AN IMPROVED DATA LEAKAGE DETECTION SYSTEM IN A CLOUD COMPUTING ENVIRONMENT

BY

OKOCHI, PRISCA IJEOMA
B.Sc (MOUAU)
Reg. No: 20164023288

A THESIS SUBMITTED TO THE POST GRADUATE SCHOOL

FEDERAL UNIVERSITY OF TECHNOLOGY, OWERRI (FUTO)

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE (M.Sc) DEGREE IN COMPUTER SCIENCE

JANUARY, 2023.

CERTIFICATION

This is to certify that his work "AN IMPROVED DATA LEAKAGE DETECTION SYSTEM IN A CLOUD COMPUTING ENVIRONMENT" was carried out by Okochi Prisca Ijeoma (20164023288) in partial fulfilment for the award of the degree of Master of Science in Computer Science in the Department of Computer Science of the Federal University of Technology Owerri Imo state, Nigeria.

.....	27/01/2023
Dr. S. A. Okolie Supervisor 1	Date
.....	27/01/2023
Dr. (Mrs.) J. N. Odii Supervisor 2	Date
.....	27/01/2023
Dr. (Mrs.) J. N. Odii HOD, Computer Science	Date
.....	.....
Prof. (Mrs.) U. F. Eze Dean, SICT	Date
.....	.....
Prof. B. O. Esonu Dean, Postgraduate School	Date
.....	27/01/2023
Prof. Adebayo O. Adetunmbi External Examiner	Date

DEDICATION

This thesis is dedicated to almighty God for endowing me with his grace, wisdom, knowledge, understanding and strength to carry out this work and also to my beloved husband, kids, siblings and parents.

ACKNOWLEDGEMENTS

I give my profound gratitude to Almighty God our benefactor, who endows me with the precious gift of life, strength, wisdom and knowledge to successfully carry out this work. To him alone be praised and glorified forever and ever!

My supervisors, Dr S. A. Okolie and Dr J.N. Odii have not been only supervisors to me but, also a father and mother that guided me throughout the period of this thesis. I so much appreciate all their advice, insight criticisms, encouragement and patience to me throughout the course of writing this thesis.

My humble gratitude also goes to the Head of Department, Dr J.N. Odii, my former Head of Department Prof. Onyeka, Dr (Mrs) E. C. Nwokorie and all my wonderful lecturers; Prof. P. O. Asagba, Dr Uzor, Dr C.N. Njoku, Dr C.G. Onukwugha, etc. and whom God have also used as a source of inspiration to me in my entire period of M.Sc programme in the Department of computer science, FUTO.

I will forever appreciate you my lovely, caring and supportive husband; Mr Tochukwu Okochi, My kids; Sommy, Kaima, Zitel and Kossy thanks so much for cooperating with Mum. My parents; Sir & Lady Nestor Nnokwe, and my siblings. I thank you all for your patience, moral support you unceasingly offered to me throughout the period of my M.Sc program.

I must not forget to mention you my honourable colleagues and friends; Vitalis, Blessing, Charles, Emeka, Nnaemeka, Moses, Faith, Ozioma, JohnPaul, Ojukwu and others who are not mentioned here put, in one way or the other contributed to the success of this thesis and my entire M.Sc programme. **May God bless you all!**

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

An Improved Data Leakage Detection System is designed to mitigate the leakage of crucial and sensitive data in a cloud computing environment. Generally, leakage of data in computing system has caused a lot of irreparable damage or catastrophe to various institutions and organisations worldwide. Therefore, this research aims at detecting and preventing any intentional or non-intentional data leakages using dynamic password for data security. To achieve this the Object Oriented Analysis and Design Methodology (OOADM) was adopted. The new system was implemented using ASP.net MVC and Microsoft SQL Server Management Studio as the backend. And by incorporating an Audit trail/Transaction log mechanism, the new system monitors the activities within and outside the computing environment with date and time stamp. Hence, the system can be applied in any environment for the prevention and detection of any data leakage.

Keywords: Data Leakage, Audit Trail, Transaction Log, Watermarking, Perturbation.

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Data leakages have gained widespread attention as businesses of all sizes become increasingly reliant on digital data, cloud computing, and workforce mobility. With sensitive business data stored on local machines, on enterprise databases, and on cloud servers, breaching a company's data has become as simple or as complex as gaining access to restricted networks. Data leakages didn't begin when companies began storing their protected data digitally. In fact, data leakages have existed for as long as individuals and companies have maintained records and stored private information. Before computing became commonplace, a data breach could be something as simple as viewing an individual's medical file without authorization or finding sensitive documents that weren't properly disposed of. Still, publicly-disclosed data leakages increased in frequency in the 1980s, and in the 1990s and early 2000s, public awareness of the potential for data leakages began to rise. Most information on data leakages focuses on the time period from 2005 to today. This is largely due to the advancement of technology and proliferation of electronic data throughout the world, making data leakages a top concern for both enterprises and consumers. Today's data leakages can impact hundreds of thousands often millions of individual consumers, and even more individual records, all from a single attack on one company.

Data leakage happens daily when confidential business information like customer or patient data, ASCII text file or design specifications, tariffs, property and trade secrets, and forecasts and budgets in spreadsheets are leaked out. When these are leaked out it leaves the corporation unprotected and goes outside the jurisdiction of the corporation. This uncontrolled data leakage puts business during a vulnerable position. Once this data is not any longer within the domain, then the company is at serious risk. When cybercriminals live or sell this data for profit it costs our organization money, damages the competitive advantage, brand, and reputation and destroys customer trust. Data leakage which by classification is uncontrolled, unauthorized transmission of classified information from a data center or a computer system to the outside is causing companies or organizations huge sums of money and also breach of trust. Data leakage are often accomplished by simply mentally remembering what was seen by physical removal of tapes, disks and reports or by

subtle means like data hiding. Data leakages are an enormous challenge to especially organizations and institutions.

Though there are variety of systems and methods designed to make sure data is secured by using various different encryption algorithms there's an enormous issue of integrity of the users themselves. The data leakage industry is very heterogeneous as it evolved not of the product lines of leading IT security venders. The usage of enabling technologies like firewalls, encryption, access control, identity management, machine learning/context -based detectors Ajinkya et al (2013). have already been incorporated to supply protection against various facts of the data leakage threat. Classification of data Leakage: Data leakage can be classified into three types which means a document may contain confidential data that may be classified as the unintentional leak, intentional leak, and malicious leak. Unintentional Leak include activities such as: attach document, Zip and send and Copy & Paste. The unintentional leakage usually occurs when a user mistakenly sends a confidential data and information to wrong recipient, user or a third party agent. This happens without any personal intention. For instance, if an employee sends an email attaching document mistakenly this contains confidential data to a wrong person.

Intentional Leak: The intentional leakage usually occurs when the user tries to send a confidential document without conscious of company rules and policy and finally sends anyhow to anywhere. This is usually done when a user bypassing the security rules and regulations or devices without trying to gain personal benefits. For instance, when an employee renames a document folder and partially copies data from it the following activities occurs during intentional leak: Document renames, Document type change, Partial data copy, Remove keyword

Malicious Leak: Malicious leakage is usually happened when a user deliberately sneaks the confidential data or information past security policy. Malicious Leakage is done using the following activities: Character encoding**,** Print screen**,** Password protected**,** Self-extracted archive**,** Hide data**,** Policies or product. For instance, when an employee sneaks a confidential data from enterprise system and sends them through email and even cause vulnerability to the system.

In order to stop this problem different methods of data leakage prevention has been made like watermarking method. Watermarking means a technique/method in which a

singular code is embedded in each distributor's copy. It's basically encryption on a specific data which is to be distributed (Abhijeet and Abhineet 2017). Ajinkya et al (2013), listed variety of techniques which may be used to detect data leakages as follows; Perturbation technique

Perturbation is where data is modified and make less sensitive before being transfered to agents. But in some cases it's good to leave the data unaltered. Examples being an outsourcer doing payroll, he must have the precise salary and customer checking account numbers. And also Unobtrusive Techniques. Ajinkya et al (2013), developed a model for assessing the 'guilt' agents and also attempt to present an algorithm for distributing objects to agents, in a bit to improve the probabilities of identifying the leaker.

Monali et al, (2019), developed a system with three tier application that creates data access secure through a secure channel monitoring system. This leads to the development of this research work to incorporate an audit trail mechanism and dynamic code to detect and prevent data leakage of the organization's sensitive and crucial information.

1.2 Problem Statement

Data security is one of the main issues to be considered when the transmission is via wireless communication. The problems that necessitated this research are:

1. Most Organizations and individuals have shifted to cloud computing and the global concern is how their data will be protected.
2. Problems of confidentiality and integrity. Data leakage can be caused by internal and external parties, either intentionally or accidentally.

In 2018, InfoWatch Analytical Center registered 1,039 data leaks, which is 12% more than in 2017 (925 leaks). Among the logged data leaks, 651 (64.5%) are caused by internal offenders, while 358 (35.5%) of the cases are triggered by intruders from the outside (InfoWatch Analytical Center). Efforts in this work will be to mitigate or eliminate these problems.

1.3 Objectives of the Study

The main objective of the research is to develop an improved data leakage detection system in a cloud computing environment. The specific objectives are to:

i. develop an application that profiles user access activities into a computing resource.
ii. track user activities for the detection of possible data breaches using application service from (i) above .
iii. implement an email alert system for the owner of the data (the Administrator) in the event of data leakage detection.

1.4 Justification of Study

This study will be useful in the protection of data stored in organizational databases. Data/information transmitted via insecure media will be received without interference or loss of data or information of any kind. It also reduces theft to its barest minimum. The data leakage detection system will help individuals and organizations to protect their information from unauthorized users. Securing sensitive data from illegal access, theft and forging becomes a lesser challenge for different organizations, like security agencies, higher institutions and private sectors. This will make information stored or shared confidential. This study will be of significant importance to any establishment because it will forestall tampering of records but rather ensure its integrity. Security and privacy are very important in managing sensitive data and so, the stored data will be intelligible to only intended user thereby reducing the possibility of fraud (Yunchuan & et al, 2014).

1.5 Scope of the Study

This study is streamlined to data leakage detection and its security system as it relates to how user activities can be monitored, through system transaction logs, secret key request, and email alert system to determine data leakage channels for confidentiality, integrity and reliability of data in the database.

CHAPTER TWO

LITERATURE REVIEW

2.1 Conceptual Framework

Audit trails are the manual or electronic records that chronologically catalog events or procedures to produce support documentation and history that's been used to authenticate security and operational actions, or mitigate challenges (Andy, 2017). Numerous industries use versions of an audit trail to give a historical document of progression supported a sequence of activities. The sequence of records provide proof of compliance and operational integrity. Audit trails can also identify areas of non-compliance by providing information for audit investigations. Whether it's logging the planning changes of a product build, keeping the record of monetary transactions for an electronic-commerce site, communication transactions, healthcare activity, or legitimizing the result of an election, an audit trail validates actions and results. Audit trail records contain details including date, time, and user information related to the transaction.

IT plays a crucial role within industry or regulation specific audit logs and trails. However, the department itself features a unique and densely populated log process where the various and varied activities of users, systems, and applications are constantly monitored to stop misuse, hacking, or corruption of data. IT professionals use this technique for validation as an important tool to research operations and technical controls for computer systems. An audit trail provides a tool to take care of information and system integrity.

2.2 Types of Audit Trail Activities

An audit trail provides basic information to backtrack through the whole trail of events to its origin, usually the first creation of the record. This may include user activities, access to data, login attempts, administrator activities, or automated system activities. Audit records contain elements defined by the organization which include: What the event was, what user, system, or application launched the event (this information may include IP address and device type), and the date and time the event or activity happened.

Some audit trails look more closely at actions within certain applications to chronicle quite an easy system or application launch. These logs may pinpoint elements like specific changes to a database or data contained therein, and also detect improper web-browsing or electronic mail use. Various IT departments have one audit trail which will be system, application, or event-defined. Listing abnormal activities or use deemed "out of the ordinary" can initiate an investigation. An accurate and well-defined audit trail gives the evidence to sort out answers and solve challenges or issues.

Long term maintenance of audit logs could be difficult for several organizations because the logs can occupy extensive space for storing which will not be readily available. However, it can be possible to maintain the audit trail for the lifetime of the records. These are often extremely useful in historical reporting and future solutions to problem solving.

2.3 Types of Industries that Rely Audit Trails

The need to support compliance, security, and operations is found in most (if not all) organizations. The rule both mandate and regulate the utilization of electronic records make audit records a crucial element in defending against security breaches, supporting compliance reporting, and ultimately passing numerous sorts of internal and external audits. Industries that have provisions to trace information integrity include government agencies and universities who maintain sensitive, tip, and any company that uses electronic records containing tip. Each industry, whether tracking records or transactions, will enjoy maintaining accurate audit logs (Andy, 2017).

The organizations that make use of audit trails are: Financial, accounting, and billing records, Manufacturing design controls, Health information, Clinical research data, IT helpdesk records, Content management version control, University student records, E-commerce sales, Legal and research investigations, Nursing medical records, Ballot-keeping and voting records

To that infact, there's a growing need for industries also as government and academic agencies to take care of and supply accurate and auditable information. If a corporation is utilizing a management system to manage records, likelihood is that high that they record audit trails.

2.4 Benefits of an Audit Trail

The capability to follow records back to their origin gives numerous advantages, which consist of transparency, integrity and accuracy, system protection from misuse or damage, and security of sensitive or vital information. These are achieved through these four areas according to (Andy, 2017):

User Accountability: A user is one who has access to the system. Implementation of audit trails promotes appropriate user behavior/attitude, which prevents the introduction of viruses, improper use of data, and unauthorized use or modifications. In addition, the user knows that their actions are automatically recorded against their unique identity.

Reconstruction of Events: When an investigation is warranted or triggered, the primary step to remediate a drag is knowing the "when," the "how," and therefore the "what" of the event. Visibility into this information can help in problem detection and stop future occurrences of things like hacking, system failures, outages, or corruption of data.

Detection of Intrusion: Audit trails help in identifying suspicious behavior or actions. Unauthorized access can be a significant issue for many systems. Many regulations now have mandates for the safety of data and maintaining confidentiality. Protection also extends to property, designs, personnel information, and financial records.

2.5 Dynamic Password

Dynamic password may be a password that's valid for less than one login session or transaction, on a computing system or other digital device (Gun, 2011). Dynamic passwords avoid several shortcomings that are related to traditional (static) password-based authentication; variety of implementations also incorporate two-factor authentication by ensuring that the one-time password requires access to something an individual has (such as a little keying job device with the Dynamic passwords calculator built into it, a smartcard or specific cell phone) also as something an individual knows (such as a PIN). Dynamic passwords generation algorithms typically makes use of pseudo randomness or randomness, making a prediction of successor dynamic passwords by an attacker difficult, and also cryptographic hash functions, which may be wont to derive a worth but are hard to reverse and thus difficult for an attacker to get the info that was used for the hash. This is

often necessary because otherwise, it might be easy to predict future dynamic passwords by observing previous ones.

Dynamic passwords are discussed as a possible replacement for, also as enhancer to, traditional passwords. On the downside, Dynamic passwords are often intercepted or rerouted, and hard tokens can stray, damaged, or stolen. Many systems that use Dynamic passwords do not securely implement them, and attackers can still learn the password through phishing attacks to impersonate the authorized user (Gun, 2011).

2.6 Theoretical Framework

Data leakage may be defined as the accidental or unintentional distribution of personal or sensitive data to unauthorized entity (Sandip and Kulkarni, 2012). Sensitive and crucial data of companies and organizations includes intellectual property (IP), financial data, patient information, personal credit-card data, and other information counting on the business and therefore the industry. In some cases, sensitive data is shared among various stakeholders like employees performing from outside the organizational premises (e.g., on laptops), business partners and customers (Sandip and Kulkarni, 2012). This may increase the danger of data falling into unauthorized entity. Whether it is caused by malicious intent, or an inadvertent mistake, by an insider or outsider, exposed sensitive or crucial information can seriously hurt a corporation. The potential damage and adverse consequences of data/information leak incident are often classified into the subsequent two categories: direct and indirect loss. Direct loss means tangible damage that's easy to live and estimate quantitatively. Indirect loss, on the other hand, is far harder to quantify and features a much broader impact in terms of cost, place and time (Bunker, 2009). Direct loss can include violating regulations (such as those protecting customer privacy) leading to fine/settlement/customer compensation fees; litigation of lawsuits; loss of future sales; costs of investigation and remedial/restoration fees. Indirect loss includes reduced share-price as a results of the negative publicity; damage to company's goodwill and reputation; customer abandonment; and exposure of property (business plans, code, financial reports, and meeting agenda) to competitors.

An enterprise data leak may be a scary proposition. Data leakage issues that arise from email and other internet channels have always been affected by data security practitioners. It is

easier for data loss to occur whether accidentally or maliciously with the utilization of mobile technology.

The literature reviews on how cloud computing evolved and the architecture of cloud computing due to the motive of detecting data leakages in cloud computing environment.

2.6.1 Data leakage

Organizations not only battle to be free from viruses, intrusion and spam, but are now struggling to be free from leakage of data and information both for intentional or accidental exposure of crucial data from legally protected personal information to property and trade secrets (Miller, 2009). Lately, the exposure of data has become one threat or the other to several enterprises.

Davis (2009), explained data leakage as when information is transferred illegally to the outside world.

Data leakage are sometimes achieved by simply remembering what was seen, by physical removal of disks, tapes, and reports or by data hiding.

2.6.2 Detection of data leak

Most times a data distributor gives out crucial data to at least one or more third parties. Sometime later, a number of data/information is found in an unauthorized place (e.g., on internet or on a user's laptop). The data distributor must have to investigate the source of the leak. Data leakage is stated because the unauthorized transfer of classified information from a computer or datacenter to the surface world. Data leakage are often accomplished by simply mentally remembering what was seen, by physical removal of tapes, disks and reports or by subtle means such as data hiding.

2.6.3 Data Transformation

The exposed data within the content could also be unpredictably transformed or modified by users, and it may not be similar to the initial sensitive data, e.g., insertions of metadata or formatting tags, substitutions of characters, and data truncation (partial data leak). Thus, the detection algorithm has to recognize numerous form of sensitive data variations.

2.6.4 Scalability

The heavy workload data/information leak screening is because of two reasons. Long Sensitive Data Patterns: The sensitive data as an example customer information, documents, ASCII text file may be of arbitrary length great amount of Content: The detection has to rapidly screen content instance gigabytes to terabytes. Traffic scanning is time sensitive than storage scanning, because the leak has to be discovered before the message is transmitted. Directly applying automata-based string matching to data leak detection is inappropriate and inefficient, because automata aren't designed to support unpredictable and arbitrary pattern variations. In data leak detection scenarios, the transformation of leaked data isn't known to the detection method. Creating comprehensive automata models covering all possible variations of a pattern is infeasible, which results in O(2n) space complexity or O(2n) time complexity where n is the number of automaton states. Therefore, automata approaches can't be used for detecting long and transformed data leaks. The proposed work is predicated on two algorithms i.e., RTU and DTU.

2.6.5 How data leakage occurs?

In the course of business, data must be handed over to trusted third parties for a few operations. Sometimes these trusted third parties may act as points of knowledge leakage. As an example, a hospital may give patient records to researchers who will invent new treatments. Similarly, a corporation may have partnerships with other companies that need to share customer data among them.

Network security consists of the policies and practices adopted to stop and monitor unauthorized access, misuse, modification, or denial of a network and network-accessible resources. It consist the authorization of access to data during a network, which is controlled by the network admin. Users select or are assigned an ID and password or other authenticating information that permits them access to information and programs within their authority. Network security covers a spread of computer networks, both public and personal, that are utilized in everyday jobs; conducting transactions and communications among businesses, government agencies and individuals. Networks are often private, like within a corporation, et al. which could be hospitable public access. Network security is involved in organizations, enterprises, and other sorts of institutions. It does as its title explains: It secures

the network, also as protecting and overseeing operations being done. The foremost common and straightforward way of protecting a network resource is by assigning it a singular name and a corresponding password.

