Using The Reduplication Technique, Cloud Computing Data Storage Management

Mr. Prasanna Kumar Mishra, Mr. Sandip Kumar Bala

Department of Master in Computer Application, College Of Engineering Bhubaneswar, Odisha, INDIA.

Abstract:

One advantage of cloud storage is that it allows users to store and retrieve their data from any location at any time. The idea of cloud computing is crucial to information technology since it plays a vital role in data processing and storage. However, because all data are saved in cloud storage, it creates concerns about privacy and data secrecy, which makes customers less likely to use cloud storage services. The data kept on the cloud should be safeguarded to deter unauthorized users from wanting to access it illegally. There is a new idea in data security known as encryption. However, data security and privacy are preserved in the cloud using cipher text or encrypted data. Because storage space is limited, a cloud storage server usually uses a specific data compression technique called data deduplication to remove duplicate data in order to maintain user volume. This significantly lowers the amount of storage devices needed because only encrypted data will be kept in the cloud.

Keywords – Cloud Computing, Encryption, Data storage, Decryption, Deduplication management, Cipher-text, Security, Perfect hashing, etc.

1. INTRODUCTION :

There have been people eager to forecast the future of technology ever since the first technologies were flourished. Since the formation of computers, however, the outbreak in both the pace of alteration and the volume of prognostication on all sorts of technology trends is seen. The digital transformation is far more significant wherever the cloud computing is concerned. Cloud computing is the third wave of digital revolution.
the duplicate data, but the main issue with those plans is poverty of security and poverty of tractability for the safe data access control. Due to these two issues, very few of them are taken into practice. In this, to deduplicated the encrypted data to allow secure data access control we used a scheme known as Attribute-based Encryption. Data deduplication technique provides the cloud users to control their cloud storage space virtually by avoiding storage of regular data's and save bandwidth. The data are finally stored in cloud server namely Cloud Me. To make certain data confidentiality the data are stored in an encrypted type using Advanced Encryption Standard (AES) algorithm. Data deduplication, which makes it possible for data owners to share a copy of the same data, can be performed to reduce the consumption of storage space. Due to the above issues, there is research on encrypted data deduplication. In this manuscript, we propose an encrypted data de-duplication mechanism which makes the cloud storage server be able to eliminate duplicate cipher texts and improves the privacy protection.
computing has completely changed the way of business and their consumers-for storing and accessing the data. Cloud computing is empowering the companies to leverage the cloud to innovate cheaper and faster. There are 3 types of cloud-high, low & medium level. Based on this cloud computing can be broken up into three main services: Software as a service (SaaS), Infrastructure as a service (IaaS), and platform as a service (PaaS). These three services make up a rack space calls the cloud computing stack. The expeditious development in big data and cloud computing has changed the user's terminology to the tackle the vast information. cloud computing immensely opens the doors for data providers who want to outsource their data to the cloud without revealing their precise data to the foreign parties . One threatening challenge of today's cloud storage services is the management of gross amount of data. The estimated amount of data generated in 2017 and 2018 is 30 zettabytes and 33 zettabytes respectively . And there is a threat of expanding more amount of data in upcoming years. It is

