Aisha Abdulrahman Abba

SECURING BIG DATA IN CLOUD WITH INTEGRATED AUDITING

ayeeshahabba@gmail.com

Department of Information Technology
Middlesex University Flic -en-flac, Mauritius

The link to this publication is *https://ajoeer.org.ng/otn/ajoeer/2023/qtr-1/01.pdf*

A SURVEY ON SECURING BIG DATA IN CLOUD WITH INTEGRATED AUDITING

Aisha Abdulrahman Abba; Dr Girish Beekaroo

Department of Information Technology
Middlesex University Flic -en-flac, Mauritius
***Corresponding author:** ayeeshahabba@gmail.com

ABSTRACT

In this paper, the review of the features of big data characteristics of cloud protection problem and research various security regulations and cloud domains were carried out. In order to perform security studies on availability and average time to failure of security, the stochastic process model was applied to analyze security. Based on this research, the use of integrated auditing for safe data storage and transaction logs is highly recommended, and transaction logs, enforcement and security reporting in real time, data climate, compliance with legislation, auditing of facilities, privacy, legality, management of identity and access, cyber threats, and granular auditing to achieve big data security advocated. The purpose of this research is to enforce the use of big data analytics for decision making and policy formulation in corporate and private organizations as well as create room for further research in cloud computing discipline.

Keywords: Big data; security regulation; integrated auditing; cloud computing.

1.0 INTRODUCTION

According to (Manyika et al). Due to recent technical advances, the volume of data generated by social networking sites, sensor networks, the Internet, healthcare apps and many other businesses is growing rapidly daily (Q. Duan, 2020). All the large amount of data produced from various sources in multiple formats at an exceedingly high speed is referred to as big data. Big data has been an immensely popular area of study for the last few years. The rate of data generated is increasing so exponentially that it is becoming increasingly difficult to manage using conventional approaches or structures. However, the broad data system may be organized semi-structured or UN-structured, which adds more complexities when conducting data collection and processing activities. Electronic data from computers is increasingly growing as computer applications become more popular. Big data is created by organizations, especially patient health information, student information, and bank information. According to the Computer Science Corporation report, data accumulation will increase by 4300 per cent per year by 2020. It is a task to handle such a huge volume of data (Q. Duan, 2020).

Basically, Cloud is an internet-based service provider that allows users to share services on demand, such as information, data, framework (V. Lalitha, 2017). Since 2009, there has been a strong demand for cloud computing because of the popularity of powerful data networks and inexpensive computers and storage devices. Service-oriented architecture, hardware virtualization, and autonomous and utility computing have improved availability. A 50 percent rise every year has been seen by some cloud sellers. Cloud computing, however, is only in its early development stages. In cloud computing, we still face many challenges, specifically the achievement of more stable, efficient, and user-friendly computing (Q. Duan, 2020).

The popularity of the Internet and cloud computing and the prevalence of mobile devices uses requires auditors to perform in a global environment. The most common type of cloud data is generated from bank transactions, purchases, payments, inventory, etc. In a white paper, Murphy stated that technology can be used to change auditing and improve (Q. Duan, 2020). Murphy also found out that deeper evaluations should be carried out by big data auditors, the audit process must be ongoing, and auditing practices should be present. Big data auditors need to report further vulnerability incidents, functionality, adjustments, and liability permissions.

In the past, auditors of information technology have focused on fundamental research tools to carry out assessments and draw conclusions. As the amount of data from various application contexts has recently risen, auditing tools need to be revised. Big data employs data sets that are so large that the use of accessible information management systems and conventional data processing methods makes them impossible to handle. We need to switch with big data to tackle structure data (SQL) and not-only SQL (NoSQL), business data warehouses, NewSQL, and data base management system (MPP) mostly parallel processing (R. Vargheese, 2021). Rawal et al. showed an un-conventional way of dealing with big data security problems in the cloud (A. Mehmood, 2020). It is imperative that we set up innovative methods to strengthen the cloud protection of big data. Many big data flaws cannot be removed by a single approach to auditing. We will significantly increase the security of big data in the cloud with automated auditing of our suggested categories.

