iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XI **Month of publication:** November 2023

DOI: https://doi.org/10.22214/ijraset.2023.57098

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

Analysis of Data Leakage Detection Algorithms

Sree Vagdevi Kandukuri[1], Shivang[2], Chirag[3]
AIT-CSE (Cloud Computing), Chandigarh Univeristy, Gharaun, India

Abstract: ***Data refers to information about a particular mention. There are several types of data and they can be stored either physically or virtually. The use of Cloud Computing has gained popularity in many IT organizations in order to store data effectively. With the growth of the population, data management becomes more critical. A management system relies heavily on the protection of data and preventing leakage. A data leak not only damages the company's reputation but also damages its customers' trust. It is high time to address the issue, especially considering previous incidents like Yahoo's data breach that compromised over 2 billion accounts, Friend Finder Networks' leak that caused 412 million accounts to be compromised, and more. In this paper, we present methodologies for detecting and preventing data leakage in cloud computing. The cloud platform is being used to perform various algorithms, which will lead to an efficient detection of data leakage, as well as an immediate identification of guilt.***

Keywords: (Data Leakage, cloud computing, advanced encryption technique, watermarking)

I. INTRODUCTION

Cloud computing is the latest and most intriguing technology in information technology with almost every IT company looking to get involved. The principle of cloud computing is the sharing of information resources and software on demand among devices. Cloud services include data storage as one of their most basic functions. Data can be stored and monitored remotely with cloud computing, which frees employees from the burden of storing and monitoring it on-site. However, it also exposes files to a significant level of privacy risk. In cloud computing, virtualization plays a central role. The reason is that it abstracts resources into isolated virtual computing environments, thereby releasing applications, servers, storage and desktops from their dependence on physical hardware layers. In a cloud platform, virtual machines are called virtual machines (VMs), and replication and migration are common operations for VMs. There is a risk of data leakage during those operations due to misconfiguration, software bugs (e.g., hypervisors), or poor management practices.

A. What is Data Leakage?

Data leakage occurs when unauthorized data is transmitted from an organization to an external source. Data may be leaked physically or electronically via hard drives, USB devices, mobile phones, etc., and could be exposed publicly or fall into the hands of hackers. An organization's firewall may be breached allowing unauthorized access to data or information. Information leakage can refer to the unauthorized passage of data or information inside the organization. Leakage of data can occur either electronically, such as data transmitted over the internet, or physically, such as on USB flash drives or hard drives. Businesses today must consider data leakage when considering cybersecurity and can prevent it through the use of tools and education.

It is impractical to simply download all the data from the cloud for data integrity verification because of the high storage and communication costs. The issue with many of these organizations adopting these updated tech solutions was that, from a security standpoint, a lot of companies inadvertently created vulnerabilities in their data and information by not establishing comprehensive protocols for their cybersecurity.

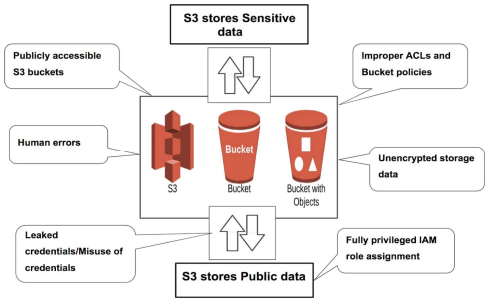


Fig.1. Causes of cloud storage breaches (www.google.com)

The problem isn't necessarily related to the business itself; it is rather an indication of the challenges organizations have faced due to the new working circumstances -- especially those that were forced to develop their communication stacks on the fly. The term cloud leak refers to data stored in a private cloud instance that is accidentally exposed to the internet. The cloud is part of the internet. Cloud computing is different from traditional computing because it offers pockets of privatized space where enterprise-scale IT operations can take place.

