Vandana Publications
IJEMR

Volume-7, Issue-2, March-April 2017
International Journal of Engineering and Management Research
Page Number: 24-28

Authorized Data Deduplication Using Hybrid Cloud

Muhammed Noushad K[1], Ashwini Sasi[2], Jusammila MK[3], Megha CM[4], Happy A[5]
[1,2,3,4]B.Tech Students, Department of computer science and Engineering, Ammini college of engineering, Mankara, Palakkad, Kerala, INDIA
[5]Assistant Professor, Department of Computer Science And Engineering, Ammini College Of Engineering Palakkad, Kerala, INDIA

ABSTRACT

Now a days use of cloud computing is increasing rapidly. Cloud computing is very important in the data sharing application. Daily use of cloud is increasing. But the problem in cloud computing is every day data uploaded on the cloud, so increasing similar data in cloud. Therefore it can be reduce the size of similar data in cloud using the data Deduplication method. These method main aim is that remove duplicate data from cloud. It can also help to save storage space and bandwidth. This proposed method is to remove the duplicate data but in which user have assigned some privilege according to that duplication check & each user have their unique token. Cloud Deduplication is achieve using the hybrid cloud architecture. This proposed method is more secure and consumes less resources of cloud. Also it shown that proposed scheme has minimal overhead in duplicate removal as compared to the normal Deduplication technique. In this paper Content Level Deduplication as well as File Level Deduplication of file data is checked over the cloud..

Keywords**—** Authorization; data security; privilege; deduplication, credential; hybrid cloud

I. INTRODUCTION

Current era is cloud computing era. Now a days cloud computing has wide range of scope in data sharing. Cloud computing is provide large amount of virtual environment hiding the platform and operating systems of user. Users use the resources for sharing data. But users have to pay as per the use of resources of cloud. Now cloud service providers are offering cloud services with very low cost and also with high reliability. User can upload the large amount of data on cloud and shared data to millions of users. A cloud provider is offer different services such as infrastructure as a service, platform as a service, etc. Users not need to purchase the resources.

As the data is get uploaded by the user every day it is critical task to manage this ever increasing data on the cloud. Deduplication is best method to make well data management in the cloud computing. This method is becoming more attraction for data Deduplication. This method send the data over the network required small amount of data. This method have application in data management and networking. Data duplication is the technique of reducing the size of data Also it is the best compression method for the data Deduplication.

II. LITERATURE SURVEY

There are so many researches have been done to secure duplication check of data on cloud. The cloud storage and data Deduplication are two methods present in existing system. First method of the data Deduplication is perform as post processing method In this which data is first store on the storage device and then duplication check is applied on the data. The use of this method is there is no need to wait for calculating the hash function and the speed of storage not get downgrade. The main drawback with this system is that if storage capacity of the device is low then the file storage may get full. Some problem of this the post processing method is not useful at all because it checks the file after storing it on the cloud server.

Duplication check is the inline duplication check. It is check when new entries are to be added to the database the duplication of the file. It will checks for the block level duplication of the file before adding the new entry or new data to the database. This method have some drawback such as each time need to calculate the hash function which may lead to slower throughput of the storage device. But the some of the vendors have proof that data duplication check have same output in the inline and post processing method.

Duplication check is source duplication check in which the file duplicate contents are checks for duplication after storing it on the cloud server. This method of Deduplication is source data Deduplication in which data duplication is done at the side of the source. The file duplication is check before it get uploaded on the cloud server. The duplication is checked at the target level in which file get scanned periodically and hash get generated for the software can check for the hash value if both value get new matched with the existing hash value then the new file not get uploaded on the cloud server only link to that data is to be provide to the file user. If new file is to be added to the cloud server and it get match the hash function of the old file then it only remove the new file and just provide hard link to the old file resides on the cloud server.

III. SYSTEM ANALYSIS

System analysis is the term used to describe the process of collecting and analyzing facts in respect of existing operation of the solution of the situation prevailing so that an effective computerized system may be designed and implemented of proved feasible. It also diagnosis the problems and using that information recommends improvement to the system.