Traditionally, leakage detection is handled by watermarking, e.g., a singular code is embedded in each distributed copy. If that replicate is later discovered within the hands of an unauthorized party, the leaker are often identified. Watermarks are often very useful in some cases, but again, involve some modification of the first data. Furthermore, watermarks can sometimes be destroyed if the info recipient is malicious. E.g. A hospital may give patient records to researchers who will devise new treatments. Similarly, a corporation may have partnerships with other companies that need sharing customer data. Another enterprise may outsource its processing, so data must tend to vary other companies. We call the owner of the info the distributor and therefore the supposedly trusted third parties the agents.

2.6.6. Reason of data losses

Nowadays, Data are leaked by some ways. There are a number of common reason of information/data losses (Periyasamy and Thenmozhi, 2017):

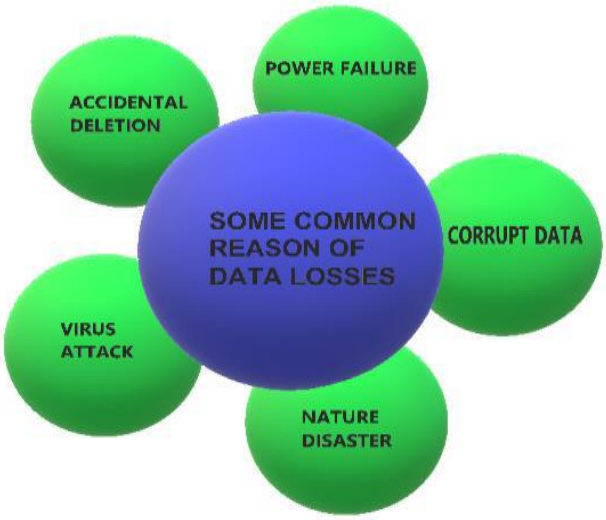

Fig. 2.1: Reason of Data Losses (Source: Periyasamy and Thenmozhi, 2017)

Corrupt data: If the database is corrupted, then chances arise to lose the data. Somehow it's possible to recover the lost data from a corrupt file with the proper applications (Periyasamy and Thenmozhi, 2017).

Power failure: If you're performing on any electrical devices like laptops, PC, etc. Suddenly power off while working then have the utmost chance to lose their confidential data. So, avoid this problem by saving your work regularly.

Natural disaster: Suppose your file database is damage by some natural disasters like fire, flood, and another unpredictable disaster. We will resolve this problem by storing the information in another place.

Accidental deletion: Sometimes the information is deleted accidentally from your storage drive. That data isn't be notice by the user for an extended time.

Virus attack: Sometimes the machine is deeply infected by some hidden virus, therefore the data base attack by that virus which results in corrupt the information.

2.6.7 Challenges of data Leakage Detection and Prevention

With the recognition of data leakage prevention/detection, there are some major challenges also occur. Generally, these varieties of problem arise within the data storing and therefore the data transaction from one node to a different node.

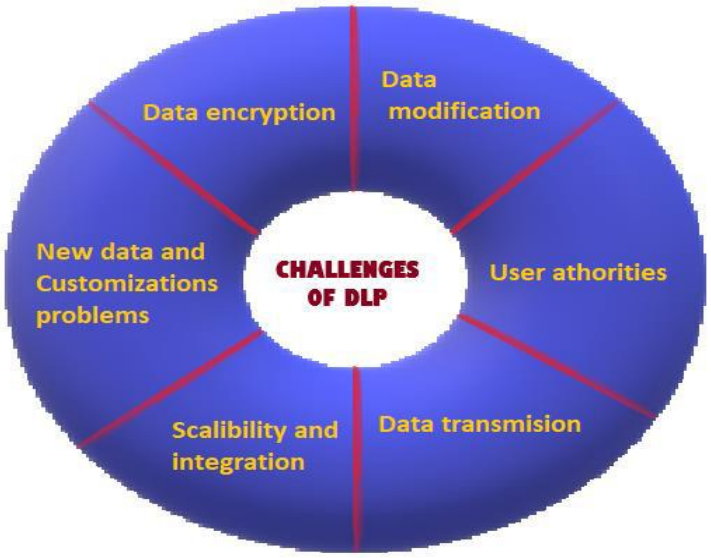

Fig.2.2: Challenges of Data Leakage Detection and Prevention (Source: Periyasamy and Thenmozhi, 2017)

Data encryption: With the encryption technique, the info are often shielded from a malicious user. By the encryption technique, we will achieve the integrity, confidentiality, and authenticity. But the strong encryption technique is employed then it's difficult to research the info file.

Data modification: Sometimes the data is modified then are often automatically partially leaked to the users. It is a big challenge for the administrator to secure their data at the time of modification.

User authorities: In times, there's maximum data which is storing within the cloud but only the authority user can access their data, if there are not any limitations on the users then any user can access all the info in order that there are challenges that face while storing data within the cloud. The administrator gives the precise rights to the precise user to access their file to extend the safety level.

Data transmission: Data are often transferred differently, if the info is transmitted over a fanatical or specific channel then it gives the confidentiality to the info but if channels

are aside from specific ones like USB, email and another format then it makes it difficult to secure their data.

New Data and customization: Every user may have differing types of knowledge formats. Sometimes which isn't suitable for a store within the cloud in order that it's difficult to manage by the administrator to store their data. Whenever new user information is stored within the cloud User must provide all the knowledge which is said to the documents.

Scalability and integration: These days, the main company has a vast amount of knowledge which they typically like better to store their data into the cloud. this is often a secure thanks to preserve their data but even have their disadvantages during which data size is large in order that monitoring, matching and accessing their data becomes harder .

2.6.8 Benefits of data Leakage Detection

DLP innovation enables ensure to user and control of all kinds of data, including client information, monetary information, and guarded innovation. With the assistance of data leakage prevention, we protect the information from unauthorized users

.

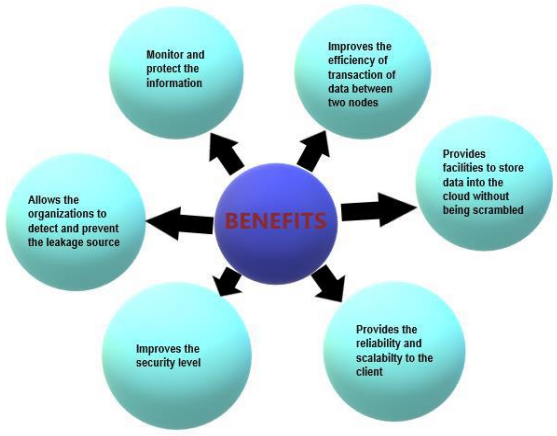

Fig. 2.3: Benefits of data Leakage Detection (Source: Periyasamy and Thenmozhi, 2017)

Data Leakage Prevention perceives classified information, guarantees that it doesn't advance into the cloud without being scrambled, and is simply sent to approve cloud applications. DLP offers total information to perceive ability and control, guaranteeing that representatives,

outsider sellers, contractual workers, and accomplices are kept from releasing your information purposefully or incidentally.

DLP allows the organization to detect the info leakage source and reduce the leakage source by blocking it to avoid future risk.

Data leakage methods can improve the safety principles by differing types of methods like look based data, guilt probability and faux objects.

DLP methods also help to enhance the value of the transaction between the one node to a different node and therefore the method improves the efficiency to offer the information from distributor to the client. Within the model used the minimum overlapping process of data file between the users.

DLP method gives reliability and scalability to the client. The client can easily store their data on the cloud and protect it from the malicious user.

2.6.9 Cloud computing

Cloud computing is one among the foremost important emerging and promising field in Information Technology. It provides services to varied organization over an online with the power to proportion or scale down their service requirements. Cloud computing refers to the delivery of on-demand computing services from applications to storage and processing power typically over the web and on a pay-as-you-go basis.

There are five key properties of cloud computing:

1. Cloud Computing is Flexible.
2. Cloud computing is Innovation.
3. Cloud computing is Accessibility.
4. Cloud computing is Savings.
5. Cloud computing is Efficiency.

Simply put, cloud computing means the delivery of computing services—including servers, storage, databases, networking, software, analytics, and intelligence—over the web (“the cloud”) to supply faster innovation, flexible resources, and economies of scale. you

sometimes pay just for cloud services you employ , helping you lower your operating costs, run your infrastructure more efficiently, and scale as your business needs change.

2.2.10 Evolution of cloud computing

Stedum (2013), cloud computing concept dates back to the 1950's when mainframes were made available to institutions. Server Rooms were created with the mainframe to cater for multiple users.so thanks to high cost of obtaining and maintaining mainframes, it became the practice of sharing computer resources. The introduction of Virtual Machines (VMs) by IBM within the 1970s also took communication and computing to a better level and therefore the Telecommunication Companies also allowed users shared access to an equivalent physical infrastructure and also to shift traffic to permit for better network balance and more control over bandwidth usage within the 1990s. By installing and configuring simple software called hypervisor across multiple physical nodes, a system will appear as if it's one physical node. thanks to this Technologist started using the term "Utility Computing" and "Cloud Computing".

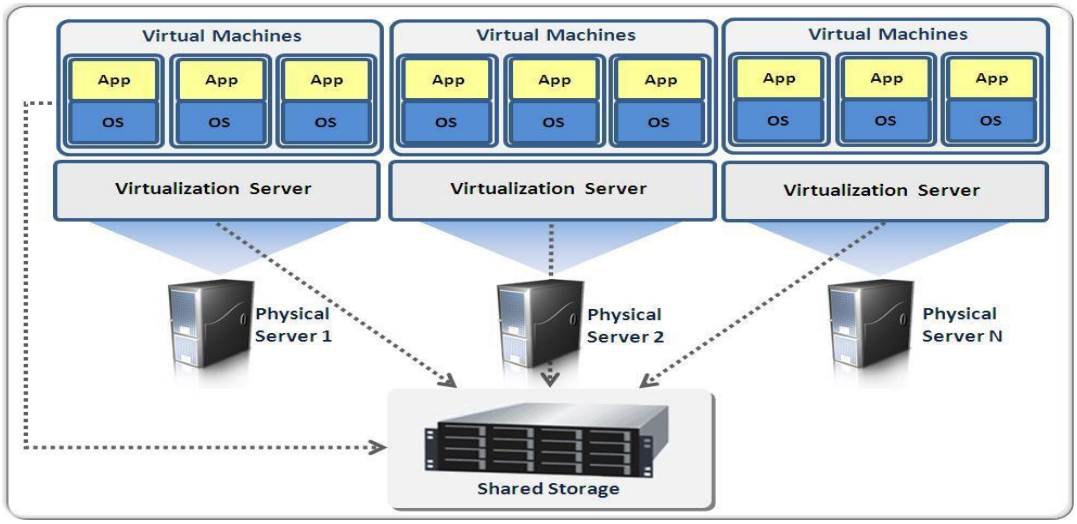

Fig.2.4: Virtualization Architecture of a Cloud. (Source: www.itworksite.com/virtualization, Karikari, 2015)

As technologies and hypervisors got better at reliably delivering and sharing resources, many enterprising companies decided to start out carving up the larger environment to form the cloud's benefits to users who don't happen to possess an abundance of physical servers

available to make their own cloud computing infrastructure. Those users could order "cloud computing instances" (also referred to as "cloud servers") by ordering the resources they have from the larger pool of obtainable cloud resources, and since the servers are already online, the method of 16 "powering up" a replacement instance or server is nearly instantaneous. Because little overhead is involved for the owner of the cloud computing environment when a replacement instance is ordered or cancelled (since it's all handled by the cloud's software), management of the environment is far easier. Figure 2.1 depicts virtualization architecture of cloud.

2.6.11 Top benefits of cloud computing

Cloud computing may be a big shift from the normal way businesses believe IT resources. Here are seven common reasons organizations are turning to cloud computing services:

1. Cost: Cloud computing eliminates the capital expense of shopping for hardware and software and fixing and running on-site data centers the racks of servers, the round-the-clock electricity for power and cooling, and therefore the IT experts for managing the infrastructure. It adds up fast.
2. Speed: Most cloud computing services are provided self service and on demand, so even vast amounts of computing resources are often provisioned in minutes, typically with just a couple of mouse clicks, giving businesses tons of flexibility and taking the pressure off capacity planning.
3. Global scale: the advantages of cloud computing services include the power to scale elastically. In cloud speak, meaning delivering the proper amount of IT resources for instance , more or less computing power, storage, bandwidth right when they’re needed, and from the proper geographic location.
4. Productivity: On-site data centers typically require tons of “racking and stacking” hardware setup, software patching, and other time-consuming IT management chores. Cloud computing removes the need for many of these tasks, so IT teams can spend time on achieving more important business goals.
5. Performance: The biggest cloud computing services run on a worldwide network of secure datacenters, which are regularly upgraded to the latest generation of fast and efficient

computing hardware. This offers several benefits over a single corporate datacenter, including reduced network latency for applications and greater economies of scale.

6. Reliability: Cloud computing makes data backup, disaster recovery, and business continuity easier and less expensive because data can be mirrored at multiple redundant sites on the cloud provider's network.
7. Security: Many cloud providers offer a broad set of policies, technologies, and controls that strengthen your security posture overall, helping protect your data, apps, and infrastructure from potential threats.

2.6.12 Types of cloud computing

Not all clouds are similar and not one kind of cloud computing is right for everybody. Several different models, types, and services have evolved to assist offer the proper solution for your needs.

First, you would like to see the kind of cloud deployment, or cloud computing architecture, that your cloud services are going to be implemented on. There are three alternative ways to deploy cloud services: on a public cloud, private cloud, or hybrid cloud.

Public cloud: Public clouds are owned and operated by a third-party cloud service providers, which deliver their computing resources, like servers and storage, over the web. Microsoft Azure is an example of a public cloud. With a public cloud, all hardware, software, and other supporting infrastructure is owned and managed by the cloud provider. You access these services and manage your account employing a browser, (Mell and Grance, 2011).

Private cloud: A private cloud refers to cloud computing resources used exclusively by one business or organization. a private cloud are often physically located on the company's on-site datacenter. Some companies also pay third-party service providers to host their private cloud. A personal cloud is one during which the services and infrastructure are maintained on a personal network.

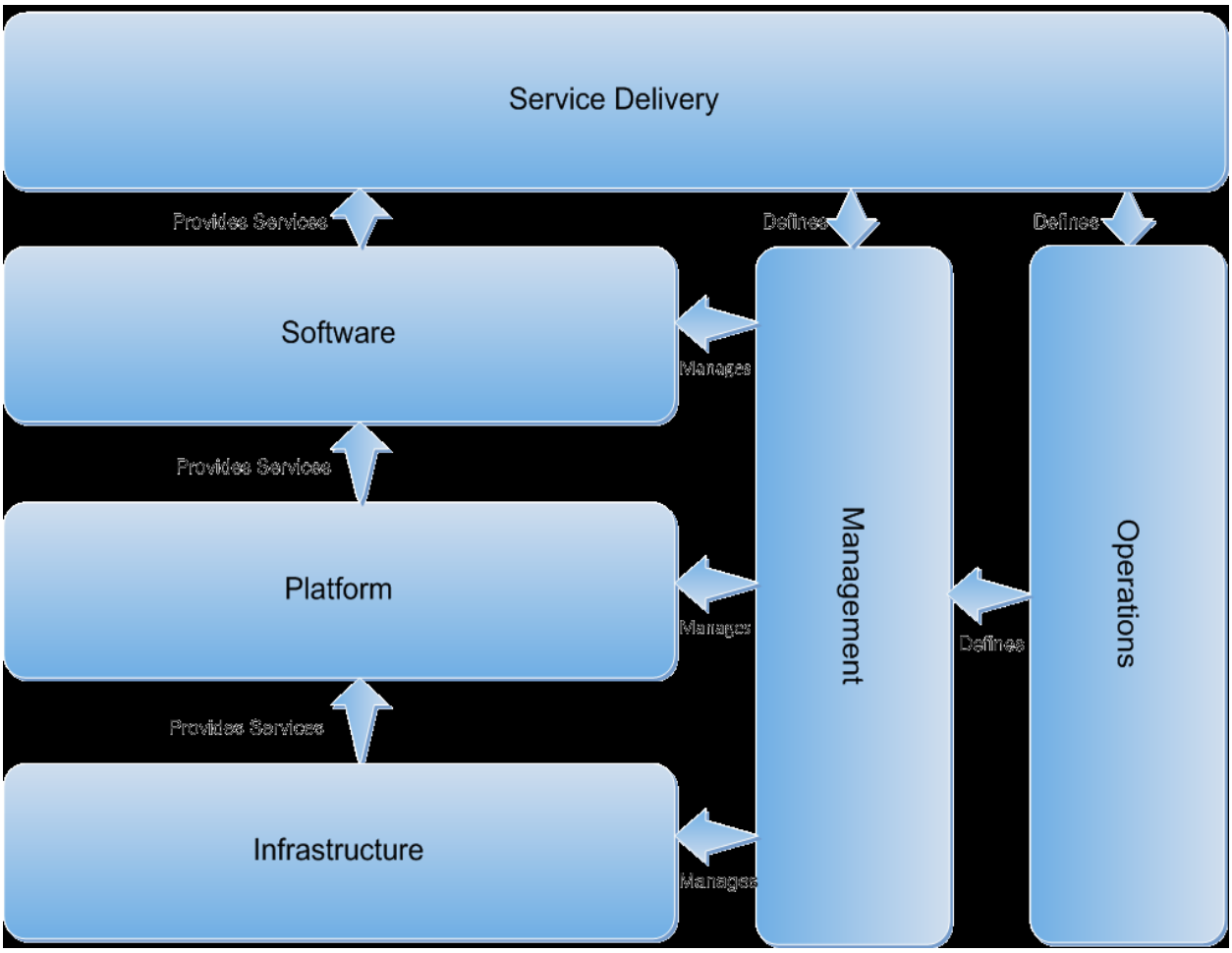

Fig. 2.5: Illustration of private cloud Architecture (Source: Karikari, 2015)

Hybrid cloud: Hybrid clouds combine public and personal clouds, bound together by technology that permits data and applications to be shared between them. By allowing data and applications to maneuver between private and public clouds, a hybrid cloud gives your business greater flexibility, more deployment options, and helps optimize your existing infrastructure, security, and compliance.

Shaw (2013), hybrid cloud is that the better of breed because it combines the comfort level of personal cloud with the pliability and flexibility of the general public cloud. Hybrid platforms use either public clouds or off-site Hosted Virtual Private clouds for a few applications and processes.

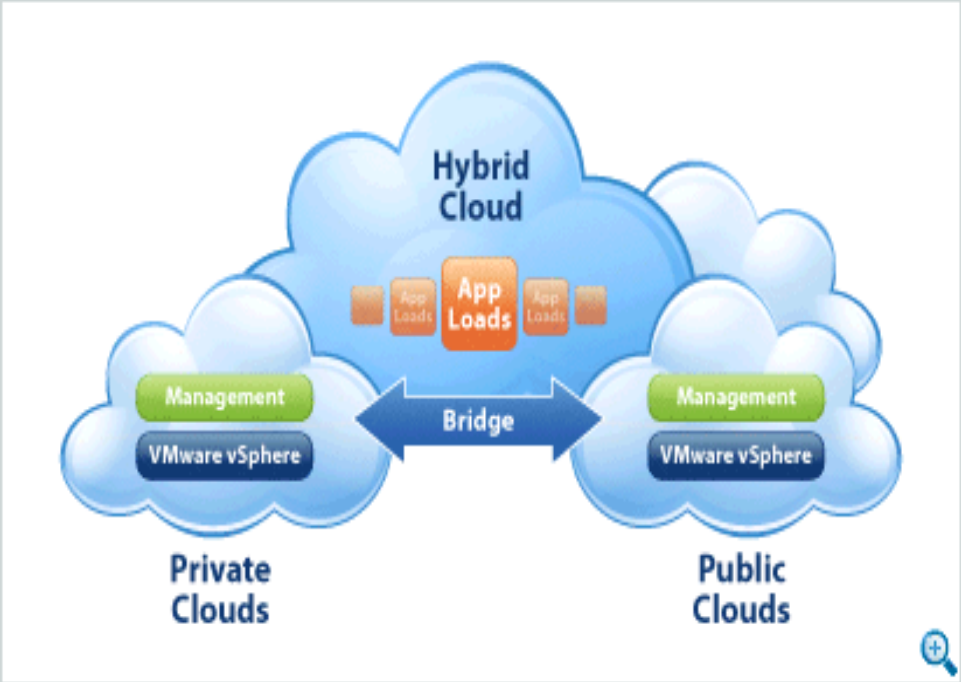

Fig. 2.6: VMWare's Hybrid Cloud Architecture (Source: Karikari, 2015)

Hybrid cloud may be a combination of a minimum of one private cloud and a minimum of one public cloud of which the private are often on-premises or virtual private cloud located outside the enterprises' data center. Figure 2.6 shows how hybrid cloud architecture seems like, (Claybrook, 2010).

