Data Leakage Detection and Prevention Using Cloud Computing

Vanshika Singh, Manish Raj, Indrajeet Gupta, and Mohd Abuzar Sayeed

1 Introduction

Data leakage refers to the accidental disclosure of critical data that can cause it to be exposed to cyber-attackers and criminals which can cause huge damage to the organization and data subjects. Intellectual property (IP), monetary details, healthcare documents, credit report data, and other details related to the organization and market are examples of critical data held by corporations and businesses. Furthermore, the critical data is frequently shared across a range of consumers, including workers operating from outside the corporation (e.g., on computers), corporate associates, and consumers. Any unauthorized access to the organization-specific data can result in leak of sensitive information, which can cause certain implications for the business model itself. Software as a Service (SaaS) is defined in cloud computing as technology that acts as a middleman among both the consumer and the web and deploys via the internet. Where a provider authorizes a software to consumers as a service on request, through a membership, in a "pay-as-you-go" arrangement, or (progressively) for no cost when there are opportunities to make money from sources other than the customer, it is known as SaaS [1]. Because of this tremendous expansion of SaaS, it will soon be ubiquitous within every business establishment. It is critical that software purchasers and consumers comprehend the applicability of SaaS. In information security and computer security, cloud computing security is one of the major concerns.

It is usually recommended that Information Security Controls (ISCs) be established and deployed in accordance with and in proportion to the vulnerabilities, and repercussions [1]. Cloud computing is provisioning and assigning IT Services

V. Singh · M. Raj (✉) · I. Gupta · M. A. Sayeed
Bennett University, Greater Noida, Uttar Pradesh, India
e-mail: e19cse013@bennett.edu.in; manish.raj@bennett.edu.in; indrajeet.gupta@bennett.edu.in; abuzar.sayeed@beneett.edu.in

S. Awasthi et al. (eds.), *Sustainable Computing*,
https://doi.org/10.1007/978-3-031-13577-4_9

and resources over the internet to individuals and enterprises. The main advantage of cloud computing is you only pay for the resources that you use. You do not need to set up datacenters and operate them, which results in reducing the capital expenditure. Due to the growing popularity of cloud computing, we can find examples of many tasks that require high computing power. Examples may include Research Labs, Artificial and Machine Learning use cases, and Complex System Simulations [1]. These tasks may lead to cloud resources being unutilized, which in turn creates cloud wastage.

We observed that many people who are new to cloud computing are not handling there resources properly, which eventually leads to wastage of lots of resources and many resources being unutilized. First of all, our innovation is predicting cloud resource utilization using a machine learning model, using the compute resources dataset of specific tasks. Even the cloud platforms are currently not exploring this problem statement which makes it more unique. Apart from that, we will also predict cloud wastage utilization which will help individuals, companies, and other institutions to manage their cloud utilization systematically and cost-efficiently. we will create a new model from scratch which will help our solution be revolutionary in this sector.

The billed amount for clouds is frequently more than expected. It's because people don't understand how to handle cloud resources. People have yet to fully grasp cloud computing, and their inexperience shows. Several corporations have clouds all over the world but fail to make optimum use of the cheapest ones. It contributes to cloud waste and latency difficulties. You will suffer a greater fee if you use an Indian server for any operation in the United States. If we don't act quickly, the prices of these unutilized units will rise over time. Because of shifting demand, the cloud manager must be able to utilize resources by providing and de-provisioning resources to match the need. Inadequate resource provisioning results in Service Level Agreement (SLA) violations, poor Quality of Service (QoS), and performance deterioration, which leads to customer dissatisfaction [1]. Overprovisioning, on the other hand, results in resource waste, which raises the cost and energy.

A careful examination of the dynamic and precise resource provisioning is required for the system's smooth operation [2]. Accurate prediction should be used to determine the proper quantity of resource to meet the needs. To accurately anticipate future workload, it is necessary to use a trustworthy and exact prediction model. Typically, in Cloud Data Centers, user tasks occur in an irregular schedule with varying resource requirements. This scenario makes predicting the specific workload extremely difficult. Cloud computing is provisioning and assigning IT Services and resources over the internet to individuals and enterprises. Due to the growing popularity of cloud computing, we can find examples of many tasks that require high computing power. Examples may include Research Labs, Artificial and Machine Learning use cases, and Complex System Simulations.

