Secure De-Duplication using Convergent Keys (Convergent Cryptography) for Cloud Storage

Madhuri Kavade
PG student
JSPM's B.S.I.O.T.R
Wagholi, Pune, India

A.C. Lomte
Assistant Professor
JSPM's B.S.I.O.T.R
Wagholi, Pune, India

ABSTRACT
One very important challenges of today's cloud storage services is that the management of the ever-increasing amount of data. To create data management scalable, de-duplication has been a widely known technique to condense space for storing and transfer information measure in cloud storage. Rather than keeping multiple data copies with an equivalent content, de-duplication eliminates redundant data by keeping just one physical copy and referring different redundant data to it copy. Convergent Encryption, additionally called content hash keying, could be a cryptosystem that produces identical ciphertext from identical plaintext files. Data de-duplication could be a technique for eliminating duplicate copies of data, and has been wide employed in cloud storage to scale back space for storing and transfer information measure. Promising because it is, associate degree arising challenge is to perform secure de-duplication in cloud storage. Currently daily the foremost arising challenge is to perform secure de-duplication in cloud storage. Though Convergent Encryption has been extensively adopted for secure de-duplication, a essential issue of creating Convergent Encryption sensible is to with efficiency and dependably manage a large range of convergent keys. we have a tendency to initial introduce a baseline approach during which file level de-duplication is performed and it doesn't support on-line approach. For that purpose the current context have a tendency to formally address the matter of achieving economical and reliable key management in secure de-duplication. Within the planned design Dupkey, users don't ought to manage any keys on their own. It implements de-duplication at block level and provides a lot of security. Dupkey is on-line approach and supply de-duplication and reduces DB size on cloud.

Keywords
De-duplication, convergent encryption, convergent key, cloud Storage.

1. INTRODUCTION
Cloud computing is obtaining a lot of and a lot of well-liked because it will offer affordable and on demand use of vast storage and process resources. With the explosive growth of on-line digital contents, cloud storage focuses on effectively coalescing storage resources for higher power utilization and value effectiveness. Because of the volume of data grows, additionally increasing storage infrastructure price, management price and human administration price. So in cloud storage systems, reducing the number of data that require to be transferred, stored, and managed becomes a vital, and it additionally advantages for application performance, storage prices and body overheads.

Since the information is keep and processed on the cloud infrastructure or platform, that the data owner doesn't have full management of. just in case of the user data being peeked, leaked or changed by the cloud supplier or different adversaries, cryptography become a necessary before change data into the cloud. However, de-duplication and cryptography, to an excellent extent, conflicts with one another. De-duplication takes advantage of data similarity so as to realize storage reduction. whereas cryptography makes ciphertext indistinguishable from in theory random data, i.e., encrypted data are invariably distributed every which way, thus identical plaintext encrypted by every which way generated crypto logical keys will terribly doubtless have totally different ciphertexts that cannot be de-duplicated. data De-duplication could be a technique for decreasing quantity of space for storing for organization must save its data. In several organizations, the storage system have duplicate copies of the many items of {information} and its information. as an example, a equivalent file is also saved in many places by several users or 2 or a lot of files a lot of files that aren't equivalent should embody abundant of an equivalent data. De-duplication removes these additional copies by saving only 1 copy of the information.

A method that has been wide adopted is cross-user De-duplication. the straightforward plan behind De-duplication is to store duplicate data (either files or blocks) just once. Therefore, if a user needs to transfer a file (block) that is already keep, the cloud supplier can add the user to the owner list of that file (block). to create data management scalable in cloud computing, De-duplication has been a data technique and has attracted a lot of and a lot of attention recently. Data De-duplication could be a specialized data compression technique for eliminating duplicate copies of repetition data in storage. The technique is employed to enhance storage utilization and may even be applied to network data transfers to scale back the amount of bytes that has got to be sent. rather than keeping multiple data copies with an equivalent content, De-duplication eliminates redundant data by keeping just one physical copy and referring different redundant data to it copy. De-duplication will happen at either the file level or the block level. For file level De-duplication, it eliminates duplicate copies of an equivalent file. De-duplication may also happen at the block level, which eliminates duplicate blocks of data that occur in non-identical files. though data De-duplication brings plenty of advantages, security and privacy issues arise as users sensitive data are vulnerable to each business executive and outsider attacks. ancient cryptography, whereas providing data confidentiality, is incompatible with data De-duplication. Specifically, ancient cryptography needs totally different users to write in code their data with their own keys. Thus, identical data copies {of totally different or of various} users can cause different cipher-texts, creating De-duplication not possible.to resolve the higher than conflict, most of current work use Convergent

Encryption to induce identical ciphertext from identical plaintext, however the data outflow in such cryptography theme will be unacceptable. What's worse, the Convergent Encryption provides one settled rework from a selected plaintext to the ciphertext, which exposes a lot of vulnerability.