expected that volume of the data will reach a peak of 40 trillion gigabytes, concerning report of IDC . To resolve the above issue data deduplication is broadly practiced in cloud services providers, which eliminate multiple duplicate data to improve storage utilization. Paradisiac ally, the increasing value of data needs cost-efficient storage, to end this mutually the deduplication technique can be used. The method of Deduplication is holds unique information compression strategy to remove excess information. This decreases the rate of transmission and loading space in framework. This method of Deduplication finds out the replica of the data. It spares just one duplicate of the information and emphatically utilize consistent pointers for copied information. This whole technique actually removes the unnecessary copies of encrypted data in the cloud. For feasible cloud storage such as Dropbx, Mozy, Memopal and lessened maintenance cost this deduplication is the technique applied to user data . Actually, there are two kinds of data duplication which are: file and block. From a meta context, block - level duplication is achieved at the aggregate level by duplicating the blocks of data that subsume the volume. While file-deduplication works, as its name implies, at the file level. If the duplicate files are in the duplication domain, they are single-instanced. The file duplication is generally considered as coarse level of duplication and block level, fine grained. As far as such type of block level deduplication is bothered, it often yields more considerable outcomes than file level dedupes. Block level block dedupe works even on just similar file but file dedupe works on whole identical file. Hence we have implemented block level deduplication in our paper. For example, in case where multiple edits of document are maintained, a file may have several copies each of with few words changed. File deduplication wouldn’t work on these but block deduplication may be able to duplicate at block level. For better security and efficient handling of data, two models are used: client-side and server-side. Both can be applied to single server storage and distributed storage [8]. Both models use uniform scenario which depends on number of basic security techniques. For this we use block encryption to enable encryption which acquiesce deduplication of common chunks. A hash function is used by convergent encryption for getting the key value by converting the plain text into block of chunks. Any client encrypting the data will use the same key to decrypt the document. When the users utilize cloud services, they hand over the control of their confidential data to the cloud service providers which can cause the risk of privacy leakage. Encrypted data is more secure to transmit over insecure network. To read the encrypted data he/she needs to have a secret key to decrypt message. Encrypting data has additional advantages assuring that messages should not be revoked during transmit of data and verifying the identity of the sender other than providing the confidentiality and privacy of a message. For the encryption and decryption, we use RC6 Algorithm. RC6 is a symmetric key block cipher. The improvements of RC6 over RC5 include using four w-bit word registers, and introducing a quadratic equation into the transformation, integer multiplication as an additional primitive operation. The evolution has provided a simple cipher yielding numerous evaluations and security in a small package. RC6 uses 128 bit block size and supports key sizes of 128, 192 and 256 bits. Also it uses fewer rounds and offers a higher throughput, data-dependent rotations, modular addition, and XOR operations. Hashing is the practice of using an algorithm to map data of any size to a fixed length. This is called a hash value. For every block, hash value is unique. Now, whereas encryption is meant to protect data in transit, hashing is meant to verify that a file or piece of data has not been altered that it is authentic. Hash functions are useful and such functions are important cryptographic primitives used for things such as digital signatures and password protection and appear in almost all information security applications. SHA stands for “Secure Hash Algorithm” it is a fingerprint that specifics the data. SHA-2 is a family of hashes and is available in a variety of lengths, the most popular being 256-bit. The hash function compares the computed "hash" to a known and expected hash value through which a person can also determine the data's integrity. As we are using SHA-256 that means that the algorithm is going to output a hash value that is 256 bits, usually represented by a 64 character hexadecimal string. Whereas encryption is a two-way function hashing is a one-way function. A one-way hash the data cannot be generated from the hash, but can be generated from any piece of data. SHA-256 consists of bitwise operations, modular additions, and compression functions. This is helpful in case an attacker hacks the database. Additionally, SHAs exhibit the avalanche effect where, when an input is changed slightly the output changes significantly. To develop of the web application, we used ADO.NET this is a module used to establish connection between application and data sources. .NET is a framework to develop software applications. Moreover, it provides a broad range of functionalities and support. ADO.NET Data sources can be such as SQL Server and XML. To connect, retrieve, insert and delete data ADO.NET consists of classes these classes. The ADO.NET classes are integrated with XML classes located into System.Xml.dll which are located into System.Data.dll and the components that are used for accessing and manipulating data are the .NET Framework data provider and the Data Set. As we have developed a web application, the web pages are created using ASP.NET. It actually executes code on the server, code that can use databases and then produce html to the browser. ASP.NET with C# is used to develop the re-engineered supermarket management system, where ASP.NET is a reworking of the original Active Server Pages technology. Along with ADO.NET and ASP.NET, we have used stored procedures in SQL. Stored procedures in SQL Server, stores the procedures program statements to perform operations in the database and return a status value to a calling procedure or batch. Structured Query Language (SQL) statements are set of stored procedure with an assigned name and are stored in a relational database management system as a group.