2.6.13 Types of cloud services: IaaS, PaaS, Serverless, and SaaS

Most cloud computing services fall under four broad categories: infrastructure as a service (IaaS), platform as a service (PaaS), serverless, and software as a service (SaaS). These are sometimes called the cloud computing "stack" because they repose on top of 1 another. Knowing what they're and the way they're different makes it easier to accomplish your business goals.

Infrastructure as a service (IaaS): The most basic category of cloud computing services. With IaaS, you rent IT infrastructure servers and virtual machines (VMs), storage, networks, operating systems from a cloud provider on a pay-as-you-go basis.

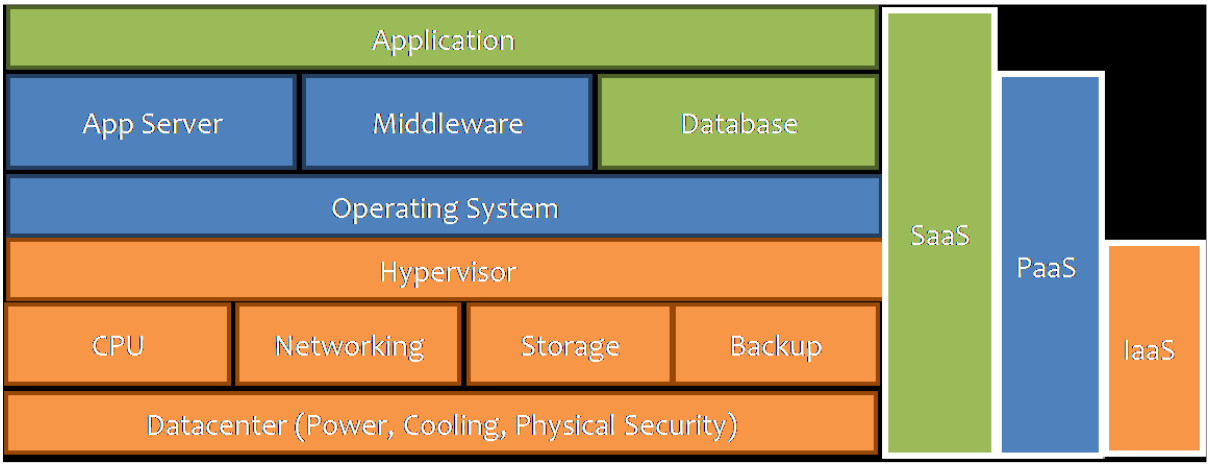

Fig. 2.7: Infrastructure-as-a-Service (IaaS) (Source: Karikari, 2015)

Figure 2.7 describes among other services how the stack of Infrastructure –as-a Service (IaaS) seems like . The physical resources are abstracted by virtualization, which suggests they will then be shared by several OS and user environment on the virtual resources- ideally, with none mutual interference. These virtualized resources usually comprise CPU and RAM, data storage resources (elastic block store and databases) and network resources.

Platform as a service (PaaS)

Platform as a service refers to cloud computing services that provide an on-demand environment for developing, testing, delivering, and managing software applications. PaaS is meant to be easier for developers to quickly create web or mobile apps, without fear about fixing or managing the underlying infrastructure of servers, storage, network, and databases needed for development.

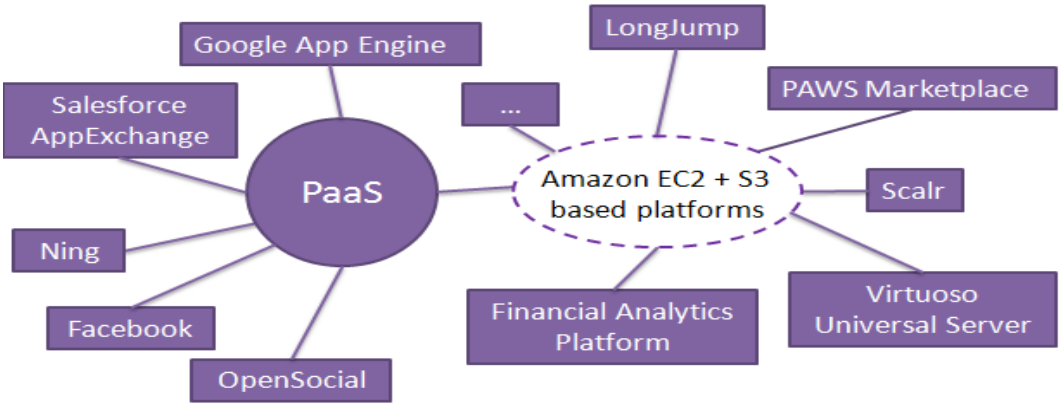

Fig. 2.8: Platform-as-a-Service (PaaS) (Source: Karikari, 2015)

Figure 2.8 describes Platform as-a-Service (PaaS) which comprises the environment for developing and provisioning cloud applications. The principal users of this layer are developers seeking to develop and run a cloud application for a specific platform.

Serverless computing: Overlapping with PaaS, serverless computing focuses on building app functionality without spending time continually managing the servers and infrastructure required to try to to so. The cloud provider handles the setup, capacity planning, and server management for you. Serverless architectures are highly scalable and event-driven, only using resources when a selected function or trigger occurs.

Software as a service (SaaS): Software as a service may be a method for delivering software applications over the web, on demand and typically on a subscription basis. With SaaS, cloud providers host and manage the software application and underlying infrastructure, and handle any maintenance, like software upgrades and security patching. Users hook up with the appliance over the web, usually with an internet browser on their phone, tablet, or PC.

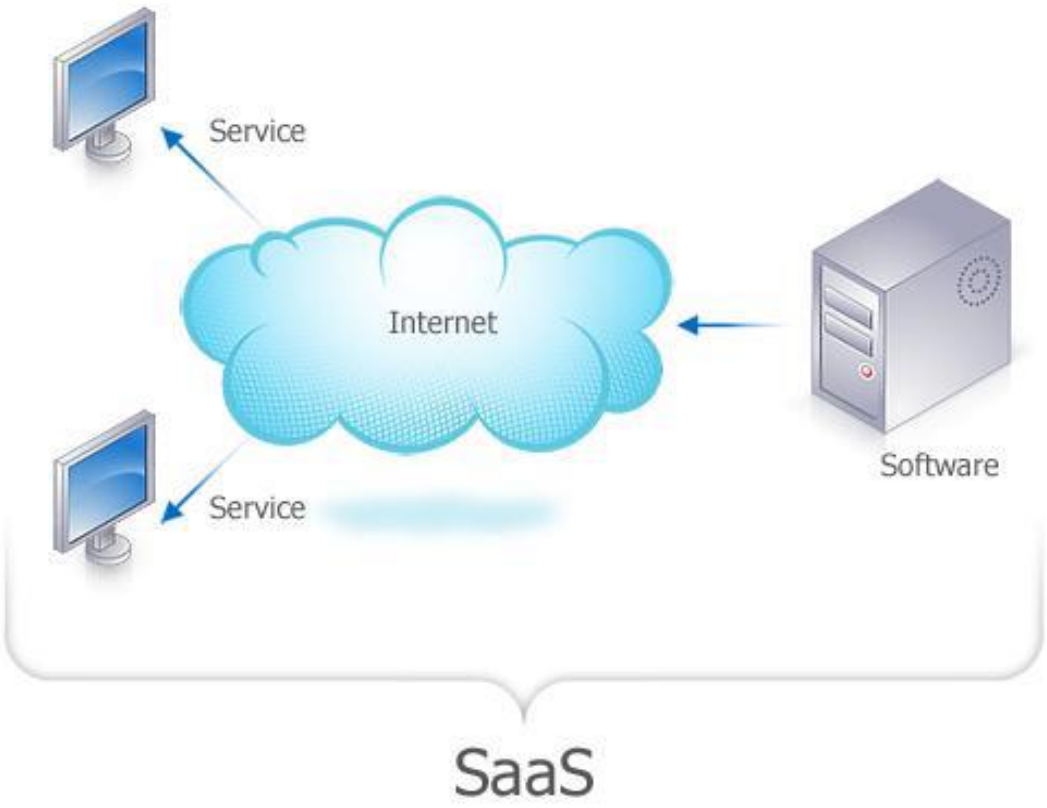

Fig. 2.9: Software-as-a-Service (SaaS) (Source: Karikari, 2015)

2.6.14 Uses of cloud computing

If you employ a web service to send email, edit documents, watch movies or TV, hear music, play games, or store pictures and other files, it's likely that cloud computing is making it all possible behind the scenes. the primary cloud computing services are barely a decade old, but already a spread of organizations from tiny startups to global corporations, government agencies to non-profits are embracing the technology for all kinds of reasons.

Here are couple of samples of what's possible today with cloud services from a cloud provider:

Create cloud-native applications: Quickly build, deploy, and scale applications web, mobile, and API. Cash in of cloud-native technologies and approaches, like containers, Kubernetes, micro-services architecture, API-driven communication, and DevOps.

Test and build applications: Reduce application development cost and time by using cloud infrastructures which will easily be scaled up or down.

Store, back up, and recover data: Protect your data more cost-efficiently and at massive scale by transferring your data over the web to an offsite cloud storage system that's accessible from any location and any device.

Analyze data: Unify your data across teams, divisions, and locations within the cloud. Then use cloud services, like machine learning and AI, to uncover insights for more informed decisions.

Stream audio and video: Connect together with your audience anywhere, anytime, on any device with high-definition video and audio with global distribution.

Embed intelligence: Use intelligent models to assist engage customers and supply valuable insights from the info captured.

Deliver software on demand: Also referred to as software as a service (SaaS), on-demand software allows you to offer the newest software versions and updates around to customers anytime they have , anywhere they're.

Abhijeet and Abhineet (2017), the cloud technology is completely hooked in to the web where the info is stored in its data centers of the service providers. The infomation security is one among the main challenges within the cloud computing technology. Less control over data may cause some serious security issues and threats which can cause data leakage, data insecurity and attacks on data by an insider or an outsider. so as to save lots of data from being leaked every IT company must specialize in security problems with securing their data from different third parties. Sometimes the leakage is completed by an insider mainly existing employees of the corporate, therefore the security must be beyond their employee's knowledge in order that they might not have clue to crack it. There's no specific time of knowledge leakage it could happen at any time. The quantity of injury done by a data leakage only depends on the standard of sensitive data leaked by the person. If the info which is leaked is extremely much important to the institution. It is going to leave the institution during a helpless state. The leakage could lower the business and should end in the downfall of the corporation.

In order to stop this problem different methods of knowledge leakage prevention has been made like Watermark method.

Watermark Method: Watermark may be a technique to stop the copyright of the owner of the data. It's a way during which a singular code is embedded in each distributor's copy. It's basically encryption on a specific data which is to be distributed.

The data are often within the sort of image, video or any important file. The watermark helps the corporate to say the ownership of particular information. During this technique, little pattern is added within the data mainly the tuples and therefore the subset of data. The attributes of the tuple and subset are algorithmically coded in order that they're controlled by a personal key to be accessed only by the owner of the data. This pattern represents the watermark. The info are often only accessed only if an individual has the key. For detection of the watermark, access to the first data isn't required. The watermark are often detected even during a small subset of knowledge as long because the data contains a number of the marks in it. The watermarking is completed through a software which embeds watermarks by watermarking algorithms .the software introduces a couple of errors within the data. These errors are referred to as marks and every one these marks together makes up a watermark.

Abhijeet and Abhineet (2017) discussed different Techniques of Watermark they include:

1. Watermarking by DCT (Discrete Cosine Transforms): Discrete cosine transform (DCT) may be a method for converting a sign into elementary frequency components. During this method, the image is first converted into 8x8 blocks of pixels. After DCT conversion, the mid-frequency range are selected which is predicated on Gaussian network classifier. Now the mid-frequency DCT coefficients are used for embedding. The DCT coefficients are modified by a linear DCT constraints. This may don't affect the visibility of the image and therefore the watermark won't be removed by compression.
2. Watermarking by DWT (Discrete wavelet transform): This is often a contemporary method which is widely used for watermarking, image compressions etc. this system uses wavelet filters to rework the image. Wavelets are small waves of varying frequency and limited duration. The wavelet transform decomposes the image

into three spatial directions horizontal, vertical and diagonal. the essential idea of DWT is to multi-differentiate decompose of the image into sub-image of the various spatial domain and independent frequencies.

3. LSB (Least Significant bit): In this method, the watermark is embedded within the LSB of pixels. This method is straightforward to implement but it's not very secure against attacks i.e. the watermark are often destroyed easily. The watermarking is completed by choosing a subset of image pixels then substituting the LSB of every of the chosen pixels with watermark bits. -Watermarking by Embedding and Extraction during this method, the insignificant a part of "> a part of the fractional part of the pixel intensity value of the most cover image is encoded to supply watermark. The watermark within the insignificant maintains the accuracy of the image. During this method the watermark is imperceptible. An outsized amount of watermark are often easily embedded and extracted using this method which can help companies and firms involved in digital information security products. It's another advantage of this method. Various algorithms for embedding and extraction are utilized in this system.
4. Wavelet Based Watermarking: In this method, the multi-resolution data fusion is embedded where the image and watermark are both used and transformed into discrete wavelet form. The watermark is embedded into each wavelet level. The typical of the estimates from each resolution level of wavelet decomposition is taken to detect the watermark. This algorithm works for JPEG compression, additive noise and filtering operations.
5. Secure Spread Spectrum Watermarking: In order to save lots of multimedia data like audio, video and image, a watermark must be added in significant components of a sign if it's to be robust to common signal distortions and malicious attacks. But, the modifications of those components may cause degradation of the info signal. So, watermark must be added to the spectral components of the info using techniques almost like spread spectrum communications, hiding a narrow band of the signal during a wide band signal. This watermark is very difficult for an outsider to remove, even many outsiders collude with different copies of the watermarked data.
6. Robust Watermarking Technique: This watermarking technique can't only survive general operations like compressions, adding noise, filtering then forth but also

geometric attacks like rotation, scaling translation etc. it's often used for ownership protection. This method is employed to encode ownership information directly into the data, so whenever the rightful ownership is at issue, information/data are often extracted and be used to find the rightful owner. The watermark should be stable during extraction and must not degrade information/data.

• Watermarking of Digital Audio and Image using Mat lab: A watermark is encrypted using RSA algorithm and is embedded within the audio file using LSB technique. LSB is an old method which isn't very vigorous against attacks. In this, the primary watermark is encrypted then embedded within the audio file, because of which removal of watermark becomes very difficult. This provides very high robustness. During retrieval of data/information, the embedded watermark is first retrieved then decrypted. Similarly, for image watermarking, DWT technique is employed. The watermark is embedded as a pseudo-noise sequence. This method makes image and audio very secure because the original watermark, must be known in order that embedded watermark are often far away from the watermarked image or audio.

7. Invisible Watermarking: This method provides an invisible robust watermarking scheme for embedding and extraction of a digital watermark in a picture. One among the most features of this algorithm is that during this method a sub image used for watermarking. The watermark is embedded within the most vital region of the host image such the modification of that region will corrupt the standard of the image. The watermark is made by two phases. The primary phase involves synthesizing of a picture from the sub-image of the image. Then a compound watermark is made by embedding a logo (watermark) to the synthetic image by employing a visible watermarking technique. This compound watermark is then invisibly embedded into the most block of the host image. This method proved its robustness and effectiveness under experimentation.

Ajinkya *et al* (2013), listed variety of techniques which may be used to detect data leakages as follows;

Perturbation technique: Perturbation means, where data is modified and make less sensitive before being handed to agents. But they claim that in some cases it's good to depart the

data unaltered. Examples being an outsourcer doing payroll, he must have the precise salary and customer checking account numbers.

Unobtrusive Techniques: Ajinkya *et al* (2013), developed a model for assessing the 'guilt' agents and also attempt to present an algorithm for distributing objects to agents, in a bit to improve the probabilities of identifying the leaker. With the heading Entities and Agents, the writers said "A distributor owns a group T = (t1.....tn) of valuable data objects. The distributor wants to share a number of the objects with a group of agents U1, U2 ...Un but doesn't wish the objects to be leaked to other third parties. The objects T might be of any type and size, e.g. they might be tuples during a relation, or relations during a database "An agent Ui receives a subset of objects, determined either by a sample request or a particular request.

i. Sample request

Ri = SAMPLE (T1, mi); Any subset of mi records from T are often given to Ui.

ii Explicit request

Ri=EXPLICIT (T, condi): Agent Ui receives all T objects that satisfy condi.

Guilty Agents: Suppose that after giving objects to agents the distributor discovers that a group S_T has leaked. This suggests that some third party, called the target has been caught in possession of S. for instance, this target could also be displaying S on its website, or perhaps as a part of legal discovery process, the target turned over S to the distributor. Since the agents Ui .....Un have a number of the info , it's reasonable to suspect them leaking the info .But during this case the agent can feign innocence or argue they need no hand within the leakage but rather the distributor secured it through other means. Experiments are often conducted and ask an individual with approximately the expertise and therefore the resources of the target to seek out the e-mail of say 100 individuals. If this person can find say 90 emails, then we will reasonably guess that the probability of finding one email is 0.9. On the opposite hand, if the objects in question are bank accounts numbers the person may only discover say 20, resulting in an estimate of 0.2. If the Model T was taken because the total content of objects and therefore the R's are used because the set of objects given to the agents and S is that the target set which contains the leaked objects. There are T= (t1, t2,

t3,), R1 =(t1,t2); R2= {t2,t3} ; S=(t1, t2, t3). By the design of the sets, it are often deduced that each one the objects given are leaked and thus appear in S.

To find the probability that an agent Ui is guilty given a set S, first, we compute the probability that he leaks a single object t to S. To compute this, we need to know which agents were given t , that set is Vt.

Pr {some agent leaked t to S} = 1-p..... (1)

All the agents possessing t have equal chances of leaking, thus

Pr {Ui leaks t to S} = $(1-p)/|Vt|$.....(2)

Probability that an agent leaked an object will be equally divided by number of agents who were given that object. If no agent has that object, i.e. Vt=0, then Pr (Ui leaks t to S) =0.

To compute probability that an agent Ui leaked is calculated by multiplying probability of each Object leak that are in both S as well as Ri, i.e.

Pr {G1 | S} = $1 - \prod t \in S \cap (1-(1-p)/|Vt|)$..... (3)

In the model, you can't make certain who leaked the objects because the objects are often obtained through guess work.

In the end agent can't be blamed for a leakage when the objects are often secured by guessing.

Fake Object Module

The distributor could also be ready to add fake objects to the distributed data so as to enhance his chances of detecting the guilty agents. The idea of perturbing data to detect leakage isn't new, consistent with the writers, e.g. by adding watermark to a picture or a noise to patients' records. Maximum care therefore should be taken in perturbing data because medical records for instance are sensitive data. In many cases, the distributor could also be limited in what percentage fake objects he can create .For example objects may contain email addresses, and every false e-mail address may require the creation of an actual inbox in other to not arouse the suspicion of the agent. The inboxes can actually be monitored by the distributor. If email is received from someone aside from the agent who was given the address, it becomes evident that the address was leaked. E-mail monitoring consumes resources therefore the distributor should have a limit on the amount of faux object.

