BOOK CHAPTER | Data Collection With a Conscience

Collecting Digital Data with the Assurance of Integrity.

Gilbert Sam
Digital Forensics and Cyber Security Graduate Programme
Department of Information Systems & Innovations
Ghana Institute of Management & Public Administration
Greenhill, Accra, Ghana
E-mail: gilbert.sam@st.gimpa.edu.gh
Phone: +233244044354

ABSTRACT

It is very common today that systems collect and store sensitive information. The database administrators of these types of systems have access to this sensitive information and can manipulate it. Therefore, data integrity is of core importance in these systems, and methods to detect fraudulent behavior need to be implemented. The objective of this article is to evaluate the features and performance impact of different methods for achieving and implementing data integrity in a database during data collection to improve assurance. Five methods for achieving data integrity were tested. The methods were tested in a controlled environment. This paper evaluates traditional Digital signature, Linked timestamping applied to a Merkle hash tree, and Auditing performance impact and feature impact wise. Two more methods were implemented and tested in a controlled environment, Merkle hash tree and Digital watermarking. In the evaluation, the researcher proved that traditional Digital signature is faster than Linked timestamping. In this study, it was concluded that when choosing a data integrity method to implement, it is of great importance to know which type of operation is more frequently used. The experiments show that the Digital signature method performed better than Linked timestamping and Auditing.

Keyword: Data Collection, Integrity, Assurance, Security, Cyber space, cybercrimes,

BOOK Chapter | Research Nexus in IT, Law, Cyber Security & Forensics. Open Access. Distributed Free

Citation: Gilbert Sam (2022): Collecting Digital Data with the Assurance of Integrity.
Book Chapter Series on Research Nexus in IT, Law, Cyber Security & Forensics. Pp 139-146
www.isteams.net/ITlawbookchapter2022. dx.doi.org/10.22624/AIMS/CRP-BK3-P23

1. INTRODUCTION

When it comes to assuring data integrity, the situation is more complex because words mean different things to different people (Roy, 2017). To the IT Security group, it is the assurance that information can be accessed and modified only by those authorized to do so. To the Database Administrator, it is about data entered into the database being accurate, valid, and consistent. To the Data Owner, it is a measure of quality, with the existence of appropriate business rules and defined relationships between different business entities, and to the Regulator, data

integrity is the quality of correctness, completeness, wholeness, soundness, and compliance with the intention of the creators of the data (Roy, 2017). This difference in meaning creates a fertile ground for miscommunication and misunderstandings, with the risk that the activity will not be done well enough because of unclear accountabilities. Notwithstanding the impossibility of eliminating all vulnerabilities to data integrity in the organization, controls should be established to reduce the propensity for data integrity errors and vulnerabilities. Such controls should integrate and coordinate the capabilities of people, operations, and technology through a data integrity assurance infrastructure. *It hinges upon a multi-faceted approach consisting of the following triad components which are* management controls, procedural controls, and technical controls

1.1 Background To The Study

Most organizations have a variety of techniques and tools they employ to identify and correct data quality issues when amassing data. Unfortunately, these techniques assume that compromised data exists and waits to clean the data at the point where it is about to be used and this reduces assurance of quality (Batini, et al. 2009). Reporting applications filter out suspect data to report on a "clean subset" of actual records. Similarly, extract, transform and load (ETL) processes "clean" data as they take disparate data from source systems into a common representation in a data warehouse. Source systems attempt to validate data on entry but often must live with compromised data.

There is a large body of work on the use of integrity constraints to ensure logically consistent data within a database (Fan & Jia, 2008). This includes significant work on the use of dependency relationships within relational databases to flag inconsistencies. Implementing dependency relationships with triggers is an effective mechanism for rejecting any transaction that might result in a loss of integrity. Unfortunately, it is not always practical to simply reject transactions, nor is it possible to automatically fix transactions. Human interaction is often required to resolve issues of semantic accuracy that are the root cause of inconsistencies. The source system is usually the best place to resolve such issues but, even there, it is often necessary to first accept the record as it can take time to determine the correct resolution.