2.0 Materials and Method

The Four cloud infrastructure frameworks exist in cloud computing. Software: Major businesses that need money to build and confirm various new applications use Platform as a Service (PaaS).

Technology: Small and medium sized enterprises need Software as a Service (SaaS). For apps, companies see the merit of a use-per-pay model.

Infrastructure: A self-service plan is Infrastructure as a Service (IaaS). Clients will pick what they want and do not have to order hardware.

Private/Hybrid: Big companies that choose to combine technologies and cloud offerings favor these solutions. A higher performance than a single algorithm can be obtained by ensemble or boosting strategies in the experiments (C. Zhang, 2021).

The 10 protection and privacy issues have been identified for big data (C. Zhang, 2021).

1. Assure the world of computing in a distributed model.
2. Root of Data
3. Execute good non-relational data applications.
4. Perform validation of end inputs.
5. Encrypting information
6. Keep the transaction and collection of data free of security violations.
7. Ensure periodic monitoring of real-time security.

8. Conduct study of mining and data
9. Authorization of access authentication
10. Granular checks

The lists above can be divided into four large lists.
Ingredients: (S. Bahulikar, 2016).

1. Infrastructure security.

1). In distributed programming systems, stable modules.
2). Stable data stores that are non-relational.

2. Privacy of data.

1). Test the protection of data privacy.
2). Encrypt data at rest and activity.
3). Apply Big Data granular access management.

3. Control of data.

1). Control of Data
2). Maintain routine oversight and monitoring of data transfers and stocking.
3). Origin of data
4). Inspection of granular data

4. Integrity Inside

1). Observe and validate protection in real time.
2). Checking the accuracy of end output

The security of big data should cover four areas (A. Manekar, 2021):

1. Security-required perimeter to secure entry to the device. And the use of LDAP/active, box directory may be helpful.
2. Access and authorization-required to handle data access and control. This includes authorization for files and folder.
3. Data protection-required to monitor access to confidential data that is unauthorized. This includes Rest and Motion encryption, tokenization, and data masking.
4. Audit and report-required for the system to maintain and document operation. In order to handle enforcement and other obligations, including auditing data and audit reports, auditing is required.

Data itself is the product of big data protection problems (Q. Duan, 2020). Capture, curation, collection, scan, exchange, and conversion require big data processes. There are four V characteristics of big data: variety, volume, velocity, and veracity.

Volume: To manage it, big data needs security strategies. This is different from standard settings.

Velocity: Big data speeds needs to be reached by protection technologies. This needs to concentrate more on the throughput of data parsing and collection.

Veracity: Robust audit capabilities are required for several categories of data with different access permissions. Online and social media, large transaction data and human- generated data come from big data (A. Manekar, 2021). With certain types of big data, precisely and durability are less controllable, e.g., Instagram posts and twitter publishes a series of tweets and hash tags, typos, abbreviation, and colloquial expression. Big veracity is a major characteristic of big data [14].

The security challenges of big data also arise from its environment.

1. Multiple instances: often major data environment have multiple instances or variation of the same key building segments, which ensures the security tools solve various diversity and complexities.
2. Multiple Layer: the open-source Hadoop projects, for example, has various stack layers for different purposes, and in table and scheme management, storage is distributed at the bottom, with different programming.
3. Different technologies: large data environment use multiple data storage and retrieval technologies in general. To support an analytical workload and non-relational technology, it is popular to support relational store and query methods.
4. Multiple scattered data stores: Various geographically spread data stores have big data implementations, but even physical nodes need protection.

An audit is an official review of an account of a person or agency. There are two types of audits in information technology (Y. Wang, 2021). An internal evaluation should be conducted by corporate staff, who can include risk and management evaluations and enhance organizational processes. An outside service can carry out an external audit and comply with strict regulations, both government-oriented and industry-oriented. The cloud infrastructure must pass rules and follow legislation for cloud service providers through an external review (CSPS). These are commonly used in conventional and non-traditional

Audit regulation enforcement is required for big data protection. For the following cloud domains, we will analyze certain major big data Protection policies.

