Data labeling for data security in data lifecycle: A state of the art and issues

Kenza Chaoui*, Nadia Kabachi**
Nouria Harbi**, Hassan Badir *

*IDS Team ENSAT, Abdelmalek Essaadi University Tangier
Kchaoui1994@gmail.com
hbadir@uae.ac.ma
**ERIC Laboratory,Lumière Lyon 1 University Lyon, France
nadia.kabachi@univ-lyon1.fr
nouria.harbi@univ-lyon2.fr

Résumé. One of the most serious issues with cloud computing is data security, As businesses start on digital transformation, there is a clear requirement for privacy and data protection. Organizations today have more data, applications, and websites than they have ever had before. Data security has risen to the top of the priority list for cloud computing security. Despite the fact that a variety of solutions have been proposed, the majority of them only address one stage of the data life cycle, such as storage, which is insufficient to handle the cloud data security challenge because threats appear at all stages of the data life cycle. During the data life cycle process, any stage's security breaches could affect data security. Therefore, data security must be considered throughout the data lifecycle. The main contribution of this article is a new perspective on data security solutions based on the data lifecycle, which is crucial and can be used as a guide to create a complete security solution.A literature review on the entire data life cycle is carried out and a research gap of unresolved issues that may be research questions for our future work is presented.Also a proposed solution on data labeling used for data tracking to secure data in all stages in data life cycle is presented.

1 Introduction

Cloud computing is a virtualized system that allows users to access compute, storage, and software resources as well as servers from a single platform. Data management services are currently provided in the user's local environment, however CSP Cloud Service Providers provide them remotely. Users may not know where, when, how, why, or by whom their data is seen or modified in a cloud environment because services are supplied in abstract form. Cloud computing, on the other hand, has several security risks. CSPs are also more vulnerable to adversaries and hackers who can take advantage of these benefits. The cloud is vulnerable from a security and data privacy perspective, as sensitive user data is stored in a third-party CSP. In Michelin et al. (2018) all these weaknesses and mistrust build a common element which is the issue of trust and safety between providers and consumers, because Cloud Computing requires the trust of providers of cloud services. As a result, trust is one of the main factors for the adoption of this new paradigm. On the other hand, data can be stolen and used maliciously by the provider itself. The ultimate cloud challenge is data level security, and sensitive data will need to be protected at the enterprise level, not at the cloud provider level. Security will need to move to the data level so that organizations can ensure that their data is protected wherever it is located. Thus, it is advantageous for each customer to take their own security measures, regardless of what the providers are offering. When we talk about data security we must ensure the triplet of confidentiality, integrity and availability.

Losses of confidentiality, integrity and availability (CIA) can have a big impact on cloud computing operations because data is the backbone of any business. One of these aspects may prevail over the others depending on the environment, application, context or use case. Many data security solutions have been proposed, however, most of them do not cover all of these security aspects. On the other hand, there is the challenge of data traceability which is used to trace the state and movements of the data. Without it, there is no chance of being sure that the other three criteria (CID) are met. Our work will be structured as follow : We will start by presenting a state of the art on data protection and security methods in the data life cycle. Next, we will describe in detail a discussion or the research gap of the previous work. Finally,we will present the conclusion and the future work.

2 Data life cycle

There are several different data life cycles that have been suggested in the literature to help businesses to choose the most suitable and appropriate solutions for their context.

CRUD life cycle Demchenko et al. (2014), Life cycle for Big DataEl Arass et Souissi (2018), IBM life cycle El Arass et Souissi (2018) ,DataOne life cycle El Arass et al. (2017) ,Information life cycle Lin et al. (2014), CIGREF Vidgen et al. (2017) ,DDI life cycle Ma et al. (2014), USGS life cycle El Arass et Souissi (2018) ,PII life cycle in Michota et Katsikas (2015), Enterprise data life cycle Chaki (2015)and Hindawi life cycle Khan et al. (2014). All the mentioned life cycles are described and discussed in El Arass et al. (2017). The most important phases that are typical are : create, store, use, share, archive, and destroy.El Arass et Souissi (2018) Data should be protected in all the phases of the life cycle, from initial creation through the destruction. The store and archive phases are defined as data-at-rest, the use phase is referred to data-in-use, the sharing phase is called data-in-transit and the destruction phase can be referred to data-after delete.

The six stages are summarised in 1.