Archana et al (2013), developed a model for locating the guilty agents. Also presented algorithms for distributing objects to agents, during a way that improves our chances of identifying a leaker. Finally, we also consider the choice of adding fake objects to the

distributed set. Such objects don't correspond to real entities but appear realistic to the agents. If it seems that an agent was given one or more fake objects that were leaked, then the distributor are often more confident that agent was guilty. They also considered optimization during which leaked data is compared with original data and accordingly the third party who leaked the info is guessed. They also used approximation technique to encounter guilty agents. They proposed one model which will handle all the requests from customers and there's no limit on number of consumers. The model gives the info allocation strategies to enhance the probability of identifying leakages. Also there's application where there's a distributor, distributing and managing the files that contain sensitive information to users once they send request.

Data leakage happens always when confidential business information like customer or patient data, ASCII text file or design specifications, tariffs , property and trade secrets, and forecasts and budgets in spreadsheets are leaked out. When these are leaked out it leaves the corporate unprotected and goes outside the jurisdiction of the corporation. This uncontrolled data leakage puts business during a vulnerable position. Once this data is not any longer within the domain, then the corporate is at serious risk.

At now the distributor can assess the likelihood that the leaked data came from one or more agents, as against having been independently gathered by other means. If the distributor sees enough evidence that an agent leaked data then they'll stop doing business with him, or may initiate legal proceedings.

Priyanka et al. (2013) present concept of knowledge leakage, its effects of leakage and various techniques to acknowledge the info leakage. The worth of the info is incredible, so it shouldn't be leaked or changed. Huge database is being utilized in IT field. This database is shared with multiple people at a time. But during this sharing of the data, there are huge chances of knowledge vulnerability, leakage or alteration.

So, to stop these problems, a data leakage detection system has been proposed. This paper includes brief idea about data leakage detection and a strategy to detect the info leakage persons.

Sandip *et al.* (2012) present the results of implementation of data Leakage Detection Model. Currently watermarking technology is being utilized for the info protection. But this technology doesn’t provide the entire security against data leakage. This includes the

difference between the watermarking & data leakage detection model's technology. This leads for the new technique of research for secured data transmission & detection, if it gets leaked. Malsoru V. (2016), author discussed a further scenario that shows how the sharing of S objects by agents affects the possibilities that they're guilty. The scenario conclusion matches our intuition: with more agents holding the replicated leaked data, it's harder to get the blame on anybody agent.

Sion *et al* (2003) deals with the thought of generating bit patterns on the file at certain location and every one the bit patterns merged and make a watermark. The bits inserted are set of numbers which give right protection to the info that's present within the data base. This also deals with the event of watermark detection application which reads the algorithms of the bit pattern by locating the markings and retrieves the first data at the client side. This implemented on the technique of watermarking (Hartung and Kutter, 2003) the info utilizing multi-media watermarking technology to stop the digital content going vital on net by disabling the copy facility. Encryption of the info has its own constraint from protecting the data. If the rights are decrypted then the info can't be shielded from illegally replicating the digital content. But this encryption issues is resolved by sung digital watermark which is embedded on the host data and can't be eliminated and it includes the copyrights, data protection and monitoring and tracking.

Sweeney (2002) deals with generalization and suppression techniques to guard the info from leakage using K-anonymity privacy protection. Where every a part of the info is categorized into k numerous subsets and each subset is linked with specific set of details and therefore the final data is obtained at the external source. This system may be a failure because it lacks in clear description on how the info is being secured and what happens to the info if they're not systematically liked to at least one another.

Patil and Sangve (2015), proposed improved Remote Data Possession Checking protocol supported homomorphic hash algorithm. Author says proposed system supports secure and efficient dynamic operations at block level. Dynamic operation consist insert, delete, update, and modify. They utilized Merkle Hash Tree to seek out the situation of every data. a 3rd party auditor also can be called as trusted party auditor (TPA) checks the user's data stored in cloud storage for its correctness and accuracy. A 3rd party ensures

correctness of user's data. Repeatedly verification is allowed without the requiring the verifier to match against the first data.

Monali *et al,* (2019), developed a system with three tier application that creates data access secure through a secure channel monitoring system. The resource manager makes the precise calculation of the RAM and CPU usage for the one data user and thereby predicts the usage for registered users. If the number of users are over the registered users, data leakage is detected and therefore the data service is stopped.

Also used the subsequent algorithms:

1) AES Encryption algorithm

Step 1: Byte Substitution (Sub Bytes)

The 16 input bytes are substituted by looking up a set table (S-box) given in design. The result's in a matrix of 4 rows and 4 columns.

Step 2: Shift rows

Each of the four rows of the matrix is shifted to the left. Any entries that 'fall off' are re-inserted on the proper side of row. Shift is administered as follows: First row isn't shifted. - Second row is shifted one (byte) position to the left. - Third row is shifted two positions to the left. - Fourth row is shifted three positions to the left. - The result's a replacement matrix consisting of an equivalent 16 bytes but shifted with reference to one another

Step 3: Mix Columns

Each column of 4 bytes is now transformed employing a special function . This function takes as input the four bytes of 1 column and outputs four completely new bytes, which replace the first column. The result's another new matrix consisting of 16 new bytes. It should be noted that this step isn't performed within the last round.

Step 4: Add round key

The 16 bytes of the matrix are now considered as 128 bits and are XORed to the 128 bits of the round key. If this is often the last round, then the output is that the ciphertext. Otherwise, the resulting 128 bits are interpreted as 16 bytes and that we begin another similar round. Advantages of AES Algorithm: - Symmetric key symmetric block cipher - 128-bit data, 128/192/256-bit keys - Stronger and faster than Triple-DES - Provide full specification and style details.

Sandip and Kulkarni (2012) developed a model for assessing the "guilt" of agents. Also present algorithms for distributing objects to agents, during a way that improves our chances of identifying a leaker. Finally, also consider the choice of adding "fake" objects to the distributed set. Such objects don't correspond to real entities but appear realistic to the agents. During a sense, the fake objects acts as a kind of watermark for the whole set, without modifying a person members. If it seems an agent was given one or more fake objects that were leaked, then the distributor are often more confident that agent was guilty.

Modules of knowledge Leakage Detection System by Sandip A. Kale and Prof. S.V.Kulkarni

A. Data Allocation Module

The main focus of the project is that the data allocation problem as how can the distributor "intelligently" gives data to agents so as to enhance the probabilities of detecting a guilty agent, Admin can send the files to the authenticated user, users can edit their account details etc. Agent views the key details through mail. So as to extend the probabilities of detecting agents that leak data.

B. Fake Object Module

The distributor creates and adds fake objects to the info that he distributes to agents. Fake objects are objects generated by the distributor so as to extend the probabilities of detecting agents that leak data. The distributor could also be ready to add fake objects to the distributed data so as to enhance his effectiveness in detecting guilty agents. The utilization of faux objects is inspired by the utilization of "trace" records in mailing lists. Just in case the incorrect secret key's given to download the file, the duplicate file is opened, which fake details also send the mail. Ex: The fake object details will display.

C. Optimization Module

The Optimization Module is that the distributor's data allocation to agents has one constraint and one objective. The agent's constraint is to satisfy distributor's requests, by providing them with the amount of objects they request or with all available objects that satisfy their conditions. His objective is to be ready to detect an agent who leaks any portion of his data. User are often ready to lock and unlock the files for secure.

D. Data Distributor Module

A data distributor has given sensitive data to a group of supposedly trusted agents (third parties). A number of the data is leaked and located in an unauthorized place (e.g., on the

online or somebody's laptop). The distributor must assess the likelihood that the leaked data came from one or more agents, as against having been independently gathered by other means Admin is in a position to look at the which file is leaking and faux user's details also.

E. Agent Guilt Module

To compute this PrfGijSg, there's need for an estimate for the probability that values in S are often "guessed" by the target. As an example, say a number of the objects in T are emails of people. An experiment are often conducted and ask an individual with approximately the expertise and resources of the target to seek out the e-mail of say 100 individuals. If this person can find say 90 emails, then we will reasonably guess that the probability of finding one email is 0.9. On the opposite hand, if the objects in question are checking account numbers, the person may only discover say 20, resulting in an estimate of 0.2. We call this estimate pt, the probability that object t are often guessed by the target. To simplify the formulas presented within the remainder of the paper, we assume that each one T objects have an equivalent pt, which we call p. Our equations are often easily generalized to diverse pt's though they become cumbersome to display. Next, make two assumptions regarding the connection among the varied leakage events. The primary assumption simply states that an agent's decision to leak an object isn't associated with other objects.

2.7 Comparison study of data leakage

Paper Title	**DESCRIPTION**	**TECHNIQUES**	**ALGORITHM USED**	**ADVANTAGE**
A Distribution model for Data Leakage Prevention,2013	Data allocation strategy to distribute the data to customers with the minimum transaction for fulfil their requirement.	Overlap of the data between minimum number of clients for easily detection. -	First Dis Algorithm Next Dis Algorithm, Minimum Sum Algorithm	In the minimum overlap process, The file transfer with the minimum number of users.
Assessing the Guilt Probability in Intentional Data Leakage, 2012	Every user has a different guilt probability (G) to leak the file. Guilt probabilities are calculated from how much time	Used guilt probability ratio of each client for detect user. -	Guilt Assessment Algorithm with data allocation strategy	With the highest guilt probability of user we easily detect the user Who have leaked the file.

	user access and transaction the file.			
The Guilt Detection Approach in Data Leakage Detection, 2015	Distributor adds some fake object with the data and send to the agent. The fake object helps to detect the user who have leak the file.	Add some fake record in original data to trace the leaker.	Fake object Algorithm in Sample Data Request and Explicit Data request	The fake tables maintain with the record of every user. The fake table helps detect the user who have leak the file.
A Novel Model for Data Leakage Detection and Prevention in Distributed Environment, 2016	Data allocation with fake objects and the reliability checker.	Used both fake object and Least overlap to detect guilt agent.	Least Reliable Agent [LRA] Algorithm and Agent reliability check in Hadoop framework (Data Leakage Prevention)	With the help of the least overlap method and some fake record to easily detect the leaker.
Data Leakage Detection and Data Prevention Using Algorithm	Distributor used the minimum Overlap process in this algorithm. Minimum overlap process used the minimum transaction of file to fulfil the client request.	Used the guilt probability and the least intersection of the file between the clients.	Allocation for Explicit Data Requests and Agent Selection for e- random with minimum overlap	In the base cover process, the document is move with the base number of the client. In the event that the record is released, at that point its simple to identify the leaker who has release the document.

CHAPTER THREE

METHODOLOGY

3.1 Research Methodology

This is a systematic approach towards the collection of data, the procedure of analyzing these data in order to test the hypothesis and then the conditions for acceptability of this hypothesis. The description of this subject matter will include the definition of the population sample that was selected from the collected data of various sources.

In this section of the research, the data collected by the researcher for this project work was secondarily from textbooks, internet, related papers and journals on towards an effective data leakage detection system and then primarily through interviews. Hence the data is pertained to the old existing method of data leakage detection.

This therefore, allowed the researcher to clearly identify that the important area that required changes, upgrading and collectively a computerized system is on the part of system. The researcher arrived at this conclusion by employing various types of research findings method on which she performed an in depth and comprehensive study to put together a vital and relevant fact which will be an embodiment in the development of newly and improved system upon the existing approach.

Sources of data;

- Primary source of data
- Secondary source of data

Primary source of data is the type that permits the researcher to derive first-hand information through observation, experience, investigation and survey (questionnaire and interview). Hence the researcher for special purpose collected data in form of facts and figures in relation to the population to provide a primary data, which gave the exact information, needed. The terms were carefully defined to be natural and avoid misunderstanding.

Secondary sources of data are the type that permit the researcher to have a second hand information from other materials like journals, magazines, textbooks, dictionaries which the researcher finds useful on the subject matter.

The main purpose of the research work was to show how user accessibility and activity to a computing resource could be tracked, monitored and audited to protect the integrity, accessibility and availability of data to authorized and authenticated users in cloud applications. The strategy was to adopt Data Leakage Detection Software. This is because these software hosted in cloud environments have no local databases, different access levels and remote user authentication and authorizations.

3.2 Methodology Adopted

The methodology adopted for this research work is the Object Oriented Analysis and Design methodology (OOAD). It improves the quality of the system due to its property of program reuse and maintenance. Object oriented programs (OOP) methods make code more maintainable and identifying the source of errors are very easier because objects are self-contained (encapsulation). In this Object-Oriented-Analysis and Design (OOAD) model, there is no separation between the analysis and the design phases, which improves communication with the users throughout the project development. Object-Oriented-Analysis and Design (OOAD) model is the industry-proven methodology for developing high-quality object oriented systems. The software life cycle is typically divided into stages going from abstract descriptions of the problem to designs then to code and testing and to deployment. The phases are: Requirements, Design, Implementation, Verification and Maintenance.

1. **Requirements:** During the requirement phase, the researcher asked questions on the software and hardware required in developing the new system. It was achieved through observation, experience, investigation, materials from journals, textbooks, etc. The key requirement is the user inputs**.** Three users were required to log into the system; the first is the administrator who is responsible registering the user and assign roles to them, while the second user is the distributor who uploads files in the application, shares to the users, then sends/approves key request to the user. Then, the third is the user, who request for file and also downloads file. The users were required to provide details such as Username, phone number, email address and a password before they can be allowed access to the system.
2. **Design:** During this phase, the framework (algorithm) of the new system was designed to also accommodate existing components that can be reused (or reintegrated) into our own

system. Different modules were designed such as audit trail module, registration module etc. The platform for uploading and sending files were incorporated to our new system to efficiently prevent the leakage of data. Also, the encryption and decryption platforms which are a very important part of the design were added to the new system.

3. **System Implementation:** To achieve this an object oriented programming language was used such as C# embedded in ASP.Net MVC method. This stage involves implementation of the conceptual model produced during object-oriented analysis. In OOD, concepts in the analysis model, which are technology−independent, are mapped onto implementing classes, constraints are identified and interfaces are designed, resulting in a model for the solution domain, i.e., a detailed description of how the system is to be built on concrete technologies.
4. **Verification/Validation**: The new system was tested with different users (about 50 users) logging in and accessing files sent by the admin. The administrator also has administrative privileges different from that of the users. As a test mechanism, only one administrator was used while all registered users appeared on the admin's page. A local server domiciled on the admin's computer stored all files for both administrator and users.
5. **Maintenance:** The maintenance phase involves making changes to hardware, software, and documentation to support its operational effectiveness. It includes making changes to improve a system's performance, correct problems, enhance security, or address user requirements.

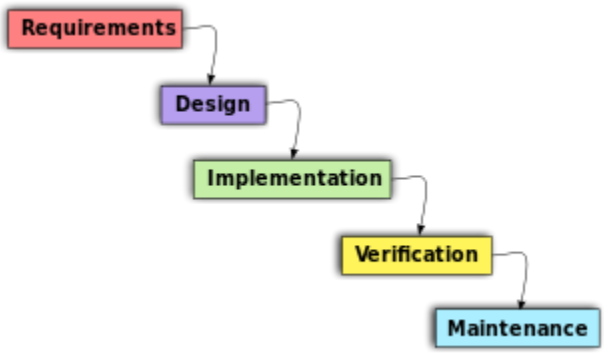

Fig. 3.1: Waterfall Model (Researcher, 2021)

The advantages of OOADM include:

Easy maintenance of the objects.

Objects may be understood as stand-alone entities.

Objects are reusable components.

For some systems, there may be an obvious mapping from real entities to system objects.

Encourages encapsulation

Easy to understand because it is based on real world objects.

3.3 Analysis of Existing Systems

For the development of security solutions vast challenges have been faced in the cloud computing environment. In particular, an organization faces challenges on files security and it tries maximum authentication over files while distributing to the user. Especially in software as a service the administrator has to distribute files to the employees to process the tasks and client process those request and resend to the admin when completed. When retrieving data, user is not a trusted party and any third party can access data with that user id too. This may lead to data leakage and we cannot find out that exact user who communicates with third party that is through any guilt agent (Papadimitriou and Garcia-Molina, 2011). The old system made of watermarking technique and data allocation strategy.

3.3.1 Watermarking Technique:

This is a technique that implements a unique code in every copy of data. If later, a distributed copy of data is found at some unauthorized location, the guilty agent can be detected very easily (Bhatt and Sharma, 2014). This technique seem to be very useful at some places, but it causes some changes into the real data. Furthermore, watermarks can be damaged sometimes if the data receiver is malicious. For example, a hospital supplies/provides its patient records to the researchers who use this data to formulate new more effective treatments. In the same manner, a company running in partnerships with some other companies requires sharing of its client data. Major disadvantages of watermarking technique can be listed as:

1. It causes some variation/changes into data by modifying some of the data attributes and hence makes data less sensitive. This modification of data is known as the perturbation. On the other hand in some situations, real data can't be modified at any level. For

example, if an agent requires the exact salary to perform payroll. Salary can't be modified here

2. Next problem is that, if the recipient is nasty/malicious, it can be damaged very easily.

3.3.2 Data Allocation Strategy

This is a technique that formulates various strategies to allocate the data among the agents so that the probability to detect the guilty agent(s) can be improved. These methods are not based on the modification of the data as watermarks do. In some situations, there is need to embed "realistic but fake" data records with real data so that the chances to identify data leakage and hence to detect the guilty agent can be improved. A large number of algorithms are been used to distribute objects to agent effectively. The goal is to find that is critical data been leaked by any of the agents, and next target is to detect the guilty agent who caused data leak.

Architecture of the existing system

Monali et al (2019), developed a three tier application that makes data access secure through a safe channel monitoring system. The resource manager makes the exact calculation of the RAM and CPU usage for the single data user and thereby predicts the usage for registered users. If the number of users are more than the registered users, data leakage is detected and the data service is stopped.

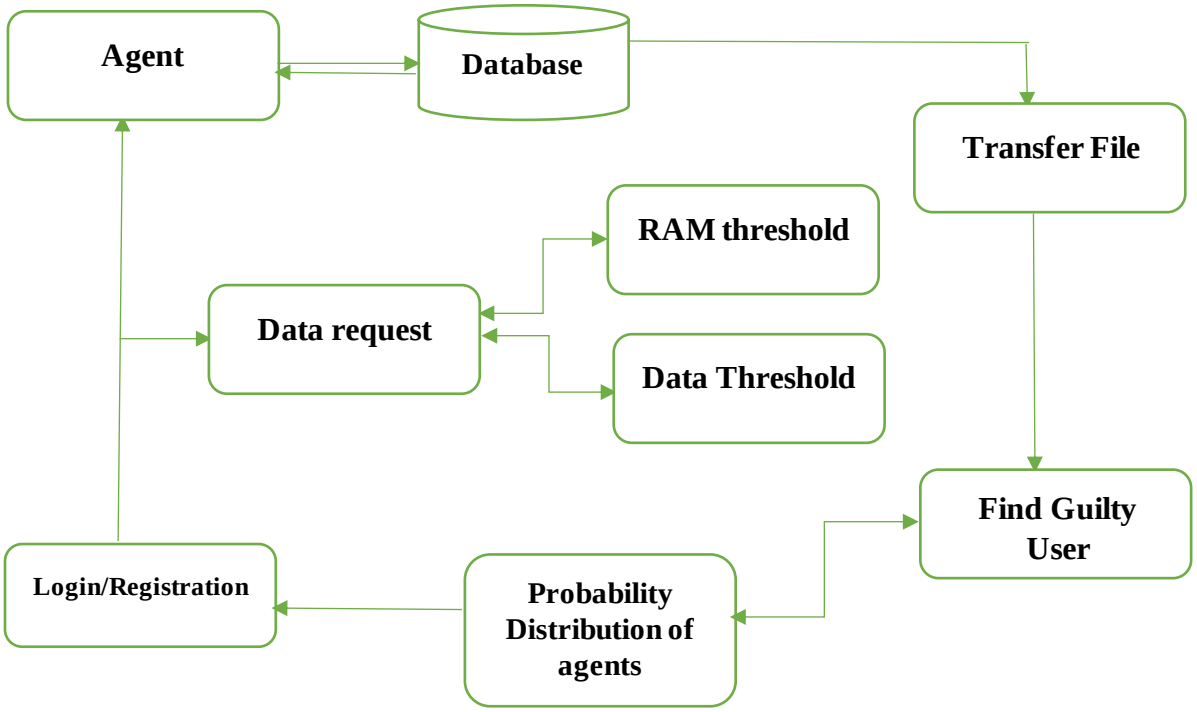

Fig. 3.2: Architecture of the existing sytem (Monali et al, 2019)

3.4 Challenges of the existing system

The present models of data leakage detection has been faced with many challenges, most of which are based on security issues. Watermarks can be very useful in some cases, but, involve some modification of the original data. Watermarks can sometimes be destroyed if the data recipient is malicious. If the object to be watermarked cannot be modified, then a watermark cannot be inserted.

