent of their unpredictability, to authenticate. A suspension-chain model (SCM) as a trust model is proposed where certificate authority (CA) based trust suspends the neighbor-trust-based certification chain. It allows data-centered authentication to be carried out [13]. The proposal to build a reliable SCM-based suspension certificate chain is to deliver integrated hop-by-hop collection along with adaptive replacement of the chains with reliable certificates. The delay enlargement problem is resolved by avoiding dependence on centralized servers for the construction of a chain. DCAuth extends physical entities' authentication to the logical entities smoothly. It breaks the barrier that follows the data center approach between networking and big-data applications. Extensive simulations have been performed to demonstrate that DCAuth can reduce certificate collection delays compared to PCI-NDN and effectively prevent malicious requests and data poisoning attacks.

In the third-party auction platform, W. Gao et al. addressed the issue of protecting information privacy during the data auction. The concept of homomorphic encryption has been leveraged to create a confidentiality auction system (PPAS). The authors have chosen a set of crypto primitives and designed algorithms in their system, to make the auction process efficient, to perform a private auction. They proposed an enhanced privacy protection auction to

further improve the safety and resistance of PPAS attacks (EPPAS). For a comprehensive experimental review, the prototype system of the auction scheme has been implemented [14]. The experimental results show that the proposed scheme is able, under normal operations and without private information leakage, to ensure the determination of an auction winner with a 100% correct rate. In addition, they examined and deployed several attack scenarios against the auction in their review, and they were correctly detected and avoided.

Y. Zhang et al. presented a new system to perform a highly efficient storage audit, irrespective of the overall number of file blocks the revoked user possesses in the cloud. To do this, they explore a new strategy and a new private key update technique for key generation. With this strategy and the technique, they can repeal users simply by updating their private keys instead of the revoked user's authenticators. When the authenticators are not updated, the integrity check of the revoked user data may still be performed correctly. The proposed scheme is based on identity-based cryptography, eliminating the complex management of certificates in traditional PKI systems. The safety and efficiency of the proposed scheme are validated through both analysis and test results [15].

3. CONCLUSION AND FUTURE SCOPE

The methodology for an effective approach for securing and validating the Big Data and its integrity has been envisioned in this survey article. The term "big data" refers to the massive amounts of data that businesses process, analyze, and store. Big data was created as a result of the increased usage of information assets and the demand for improved data processing technology. Big data analytics provides service tools like the Hadoop Distributed File System, which helps manage and store large amounts of data, make quick automated choices, and reduce the danger of human predictions. However, the widespread use of big data poses significant security and privacy issues, especially when dealing with sensitive data such as proprietary records confidential info, and medical information. Therefore, to achieve effective preservation of the integrity of the big data, a number of related researches have been analyzed in this survey article. This has been highly useful In reaching our approach for big data preservation which will be effectively elaborated in the upcoming editions of this research article.

4. REFERENCES

[1] C. Chen et al., "An Efficient Privacy-Preserving Ranked Keyword Search Method," in IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 4, pp. 951-963, 1 April 2016, DOI: 10.1109/TPDS.2015.2425407.

[2] X. Xiong, D. Jiang, Y. Wu, L. He, H. Song, and Z. Lv, "Empirical Analysis and Modeling of the Activity Dilemmas in Big Social Networks," in IEEE Access, vol. 5, pp. 967-974, 2017, DOI: 10.1109/ACCESS.2016.2626079.

[3] K. Yang, Q. Han, H. Li, K. Zheng, Z. Su, and X. Shen, "An Efficient and Fine-Grained Big Data Access Control Scheme With Privacy-Preserving Policy," in IEEE Internet of Things Journal, vol. 4, no. 2, pp. 563-571, April 2017, DOI: 10.1109/JIOT.2016.2571718.

[4] L. Watkins et al., "Using semi-supervised machine learning to address the Big Data problem in DNS networks," 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), 2017, pp. 1-6, DOI: 10.1109/CCWC.2017.7868376.

[5] Jin Kim, Nara Shin, S. Y. Jo, and Sang Hyun Kim, "Method of intrusion detection using deep neural network," 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), 2017, pp. 313-316, DOI: 10.1109/BIGCOMP.2017.7881684.

[6] H. Xiong, K. R. Choo, and A. V. Vasilakos, "Revocable Identity-Based Access Control for Big Data with Verifiable Outsourced Computing," in IEEE Transactions on Big Data, DOI: 10.1109/TBDATA.2017.2697448.

[7] Y. Wang, Y. Shen, H. Wang, J. Cao, and X. Jiang, "MtMR: Ensuring MapReduce Computation Integrity with Merkle Tree-Based Verifications," in IEEE Transactions on Big Data, vol. 4, no. 3, pp. 418-431, 1 Sept. 2018, DOI: 10.1109/TBDATA.2016.2599928.

[8] J. Wang, X. Zhang, J. Yin, R. Wang, H. Wu and D. Han, "Speed Up Big Data Analytics by Unveiling the Storage Distribution of Sub-Datasets," in IEEE Transactions on Big Data, vol. 4, no. 2, pp. 231-244, 1 June 2018, DOI: 10.1109/TBDATA.2016.2632744.

[9] B. Hong, H. Kim, M. Kim, T. Suh, L. Xu, and W. Shi, "FASTEN: An FPGA-Based Secure System for Big Data Processing," in IEEE Design & Test, vol. 35, no. 1, pp. 30-38, Feb. 2018, DOI: 10.1109/MDAT.2017.2741464.

[10] H. Cui, R. H. Deng, Y. Li, and G. Wu, "Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud," in IEEE Transactions on Big Data, vol. 5, no. 3, pp. 330-342, 1 Sept. 2019, DOI: 10.1109/TBDATA.2017.2656120.

[11] W. Liao, C. Luo, S. Salinas, and P. Li, "Efficient Secure Outsourcing of Large-Scale Convex Separable Programming for Big Data," in IEEE Transactions on Big Data, vol. 5, no. 3, pp. 368-378, 1 Sept. 2019, DOI: 10.1109/TBDATA.2017.2787198.

[12] J. H. Abawajy, M. Chowdhury and A. Kelarev, "Hybrid Consensus Pruning of Ensemble Classifiers for Big Data Malware Detection," in IEEE Transactions on Cloud Computing, vol. 8, no. 2, pp. 398-407, 1 April-June 2020, DOI: 10.1109/TCC.2015.2481378.

[13] R. Li, H. Asaeda and J. Wu, "DCAuth: Data-Centric Authentication for Secure In-Network Big-Data Retrieval," in IEEE Transactions on Network Science and Engineering, vol. 7, no. 1, pp. 15-27, 1 Jan.-March 2020, DOI: 10.1109/TNSE.2018.2872049.

[14] W. Gao, W. Yu, F. Liang, W. G. Hatcher, and C. Lu, "Privacy-Preserving Auction for Big Data Trading Using Homomorphic Encryption," in IEEE Transactions on Network Science and Engineering, vol. 7, no. 2, pp. 776-791, 1 April-June 2020, DOI: 10.1109/TNSE.2018.2846736.

[15] Y. Zhang, J. Yu, R. Hao, C. Wang, and K. Ren, "Enabling Efficient User Revocation in Identity-Based Cloud Storage Auditing for Shared Big Data," in IEEE Transactions on Dependable and Secure Computing, vol. 17, no. 3, pp. 608-619, 1 May-June 2020, DOI: 10.1109/TDSC.2018.2829880.