Data De-duplication methods will be classified in line with the fundamental data units they handle. during this respect there are 2 main data De-duplication strategies: (1) File-level De-duplication, during which solely one copy of every file is keep. 2 or a lot of files are known as identical if they need an equivalent hash price. this can be a awfully well-liked style of service offered in multiple products; (2) Block-level De-duplication, that segments files into blocks and stores solely one copy of every block. The system might either use mounted sized blocks or variable-sized chunks. In terms of the design of the De-duplication answer, there are 2 basic approaches. within the target-based approach De-duplication is handled by the target data-storage device or service, whereas the shopper is unaware of any De-duplication that may occur. This technology improves storage utilization, however doesn't save information measure. On the opposite hand, supply primarily based De-duplication acts on the information at the shopper before it's transferred. Specifically, the shopper computer code communicates with the backup server (by causing hash signatures) to envision for the existence of files or blocks. Duplicates are replaced by pointers and also the actual duplicate data is rarely sent over the network. The advantage of this approach is that it improves each storage and information measure utilization.[7]

2. MOTIVATION

With the possibly infinite space for storing offered by cloud suppliers, users tend to use the maximum amount house as they will. The vendors perpetually explore for techniques aimed to reduce redundant data and maximize house savings. that the technique which may offer each of those options should have to be compelled to implement. This motivates United States to introduce a method known as data de-duplication. the straightforward plan behind de-duplication is to store duplicate data (either files or blocks) just once. Therefore, if a user needs to transfer a file (block) that is already keep, the cloud supplier can add the user to the owner list of that file (block).De-duplication has evidenced to realize high house and value savings and lots of cloud storage suppliers are presently adopting it. On the opposite hand, de-duplication introduces new security risks. thus there's want of secure de-duplication. Many systems are developed to produce secure storage however ancient cryptography techniques aren't appropriate for de-duplication functions. Deterministic cryptography, particularly Convergent Encryption, could be a sensible candidate to realize each confidentiality and de-duplication

3. BACKGROUND AND RELATED WORK

The idea depend upon ancient cryptography during which user encrypts data with associate degree freelance secret key. a number of the studies propose to use threshold secrete sharing to take care of strength of key management however they are doing not think about de-duplication commonplace cryptography makes de-duplication not possible. [2, 3, 4]

It proposes the proofs of possession for de-duplication system specified shopper will with efficiency persuade the cloud storage server that he/she owns a file while not uploading file itself. However these approaches don't think about data privacy. [10, 7]

The systems formalize a replacement cryptological primitive, Message-Locked cryptography (MLE), wherever the key beneath that cryptography and cryptography are performed is itself derived from the message. MLE provides the simplest way to realize secure de-duplication, a goal presently targeted by varied cloud storage suppliers. This work shows that MLE could be a primitive of each sensible and theoretical interest. It provides definitions each for privacy and a style of integrity that we have a tendency to decision tag consistency. It additionally provides space-efficient secure outsourced storage. However they are doing not address a way to minimize the key management overhead. [6]

This study focuses on the privacy implications of cross-user de-duplication. they need incontestable however de-duplication will be used as a aspect channel that reveals data regarding the contents of files of different users. in an exceedingly totally different situation, De-duplication will be used as a covert channel by that despiteful computer code will exchange a couple of words with its center, despite any firewall settings at the attacked machine. as a result of the high savings offered by cross-user de-duplication, cloud storage suppliers are unlikely to prevent mistreatment this technology. we have a tendency to so propose easy mechanisms that alter cross user de-duplication whereas to an excellent extent reducing the danger of data outflow. [8]

