Proposed Design for Improvement in De-duplication with Integrity Verification

[1]Minal B.Pokale, [2]Dr.S.M.Chaware
[1]Student, [2]Guide
[1]Computer Engineering,
[1]Marathwada Mitra Mandal's College of Engineering, Pune, India

***Abstract*— This paper presents improvement in de-duplication and checking integrity of the de-duplicated document using Third Party Auditor. In the era of digital world, everything becomes online, we store, upload and download the document and outsourced the services. Security is one of the important issues in every aspect and field. Now a days, vast amount of information gets stored over cloud, which include personal file, images, pdf, text, multimedia data etc. Everyone just want to upload, store its data without worrying about the security. IT organization should look after to provide security to the document. Documents content get altered by some outsider attacker, so it is necessary to verify the integrity. Suppose intentionally some document get corrupted, then recovery of it is one of the difficult task. As huge amount of data get uploaded and stored over cloud server, data gets duplicated. Unnecessary Space get waste because of data duplication. The Proposed system helps to improve usage of space and provide assurance about document confidentiality.**

***Index Terms*— Cloud server, Secure Hash Algorithm(SHA-1), Chunk, Private Server, Hash table, Third Party Auditor.**

I. INTRODUCTION

IT companies are focusing on providing a high quality service. IT become part of daily business activities. Infrastructure of cloud computing is very powerful and reliable. It is one of the growing technology in IT sectors. Daily, so many users get connected to the cloud server. Cloud computing is popular because of the services and facility it provide to the user. Cloud computing provide on demand services like social networking site, mails, online storage of file (e.g. Amazon, Google app engine, Google Docs, Salesforce, Dropbox) [2]. Today, quantity of data that is uploaded on daily bases goes on increasing. From the analysis about 90% of the user data get store over cloud. Space requirement is increased. Data may include text file, pdf, images and multimedia file etc. As people depends totally on the services that are provided by the cloud computing. In cloud computing every organization and user uses the various application, software, hardware and storage space of cloud. Security issues in cloud computing:

- Information leakage
- Unauthorized secondary usage
- Types of attacker and their capability
- Any unauthorized user can modify data using its authority [11].

Fig1. Concept of Cloud computing, source: ref [8]

I. De-duplication

Data De-duplication is one of the major challenge in front of IT organization. It becomes the interesting trend. De-duplication is nothing but elimination or removal of duplicate of the file or data that is store over the cloud server. For de-duplicating data, various version of the message digest is used.

Previously various version of message digest algorithm which is hash function are used to generate the hash value. These algorithm have some vulnerabilities like not resistant to collision, cracked by brute-force attack. Secure Hash Algorithm is developed that overcome these vulnerability it is strong, it is against collision attack. In this if data get changed then the respective hash value for the file will also get change.

II. Third Party Auditor

Auditing is nothing but the checking or examination or analysis of process or system or document that is store. Auditor is first, second or third Party each having its separate purpose. There is internal and external auditor where internal audit that is work within an organization, External audit is any outsider that audit the process or document or system.

In proposed system third party auditor is used, that check the integrity of the documents that is store over the cloud. One of the important advantage of this is that after auditing if any file get corrupted or its content get changed, then the auditor gives exactly which file get corrupted. It informs to the user and private server by sending them notification. We can improve the deduplication using Secure Hash Algorithm (SHA1) to generate the hash of the de-duplicated partition. In section I we are given introduction about De-duplication and Third Party Auditor, security issues in cloud computing. In section II given the literature survey. In section III we are given the motivation about proposed system. In section IV we describe the system architecture and working of entity involve in it. In section V, mentioned the mathematical model and section VI, conclusion is given.

II. LITERATURE SURVEY

Cloud computing becomes one of the vast research area for the scientist and IT enterprise. In de-duplication lot of research had been done and it is keep on going. De-duplicating document reduces the unnecessary consumption of space. Third Party Auditor is used in IT-enterprise to check the integrity of the document from the outsider attacker. Much of the research also done in this area.

Junbeom et.al. Proposed dynamic ownership management in cloud storage, it proposed for encrypted data using novel server-side de-duplication scheme. Most of the user encrypt their data before uploading them to the cloud server to make data secure but it hampers the deduplication because of the randomization property of the encryption, to solve this problem author use the convergent encryption. With convergent encryption there is no "password reset", so if you forget you will loss access to the cloud storage. Tag consistency problem occur in convergent encryption [1][6].

Jingwei et.al. Achieve the secure auditing and de-duplication, in this file uploading protocol, integrity auditing protocol and proof or ownership protocol is mentioned. First auditing is done then secure deduplication is done. File level de-duplication is achieved [5].

Reshma N.S et.al. Author point out that as cloud services are increased day by day and user connected to it is also get increased. So demand of storage is get increased. De-duplication is necessary to make space available for storing more document. Author proposes the system in which it assures the de-duplication of data and confidentiality of data. In this block level de-duplication is achieved. It distributed the data over multiple machine host. Bilinear pairing algorithm, weil pairing and its property, miller's algorithm are used, each solving its particular drawback [7].

