# SECURE AUDITING AND DEDUPLICATING DATA IN CLOUD

# ABSTRACT

Cloud computing has emerged as a computing paradigm to solve large scale problems in every sector like industry, e-commerce. Cloud computing provides the facility of on-demand resource provisioning and data storage too. Datacenter is the core component of cloud computing environment. An industry, academic users have started to store their data on cloud, there is a need to consider data management issues in cloud. As cloud storage systems are gaining importance as datacenter is the core-component. In order to provide reliable, secure data on time, there is need to focus data integrity, data deduplication, data reliability. An issue in cloud data management has become an important research in cloud computing.

Cloud computing made a greater impact on the storage systems from last decades. It allows the users to store huge amount of data in a platform where maintenance of data is easier. This technology made a greater impact on the storage management and maintenance. Since the data stored in cloud is not so safe it raises some security concerns. One of the concern is about the duplicate data storage in cloud while achieving integrity auditing. The aim of this project is to achieve integrity auditing and deduplication. Two systems are used namely SecCloud and SecCloud+. SecCloud system achieves data deduplcation by attaching a data tag along with the file need to be transmitted. This data tag helps to check the integrity of the transmitted data also the SecCloud system produce an entity which is capable of performing auditing job on the incoming data and maintains a MapReduce cloud for the duplicated data. Since more number of users is accessing the same copy of file, the bandwidth required to transmit data is saved greatly as well as disk space is also saved. The SecCloud+ system is the evolution of previous system, where cryptographic technology is applied to each data in the system. The system allows the user to encrypt their data before uploading to the cloud and decrypt the data before downloading it from the cloud. This secure technique helps the client to protect the data from the intruders especially from the cloud server which can cause an attack known as dictionary attack. This deduplication process increases the storage utilization by storing only a single copy of encrypted data in cloud. Also it maintains confidentiality of the stored data by cyphering and deciphering it.

i

**CHAPTER 1**

# INTRODUCTION

## Cloud Computing: An overview

A.Moiz Qyser et al [21] defined Cloud computing as a distributed computing paradigm, which allows the users to make use of computer resources in an efficient way. Its low cost and on demand service providing technique renders many developing company’s to explore their ideas. The main aim of this technique is to outsource the client’s data to the cloud server [27]. In traditional data storage system it is not possible to keep the huge amount of data for a long time, because it requires high investment in purchasing and maintaining the disk systems [22]. Through this technique anybody within the organization can have access to the cloud data easily. This technology is mainly useful in company’s where people need to interact with the other who are residing across the globe.

### Characteristics of cloud computing

The characteristics of cloud computing are mentioned below:

* + - * **On demand capability**: Cloud users can have the access to the cloud resources whenever they require it.
      * **Broad network access**: By allowing the clients to store their data, cloud can connect more people which will help them to establish a new business strategies.
      * **Resource pooling**: Cloud allows the client to utilize its services anywhere at any point of time. This will greatly reduce the space required to keep the data as well as the cost [42].
      * **Rapid elasticity**: Any changes required by the client has to be done very quickly without waiting for any environmental setup. This characteristic is known as rapid elasticity.
      * **Measured service**: Cloud client has to pay only for the services what he had used before. No other extra maintenance charge will be issued to client [42].

### Types of clouds

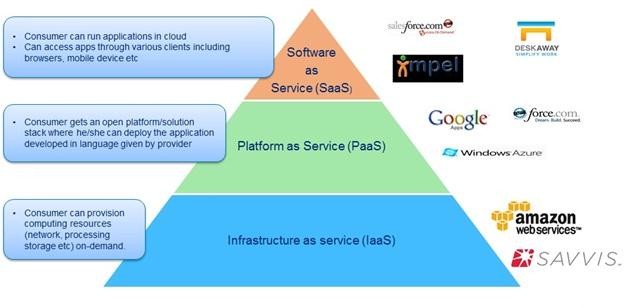
There are mainly three types of clouds are depicted as below:

* + - * **Public cloud**: This type of cloud is maintained and managed by entity known as third parties. These third parties provide services to all kind of users around the global. Since it is easily accessible by any cloud user, there exists lot of issues regarding the security.
      * **Private clouds**: Private clouds are managed by a particular personnel or an organization. Here the services provided by the cloud are restricted to only authorized user of the cloud.
      * **Hybrid clouds**: It is a combination of both private as well as public clouds. It is inherited the feature of both private and public cloud.

### Cloud deployment models

Cloud offers some services to the user, which are categorized into three types as follows [39]:

* + - * **Software as a service (SaaS)**: Here the cloud provides the different kinds of softwares to the cloud users. Instead of buying the costlier softwares, the cloud server lease some softwares for a particular time period to the users.
      * **Platform as a service (Paas)**: Here instead of providing a software, a set of softwares are provided as a complete platform to develop the entire component of a system. For example the platform such as java, C++ are made available to the cloud users for their further utilization.
      * **Infrastructure as a service (Iaas)**: In Iaas, cloud providers do not lease software resources, they also give hardware resources on lease [39].



**Fig 1.1 Cloud deployment models [39]**

## Cloud computing key areas

Cloud computing areas are described as below:

* **Resource management:** Cloud providers offer various kinds resources to cloud clients in order to fulfil the need of clients. These resources include virtual machines, computer hardware devices, network devices, firewalls. Resource management issues occur due to heterogeneity of resource types and their inter dependencies and also due to unpredictable loads [23]. The problem of resource management includes assigning resources to cloud clients, resource provisioning, resource adapting. The most common problem in resource allocation is to assign virtual machine to physical machines which are resided in the cloud data centers. This problem is known as virtual machine placement.
* **Data management:** Cloud allows multiple users to store their data on a single platform. Since huge amount of data are transferred between the client and the server there is a need of efficient techniques to handle the collected information. There occur many issues in handling the data. Some of them are explained and resolved in this work. Such problems include data duplication and data integrity.. Since data is stored at an untrusted host there occur an issue of data integrity [32].
* **Security:** The cloud users outsource their data to some third parties in order to communicate with other clients efficiently. Since data is transmitted through the internet, there is a risk in transmitting confidential data. Attackers exists inside and outside the cloud centre. Security concerns must be taken in order to protect the confidential data from the attackers. Some of the concerns such as cryptographic technique is used in this work.

## Cloud storage

Cloud storage is a space where user can store huge amount of data for further use. The cloud space is attracting more clients day by day because of high availability of data as well as other software tools, software platforms and also low cost required to store huge amount of data [29]. It is estimated that the usage of cloud storage would reach the forty trillion gigabytes in 2020 [25][43].

