A Review on Improving big data security and Service

[1]Ayush Raju Gaikwad, [2]Tejas Bharat Dhotre, [3]Ashutosh Rajaram Barkade,
[4]Pankaj Subhash Chavan,

Computer dept.
Dhole Patil College of engineering, Pune
Guide name -**Prof. Manisha Mehrotra**

***Abstract*- The paradigm of big data has been one of the most interesting concepts that have been evolving in the recent years. The big data is extremely difficult to analyze and effectively process to achieve insightful information. The security management of big data that has extremely high velocity and volume of data becomes a mammoth task. The mismanagement of this data can lead to the theft of confidential or sensitive information that can be highly problematic in the event of a data leak. A number of approaches has been designed to combat this effect but there has been a lack of efficiency in these approaches. Therefore, an effective methodology has been implemented that it effectively classifies and partitions big data queries which are mapped for parallel computation on to a mongo DB database. The approach specified in this research article also utilizes the concept of bilinear pairing through implementation of hashing and the detection of avalanche effect for the purpose of tamper detection and the forensic analysis for effective security report generation.**

***Keywords:* Big data, Parallel Computation, Bilinear Pairing, and Avalanche effect, Data Mapping.**

I. INTRODUCTION

Ever since the introduction of the internet platform the amount of data has skyrocketed. This is due to the fact that the internet platform allows the sharing and integration of a large amount of data all over the globe. This has been significant in the re cent decades which has been referred to as the information age. There is large amount of information on various different fields o n the online platform which has been disseminated copiously over web pages and web applications. This has been significant enough that the large amount of data that is being generated is very difficult to manage and store effectively.

This large amount of data contains information in different formats. These formats can be images media or textual information. As this data is in the unstructured format it becomes highly difficult to analyze and effectively store this large amount of data easily. This data is nowadays referred to as big data which has significant problems for the purpose of storing managing and utilizin g the big data. The big data paradigm has been improving over the past years where different techniques have been developed to store and analyze this data through different tools such as HDFS which allows scalability redundancy parallel processing and reliab ility to effectively manage the big data.

This data is being mapped and reduced to understand and evaluate any insightful information present in it. As there have been various issues in effectively processing this data which can be problematic. The cloud system has been effective in providing the much-needed mobility to this data which can be highly useful. But this data is subjected to and can be the target of various different threats that can lead to the privacy of the individuals of the organizations being at a potential risk, due to the inability of the current security platforms to provide effective and reliable security to big data.

There has been an increased interest in the internet platform and their respective services that are enabled through this platform. These services help the user in achieving an ease of access and increased convenience that can be highly useful for the organizations at the same time. The cloud service offers computational capabilities as well as resources for the purpose of storage and remote access. The services offered by the cloud are effectively categorized into service modules, such as IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service). These services are highly useful and utilizes the internet platform as for the purpose of providing these services to the users.

The data stored on the cloud platform is usually confidential data of the users that is readily required at various times of the day. The most effective and useful strategy the cloud uses is to effectively store the user’s data on a remote server. This data then can be accessed by the user, using the valid credentials anywhere in the world through an internet enabled device. This offers increased mobility to the user while reducing the costs of local storage and its maintenance. The data that is being stored on the plat form needs to be secured, and as there is a large amount of data being stored on these platforms, it becomes increasingly difficult.

The big data, therefore, can be understood as a collection of a large amount of data. This large data is in different formats , such as images, video, text, etc. The data is also coming in at a large speed, which leads to problems processing the data effectively. There are various techniques that allow this processing, but have been ineffective in achieving robust security to the data while processing. This research is focusing on the security of the big data through the use of forensic analysis and bilinear pairing.

There have been a large number of techniques that have been utilized for the purpose of managing and securing the data on the big data cloud platform. There have been a couple of problems in handling the large size and the increased velocity of the data efficiently. The big data queries mirror the size of the dataset as numerous queries need to per processed in order to maintain and manage the data contained in the database. This data needs to be protected to ensure that no tampering is being done on the data which can problematic to identify and manage with efficiency due to the large size.