— **Creation** is the generation of a new item or the modification of a digital data item exists. It can therefore also be called as a creation / update phase. It can be any type of content, not just a document or a database i.e. it can be structured or unstructured. In this phase, the data is classified and the appropriate rights are determined. Data may be generated in client or the server in cloud.

— **Storage** is the action to form digital data according to a structured or unstructured storage repository type (database or file). This usually happens at the same time as creation. Here the classification and rights of the security controls must be mapped, including access controls, encryption, and rights management.

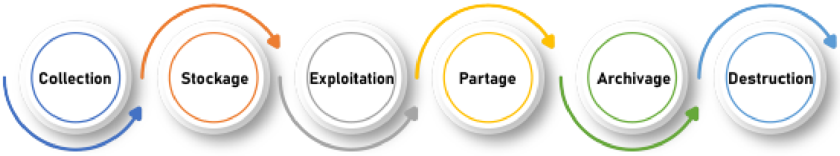


FIG. 1 – *A Data life cycle .*

- **Use** The data is viewed, processed or used in a way where that original data is not altered. These activities generally apply to data stored at the time of use from a user's PC or application. To ensure this type of activity, there are detection controls such as activity monitoring, preventive controls such as rights management, and logical controls that are typically applied in databases and applications.
- **Sharing** Data is made accessible to others, and it is exchanged between users, customers and partners. Controls in this phase include a combination of detection and prevention, encryption for secure data exchange, logical controls as well as application security.
- **Archiving** Data remains idle and goes into long-term memory to be archived, here data protection and availability is ensured by a combination of encryption management and profit management.
- **Destruction** The data is permanently destroyed using physical or logical means. The data must be deleted in a secure manner and tools must be used to find the permanent copies.

3 Literature Review

The data must be secure throughout the life cycle of the data here are the existing works that are proposed in some phases oh the data life cycle :

3.1 Secure Data Creation

InFu et al. (2018),The authors proposed a proxy re-encryption scheme with the support of a cloud, the authors proposed a protected cloud-assisted IoT data management system to preserve data confidentiality while collecting, storing, and accessing IoT data, taking into account the increase of users. Therefore, a secure IoT could handle most attacks from both IoT insiders and outsiders to break data confidentiality, in the meantime with communication's constant costs for IoT re-encryption anti-incremental scale.

A safe, flexible and efficient data storage and retrieval system based on both fog computing and cloud computing techniques is designed in Liu et al. (2020). In terms of data refinement, data organization, searchable encryption and dynamic data collection, the main challenges are summarized, and appropriate solutions are also offered. A tree of retrieval functions is designed to support effective and efficient privacy preserving data search, precise data retrieval and an index encryption scheme based on the secure kNN algorithm are suggested. From a broader view, a flowchart including data mining and remote control is also presented.

According toLiu et al. (2020), the authors proposed a blockchain framework in MANETs for security-related data collection. The collector can restrict its payment by controlling the scale of Route REQuest (RREQ) forwarding in route discovery and at the same time make every forwarder of control information (namely RREQs and Route REPlies, short RREPs) receive rewards as much as possible to ensure fairness. In parallel, the system avoids collision threats by implementing a secure digital signature with cooperative receipt reporting and spoofing attacks. The system not only offers rewards for all participating nodes but also prevents forking and ensures high efficiency and true decentralization, based on a novel Proof-of-Stake consensus mechanism by accumulating stakes by message forwarding.

A new data collection scheme called Secure Data was proposed in Tao et al. (2018) to provide data protection and to protect the rights of the personal data of patients. The authors presented the KATAN secret cipher algorithm to secure communication and implement it on the FPGA hardware platform.Authors apply secret cipher sharing and share repairing for the privacy of the KATAN cipher. The evaluation shows that the Secure Data scheme when applied to attacks,can be effective in terms of frequency, energy cost, and overall computational cost.

In Zhang et al. (2018), A secure data collection system based on compressive sensing (SeDC) was proposed , to improve data privacy through the asymmetric semi homomorphic encryption system and reduces the cost of computation through a sparse compressive matrix. The asymmetric mechanism reduces the complexity of distribution and control the secret key. Homomorphic encryption allows cipher domain in-network aggregation while improving security and achieving network load balancing. The sparse measuring matrix reduces both the computation cost and communication cost, which compensates for the increasing cost caused by homomorphic encryption.