It is also possible to assess the likelihood that an agent is responsible for a leak, based on the overlap of his data with the leaked data and the data of other agents, and based on the probability that objects can be "guessed" by other means.

Adding fake objects can be very useful in some cases, but in some cases data need to be exact. E.g. A hospital may give patient records to researchers who will devise new treatments. This case requires the exact history of the patient.

The three tier application that calculates the CPU and RAM usage for single user only detects the number of registered user and the guilty user, some users may be offline at the moment and can be traced.

3.5 PROPOSED SYSTEM

The proposed system provides security to overcome issues related to cloud environment which provides access to store and distribute data from cloud administrator to user or employees of a company. It helps to protect the data from leakage by tokenization of files before distributing and set the time bound for each and every file that particular user needs to download from the cloud storage. This formula protects the data leaked from guilty agent who act as a third party and security is provided using Key generation which is an auto generated random unique number for every file when user or an employee make attempts to view the content of file. Incorporated in it also is an Audit Trail/ Transaction log which profiles user activities in the system as against the old system that does not have an Audit trail. The audit trail will monitor when a user sends out organization's information with date and time stamp.

When these data is leaked out, then the companies are at serious risk. This system presents the data leakage detection from trusted third parties and provides an administrator to easily identify the third party or guilty user. This will help user to make better understanding of security issues in cloud computing environment.

3.5.1 Advantages of the proposed system

The proposed system is very useful as compared to watermarking and data allocation strategy in the sense that it can provide security to our data during its distribution or transmission and even can detect if the data gets leaked. Thus using this methodology security of data is ensured and detection technique is provided. This model is very helpful in various industries, where the data is shared through any public or private channel and shared with third party. This system is simple, but it captures the essential trade-offs.

3.5.2 Architecture of the Proposed Model

The proposed system has in it an audit trail/transaction log system which monitors user activities/transactions to computing resource at every stage of the process. The system administrator registers each user and assigns roles to them. The roles include an administrator, a distributor and a user. The distributor uploads new files into the system, approves, sends key request to the user, also shares file to each user. Figure 3.3depicts the architecture of the proposed system:

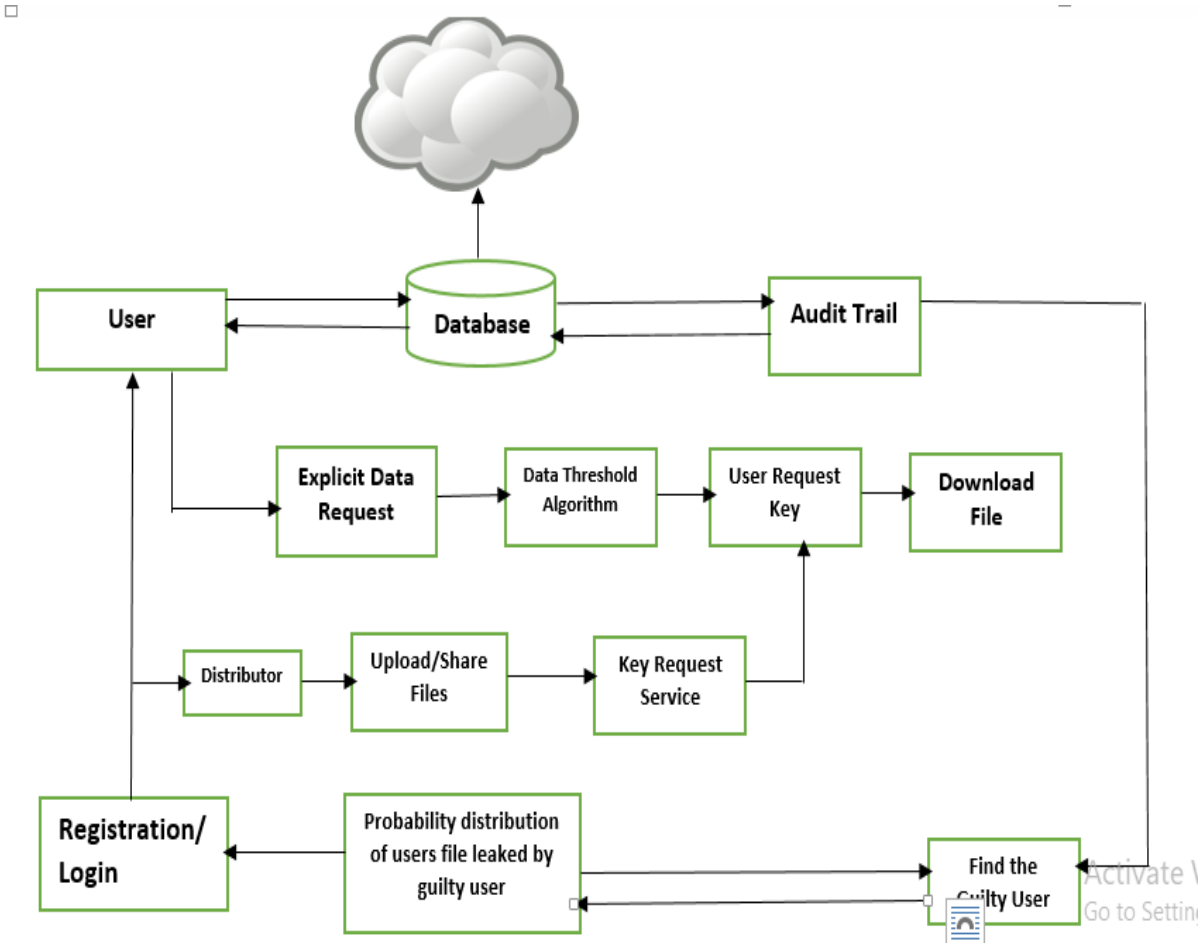

Fig. 3.3: Architecture of the proposed system. (Researcher, 2021)

3.5.3 High Level Model of the Proposed Model

Figure 3.4 depicts the high level model of the proposed system, there are different modules they are:

Registration Module: here, the owner of the data that is the Administrator logins in and registers new users, then assigns roles to them. The role could be Distributor of files or just a User.

Upload and Share files Module: In this module, the distributor uploads files and shares these files to different users.

Key Request Module: the user on viewing the shared files, request for dynamic key/code in order to download the requested file.

Download File Module: with the dynamic key/code sent by the distributor to user, the user inputs the code to download the file.

Audit Trail/Email Alert Module: the administrator always checks and monitors the audit trail /transaction log so as to detect when sensitive information is being leaked out. The Administrator also get email alert whenever any file is being sent out.

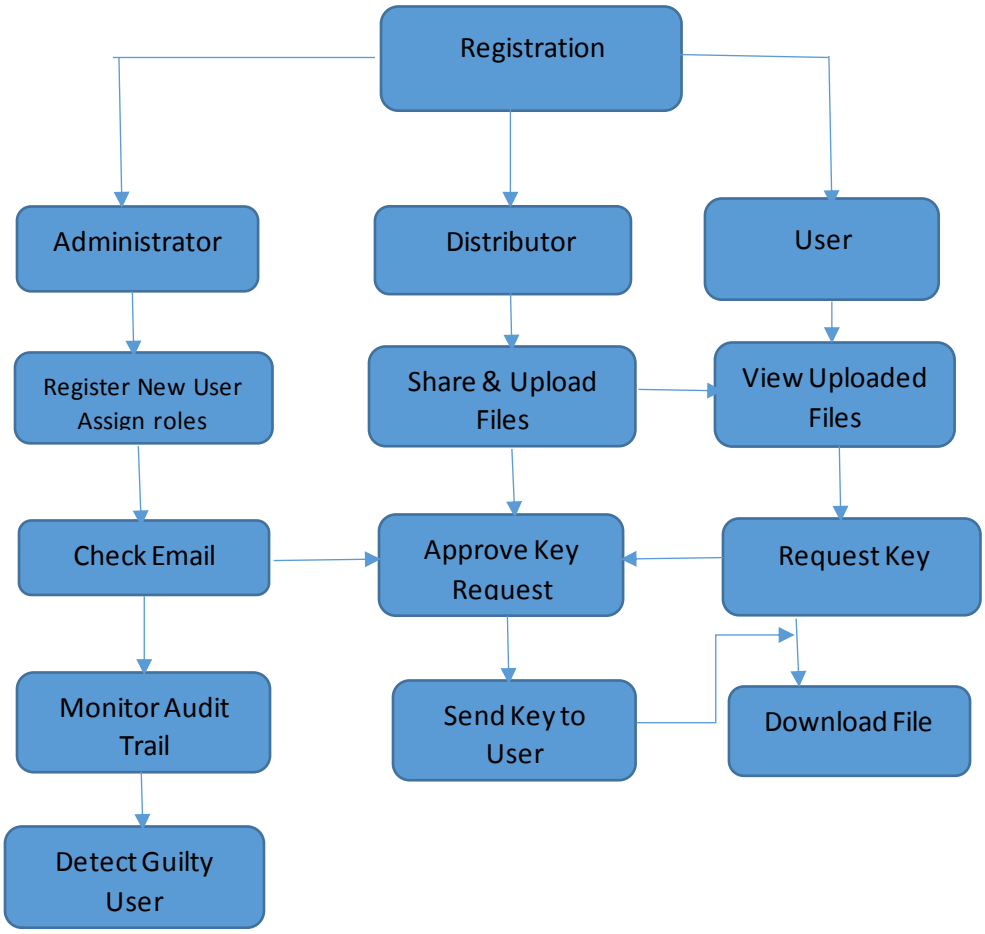

Fig. 3.4: High Level Model of the Proposed System (Reseacher, 2021)

3.6 System Flowcharts

The flowchart of the system is represented in Figure 3.5, figure 3.6, and figure 3.7. The system flowchart has three (3) processes,

- The flowchart for the Admin that registers all the users and assign roles to them.

- The distributor who uploads files, approves key request and sends the secret key to the user.

- Lastly the user

3.6.1 Admin Flowchart

Here, the Admin logins in with username and password in order to register new users and assigns roles to them. Fig.3.5 depicts flowchart for the admin.

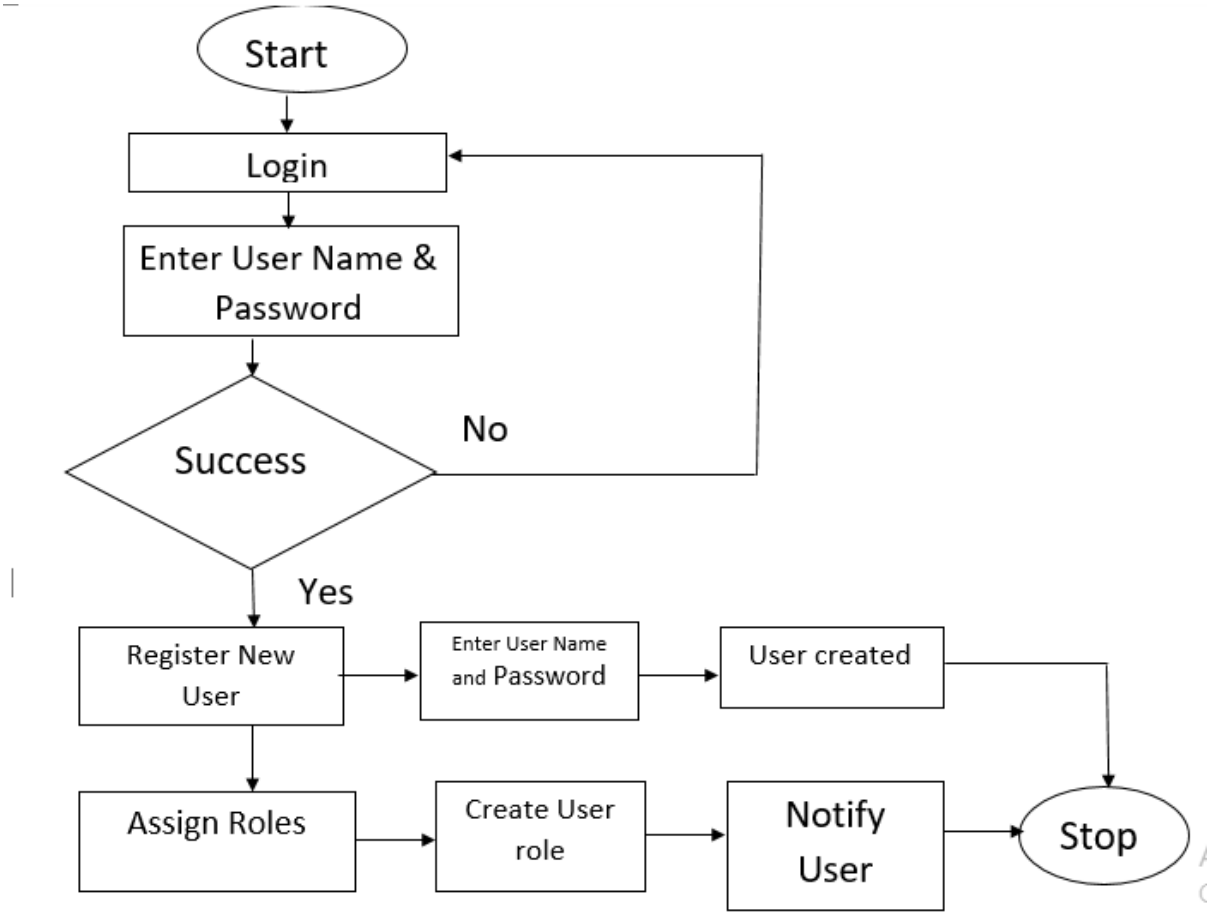

Fig. 3.5: Flowchart for the Admin

Figure 3.6 below is the flowchart the distributor. Here, the distributor login to upload and share files, also approves key request.

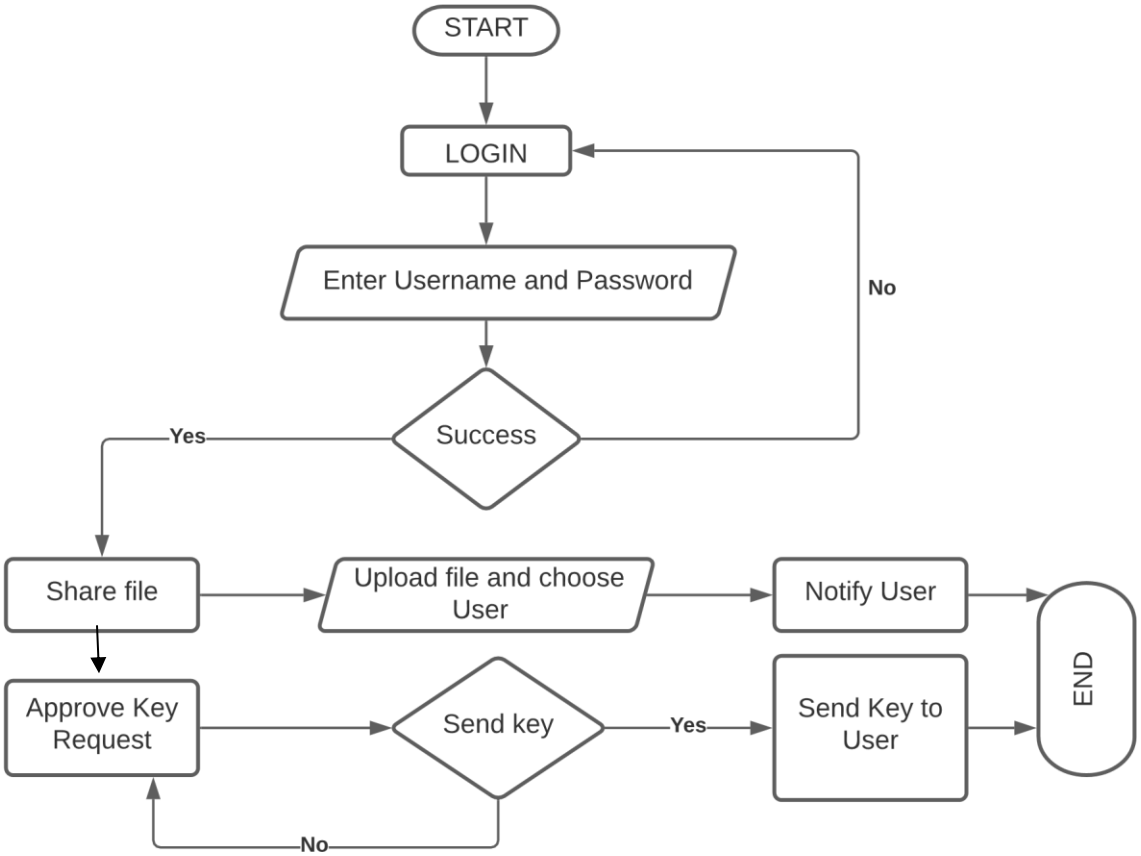

Fig. 3.6: Flowchart for the distributor

The User logins with username and password to view files uploaded and shared by the distributor, then sends out key request in order to download a file. It can be depicted on figure 3.7 below;

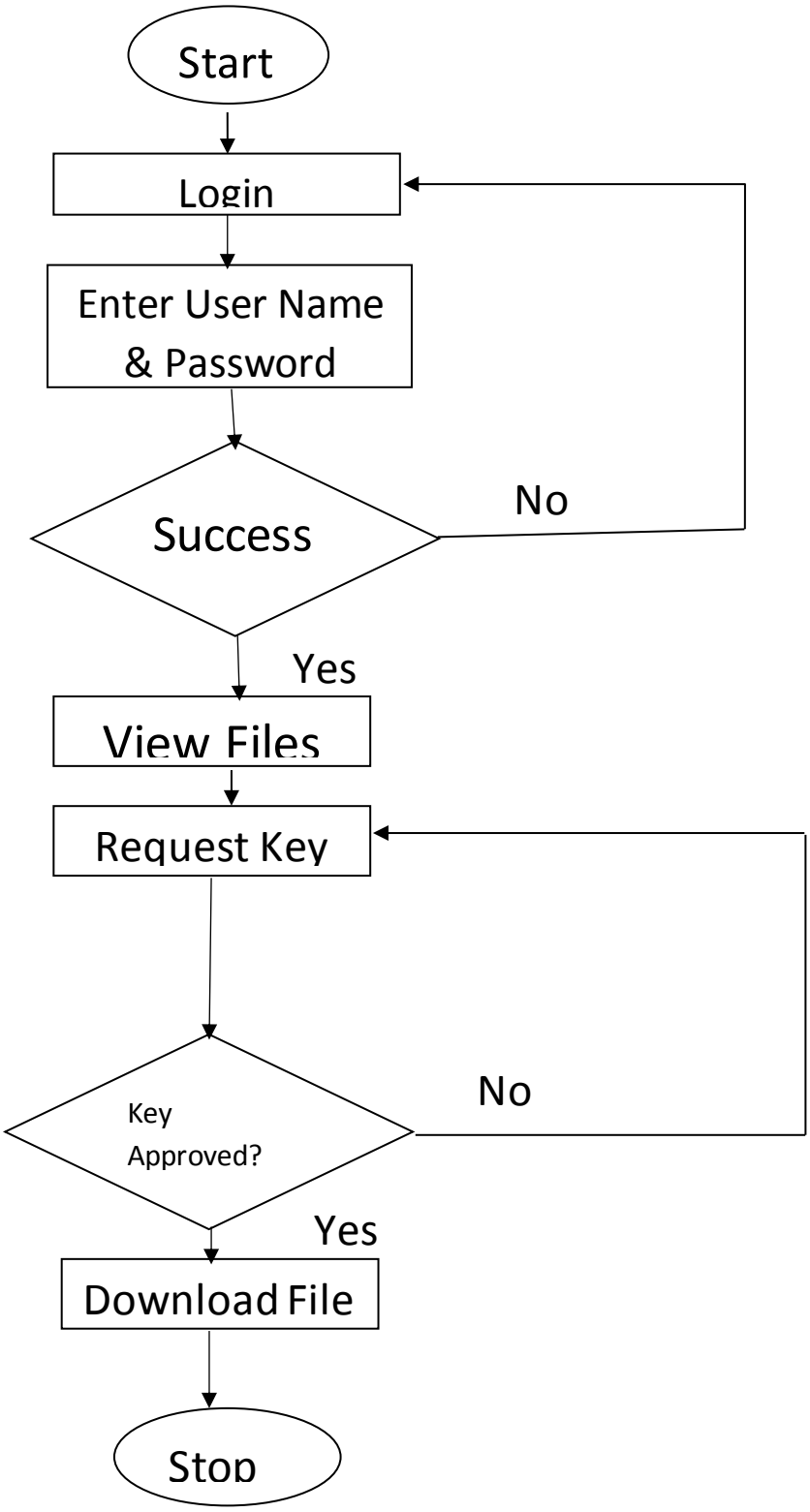

Fig. 3.7: Flowchart for the user

3.7 Database Design

The storage server is designed to store user login details and files or messages for authenticated users. It also stores users' registration details.