II. LITERATURE SURVEY

Prisca I. Okochi et.al. (2021) discuss methodologies to detect data leakage using cloud computing in the paper "An improved data leakage detection system in a cloud computing environment". For the purpose of this paper, the method "Object-Oriented Analysis and Design (OOAD)" is employed to analyse and design objects based on their properties. Objects and similar objects are grouped into classes, and their characteristics are considered properties while their behaviours are considered methods or actions within the same bundle of objects. This is a structured approach to analysing and designing systems. During the software development lifecycle, it develops a set of graphical system models using object-oriented concepts. As a result, this approach is best suited for systems with multiple objects.

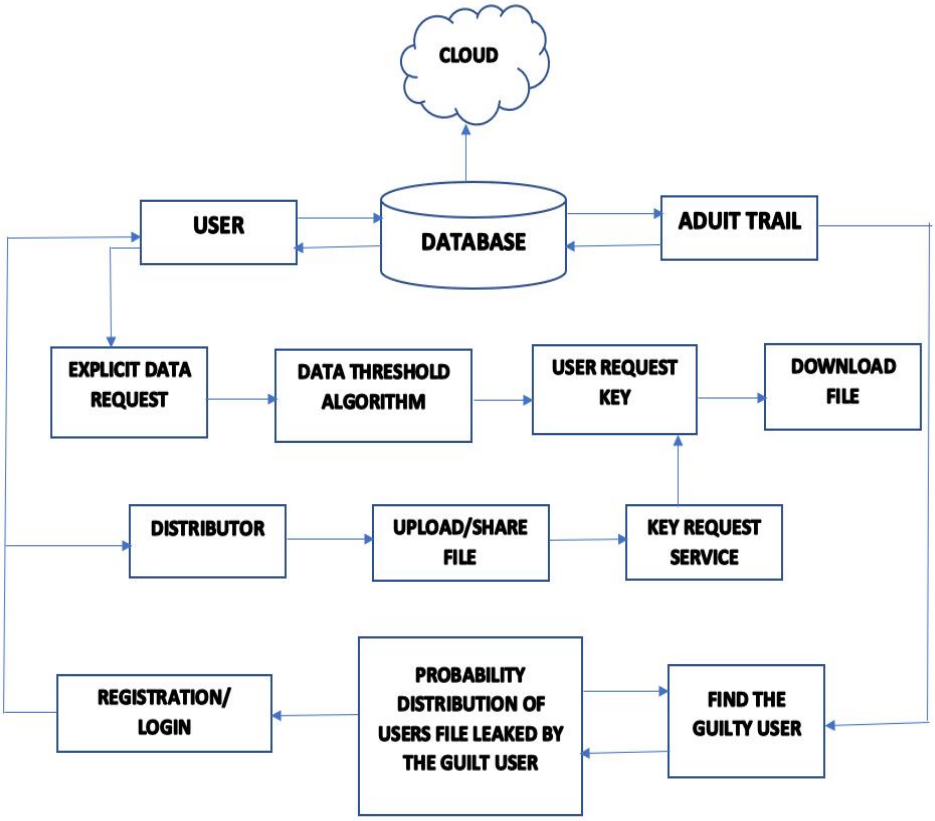


Fig.2.Architecture for OOAD.

Khadijah Ab Rahman et.al (2019) has come up with various methodologies in the paper, titled “Data Leakage Detection in Cloud Computing”. The concept of covering data leakage due to vulnerabilities of the hypervisor and dashboard of the cloud management software is carried out. Moreover, the two main research approaches used for the demonstration are VM Migration Process, and the Cloud Dashboard Authentication experiment. In the VM migration process, the experiment will detect if sensitive information will leak out while replicating and migrating AD and text files. As a result of the second experiment, a user authentication process is undertaken between the cloud user and the web dashboard login page, which demonstrates that data is being leaked as the communication session in OpenStack software is not encrypted. Thus, the method proves to be effective in detecting data leakage on cloud platforms. In comparison with related methods, other methods focus more on application-level data leakage. Future investigations will focus on detecting data leakage on other cloud components, such as block storage, API requests, and telemetry.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