The integrity in this research article is being maintained through the use of Bilinear Pairing. This forms pairs of hash keys that are generated for the data on regular intervals. These intervals facilitate the state of the data effectively and can be used to identify any problems or modifications in the database. The bilinear pairs are used to identify the avalanche effect that occurs when the data is modified or tampered even in the slightest. This slight variation can lead to extraordinary changes in the Hash keys that can be detected to prove the tampering being performed on the database. This allows for effective security report generation that can be useful for the forensic analysis of the data for the purpose of tamper detection. This approach is further elaborated in much detail in the later sections of this research article.

The next step of this research article details the related researches that have been referred in section 2. The section 3 is reserved for the detailing of the presented approach whereas the section 4 lists the various experimental outcomes and the results and the section 5 provides the conclusion and the future directions for research.

II. LITERATURE SURVEY

Hadeal Abdulaziz Al Hamid et. al. [1] introduced the decoy strategy along with a fog computing facility to protect patients' MBD in the healthcare cloud. It acts as a secondary gallery for decoy MBD (DMBD) that fools the intruder into thinking it's the real thing (OMBD). Unlike other approaches, which call the decoy files when an attacker tries to enter the device, the proposed techniqu e retrieves the decoy files from the start to ensure better security. If an intruder knows that he or she is working with a decoy gallery, it employs a double encryption tactic by encrypting the original file; the attacker will then have to work out how to deciphe r the original gallery. As a result, the presented approach guarantees that the user's MBD is fully safe while still speeding up the operation. There's no reason to be concerned whether the user is an attacker because the decoy big data gallery is open to all by default, while the original one is secret and only made available to a genuine user after successful authentication. An effective tri-party authenticated key agreement protocol DPG, and the OPG, based on paring cryptography has been proposed among the consumer, to facilitate the above process.

Abdur Rahim Mohammad Forkan et. al. [2] developed BDCaM, a groundbreaking architecture framework for context-aware surveillance based on cloud computing. Every AAL system's created background is sent to the cloud. The presented novel approach utilizes a variety of distributed servers in the cloud storage and proceeding certain contexts to retrieve necessary data for decision-making. The authors devise a two-step learning strategy. In the first step, the method looks for associations between background attributes and vital sign threshold values. Using the MapReduce Apriori algorithm, the method produces a series of associatio n directives that are unique to each patient based on long-term background details. In the second step, the approach employs supervised learning over a new large set of context data generated using the rules discovered in the first step. As a result, the system improves its ability to predict any medical condition with greater accuracy. The experimental evaluation of the presented system in a cloud model for patients with differing HR and BP levels has demonstrated that the system can predict precise irregular conditions in a patient with great accuracy and in a short period when properly fitted with large samples.

Yue Deng et. al. [3] proposed a merged fuzzy deep neural network (FDNN) that collects data at the same time from both fuzzy and neural presentation. The information gained from these two perspectives is combined in a fusion layer to form the final data classification representation. Fuzzy representation eliminates uncertainties, thus neural representation extracts noise from the authentic results. For final grouping, the introduced FDNN uses these two superior representations to form a fused representa tion. As a consequence, FDNN could be ideal for more complex pattern recognition tasks containing unclear data and noise. The contrasts with other non-fuzzy deep neural networks demonstrate that fuzzy learning is a possible way to improve deep learning's results.

Jianguo Chen et. al. [4] introduced a parallel random forest algorithm for huge data. The PRF algorithm's consistency is impr oved by using a weighted vote strategy and lessen the number of proportions. Following that, a hybrid parallel PRF approach implemented a data-parallel and task-parallel optimization framework based on Apache Spark. The training dataset is reutilized, and the data capacity is greatly decreased, due to data-parallel optimization. The data processing cost is beneficially minimized, and the algorithm's efficiency is clearly increased, due to task-parallel optimization. The outcome of the experiments shows that PRF performs better than other algorithms in a condition of classification precision, efficiency, and scalability.