3.2 Secure Data Sharing

IN Michelin et al. (2018) ,The authors proposed an advanced secure and privacy-preserving data sharing system for smart cities based on blockchain. The proposed system en- sures that personal user data is protected, safely stored, and accessible to stake- holders on the need to know the basis of smart contracts embedded in user- defined ACL laws.Besides, they developed a system called ”PrivyCoin” in the form of a digital token for users to share their data with stakeholders/third parties,Also a ”PrivySharing” was presented to satisfies some criteria of the EU GDPR, such as data asset sharing, usability, and data owner consent purge. The experimental results in the paper confirmed that a solution for multi-Ch blockchain scales better than a single-Ch blockchain system .

In a similar endeavorEltayieb et al. (2020), A blockchain-based data usage auditing architecture that provides the data controllers with unforgeable evidence of users’ consent was presented . The researchers claim to provide user anonymity by allowing data owners (delegated to PKG) to create a distinctive public-private key pair for each smart contract they enter into with a service provider or data processor to share data.In addition, hierarchical ID-based encryption was used to avoid unauthorized disclosure. The data stored on off-blockchain storage, while blockchain smart contracts are used to store the hash of data and data usage policy.There is also a particular smart contract between the data holder and Any other provider of services or processor of data.

To secure data sharing in the cloud environment, the authors proposed in Lee (2020) a new blockchain-based attribute-based signcryption scheme (BABSC). The suggested BABSC has the benefit of using the blockchain and attribute-based signcryption To guarantees both confidentiality and unforgeability of data. The evaluation part shows that the BABSC not only minimizes the communication overhead but also gives quick designcryption on the user side. BABSC also en- forces user access control and it is appropriate for cloud computing.

In Wang et Song (2018),a blockchain-based data-sharing platform ”SpeedyChain” for a smart city ecosystem was proposed .The ”SpeedyChain” architecture focuses on reducing the time of TX settlement for real-time applications such as smart cars and also aims to guarantee user privacy. Also, it guarantees data integrity, tolerance to tampering, and non-repudiation.

In Zhou et al. (2016),the authors proposed a revocable-capacity personality based encryption called RS-IBE, which support identity revocation and cipher text update simultaneously to create a cost-effective and stable data sharing system in cloud computing, such that Access to previously shared data, as well as subsequently shared data, is blocked by a revoked user.

In Shao et al. (2011),The authors proposed a protected data group sharing and dissemination framework in the public cloud, based on attribute-based and timed-release conditional identity-based broadcast PRE. The framework allows users to share data with a group of recipients using identification such as email and username at one time, ensuring protection and convenience for data sharing in the public cloud. Also, with the use of fine-grained and timed-release CPRE, the framework enables data owners to configure ciphertext access policies and time trapdoors that could limit the conditions of distribution while outsourcing their data. The CSP can only successfully re-encrypt the ciphertext when the data disseminator attributes associated with the re-encryption key satisfy the access policy in the initial ciphertext and are exposed to the time trapdoors in the initial ciphertext. The results of the experiment based on a cryptography library focused on pairing have shown the system’s protection and effectiveness.

The authors in Rivest et al. (1996) propose a secure and effective cloud data sharing PRE scheme based on ciphertext-policy attributes, which helps the proxy to convert a ciphertext Under an access policy that meets the requestor’s attributes to another ciphertext under a new access policy. However, these systems do not accept a situation in which data access rights are expected to be released to various groups of users at different time points, respectively the authors in Huang et al. (2018), recommended a realistic TRE method for this problem, using a trusted time agent rather than a data owner to uniformly unlock the access privilege at a particular time.

In Huang et al. (2018),the authors proposed a secure data sharing and profile matching scheme for mobile healthcare social networks MHSN in cloud computing. Patients can outsource their secure health records with identity-based broadcast encryption for cloud storage (IBBE) and share them safely and successfully with a community of doctors. Besides, a conditional data re-encryption construction based on attributes are proposed ,to allow doctors who meet the pre-defined requirements in the ciphertext to authorize the cloud platform, to transform a ciphertext into a new ciphertext for specialists without leaking any sensitive information from an identity-based encryption scheme. Also, a profile matching mechanism is presented in MHSN based on identity-based equality testing en- cryption, which enables patients to find friends in a Privacy-preserving process and obtaining flexible authorization to resist the guessing attack keywords on the encrypted health records.

