



ANOMALY DETECTION IN HYBRID CLOUD DATA PROCESSING

Rajendra Baleshwar^{1*}, Dr. Ramesh Vishwakarma²

¹Research Scholar

²Guide, Rabindranath Tagore University, Bhopal.

*Corresponding Author

Rajendra Baleshwar

Research Scholar.

Article History

Received: 08.09.2024

Accepted: 19.09.2024

Published: 17.10.2024

Abstract: *The cloud is extensively utilized, yet it comes with its own set of concerns and threats to users' privacy. When compared to more conventional computer frameworks, the shared computing infrastructure known as the cloud presents a significant number of security vulnerabilities. It is never feasible to have complete faith in the cloud provider or tenants who store their data in a shared environment since there is always the possibility that they may leak data at any point of the compute process. Because of these problems with cloud computing, there is a requirement for a secure framework to carry out analytics activities, and cryptographic tools need to be developed in order to safeguard sensitive data while it is being stored and while it is in transit. The majority of companies have a dilemma when it comes to protecting large amounts of data while it is being stored.*

Keywords: cloud provider, data, protecting, privacy.

Cite this article:

Baleshwar, R., Vishwakarma, R., (2024). ANOMALY DETECTION IN HYBRID CLOUD DATA PROCESSING. *ISAR Journal of Science and Technology*, 2(10), 64-69.

Introduction

Cloud computing is becoming increasingly popular as a result of the numerous advantages it offers in comparison to more conventional distributed computing methods. Elasticity, extensibility, awareness, low power consumption, and decreased initial investments are some of the most outstanding characteristics of cloud technology. As a result of these benefits, cloud computing has become the dominant computing model that is commonly chosen for the storing and analysis of large amounts of data. The security of data is becoming an increasingly pressing worry as cloud storage becomes more widespread. It is becoming increasingly likely that sensitive data belonging to users will be rendered into the cloud computing environment as a result of the rapid rise of data and the progression of multifunctional cloud services. In order to ensure the safety of the data, it is encrypted before being saved in the cloud storage. In accordance with the explanation provided by Ranjan et al. (2015), cloud computing is characterized by the accumulation of a substantial pool of virtualized information and communication technology facilities. These facilities include hardware resources such as a central processing unit, unlimited storage, and network bandwidth, among others, as well as software resources such as file servers, application servers, and web servers. Cloud computing is becoming increasingly used as a means of accumulating and applying calculations to large amounts of data. As a result of its low cost, limitless network access, scalability, resource pooling, and metered service, cloud computing has emerged as the most effective instrument for the storing and analysis of large amounts of data.

In Strohbach et al., (2016), the authors discussed the challenges and issues in storing colossal datasets in a “secured” and “privacy-preserved” manner. The work highlighted the salient features on an

ideal big data storage system like the ability to store any amount and nature (homogenous and heterogeneous) of data, effective retrieval and searching mechanisms, fault tolerance, optimal data transfers, security and privacy of data etc. They also gave the overview of the “social” and “economic” implications of big data “storage-technologies” on the basis of case studies ranging from healthcare, finance and energy sectors. The work dives deep into discussing big data storage technologies, big data query platforms like “HBase”, “HIVE”, and “Impala” etc. Finally, the work puts forward some of the best practices for implementing security features in big data management systems and was concluded by highlighting the requirements for effective and efficient storage of huge datasets.

In Park and Kim (2018), the authors introduced a new prototype “large capacity storage device” using the “network direct connection storage” device in order to effectively store and process big data. The work began with a description of bottlenecks in big data storage tools and infrastructures. The prime focus was on the network and storage performances. In their proposed prototype, there was no CPU or memory provided at the network side. The storage and I/O were handled by self-designed protocol suite itself using protocol for the exchange of data and thus no network stack was present in their approach. They performed the empirical evaluation for validating the performance and scalability enhancement of their proposal. Finally, they concluded that the performance of their system is directly proportional to the number of disks in the device. The more the number of disks, the better is the performance. However, they did not mention about the security measures taken to ensure the security of the stored big data. Without security mechanisms, the storage of data involves high risks and thus their approach cannot be applicable in general big data storage systems which require secured storage of data.