System analysis is the reduction of the entire system by studying the various operations performed and the relationship with the system and requirement of its successor. A system can be defined as an orderly grouping of independent component linked together according to a plan to achieve a specific objective.

System analysis may be considered as an interface between the actual problem and computer. Before a computer can perform, it is necessary to investigations are called system analyst. System analysis also embraces system design which is an activity concerned with the design of a computerized application based on the facts disclosed during the analysis stage. The same person who knows as the system analyst carries out both activities. In feasibility study in most cases project is being driven by a problem in the business.

IV. PROPOSED SYSTEM

In the proposed system we are doing duplication check in authenticated way. For the file duplication check proof of ownership is also set at the time of file upload the proof is added with the file this proof will decide the access privilege to the file. It is decide who can perform duplication check of the file. User is need to submit his file and proof of ownership of the file before sending the request to for the duplicate check Request to the cloud. When there is file on the cloud and also privileges of the user only that time to approved the duplicate check request.

V. CONTENTS

Encryption of files:

These instructions give you guidelines for preparing papers for IJEMR online JOURNALS. Use this document as a template. We are using secrete key resides at the private cloud to encrypt the user data .his key is used to convert plain text to cipher text and again for the decryption of the user data. To encrypt and decrypt we have used three basic functions as follows:

Key GenSE: It is generate the secrete file by using EncSE (k, M): In this we have generated a cipher text security parameter. In this k is the key generation algorithm. using formulae M is the text message and k is the secrete key. DecSE (k, C): In this we have to generate plain text using C is the cipher text and k is the encryption key.

Confidential encryption:

This ensures a data confidentiality in the duplication. User check the Convergent keys from each data set or original data and encrypt the data copy with the generated convergent key.User also add the tag for the data so that the tag will helps to*A. Encryption of File* We are using secrete key resides at the private cloud to encrypt the user data [1]. This key is used to convert plain text to cipher text and again for the decryption of the user data. To encrypt and decrypt we have used three basic functions as follows:

Key GenSE: It is generate the secrete file by using security parameter. In this k is the key generation algorithm.

EncSE (k, M): In this we have generated a cipher text using formulae M is the text message and k is the secrete key.

DecSE (k, C): In this we have to generate plain text using C is the cipher text and k is the encryption key.

MD5 Algorithm:

The MD5 algorithm is a widely used hash function producing a 128-bit hash value. Although MD5 was initially designed to be used as a cryptographic hash function, it has been found to suffer from extensive vulnerabilities. It can still be used as a checksum to verify data integrity, but only against unintentional corruption.

MD5 processes a variable-length message into a fixed-length output of 128 bits. The input message is broken up into chunks of 512-bit blocks (sixteen 32-bit words); the message is padded so that its length is divisible by 512. The padding works as follows: first a single bit, 1, is appended to the end of the message. This is followed by as many zeros as are required to bring the length of the message up to 64 bits fewer than a multiple of 512. The remaining bits are filled up with 64 bits representing the length of the original message, modulo 2^{64}.

The main MD5 algorithm operates on a 128-bit state, divided into four 32-bit words, denoted *A*, *B*, *C*, and *D*. These are initialized to certain fixed constants. The main algorithm then uses each 512-bit message block in turn to modify the state. The processing of a message block consists of four similar stages, termed *rounds*; each round is composed of 16 similar operations based on a non-linear function *F*, modular addition, and left rotation. Figure 1 illustrates one operation within a round. There are four possible functions *F*; a different one is used in each round:

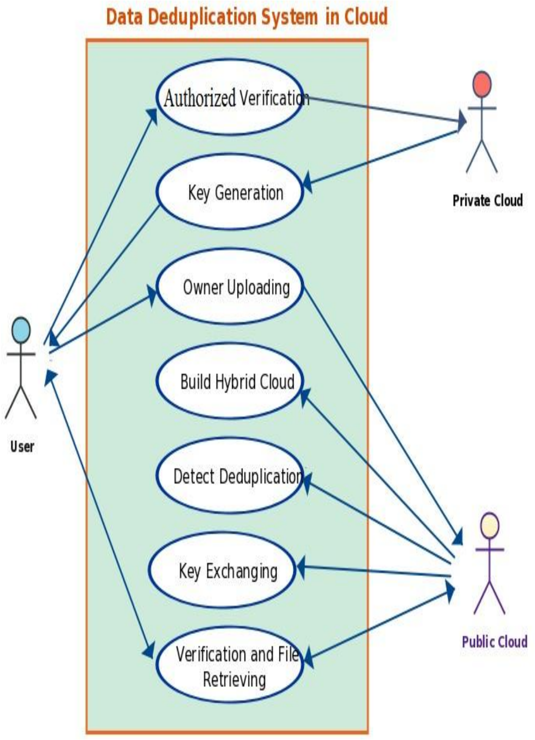
F(B,C,D) = (B AND C) OR (NOT B AND D)
G(B,C,D) = (B AND D) OR (C AND NOT D)
H(B,C,D) = B XOR C XOR DI
(B,C,D) = C XOR (B OR NOT D)

denote the XOR, AND, OR and NOT operations respectively.

VI. DETAILED DESIGN

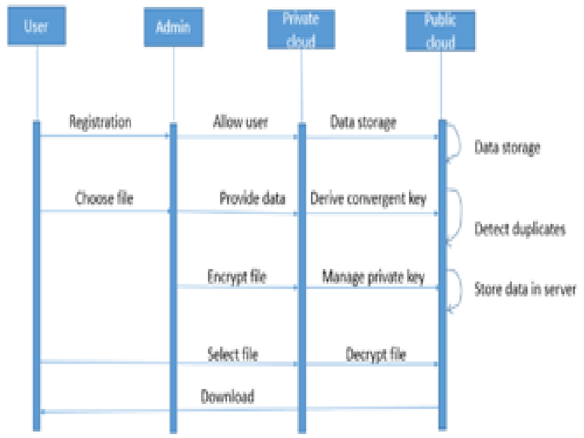
Use case Diagram:

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well.



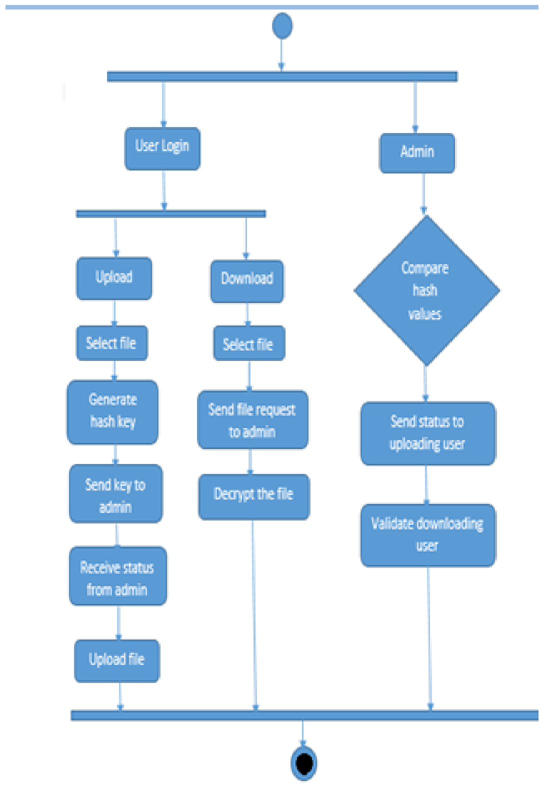
Sequence Diagram:

A Sequence diagram is an interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.



Activity diagram :

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams are intended to model both computational and organizational processes (i.e. workflows). Activity diagrams show the overall flow of control.

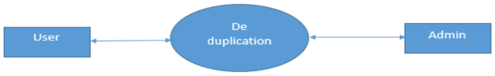


Data Flow Diagram:

A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modelling its process aspects. A DFD is often used as a preliminary step to create an overview of the system, which can later be elaborated. DFDs can also be used for the visualization of data processing.