Common domains: ISO27001:

In various organizations, the ISO family is the standard that preserves information asset protection. Employee data, financial records, intellectual property, and information entrusted by third parties are protected by the program domains. ISO/IEC 27001 is the family's best-known standard and offers information security management scheme specifications (ISMS). The 27000 family have dozen standard.

An ISMS is a general approach to the protection of confidential business data so that the information structure remains secure. By applying risk control, the approach embraces individuals, procedures, and information security infrastructure. It will ensure the retention of information assets for small, medium, and large corporation in various industries.

Domain in Medicine:

According to this domain, the portability of health Insurance Accountability Act (HIPAA), of which the main materials forbid the protection of patient records. This ensure that the medical records be kept private and available only to the approved officials. Health Information System is the other example.

To meet these ethical requirements, medical realms require a special audit methodology. A cloud protection audit of the medical sector involves reviewing the medical institution and the data-containing cloud.

Domain financial:
The banks have several transfers every day for data storage. The Sarbanes-Oxley (SOX) Act is most relevant legislation for financial firms. SOX is also known as the Act on Accounting Reform and Investor Security in Listed Corporation and the Act on Corporate and Audit Liability,

Disclosure and Transparency.
For all U.S public finance companies, especially for management and public accounting firms, this Act has essential criteria. The Act outlines several specifications for privately owned enterprises.

SOX comprises 11 parts. It describes laws in relation to a range of big corporate activities. The parts of the bill define the duties of the board of trustees of a public company and discuss the criminal consequences for fraud. The bill includes rules outlining how the legislation can be practiced by public companies (Y. Wang, 2021).

Achieving perfect protection is hard to achieve. Protection systems must respond d to violations. The main challenges are to ensure that the knowledge about the consumer is protected and accurate. The first solution is to encrypt data in the cloud to accomplish these objectives. The second solution is to limit the access of various bank clients to the data specified by the assigned access privileges.

Domain of Government:

The cloud computing domain has continued to be used by the government in recent years. Because of the delicate nature of the records, auditing protection here is more extensive. Different government agencies apply the Federal Risk and Authorization Management Program (FedRAMP) to analyze cloud resources to approve, authenticate, and inspect cloud system.

Three main aspects are audited: Method of change management, incident response, and awareness of the operation. New bugs and information leaks of cloud networks are treated by incident management. Visibility of operation CSPs to give organization with cloud system output at defined interval with modified data. The mechanism of change management restricts the capacity of CSP to make policy different. In specific, it is important to give attentive attention to the control mechanisms concerning FedRAMP requirements**.**

Domain of higher Education (J. Soria-Comas, 2021).

The family Educational Right and Privacy (FERPA) offers access for parents to the university records of their child. In a few exceptions, prior to making education documents known until the university student is eighteen years old, schools require student consent. In addition, the university needs to have consent from the student parents or qualifying pupil in order to be exposed to enforce material from the educational background of a student.

Auditing the protection of big data, relying on current policy and regulations, is also relevant. Big data protection is also influenced by data environment and user habits.

3.0 Integrated Audit for Big Data Security

In order to boost cloud protection, we summarize what we will audit in big data for the following content.

1. Logs of data collection and transaction:

Storage processing is a major component of the security of big data. Remote data auditing was proposed by Sookhak et al for the security of big data storage in cloud computing. They suggested an appropriate strategy for remote data auditing that relies on algebraic signature characteristics for storing large data.

To satisfy authentication and dynamic data criteria, (Wang et al, 2020) proposed a new method that enhances the data block index so that all dynamic data can be supported. They have widened their approach to batch auditing assistance. Their methodology increases the quality of auditing.

In database stability, transaction play a significant role (Chen, et ai, 2007). They directly influence the ultimate achievement of the system. The efficiency of transaction logs can be improved by system arrangements.