These tasks may lead to cloud resources being unutilized which in turn creates cloud wastage. Accurate prediction should be used to determine the proper quantity of resources to meet the needs. Hence, we have decided to use cloud computing data

to create a model which can accurately predict the usage and wastage of computing resources to demonstrate a system for predicting resource utilization in real time.

The system monitors resource consumption in real time and caters utilization value to various buffers based on the type of resource and time interval. The data in these buffers is checked to see if it follows a Gaussian distribution. Autoregressive Integrated Moving Average is used when the distribution is Gaussian; or else, Autoregressive Neural Network is used. A model is chosen in the ARIMA process based on small Akaike Information Criterion value. On the other hand, if Network Information Criterion value is low, then the Autoregressive Neural Network method is used. We tested our approach using real-time CPU utilization data from an IaaS cloud.

1.1 Motivation

Today's world is heavily reliant on data transmission, or the flow of information from one individual to another. The information sent by the supplier should be safe, secret, and unique, as the information exchanged with authorized third parties is sensitive and proprietary. Leakage occurs whenever information is accessed, viewed, or tampered with, while being uploaded to the cloud [2].

Based on the Ponemon Institute's Cost of a Data Breach Report, a yearly compilation of data leak patterns that has become something of a touchstone for the data protection sector over the years, security breaches have caused approximately 3.86 million dollars of loss in 2020. Healthcare industries have experienced the greatest expenditures connected with a data leak for the tenth year in a row. According to an IBM report, healthcare data theft ended up costing the companies $7.1 million on aggregate, up a smidgeon from previous year's figure ($6.45 million).

The energy sector, which was the second most expensive, caused businesses an average of 6.39 million dollars. According to Ponemon, sectors with more stringent legal requirements had greater information theft expenses this year. The more serious a data breach is, the more likely a company would lose revenue, which might illustrate why the healthcare, energy, finance, and pharmaceutical enterprises were among the most impacted [3].

2 Data Loss Prevention

In a cloud setting, a virtual machine may be deployed to operate a security engine that manages all the remaining virtual machines on a defined range of virtual servers using virtual machine management (VMM) architecture. Client software with a DLP engine may then be launched on virtual computers, which will scan, detect, and stop the flow of confidential data. The VMM may combine them into a single virtual