As most data collected from various sources are entered by users alike and can be altered deliberately or unknowingly, the study tends to evaluate various cryptography used to promote assurance of data integrity when collecting these data. Cryptography is an important tool to manage integrity even though most people associate it with secrecy. Data integrity and authentication/assurance are only some of the examples where cryptography is not used to achieve secrecy, but for data fingerprinting, indexing in hash tables, in combination with a secret cryptographic key to verify data integrity and message authentication, etc (Shingari & Verma, 2013). With technology development and large availability and accessibility of data, cryptography has become crucial to its security.

2. LITERATURE REVIEW

Data integrity is a sensitive and important issue, yet there is a minimal amount of literature review available in this context. Authors have tried to incorporate all quality research initiatives in the literature and provide a Scientometric analysis of the selected studies (Reyes-Menendez, Saura & Filipe, 2019). In order to conduct a literature review, the author surveyed various studies on the subject matter. Some of these have discussed administrative qualities and needs

and some are about the various privacy and data security approaches. Authors also found that data integrity management to improve assurance is the most crucial and challenging topic for current security. Pashazadeh and Navimipour (2018) provide an exhausting review of big data handling mechanisms. The paper provides a brief and comprehensive knowledge related to big data handling mechanisms in healthcare through various aspects. The study categorizes the mechanisms into various fields for easy and comparative analysis.

Biancone *et al.* (2019) discussed healthcare data quality in their paper. Their study illustrates the various data quality assurance methodologies of different research work done and published from 2014 to 2018. The paper chooses various quality research initiatives and analyzes their respective results on various standards. The paper contributes effective information and the current state of data quality methods in healthcare for future researchers. Hussien *et al.* (2019) provide a brilliant review on the current situation of healthcare for developing a roadmap for blockchain technology. The paper discusses various issues like interoperability, accountability as well as law-related implications of healthcare in order to analyze the blockchain technology. The study also describes the roadmap to enable the healthcare industry for blockchain technology and prepare a taxonomy. The paper contributes some very significant and effective information for the healthcare industry.

Behrouz *et al.* (2017) provide a review on the Internet of Things (IoT). IoT is the most significant part of smart hospitals. The authors discuss the data aggregation mechanisms of IoT for better communication and effective use. The study contributes to various aspects as it provides a comprehensive study of various data aggregation mechanisms. The above-discussed research initiatives provide some significant knowledge for various industries. However, the authors found that there is a need for a literature review that focuses on various data integrity techniques and provides a roadmap for future researchers to illustrate their research initiatives. To achieve this goal, the proposed research endeavor discusses the various data integrity management techniques such as digital signature, Auditing, Merkle hash tree, timestamping, and Digital watermarking. Data integrity attributes are shown in Fig 1.

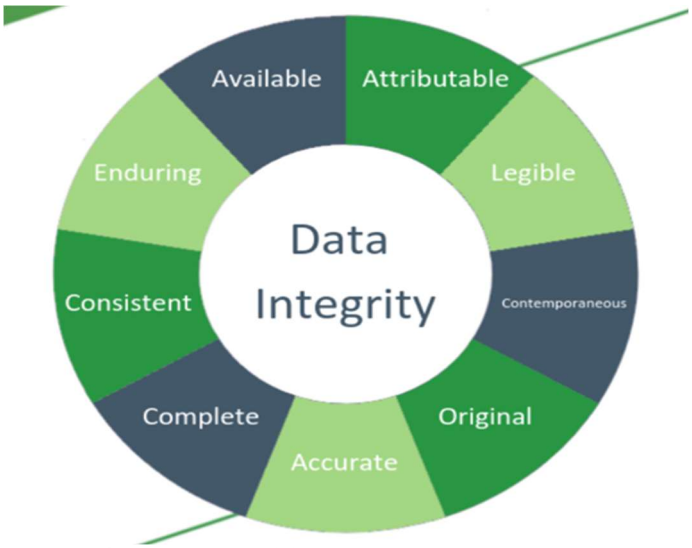

Fig 1: The The ALCOA+ Data integrity Attributes

The ALCOA+ is a framework that ensures data integrity. It has relevance in a range of areas particularly in relation to pharmaceutical research, manufacturing, testing, and the supply chain. https://slcontrols.com/en/what-is-alcoa-and-why-is-it-important-to-validation-and-data-integrity/