S. Bhattacharjee et. al. [5] proposed a new automated method for dealing with the problems of moving large files in a combinatorial manner The suggested process involves SDES (Simplified Data Encryption Standard) and a modern and enhanced data encryption standard to improve the pattern generation, use a less complicated pattern generation strategy. To maintain data secrecy and to provide a backup in the event of an unintentional data loss there is a data failure that makes use of a new pattern generation bench, distinguishes it from the rest of the strategies. The proposed strategy uses a novel dual round error management strategy that fights different types of transmission errors. By accommodating any amount of error bits, this approach overcomes the shortcomings of current error management techniques. The current work proposed a unique lossless compression strategy for reducing data size. Furthermore, this approach can address existing problems in both lossy and lossless data compression techniques. Since high compression performance usually compromises data secrecy, this analysis includes advanced LSB-based audio steganography as a key component. The outcome shows that the presented combined technique has higher SNRs, avalanche effects, and entropy values, as well as lower amplitude variations between the steganographic and cover audio files when differentiating from other existi ng steganographic techniques.

A. Rosa et. al. [6] introduced new online prediction models that can accurately classify missed executions in big data networks, with an emphasis on complex interdependencies between jobs and incidents. Besides, the authors have developed two independent NNs that split jobs and events into four types, as well as a nested NN that forecasts all jobs and events. The writers analyze the proposed models using a variety of features and data reduction techniques. Compared to other classifiers, such as LDA, ELDA, QDA, LR, and SVN, the results of the experimental evaluation demonstrated a substantial improvement inaccuracy. Overall, the study provided does not only identify the root causes and observations required to develop a fault and latency -aware scheduling policies, but also versatile prediction models that enable the system to make proactive decisions.

S. Atiewi et al. [7] proposed a scheme by using multifactor authentication and lightweight cryptographic to create a stable c loud–IoT setting. The IoT systems are divided into responsive and nonsensitive devices using the suggested process. The authors advocate the use of a hybrid cloud that blends public and private cloud services. The RC6 and Fiestel encryption algorithms are used to divide and encrypt sensitive computer data. To have high confidentiality, these data are stored in a private cloud and accessed via a gateway system. The nonsensitive user data, on the other hand, is encrypted with AES and saved in the public cloud through a gateway device. The TA offers multifactor authentication. The user goes through three layers of authentication in this process by presenting their credentials, such as their user ID, password, and biometrics. The authors use metrics that include processing time, security power, encryption time, and decryption time to test the proposed method's efficiency. The experimental outcome demonstrates that the proposed approach outperforms FCS, CP-ABE, and MCP-ABE based on the comparison findings.

R. Jiang et. al. [8] uses brainstorming, Delphi, questionnaires, interviews, and field testing to investigate the security and privacy risk metrics of medical big data processing, delivery, storage, use, and distribution. Subjective variables can be mitigated to a degree using a variety of techniques. Then, using a combination of the GI method and the entropy weight method, the detailed weight of secondary indicators is determined. The combination of these two approaches reduces not just the effect of contextual variabl es, but also the lack of evidence hidden in quantitative data, resulting in more reliable assessment outcomes. The fuzzy comprehensive assessment model is then applied. The fuzzy robust assessment model is then used to assess the degree of risk associated with the protection and privacy exposure of medical big data. Finally, risk mitigation solutions for the four phases of data acquisition, transmission, preservation, utilization, and sharing in the medical big data life cycle are introduced.

J. Moreno et. al. [9] introduced an idea for a mechanism to integrate protection into the formation of a big data ecosystem. From research to execution, this method spans the usual stages of a production process. Furthermore, this process was developed in light of the new business environment, in which many businesses are evolving their internal cultures to embrace principles such as agile methodologies. This process is aided by an SRA, which serves as a meta-model for the various components that make up a big data ecosystem, enabling abstraction and thereby facilitating the creation of such a complex world. Finally, to explain how to utilize the SRA, the authors created an example that demonstrates the key components of SRA as well as how the protection trends can be used to combat the various threats that the presented environment faces.

Y. Gao et. al. [10] has implemented the BDPM, a large data source architecture built on the PROV-DM data security monitoring model. The BDPM model refines the definitions of entities, activities, and agents according to the data forms of the big data and the elements of the big data architecture and expands the original relationship to the enrichment of origin analysis features while retaining the core framework of the PROV-DM. The BDPM paradigm in the Big Data Architecture enables the representation of the source of different data types in several levels of data organization, as well as the representation of the source of the whole data transition process across different storage, processing and communication components. The authors present data protection supervision approaches focused on provenance graph analysis, such as inference rule-based analysis, vertical provenance analysis, and horizontal provenance analysis, using the BDPM model to enforce the restrictions that a valid provenance graph can satisfy.