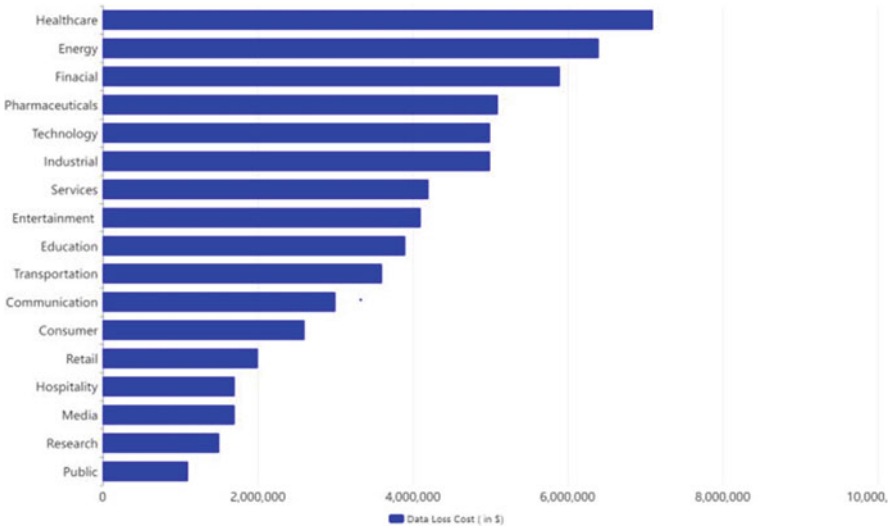

Fig. 1 Data breach cost 2020

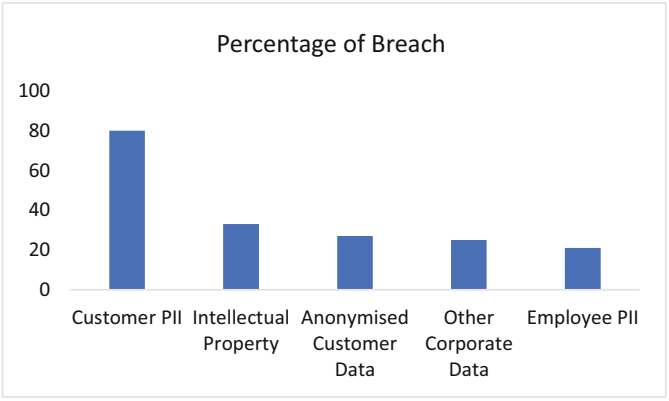

Fig. 2 Percentage of breach involving data loss

machine, allowing the DLP engine to monitor and control all the virtual machines that operate a customer, as well as observe information in transit [4] (Figs. 1 and 2).

This expands the potential for confidential information protection standards such as PCI DSS, PII, and others. DLP is a technology that may be turned on or off for virtual machines in the cloud data center. A DLP system, like a cloud infrastructure, is dynamic, since it can be extended and operated [5]. APIs can be used to orchestrate controls in a DLP solution, such as creating a regulation that moves a virtual machine with confidential information behind a firewall or puts it on shutdown. Some of the key advantages of major cloud DLP solutions are listed below:

1. Develop with cloud storage services to inspect systems for confidential information, detect it, and encrypt it before it is transmitted in the cloud
2. Analyze and verify information that has previously been saved in the cloud at any moment
3. Retrieve confidential information from the cloud with accuracy. Inspect transmitted documents on a regular basis
4. Instantly deploy restrictions to confidential information in line with corporate policy (prompt, block, or encrypt)
5. Whenever the information is exposed or an unauthorized access is detected, immediately notify the relevant authorities and data proprietors
6. Retain the transparency and authority required to adhere to confidentiality and information security laws

2.1 Limitations of Cloud DLP

If the cloud infrastructure is public, each instance may only have one network connection, necessitating the use of a virtual DLP version that may detect, transmit, or prohibit traffic with restrictions. While utilizing DLP to detect data migration to the cloud and for content discovery on cloud-based storage has a lot of value, adopting DLP in a public cloud may not. The application architecture, which relies more on application security and encryption, is perhaps the most secure part of any cloud implementation in accordance with DLP. DLP is a fantastic tool for improving information security in the cloud. Given that it is adjusted appropriately, it may be utilized to monitor information transferring to the cloud, find critical data stored on the cloud, and safeguard cloud-based applications.

3 Related Work

3.1 Watermarking Technique

Watermarking, in which a distinctive code is included in each disseminated duplicate, has historically been used to identify breaches. The leaker can be determined if that duplicate is subsequently acquired in the possession of an unauthorized person. Watermarks are effective in some situations, although they do need some change of the initial material. Moreover, if the data consumer is hostile, watermarks can be removed [6]. A healthcare institution, for example, may provide patient details to researchers who may develop new medicines. Likewise, a business may form alliances with other businesses that need the exchange of client information. Because other company may subcontract data processing, data must be handed to a number of different firms. The data proprietor is referred to as the distributor, while the ostensibly trustworthy third parties are referred to as agents. Watermarking Technique [7].

3.2 Fake Objects Method

Fake objects may create less issues in some applications than genuine things. Consider health documents as dispersed data objects, and hospitals as the agent. In this situation, even slight alterations to the data of genuine patients might be hazardous [8]. The inclusion of certain fictitious health documents, on the other hand, may be allowed because no patient resembles this information, and so no one will ever be diagnosed relying on fictitious information. In this scenario, corporation A provides a mailing list to corporation B that will only be utilized once (e.g., to send advertisements). Corporation A adds trace data with corporation A's addresses. As a result, whenever corporation B utilizes the acquired mailing list, duplicates of the mailing are sent to A. These records are a form of fictitious object, that aids in the detection of data misuse [9].