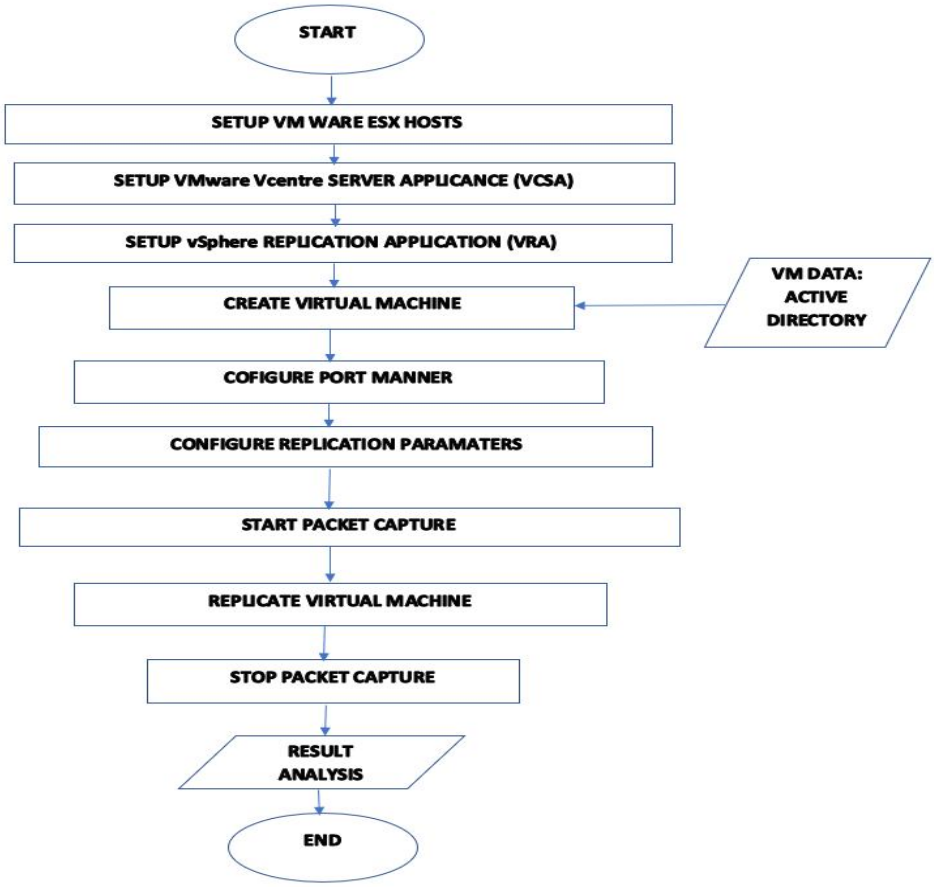


Fig.3.Flowchart for VMware migration process experiment

Sushilkumar N. Holambe et.al(2015) in the paper "Data Leakage Detection Using Cloud Computing" has discussed data leakage that is an increasingly important issue in the field of cloud. The methodology of data allocation that the distributor can use to determine whether there was a data leak or a breach when transmitting the data has been explained by the author. Techniques for implicit and explicit data allocation are employed, both with and without the usage of fictitious data objects. The calculation of guilt probability by comparing the distributed data set to the leaked data set comes after the distribution of the data by the distributor and is a crucial step in the identification of data leaks. The findings add to our understanding that watermarking along with data allocation and guilt assessment can become one of the best approaches in data leakage detection.

S. Visnu Dharsini et.al (2019) in the research paper "Data Leakage Detection and Security in Cloud Computing Environment" suggested a model to detect data leaks and improve data security during the transmission of data. The proposed paradigm combines data encryption algorithms with data allocation strategies. Secure Hashing Algorithm (SHA) is the encryption algorithm employed in this case; it compresses and encrypts the input data before returning encrypted information known as hash values. Fake objects are utilized in data allocation techniques to identify the leaking point, and the idea of cryptographic keys is applied to identify the guilty party. Faster cryptographic algorithms may be utilized in the future to accelerate encryption and decryption during data transmission, even though the model is capable of addressing the issues of data leakage and security.