W. Zhong et. al [11] introduced the BDHDLS technique that use to learn the distinct data distribution of unique invasive attacks belonging to specific families. This technique is particularly useful for identifying subtle data patterns in disruptive atta cks with a restricted number of samples. Both behavioral and material attributes are used by the BDHDLS. The BDHDL S can interpret disruptive attack samples using both network traffic attributes and payload contents because it takes both behavioral and mat erial aspects into account. Since previous methods never combined all types of functionality, this technique will help IDS work better. The presented research also demonstrates that using big data techniques and concurrent methods for feature discovery, cluster ing, and training will greatly reduce model development time. Researchers can iterate more easily to find the right model parameters for their theoretical problems. The contribution of different deep learning models in the cluster is combined using a basic decis ion fusion technique in this research.

S. Fiore et al. [12] presented the city administration dashboard, a general public transportation analytics program construct on top of the EUBra-BIGSEA system to satisfy the needs of the Brazilian municipality of Curitiba. A massive and quick data analytics infrastructure, a scalable and dynamic cloud architecture, data quality-aware modules, security and privacy strategies, a rich programmable interface layer, and a web GUI with multiple views are all defined in the proposed approach. Furthermore, the combination of PyOphidia and PyCOMPSs allows for quick parallel program creation while hiding the underlying infrastructure sophistication, parallelization, and resource management aspects. Furthermore, the whole application was created with simplicity and generality in mind to facilitate replicability in other cities. It is a significant output of the EUBra-BIGSEA program, which has been deployed and demonstrated in a true trans-Atlantic testbed across Europe and Brazil.

S. Sirapaisan et. al. [13] introduced novel Communication Pattern-based Data Authentication (CPDA) architecture, which designed to provide the highest degree of data identity security at the smallest granularity with the least amount of overhead expense . The CPDA's architecture combines many concepts: first, communication pattern-based authentication data and communication aggregation, which employs multiple aggregation methods, each for a distinct communication pattern, second a dual-use of different message authentication mechanisms, such as hash functions, Message Authentication Codes (MACs), and digital signatures, in which computationally less expensive mechanisms are used to protect individual objects that are transferred between untrustworthy and trustworthy entities, and computationally more expensive mechanisms are used to protect individual objects that are transferred between untrustworthy and trustworthy entities and third each object's authentication data is tailored in such a way that var ious objects can be checked independently. Regardless of which item is generated by which manufacturer or consumed by which user, the authors can increase the degree of security while minimizing overhead costs in this manner.

Y. Chen et. al. [14] introduced a cloud network and mathematical model measurement scheme for large data base plane dynamics that is, building a marginalized fog computing layer between the medical service terminal and the computing center to detect clinical medical equipment and input the calculated medical data into this computing layer. To resolve the issue of lengthy waits, reduce data processing time, increase clinical patient care quality, and enhance customer experience. Simultaneously, the authors su ggest a fog layer network mode in which they can spread computing activity processing, based on the CPSO-LB load balancing approach to manage the fog network's load. This can help minimize clinical time spent and needless expenditures. The experimental outcome show that by utlizing the constrained particle swarm optimization load balancing (cpso-lb) algorithm, the distributed computing scheme will accomplish the goal of minimizing task processing delay. The experimental outcome demonstrate that the complex big data cloud network can satisfy the needs of current big data, and that the consumption rate of big data can be increased by up to 90%.

III. CONCLUSION AND FUTURE SCOPE

The proposed methodology for effective security report generation and management of big data queries through parallel computation on mongo DB database has been specified in this research. In this approach the system takes the input big data queries that are numerous in nature that are used for effectively handling and managing the big data. These queries are processed and classified into various methods according to the queries such as insert update or delete. These queries are effectively partitioned and are clubbed together for the purpose of mapping. This mapping ensures that the approach can be translated in two threads for the purpose of parallel computation consecutively. This allows the effective storage and management of data onto the mongo DB database. The integrity of the big data is realized through the utilization of hashing and implementation of bilinear Pairing. Through this approach pairs of hash keys are made and evaluated at regular intervals to notice any indication of an avalanche effect. Once this avalanche effect has been detected which indicates any tempering done on the database, a forensic analysis report is generated and provided to the user.