The database design is divided into three stages: Registration Design, Login Design User Role Design and File Upload Design. These are represented in Tables 3.1, 3.2, 3.3 and 3.4.

Table 3.1: User Registration Design Database Structure

S/N	FIELD NAME	DATA TYPE	FIELD SIZE	DESCRIPTION
1	UserID	NVARCHAR	20	User's ID
2	UserName	NVARCHAR	20	User's name
3	Email	NVARCHAR	30	User's Email address
4	Phone_Number	NVARCHAR	20	User's phone number
5	Password	NVARCHAR	20	New user's password

Table: 3.1 illustrates the field name and data types of the user registration module

Table 3.2: User Login Database Structure

S/N	FIELD NAME	DATA TYPE	FIELD SIZE	DESCRIPTION
1	UserID	NVARCHAR	20	User's ID
2	UserName	NVARCHAR	20	User's name
3	Email	NVARCHAR	30	User's Email address

Table 3.3: User Role Database Structure

S/N	FIELD NAME	DATA TYPE	FIELD SIZE	DESCRIPTION
1	UserID	NVARCHAR	20	User's ID
2	RoleID	NVARCHAR	20	Role's ID
3	Email	NVARCHAR	30	User's Email address
4	Username	NVARCHAR	20	User Name
5	Role Name	NVARCHAR	30	Role Type

Table 3.4: File Database Structure

S/N	FIELD NAME	DATA TYPE	FIELD SIZE	DESCRIPTION
1	UserID	NVARCHAR	20	User's ID
2	UserName	NVARCHAR	20	User's name
3	FileName	NVARCHAR	MAX	File Type
4	DateupLoaded	NVARCHAR	MAX	Data file was uploaded

3.8 Logical Design

Logical database design was constructed which is based on the conceptual data model, this allows for the insertion, modification and deletion of rows without any inconsistencies and a coherent relation was constructed. Figure 3.8 illustrates the relationship diagram of the database.

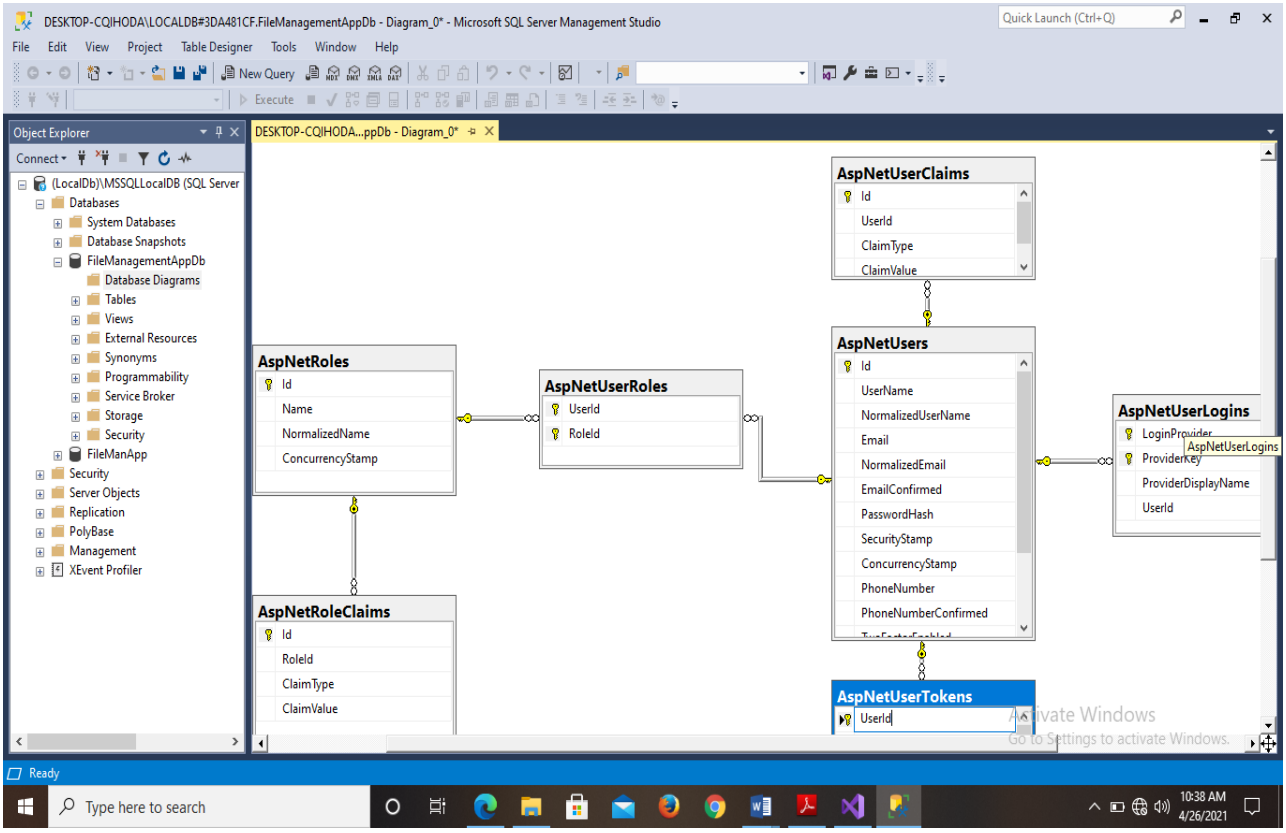

Fig. 3.8: Relationships in the database design (Researcher, 2021)

3.9 Use case diagram

Figure 3.9 is the use case diagram of the model. It shows the different activities that begin with the administrator who creates new users and assigns role to them. It is depicted by rectangle. The actors which are individual concerned with the system have been defined in concordance to the role they play in the system. The designated function carried out by each actor is depicted within and around the system. Finally, the relationships between the actors and use cases is defined by arrow.

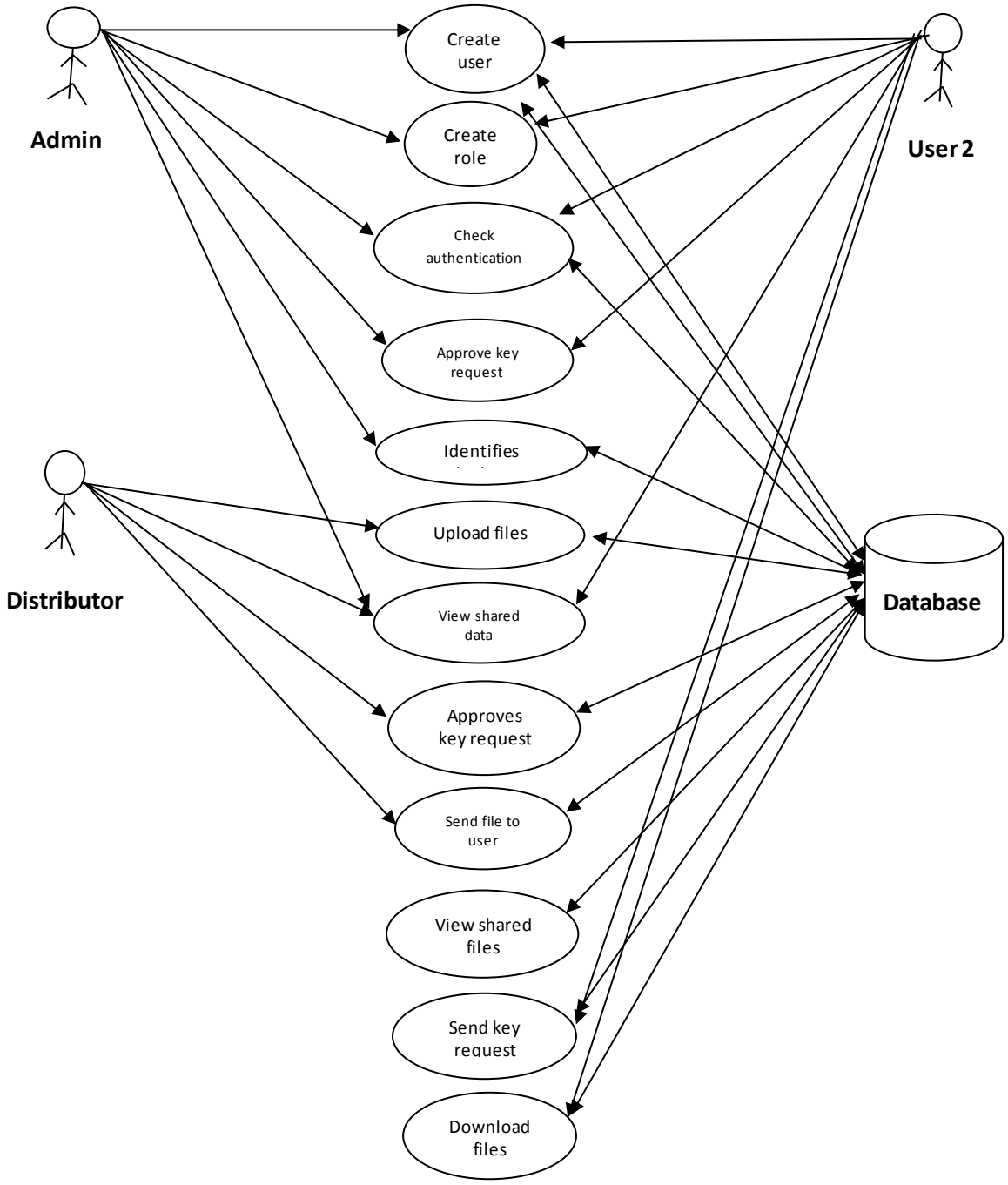

Fig. 3.9: Use case diagram (Researcher, 2021)

Evaluation metrics: Many models were evaluated to better understand the system and what it is supposed to do and also to demonstrate the kind of relations that should exist between the various components of the project. Among the models evaluated in this project are the entity relational diagrams, system flowcharts , context diagram, requirement use case diagram and a key based diagram but the one which most fits into this development model was the entity relational.

Entity Relational Diagrams: In the entity relational relationships in this project, there is a direct link between the User and program Accessibility options. A user's authentication determines his/her accessibility into the software program. If authenticated, it shows that the user has been duly verified based on credentials contained in the repository database system.

3.10 Choice of Programming Language

The programming language used in the design of the proposed system is Asp.net MVC method. ASP.NET MVC is a web development framework from Microsoft that combines the features of MVC (Model-View-Controller) architecture.

The MVC (Model-View-Controller) design pattern actually has been around for a few decades, and it's been used across many different technologies. From Smalltalk to C++ to Java, and now C Sharp and .NET use this design pattern to build a user interface.

Some salient features of the MVC pattern are as follows:

1. Initially it was named Thing-Model-View-Editor in 1979, and it was later simplified to Model- View-Controller.
2. It is a very powerful and elegant means of separating concerns within an application (for example, separating data access logic from display logic) and applies itself extremely well to web applications.
3. Its explicit separation of concerns adds a small amount of extra complexity to an application's design, but the extraordinary benefits outweigh the extra effort.

The MVC architectural pattern separates the user interface (UI) of an application into three major parts.

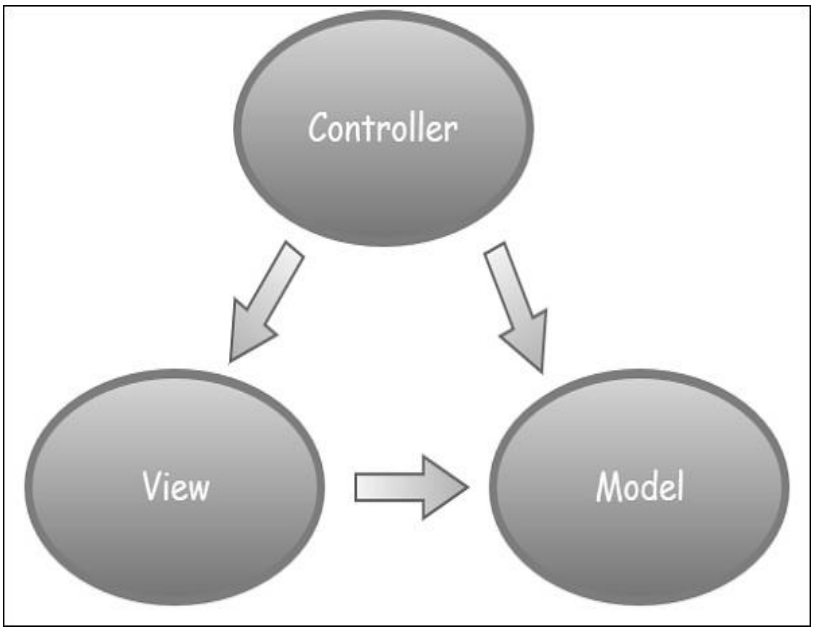

Fig. 3.9: MVC Architectural Pattern (Reseacher, 2021)

The Model – This is a set of classes that describes the data you are working with as well as the business logic.

The View – this defines how the application's UI will be displayed. It is a pure HTML, which decides how the UI is going to look like.

The Controller – This is also a set of classes that handles communication from the user, overall application flow, and application-specific logic.

Benefits of ASP.NET MVC

The benefits of using ASP.NET MVC are −

1. ASP.NET makes it easier to manage complexity by dividing an application into the model, the view, and the controller.
2. It enables full control over the rendered HTML and provides a clean separation of concerns.
3. Directs control over HTML also means better accessibility for implementing compliance with evolving Web standards.
4. Facilitates adding more interactivity and responsiveness to existing apps.

5. Provides better support for test-driven development (TDD).
6. Works well for Web applications that are supported by large teams of developers and for Web designers who need a high degree of control over the application behavior.

Visual web developer 2019 (VISUAL STUDIO 2019) was used as the web authoring tool because of its flexibility, bend ability and very easy deploying site.

3.11 System Specifications

These are the minimum specifications or system requirements without which the new model cannot be successfully implemented. It is divided into two categories; hardware specifications and software specifications.

3.11.1 Hardware Requirements

Being a server/client platform that runs on a network, the system is not a bulky and space-consuming one. The system will take the following minimum hardware requirements for its successful implementation;

- **Primary Memory:** 1GB of RAM and above
- **Secondary Memory:** 200GB of Hard disk space and above
- **CPU:** Duo Core Intel Processor with frequency of 2GHz or higher
- **Monitor:** VGA / SVGA
- **Mouse:** Digital mouse

3.11.2 Software Requirements

The software requirements of the system are as follows:

- **Operating System:** 64bit Windows 8 and above
- **Programming Frontend:** Visual Studio 2012 IDE or above
- **Programming Backend:** Microsoft SQL Server Management Studio version 18

Browser: Windows Explorer 7.0 / Moxilla Firefox 58.0 or above

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 User Interfaces

The user interface (UI) is the point at which users of the system interact with a computer, website or application. The goal of is to make the user's experience easy and intuitive, requiring minimum effort on the user's part to receive maximum desired outcome.

User Interface (UI) is created in layers of interaction that appeal to the human senses (sight, touch, auditory and more). They include both input devices like keyboard, mouse, trackpad, microphone, touch screen, fingerprint scanner, e-pen and camera and output devices like monitors, speakers and printers. Devices that interact with multiple senses are called "multimedia user interfaces". For example, everyday UI uses a combination of tactile input (keyboard and mouse) and a visual and auditory output (monitor and speakers).

Other types of user interfaces can include:

Form-based user interface: this is used to enter data into a program or application by offering a limited selection of choices. For example, a settings menu on a device is form-based.

Graphical user interface (GUI): A tactile UI input with a visual UI output (keyboard and monitor).

Menu-driven user interface: this is a type of UI that uses a list of choices to navigate within a program or website. For example, ATMs use menu-driven UIs and are easy for anyone to use.

Touch user interface: User interface through haptics or touch. Most smartphones, tablets and any device that operates using a touch screen use haptic input.

Voice user interface: it is the interactions between humans and machines using auditory commands, examples include virtual assistant devices, talk-to-text, GPS and much more.

User interface is a module in the encryption system through which users interact with the system. User information required by the system is gotten through the modules and any desired outputs are also shown through the same modules. These completely make up the input and output interface.

4.1.1 Input Interface: Input refers to the information or requirements fed into the computer which is processed to produce the desired output. In this work, data is supplied by the user during registration which in turn is used to create login details and the secret key of the user. The interface makes use of GUI components such as radio buttons, textboxes, text boxes to accept input from the users into the system.

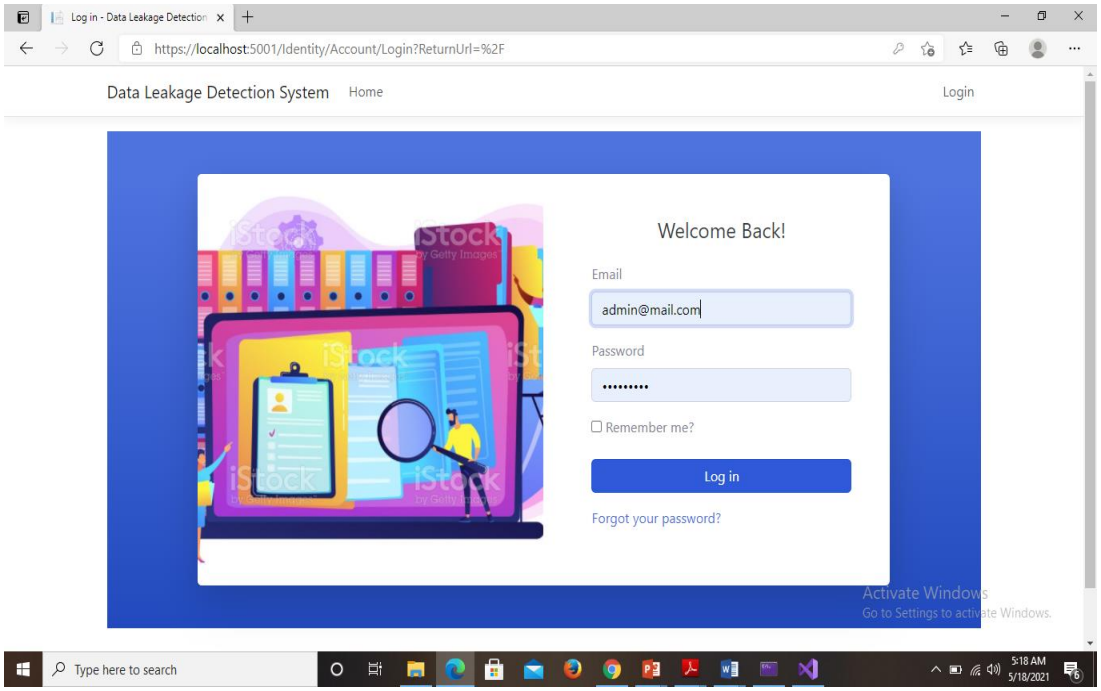

Fig. 4.1: Admin sign in Module (Researcher, 2021)

Figure 4.1 shows the module where the admin logs in to the system in order to create new users and assign role to the users.