Cong Wang et.al. proposed privacy-preserving public auditing for secure cloud storage system, uses the homomorphic linear authenticator and random masking technique is used that guaranteed the data uploaded by the user is not exposed to the outsider and Third Party Auditor would not learn any knowledge about the content of the data. Here it unable to achieve the batch auditing, that is many request cannot verify at one time [3].

Boyang et. al. proposed new public auditing, it used key-gen, re-Key, sign, re-sign, proof-gen, proof-verify algorithm. Uses the proxy re-signature, means it is used to convert the key of one user into the same as another user. This scheme allow, on behalf of the existing user, the cloud to re-sign the block. Verification of integrity of the shared data is done by auditor without retrieving the entire data from the cloud. This scheme does not taken into account the collision and assumption of presence of secure channel between the entities. [4].

Bhale et. al. Proposed system in which for securing user document Identity Based Encryption technique is used and for generating respective hash value of the data store over cloud using MD5. Three entities are involve in this client, cloud admin, Third Party Auditor. There is no practical method for authentication of the user, it is higher exposure to the risk and it is less secure [2].Anirudha et.al. Focus is given on integrity verification of the data stored on cloud storage server. Existing Third Party protocol is optimized to make it resistance to replay, replace and force attack launch by malicious insider. Dynamic data update over fine and block level is also achieve using a modified Chameleon Authentication Tree which is generalize merkle hash tree [11].

In this, static and dynamic protocol is also mentioned in which static means data that is stored over cloud is check for integrity verification and in dynamic means verifies the integrity of the data that is updated or modified by the client on cloud [9].

Salah et.al. Author used hash message authentication code algorithm, for verification of information passed between applications or store in a potentially vulnerable location. It uses the SHA -256 which operates in the same way as MD-4, MD-5 and SHA1. This research ensures the privacy protection, data storage correctness. Problem with HMAC algorithm is that it is difficult to convince the third party, it is open to existential forgery attack and reversible encryption is sometimes necessary[8].

Zaid et.al. Uses the AES to secure the data during auditing that means semi –trusted Third Party Auditor does not reveal the content of the file. Use of Elliptical Curve Cryptography provide the confidentiality to the data during transmission, hence file is encrypted before store to the cloud server [10]. It also focuses on to reduce the auditing cost, author used the modern cipher cryptography with cryptographic hash function. But in this total length of the request is less than thread in bilinear map-based scheme. This achieves the dynamic data auditing and batch auditing too [11].

III. MOTIVATION

Tremendous amount of data, personal file get stored by user on cloud. There is possibility that the hacker or attacker will attack the server, also because of some software and hardware failure the integrity is violated. The data store or uploaded by user on cloud will get duplicated, so it will consume large amount of unnecessary space over cloud. Security is get reduced and large amount of space will get consume so this system is design.

IV. PROPOSED SYSTEM ARCHITECTURE

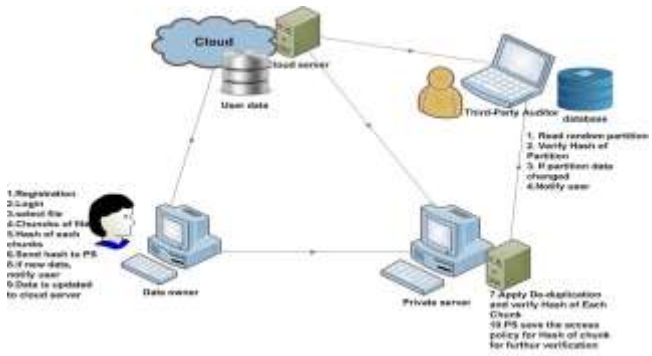

Fig 2. Proposed System Architecture

There are four modules are involve. Each module has its separate working. Secure Hash Algorithm is used to generate the hash table that is the unique identity number for each chunk that is generated during the file (.txt extension) selected for upload by the user. Third Party Auditor is used to check any integrity of the partition that is store over the cloud server. If any content of that file will get change, then it will give alert to the private server and user. The input to the system is the text file (file with .txt extension) which is uploaded, during uploading the file get partition using algorithm. Secure Hash Algorithm is used, which is secure, used to remove the duplicate of the file.

Third Party Auditor which takes hash value from private server. Third Party Auditor randomly read the partition from cloud server and corresponding hash value for the respective partition is selected from the private server. Verification of the content will be done by Third Party Auditor. Advantage of Third Party Auditor is, it will give the user exact which partition is going to corrupt.

I. Working of Proposed system

There are mainly four entity are involved in proposed system which perform de-duplication with integrity verification using Third Party Auditor. Varieties of task perform by each entity is given below.