2. Real time enforcement security monitoring (Talamo et al, 2021)..

For companies, compliance is still difficult. It is easier to approach it with real-time analytics and security at each level while managing big data. The Cloud Protection Alliance (CPA) recommends that companies use multiple instruments to manage big data analytics. For real time details, these include internet protocol, Kerberos, and stable shell. If these are approached, this is easy to hack logging events and add front-end applications. Which involve routers and firewalls. In mobile heterogeneous cloud computing, Gai et all [4] proposed security-aware effective data transfer for Intelligent Transformation System (ITSs). They aimed to achieved safe real-time data exchange and conversion in their approach.

3. Formance with legislation

In this strategy, enforcement criteria need to be audited, periodically including to satisfy the needs for FERPA, SOX, and ISO 2700 family, and others. Auditing should analyze the vendor's responsibility to the third party accessing the service to comply with compliance, as well as address the commitment of the cloud provider service (CCP) to comply with safety compliance and policy changes.

4. Data Environment (Xiaoyan et al, 2021).

Can CSP satisfy the basic criteria for privacy? Operating systems should be evaluated and a review of the infrastructure component which have been modernized should be given.

5. Identification and Management Access

Information regarding authentication, entry, and installation for cloud service workers should be presented here, as well as descriptions of physical security measures in CSP data centers. This provides server and network machine control.

Accessible forms of connectivity should be reviewed: single sign in and verification for the security software for client identities. Different consumers should be tested for correct access rights and controls.

Clear-text files are typically designed for function Based Access Control (RBAC) policy files and access control lists (ACLs) for components. MapReduce and HBase contain these. Such files can be modified by separate privileged accounts (root and others) [8]. (Li et all 2016) recommended the use of Semantic Based Access Control (SBAC) to achieve a stable Big Data financial service (H .Kupwade, 2021).

6. Security and Integrity of facilities.

Asking the following questions in this category is necessary and the following checks (Gupta, s2020)
How to handle the sensitivity of staff or third parties to consumer data?
Control of vulnerability: Virtual Machine Models patch application.

System security: analysis of networks in pursuit of compromised system related to cloud applications.

Analysis on how correspondence between different virtual machines is managed by the CSP (cloud service provider). Discovering of cloud technology change management, including intrusion prevention, firewall, device repair, and control of the virtual world.

(Gaddam, 2020) point out that infrastructure and credibility must be audited. Auditing is necessary for all system/ecosystem improvement specific to Hadoop.

For e.g., modifications to the scheme include.

1. Ensuring distinctions in the states of control nodes. This provides nodes for job monitoring.
2. Changes are made when removing and data inserting.
3. Reducing knowledge exchanged. When the distribution or encryption approach is added to the cloud, which prohibits unwanted nodes from happening.

Database administrators (DBAs) are responsible for nodes, board, column, and cell protection in general business. (DBAS) make machine arrangements and maintain monitoring of security access in a granular fashion.

7. Availability.

The customer wants to check the arrangement on facilities standard. It is also important to review the storage option, the storage area network, and connectivity to cloud client services.

After system modifications and upgrades, does the CSP have the potential to rebound rapidly (i.e., cluster network and replication, etc.)? where is the storage backup located (onsite, other location)? What utilities for the system are available? Where will full loads occur? Do the CSPs have the potential to cope with this?

8. Privacy.

How in cloud systems are automated identities and password stored securely? Where do consumers expect information to be stored and used? Under what cases do third parties have access to classified information? Will access to shared information from third parties reveal personal data? (Gai et all 2020) proposed a data encryption technique for big data that protects privacy.

9. Auditing in granules

For big data protecting, specifically after an attack on the infrastructure, granular auditing is needed. After any attack, the CSA advises that organizations develop a cohesive audit to ensure that they have a complete audit while providing clear access to the records

Integrity and security of audit records are also important. The integrity of audit data and secrecy must be stored separately. With granular access, audit data is secured (Chen et al, 2007). When setting up auditing, it is important to keep audit data separate and have all the appropriate logs. For auditing, the Elasticsearch method can be used.