2. Cloud LITERATURE REVIEW

In this project, the notion of authorized data deduplication was proposed by Li et al. to protect the data security by including differential advantages of users in the duplicate check. In this project we perform several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture where the duplicate-check tokens of files are generated by the private cloud server with private keys. In order to tackle the problem that the unauthorized users can access the user information only by supplying the hash value. Halevi et al. proposed the proof of ownership (PoW), which is an interaction protocol between client side and server side to verify the ownership of that client. In, the client and server create a Merkle Hash Tree (MHT) based on the source _le, and use a challenge-response model to verify the correct of MHT path provided by the client. Blasco et al. proposed a system which is a bf-PoW scheme based on the bloom filter, to achieve the proof of ownership systematically which has the requirements of the certain tokens from the substantiated client. Through the security analysis, a wide range of benchmark tests and comparison of the already existing schemes, the proposed scheme greatly decreases the amount of both the client and the cloud server. D. Harnik et al. proposed a procedure that provides higher secure guarantees while slightly decreasing bandwidth savings, since deduplication offers considerable savings in both disk capacity and network bandwidth. Xia et al. review the differences between data deduplication and the background of data deduplication and traditional data compression. Ng et al. proposed a private data deduplication in data storage, where a client held a private data proves to a server stored a summary string of the data that he/she is the owner of that data without revealing further information to the server. In this paper, Xu et al. offers a cryptographic antediluvian to enlarge the security of client- side deduplication in the bounded percolate setting where the certain amount of efficiently-extractable information about any file is oozed. Encrypted deduplication has been deployed in commercial cloud environments and extensively analyzed in the literature to simultaneously achieve both data security and storage efficiency. Li et al. proposed how the deterministic nature of encrypted deduplication makes it vulnerable to information leakage caused by analysis of the frequency. In the Storer's et al. paper that have developed two models for secure deduplicated storage which are anonymous and substantiated. These two of the models demonstrate that the security can be amalgamated with deduplication in a way that provides a multiple security characteristics range. The security is provided through the use of convergent encryption in the models that they have presented. A map is created for each file that narrates how to rebuild a file from blocks in both the anonymous and authenticated models. To prevent information leakage, several solutions have been proposed. However, these solutions are based on a strong assumption that all individual files are independent of each other. Shin et al. proposed a storage GW-based secure client-side deduplication protocol. A storage GW is a network appliance that provides access to the remote cloud server and simplifies interactions with cloud storage services, and is used in various cloud service delivery models such as public, private, and hybrid cloud computing. The proposed solution, by utilizing the storage GW as an important component in the system design, achieves greater network efficiency and architectural flexibility while also reducing the risk of information leakage. SHOBANA et al. offers system that compress the data by removing the duplicate that is same data, to assure the privacy of the confidential data during deduplication. To encrypt the data before deploying encryption method is used. MihirBellare et al. proposed a paper on Duples: Server helped encryption for de-copied capacity. This paper for the most part highlights on Data trustworthiness and capacity which are two principles essential requirement for distributed storage. Evidence of Retrievability (POR) and Proof of Data Possession (PDP) systems assures information uprightness for distributed storage. The individual examining of these developing undertakings can be blundering. This plan of open key in view of homomorphism direct authenticator, which permit TPA to play out the evaluating without requesting the nearby duplicate of information and along these lines drastically decreases the correspondence and calculation expenses when differed with the clear information examining perspective. Jewie Yuan et al. proposed a paper on Secure and Constant Cost Public Cloud Storage Auditing with De-duplication. This paper conveys the arrangement that bolster skilled and secure information trustworthiness evaluating with capacity deduplication for distributed storage. This sorts out issue with novel plan in light of strategies including polynomial-based confirmation labels and homomorphic straight authenticators. Evidence of Retrievability (POR) and Proof of Data Possession (PDP) are the course of actions for information honesty and capacity proficiency for distributed storage. Verification of Ownership (POW) is the method that exile redundantly copied information from the server intensifying capacity proficiency in a secured way. This security of the proposed plot in view of the computational Diffie-Hellman issues, the Static DiffieHellman issue and t-solid Diffie-Hellman issue. Yang et al. another secret method for more reliable possession. Provable duty for in de-duplication in distributed storage by utilizing remote information checking method. The far site circulated document framework gives approachability by repeating each record onto different desktop PCs. Since this replication demolish huge storage room, it is essential to recover utilized space where feasible. In the following project Zhang et al. offered two anonymous CP-ABE schemes in which a similar part is added before the decryption part with fast decryption by introducing a new technique called match-thendecrypt into the decryption. For the purpose of the fast decryption, to allow aggregation of pairings during decryption, special attribute secret key components are generated, and hence the unidentified decryption only involves small and constant number of pairings Gigi's et al. this paper widens data flow testing techniques to Web applications, and presents a proposed perspective to data flow testing of ASP.NET Web applications. It considers the data flow analysis of ASP.NET Web applications, which have different structure than traditional programs. Yingyang et al. design has been fruitfully applied in system design. It brings a great ease to secondary development staff and escalate the system design process, improve the system design performance. Stored procedure router has a certain reuse. When invoking stored procedure coded by PL/SQL, it is convenient to use stored procedure router. In this article, Zhang et al. mainly discusses the application of distributed database technology in research-oriented platform building. Based on the .NET Remoting technology in C# solves the remote communication problems. Then solve the system's data consistency problem combing Based on the above research, the developed comprehensive research-oriented platform can effectively manage various branch course data and improve users' using efficiency and accuracy. Bhaskar et al. this paper comprises of a straightforward outline to accomplish a secured deduplication system in half and half cloud. The outline comprises of couple of modules and two stages. The main stage is encryption and second stage is deduplication. However, the above data deduplication schemes do not take into account the key updating and user revocation. Kwon et al. proposed a new deduplication scheme with multimedia data, which is based on randomized convergent encryption and privilege-based encryption to achieve authorized reduplication and user revocation. Hur et al. proposed a novel data reduplication scheme for the server-side, which uses the randomized convergent encryption algorithm and ownership group key distribution technique to achieve the authorized access and support security reduplication with ownership changes dynamically. Ding et al. proposed a secure encrypted data reduplication scheme, which exploits the homomorphic encryption algorithm to achieve security data reduplication, and supports ownership check and user revocation. However, the above schemes exploit the homomorphic encryption and proxy re-encryption with high computation cost.