Riya Naik.et.al (2019), in the paper titled "Data Leakage Detection in cloud using Watermarking Technique" presented an approach for detecting data leakage in clouds. Researchers have developed a variety of methods for detecting leakages in the field of data security. To identify tampering and data leaks in the cloud, several techniques have been documented in literature. For the purpose of detection, the algorithm is known as "Advanced Encryption Standard (AES)". A frequency domain method is used in the watermarking algorithm, which increases efficiency and resilience. It is primarily information retrieval and encryption that constitute the data transfer phase. In order to uniquely identify the data, information is obtained about it. After retrieving the information, the message is created using that information and the recipient's client ID. Different steps are involved in generating QR codes.

Chaoshun Zuo.et.al (2019) proposed to detect data leakage in applications built on cloud in the paper titled “Why Does Your Data Leak? Uncovering the Data Leakage in Cloud from Mobile Apps”. To detect breaching in data we use a technique which relies on the concept of API’s. More precisely we use generating signatures for each and every function present in the application. Increasing the coverage of analysis is the 3rd avenue for future work.

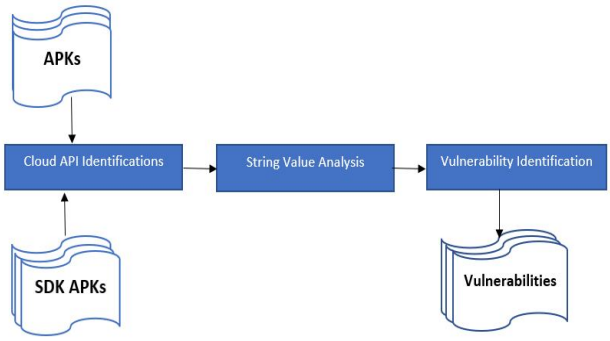


Fig.4. An Overview of Our Leak Scope

Jianbing Ni.et.al, Kuan Zhang.et.al (2020) proposed a method in their paper named “Identity-based Provable Data Possession from RSA Assumption for Secure Cloud Storage" for detecting data leakage which is based on RSA assumption. It has been found that an identity-based privacy-preserving provable data ownership scheme (ID-P^{3}DP) has excellent performance for the purposes mentioned above. Cloud users can generate identity-based homomorphic authenticators using the outsourcing file and a global parameter over a period of time using ID-P^{3}DP, and third-party auditors (TPAs) can verify the validity of homomorphic authenticators to verify the integrity of outsourced files. As a result, we can provide an alternative realization for remote data integrity verification under different cryptographic assumptions.

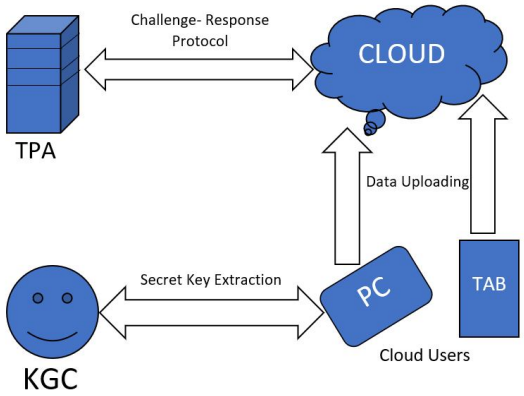


Fig.5. Identity-based privacy-preserving provable data possession scheme (ID-P^{3}DP)