In this specific context study of on-line file storage services are introduced. whereas varied of those services offer basic practicality like uploading and retrieving files by a selected user, a lot of advanced services provide options like shared folders, time period collaboration, and diminution of data transfers or infinite space for storing. they need given an outline of existing file storage services and examine Dropbox, a sophisticated file storage answer, in depth. By discussing security enhancements for contemporary on-line storage services normally, and Dropbox particularly. To forestall our attacks cloud storage operators ought to use data possession proofs on shoppers, a method that has been recently mentioned solely within the context of assessing trust in cloud storage operators. [9]

3.1 Traditional cryptography

To protect the confidentiality of outsourced data, varied cryptological solutions are planned within the literature (e.g., [2], [3], [4]). Their plan builds on traditional(symmetric) cryptography, during which every user encrypts data with associate degree freelance secret key. Some studies [10], [11] propose to use threshold secret sharing [12] to take care of the strength of key management. However, the higher than studies don't think about de-duplication. mistreatment ancient cryptography, totally different users can merely write in code identical data copies with their own keys, however this may cause totally different cipher-texts and therefore build De-duplication not possible.

3.2 Convergent Encryption

Convergent cryptography [5] provides a possible choice to place operative data confidentiality whereas realizing de-duplication. It encrypts/decrypts a data copy with a convergent key that comes by computing the cryptological hash price of the content of the information copy itself [5]. When key generation and encryption, users retain the keys and send the cipher-text to the cloud. Since cryptography is settled, identical data copies can generate an equivalent

convergent key and also the same cipher text. this permits the cloud to perform de-duplication on the cipher-texts. The cipher-texts will solely be decrypted by the corresponding data homeowners with their convergent keys.

4. IMPLEMENTATION DETAILS

4.1 Existing System

In existing system they're mistreatment commonplace cryptography theme for distinguishing duplicate blocks of data in cloud storage. In Cloud Storage, commonplace cryptography of identical files produces same key and same cipher text. thus data First State duplication in encrypted data is not possible. once user lost the key, there was not possible to revive the initial content of the file. it's compromised by attackers, then the user data are going to be leaked. the present cryptography rule doesn't maintain the key management theme.

Disadvantages

1. Data confidentiality isn't achieved.
2. Data First State Duplication isn't attainable in commonplace cryptography theme.
3. User cannot restore the initial content of the file.

4.2 Proposed System

We propose Dupkey, a replacement construction during which users don't ought to manage any keys on their own however instead firmly distribute the convergent key shares across multiple server instances. it's not possible to use physical key servers to distribute the keys. rather than mistreatment physical servers we've got created multiple instances for the servers. This makes our system better. Our approach supports each file-level and block level de-duplication. aside from that our main contribution during this project is we've got else another module to envision the information size needed for the approaches. and that we prove that in our planned system the information size needed is a smaller amount than that of the present system.

4.3 Algorithmic Explanation

4.3.1 Load Equalization Algorithm

We have used the load equalization rule to store our keys on distributed server. Our controller module takes a call wherever to store key. it'll additionally keep the data entry for the keys that we have a tendency to store. we have a tendency to use choose the Random storage server from the gathering of storage server.

GetServerForData()
This technique can come back the sever wherever data can get store. It'll browse all servers that are listed on to the "serverCfg". serverCfg is that the Property information gift at the Controller Server which can hold all accessible server details.This technique can browse unmarshal "serverCfg" information with the assistance of Java XML information Binding. JAXB framework can come back United States the List of obtainable server. we have a tendency to used the Random rule to induce the random range at intervals vary of "0 to (serverList.Size() -1)".Once we have a tendency to generate the random range we are going to get the Server store at specific index and selected an equivalent server to store the information

4.3.2 Random Class statistics

An instance of this category is employed to come up with a stream of pseudorandom numbers. the category uses a 48-bit seed, that is changed employing a linear congruential formula. If 2 instances of Random are created with an equivalent seed, and also the same sequence of technique calls is formed for every, they'll generate and come back identical sequences of numbers. so as to ensure this property, specific algorithms ar nominal for the category Random. Java implementations should use all the algorithms shown here for the category Random, for the sake of absolute immovability of Java code. However, subclasses of sophistication Random are permissible to use different algorithms, see you later as they adhere to the final contracts for all the ways. The algorithms enforced by category Random use a protected utility technique that on every invocation will offer up to thirty two pseudorandom solely generated bits.