3. RESEARCH GAPS

Digital signatures are relatively easy to integrate into the system as all standardized methods are already implemented in most programming languages. As the method only computes digest representation of data it is easy to insert, update and verify entries. However, if the verification fails, you cannot know which data has been manipulated, the one coming from the verifier or the one in storage. And with access to the database, manipulations can easily be made to single entries which will not be detected until verifiers try to check their integrity. Some of the benefits include easy implementation, fast verification, and easy insert and update values. However, a mismatch in verification only means that one of the values has been manipulated and there can be a possible manipulation of data in storage without affecting other entries.

Auditing is easy to use since it is already supported in the Oracle database. All one needs to do is enable it in the Oracle database and decide whether to save the audits on a file or in the database. However, there are some problems with this. The problem is that Oracle will audit so much that it is almost impossible to go through the file by hand and more sophisticated methods must be created to filter through the file (Oracle, 2015). Auditing is already supported in Oracle databases, capable of detecting and tampering and the rules for auditing are simple to define. The Merkle tree was both easy to understand and implement. The biggest issue with this method is how one should handle the insertion of new nodes. For example, if one creates a Merkle tree with four leaves, how should one most effectively insert a fifth element? One solution tried that works, in theory, is to initialize the tree with a very large amount of leaves. However, this solution only postpones the problem to a later time; it could also create a future problem with memory since the empty nodes still need to be saved. One of the newer implementations available is discussed by Becker (2008) where he mentions several different implementations used to improve the original scheme and talks about the major weaknesses of these. But this study implements a scheme toward an Oracle database and measures the performance which the previous study didn't implement

The Linked timestamping method was relatively easy to implement because it is built on the structure of Merkle trees which we have already implemented earlier but has higher integrity due to linked timestamping. The problematic part is that entries cannot be updated so all updates become inserts and old nodes become deprecated but still take up space (Buldas, Kroonmaa, & Laanoja, 2013). Buldas et al. (2013) developed a method based on Merkle Signature Scheme (Becker, 2008). Their method builds small binary hash trees that are time-dependent and connected to each other. Their tops are placed in a publicly available "calendar" which makes the verification of data easier. This method is an improvement over the original Merkle tree method because it dynamically allocates memory and grows to depend on how much data needs to be stored, unlike its predecessor which is static. The company behind the method, states: "...signatures are cryptographically linked to the underlying data such that assertions can be made at a later date regarding the time, integrity and provenance of the underlying data." (Guard, 2015). Hence, the study will address these linkage issues.

The Digital watermark was the hardest to implement and it is questionable if it will work efficiently in a system with this many updates since the watermark generation is quite costly, they also say that the method should work best when used only a couple of times a day and that future work should focus on optimizing the method. According to Camara et al. (2014), the security in the watermark lies in that it is hard to change values in the data set so that a group is generated with the same determinant and diagonal minor values. There are many different methods to generate watermarks on databases but an interesting article is done by Khan and Husain (2013) who state that most watermark schemes introduce distortion in the data. Data distortion makes this method unusable in this study due to the fact that we are trying to protect the currency and distorting that data is not advised. They introduce their own method that stores the watermark without introducing distortions in the data. However, according to Camara et al. (2014), their scheme is vulnerable to certain attacks. This study follows this up by introducing its own scheme that should fill this vulnerability.

4. FINDINGS OF THE STUDY

Digital signature has fast verification because it is a simple comparison between two hashes. Its integrity is somewhere in the middle, compared to the other methods, because of the verification results. Data inserts and updates during data collection are fast and easy to perform as you only need to change stored database hashes. All standardized digital signature algorithms are available online and easily integrated. Size is not affected.