Hanyi Zhangl.et.al, Liming Fangl.et.al (2020) came up with a scheme to protect data present in the cloud in the paper “Secure Door on Cloud: A Secure Data Transmission Scheme to Protect Kafka's Data”. In Yahoo, Facebook, Alibaba, and other companies, Kafka is widely used as an asynchronous message queue due to its scalability, reliability, and high performance. Apache Kafka is currently being integrated with more and more open-source distributed processing systems. A large data stream processing platform can make use of the characteristics of its message systems. To prevent data leakage in Kafka, a fine-grained data transmission scheme called Secure Door on Cloud (SDoC) is used. In addition to being more secure than the built-in security mechanism in Kafka, SDoC can also effectively prevent unauthorized users from stealing plaintext information. When consumers need to consume messages, they need to apply to A, so that A can produce a re-encrypted key for the corresponding ciphertext. According to the authorization list, A determines the rights of access to ciphertext when generating re-encrypted keys. Kafka is used to execute the re-encryption algorithm so that Kafka can fully utilize its computing resources.

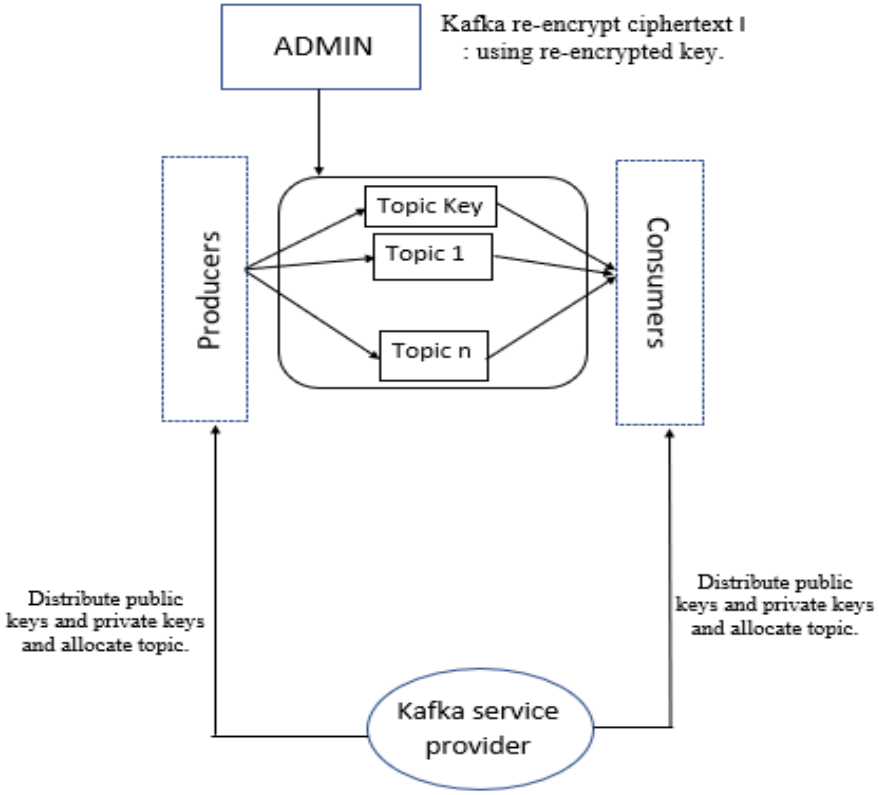


Fig.6.The execution process of SDoC.

Neeraj Kumar.et.al, Vijay Katta.et.al (2014) proposed a method to detect data leakage in the cloud. The method was described in the paper labelled as "Detection of Data Leakage in Cloud Computing Environment". For achieving the same Bell-LaPadula model is used which is based on the concept of a state machine with a set of allowable states in a computer system. AES algorithm and RSA algorithm show good performance among different symmetric and asymmetric encryption techniques based on different performance factors such as key value, computational speed and tenability. We are using the concept of Bell-LaPadula Model for providing secured infrastructure. In this model server will add an image logo to all the stored documents and this image logo represents the organization. As we know that each intensity value in the image ranges from 0 to (224 − 1), and for each of the three components of the color image as RED, GREEN and BLUE ranges from 0 to (28 −1). Each character has their ASCII values ranges from 0 to (28 − 1). Therefore, this can be useful in a distributed computing environment to protect data from data leakage. The proposed technique is based on a symmetric algorithm; therefore, it is infeasible to extend this model for web environments where multiple users frequently access the data object. We can also implement this technique for asymmetric cryptography.