Similarly, Wang et Song (2018) ,Xu et al. (2015) and Huang et al. (2018) have implemented IBBE to secure sharing con- fidential data with a community of users. The comparative schemes will share ciphertext with other users using the PRE technique by re-encrypting the ciphertext through Proxy.

However,Wang et Song (2018) ,Michelin et al. (2018) and Liang et al. (2015) Since all the health data of this patient can be re-encrypted by the doctor authorized by a patient,Xu et al. (2015) , Yang et al. (2016) and Huang et al. (2018) support re-encryption of conditional data. In particular,Yang et al. (2016) and Huang et al. (2018) follow the ABE approach that supports complex operations to describe the MHSN flexible condition. Further, Zhou et al. (2016) , Qiu et al. (2015), andMa (2016) and Huang et al. (2018) All help pro- file matching on ciphertexts. Although Qiu et al. (2015) have obtained flexible authorization, two negotiated users generate the authorization token, which may not be valid in the MHSN. By identifying various trapdoors Huang et al. (2018) andMa (2016), the user may pick the data that will be matched according to their wishes.

3.3 Secure Data deletion

Authors proposed in Xue et al. (2019) an effective attribute revocation scheme based on Merkle Hash Tree for assured data deletion. When the cloud server receives the deletion request from a user,the associated files will be re-encrypted using the re-encryption key created by the trust authority. In parallel, a new root of the Merkle Hash Tree will be sent to the data owner according to attribute re- vocation, so that he can check that the data has been successfully deleted. In addition,for Data Deletion Validation,cloud data can also be accessed by other users.

A fine grained data deletion system was built by Yang et al. (2019) to prevent fraudulent tampering with data from cloud servers and hackers as well as incomplete data deletion of cloud service providers. Also,the rank-based Merkle Hash Tree chain is added to verify if the data block is altered or removed on behalf of the user.

InHao et al. (2015),authors proposed a data deletion system in cloud computing. The proposed framework is based on a ”trust-but-verify” mechanism that enables users to verify the accuracy of encryption and deletion operations. According to the authors,it is difficult to guarantee the complete deletion of data using software, so they prefer to return a digital signature that is bound by a commitment to delete the corresponding secret key. If the deleted key re-emerges later. The signature can be used as evidence to call to account for the cloud service provider’s liability.

In Yang et al. (2020) ,They proposed a fine-grained outsourced data deletion scheme based on invertible Bloom filter, which can achieve both public and private verifiability of the storage and deletion results.Users can easily recognize the malicious activities of the cloud server with an overwhelming probability if the cloud server does not honestly maintain/delete the data and produce corresponding evidence. Meanwhile, the computational complexity is independent of the number of out- sourced data blocks in data deletion and deletion outcome verification procedures, which makes the proposed scheme ideal for the large-scale data deletion scenario.

InYang et al. (2018),a new blockchain-based data deletion framework was proposed in this paper, which can make the process of deletion more transparent. in the solution,no matter how malevolently the cloud server behaves, the data owner can check the deletion result. In addition, the proposed scheme will achieve public verification through the use of blockchain without any trusted third party.

The authors Yu et al. (2018) suggested an assured system of data deletion that meets verifiable deletion of data as well as flexible control of access over sensitive data. When deleting cloud data and validating the deletion of such data, only data owners and fog devices are involved, which makes the protocol practical due to the features of low latency as well as real-time interaction with fog.

3.4 Data security in the creation phase

Here is a table that summarizes all the security issues found in the collection phase. What we can summarize is that, in the stage of collection, the data must be protected. If data is collected indiscriminately, then the source of the data is unclear and noisy data is collected. As a result, these datasets affect the data exploitation and analysis phases and can have serious consequences in terms of reliability and the results obtained. Also, as a lot of unstructured data is collected, there is a need to classify it properly, as mentioned in the articles Rahul et Banyal (2020)Binjubeir et al. (2019). Therefore, at the collection stage, one of the main challenges is to properly filter and classify the data so as not to compromise its reliability .

Also, when the data provider provides the data, control of the data moves away from the provider. Therefore, it is necessary to guarantee data protection Tabrizchi et Kuchaki Rafsanjani (2020) by ensuring that users know whether they are correctly used. What attracted us to this literature is the problems related to CID (confidentiality, availability, and integrity Rahul et Banyal (2020)Dissanayake (2021) Kumar et al. (2018)Yadav et Behera (2020) and a new problem of traceability mentioned in the article Binjubeir et al. (2019) which has a strong relationship with our research. The latter plays a major role in data protection. Data traceability is the process of tracking data access, values, and changes as it moves through its life cycle.without it we cannot ensure that the security criteria are respected or the data is not altered during its life cycle. Indeed, this literature pushed us to do research in this direction and especially to see how we can trace to secure the data and process the traceability of the data in our future solution.

Articles	Authentication	Access control	Integrity	Confidentiality	Availability	Traceability	Unstructured
Tabrizchi et Kuchaki Rafsanjani (2020)			✓	✓	✓		
Rahul et Banyal (2020)	✓				✓		✓
Dissanayake (2021)	✓	✓		✓			
Kumar et al. (2018)	✓	✓	✓	✓	✓		
Ali et al. (2022)	✓		✓	✓	✓		
Yadav et Behera (2020)			✓	✓		✓	
Aswini et Mervin	✓		✓	✓	✓		
Mandal (2021)			✓	✓	✓	✓	
Binjubeir et al. (2019)							✓

TAB. 1 – *Data security in the creation phase.*

4 Research Gap in Data Life Cycle

When we talk about data security we must ensure the triplet of confidentiality, integrity and availability. Losses of confidentiality, integrity, and availability (CIA) can have a big impact on cloud computing operations because data is the backbone of any business. One of these aspects may prevail over the others depending on the environment, application, context or use case. Many data security solutions have been proposed, however, most of them do not cover all of these security aspects. On the other hand, there is the challenge of data traceability which is used to trace the state and movements of the data. Without it, there is no chance of being sure that the other three criteria (CID) are met.

In addition, thanks to traceability, we can control and monitor what is happening and where the data is located, who has access to it, in what state it is, what process has it undergone, has it arrived ? to destination ? On the other hand. The current proposed data security solutions that address all three aspects of security, have the common feature that they focus on a single stage of the data life cycle, which is the entire process from data creation to data destruction. In this process, a security vulnerability at any stage could break the data security state. In a cloud computing environment, data can move from one place to another. In addition to cloud storage, data can frequently be transferred to the customer through an insecure network. Everywhere data is stored, there is a risk for security issues. Data security must therefore be taken into account in all phases of the data life cycle. In addition, in our study we claim that is one of the major challenges in the data life cycle is data after-delete which called as data remanence.

After a storage media is deleted, there may be some physical characteristics that allow the data to be reconstructed. Tracing the data path (data lineage) is important for auditing in cloud computing, especially in the public cloud apart from the above stages. There are a challenges in that investigates securing data during the data restore operation and after restoring. Additionally,during investigating these challenges,the confidentiality, integrity and availability (CIA) are considered, The challenges addressed can be summarized as follows : most of the exsisting SWs that are used to restore data deleted, retrieves a part of the deleted data and the cloud service consumer (CSC) can construct the reminder, and there- fore all data are retrieved. Besides, the CSC can perform illegible operations (edit,delete) on the data retrieved which can lead to crisis for other consumers, and these operations represent security breach issues. Also, it is critical to note, that there is no clear way that ensure the integrity of the data deleted. Second challenge is preserving the privacy for the restored data, for example ; when the data deleted, the access roles for authenticated users are also deleted, so when the data is restored, these roles must be maintained to maintain the privacy of the data.Finally, in the data-in-use stage, there is unauthorized users who can guess the data creation standards and generate data that can be used in real operations.

5 Data labeling for data security : proposed approach

The traceability of products, otherwise known as tracing, makes it possible to qualitatively ensure the journey of the products. Tracing makes it possible, in the event of a quality problem, to find the causes and the origin of the problem. Product traceability concerns all sectors, whether food, chemicals, medicines, children's toys, etc. This is a guarantee of quality for the consumer and is also part of the company's quality approach.

Also, product traceability consists of :

— trace a product and control its quality throughout its journey ;

— identify the causes of a quality problem.