Auditing on the other hand has a low verification speed because the auditing is not well suited for this.It has to be done after the fact and must have some tool to go through the logs. The integrity is average since depending on the filters used on the logs one can detect tampering. For example, tampering done by someone with access to the database will audit the name of the user, what kind of operation the user performed, when it happened, and where the user did it. The data inserts and updates are easily implemented since it is done in the database. However, auditing everything will create large files that need to be managed somehow.

With the **Merkle hash tree**, the verification speed is somewhere in the middle compared to the other methods as many hash computations need to be performed to be able to verify data. Data integrity is also somewhere in between the other methods because it is hard to pinpoint to which entry has been manipulated during data collection and entry. Inserts and updates are of medium/high complexity based on the fact that authentication paths need to be recomputed and tokens need to be redistributed for all leafs. Implementation is of medium complexity because it is a standard binary tree implementation. Size increase over time is high as when a tree is full a completely new tree of twice the size needs to be built.

The verification speed of Linked **timestamping** is almost the same as Merkle tree hash as it is built on the same method. Integrity is a bit higher than Merkle tree hash because of the timestamping that is applied to each calendar node. Inserts are very easy/fast as you are always appending new trees to the hash chain. Data updates are impossible due to the fact that every tree was created with a timestamp and that time is unknown after computation and can't be reused. It is not very complex to implement as it uses Merkle trees and hash chaining to create its structure. The size increases over time as updates are impossible and they become new inserts so over time the calendar can become relatively large.

The verification speed in **Digital watermarking** is quite slow due to its need to compute several matrix calculations every time it needs to verify. Since the analysis manage to test the speed, this value is based on the view of the author. It does however give high integrity since it is very hard to tamper with the data undetected and it is possible to detect what and where the tampering has occurred. Both inserts and updates of data are expensive in watermarking due to the need to compute several matrix calculations each time.

5. CONCLUSION

In this article, different methods for achieving data integrity in a database have been suggested. The purpose is to solve the data integrity problem that arises from the presented scenario where a large data collection system is the victim of internal data alteration, deletion, or manipulation. The methods tested were traditional Digital signature, Merkle hash tree, Oracle auditing, linked timestamping, and Digital watermarking. They were implemented and tested to see if they truly do detect unfiltered data and three of the methods were also integrated and performance evaluated by a performance evaluation tool. An empirical analysis was also performed to see the benefits and drawbacks of each method. Based on this work a conclusion was reached that a system such as the Merkle trees works well to detect integrity fraud and looking at previous work and on this current study implementation, the previous studies do provide adequate verification and signature times. Of the three methods tested Linked timestamping works well since it creates quite small trees consistently compared to the standard Merkle hash tree that would eventually need to make a large increase in size to fit more leaves.

6. RECOMMENDATION FOR POLICY AND PRACTICE

When an organization is collecting large amounts of data, it can be difficult to keep track of changes that affect data integrity. New data coming in may not necessarily satisfy existing constraints. Therefore, the idea behind our approach of examining various hashing methods is to improve the assurance of clean data in the database in the process of collecting them, so that compromised data is flagged and segmented from clean data for follow up and a separate view of all data is also maintained for processing based on priority. By using our approach, we envision that a database will never be allowed to enter an inconsistent state. In other words, once an area of clean data within the database is established, it stays clean. Hence, firms can refer to the study to know how best to combine methods to achieve a cleaner data collection. A gap analysis of existing approaches to managing data integrity in and in the academic literature clearly identifies where these approaches fall short. This gap is articulated in a set of criteria that can be used to evaluate approaches for managing data integrity.

7. DIRECTION FOR FURTHER WORK

Due to limited time, the author was not able to integrate the Merkle hash tree and Digital watermarking with the performance evaluation tool and optimize the methods for best performance. Therefore, suggested future work is integration and performance analysis of the remaining data integrity methods. As we stated earlier Auditing cannot automatically detect fraudulent behavior it can only log database activity. Possible future work could be to implement an AI, script, or rule-based system that can go through the logs and flag the possible breach of integrity in data collection.