3.3 Adversary Model

This approach captures all types of information security concerns, and cloud data is not designated at the cloud customer level, but rather at the cloud service level [10]. There are two forms of attacks capable of causing this:

1. Internal Attack: The cloud service provider might not be trustworthy
2. External Attack: The attacks launched by unauthorized users

4 Secure Hashing Algorithm (SHA)

SHA is an abbreviation for Secure Hash Algorithm [11]. Hashing techniques are computationally intensive routines that squeeze, encode, and encrypt the ingress data and generate hash or hash values, which appear to be random. These seemingly random numbers are really the input data in an encrypted or coded format. Hash values of data are relatively easy for computers to deploy than the actual data because hashing allows the system to execute numerous procedures or calculations over directories and data chunks. A hashing algorithm must be predestined, which means it must return symmetrical results for each ingress value. The SHA secure algorithm provides a hash function during data transfer with a reference to a third-party agent and the actual cloud; this authentication mechanism is introduced to help secure the transfer of essential information, over the internet. The SHA algorithm is utilized because message digests are needed to ensure data security and authenticity. In this method, the data is secured using SHA-2 (Secure Hash Algorithm 2), a subset of the cryptographic primitive family Keccak [1] SHA-2 was designed with an internal block size of 1024 and can perform And, Xor, Rot, Add (mod 264), Or, Shr, and contains 256–128 security bits [12].

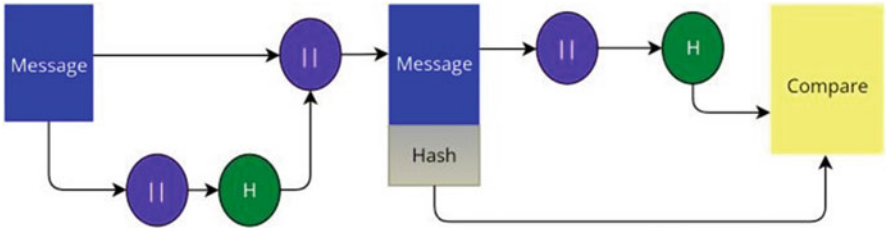

Fig. 3 Hash function working

The Third-Party Auditor (TPA) ensures performance, integrity, and impartiality in necessary auditing operations and acts as a negotiator between CSUs and CSPs [13]. Corporations and enterprises do not have their own storage places for their databases and other confidential information in today's reality. They procure disk space from other entities on bill per use basis. TPA serves as an administrator in this case, governing user access control and other administrative rights such as uploading incoming data or modifying current data to entities in the data communication network (Fig. 3).

5 Proposed Model

The proposed methodology addresses the persistent problem of data leakage by providing a conceivable attestation against the rogue operator and the unauthorized data released by the same. The modules included in the proposed model are listed below.

5.1 Data Allocation

This approach emphasizes on a secure and ingenious methodology for a distributor to distribute data to its stakeholders to identify the "guilty" workforce. With administrative credentials, a checked-in user can rescript and alter their files and upload it to the cloud. Admin sends the necessary info to the user and the authentication data is sent between agents and users through email.

5.2 Fake Object Module

To discover the "site of leakage," the data distribution mechanism will append redundant information to the credible data stream. Fake or redundant objects are objects that appear to be credible data being delivered, but are different and differentiable from real records. The data leaker believes he has the actual record

due to a false item. It will show the bogus record and deliver the message if any record is downloaded using counterfeit measures using the "wrong key." The term "wrong key" relates directly to the duplicate key known as Public Key, which is shown as the primary key. It safeguards the private key (the primary key).

5.3 *Optimization Module*

This is a mandatory module for the data distribution entity. The entity has binary objectives that it must achieve. These are:

1. To give User-requested data and to conduct the User-requested changes
2. To recognize information breach and discover the root of the incident

5.4 *Data Distribution Module*

A data distributor is tasked with transmitting confidential information to one or more of his entrusted stakeholders or workers. He must guarantee that none of this information gets disclosed. However, if it is released, it is rumored that admin will find out about the particular files which are being disclosed. Also, for picking out the guilty employee distributor, data must be analyzed and scrutinized, assuming that it has been disclosed by one or more of its own employees. The information that is transmitted might be in any form or magnitude.

The entities and flow of data for constituting the cloud computing environment, as well as the data flow paradigms, are depicted in the architectural Fig. 4.