These traceability techniques have inspired us in our field of security and we have considered that a data is like a product and traceability makes it possible to capture, store and manage all the information of the data. We have proposed labeling to trace the data, labeling makes it possible

— Identify the level of data sensitivity

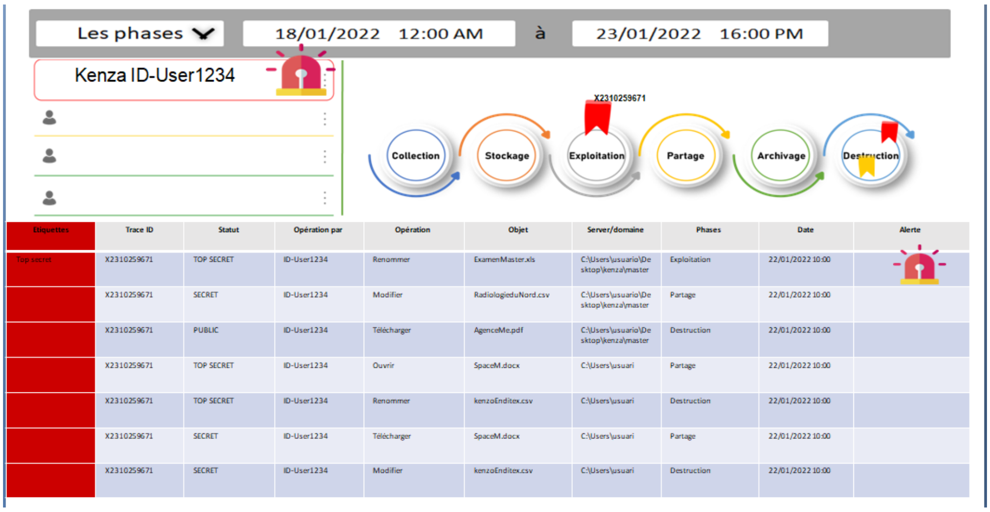
— Secure Data Tracking

— Control, classify and track data by setting specifications and access

— Track every file access and permission change

— See exactly which users have access to sensitive files

The labeling proposed in the figure2 is intended to plot and track data by setting specifications. For each label, we will have detailed information on its traceability in the data life cycle.



Etiquettes	Trace ID	Statut	Opération par	Opération	Objet	Server/domaine	Phases	Date	Alerte
Top secret	X2310259671	TOP SECRET	ID-User1234	Renommer	ExamenMaster.xls	C:\Users\usuario\Desktop\kenza\master	Exploitation	22/01/2022 10:00	🚨
	X2310259671	SECRET	ID-User1234	Modifier	RadiologieduNord.csv	C:\Users\usuario\Desktop\kenza\master	Partage	22/01/2022 10:00	
	X2310259671	PUBLIC	ID-User1234	Télécharger	AgenceMe.pdf	C:\Users\usuario\Desktop\kenza\master	Destruction	22/01/2022 10:00	
	X2310259671	TOP SECRET	ID-User1234	Ouvrir	SpaceM.docx	C:\Users\usuari	Partage	22/01/2022 10:00	
	X2310259671	TOP SECRET	ID-User1234	Renommer	kenzoEnditex.csv	C:\Users\usuari	Destruction	22/01/2022 10:00	
	X2310259671	SECRET	ID-User1234	Télécharger	SpaceM.docx	C:\Users\usuari	Partage	22/01/2022 10:00	
	X2310259671	SECRET	ID-User1234	Modifier	kenzoEnditex.csv	C:\Users\usuari	Destruction	22/01/2022 10:00	

FIG. 2 – *Data labeling for data tracking in data life cycle.*

For example, the operations carried out, the users who have the right of access, the validity date of each file, as well as the phase in which the data is located. These labels flowing through the life cycle of the data give us the ability to make sure the data is well secured and if there's a label change or, for example, a label that's been tampered with by a hacker, it's going to be very clear and easy to manage. We can take action on them immediately because the data is tagged. We know where the data resides, the changes are found and this will help the administrator to control the data well.

6 Conclusion

Data security has risen to the top of the cloud computing security priority list. Though many solutions have been proposed, many of them only consider one aspect of security, We proposed that data security in the cloud must considered throughout the data life cycle. This paper's main contribution is a new angle on data security solutions based on the data life cycle, which is crucial and may be used as a guide for creating a complete security solution. The future work is to work on the data labeling for data tracking and implement this part.

Références

Ali, I., I. Ahmedy, A. Gani, M. U. Munir, et M. H. Anisi (2022). Data collection in studies on internet of things (iot), wireless sensor networks (wsns), and sensor cloud (sc) : Similarities and differences. *IEEE Access 10*, 33909–33931.

Aswini, G. et R. Mervin. A survey on cloud security issues and threats.