8. IMPLICATIONS FOR CYBER SAFETY IN AFRICA

Poor data quality could have a destructive impact on the social and well-being of the African continent. Periodic systematic data assessment will offer policymakers a much stronger set of findings to use in making policy decisions (Diener & Seligman,2004). African stakeholders need to improve digital data quality with adequate strategies, approaches, techniques, and tools to ensure the reliability and integrity data collection. It is vital for developing nations to have reliable and quality data for economic development and accurate decision-making. Technological advancement has created a higher demand for IT experts to improve strategies to ensure reliability in data collection (Wang & Wang,2005). These requests and demands have placed a transformed energy on quality improvements for long term survival for governments and organizations. The use of wrong data in most developing countries in Africa has resulted in the problem of inaccurate distribution of resources for national development. Decisions made with the use of unreliable data could result in serious setbacks in the economic growth of a nation (Chukwu et all., 2024)

REFERENCES

1. Batini, C., Cappiello, C., Francalanci, C., & Maurino, M. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, *41*(3), 121–130.
2. Becker, G. (2008). *"Merkle signature schemes, merkle trees and their cryptanalysis"* [Ruhr-Universisät Bochum]. http://goo.gl/KNeVhj
3. Behrouz, A. K., Tripathi, A. K., Kapil, G., Singh, V., Khan, M. W., Agrawal, A., Kumar, R., & Khan, R. A. (2017). *Current Challenges of Digital Forensics in Cyber Security* (pp. 31–46). http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-7998-1558-7.ch003
4. Buldas, A., Kroonmaa, A., & Laanoja, R. (2013). "Keyless signatures' infrastructure: How to build global distributed hash-trees". *18th Nordic Conference, NordSec 2013*, 313–320.
5. Camara, L., Li, J., Li, R., & Xie, W. (2014). "Distortion-free watermarking approach for relational database integrity checking." *Mathematical Problems in Engineering*, *14*, 235–250.
6. Fan, W., & Jia, X. (2008). "A Revival of Integrity Constraints for Data Cleaning." *Very Large Data Base*, *1*(2), 1522–1523.
7. Guard. (2015). *"Keyless signature infrastracture."* KSI Technology. http://guardtime.com/ksi-technology
8. Hussien, H. M., Yasin, S. M., Udzir, S. N. I., Zaidan, A. A., & Zaidan, B. B. (2019). "A systematic review for enabling of develop a blockchain technology in healthcare application: Taxonomy, substantially analysis, motivations, challenges, recommendations and future direction." *Journal of Medical Systems*, *43*(10).
9. Pashazadeh, A., & Navimipour, N. J. (2018). Big data handling mechanisms in the healthcare applications: A comprehensive and systematic literature review. *Journal of Biomedical Information*, *82*, 47–62.
10. Reyes-Menendez, A., Saura, J. R., & Filipe, F. (2019). The importance of behavioral data to identify online fake reviews for tourism businesses: a systematic review. *PeerJ Computer Science*, *5*, e219. https://doi.org/10.7717/peerj-cs.219
11. Roy, C. (2017). *The Data Integrity Triad*. LinkedIn. https://www.linkedin.com/in/chinmoy-roy-521736

12. Shingari, I., & Verma, S. S. (2013). *Achieving data integrity by forming the digital signature using rsa and sha-1 algorithm.* Mody Institute Of Technology and Science.
13. Diener & Seligman, (2004). Reliable Data Collection: A tool for Data Integrity in Nigeria
14. Chukwu et all., (2024). Reliable Data Collection: A tool for Data integrity in Nigeria.
15. Nir Kshetri Bryan School of Business and Economics, The University of North Carolina, GreensboroCorrespondencenbkshetr@uncg.edu View further author information Published online: 09 Apr 2019https://doi.org/10.1080/1097198X.2019.1603527 CrossMark Logo CrossMark.
16. https://www.tandfonline.com/doi/full/10.1080/1097198X.2019.1603527