10. Threats to the cyber.

How is the handling of activities done for patch and vulnerability management? How does the

CSP guarantee that these software activities do not result in violation of the client facility's security? The vulnerability remediation process and compliance testing process used by the CSP should be checked and application feedback developed.

In specific, Life cycle process will be review for product creation, as well as release and update notification of software.

4.0 Results and Discussion Security Analysis.

For the security analysis, stochastic process model was added to the security analysis. The Markov process was extended to process the model of multiple dimensions to cloud applications (Fan, et al 2019)
Web and database applications are used in protection evaluations. We need to define the imbedded Markov chain in the analyses, and then we need to measure each steady-state likelihood before measuring protection attributes in availability and average time for safety failure (MTTSF) (A Manekar, 2021).

The analysis for the imbedded Markov chain is outlined as follows:

I. Perform steady-state analysis with the following equations for irreducible sets (Moghaddam et al, 2018).

II. For transient states, NT = (I-Q)-1,

The identity matrix is where I is located. In the Markov matrix P, Q is a sub-matrix connected to the transient states, and NT is the number of Markov chain visits to fixed states.

III. **If I = j or FT(i,j) = N(i,i)/N(j,j), FT(i,i)= 1-1/N(j,j) (j,j)**

The first passage probability, where FT(i,j) is defined, is that the Markov chain eventually enters state j at least once from the starting state I [1].

IV. The fk probability can be determined by from a transient state I to the irreducible kth set with the sub-matrix bk.

$$f_k = (I-Q)^{-1} b_k.$$

5.0 Conclusion.

In this journal, the review of big data security challenges and big data characteristics was done and visit various regulations and policies for various application domains. To protect data storage and transaction records, the automated auditing for big data protection was suggested, in real-time enforcement and security reporting, data climate, regulatory compliance, infrastructure auditing, identity and access management, cyber-attacks, availability and auditing in granular terms (Y. Wang, 2020). In addition, to perform security assessments for availability and MTTSF, stochastic process model was applied to the method
In the future, as data collection become easy, further studies will be carried out on protecting big data with integrated auditing. To check the security analysis models, the focus on data collections will be more. In specific, security assessment measures for security of big data will be built. The

estimation metric must also be generalized to represent relational data and non-relational data in cloud security assessment.

6.0 Reference

Y. Wang, B. Rawal and Q. Duan, "Securing Big Data in the Cloud with Integrated Auditing", *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pp. 1-5, 2017. Available: https://ieeexplore.ieee.org/document/8118429. [Accessed 15 December 2020].

A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua and S. Guo, "Protection of Big Data Privacy", *IEEE Access*, vol. 4, pp. 1821-1834, 2016. Available: https://ieeexplore.ieee.org/document/7460114. [Accessed 23 December 2020].

V. Lalitha, M. Sagar, S. Sharanappa, S. Hanji and R. Swarup, "Data security in cloud", *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017. Available: https://ieeexplore.ieee.org/document/8390134. [Accessed 5 January 2021].

H. Kupwade Patil and R. Seshadri, "Big Data Security and Privacy Issues in Healthcare", *2014 IEEE International Congress on Big Data*, 2014. Available: https://ieeexplore.ieee.org/document/6906856. [Accessed 8 January 2021].

R. Vargheese, "Dynamic Protection for Critical Health Care Systems Using Cisco CWS: Unleashing the Power of Big Data Analytics", *2014 Fifth International Conference on Computing for Geospatial Research and Application*, 2014. Available: https://ieeexplore.ieee.org/document/6910124. [Accessed 13 January 2021].

A. Manekar and G. Pradeepini, "Cloud Based Big Data Analytics a Review", *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, 2015. Available: https://ieeexplore.ieee.org/document/7546203. [Accessed 14 January 2021].