### Key terms in cloud storage:

* + - * **Confidentiality:** Data can be attacked by the intruders in and outside the cloud. A technique called cryptography is used to maintain the data confidentiality.
      * **Integrity**: It is necessary to maintain the correctness of data throughout its transmission across the network [32].
      * **Availability**: When a data is transferred across the network , it has to be made available to the end user. Necessary precaution must be taken so that the data should not be stolen by any of the middle man [23].

## Project Objective

The main objective of this project are mentioned as follows:

* **Integrity Auditing**: Cloud server can allowed to access the uploaded file without retrieving or downloading it. Since the client sends the data after encryption, no other user can access it including the cloud server. This maintains the originality of data.
* **Secure Deduplication:** Deduplication in the sense keeping only a single copy of data in cloud. Even though the multiple users keep the same data in cloud, all the files are not uploaded to cloud, instead only one copy is saved. This reduces the space required to store multiple file, thereby saving the disk space which is then reserved for other files.
* **Price Effective**: Since the deduplication technique is being implemented in the proposed system, the disk space of a server is greatly saved. This results in minimum requirement of disk systems in the cloud storage. This is how cost is reduced efficiently.

## Project Organization

The organization of the project report is explained as follows. The project report is designed with eight chapters, references and appendices. Introductory part is explained in Chapter 1. Chapter 2 gives information regarding the existing literature survey in cloud storage area. Chapter 3 provides details of different types of requirements needed to complete the project. Chapter 4 deals with the analysis about the system. Chapter 5 includes some basic design concepts. Chapter 6 is based on the implementation techniques used in the project. Chapter 7 includes several kinds of testing such as unit testing, functional testing. Chapter 8 presents conclusion and future work.

**CHAPTER 2**

# LITERATURE SURVEY

M. Bellare [5] et al. introduced a system, where multiple clients share some secret keys known as key server which are then used to derive messages in DupLESS system. After sharing the keys, clients interact with each other using the key server with the help of a key generation protocol. These keys are further used by the clients to perform cryptography. This technique involves providing security to the data by encrypting and decrypting it. This technique ensures that DupLESS provides more security against the attackers in the system.

M. Bellare [6] et al. formalized a system based on the cryptographic technique known as Message-Locked Encryption (MLE). In this protocol the keys which are used to perform encryption and decryption are generated by the message itself. This has to be done before the transmission of message across the internet. The main goal of this approach is to achieve secure auditing

G. Neven [7] et al. proposed many identification and signature schemes which provide protection against the security attacks. These frameworks are defined in the existing literature survey. These frameworks help the user to analyze the security issues and also explains how these techniques are evolved. The knowledge of these information will lead to understand the framework easily.

S. Bugiel [9] et al. introduced a model and protocols which collects some of the computations regarding the security and make use of those computations whenever a client wants to query them. At first user interacts with the trusted cloud for resources. The cloud which uses encryption and decryption technique is known as trusted cloud. Later the user interacts with untrusted cloud.

J. li [11] et al. proposed Dekey protocol for managing the secure keys which are required for deduplication purpose. The Dekey protocol shares the secure keys among all key servers. Since the secure keys are kept in servers, it is not possible for the attackers to get those keys. Thereby providing more security to the data, which results in confidentiality of the data is well maintained.

C. Ng [13] et al. introduced RevDedup protocol, which is applicable only in the virtualization environments. There are some aims of this protocol. Firstly to give efficient

storage capability. Secondly, to provide backup for every data stored in cloud and lastly, utilizing least possible memory to store the data. The design goal of this protocol is to achieve deduplication. The existing duplicate files are removed by their work and newly encountered data are checked for the duplication.

J. Stanek[14] et al. proposed an encryption scheme which can be applicable for two kinds data, popular data and unpopular data. They proposed that for popular data weaker security semantics is provided whereas in case of unpopular data high security is provided. In this way, the process of deduplication is done easily for popular data because anybody can view the content of the file. Deduplication is bottleneck for unpopular data, thus preventing it from deduplication.

K. Zhang[19] et al. explained many commercial cloud service providers allow their customers to store their information at a low cost. But most of the time these commercial service providers do not follow security aspects whenever providing services at low cost. Since these customers cannot store their sensitive data in public cloud without any protection to the data, cloud storage requires additional security. It is not possible to expect more security from the traditional system. One solution of this problem is to share the content of cloud with some commercial outsourcing system by keeping rest of the data in private cloud itself. This framework is known as hybrid storage system where popular data is kept in private cloud and unpopular data is kept in public cloud. A new system Sedic is developed which consists of special characteristics of MapReduce cloud. The data must be scheduled according to the planned data management, which comprises of many tasks need to be performed. These tasks are scheduled based on the popularity of the data . In private cloud, each task is performed on the most secured data and the task within the public cloud works on the least secured data. At the end unwanted data are moved to the commercial cloud system.

J. Xu [20] et al. proposed the intended needs of cloud storage and the properties of cryptographic technique to evolve a new technology to outsource the data and balancing security in it. In cryptographic technique there are several types are evolved such as symmetric and asymmetric encryption. Symmetric encryption make use of small data sets whereas asymmetric encryption make use of meta data sets to perform cyphering of data. Along with cryptographic technique, additional properties such as Merkle tree property has been used to search for a duplicated files in the cloud server prior to the data uploading process. The bandwidth required

to save the data on cloud is greatly reduced because of identifying duplicate files using Merkle property. This property also do not permit the unauthorized access to cloud resources by providing two levels of verification.

**CHAPTER 3**

# SYSTEM REQUIREMENTS

## Software Specification

* + - Operating System : Windows 7
    - Technology : Java
    - Web Technologies : Html, JavaScript, CSS
    - IDE : My Eclipse
    - Web Server : Tomcat
    - Database : My SQL
    - Java Version : J2SDK1.7

## Hardware Specifications

* + - Hardware : Dual Core Processor
    - Speed : 2.80 GHz
    - RAM : 1GB
    - Hard Disk : 2 GB
    - Key Board : Standard Windows Keyboard
    - Mouse : Three Button Mouse
    - Monitor : SVGA

## Functional Requirements

In system specification, the activities of the system is depicted as its functional requirement. A function is explained as a set of inputs, the characteristics of entities, and outputs.