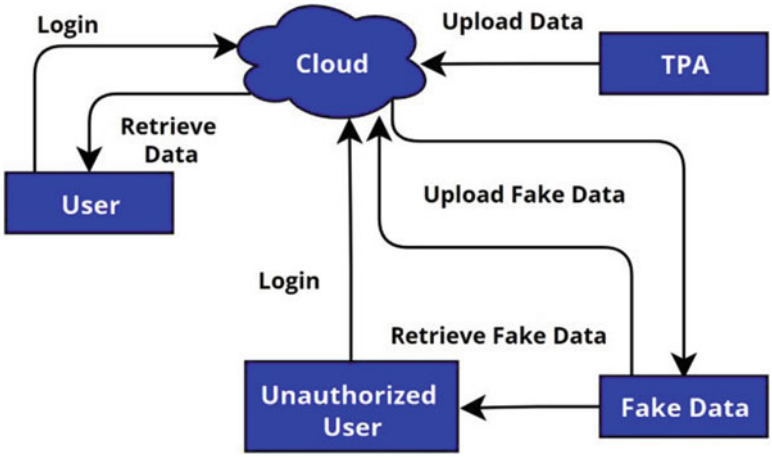

Fig. 4 Proposed architectural diagram of data leakage detection

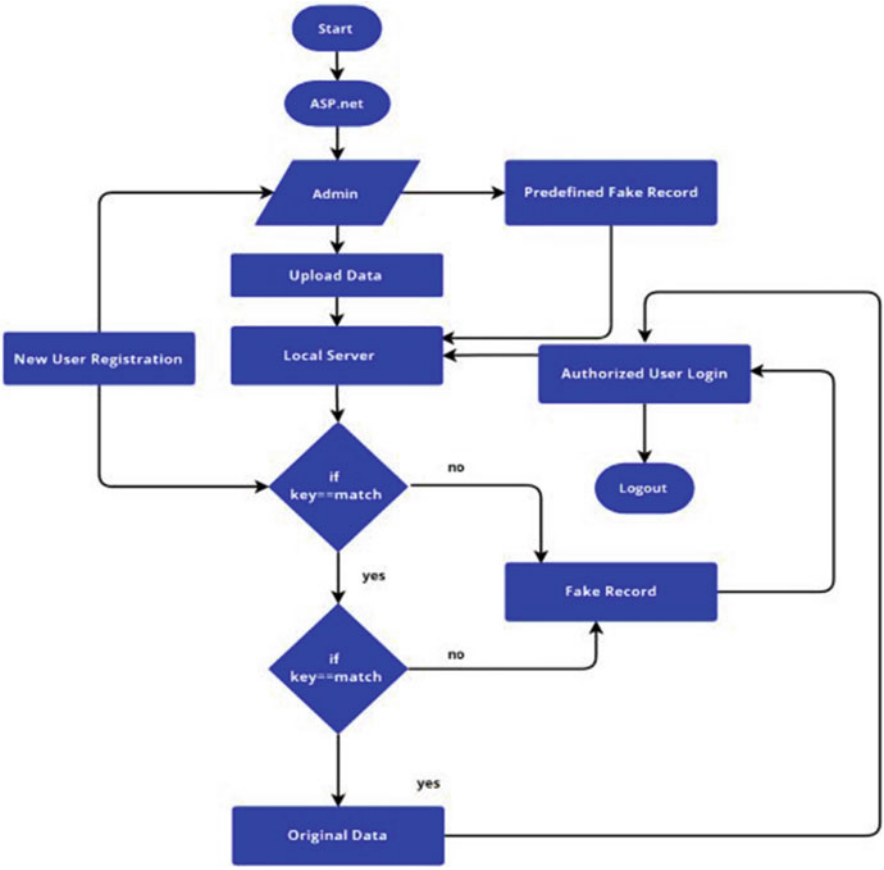

Fig. 5 Workflow of proposed diagram

The suggested model's control flow is depicted in the flow diagram in Fig. 5. The new customer creates an account and then uploads their information to the local server with administrative capabilities granted by admin. For safety reasons, the admin concurrently regulates the placement of pre-defined bogus information on the local server. Any information delivered to a receiver now includes the phony redundant information that was appended or concatenated to the authentic data using Wrong Key. If someone gains access to the transmission apparatus or channel and holds the correct key to decipher the transmission (the recipient), will be facilitated with access privileges to the original transmission. However, if an unauthorized access is detected, the information sent to the customer will be bogus owing to a key match error.