C. Zhang, X. Shen, X. Pei and Y. Yao, "Applying Big Data Analytics Into Network Security: Challenges, Techniques and Outlooks", *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, 2016. Available: https://ieeexplore.ieee.org/document/7796195. [Accessed 18 January 2021].

Z. Xing, S. Yuan and C. Xiongzhi, "Study on the Impact of Big Data Technology on the Audit and its Application", *2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2020. Available: https://ieeexplore.ieee.org/document/9202383. [Accessed 3 February 2021].

Y. Wang, M. Ding, S. Kan, S. Zhang and C. Lu, "Deep Proposal and Detection Networks for Road Damage Detection and Classification", *2018 IEEE International Conference on Big Data (Big Data)*, 2018. Available: https://ieeexplore.ieee.org/document/8622599. [Accessed 2 February 2021].

S. Bahulikar, "Security measures for the big data, virtualization and the cloud infrastructure", *2016 1st India International Conference on Information Processing (IICIP)*, 2016. Available: https://ieeexplore.ieee.org/document/7975336. [Accessed 3 February 2021].

Q. Zhao, W. Liao, L. Wei and H. Shu, "An Equipment Cloud Service Storage Scheme Based on Domain Division in Cloud Manufacturing Environment", *2020 International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*, 2020. Available: https://ieeexplore.ieee.org/document/9258545. [Accessed 4 February 2021].

J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez and D. Megias, "Individual Differential Privacy: A Utility-Preserving Formulation of Differential Privacy Guarantees", *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 6, pp. 1418-1429, 2017. Available: https://ieeexplore.ieee.org/document/7839941. [Accessed 6 February 2021].

T. Nguyen, "A Framework for Five Big V's of Big Data and Organizational Culture in Firms", *2018 IEEE International Conference on Big Data (Big Data)*, 2018. Available: https://ieeexplore.ieee.org/document/8622377. [Accessed 7 February 2021].

R. Tardio, A. Mate and J. Trujillo, "An iterative methodology for big data management, analysis and visualization", *2015 IEEE International Conference on Big Data (Big Data)*, 2015. Available: https://ieeexplore.ieee.org/document/7363798. [Accessed 7 February 2021].

Talamo, M., Povilionis, A., Arcieri, F. and Schunck, C., 2015. Providing online operational support for distributed, security sensitive electronic business processes. *2015 International Carnahan Conference on Security Technology (ICCST)*, [online] Available at: <https://ieeexplore.ieee.org/document/7389656> [Accessed 8 February 2021].

Gupta, S. and Sharma, K., 2020. A Review on Applying Tier in Multi Cloud Database (MCDB) for Security and Service Availability. *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, [online] Available at: <https://ieeexplore.ieee.org/document/9132931> [Accessed 10 February 2021]

Fan, W., Ziembicka, J., de Lemos, R., Chadwick, D., Di Cerbo, F., Sajjad, A., Wang, X. and Herwono, I., 2019. Enabling Privacy-Preserving Sharing of Cyber Threat Information in the Cloud. *2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/ 2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, [online] Available at: <https://ieeexplore.ieee.org/document/8854026> [Accessed 10 February 2021].

Moghaddam, F., Wieder, P., Yahyapour, R. and Khodadadi, T., 2018. A Reliable Ring Analysis Engine for Establishment of Multi-Level Security Management in Clouds. *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, [online] Available at: <https://ieeexplore.ieee.org/document/8441183> [Accessed 4 February 2021].

Chen, B., Fu, X., Zhang, X., Su, L. and Wu, D., 2007. Design and Implementation of Intranet Security Audit System Based on Load Balancing. *2007 IEEE International Conference on Granular Computing (GRC 2007)*, [online] Available at: <https://ieeexplore.ieee.org/document/4403168> [Accessed 5 February 2021].

Xiaoyan, H., Rong, J., Xiaoming, H. and Luxiao, W., 2020. Thoughts On The Ecological Environment Management Innovation Driven By Big Data. *2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, [online] Available at: <https://ieeexplore.ieee.org/document/9196489> [Accessed 12 February 2021].