Different modules used in this system are given below:

* + - **User Login**: Cloud client can have access to the cloud by providing his username and password.
    - **Auditor Login**: Auditor is the person who is having the authority to upload, download and update the file onto the cloud. Only authorized person can do these things by providing his username and password.
    - **File Upload**: User can upload his file onto the cloud only after it is being encrypted by using the encryption key.
    - **File Download**: User can download files by providing decryption key from cloud only after it is being checked by the auditor for duplication.
    - **File Mapping**: Once the file is uploaded to the cloud successfully, it is then checked for the duplication. If the file is not duplicated, then it has to be taken for the further activities by the user, otherwise the duplicate file is mapped with its original file in cloud.

## CHAPTER 4

* 1. **Existing System**

# SYSTEM ANALYSIS

The cloud storage requires lot of data management and maintenance which is done by the cloud servers. . The main difference between the cloud storage and traditional data center storage is that data is taken to the cloud through internet, so there is a possibility of attacking the data by the intruders, where as in case of traditional cloud storage data is resided in the system itself, so attacking data is not possible. It is difficult to maintain data integrity in cloud storage. Data threats can occur both inside the cloud as well as outside the cloud[1]. Cloud server can hide the data loss from the cloud clients in order to maintain their identity. The rarely accessed files or data can be discarded by the cloud server which results in data loss.

The demand for cloud storage has greatly increased when clients want to store their data remotely by any local server. Since there is a possibility of storing same data by multiple user, duplication ratio increases day by day. The company called EMC had done a survey on the duplicate files stored in cloud, around 75% of the cloud data is duplicated one. Increase in duplicated files will be bottleneck to the cloud storage[3][2]. A client can’t directly get the information about others data in server, so he will insist the server to tell the information about the duplicated files, thereby he will get the confidential data from the server [15].

## Limitations of Existing system

* + - Auditing huge amount of data for duplication is very difficult in traditional cloud storage system and also it requires lot of time for computation.
    - Using integrity auditing, it is very difficult to audit large amount of data in traditional cloud storage.

## Current System

In this project there are mainly two systems are defined: SecCloud as well as SecCloud+ systems. When a user uploads a file onto the cloud, he cannot prevent the server from reading the content of file. In Seccloud system user cannot provide more security to the cloud data. To

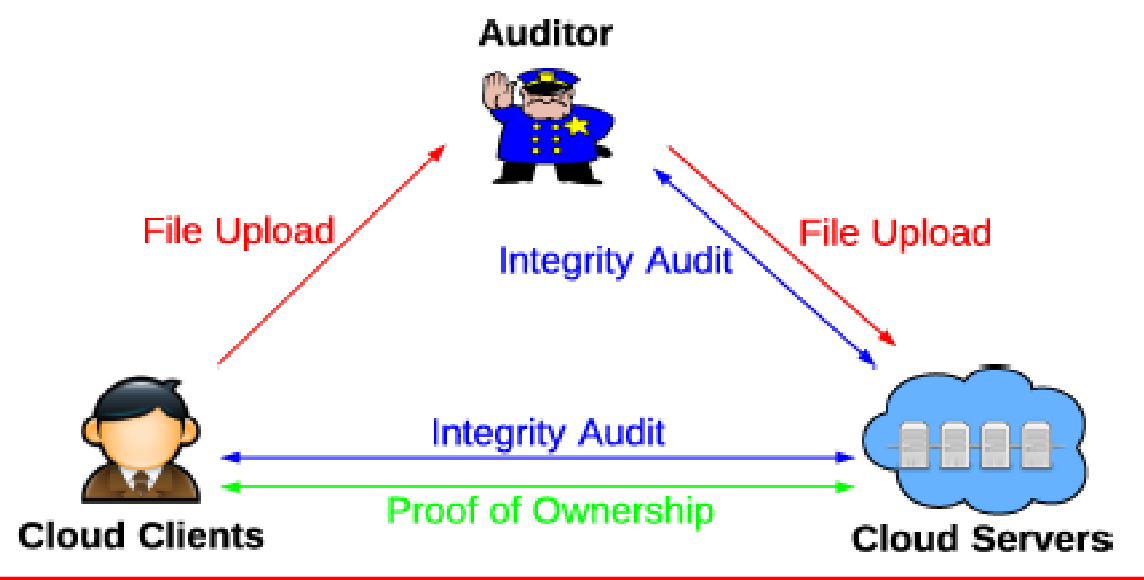
overcome this disadvantage another system known as SecCloud+ is used. In this system, before uploading the data, it has to be encrypted by the client, thereby providing security to the data. In this way secure auditing is done.

Cloud servers exhibit the cloud information to the clients whenever a client wants to have an access to it. The another main purpose of this project is to achieve the data deduplication. Usually cloud clients buy or lease a space from the server and server allocates required storage space to the client for his computation. There exist another entity in the system, the auditor.

Auditor is the one who has an authority to check each and every content of a cloud. He will be given a unique use rid and password, using which he can view the content of a file, update it and manipulate it whenever it is required. Auditor maintains a MapReduce cloud which is used to map the duplicate files with the existing one in cloud. Using this auditor can identify the duplicated files in cloud.

Another design goal of this project is to reduce the cost required to store the data in cloud. This is done by keeping only one copy of a file in cloud by identifying duplicate files. This is the responsibility of auditor. Since disk space is saved, server can store all the data request sent by the clients. This increases the effectiveness of the system in terms of the cost. This is how cost required to save the data is reduced.

## System Model



#### Fig 4.1 System Model [12]

The three entities used in the system are depicted below:

1. Cloud clients are the one who have large data to store on the cloud. Usually these clients include individual users of the cloud or any organization which is intended to use different software’s stored on the cloud.
2. Cloud servers are the one who can store the data of a client in his system. Usually client purchases some space on the servers system, or the server can lease the space for a particular time period. The data stored on the servers system is secured by providing an appropriate security technique such as encryption and decryption.
3. Auditor is an entity which helps the client to upload the file by performing appropriate auditing of the file. Auditor has many job to perform, he can audit the file for duplication, if the file is duplicated one then such file is not uploaded to cloud, instead auditor maintains a separate MapReduce cloud which maps the duplicate files with its original file. Only non-duplicate files are directly uploaded to cloud by the auditor.

The following protocols described below represent the file level deduplication shown in three colors red, blue and green colors as depicted in fig 4.1.

**File Uploading Protocol:** This protocol describes how data is uploaded to cloud by the client. The process of file uploading is done in three steps, each step is explained as below:

* + - Step 1: Cloud client communicate with the cloud servers for duplicated files by sending its identity to the server before uploading data. Server replies to the client with appropriate message. If the user is an unauthorized one, then proof of ownership protocol is executed between client and server. Otherwise step 2 and step 3 are run between these two entities.
    - Step 2: Client sends file along with its tag to the auditor and receives a reply from the auditor.
    - Step 3: Auditor verifies the tags attached the file to check for the duplication and sends the file to the server for uploading after verification.