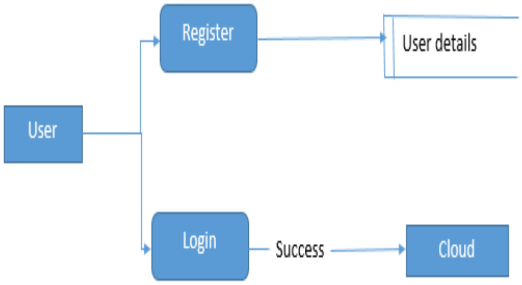
DATA FLOW DIAGRAM

Level 0 :



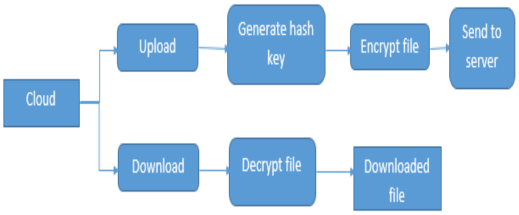
DATA FLOW DIAGRAM

Level 1 :



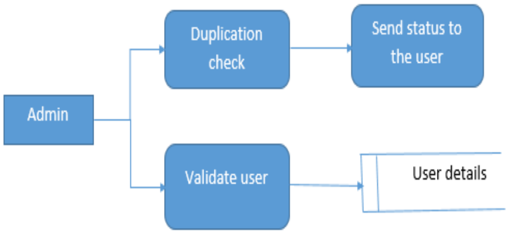
DATA FLOW DIAGRAM

Level 1.1 :



DATA FLOW DIAGRAM

Level 2 :

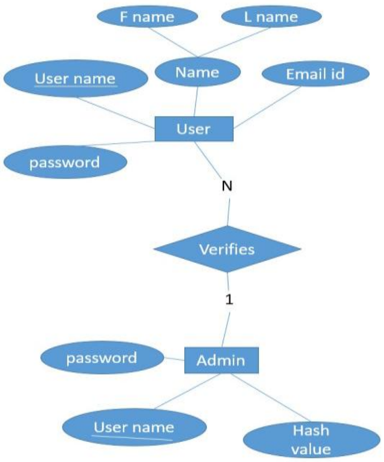


ER Diagram:

An entity-relationship diagram (ERD) is a graphical representation of an information system that shows the relationship between people, objects, places, concepts or events within that system. An ERD is a data modeling technique that can help define business processes and can be used as the foundation for a relational database.

While useful for organizing data that can be represented by a relational structure, an entity-relationship diagram can't sufficiently represent semi-structured or unstructured data, and an ERD is unlikely to be helpful on its own in integrating data into a preexisting information system. Three main components of an ERD are the entities, which are objects or concepts that can have data stored about them, the relationship between those entities, and the cardinality, which defines that relationship in terms of numbers.

ER DIAGRAM



VII. CONCLUSION

Here It is conclude that this proposed system data

Deduplication of file is done with authorized way and securely perform all operations. In this system it also proposed new duplication check method which generate the token for the private file and check content level deduplication. User need to submit the privilege along with the convergent key as a proof of ownership. It solved more critical part of the cloud data storage which is only tolerated by different methods. Proposed methods ensure the data duplication securely. Performance of this system is 98 % more than existing system.

REFERENCE

[1] IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEM VOL:PP NO:99 YEAR 2014,A Hybrid cloud approach for secure authorized deduplication, Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou

[2] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. InUSENIX Security Symposium, 2013.

[3] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.

[4] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2013.

[5] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups.Proc.of APSYS, Apr 2013.

[6] 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), 2012.

[7] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on AppliedComputing, pages 441–446. ACM, 2012.

[8] R. D. Pietro and A. Sorniotti. Boosting efficiency and security in proof of ownership for deduplication. In H. Y. Youm and Y. Won, editors, ACM Symposium on Information, Computer and communications Security, pages 81–82. ACM.

[9] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.

[10] C.-K Huang, L.-F Chien, and Y.-J Oyang, “Relevant TermSuggestion in Interactive Web Search Based on ContextualInformation in Query Session Logs,” J. Am. Soc. for Information science and Technology, vol. 54, no. 7, pp. 638-649, 2003.