The authors in Zhou et al., (2018) proposed and implemented a novel "Preference Aware HDFS (PAHDFS)" system with an aim to improve the performance of classical Hadoop system in hybrid environments. The primary concern addressed in their work was the imbalance of storage and its effects on the performance of the system when the data is stored in a distributed manner. They first described the properties of "Hard Disk Drives (HDDs)" and "Solid State Drives (SSDs)". The limitations of Hadoop that it considered all storage devices having the same I/O throughput and distributes the data across all clusters in an even manner were highlighted. They also mentioned that a distributed system may have multiple kinds of storage devices and each of these devices largely differ in the format, capacity and capabilities of handling data. Thus it is necessary to store the data in the most appropriate storage device in order to achieve optimal performance. The proposed model works by tracking the read/write characteristics of the data and optimally chooses the most suitable device in accordance with the matching read/write capacities of the device thereby increasing the read/write efficiency.

Faraway and Augustin (2018) argued that big data is not always useful as compared to small data. They provided several situations wherein small data is preferred over big data to draw much better inferences. The work started with defining big and small data and then the situations when small data can overpower big data were highlighted. The authors claimed that a well-organized small data is more useful than an unorganized big data in many cases. Secondly, the cost involved in managing (acquiring, storing, securing) big data is much higher as compared to managing small data. Thirdly the heterogeneity and uncertainty attached with big data makes it intricate to handle. However, no such problems are associated with small data. Finally, the work concluded with describing the problems and challenges associated with mining the relevant information for the huge datasets.

Xuan et al., (2017) proposed a new two-level storage system by incorporating the "upper-level in-memory file system" and "lower-level parallel file system". They used "Tachyon" and "OrangeFS" to build their prototype. Tachyon is an in-memory file system developed in Java while OrangeFS is an open source file system which supports parallelism at its core. To describe their approach the authors first highlighted the basic functioning and know-how about data intensive High Performance Computing (HPC) paradigms and the related tools and technologies used in HPC. Their work was aimed at improving the performance of HPC when integrated with Hadoop System. There were two prime categories of nodes in the conventional HPC system. First are "data-nodes" which store the actual data contents and the other are "compute-nodes" which perform the processing on these data. In their approach, Tachyon was deployed on "compute nodes" while OrangeFS was deployed on "data nodes". OrangeFS has the responsibility of storing a replica of the data to recover in case of a fault or data losses. A tabular comparison of access performances of the prominent HPC facilities was provided in their work. The experimental evaluation carried out by them, claimed to improve the throughput of the system.

The work Subbu and Vasilakos (2017) surveyed the various context-aware big data processing systems. They discussed the various challenges and issues that arise in context-aware

computing environment. In order to provide a comprehensive review of the state-of-the-art approaches in context-aware computing techniques the authors first gave a brief insight about big data and its properties and domains of existence. Furthermore, the meaning of context was defined in terms of characteristics like computing, users, time and physical etc. The primary aim of the work was to characterize context-aware systems with respect to sensing criteria of the sensors along with highlighting the bottlenecks of the present context-aware system. The work also puts forward two case studies of context-aware systems in healthcare and agricultural domain.

Zhang et al., (2018) reviewed the recent researches involving deep learning models for big data. The work covered several application domains of deep learning technology in conjunction with IoT ecosystem like healthcare, transportation, utilities for modern consumer conveniences etc. The various applications of deep learning in conjunction with big data were also highlighted in their work. The authors stressed upon the ability of deep learning technology to bring forth unconventional and previously undiscovered knowledge for the huge datasets which can be crucial in taking better and informed decision at every level. The work further discussed various deep learning models and finally, the challenges in adopting deep learning techniques with big datasets were highlighted.

Rahman and Esmailpour (2016) proposed a hybrid architecture using an integration of both electrical and optical networks components to leverage cloud computing and Big Data applications. They argued that conventional electrical signal links were not suitable for handling huge datasets and frequent I/O requests as they consume significant amount of energy. With the emergence of IoT technology, the rate at which the data is produced, stored and retrieved has increased exponentially and conventional electrical links were unable to provide required bandwidths with minimal latencies and losses. Another limitation of electrical links highlighted by them include the significant amount of cost involved in cooling the system which gets heated while transferring huge datasets from one link to another in the distributed network. Thus in order to overcome these limitation their work provided a means of optical links between the network components so as to make the data transfer faster while consuming less amount of energy. In order to do so, they examined the traffic pattern in the data centers running on cloud technology on the basis of crucial parameters like packet size, data transfer window size, arrival rate etc. Finally, they evaluated the performance by comparing their proposal with the classical and other similar architectures. The experimental results as claimed by them can reduce the overall costs of data centers. Specifically the work claimed to improve the delay rate by reducing it to nearly 39% while significantly decreasing the cost incurred in cooling the system from 49% down to 27%.