REFERENCES:

1. Hadeal Abdulaziz Al Hamid, Sk Md Mizanur Rahman, M. Shamim Hossain, Ahmad Almogren and Atif Alamri, "A Security Model for Preserving the Privacy of Medical Big Data in a Healthcare Cloud Using a Fog Computing Facility With Pairing-Based Cryptography", IEEE Access (Volume: 5), 28 September 2017.
2. Abdur Rahim Mohammad Forkan, Ibrahim Khalil, Ayman Ibaida and Zahir Tari, "BDCaM: Big Data for Context-Aware Monitoring—A Personalized Knowledge Discovery Framework for Assisted Healthcare", IEEE Transactions on Cloud Computing (Volume: 5, Issue: 4, Oct.-Dec. 1 2017), Dec 2017.
3. Yue Deng, Zhiquan Ren, Youyong Kong, Feng Bao, and Qionghai Dai, "A Hierarchical Fused Fuzzy Deep Neural Network for Data Classification", IEEE Transactions on Fuzzy Systems (Volume: 25, Issue: 4, Aug. 2017), Aug 2017.
4. Jianguo Chen, Kenli Li, Zhuo Tang, Kashif Bilal, Shui Yu, Chuliang Weng, and Keqin Li, "A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment", IEEE Transactions on Parallel and Distributed Systems (Volume: 28, Issue: 4, April 1, 2017), April 2017.
5. S. Bhattacharjee, L. B. A. Rahim, J. Watada, and A. Roy, "Unified GPU Technique to Boost Confidentiality, Integrity and Trim Data Loss in Big Data Transmission," in IEEE Access, vol. 8, pp. 45477-45495, 2020, DOI: 10.1109/ACCESS.2020.2978297.
6. A. Rosa, L. Y. Chen and W. Binder, "Failure Analysis and Prediction for Big-Data Systems," in IEEE Transactions on Services Computing, vol. 10, no. 6, pp. 984-998, 1 Nov.-Dec. 2017, DOI: 10.1109/TSC.2016.2543718.
7. S. Atiewi et al., "Scalable and Secure Big Data IoT System Based on Multifactor Authentication and Lightweight Cryptography," in IEEE Access, vol. 8, pp. 113498-113511, 2020, DOI: 10.1109/ACCESS.2020.3002815.

8. R. Jiang, M. Shi and W. Zhou, "A Privacy Security Risk Analysis Method for Medical Big Data in Urban Computing," in IEEE Access, vol. 7, pp. 143841-143854, 2019, DOI: 10.1109/ACCESS.2019.2943547.
9. J. Moreno, E. B. Fernandez, M. A. Serrano, and E. Fernández-Medina, "Secure Development of Big Data Ecosystems," in IEEE Access, vol. 7, pp. 96604-96619, 2019, DOI: 10.1109/ACCESS.2019.2929330.
10. Y. Gao, X. Chen and X. Du, "A Big Data Provenance Model for Data Security Supervision Based on PROV -DM Model," in IEEE Access, vol. 8, pp. 38742-38752, 2020, DOI: 10.1109/ACCESS.2020.2975820.
11. W. Zhong, N. Yu, and C. Ai, "Applying big data-based deep learning system to intrusion detection," in Big Data Mining and Analytics, vol. 3, no. 3, pp. 181-195, Sept. 2020, DOI: 10.26599/BDMA.2020.9020003.
12. S. Fiore et al., "An Integrated Big and Fast Data Analytics Platform for Smart Urban Transportation Management," in IEEE Access, vol. 7, pp. 117652-117677, 2019, DOI: 10.1109/ACCESS.2019.2936941.
13. S. Sirapaisan, N. Zhang and Q. He, "Communication Pattern Based Data Authentication (CPDA) Designed for Big Data Processing in a Multiple Public Cloud Environment," in IEEE Access, vol. 8, pp. 107716-107748, 2020, DOI: 10.1109/ACCESS.2020.3000989.
14. Y. Chen and Z. Qiu, "Cloud Network and Mathematical Model Calculation Scheme for Dynamic Big Data," in IEEE Access, vol. 8, pp. 137322-137329, 2020, doi: 10.1109/ACCESS.2020.3009675.