Tushar Aggarwal.et.al, Narinder Kaur.et.al (2019) introduced an innovative method for detecting data leakage in the paper named as "Data Leakage Detection using Cloud Computing". Data is mainly sent by the distributors which are generally the owner of data to the user who wants the information, mainly the trusted third parties. The digital information shared by the distributor should be confidential and must be shared by a secure way. In some scenarios, the data distributed by the organization is copied by different agents who cause a huge damage to the organization and this is termed as data leakage. The data leak must be detected as early as possible in order to prevent the confidential data from being public or for some malicious use. To achieve the same data can be protected by giving a special inscription to the sensitive data so that it cannot be reproduced. They are spatial domain and frequency domain. Out of these two, frequency domain is more used than spatial domain watermarking.

Table I
DIFFERENT METHODOLOGIES TO DETECT DATA LEAKAGE

Author	Year	Technology	Findings	Limitation
Khadijah Ab Rahman	2021	VM Migration Process, and the Cloud Dashboard Authentication experiment	The experiments detect If sensitive information will leak out while replicating and migrating AD and text files along with the user authentication. The method proves to be effective in detecting data leakage on cloud platforms.	The experiments detect if sensitive information will leak out while replicating and migrating AD and text files along with the user authentication. The method proves to be effective in detecting data leakage on cloud platforms.
Jianbing Ni, Kuan Zhang	2020	Data leakage can be detected by provable data possession method which is based on RSA. It uses the ID-P3DP scheme.	It has been found that an identity-based privacy-preserving provable data ownership scheme (ID-P^{3}DP) has excellent performance. It can be used with any type of data. Moreover, ID-P^{3}DP can be extended to support other desirable properties, including data updating, data deduplication, and user revocation, based on the mature techniques.	-
Hanyi Zhangl, Liming Fangl	2020	To prevent data leakage in Kafka, a fine-grained data transmission scheme called Secure Door on Cloud (SdoC) is used.	Secure Door on Cloud (SdoC) is a scheme which can be used to detect leakage in Kafka Data. In order to evaluate the performance of SdoC, we can simulate normal inter-entity communication. By comparing Kafka with SdoC integration with Kafka with built-in security mechanisms, we find that Kafka with SDoC integration has a lower data transfer time overhead.	The time cost is slightly higher than that of the Kafka without security mechanism enabled, which is caused by the time loss of encryption and decryption. In the SDoC scheme, the queue congestion becomes more and more serious due to the limitation of participants' computing capacity.
Prisca I. Okochi	2019	Object-Oriented Analysis and Design (OOAD)	The introduced methodology performed well in analyzing and approaching the detection of data. During the software development lifecycle, it develops a set of graphical system models using object-oriented concepts. As a result, this approach is best suited for systems with multiple objects.	The technique could not identify the leakage due to vulnerabilities of the hypervisor and dashboard of the cloud management software.
Riya Naik	2019	Advanced Encryption Standard using watermarking technology	The algorithm detects if there is any tampering in the images or videos. It generates a QR code for each and every image which can be analysed effectively.	The algorithm cannot work effectively with text and needs a QR code generator support at both the ends.