Need for Cloud Storage in Big Data Analytics

Cloud computing offers a number of advantages, including the fact that services and cloud storage are favored for the storage and analysis of large amounts of data. The following table provides a description of some of the qualities that should be considered when selecting cloud storage for big data analytics.

Table 1: Cloud Characteristics

Cloud Characteristics	Description
Pay-as-you-use	The use of cloud computing is preferred because it may be able to use resources as services, receive additional resources when the demand for them develops, and payments may be required based on the amount of resources and computing platforms that are utilized.
Extensive network	The owner of the data that is stored in the cloud may be utilizing a wide range of different devices, including but not limited to personal computers, digital assistants, smart phones, and other electronic devices. In order to receive data from all of these devices via the internet, the network protocol has to be a common one.
Pooled resources	A pool of shared resources that may be accessed whenever they are needed is provided by cloud computing. When using a multi-tenant architecture, several users are able to share resources, and no one user is able to lock in any of those resources. It is necessary for the cloud's resource allocation strategy to have the capability of assigning resources to users, subtracting resources from users, and reassigning resources to users who have requested them.
Scalability	The potential of cloud computing to increase in size is likely the most important capability it possesses. Customers that utilize cloud computing are provided with the guarantee that they will have access to services that are acceptable and have an endless capacity. Evidently, the cloud is responsible for locating and assigning resources to users.
Accounting	The cloud services are supplied on a metered charge basis, and the user of the cloud is paid based on the amount of resources that are utilized. This results in cost reductions. This resulted in the requirement for accounting as well as the provision giving users the ability to monitor consumption patterns, calculate costs, and access accessible services.

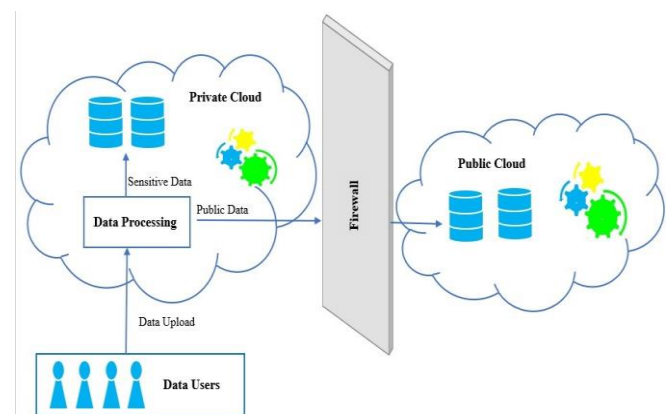
Need for Hybrid Cloud

For the purpose of storing and processing large amounts of data, many cloud computing models, including private cloud, public cloud, and hybrid cloud, are utilized, depending on the requirements. When it comes to storage and analytics, private clouds are preferred by enterprises that deal with very sensitive data. Due to the fact that private cloud is designed for usage by a single business, the danger to the organization's security and privacy is significantly reduced. Despite this, private cloud computing does have some restrictions when it comes to the processing of data. The scalability concern is the first limitation that must be considered. The private cloud that is run on-premises does not take into consideration the flood of data, and as a result, it has limited scalability. For the purpose of constructing private clouds that are highly scalable, it is necessary to make a substantial initial expenditure with the intention of acquiring and installing computing resources. When it comes to precisely planning the storage limit of a private cloud, however, the dynamic properties of big data, such as increasing volume, velocity, and diversity, result in issues. As a consequence of this, the capacity of private clouds is perpetually either fully or partially provided. Developing a private cloud that has restricted scalability is being done primarily for the purpose of lowering the amount of money that is being invested in its creation.