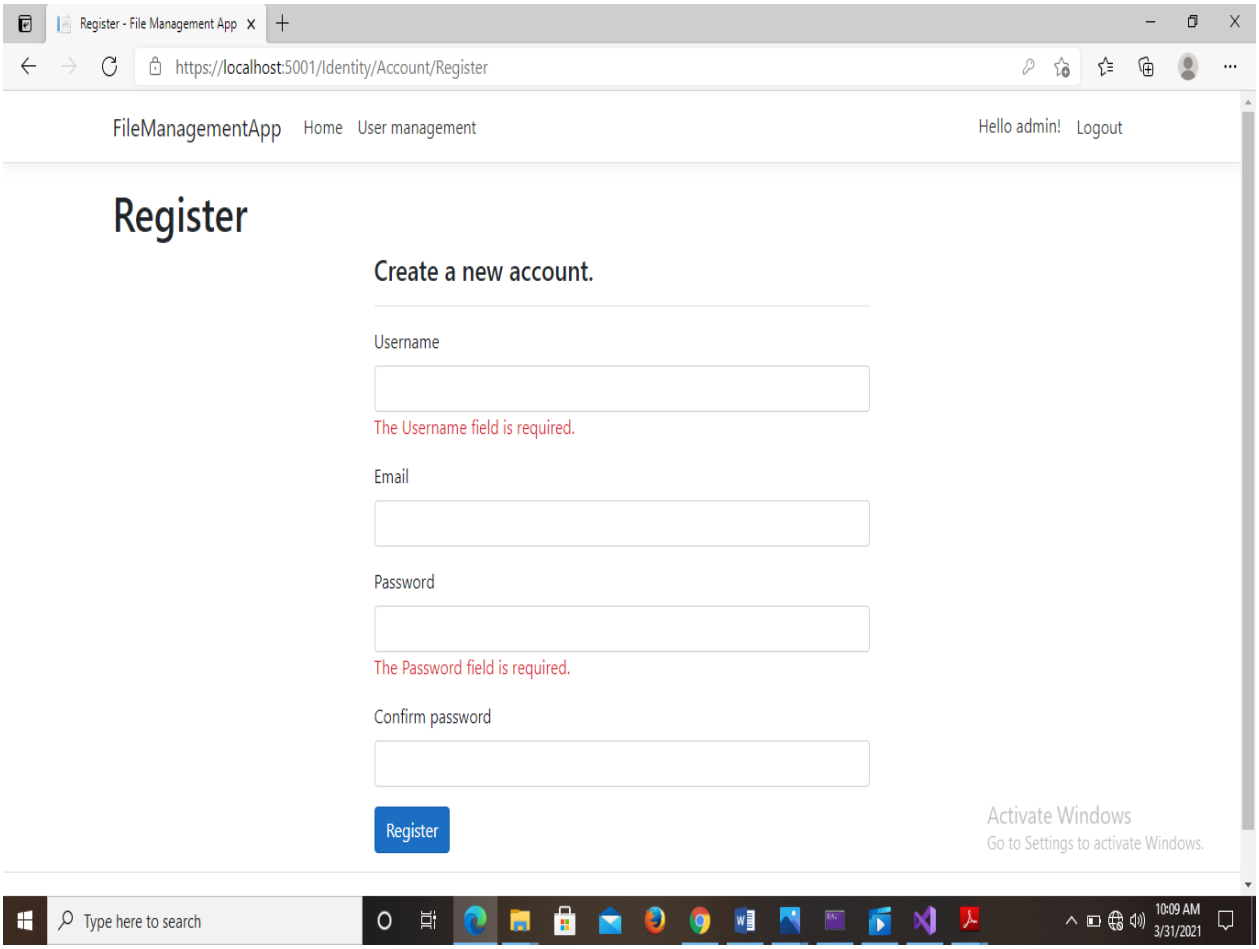

Fig. 4.2: Admin registers a new user (Researcher, 2021)

Figure 4.2 illustrates where the administrator creates a new user. The "Register" button is clicked after entering the fields.

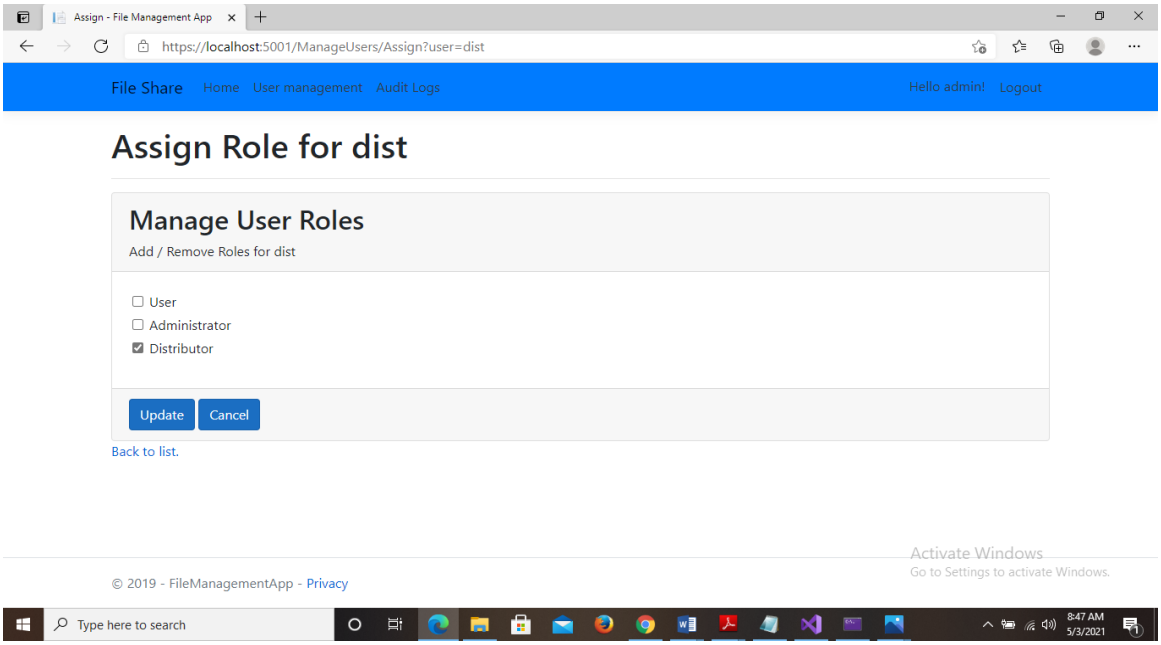

Fig. 4.3: Admin assigns roles (Researcher, 2021)

Fig 4.3 shows where the administrator assigns role to the users, one user can be a distributor of files, or another administrator or only a user.

4.1.2 Output Interface/Simulation Test Results: Firstly, the administrator logs into the system and has the privileges of registering users, viewing the list of all registered users, assign roles to members, and checking the audit trail to identify a leaker.

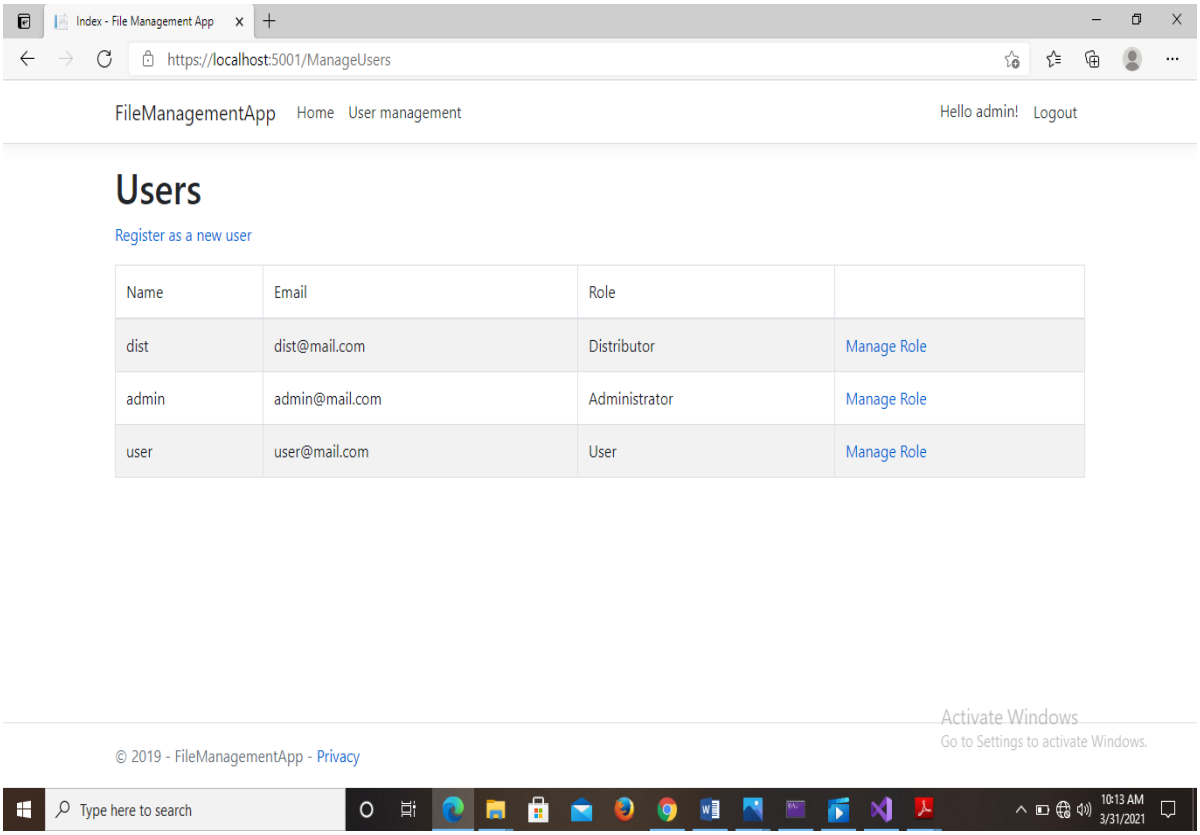

Fig. 4.4: Admin dashboard (Researcher, 2021)

In Figure 4.4, the admin views the set of all users that have registered and also can assign roles to them.

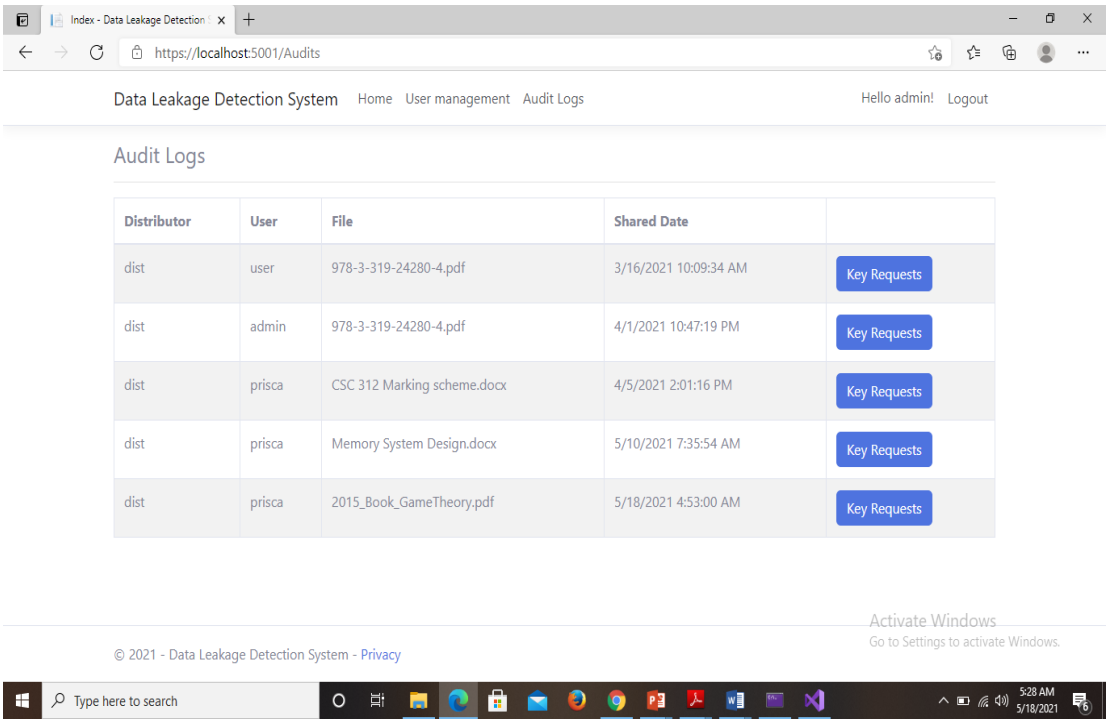

Fig. 4.5: Audit Trail/Transaction log module (Researcher, 2021)

Figure 4.5 shows the audit log or trail that tracks user activities/transactions in the system. It shows the various transactions going on or performed by the user including the date and time and actual activity performed at each stage of the application process. The audit trail table has no link or dependency to the other tables in the program making it safe from manipulations from other tables. It is intended to be activated at every login stage for user identification and verification purposes. It has been tied to application accessibility, which means a user must be successfully verified and authenticated before access is granted into any software application.

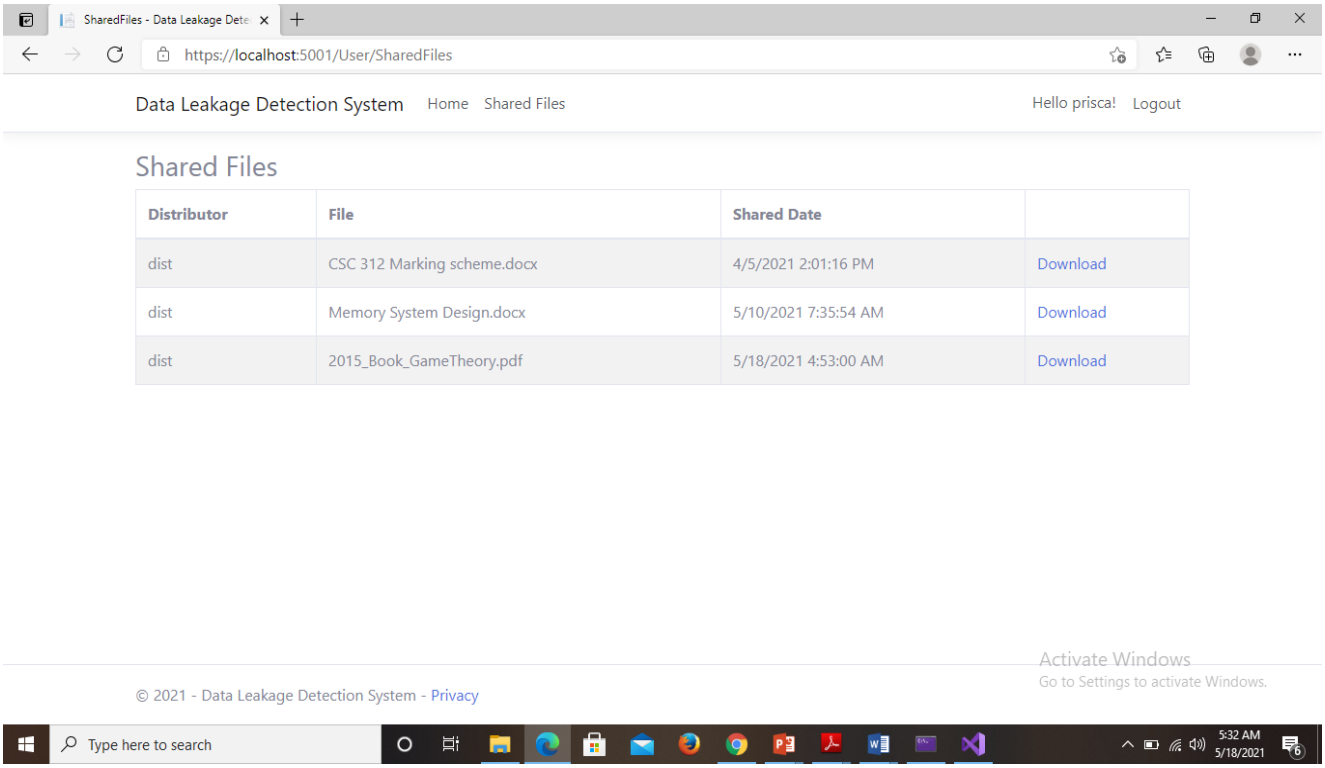

Fig. 4.6: Shared file module (Researcher, 2021)

Figure 4.6 shows the files shared to a particular user and can be downloaded through the approval of key request by the distributor or administrator.

4.2 System Testing

This is ensuring that the program runs as expected. Free of errors. The system developed was not free of bugs. We therefore employed the following testing and debugging method to checks for errors.

a) Desk Checking
b) Unit Testing
c) Integration Testing
d) Alpha Testing
e) Beta Testing

Desk Checking: This means reading through or checking the programs to make sure that it is free from errors and that the logic works well (correctly) before it is entered into the computer.

Unit Testing: Different modules are tested and the specifications are produced during design for the modules. It is essential for verification of the goal and to test the internal logic of the modules. Unit testing was conducted to the different modules of the project. Errors were noted down and corrected down immediately and the program clarity as increased.

Integration Testing: Integration testing is a systematic testing of constructing structure. At the same time tests are conducted to uncover errors associated with the interface. It need not be the case, that software whose modules when run individually and showing perfect results will also perfect results when run as a whole.

Alpha Testing: Some errors were not detected during desk checking, so we prepared some test data with known output to test the program output if it tallies with the expected result.

Beta Testing: This testing is done with real life data and real users. At this stage, we test all possibilities that may lead to failure of the program. After testing of the program and we are now sure that it is free from errors we preceded to the next phase of System Implementation.

4.3 Analysis of Results

The results of the new model were analyzed and compared with the previous model presented by previous researchers.

Registration: in our new model, the administrator creates or registers all the users and assigns roles.

Dynamic Code/Key: in the new system, the user request for key/code in order to download any file. This code expires after two hours from the time it was sent to the user and cannot be used.

Audit trail: the admin always checks and monitors the audit trail to detect possible data breaches. Any file been shared, uploaded and sent are seen in the trail by the administrator.

Database: our database is in the cloud.

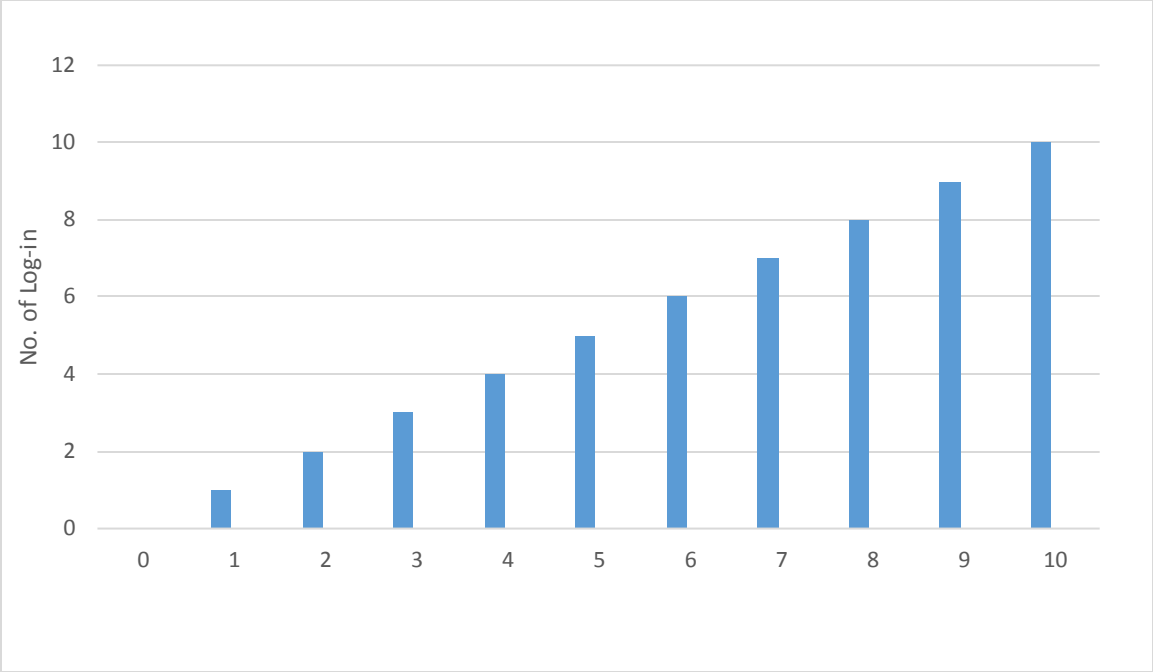

Fig. 4.7: Graph of Entries without Transaction Log/Audit Trail (Researcher, 2021)

Table 4.1: Table of Entries without Transaction Log/Audit Trail

No. of Users	**Audit Trail**
0	**nil**
1	**nil**
2	**nil**
3	**nil**
4	**nil**
5	**nil**
6	**nil**
7	**nil**
8	**nil**
9	**nil**
10	**nil**

Figure 4.7 depict graphically, 10 clients entered into the system without transaction log/Audit trail system. The graph shows that for any member of clients entered there is not any mechanism to profile user activity/transaction in the system.