				There are chances of little error as QR codes cannot be 100% accurate.
Chaoshun Zuo	2019	Concept of API i.e REST	This concept works by generating signatures for every function uniquely	The system works perfectly but needs regular updates as the number of API's increases. It can only be used with the apps that are built by using cloud API's. Increasing the coverage of analysis is the 3rd avenue for future work.
S. Visnu Dharsini	2019	Cryptographic algorithms with the fake object model	It relies on transmitting messages in encrypted form and using fake objects to look for points of leakage while also protecting the data.	Data transfers can take longer when encryption and decryption are involved.
Tushar Aggarwal, Narinder Kaur	2019	Watermarking which is based on steganography	Data can be protected by giving a special inscription to the sensitive data so that it cannot be reproduced. Watermarking can be done by various techniques like spatial domain watermarking, least significant bit, Frequency domain watermarking, Discrete cosine transform, Discrete wavelet transform etc. Out of all methods in frequency DWT is the most effective watermarking algorithm.	This technique can be used to protect audio, video and image. But this kind of modification can lead to degradation of the data signal.
Sushilkumar N. Holambe	2015	Data allocation with calculation of guilt probability	The study offers a methodology based on sending fictitious objects to locate guilty agents and then computing the guilt probability to identify the leakage.	When we supply fake objects, agents may obtain inaccurate information, which may lower the level of service.
Neeraj Kumar,Vijay Katta	2014	Bell-LaPadula model is used for detection which is based on AES and RSA algorithms.	We use the concept of Bell-LaPadula for infrastructure.It is efficient with any size of document. This can also be used with any type of images.It also protects different types of attacks like passive and active attacks.	RSA leads to two major problems. First its key length is very high. Second it will be more difficult to calculate exponential computations than simple computations.