Limitations in analytics are yet another problem associated with private cloud computing. In order to avoid incurring greater operating expenses, private cloud computing does not provide widespread support for the processing models and platforms that are required for analyzing and preserving heterogeneous data. The third disadvantage of private cloud computing is that it does not allow for the efficient sharing of data among users who have been permitted to access it. The exchange of information with is always required. Colleagues who do not have permission to utilize private clouds are required to have access to the data provided. In spite of the fact that private cloud computing offers certain advantages in

terms of security, it is not frequently selected for big data analytics due to the aforementioned problems.

When contrasted with private clouds, public clouds offer support for processing large amounts of data and facilitate the efficient sharing of data among users who have been permitted to access it. Furthermore, the public cloud service providers that are now in operation and the data analytics capabilities that are currently accessible do not have any means for recognizing or monitoring data outflow. In recent years, the mechanisms for auditing cloud data by a third party auditor, data protection, and privacy preservation have emerged as key methods for preserving data stored in the cloud. In order to develop a reliable framework for big data analytics, it is crucial to include the main features that are offered by both public clouds and private clouds. The purpose of this study is to offer a technique for moving data to hybrid infrastructure, which is comprised of both public and private clouds, as seen in Figure 1.

**Figure 1: Data Distribution in Hybrid Cloud**

Challenges of Existing Schemes

Despite the fact that cloud computing is the most popular option for big data analytics, there are a lot of difficulties that are related with cloud characteristics. The outsourcing of user data presents

the most significant challenge. Through the process of outsourcing, the obligation of carrying out computations and ensuring that data storage is maintained is delegated to a reliable third party organisation. The primary advantage of outsourcing is that it frees users from the responsibility of generating and controlling data storage. This is the most significant advantage. Therefore, this is a better option for storing a vast amount of different types of data. As a result of outsourcing, the data owner's control over the data is diminished, and the data becomes more vulnerable to security concerns. When the user's sensitive data is outsourced, it is possible that it will be exposed to assaults such as modification, deletion, and replay. Many of the currently available security architectures are not capable of supplying data with assured security. Public cloud storage is not the best option for storing and processing sensitive data because of the security concerns that are associated with it. There is a requirement for the protection of big data that has been outsourced, and the architecture that has been provided offers a reliable hybrid cloud architecture for the processing of big data.

Computing in the cloud allows for multi-tenancy, which refers to the capacity for different users to coexist while sharing shared resources. In order for cloud suppliers to get the financial benefits of cloud computing, the data of all rent users is stored on the same storage environment. In spite of the fact that multi-tenancy offers advantages to cloud providers, it poses security risks to the data that is kept in the cloud. Users who are not permitted to use the cloud storage service obtain data by utilizing the multi-tenancy attribute. In order to address this problem, the work that is being presented offers identity-based encryption as a prevention measure against unwanted access.

A great number of firms are currently in the process of developing their very own storage arrangements in order to mitigate the dangers that are linked with cloud storage. The process of developing and maintaining data storage on one's own premises is both challenging and expensive. Cloud storage is favored because it alleviates the agony of handling the tiresome tasks associated with cloud storage. In any case, the implementation of security procedures is necessary in order to move data to the cloud, which is managed by many providers.

Attribute Based Encryption

In order to offer a system for access limitation that is dependent on the properties of the environment, the user, or the actual data, attribute-based encryption, also known as ABE, is frequently utilized. As a form of asymmetric encryption, ABE is seen as a method that offers protection to cloud storage applications that cannot be trusted. ABE approaches do not only provide well-defined data access restrictions, but they are also efficient. For the purpose of implementing ABE, the count of characteristics is used to determine the level of difficulty associated with encryption and key organization. The performance of encryption is unaffected by the number of users that are added or removed from the security system. When it comes to ABE, the difficulty of encryption and key management is directly proportional to the number of clients, and it is only related with the total number of system characteristics. Whenever the owner of the data makes the decision to revoke a user by withdrawing the delegation that was given, it causes problems in the system that controls access. In ABE, ciphertext is not simply generated for a single user; rather, secret

keys and encrypted text are associated with a collection of characteristics or a rule that is positioned above the attributes. However, in order to decode the ciphertext, it is necessary for there to be a connection between the private key and the ciphertext. Searching over encrypted data is made possible by this approach, which is mostly utilized for the purpose of accomplishing and distributing data in cloud environments. It is possible to think of attributes as keywords, and they allow for keyword-based searches to be performed on encrypted data.

It is difficult to effectively execute user revocation in ABE due to the fact that every property is shared by a large number of users. The owner of the data must re-encrypt all of the details connected with the retracted attributes, and the secret keys of all of the current users must be changed. Following these steps, the data owner will be able to retract attributes from permitted individuals. Not only does this result in the execution of complicated procedures, but it also results in significant processing and transmission costs, and it requires the data owner to remain online continually without interruption. Proxy Re-Encryption is the preferable method and the one that is utilized in this study in order to overcome the disadvantage of ABE for the purpose of securing cloud storage and supporting secure data sharing.

Homomorphic Encryption

Because of virtualization and the multi-tenancy idea, public clouds are more likely to have greater privacy concerns than private clouds. Due to the fact that cloud storage is mostly shared by a large number of users, the likelihood of data loss is actually increased. Encryption and storage of data in cloud storage are both accomplished through the utilization of a method known as fully homomorphic encryption. In order to avoid the necessity for decrypting data across the whole life cycle of data, calculations may also be performed on encrypted data. This eliminates the requirement for data to be decrypted. Due to the fact that homomorphic encryption allows for straightforward deterministic computations to be performed on encrypted data, it is an appropriate method for the development of privacy-preserving protocols. Homomorphic encryption, despite the fact that it offers complete security throughout the life cycle, is associated with a high computing cost and is extremely challenging to implement in terms of both complexity and space. Homomorphic encryption has a number of drawbacks, including a high computing cost and a space complexity. The work that is described in this thesis addresses these issues.

Data Auditing Process

When large amounts of data are transferred to cloud storage, the owner of the data automatically loses control over the data. Because there is a lack of complete trust in the cloud server, the dangers that are linked with data that is outsourced are quite significant. An assurance must be provided to the owner of the data that the data saved in the cloud is stored correctly and is not altered while it is stored in the cloud. Additionally, in order to prevent unwanted access, cloud storage should place greater emphasis on access control techniques. Data auditing may be performed by a third party auditor, as seen in Figure 2, in order to ensure that the data that is kept in the cloud is accurate and complete. The procedure in question requires the downloading of data for the purpose of integrity verification, which is not appropriate for big data applications that include a significant amount of data.

The current cryptographic techniques may not be enough for the storage systems that have been developed recently, despite the fact that they guarantee security. They are unable to use existing schemes to safeguard large volumes of outsourced big data because of the restrictions that are inherent to those schemes. The efficacy and availability of cloud-based services are contingent upon the substantial network, memory space, resources, and power consumptions that are required. It is necessary that the traditional cryptographic tools chosen for cloud security also meet the requirements. On account of the cloud's lack of trustworthiness, the owner of the data typically opts to encrypt the data before outsourcing it to cloud servers. The utilization of typical asymmetric algorithms for the purpose of encrypting enormous amounts of massive data creates a great deal of complexity due to the limited computational capabilities of the user side. In addition, the algorithms that are now in use necessitate the use of public key infrastructure and intricate certificate management operations in order to safeguard data from being accessed by unauthorized individuals.



Conclusion

Cloud computing offers software, platform, and software as services on demand. Cloud's flexibility, scalability, and data sharing make it ideal for big data analytics. Three basic cloud deployment models—public, private, and hybrid—have pros and cons. On-demand public cloud prices are dependent on consumption. Private cloud is for data-secure organizations. Private and public clouds work together to distribute data and process it in parallel. Hadoop MapReduce framework executes processing mechanisms using map and reduce operations. Apache Spark outperforms Hadoop due to in-memory processing. It also executes fast-finishing jobs in clusters.

References

1. Krishna Reddy and Reddy L.S.S (2011) Security Architecture of Cloud Computing, International Journal of Engineering Science and Technology (IJEST), 3(9), pp.7149-7155.
2. Mahbub Ahmed, Yang Xiang, Shawkat Ali, (2010) Above the Trust and Security in Cloud Computing: A Notion towards Innovation, International Conference on Embedded and Ubiquitous Computing, pp.723-730.
3. McKinley, Samimi, Shapiro and Chiping T (2006) Service Clouds: A Distributed Infrastructure for Constructing Autonomic Communication Services. Dependable, Autonomic and Secure Computing, IEEE, pp.341-348, 2006.
4. Mohamed Firdhous, Osman Ghazali and suhaidi Hassan (2014) Statistically Controlled Robust Trust Computing Mechanism for Cloud Computing, Journal of Information Communication Technology, pp: 21–36.
5. Monsef, Gidado, (2011) Trust and Privacy Concern In The Cloud, European Cup, IT Security for the Next Generation, pp.1-15.
6. Pearson S, Casassa Mont M, Crane S (2005), Persistent and Dynamic Trust: Analysis and the Related Impact of Trusted Platforms. In: Trust Management, Proc. iTrust, LNCS 3477, 355-363.
7. Philipp Zech (2011) Risk-Based Security Testing in Cloud Computing Environments, Fourth IEEE International Conference on Software Testing, Verification and Validation, pp.411-414,
8. Rana, Warnier, Quillinan, and Brazier (2008) Monitoring and Reputation Mechanisms For Service Level Agreements. In Proceedings of the 5th International Workshop on Grid Economics and Business Models.
9. Sachdeva (2011) Cloud Computing: Security Risk Analysis and Recommendations, Master Thesis, University of Texas, Austin.
10. Sajjad Hashemi1(2013) Cloud Computing Technology: Security And Trust Challenges, International Journal of Security, Privacy and Trust Management 2(5), DOI : 10.5121/ijspmt.
11. Siani Pearson (2012), Privacy, Security and Trust in Cloud Computing, HP Laboratories, appeared as a book chapter by Springer, UK
12. Singh S and Morley C (2009), Young Australians privacy, security and trust in internet banking. In: Proceedings of the 21st Annual Conference of the Australian Computer-Human interaction.
13. Subashini and Kavitha(2010), A Survey On Security Issues In Service Delivery Models of Cloud Computing, Elsevier, Network and Computer Applications, Vol. 34, pp 1-11.
14. Takabi H, Joshi J, Ahn G (2010) Security and Privacy Challenges in Cloud Computing Environments, IEEE Security and Privacy 8(6), pp 24–31.
15. Talal H. Noor, Quan Z. Sheng (2016) CloudArmor: Supporting Reputation-Based Trust Management for Cloud Services IEEE transactions on parallel and distributed systems, vol. 27, no. 2, pp 367-380.
16. Tian Li-Qin and LIN Chuang (2010) "Evaluation of User Behavior Trust in Cloud Computing" International

- Conference on Computer Application and System Modeling, Beijing East, China, volume-7, pp.v7-567-v7-572.
17. Tweney, A. and Crane, S (2007) Trust Guide2: An Exploration of Privacy Preferences In an Online World, Expanding the Knowledge Economy, IOS Press.
18. Welten (2009) (Rob). Towards The Cloud-The Role Of Trust And Perceived Privacy Risk On The Adoption Of Cloud Computing, Master Thesis, Tilburg University, Netherlands.
19. Zheng Yan, Xueyun Li, Mingjun Wang and Athanasios V. Vasilakos (2015) "Flexible Data Access Control based on Trust and Reputation in Cloud Computing", IEEE Transactions on Cloud Computing.
20. Zissis, D. Lekkas, (2012) Addressing cloud computing security issues, Future Generation Computer Systems, 28(3), pp. 583–592.
21. Chen M., Mao S., Zhang Y., Leung V. C., et al., (2014b), Big data: related technologies, challenges and future prospects, volume 96. Springer, 2014b.
22. Singh D. and Reddy C. K., (2015), A survey on platforms for big data analytics, Journal of big data, 2(1), 8.
23. Chen M., Mao S., and Liu Y., (2014a), Big data: A survey, Mobile networks and applications, 19(2), 171–209.
24. Dou W., Zhang X., Liu J., and Chen J., (2013), HireSome-II: Towards privacy-aware cross-cloud service composition for big data applications, IEEE Transactions on Parallel and Distributed Systems, 26(2), 455–466.
25. Jung K., Park S., and Park S. (2014), Hiding a Needle in a Haystack: Privacy Preserving Apriori algorithm in MapReduce Framework. In Proceedings of the First International Workshop on Privacy and Security of Big Data, 11–17. ACM.
26. Mohammadian E., Noferesti M., and Jalili R. (2014), FAST: fast anonymization of big data streams. In Proceedings of the 2014 International Conference on Big Data Science and Computing, 23. ACM.