The proposed methodology can be accomplished by asymmetric encryption techniques. Asymmetric cryptography facilitates the encryption and decryption of information using two keys: a private key and an original key [3]. The public key,

on the other hand, is utilized to mislead unlawful access into assuming they have the original key. This approach is also faster in terms of computing. Information integrity is improved using hash functions and hashing algorithms.

6 Conclusion

The presented paper may be ended with the observation that the supplied model includes an approach to the concerns of information leakage and information transmission safety. The employment of data encrypting methods results in data being more reliably encoded, as well as the keys included inside it. Through the assistance of redundant objects functioning as outlandish marking over the authentic data, the culpable employee inside the company or any trustworthy entity of the corporation may be in the information communication network of the corporation. Also the corporation must investigate what information has been released by the responsible party. As we've seen, this suggested approach improves and extends information leakage tracking and protection, and even considers the potential of what information has been compromised. This methodology also solves the challenge of locating the data source and perpetrator of the leaks in the institution's information transmission network. Because hashing methods like SHA-2 are used in the suggested paradigm, the computational performance for encryption and decryption of information to be transferred or acquired is also faster. Though the presented model has solved the fundamental elements of the problem, more advanced and efficient algorithms and approaches can be formulated in the future. For example, in this methodology, we used the SHA-2 method, which is more efficient and safer than the SHA-1 algorithm, but there is also a variant called SHA-3 that may be employed in other methodological approaches.

References

1. R.G. Pearson et al., SPECIES: A spatial evaluation of climate impact on the envelope of species. Ecol. Model. **154**(3), 289–300 (2002)
2. M.S. Darms et al., Obstacle detection and tracking for the urban challenge. IEEE Trans. Intell. Transp. Syst. **10**(3), 475–485 (2009)
3. C. Tan et al., Understanding the nature of first-person videos: Characterization and classification using low-level features, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (2014), pp. 535–542
4. K.V. Wong, Research and development of drones for peace—High power high energy supply required. J. Energy Resour. Technol. **137**, 3 (2015)
5. H.U. Zaman et al., A novel design of line following robot with multifarious function ability, in *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, (IEEE, New Jersey, 2016), pp. 1–5
6. O. Shrit et al., A new approach to realize drone swarm using ad-hoc network, in *2017 16th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, (IEEE, 2017), pp. 1–5

7. M. Bennis, M. Debbah, H.V. Poor, Ultrareliable and low-latency wireless communication: Tail, risk, and scale, in *Proceedings of the IEEE* 106(10) (2018). Accessed on 16 Nov 2021, pp. 1834–1853
8. Rotor Riot LeDrib. *What is FPV Freestyle* (2018). https://rotorriot.com/pages/beginners-guide. Accessed on 16 Nov 2021
9. Liftoff Drone Simulator (2018) https://www.liftoff-game.com/liftoff-fpv-drone-racing. Accessed on 17 Nov 2021.
10. F. Naujoks et al., From partial and high automation to manual driving: Relationship between non-driving related tasks, drowsiness and take-over performance, in *Accident Analysis & Prevention*, vol. 121, (2018), Accessed on 16 Nov 2021), pp. 28–42
11. Get FPV Aaron Ziemann. *How to Find the Perfect Drone Racing Line* (2019). https://www.getfpv.com/learn/fpv-essentials/how-to-find- the-perfect-drone-racing-line/. Accessed on 16 Nov 2021
12. J. Delmerico et al., Are we ready for autonomous drone racing? The UZH-FPV drone rac- ing dataset, in *2019 International Conference on Robotics and Automation (ICRA)*, (IEEE, New Jersey, 2019), pp. 6713–6719
13. A. Loquercio et al., Deep drone racing: From simulation to reality with domain randomization. IEEE Trans. Robot. **36**(1), 1–14 (2019)