**Integrity Auditing Protocol:** This protocol is aimed at achieving data integrity. Integrity of the data is verified and this process is initiated by either cloud client or an auditor. In this protocol, the cloud server becomes the prover, while the auditor or client works as the verifier. This protocol includes two steps:

* + - Step 1: Client generates set of tags and sends them to the auditor for integrity verification.
    - Step 2: Upon receiving the file, auditor tries to verify for its integrity and notify to client if any malfunction is done. The verifier returns true if the integrity of data is passed otherwise returns false.

**Proof of Ownership Protocol:** The protocol is developed to check the authority of the client and the verification is done by the cloud server. This protocol is typically called along with file uploading protocol to avoid the loss of information. This protocol also includes two steps.

* + - Step 1: Cloud server creates a set of challenges and sends them to the client.
    - Step 2: The client replies with the proof for file ownership, and cloud server finally checks the correctness of proof.

## Advantages of Current System

* + - Since cloud users send only encrypted file to the cloud, confidentiality of a file is maintained effectively.
    - Since the duplicated files are mapped with the existing files in cloud, disk space is greatly saved, which in turn increases effectiveness in cost management.

## Feasibility Study

The feasibility study involves analyzing the project to determine its functionalities for the prepared plan and to estimate the cost of it. While system analysis phase this feasibility study is to be performed. The developing system should not be bottleneck to the purchaser, that is the reason why we do this study during analysis phase itself. An analyzer should understand the requirements of a proposed system before started designing it. There are different varieties of feasibility study depending on some criteria.

Three types of feasibility studies are mentioned below:

* + - Economical feasibility
    - Technical feasibility
    - Social feasibility

### Economical feasibility

This study gives the influence of the defined system in the economic development of a buying organization. The organization cannot afford the cost of any developing system beyond its limit. The cost required to develop a component must be clearly mentioned by the developer. This leads to the cost required to develop a system will not be bottleneck to the developer, because now a days due to the modernization of technology some of the software’s are available for free of cost and only the hardware components are need to be purchased by the developer.

### Technical Feasibility

The study of technical feasibility gives the description about the technical requirements of the system. This technical requirements involves availability of software resources, hardware components etc. Since the newly emerging system should possess modern technology, the demand for these resources are been greatly increased. The organization should provide the necessary software’s and tools required to implement the proposed system.

### Social Feasibility

Once the system is developed it has to be taken to the user for further usage. The user of the system should know the advantages and disadvantages of the system in order to operate it. This study yields the percentage of acceptance by the user. If the user fails to operate it then it is the responsibility of the developer to give sufficient training to the user of the system. The user must have an access to the system without any fear in him. The main goal of the system development is to encourage the purchaser to buy the product by making them feel free to use the system.

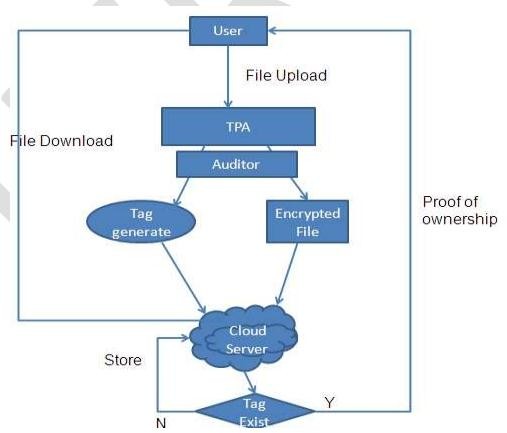
**CHAPTER 5**

# SYSTEM DESIGN

The designing of a system involves applying different methods and techniques to obtain a blue print of a desired system. Various methods can be applied to get solution for each module in the system and finally these solutions are integrated to obtain the solutions for the desired system. Design features include different components of the system such as different entities of the system and the interaction between these entities.

## System Architecture

Any architecture of a system includes the different entities used in the system, their properties and interaction between entities in the system. This architecture is depicted in a structural way, which gives the all possible ways of getting the required solution for the problem as shown in figure 5.1.



#### Fig 5.1 System architecture [12]

Fig 5.1 refers the architecture diagram of the system. It includes entities such as user, cloud server and auditor.

The architecture involves three main entities:

* + 1. Cloud Clients
    2. Cloud Server
    3. Auditor
* Cloud clients are the one who have large data to store on the cloud. The clients refers to individual users of the cloud or any organization which is intended to use different software’s stored on the cloud. In this project client has to cipher the data before it is being uploaded and also decipher it before the user can download it by using the appropriate keys generated by the user.
* Cloud servers are the one who can store the data of a client in his system. The server can lease the space for a particular time period. The data stored on the servers system is secured by providing an appropriate security techniques such as encryption and decryption.
* Auditor is an entity which helps the client to upload the file by performing appropriate auditing of the file. Auditor has many job to perform, he can audit the file for duplication, if the file is duplicated one then such file is not uploaded to cloud, instead auditor maintains a separate MapReduce cloud which maps the duplicate files with its original file. Only non-duplicate files are directly uploaded to cloud by the auditor.

## Data Flow Diagrams

The entire system can be represented with the help of a flowchart known as the DFD (also called as the bubble chart). This chart is written for different entities of the system separately, which gives us a detailed information about a particular entity. The flowchart specifies various tasks and functionalities performed by the entity. This technique of representing functionalities of each entity has become widely used because it is a simple technique, which does not involve complex activities.

#### Activities of User:

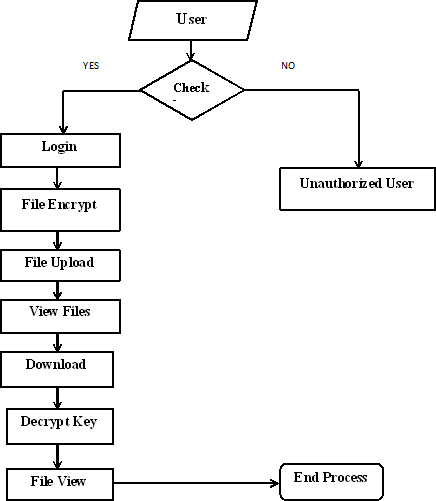
Initially the user is checked for his authorization by the auditor. If the user is not an authorized user, then such user is not be considered for further processing. If the user is an authorized user then user can upload file on to the cloud. Each user is allocated a unique username and password by the server. Using that a user can log on to the system. In SecCloud system, once the login is

done successfully, user can upload file onto the cloud. But in SecCloud+ system user has to encrypt the file using encryption key created by the user, before sending the file. Once the file is uploaded user can download it by using decryption key sent to user’s mail and later user will decrypt the file before viewing it.

#### Activities of auditor:

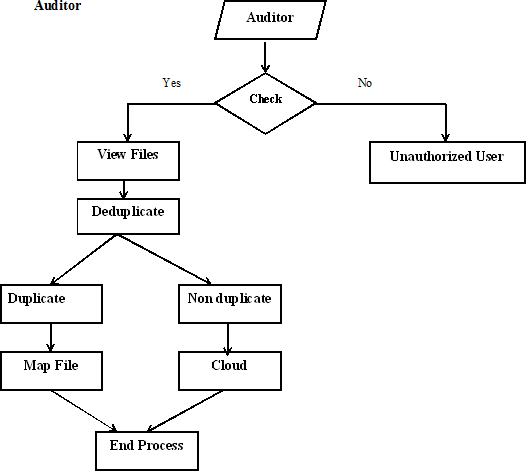
Once the user sends the encrypted file to the cloud, it is not actually uploaded to cloud. Rather than it is forwarded to auditor for checking the uniqueness of it. Auditor maintains a MapReduce cloud to map the duplicate files. If the received file is duplicated one, auditor maps the file with its original file , otherwise auditor will directly send the non-duplicate file to the cloud.

## User:



#### Fig 5.2 Data flow diagram for User

Fig 5.2 depicts the data flow diagram for user. Each cloud user is checked for authorization. The authorized user is allowed to upload file to the cloud, download data from the cloud and also view the files in cloud. The authorized user can be able to do file encryption and decryption. The unauthorized user cannot perform any of these operations.



#### Fig 5.3 Data flow diagram for Auditor

Fig 5.3 depicts data flow diagram for auditor. The authorized auditor can view the content of cloud and also checks for the duplicated files. The duplicated files are mapped with the existing one and non-duplicate files are uploaded to cloud by the auditor.

## CHAPTER 6

* 1. **Configuration**

# IMPLEMENTATION

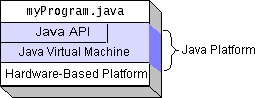
Implementation is a phase where the developed designs are put into action by the developer. This stage is most important in the system development life cycle because the aim of this phase is to get the most desirable successful system as predicted by the purchaser. This phase involves investigating existing system, identifying drawbacks of previous system. By these information the system developer plans for the proposed system and acts according to that to achieve changeover system.

### Selection of platforms

#### The Java Platform

The java platform consists of four main components:

* + - * Java programming language
      * Java class file format
      * Java application programming interface (Java API)
      * Java virtual machine (JVM)



#### Fig 6.1 Java platform

Fig 6.1 describes the components of java. It is depicted as four layers.

The following steps are used to write, compile and interpret any java program:

* + - * Firstly we have to make sure that the code must be written in java programming language.
      * After the compilation of java program the compiler generates a class file which consists of byte codes.
      * The obtained class file is always run on the JVM.
      * During the execution of program, the file makes a call to the functionss used in java API.

#### Java virtual machine:

Java programs are not directly translated into machine instructions, rather than they are converted into intermediate language known as byte codes by the compiler. This code does not exists physically. Therefore this machine is known as Java Virtual Machine (JVM). This machine exists only inside the computer memory.



#### Fig 6.2 Process of Compilation



**Fig 6.3 Process of Execution Application Programming Interface:**

The java standard library includes huge class files and functions which are collectively called as application packages.

The packages used in the program are described below:

* + - * Programming Language support packages: This includes different classes and functions required to implement basic characteristics of java.
      * Utilities packages: The methods which are used to indicate date and time functions are defined in this.
      * Input/ Output packages: The packages reserved for input and output operations belong to this category.
      * Networking packages: The collection of packages used to communicate with other components via internet.
      * AWT packages: This includes packages which consists of classes that implement platform independent graphical user interface.
      * Applet packages: It consists of classes that support java applet.

#### Introduction to SQL:

The SQL (Structured Query Language) is used to access, store and change the data. In this project SQL is required to store users data on cloud.

The following commands are used in SQL:

* + - * Create: The command used to create a table in database is given below: CREATE TABLE tablename (list of values);
      * Insert: Insert command is used to add rows in a table. INSERT INTO tablename VALUES(list of values);
      * Select: This command is used to get the information in the table. SELECT \*FROM tablename;
      * Update: The purpose of this command is to alter the values in the columns. UPDATE tablename SET column name = value WHERE condition.
      * Delete: The command used to delete table is given below: DELETE FROM tablename WHERE condition.

#### Database connectivity:

An application should possess the following information when it needs to be connected to database.

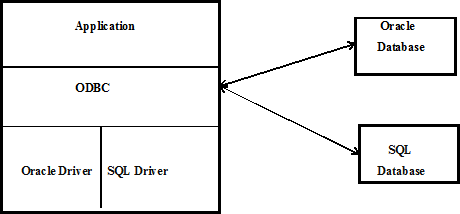
* + - * The RDBMS product is used.
      * Location of the database.
      * Name of the database.

The JDBC calls are used to retrieve or update data from database. It consists of following programs:

* + - * Opening and connecting to database.
      * Send SQL queries.
      * Process the obtained results.
      * Terminate the database connection.

#### ODBC API:

ODBC (Open Data Base Connectivity) is an API used to access RDBMS.ODBC is entirely written in C language. This makes a call to database by sending SQL statements. The following figure shows the design features of ODBC.

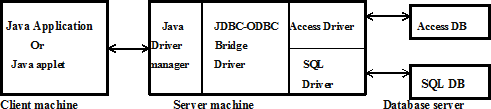


**Fig 6.4 ODBC database design**

**JDBC API:**

Connecting java programs to database needs the APIs of java programs. Java programs cannot directly interact with the ODBC driver. A company called Sun microsystem developed an interface which converts JDBC to ODBC. Several JDBC drivers are listed below:

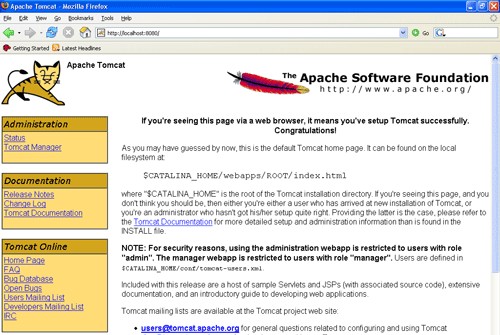
* + - * JDBC-ODBC bridge + ODBC driver
      * API java installer
      * JDBC-Net java installer
      * Native protocol java installer



**Fig 6.5 JDBC Application database**

**Tomcat 6.0 Web Server**

Tomcat is a kind of website server prepared by the apache software foundation. This server is mainly designed to implement java servlets and java application packages. Apache server supports only web pages and other application servers support both web and business components. Any application written in java or C# languages use this tomcat server for execution.



**Fig 6.6 Tomcat Webserver**

### Selection of Programming Language

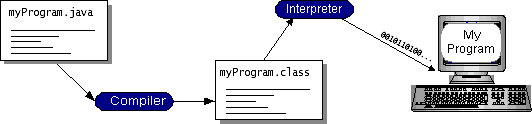
Some of the features of java programming language are given below:

* + - * **Simple**: The reason why this programming language is called simple because it can be learnt even by the beginner very easily.
      * **Object Oriented language**: The entire structure of the java program is written inside a class. Also it consists of object, class.
      * **Distributed**: Java applications and applets can access data from the web through the resources as they can access a local file system.
      * **Interpreted**: In java program errors are rectified using the compiler known as javac. Java programs are translated into an intermediate language called java byte code (.class file).
      * **Robust**: Java is more reliable language. It does not make use of entire memory of the system and also it handles exceptions very well. So it is less addicted to errors.
      * **Secure**: Java contains inbuilt security aspects. A user can access the java code from the internet thinking that the downloaded code cannot attack his system.
      * **Architecture neural**: Java program produces a byte code, which can be executed on different operating systems such as windows xp, windows 7 ,linux etc.
      * **Portable**: Java programs can run on any platforms. The byte codes can be loaded on clients machine by using a special softwares known as Java Virtual Machines(JVM).
      * **High performance**: Java byte codes are designed such that it would be easy to translate directly into clients machine using a compiler Just-in-time(JIT).
      * **Multithreaded**: Multithreading means ability of an application to perform multiple task at as time. This feature is implemented in java.
      * **Dynamic**: It is very easy to maintain the different versions of an applications in java.

#### Building the java program:

The following four points need to be remembered when we write a java program:

* + - * Java is sensitive to case.
      * Each java program must have its name as that of its class name.
      * Every statement ends with semicolon in java program.
      * Every program has a function called main.



#### Fig 6.7 Compiling Java program Compiling and Running Java programs:

* + - * Initially the java source code must be compiled in a compiler known as javac before executing it.
      * After compilation, the compiler generates a file with .class extension which contains the byte code of the program.
      * This generated class file is fed to the interpreter to run the application.
      * The interpreter interprets the byte code into machine code.

**Java Program**

**Compilers**

**My Program**

**Interpreter**

**Fig 6.8 Running Java Program**

**CHAPTER 7**

# SYSTEM TESTING

The goal of testing is to detect faults in the developed system. Testing is the process of attempting to find each identified errors in an exceedingly work product. It gives the clear idea about various parts, assembly units of the system. It is a main phase in the software development life cycle which results expectable, desirable system unit developed for the user. There are verities of testing demands which gives the different views of a system.

## TYPES OF TESTS

### Unit testing

Unit checking refers to writing the test cases for different entities in the system which ensures that the built components works according to the designed plan. It also involves testing the different software components before the integration of each unit. Unit checking is done at component level and checks for a unique business methods, different applications, and system configuration.

### Integration testing

Integration checks are developed to check combined software system which is obtained by combining all software units with expected results. Testing mainly involved with outcome of every units. Though the outcome of unit testing is satisfactory, the aim of this testing is to get correct, specific and desired result. Integration testing mainly focus on the issues that arise while combining the outcomes of different parts.

### Functional testing

Functional testing is based on the following items:

Correct Input : Detected varieties of correct input should be taken. Incorrect Input : Detected varieties of incorrect input should not be accepted. Functions : Detected functions should be worked out.

Output : Detected varieties of application outputs should be worked out. Systems/Procedures : Systems or procedures should be called.

.

### System Testing

System testing involves checking whether the whole combined software system satisfies the requirements or not. It mainly checks for the configuration method used in the system. A system can be built using more than one configuration, but some of them yield better performance. System testing depends on the system description, system flow, varieties of platforms used to implement and test inputs.

### White Box Testing

White box testing is used to test the software based on its inner working. The tester checks the software for the parameters such as its versions, validity, the language used or at least its purpose. The requirements that are not fulfilled by black box testing are focused here.

### Black Box Testing

Black box testing is done on the software without knowing the parameters such as its inner work, structure and module. Black box testing is done on some specification by the system developer. The reason why this test is called as black box is that here the software to be tested is known as black box and tester cannot look into the black box while testing it.

## Unit Testing

Unit testing is usually done for each component of software development lifecycle and later the result of each component is integrated to get the expected result. The testing is done with the help of test case data. The valid input is given to each module and later test result is compared with the expected result. If both the results matches then system is said to be working properly.

#### Test strategy and approach

Unit testing is written by considering possible test cases but functional testing is described briefly.

* + 1. **Test objectives**
       - The entered input must be acceptable input.
       - There should not be any delay in displaying entry screen, request and response messages.

### Features to be tested

* + - * To verify that the entered data are valid.
      * There should not be any duplicate entries.
      * The result must be desired one.

#### SUT 1 :- Test case for User Login:

|  |  |
| --- | --- |
| **Sl. No of Test cases** | **Test case to verify** |
| Name of test | User Login |
| Feature being tested | Authentication |
| Description | User has to provide ID & password to enter into the cloud |
| Sample I/P | Username & Password |
| Expected O/P | Login is successful |
| Actual O/P | Output as expected |
| Remarks | User login is successful |

**Table 7.1 Test case for User Login SUT 2 :- Test case for Auditor Login:**

|  |  |
| --- | --- |
| **Sl. No of Test cases** | **Test case to verify** |
| Name of test | Auditor login |
| Feature being tested | Authentication |
| Description | Auditor has to provide ID & password to enter into the cloud |
| Sample I/P | Username & Password |
| Expected O/P | Login is successful |
| Actual O/P | Output as expected |
| Remarks | Login is successful |

#### Table 7.2 Test case for Auditor Login SUT 3 :- Test case for File Upload:

|  |  |
| --- | --- |
| **Sl. No of Test cases** | **Test case to verify** |
| Name of test | File upload |
| Feature being tested | Integrity Auditing |
| Description | User can upload a file only after it is audited by the auditor. |

|  |  |
| --- | --- |
| Sample I/P | Encrypted file name along with encryption key |
| Expected O/P | File moved to cloud |
| Actual O/P | Output as expected |
| Remarks | File uploaded successfully |

**Table 7.3 Test case for File Upload**

#### SUT 4 :- Test case for File Download:

|  |  |
| --- | --- |
| **Sl. No of Test cases** | **Test case to verify** |
| Name of test | File Download |
| Feature being tested | Auditing |
| Description | User can download a file only after it is moved to clud. |
| Sample I/P | Filename along with decryption key |
| Expected O/P | File being downloaded from the cloud |
| Actual O/P | Output as expected |
| Remarks | File downloading is done successfully |

**Table 7.4 Test case for File Download**

**CHAPTER 8**

# CONCLUSION AND FUTURE WORK

## Conclusion

In this project we tried to achieve information integrity and data deduplication using two models SecCloud as well as SecCloud+. SecCloud system allows the clients to generate information tags along with data and this task can be done by the auditor who maintains a MapReduce cloud. Apart from this SecCloud system allows the data deduplication by using the proof of possession protocol, which allows the user to keep only a single copy of data on cloud. Whatever the computation required in file uploading and auditing phase is greatly reduced in the current system. In SecCloud system server can see the content of file being stored in the cloud. To overcome this, a new system is developed which allows the clients to encrypt the data before uploading it and also decrypt the data before downloading it. This feature achieves our project objectives secure auditing as well as secure deduplication.

## Future Work

The current way of auditing and deduplicating technique can be applied to only the files containing the text documents. We can extend this technique for different kinds of files such as image file, audio file, video files, executable files etc. In the current system the process of deduplication is applicable to only single server systems. In future if multiple servers are involved, then time required to do deduplicate the cloud content will be greatly reduced. Also the future work can be extended to save disk space of a cloud server which is then reduces the cost of the system development.

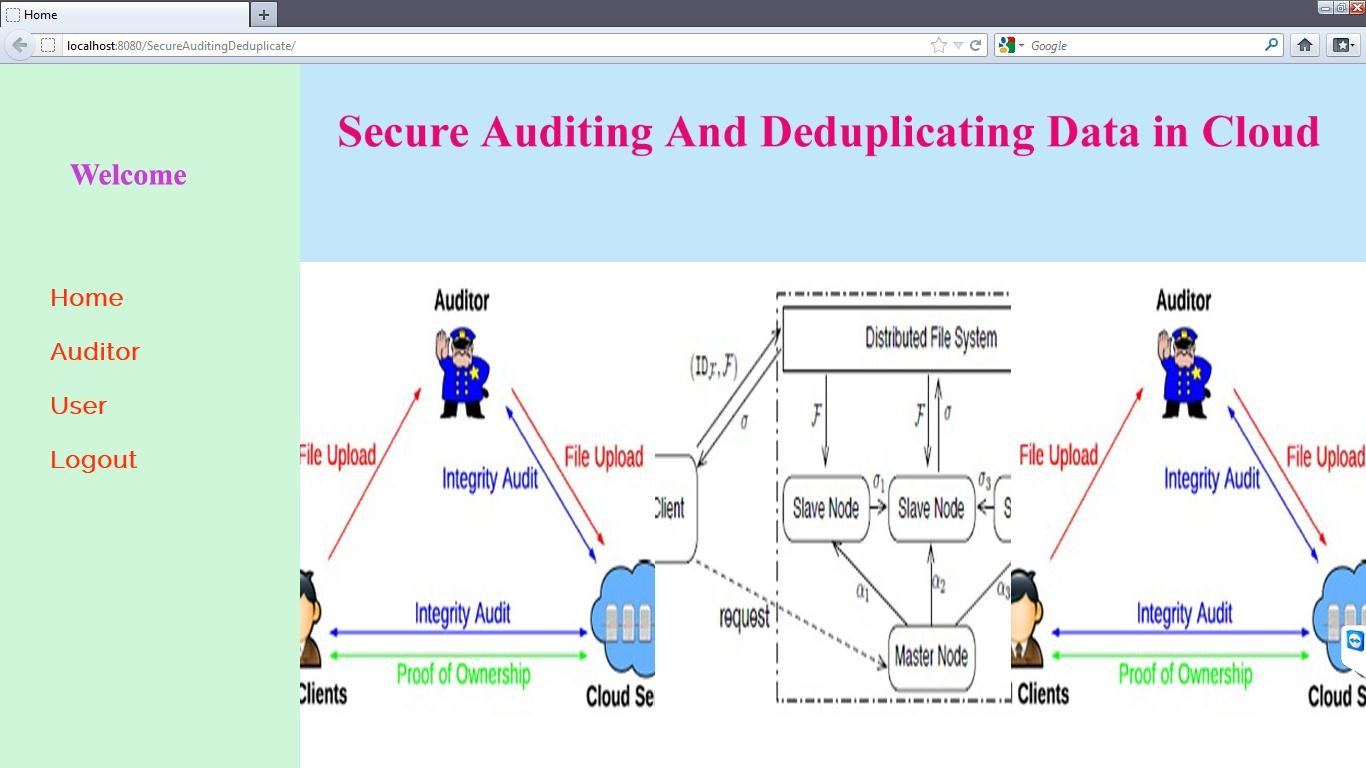
# REFERENCES

1. Komal Puri, “IDC Cloud Report 2014: Cross-Vertical Demand Side Perspective”, Nov 19, 2014.
2. S. Quinlan and S. Dorward. “Venti: a new approach to archival storage”, in Proc. USENIX FAST, Jan 2002.
3. J.R. Douceur, A. Adya, W.J. Bolosky, D. Simon, and M. Theimer, ”Reclaiming Space from Duplicate Files in a Serverless Distributed File System,” in Proc. ICDCS, 2002, pp. 617-624.
4. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman- Peleg, “Proofs of Ownership in Remote Storage Systems” in Proc. ACM Conference. Computer Communications Security, Eds., 2011, pp. 491-500.
5. M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Serveraided encryption for deduplicated storage” in USENIX Security Symposium, 2013.
6. M. Bellare, S. Keelveedhi, and T. Ristenpart, “ Message-locked encryption and secure deduplication” in EUROCRYPT, pages 296–312,2013.
7. M. Bellare, C. Namprempre, and G. Neven, “Security proofs for identity-based identification and signature scheme”, J. Cryptology, 22(1):1–61, 2009.
8. M. Bellare and A. Palacio. Gq and schnorr “Identification schemes: Proofs of security against impersonation under active and concurrent attacks”, in CRYPTO, pages 162–177, 2002.
9. S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider, “Twin clouds: An architecture for secure cloud computing”, in Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.
10. J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. “Reclaiming space from duplicate files in a serverless distributed file system”, in ICDCS, pages 617–624, 2002
11. J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, “ Secure deduplication with efficient and reliable convergent key management”,in IEEE Transactions on Parallel and Distributed Systems, 2013.
12. libcurl. [http://curl.haxx.se/libcurl/.](http://curl.haxx.se/libcurl/)
13. C. Ng and P. Lee,” Revdedup: A reverse deduplication storage system optimized for reads to latest backups”, in Proc. Of APSYS, Apr 2013.
14. J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl,”A secure data deduplication scheme for cloud storage”, in TechnicalReport, 2013.
15. R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, “Role-based access control models”, IEEE Computer, 29:38–47, Feb 1996.
16. M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, “Secure data deduplication”, in Proc. of Storage system, 2008.
17. Z. Wilcox-O‟Hearn and B. Warner, “Tahoe: the least-authority file system”, In Proc. of ACM Storage system, 2008.
18. J. Yuan and S. Yu, “Secure and constant cost public cloud storage auditing with deduplication” IACR Cryptology ePrint Archive, 2013:149, 2013.
19. K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan, “Sedic: privacy aware data intensive computing on hybrid clouds” in Proceedings of the 18th ACM conference on Computer and communications security, CCS11, pages 515–526, New York, NY, USA, 2011. ACM.
20. J. Xu, E.-C. Chang, and J. Zhou,”Weak leakage-resilient client-side deduplication of encrypted data in cloud storage”, in ASIACCS, pages 195–206, 2013.
21. J.Srinivas, K.Venkata Subba Reddy, Dr.A.Moiz Qyser, “Cloud computing basics”, in International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 5, July 2012.
22. K. Birman, G. Chockler and R. van Renesse, "Toward a cloud computing research agenda", SIGACT News, 40, 2009, pp. 68-80.
23. K.D. Bowers, A. Juels and A. Oprea, "HAIL: A high-availability and integrity layer for cloud storage", CCS'09: Proc. 16th ACM Conf. on Computer and communications security, 2009, pp. 187-198.
24. C. Cachin, I. Keidar and A. Shraer, "Trusting the cloud", ACM SIGACT News, 40, 2009, pp. 81-86.
25. Clavister, "Security in the cloud", Clavister White Paper, 2008.
26. S.L. Garfinkel, "Email-based identification and authentication: An alternative to PKI?",IEEE Security and Privacy, 1(6), 2003, pp. 20-26.
27. E. Grosse, J. Howie, J. Ransome, J. Reavis and S. Schmidt, "Cloud computing roundtable", IEEE Security & Privacy, 8(6), 2010, pp. 17-23.
28. A. Juels and B.S. Kaliski Jr, "PORs: Proofs of retrievability for large files", CCS '07: Proc. 14th ACM Conf. on Computer and communications security, 2007, pp. 584-597.
29. S. Kamara and K. Lauter, "Cryptographic cloud storage", FC'10: Proc. 14thIntl.Conf. on Financial cryptograpy and data security,2010, pp. 136-149.
30. U. Maheshwari, R. Vingralek and W. Shapiro, "How to build a trusted database system on untrusted storage", OSDI'00: Proc. 4thConf. On Symposium on Operating System Design & Implementation, 2000, p. 10.
31. R.C. Merkle, "Protocols for public key cryptosystems", IEEE Symposium on Security and Privacy, 1980, pp. 122-134.
32. E. Mykletun, M. Narasimha and G. Tsudik, "Authentication and integrity in outsourced databases", ACM Transactions on Storage (TOS), 2,2006, pp. 107-138.
33. T. Ristenpart, E. Tromer, H. Shacham and S. Savage, "Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds", CCS'09: Proc. 16thACM Conf. on Computer and communications security, 2009, pp. 199-212.
34. F. Rocha and M. Correia, "Lucy in the Sky without Diamonds: Stealing Confidential Data in the Cloud", Proc. 1stIntl. Workshop of Dependability of Clouds, Data Centers and Virtual Computing Environments, 2011, pp. 1-6.
35. N. Santos, K.P. Gummadi and R. Rodrigues, "Towards trusted cloud computing", USENIX Association, 2009, pp. 3-3.
36. G. Brunette and R. Mogull (eds), "Security guidance for critical areas of focus in cloud computing", CloudSecurityAlliance, 2009.
37. A. Shraer, C. Cachin, A. Cidon, I. Keidar, Y. Michalevsky and D. Shaket, "Venus: Verification for untrusted cloud storage", CCSW'10: Proc. ACM workshop on Cloud computing security workshop, 2010, pp. 19-30.
38. S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing", Journal of Network and Computer Applications, 34(1), 2011, pp 1-11.
39. Cloud deployment model: http://[www.cloudcomputingpatterns.org](http://www.cloudcomputingpatterns.org/)
40. H. Takabi, J.B.D. Joshi and G.-J. Ahn, "Security and Privacy Challenges in Cloud Computing Environments", IEEE Security & Privacy, 8(6), 2010, pp. 24-31.
41. M. Van Dijk and A. Juels, "On the impossibility of cryptography alone for privacy- preserving cloud computing", HotSec'10: Proc. 5thUSENIX Conf. on Hot topics in security, 2010, pp.1-8.
42. J. Viega, "Cloud computing and the common man", Computer, 42, 2009, pp. 106-108.
43. C. Wang, Q. Wang, K. Ren and W. Lou, "Ensuring data storage security in cloud computing", ARTCOM'10: Proc. Intl. Conf. on Advances in Recent Technologies in Communication and Computing, 2010, pp. 1-9.

# APPENDIX

1. **Snapshots**

## Home page:



#### Fig A.1 Home Page

This page specifies different entities of the system such as user, auditor.

## User Registration:



#### Fig A.2 User Registration Page

The registration page allows the user to enter his personnel information in order to log on to the system

## User login:



**Fig A.3 User Login Page**

## File encryption:



#### Fig A.4 File Encryption Page

Encryption is done using the encryption key created by the user.

## File upload:



#### Fig A.5 File Uploading Page

Once the encryption is done the file is now ready for uploading.

## Files in Cloud:



**Fig A.6 Cloud Content Page**

## Auditor moves the file unix1 to cloud:



**Fig A.10 Moving Files to Cloud**

**File mapping:** If the uploaded file is a duplicated file then it has to be mapped with existing files in the cloud otherwise nonduplicate files are directly uploaded to the cloud by the auditor.



**Fig A.12 File Mapping**

## The duplicate file unix2 is mapped to unix1:



## Graph for cloud files:



Fig A.17 Analysis of Cloud Content