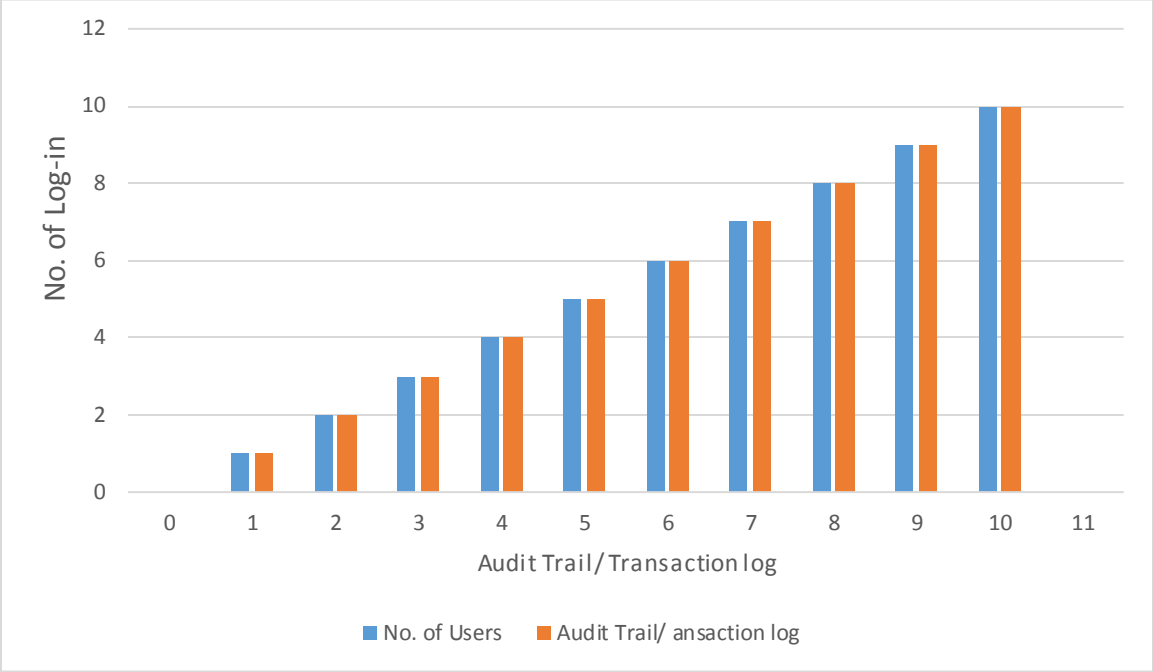

Fig. 4.8: Graph of Entries with Transaction Log/ Audit Trail (Researcher, 2021)

Table 4.2: Table of Entries with Transaction Log/Audit Trail

No. of Users	Audit Trail
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
810	10

Figure 4.8 therefore shows that for each client entered there is a corresponding transaction log/Audit Trail System that profile user activity in the system.

Table 4.3: Comparison of results of the new and previous models

S/n	Existing system	New System
1.	New registration is done by the user	Administrator registers new users with the user's details, then assign roles.
2.	In this system, no password or code to download any file	In our new model, users request for dynamic password to download any file
3.	No Audit trail/ transaction log Watermarking technique was used	There is an Audit Trail that profiles users' activities in the new system (as shown in fig. 4.5)
4.	The database is not in the cloud	Our new system database is in the cloud

CHAPTER FIVE

SUMMARY AND CONCLUSION

5.1 Summary

This research work has proposed a cloud security system and based on this concept, contributions are made in the area of authentication and authorization services for a cloud environment. Our proposed model helps various users and provides secured connection between the environments. Dynamic password is provided with the time bound for viewing the files contents, the audit trail model process with authorization and authentication. The problem has been solved and the goals have been achieved to prevent data loss with the dynamic code/password generation and efficiency of data access has been improved with the methodology (OOADM) adopted.

5.2 Conclusion

Information technology security management consists of processes to enable organizational structure and technology to safeguard an organization's IT operations and assets against internal and external threats, intentional or otherwise. These processes are developed to make sure confidentiality, integrity, and availability of IT systems. There are various aspects to the IT security in a corporation that require to be considered. These include security policies and procedures, security organization structure, IT security processes, and rules and regulations. Security policies and procedures are essential for implementing IT security management: authorizing security roles and responsibilities to numerous security personnel; setting rules for expected behavior from users and security role players; setting rules for business continuity plans; and more. The safety policy should be generally agreed to by most personnel within the organization and have support from the highest-level management. This helps in prioritization at the general organization level. The IT security processes are essentially a part of an organization's risk management processes and business continuity strategies. During a business environment marked by globalization, organizations need to remember of both national and international rules and regulations. Their information security and privacy policies must conform to those rules and regulations

Data security is not a simple task even in top leading organizations suffering fear of data leakage, this research work discussed most of the newest work administered within

the area of detection of data leakage and prevention algorithms, tools and technologies supported.

5.3 Contribution to knowledge

The contributions of this research work to ICT body of knowledge are as follows:

1. Improvement in data protection and security as unauthorized access can be easily detected if any of the users sends sensitive information to an unauthorized user or third party.
2. We successfully developed an improved Data leakage detection system that allows the organization to detect the data leakage source and reduce the leakage source by blocking it to avoid future risks.

REFERENCES

Agarwal M., Gaikwad K. G., and Inamdar V. (2012), Robust Data leakage and Email Filtering System, in IEEE International Conference on Computing, Electronics and Electrical Technologies (ICCEET), pp. 1032-1035.

Agrawal R. and Kiernan J. (2002), Watermarking Relational Databases, in 28th Int'l Conf. Very Large Data Bases (VLDB '02), Honkong, China, pp. 155-166.

Ajay K., Ankit G., Ashwani K., Navneet K. C. and Sowmya K., (2013),"Comparative Evaluation of Algorithms for Effective Data Leakage Detection", Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT). Thuckalay, India, 2013, pp. 177-182,

Ajinkya S. Y., Ravinda P. B and Shadab A. P. (2013) "Detection of Data leakage Using Unobstructive Techniques" Journal of Computer Engineering, Volume 8 Issue 4, Pp27 – 84.

Andy M. (2017), Audit Trails: Managing the Who, What, And When Of Business Transactions. Smartsheet Contributor (https://www.smartsheet.com/audit-trails-and-logs).

Archana V., Prakash L., Kiran M., Shefali K. and Nivedita P. (2012.), Data Leakage Detection, International Journal of Advances in Engineering & Technology, Vol. 3, Issue 1, pp. 315-321

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I. and Zaharia, M. (2009) Above the Clouds: A Berkeley View of Cloud Computing. Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley.

Baby and Krishnan H. (2017), "A Literature Survey on Data Leak Detection And Prevention Methods," International Journal of Advanced Research in Computer Science, vol. 8(5) pp. 2416-2418

Backes M., Grimm N., and Kate A. (2016), Data Lineage in Malicious Environments, IEEE Transactions on Dependable and Secure Computing, vol. 13, no. 2, pp. 178-191

Bhatt .C. and Sharma R. (2014), Data Leakage Detection, International Journal of Computer Science and Information Technologies, vol. 5, no. 2, pp. 2556-2558

Buneman P. and Tan W. C. (2007), Provenance in Databases, in SIGMOD ACM, Bejing, China, 2007, pp. 1171-1173.

Buneman P, Khanna S., and Tan W. C. (2001), "Why and Where: A Characterisation of Data Provenance," in International conference on database theory (ICDT), 2001, pp. 316-330.

Böhm M., Stefanie L., Christoph R.,Helmut K. of Technische Universitat, Munchen T. (2016) "Cloud Computing and Computing Evolution", p 6.

Carlin S. and Curran K. (2016), Cloud computing security, International Journal Ambient Computing and Intelligence, Volume 3 issue 1 Pp. 1-6.

Clay brook Bill (2010): Cloud vs. In-house: Where to run that App? Computer World Journal, https://www.computerworld.com/article/2520140/cloud-vs--in-house--where-to-run-that-app-.html

Cui Y. and Widom J.(2003), Lineage Tracing for General Data Warehouse Transformation, VLDB Journal, Springer-Verlag, vol. 12, no. 1, pp. 41-58.

Dahal S. (2012), Security Architecture for Cloud computing Platform, KTH, School of Information and Communication Technology (ICT).Pp24-26.

Davis Z. (2009) PC Magazine Encyclopedia, Definition of Cloud.

Ellison Larry, CEO of Oracle (2007): Analysis Conference in 2007.

Fowler A. G. and Ben W. (2009): The Internet industry Is on a Cloud – Whatever That May Mean.(ww.wsj.co/articles/SB123802623665542725, Date assessed, September, 2013.) Gordon Peter (2007), Data Leakage – Threats and Mitigation, Pp. 5-6.

Gun C. Y., Lin H. F. and Chen C. Y. (2011) A fair and dynamic password authentication system, 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Deng Feng, China, pp. 4505-4509

Hartung, F., & Kutter, M. (1999). Multimedia watermarking techniques. *Proc. IEEE, 87*, 1079-1107.

Hartung F. and Girod B.(1998), Watermarking of Uncompressed and Compressed Video, Elsevier, vol. 66, no. 3, pp. 283-301.

InfoWatch Analytics Center (2018), "A Study on Global Data Leaks in H1 2018". https://infowatch.com/report2018_half Date assessed September, 2020)

Infowatch's "Global Data Leakage Report, 2009 (www.infowatch.com Date assessed September, 2020)

Jagtap N. P, Patil S. J., and. Bhavsar A. K (2012), Implementation of data watcher in data leakage detection system, International Journal of Computer & Technology, vol. 3, no. 1, pp. 44-47.

Kadu R. S. and Gadicha V. B. (2017), Review on Seuring Data by Using Data Leakage Prevention and Detection. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 5, no. 5, pp. 731-735

Karikari A. O. Joseph K. P.,James B. H., Frimpong T. (2015), Detecting Data Leakage In Cloud Computing Environment. International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3

Karthik R., Ramkumar S., and Sundaram K. (2014), "Data Leakage Identification and Blocking Fake Agents Using Pattern Discovery Algorithm," International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, no. 9, pp. 5660-5667.

Katz G., Elovici Y., and Shapira V. (2014) A Context Based Model for Data Leakage Prevention, Information Sciences, Elsevier, vol. 262, no. 1, pp. 137-158.

Kishu G., Ashwani K. (2017) A Review on Data Leakage Detection for Secure Communication. International Journal of Engineering and Advanced Technology (IJEAT). Volume-7 Issue-1

Kumar, Goyal A., Kumar A., Chaudhary N.K., and Kamath S.S (2013), Comparative Evaluation of Algorithms for Effective Data Leakage Detection, in IEEE Conference on Information and Communication Technologies (ICT 2013), vol. 13, 2013, pp. 177-182.

Malsoru V., Naresh B. (2016) "Review On Data Leakage Detection, International Journal of Engineering Research and Applications (IJERA) Vol. 1, Issue 3, pp.1088-1091

Mell P. and Timothy G. (2011), Definition of Cloud Computing. The National Institute of Standard and Technology (NIST) Publication (Pp. 800 – 145):

Mercy C. Praba and Satyavathy G. (2017), A Technical Review on Data Leakage Detection and Prevention Approaches, Journal of Network Communications and Emerging Technologies (JNCET). Volume 7, Issue 9

Miller Lawrence C. (2009), Data Leakage for Dummies, Published by Wiley Publishing, Inc Pp. 1-10.

Mogull, R., and Securosis, L. (2007). Understanding and selecting a data loss prevention solution. Technical report, SANS Institute, 27

Monali U. P, Shraddha A. M., Snehal S. M., Siddhi N. M., Rashmi R. P. (2019) Enhancement of Data Leakage Detection Using Encryption Technique. IJSRST Volume 6 Issue 3

Noble P., Kopaee R., Melek A., and Nandy N. (2010), Data Leak Prevention, ISACA, USA, White Paper 2010.

Papadimitriou P. and Molina H. G. (2011), Data Leakage Detection, IEEE Transaction on Knowledge and Data Engineering, vol. 23, no. 1, pp. 51-63.

Patil R. and Sangve S. M (2015), Public auditing system: Improved remote data possession checking protocol for secure cloud storage. International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), *Davangere, India, 2015, pp. 75-80.*

Peneti S. and Rani B. P. (2016), Data Leakage Prevention System with Time Stamp, in International Conference on Information Communication and Embedded System (ICICES), 2016, pp. 1-3.

Plummer D., Cearley,D., and Smith , D. (2008) Cloud Computing. https://www.gartner.com/en/documents/697413

Pon P. A., Thenmozhi E. (2017) Data Leakage Detection and Data Prevention Using Algorithm. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 10, Issue 4

Prashant K., Ashish G. (2016), Data Leakage Detection and Security in cloud computing. GRD Journals- Global Research and Development Journal for Engineering Volume 1 Issue 12

Praveen S. K., Srinivas Y, Suba R. D., Ashish K. (2016), A Novel Model for Data Leakage Detection and Prevention in Distributed Environment, International Journal of Engineering and Technical Research (IJETR). pp 2454-4698, Volume-4, Issue-4,

Priyanka B., Pratibha D., Namrata K. (2013), A novel data leakage detection, International Journal of Modern Engineering Research (IJMER) Vol.3, Issue.1, pp-538-540.

Rajat V., Vipin G., Chandra P. Y., Ishu G., Ashutosh K. S. (2020); A Survey on Data Leakage Detection and Prevention. International Conference on Data Analytics and Management (ICDAM 2020) https://ssrn.com/abstract=3603736

Ramadhan M., Christian B. (2013), Information Hiding in Images Using Steganography Techniques ASEE Northeast section conference, Norwich university, reviewed paper, pp 14-16.

Raman P., Kayacık H. G., and Somayaji A. (2011), Understanding Data Leak Prevention, in 6th Annual Symposium on Information Assurance(ASIA'11), Albany, NewYork, USA, 2011, pp. 27-31.

Rechal N. and Aliyoglu S. (2012) A Survey On Data Leakage/Loss Prevention Systems (DLPs). https://doi.org/10.1007/s10586-022-03668-2

Ruanaidh J. J. K. O, Dowling W. J and Boland F. M (1996), Watermarking Digital Images for Copyright Protection, IEE Proc. - VIS. Image Signal Processing, vol. 143, no. 4, pp. 250-256.

Sandip A. K. and Kulkarni S.V. (2012) Data leakage detection, India International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 9.

Sandip. A. K. and Kulkarni S. V (2012), "Data Leakage Detection, "International Journal of Advance Research in Computer and Communication Engineering, vol. 1, no. 9, pp. 668-679.

Schneier B. (2010): Data at Rest Vs. Data In Motion. https://www.schneier.com/blog/archives/2010/06/data_at_rest_vs.html

Shabtai, A., Elovici, Y. and Rokach, L. (2012) A Survey of Data Leakage Detection and Prevention Solutions. Springer Science & Business Media, Boston. https://doi.org/10.1007/978-1-4614-2053-8

Shaj and Kaliyamurthie K. P. (2013), "A Review on Data Leakage Detection," International Journal of Computer Science and Mobile Computing, vol. 2, no. 4, pp. 577-581

Shaw J. (2013): Dynasis Blue Paper: Cloud computing Public, Private and Hybrid (www.Dynasis.com, Date assessed Nov., 2013)

Shobana V. and Shamugasundaram (2013): Data Leakage Detection Using Cloud computing (www.ijetae.com Volume 3, special edition, January 2013 Pp. 111)

Shu X. and Yao D. (2015), Privacy-Preserving Detection of Sensitive Data Exposure, IEEE Transactions on Information forensics and Security, vol. 10, no. 5, pp. 1092-1103

Shu X., Zhang J., Yao D., and Feng W. C. (2016), Fast Detection of Transformed Data Leaks, IEEE Transactions on Information Forensics and Security, vol. 11, no. 3, pp. 528-542.

Singleton T. (2010): The Minimum IT Controls to Assess in a Financial Audit (Part II), ISACA Journal, vol. 2, Issue. 3

Sion R., M. Atallah, and S. Prabhakar (2003), •\Rights Protection for Relational Data, Proc. ACM SIGMOD, pp. 98-109, 2003.

Sodagudi S. and Kurra R. R. (2016), "An Approach to Identify Data Leakage in Secure Communication," in 2nd International Conference on Intelligent Computing and Applications, vol. 467 pp. 31-43.

Stedum J. (2013) A Brief History of Cloud Computing, January, 2021 (Posted on Soft Layer Blog, Pp. 1)

Sultan, E. S., and Muthukkumarasamy V. (2016) A Survey on Data Leakage Prevention Systems. Elsevier Journal of Network and Compute Applications, vol. 62, no. 1, pp. 137-152.

Sumedha K. and Ankur S. (2019), Network security using cryptographic techniques, International Journal of Scientific Research in Science and Technology volume 2, issue 12,

Sun Microsystems 2009, Cloud Computing architecture. https://www.slideshare.net/danielfc/cloud-computing-sun-microsystems

Sushilkumar N. H., Ulhas B. S. and Archana U. B. (2015). The Guilt Detection Approach in Data Leakage Detection. International Journal of Computer Applications. Volume 119 – No.8, pp 0975 – 8887

Sweeney L. (2002). Achieving K-Anonymity Privacy Protection Using Generalization and Suppression, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), pp 571-588.

Tim M., Subra K. and Shahed L. (2009): Cloud Security and Privacy- an enterprise perspective on risks and compliance, Pp. 1-30.

Tuscano G., Kotadiya H., Bhat V., Fernandes R, and Pancha A. (2015), A Survey on Data Leakage Detection. International Journal of Engineering Research and Applications, vol. 5, no. 4, pp. 153-158.

Vasquero L. M., Luis R., Juan C. and Maik L. (2009): A Break in the Clouds: Towards a Cloud Definition. ACM SIGCOMM Computer Communication Review, Vol 39, No 1 Pp. 51.

Wikipedia: What is cloud computing? https://en.wikipedia.org/wiki/Cloud_computing (Date assessed January, 2020).

Yadav G. B., Bhaskar P. C., Kamat R.K (2012), Assessing the Guilt Probability in Intentional Data Leakage, International Journal of Computer Science and Information Technologies. Vol. 3 (3), 4075-4078

Yin F., Wang L., Yu Rongwei, Ma Xiaoyan (2013) A Distribution model for Data Leakage Prevention, IEEE International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC 2013), Shenyang, China. 978-1-4799-2565-0/13

Youseff, L, Butrico. M. and Da-Silva D. (2008). Toward a Unified Ontology of Cloud Computing, In Grid Computing Environments Workshop,

Yunchuan S., Junsheng Z., Yongping X. and Guangyu Z. (2014), Data Security and Privacy in Cloud Computing. International Journal of Distributed Sensor Networks. Vol.10 (7).