3. PROPOSED SYSTEM

Specifically, the contributions of this paper can be summarized as below:

1. We motivate to save cloud storage and preserve the privacy of data holders by proposing scheme to manage encrypted data storage with reduplication.
2. Our scheme can flexibly support data sharing with reduplication even when the data holder is offline, and it does not intrude the privacy of data holders.
3. We aim an effective address to verify data ownership and verify duplicate storage with secure challenge and huge data support.
4. We combine cloud data reduplication with data access control in a simple way, thus resign data reduplication and encryption.
5. We verify the security and assess the execution of the proposed scheme through analysis and simulation. The results show its efficiency, effectiveness and applicability.

The proposed system detects duplicate files or data stored on the cloud. Security and privacy of the file is maintained by storing the files in encrypted format. This system will result in storing only those files which have unique content. For developing this system, we have used some hashing algorithm as well as encryption algorithm. The idea behind this project is to save the space on cloud and also the time of the user. The overview of the proposed reduplication system consists of the following three phases: authorized reduplication (Phase 1), proof of ownership (Phase 2), and encryption (Phase 3). System construction shown in the flowchart describes the working of the system. The working of the proposed reduplication system is mainly divided into three phases as authorized reduplication, proof of ownership, and encryption. In Phase 1 that is in authorized reduplication, first it will check if the user is authorized or not. The authorized user can only upload or download the documents/files from the cloud. While uploading the file, the system will check whether the content of the file is similar to those files which are already present on the cloud. This is performed using Deduplication technique. Data reduplication or Single Instancing essentially refers to the elimination of redundant data. Data reduplication is one of the transpiring techniques that can be used to enhance the use of existing storage space to store a large amount of data. Basically, data reduplication is removal of redundant data. In the proposed system, reduplication is performed page wise. The system will check for the duplicate file. Following are the steps to perform reduplication:- 1. Divide the input data into blocks or "chunks." 2. Calculate a hash value for each block of data. 3. Use these values to determine if another block of the same data has already been stored. 4. Replace the duplicate data with a reference to the object already in the database.

4. CONCLUSIONS:

Managing encrypted data with deduplication is essential and significant in practice for achieving a successful cloud storage service, primarily for data storage. In this paper, we scheduled a practical scheme to manage the encrypted data in cloud with deduplication based on ownership challenge. Our scheme can openly support data update and sharing with deduplication. Encrypted data can be securely approached because only certified data holders can obtain the symmetric keys used for data decryption. Thus this paper compresses the data by deleting the duplicate copies of equivalent data and it is widely used in cloud storage to save bandwidth and minimize the storage space. Deduplication eliminates duplicate data stored on Cloud. Thus by reducing the storage usage, cost will also be reduced automatically. To secure the privacy of sensitive data during deduplication, the encryption technique is used to encrypt the data before outsourcing.

5. REFERENCES

1) S. G. Pundkar, G. R. Bamnote "Secure Sharing of Personal Records in Cloud using Encryption" Global Journal of Engineering Science and Researches May 2015.
2) J. Li, C. Qin, P. P. C. Lee, and X. Zhang, “Information leakage in encrypted deduplication via frequency analysis," in Proc. 47th Annu. IEEE/IFIP Int conf. Dependable Syst. Netw., Jun. 2017, pp. 1_12.
3) D. Harnik, B. Pinkas, and A. Shulman-Peleg, “Side channels in cloud services: Deduplication in cloud storage,"IEEE Security Privacy, vol. 8, no. 6, pp. 40_47, Nov./Dec. 2010.
4) R. SHOBANA, K. SHANTHA SHALINI, S. LEELAVATHY and V. SRIDEVI “De-Duplication of Data in Cloud” Int. J. Chem. Sci.: 14(4), 2016
5) J. Blasco, R. Di Pietro, A. Orfila, and A. Sorniotti, „„A tunable proof of ownership scheme for deduplication using bloom filters,““ in Proc. IEEE Conf. Commun. Netw. Secure. (CNS), Oct. 2014, pp. 481–489.
6) W. K. Ng, Y. Wen, and H. Zhu, „„Private data deduplication protocols in cloud storage,““ in Proc. 27th Annu. ACM Symp. Appl. Comput., 2012, pp. 441– 446.
7) J. Xu, E.-C. Chang, and J. Zhou, „„Weak leakage-resilient client-side de-duplication of encrypted data in cloud storage,““ in Proc. 8th ACM SIGSAC Symp. Inf., Comput. Commun. Secure, 205, pp. 195–206.
8) M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, “Secure data deduplication," in Proc. 4th ACM Int Workshop Storage Secure. Survivability, 2008, pp. 1_10.
9) Y. Shin and K. Kim, “Differentially private client-side data deduplication protocol for cloud storage services," Secure. Commun. Netw, vol. 8, no. 12, pp. 2114_2123, 2015.