4.3.3 Convergent Encryption

Convergent cryptography provides data confidentiality in de-duplication. A user (or data owner) derives a convergent key from every original data copy and encrypts the information copy with the convergent key. additionally, the user derives a tag for the information copy, specified the tag are going to be accustomed sight duplicates. A Convergent Encryption theme will be outlined with four primitive functions:

1. KeyGen(M)-> K is the key generation rule that maps a data copy M to a convergent key K

2. Encrypt(K,M) -> C is that the radically symmetrical cryptography rule that takes each the convergent key K and also the data copy M as inputs then outputs a cipher text C

3. Decrypt(K,C) -> M is that the cryptography rule that takes each the cipher text C and also the convergent key K as inputs then outputs the initial data copy M

4. TagGen(M) ->T(M) is that the tag generation rule that maps the initial data copy M and outputs a tag T(M). we have a tendency to permit TagGen to come up with a tag from the corresponding cipher text by mistreatment T(M)=TagGen(C), wherever C=Encrypt(K,M).

4.3.4 Design And Modules

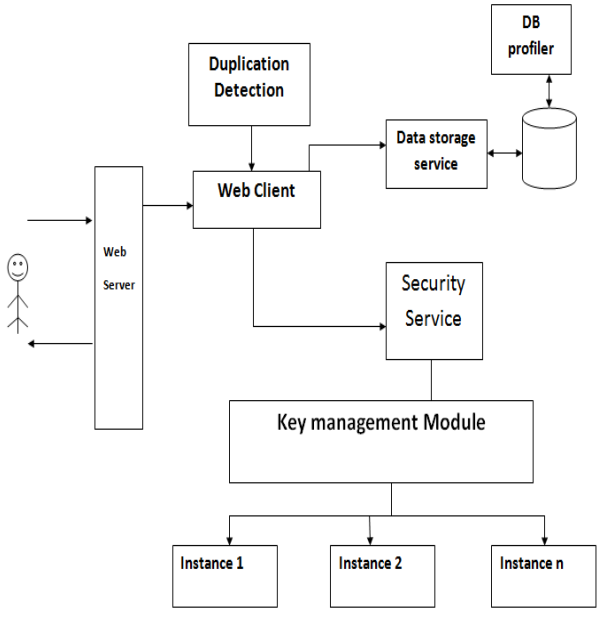

Fig. 1: Architecture of the proposed system

The Figure 1 represents the design for planned system. The flow and outline of every modules present within the system mentioned below-

1] Web Client- This module is liable for interaction with the online shopper.

2] Security Service-This module can handle cryptography and cryptography of the information. This module is additionally accountable for hash code generation.

3] Key management Module- there's one controller that will handle multiple servers. It additionally calls load equalization algorithm for distribution of keys on totally different servers i.e. instances of the server. For this purpose jaxws java framework is used which can install on every sever instance with storage for the key.

4] Data storage Service- It merely stores the information. This is the actual physical storage system wherever solely single copy of the data is keep. It additionally handles all DB operations

5] Duplication Detection- it's able to sight the duplicates of the information additionally verification of the information is taken place by this modules.

6] DB profiler- It repeatedly checks the standing of the information and generates the information report. It compares the information size needed for each the systems

5. RESULTS

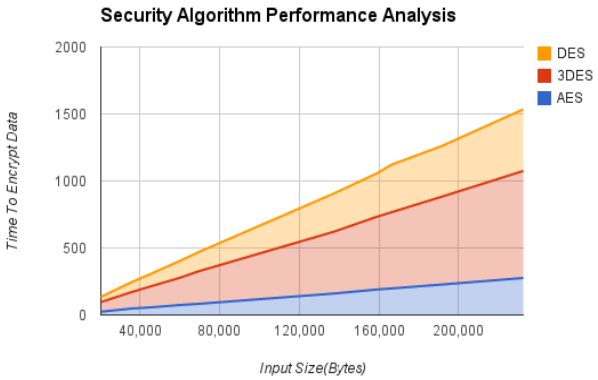

Fig. 2: Performance Analysis for security algorithm

In this section we have a tendency to represent the results of sensible work. The higher than diagram shows the performance analysis for the protection algorithms. we have a tendency to are mistreatment AES rule for encryption/decryption of the information to be uploaded on the cloud storage. From Fig.2 it's clear that AES is a lot of economical than other algorithms. The time needed to finish the method of cryptography and cryptography is a smaller amount as compared to different Algorithms. Thus on win the higher performance we've got select the AES rule.