Binjubeir, M., A. A. Ahmed, M. A. B. Ismail, A. S. Sadiq, et M. K. Khan (2019). Comprehensive survey on big data privacy protection. *IEEE Access 8*, 20067–20079.

Chaki, S. (2015). The lifecycle of enterprise information management. In *Enterprise Information Management in Practice*, pp. 7–14. Springer.

Demchenko, Y., C. De Laat, et P. Membrey (2014). Defining architecture components of the big data ecosystem. In *2014 International conference on collaboration technologies and systems (CTS)*, pp. 104–112. IEEE.

Dissanayake, A. (2021). Big data security challenges and prevention mechanisms in business.

El Arass, M. et N. Souissi (2018). Data lifecycle : from big data to smartdata. In *2018 IEEE 5th international congress on information science and technology (CiSt)*, pp. 80–87. IEEE.

El Arass, M., I. Tikito, et N. Souissi (2017). Data lifecycles analysis : towards intelligent cycle. In *2017 Intelligent Systems and Computer Vision (ISCV)*, pp. 1–8. IEEE.

Eltayieb, N., R. Elhabob, A. Hassan, et F. Li (2020). A blockchain-based attribute-based signcryption scheme to secure data sharing in the cloud. *Journal of Systems Architecture 102*, 101653.

Fu, J.-S., Y. Liu, H.-C. Chao, B. K. Bhargava, et Z.-J. Zhang (2018). Secure data storage and searching for industrial iot by integrating fog computing and cloud computing. *IEEE Transactions on Industrial Informatics 14*(10), 4519–4528.

Hao, F., D. Clarke, et A. F. Zorzo (2015). Deleting secret data with public verifiability. *IEEE Transactions on Dependable and Secure Computing 13*(6), 617–629.

Huang, Q., W. Yue, Y. He, et Y. Yang (2018). Secure identity-based data sharing and profile matching for mobile healthcare social networks in cloud computing. *IEEE Access 6*, 36584–36594.

Khan, N., I. Yaqoob, I. A. T. Hashem, Z. Inayat, W. K. Mahmoud Ali, M. Alam, M. Shiraz, et A. Gani (2014). Big data : survey, technologies, opportunities, and challenges. *The scientific world journal 2014*.

Kumar, P. R., P. H. Raj, et P. Jelciana (2018). Exploring data security issues and solutions in cloud computing. *Procedia Computer Science 125*, 691–697.

Lee, K. (2020). Comments on "secure data sharing in cloud computing using revocable-storage identity-based encryption". *IEEE Transactions on Cloud Computing 8*(4), 1299–1300.

Liang, K., M. H. Au, J. K. Liu, W. Susilo, D. S. Wong, G. Yang, Y. Yu, et A. Yang (2015). A secure and efficient ciphertext-policy attribute-based proxy re-encryption for cloud data sharing. *Future Generation Computer Systems 52*, 95–108.

Lin, L., T. Liu, J. Hu, et J. Zhang (2014). A privacy-aware cloud service selection method toward data life-cycle. In *2014 20th IEEE international conference on parallel and distributed systems (ICPADS)*, pp. 752–759. IEEE.

Liu, G., H. Dong, Z. Yan, X. Zhou, et S. Shimizu (2020). B4sdc : A blockchain system for security data collection in manets. *IEEE Transactions on Big Data*.

Ma, S. (2016). Identity-based encryption with outsourced equality test in cloud computing. *Information Sciences 328*, 389–402.

Ma, X., P. Fox, E. Rozell, P. West, et S. Zednik (2014). Ontology dynamics in a data life cycle : challenges and recommendations from a geoscience perspective. *Journal of Earth Science 25*(2), 407–412.

Mandal, M. (2021). Anonymity in traceable cloud data broadcast system with simultaneous individual messaging. *International Journal of Information Security 20*(3), 405–430.

Michelin, R. A., A. Dorri, M. Steger, R. C. Lunardi, S. S. Kanhere, R. Jurdak, et A. F. Zorzo (2018). Speedychain : A framework for decoupling data from blockchain for smart cities. In *Proceedings of the 15th EAI international conference on mobile and ubiquitous systems : Computing, networking and services*, pp. 145–154.

Michota, A. et S. Katsikas (2015). Designing a seamless privacy policy for social networks. In *Proceedings of the 19th panhellenic conference on informatics*, pp. 139–143.

Qiu, S., J. Liu, Y. Shi, M. Li, et W. Wang (2015). Identity-based private matching over outsourced encrypted datasets. *IEEE Transactions on cloud Computing 6*(3), 747–759.

Rahul, K. et R. K. Banyal (2020). Data life cycle management in big data analytics. *Procedia Computer Science 173*, 364–371.

Rivest, R. L., A. Shamir, et D. A. Wagner (1996). Time-lock puzzles and timed-release crypto.

Shao, J., G. Wei, Y. Ling, et M. Xie (2011). Identity-based conditional proxy re-encryption. In *2011 IEEE International Conference on Communications (ICC)*, pp. 1–5. IEEE.

Tabrizchi, H. et M. Kuchaki Rafsanjani (2020). A survey on security challenges in cloud computing : issues, threats, and solutions. *The journal of supercomputing 76*(12), 9493–9532.

Tao, H., M. Z. A. Bhuiyan, A. N. Abdalla, M. M. Hassan, J. M. Zain, et T. Hayajneh (2018). Secured data collection with hardware-based ciphers for iot-based healthcare. *IEEE Internet of Things Journal 6*(1), 410–420.

Vidgen, R., S. Shaw, et D. B. Grant (2017). Management challenges in creating value from business analytics. *European Journal of Operational Research 261*(2), 626–639.

Wang, H. et Y. Song (2018). Secure cloud-based ehr system using attribute-based cryptosystem and blockchain. *Journal of medical systems 42*(8), 1–9.

Xu, P., T. Jiao, Q. Wu, W. Wang, et H. Jin (2015). Conditional identity-based broadcast proxy re-encryption and its application to cloud email. *IEEE Transactions on Computers 65*(1), 66–79.

Xue, L., Y. Yu, Y. Li, M. H. Au, X. Du, et B. Yang (2019). Efficient attribute-based encryption with attribute revocation for assured data deletion. *Information Sciences 479*, 640–650.

Yadav, D. K. et S. Behera (2020). A survey on secure cloud-based e-health systems. *EAI Endorsed Trans. Pervasive Health Technol. 5*(20), e2.

Yang, C., Q. Chen, et Y. Liu (2019). Fine-grained outsourced data deletion scheme in cloud computing. *International Journal of Electronics and Information Engineering 11*(2), 81–98.

Yang, C., X. Chen, et Y. Xiang (2018). Blockchain-based publicly verifiable data deletion scheme for cloud storage. *Journal of Network and Computer Applications 103*, 185–193.

Yang, C., Y. Liu, X. Tao, et F. Zhao (2020). Publicly verifiable and efficient fine-grained data deletion scheme in cloud computing. *IEEE Access 8*, 99393–99403.

Yang, Y., H. Zhu, H. Lu, J. Weng, Y. Zhang, et K.-K. R. Choo (2016). Cloud based data sharing with fine-grained proxy re-encryption. *Pervasive and Mobile computing 28*, 122–134.

Yu, Y., L. Xue, Y. Li, X. Du, M. Guizani, et B. Yang (2018). Assured data deletion with fine-grained access control for fog-based industrial applications. *IEEE Transactions on Industrial Informatics 14*(10), 4538–4547.

Zhang, P., S. Wang, K. Guo, et J. Wang (2018). A secure data collection scheme based on compressive sensing in wireless sensor networks. *Ad Hoc Networks 70*, 73–84.

Zhou, Y., H. Deng, Q. Wu, B. Qin, J. Liu, et Y. Ding (2016). Identity-based proxy re-encryption version 2 : Making mobile access easy in cloud. *Future Generation Computer Systems 62*, 128–139.

Summary

One of the most serious issues with cloud computing is data security, As businesses start on digital transformation, there is a clear requirement for privacy and data protection. Organizations today have more data, applications, and websites than they have ever had before. Data security has risen to the top of the priority list for cloud computing security. Despite the fact that a variety of solutions have been proposed, the majority of them only address one stage of the data life cycle, such as storage, which is insufficient to handle the cloud data security challenge because threats appear at all stages of the data life cycle. During the data life cycle process, any stage's security breaches could affect data security. Therefore, data security must be considered throughout the data lifecycle. The main contribution of this article is a new perspective on data security solutions based on the data lifecycle, which is crucial and can be used as a guide to create a complete security solution.A literature review on the entire data life cycle is carried out and a research gap of unresolved issues that may be research questions for our future work is presented.Also a proposed solution on data labeling used for data tracking to secure data in all stages in data life cycle is presented.