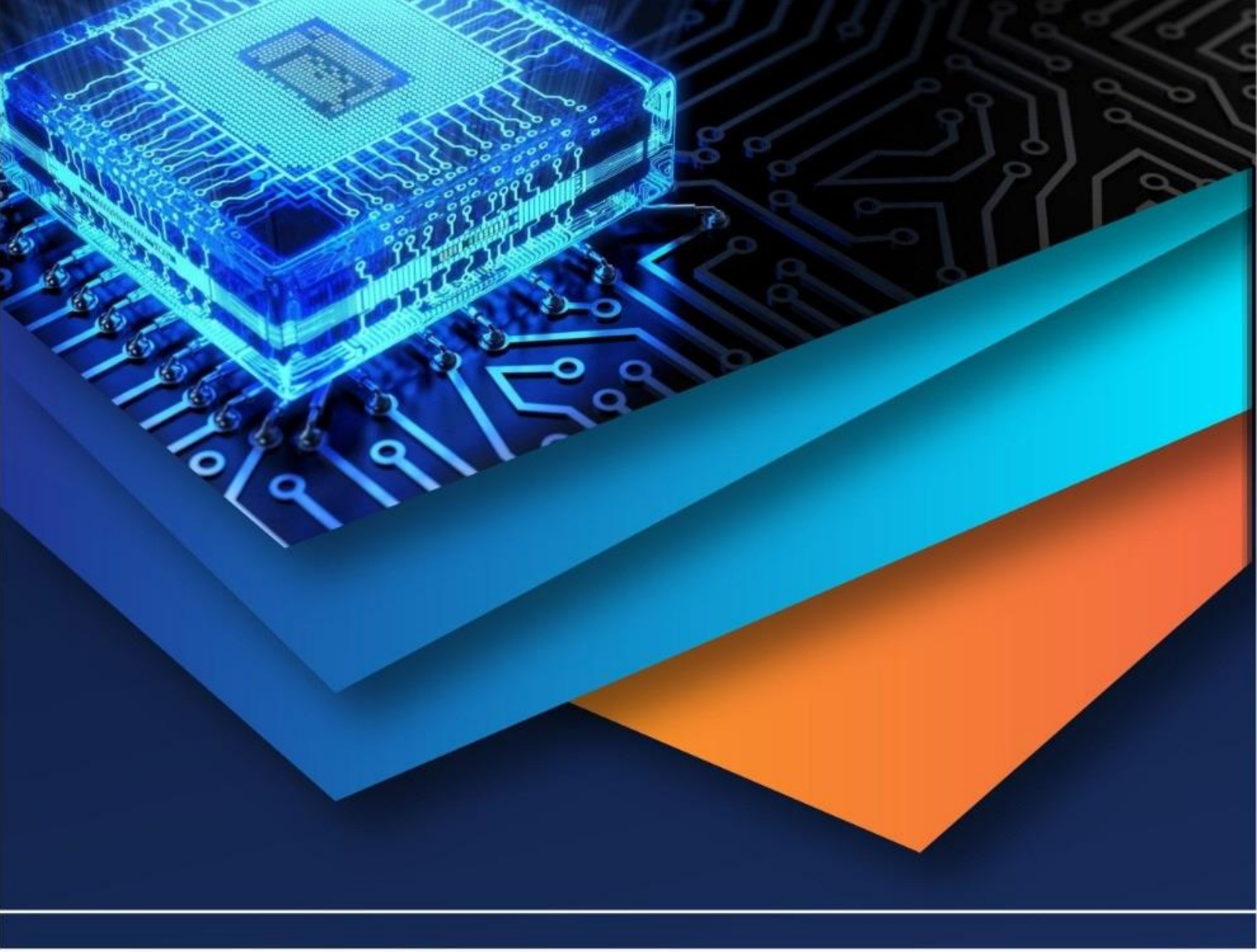
III. CONCLUSION

When a third party can access data without authorization either during data transmission or after it has been transmitted, this is referred to as data leakage. In order to prevent data leaking, we must first identify the source of the leak and secure our data during transmission so that, in the event that data is leaked, the breacher is unable to decipher the data. In this research, we analyzed a number of methods that are currently being utilized in various cloud systems to find data leaks. We also talked about leveraging encryption techniques like secure hashing, RSA, and AES to encrypt data while it is transmitted.

One way is to use the AES algorithm with watermarking, or a QR code, however, the main disadvantage is the requirement for a QR code generator at both ends. An important finding to emerge in this study is identity-based privacy-preserving provable data possession scheme (ID-P^3DP). The approach is diffused to (blood bank name) alleviate barriers and improve system efficiency access. The web application follows quantity and quality methods to build the algorithm for the whole system. Various software and languages as project tools play an essential role, providing an excellent foundation for the entire structure.

REFERENCES

[1] https://www.researchgate.net/publication/354308045: An improved data leakage detection system in a cloud computing environment
[2] International Journal of Advanced Trends in Computer Science and Engineering, Available Online at http :Data Leakage Detection in Cloud Computing Platform.
[3] International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015 1255 ISSN 2229-5518:Data Leakage Detection Using Cloud Computing
[4] International Research Journal of Engineering and Technology (IRJET) Volume: 06 Issue: 03 | Mar 2019: DATA LEAKAGE DETECTION USING CLOUD COMPUTING
[5] INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 11, NOVEMBER 2019 IJSTR©2019 www.ijstr.org Data Leakage Detection And Security In Cloud Computing Environment
[6] 2019 International Conference on Computer Communication and Informatics (ICCCI -2019), Jan. 23 – 25, 2019, Coimbatore, INDIA Data Leakage Detection in cloud using Watermarking Technique.
[7] AFOSR under grant FA9550-14-1-0119, and NSF awards 1718084, 1834213, and 1834215: Why Does Your Data Leak? Uncovering the Data Leakage in Cloud from Mobile Apps Chaoshun Zuo, Zhiqiang Lin, Yinqian Zhang The Ohio State University
[8] 2014 Sixth International Conference on Computational Intelligence and Communication Networks: Detection of Data Leakage in Cloud Computing Environment.
[9] International Research Journal of Engineering and Technology (IRJET) Volume: 06 Issue: 11 | Nov 2019: Data Leakage Detection using Cloud Computing Tushar Aggarwal, Narinder Kaur.
[10] 2020 IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS)Secure Door on Cloud: A Secure Data Transmission Scheme to Protect Kafka's Data.
[11] JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015 1 Identity-based Provable Data Possession from RSA Assumption for Secure Cloud Storage Jianbing Ni, Member, Kuan Zhang, Member, Yong Yu, Member, Tingting Yang, member, IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY