IMPLEMENTATION OF DATA DEDUPLICATION USING CLOUD COMPUTING

Bhairavi Kesalkar[1] , Dipali Bagade[2] , Manjusha Barsagade[3] , Namita Jakulwar[4]

Prof. Shrikant Zade[5]

[1][2][3][4]*Scholar (UG),Department of Computer Science and Engineering Priyadarshini Institute of Engineering and Technology, Nagpur, Maharashtra, India*

[5]*Assistant Professor, Department of Computer Science and Engineering Priyadarshini Institute of Engineering and Technology, Nagpur, Maharashtra, India*

Email:kesalkarbhairavi@gmail.com deepalibagade1@gmail.combarsagedemanjusha06@gmail.com namitajak13@gmail.com

ABSTRACT

Cloud computing has quickly become one of the most significant field due to its evolutionary services provided model of computing not only in the IT industry but also in the software and hardware industry. This mechanism came up with increasing flexibility, scalability and reliability; while decreasing the operational and support cost. Due to the cloud computing, it becomes easy for managing the stuffs related as well as provides many features which cannot be replaced by anyone. It is a way difficult as well as effective in its own. Providing security is a major concern as the cloud data are stored and accessed in a remote server with the help of by the cloud service provider. Translation of efficient storage and security for all data is very important for cloud computing. Securing and privacy preserving of data is of high priority when it comes to cloud storage. Therefore, to provide efficient storage for cloud data owners and provide high security for data this paper proposes Cloud Computing. Intrusion , detection and prevention are performed manually by network operators in the existing system. Data deduplication technique allows the cloud users to manage their cloud storage space effectively by avoiding storage of repeated data's and save bandwidth. The data are finally stored in cloud server. To ensure data confidentiality the data are stored using encryption.

Keywords: *De-duplication, Cloud computing.*

1.Introduction

Cloud Computing: Cloud computing is an information technology paradigm that enable ubiquitous access to share pool of configurable system resources and higher level services that can be rapidly provisioned with minimal management efforts, often over the internet.

Cloud computing is one of the emerging technology, which helped several organizations to save money and time adding convenience to the end users. Thus the scope of cloud storage is vast because the organizations can virtually store their data's without bothering the entire mechanism.

Fig.1.Cloud Computing

Types of Cloud: The cloud is categorized into four types:

1.Public

2.Private

3.Hybrid

4.Communit

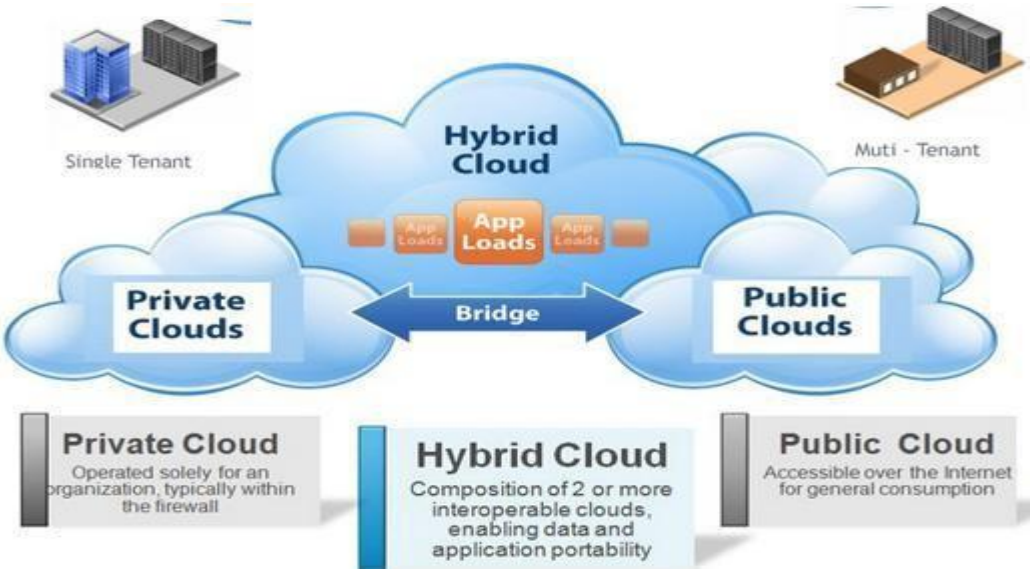

Fig.2.Types of Cloud

1. **Public Cloud**: Public cloud is a type of cloud hosting in which the cloud services are delivered over a network which is open for public usage. This model is a true representation of cloud hosting; in this the service provider renders services and infrastructure to various clients. The customers do not have any distinguishability and control over the location of the infrastructure. From the technical viewpoint, there may be slight or no difference between private and public clouds' structural design except in the level of security offered for various services given to the public cloud subscribers by the cloud hosting providers.

2. **Private Cloud**: Private cloud is also known as internal cloud; the platform for cloud computing is implemented on a cloud-based secure environment that is safeguarded by a firewall which is under the governance of the IT department that belongs to the particular corporate. Private cloud as it permits only the authorized users, gives the organisation greater and direct control over their data.

3. **Hybrid Cloud**: Hybrid cloud is a type of cloud computing, which is integrated. It can be an arrangement of two or more cloud servers, i.e. private, public or community cloud that is bound together but remain individual entities. Benefits of the multiple deployment models are available in a hybrid cloud hosting. A hybrid cloud can cross isolation and overcome boundaries by the provider; hence, it cannot be simply categorized into public, private or community cloud.

4. Community Cloud: Community cloud is a type of cloud hosting in which the setup is mutually shared between many organisations that belong to a particular community, i.e. banks and trading firms. It is a multi-tenant setup that is shared among several organisations that belong to a specific group which has similar computing apprehensions. The community members generally share similar privacy, performance and security concerns.

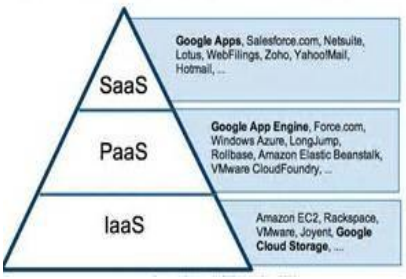

Fig.3. Services of cloud

1.3. Services of Cloud

A. SaaS (Software as a Service): SaaS model are provided with the access to the application software which are often referred as the "On-demand software". SaaS uses the Web to deliver applications that are managed by a third-party vendor and whose interface is accessed on the clients' side. Most SaaS applications can be run directly from a Web browser, without any downloads or installations required.

B. PaaS (Platform as a Service): It provides the platform which typically includes operating system, programming language execution environment, databases, web server, etc. PaaS is a framework they can build upon to develop or customize applications. PaaS makes the development, testing, and deployment of applications quick, simple, and cost- effective,

C. IaaS (Infrastructure as a service): This base layer provides the computing infrastructure, physical or virtual machines and other resources like virtual disk image library, block and file based storage, firewalls, load balances, IP addresses, virtual local area networks, etc. Instead of having to purchase software, servers, or network equipment, users can buy these as a fully outsourced service that is usually billed according to the amount of resources consumed.

2. Data Deduplication: Data deduplication or Single Instancing essentially refers to the elimination of redundant data. As the amount of digital information is increasing exponentially, there is a need to deploy storage systems that can handle and manage this information efficiently. Data deduplication is one of the emerging techniques that can be used to optimize the use of existing storage space to store a large amount of data. Basically, data deduplication is removal of redundant data [1]. Thus, reducing the amount of data reduces a lot of costs storage requirements costs, infrastructure management cost.

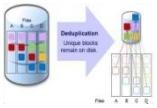

Fig.4. Data Deduplication

Implementation: For implementation we preferred ASP.NET C# language, Visual studio framework and Windows O.S. Platform as it provides inbuilt server called IIS. ASP.NET provides inbuilt MSDN managed code to support cryptographic hashing algorithm needed to perform encryption and decryption. IIS (Internet information services) server of Windows allows creating and deploying ASP.NET application, which helps in to host our prototype web application on local network as well as on public network. Data deduplication is referred to as a strategy offered to cloud storage providers (CSPs) to eliminate the duplicate data and keep only a single unique copy of it for storage space saving purpose.

Data deduplication is one of the techniques which used to solve the repetition of data. The deduplication techniques are generally used in the cloud server for reducing the space of the server. To prevent the unauthorized use of data accessing and create duplicate data on cloud the encryption technique to encrypt the data before stored on cloud server. Cloud Storage usually contains business-critical data and processes; hence high security is the only solution to retain strong trust relationship between the cloud users and cloud service providers

In this methodology we have to detect the duplicate copy of the file any type of file can be detect file .txt,.doc,.xls, ppt, .pdf. so we have to start with uploading the file when we upload the file we have to extract first 50 bytes from the file and last 50 bytes from the file. After extracting this 100 byte we have match this byte with existing byte. This comparing is done one by one byte i.e.one byte after another upon last byte. After completion of comparing this byte if this new file upload file is duplicate then we will discard the file. If this file is not duplicate, then we will upload this file. In this way the methodology is use.

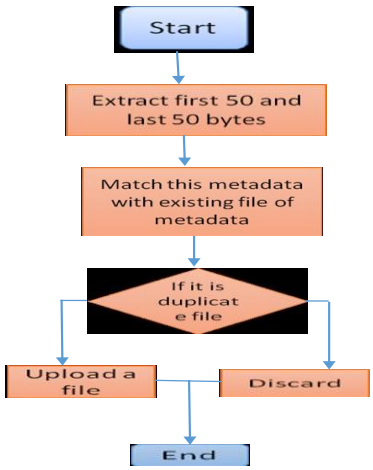

Fig.5. Flow chart

i. Encryption: Data encryption is one of the techniques where we can protect or secure our data from getting misused by the unauthorized user. Generally, there are three encryption algorithms which have been used for securing the data in cloud computing.

i. **AES Algorithm:**

The algorithm was developed by two Belgian cryptographer Joan Daemen and Vincent Rijmen. It was first published in 1998. The AES stands for Advanced Encryption Standard. It is also known as rijndael algorithm. It is symmetric key algorithm. It was first adopted by the U.S government and now is being used in the whole world.it

is having various ciphers with different keys and the block sizes [2]. In this the plain text is encrypted with the help of AES and then the cipher text which we have got will again encrypted likewise there will be various round like the AES algorithm includes 10, 12 and 14 round with the 128, 192 and 256 key bits.

ii. Framework: The term "framework" is used to loosely describe collections of anything from development tools to middleware to database services that ease the creation, deployment and management of cloud applications. Those that work at the level of servers, storage and networks are infrastructure-as-a-service (IaaS) frameworks. Those that operate at the higher level of applications are platform-as-a-service (PaaS) frameworks [7].

Client module consists of :-

1. User registration
2. Key generation
3. File-Upload: duplicate check for byte level and if file are not duplicate then encrypt and transfer it to public CSP
4. File-download: Download files uploaded by other users or admin using unique

key Admin module consists of:

1. Login operation
2. Monitoring all registered user's

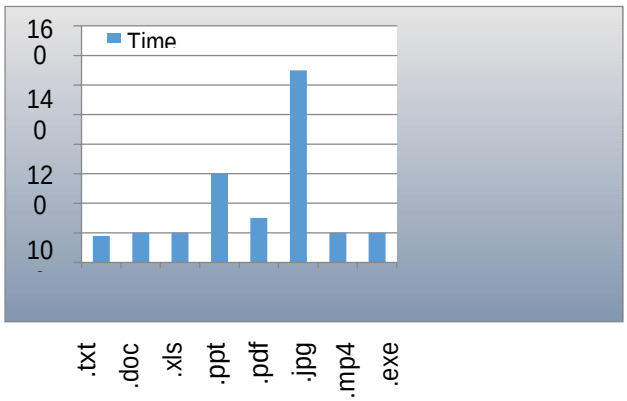

Fig.6. Deduplication time factor at Byte level

3. **Literature Survey**

In this paper they conclude that, Cloud Computing is an emerging paradigm which has become today's hottest research area due to its ability to reduce the costs associated with computing. In today's era, it is most interesting and enticing technology which is offering the services to its users on demand over the internet. Since Cloud Computing stores the data and disseminated resources in the open environment, security has become the main obstacle which is hampering the deployment of Cloud environments. Even though the Cloud Computing is promising and efficient, there are many vicinity of the data for the Cloud user. To ensure the security of data, we proposed a method by implementing RSA algorithm [8].In cloud storage service, users upload their data together with authentication information to cloud storage server. To ensure the availability and integrity of users' data stored in the cloud storage, users need to verify the cloud storage remotely and periodically, with the help of the pre-stored authentication information and without storing a local copy of the data or retrieving back the data during variation. Public variation enables a third party auditor (TPA), on the behalf of the data owner, to verify the integrity of cloud storage with the data owner's public key. In this paper, we propose a method that allows the data owner to delegate the auditing task to a potentially untrusted third party auditor in a secure manner:

(1) The data owner can verify whether the TPA has indeed performed the specified audit task; task at the right time specified by the data owner;

(2) The confidentiality of the data is protected against the TPA. Our method also enables a TPA to re-outsource the audit task [3].

Cloud computing has formed the conceptual and infrastructural basis for tomorrow's computing. The global computing infrastructure is rapidly moving towards cloud based architecture. While it is important to take advantages of could based computing by means of deploying it in diversified sectors, the security aspects in a cloud based computing environment remains at the core of interest. Cloud based services and service providers are being evolved which has resulted in a new business trend based on cloud technology. With the introduction of numerous cloud based services and geographically dispersed cloud service providers, sensitive information of different entities are normally stored in remote servers and locations with the possibilities of being exposed to unwanted parties in situations where the cloud servers storing those information are compromised. If security is not robust and consistent, the flexibility and advantages that cloud computing has to offer will have little credibility. This paper presents a review on the cloud computing concepts as well as security issues inherent within the context of cloud computing and cloud infrastructure [11].

Cloud computing is known as one of the big next things in information technology world. Unlike other traditional computing system, cloud computing paradigm that provide unlimited infrastructure to store or execute client's data/program. Cloud computing is a long dreamed vision of computing as a utility, where data owners can remotely store their data in the cloud to enjoy on- demand highly-quality application and services from a shared pool of configurable computing resources. This paper gives a brief introduction of cloud computing its types and security issue and approaches to secure the data in the cloud environment [5].

4.Conclusion

The notion of authorized data de-duplication technique is specialized data compression technique which eliminates redundant data as well as improves storage and bandwidth utilization. Convergent encryption technique is proposed to enforce confidentiality during de-duplication, which encrypt data before outsourcing. Security analysis demonstrates that the schemes are secure in terms of insider and outsider attacks. To better protect data security, we present Two Factor Authentication scheme (2FA) of user along with PoW of files, to address problem of authorized data de-duplication, in which the duplicate-check tokens of files are generated by the private cloud server with private keys.

5.Future scope:

Various open issues are identified as future scope.

1. **Secure trust based Solution for cloud computing Service:** A secure environment for execution of the cloud computing services along with overall security considerations is a challenge. A secure and trusted solution is the requirement that needs to be focused and addressed by the cloud computing infrastructure [4].

2. **Optimization of resource Utilization:** Security considerations and provisions for virtualization along with the optimum use of the cloud infrastructure also needs to be focused and addressed.

7.References:

[1]. Jin Li ,Yan Kit Li ,XiaofengChen ,Patrick P.C. Lee and WenjingLou ”A Hybrid Cloud Approach for Secure Authorized Deduplication” IEEE Transactions On Parallel And Distributed System Vol.26,No.5, May2015

[2]. Puzio, P. ; SecludIT, Sophia-Antipolis, France ; Molva, R.; Onen,M.; Loureiro,S. “ClouDedup Secure Deduplication with Encrypted Data for Cloud Storage”

[3]. Rabi Prasad Padhy, Manas Ranjan Patra, Suresh Chandra Satapathy, “Cloud Computing: Security Issues and Research Challenges” @IRACST 2011.

[4]. M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. “Secure data deduplication”. In Proc. of StorageSS, 2008

[5]. Q.Wang, C.Wang, J. Li, K. Ren, and W. Lou. Enabling public verifiability and data dynamics for storage security in cloud computing. In European Symposium on Research in Computer Security (ESORICS '09), volume 5789 of Lecture Notes in Computer Science, pages 355{370. Springer, 2009.

[6]. K Zhang, X Zhou, Y Chen and X Wang, “Sedic Privacy-Aware Data Intensive Environments”Computing” Kaaniche, N. ; Inst. Mines-Telecom, Telecom Sud Paris, Evry, France; Laurent,M.A “Secure Client Side Deduplication Scheme in Cloud Storage

[7]. Changyou Guo and Xuefeng Zheng, “The Research of Data Security Mechanism Based on Cloud Computing” @International journal of Security and its applications 2015

[8]. Iuon-Chang and Po-Ching Chien, “Data Deduplictionn Scheme for Cloud storage”. IJ3C, Vol. 1, No. 2 (2012) [10] Jin li,yan kit li,xiaofeng chen,Patrick p.c.lee,wenjing Luo. “A hybrid cloud approach for secure authorized deduplication”. IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED