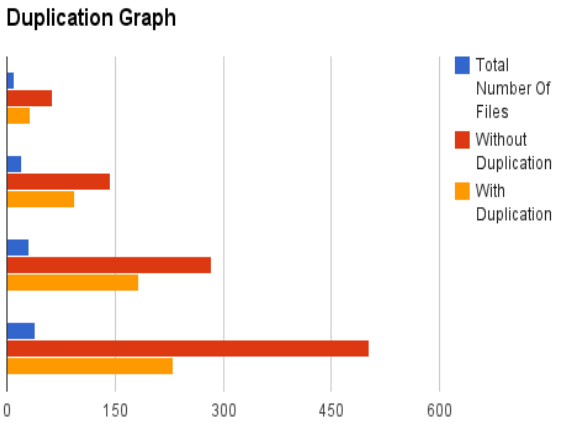

Fig. 3: Performance of application with De-duplication and without De-duplication

The fig.3 shows that application for cloud storage that states De-duplication provides us higher performance. Whereas the applications that doesn't uses the De-duplication needs large amount of files.

6. CONCLUSION

In this paper, it has a tendency to propose Dupkey, a competent and consistent convergent key management theme for secure de-duplication. Dupkey applies de-duplication among convergent keys and distributes convergent key shares across multiple key servers (key server instances), whereas protective security of convergent keys and confidentiality of outsourced data. It also demonstrates that comparison of each approaches shows that our approach would force less information size as compared to the baseline approach. The current context can be applied to only text file formats so in future it can be extended for other file formats such as audio, image, video.

7. ACKNOWLEDGEMENT

Special thanks go to authors J. Li, X. Chen, M. Li, J. Li,P. Lee, and W. Lou contributed to this paper for their valuable comments and sharing their knowledge and idea.

8. REFERENCES

[1] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure De-duplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2014.

[2] Q. Wang, C. Wang, K. Ren, W. Lou, and J. Li, Enabling Public Auditability and Data Dynamics for Storage Security in Cloud Computing, IEEE Trans. Parallel Distrib. Syst., vol. 22, no. 6, pp. 547-567, May 2010

[3] W. Wang, Z. Li, R. Owens, and B. Bhargava, Secure and Efficient Access to Outsourced Data, in Proc. ACM CCSW, Nov. 2007, pp. 66-66.

[4] A. Yun, C. Shi, and Y. Kim, On Protecting Integrity and Confidentiality of Cryptographic File System for Outsourced Storage, in Proc. ACM CCSW, Nov. 2007, pp. 67-76.

[5] 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. 69-624

[6] M. Bellare, S. Keelveedhi, and T. Ristenpart, Message-Locked Encryption and Secure De-duplication, in Proc. IACR Cryptology ePrint Archive, 208, pp. 276-32208:63

[7] J. Gantz and D. Reinsel, The Digital Universe in 2020: Big Data, Bigger Digital Shadows, Biggest Growth in the Far East, Dec. 208. [Online]. Available: http://www.emc.com/collateral/analystreports/ idc-the-digitaluniverse- in-2020.pdf.

[8] D. Harnik, B. Pinkas, and A. Shulman-Peleg, Side Channels in Cloud Services: De-duplication in Cloud Storage, IEEE Security Privacy, vol. 8, no. 6, pp. 40-47, Nov./Dec. 2010.

[9] M. Mulazzani, S. Schrittwieser, M. Leithner, M. Huber, and E. Weippl, Dark Clouds on the Horizon: Using Cloud Storage as Attack Vector and Online Slack Space, in Proc. USENIX Security, 2010, p. 5.

[10] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, Proofs of Ownership in Remote Storage Systems, in Proc. ACM Conf. Comput. Commun. Security, Y. Chen, G. Danezis, and V. Shmatikov, Eds., 2010, pp. 491-500.

[11] A.D. Santis and B. Masucci, Multiple Ramp Schemes, IEEE Trans. Inf. Theory, vol. 45, no. 5, pp. 1720-1728, July 1999.

[12] A. Shamir, How to Share a Secret, Commun. ACM, vol. 22, no. 11,pp. 612-613, 1979

[13] S. Kamara and K. Lauter, Cryptographic Cloud Storage, in Proc. Financial Cryptography: Workshop Real-Life Cryptography Protocols Standardization, 2010, pp. 16-147.