- User :

1. Registration and logging
2. Select file(text file with .txt extension)
3. Chunk of file and respective hash value is generated using SHA-1
4. Send hash to the private server
5. If new data, notify user. Encrypted data chunk is updated to cloud server

- Private server tasks :

1. Apply De-Duplication and verify hash of each chunk
2. Private server save the access policy, Checks access policy
3. If allow send requests to cloud

- Cloud server :

1. Stored the chunk of file

- Third Party Auditor :

1. Read random partition
2. Verify hash of chunk
3. If partition data changed
4. Notify user and private server

II. .Algorithmic flow of proposed architecture

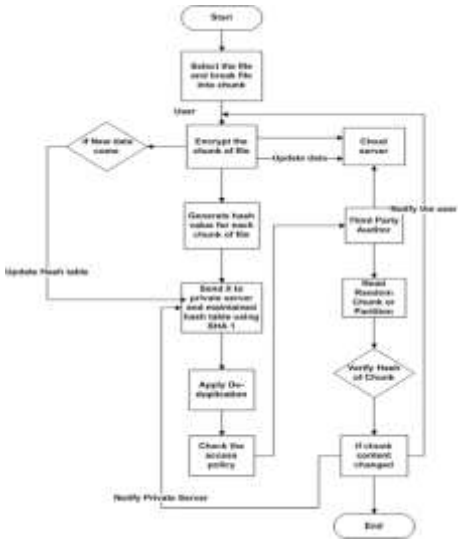

Fig 3. Algorithmic Flow of Proposed Architecture

III. Proposed Implementation Methodology

SHA-1(Secure Hash Algorithm) and AES (Advance Encryption Algorithm) are used. SHA-1 is used to generate hash value of the text document and Encryption algorithm is used to encrypt the document chunk that are formed at the time of uploading document. In this way, this provide security and de-duplicate data is stored over cloud server. Third Party Auditor is used to check the integrity of the chunk and send alert if content of that data get changed.

V. CONCLUSION

This system is design to achieve de-duplication. Integrity verification using Third Party Auditor to ensure the document uploaded by the user is secure or not. It will notify the user, exactly which file is corrupted. In future we can modify Third Party Auditor to monitor it on DoS attack. We can upload various types of file image, pdf file etc.

REFERENCES

[1] Junbeoom Hur, Dongyoung Koo, Youngjoo Shin, and Kyungate Kang,"Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage", IEEE Trans. on Knowledge and Data Engineering, 2016.

[2] Bhale Pradeepkumar Gajendra, Vinay Kumar Singh, More Sujeet ,"Achieving Cloud Security using Third Party Auditor, MD5 and Identity-Based Encryption ",[2016]

[3] Cong Wang, Sherman S.M. Chow, Qian Wang, Sherman S.M. Chow, Qian Wang," Privacy-Preserving Public Auditing for Secure Cloud Storage", IEEE TRANSACTIONS ON COMPUTERS, VOL. 62, NO. 2, FEBRUARY 2013

[4] Boyang Wang, Baochun Li and Hui Li," Panda: Public Auditing for Shared Data with Efficient User Revocation in the Cloud"2013

[5] Jingwei Li, Jin Li, Dongqing Xie and Zhang Cai," Secure Auditing and De-duplicating Data in Cloud" 2015 IEEE Transaction.

[6] P.Premkumar, Dr.D.Shanthi, "An Efficient Dynamic Data Violation Checking Technique For Data Integrity Assurance In Cloud Computing", IJIRSET, March 2014.

[7] Salah H. Abbdal, Hai Jin, Deqing Zou, Ali. A. Yassen," Secure Third Party Auditor for Ensuring Data Integrity in Cloud Storage" 2014 IEEE

[8] Anirudha Pratap Singha, Syam Kumar Pasupuletib," Optimized Public Auditing and Data Dynamics for Data Storage Security in Cloud Computing", ScienceDirect,2016.

[9] Zaid Alaa Hussien, Hai Jin, Zaid Ameen Abduljabbar, Ali A. Yassin, Mohammed Abdulridha Hussain, Salah H. Abbdal, Deqing Zou, "Public Auditing for Secure Data Storage in Cloud through a Third Party Auditor Using Modern Ciphertext",2015 IEEE.

[10] Selvamani K, Jayanthi S"A Review on Cloud Data Security and its Mitigation Techniques", ScienceDirect 2015.

[11] Tina Esther Trueman, P.Narayanasamy, "Ensuring Privacy And Data Freshness for Public Auditing of Shared Data in Cloud" 2016 IEEE, pg no. 23-27.

[12] Solomon Guadie Worku, Chunxiang Xu, Jining Zhao, Xiaohu He,"Secure and efficient privacy-preserving public auditing scheme for cloud storage" ,Computers and Electrical Engineering, 40(5):1703–1713, 2014.

[13] Reshma N.S., Greeshma N.Gopal, Sreeraag G, "A Novel scheme for Authenticated secured De-duplication with Identity based encryption in Cloud", 2016 IEEE.

[14] Swapnali More, Sangita Chaudhari," Third Party Public Auditing scheme for Cloud Storage", Sciencedirect Procedia Computer Science 79 (2016) 69 – 76.