ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Survey on Hybrid Big Data Approach For Secure Authorized DE-duplication

Gharge Nitin S[1], Bhor Ganesh G[1], Karajange Ganesh M[1], Jagdale Pravin R[1], Prof.Kadam Yogesh V[2]

B.E. Student, Dept. of Computer Engineering, Bharati Vidyapeeth Lavale, Pune, India[1]

Assistant Professor, Dept. of Computer Engineering, Bharati Vidyapeeth Lavale, Pune, India[2]

ABSTRACT: As more corporate and private users outsource their data to Big Data storage providers, recent data breach incidents make end-to end encryption an increasingly prominent requirement. Unfortunately, semantically secure encryption schemes render various cost-effective storage optimization techniques, such as data Deduplication, ineffective. We present a novel idea that differentiates data according to their popularity. Based on this idea, we design an encryption scheme that guarantees semantic security for unpopular data and provides weaker security and better storage and bandwidth benefits for popular data. This way, data Deduplication can be effective for popular data, whilst semantically secure encryption protects unpopular content.In individualized computing gadgets that depend on a distributed storage environment for information reinforcement, a fast approaching test confronting source de-duplication for cloud reinforcement administrations is the low de-duplication effectiveness because of a blend of the asset escalated nature and the constrained framework assets. Information de-duplication is one of imperative information pressure strategies for disposing of copy duplicates of rehashing information, and has been generally utilized as a part of distributed storage to lessen the measure of storage room and spare transmission capacity. To secure the secrecy of delicate information while supporting de-duplication, the merged encryption strategy has been proposed to encode the information before outsourcing. To better ensure information security, this paper makes the main endeavour to formally address the issue of approved information deduplication. Not quite the same as conventional de-duplication frameworks, the differential benefits of clients are further considered in copy check other than the information itself. We additionally show a few new de-duplication developments supporting approved copy check in cross breed cloud design.

KEYWORDS:de-duplication; data;encryption; copy; storage.

I. INTRODUCTION

Data deduplication is one of the important data compression techniques for eliminating duplicate copies data, and has been widely in cloud storage to reduce the amount of storage space and save bandwidth.to protect the confidentiality of sensitive data while supporting deduplication.Today's cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data isbeing stored in the cloud and shared by users with specified privileges, which define the access rights of the stored data. Onecritical challenge of cloud storage services is the management of the ever-increasing volume of data. To make data managementscalable in cloud computing, deduplication has been a well-known technique and has attracted more and more attention recently.Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. Thetechnique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytesthat must be sent. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data bykeeping only one physical copy and referring other redundant data to that copy. Deduplication can take place at either the filelevel or the block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can also takeplace at the block level, which eliminates duplicate blocks of data that occur in non-identical files. In such an authorized deduplication system, each user is issued a set of privileges during system initialization each file uploaded to Big Data is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Although data deduplication brings a lot of benefits, security and privacy concerns arise as users' sensitive data are susceptible to both inside and outside attacks. Traditional encryption, while providing data confidentiality, is incompatible with data deduplication. Specifically, traditional encryption requires different users to encrypt their data with their own keys. Thus, identical data copies of different users will lead to different cipher text, making deduplication impossible. Convergent encryption has been proposed to enforce data confidentiality while making deduplication feasible. It encrypts/decrypts a data copy with a convergent key, which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud.

II. GOALS AND OBJECTIVES

Goals:

To avoid duplication of files and to minimize the space of the Big Data for the data and also to fasten the search for the files so that duplication process can be catalysed.

Objectives:

- Accepting files from the user at the Big Data end.
- Hashing the file content
- Searching for the files for the duplication using bloom filtering
- Decision for file storage in Big Data servers by Big Data controller
- Encrypting the files using random key
- Key management

III. LITERATURE SURVEY

Convergent encryption has been proposed to enforce data confidentiality while making deduplication feasible. It encrypts/decrypts a data copy with a *convergent key*, which is obtained by computing the cryptographic hash value of the content of the data copy which is denoted in diagram. After key generation and data encryption, users retain the keys and send the ciphertext to the Big Data. Since the encryption operation is deterministic and is derived from the data content, identical data copies will generate.

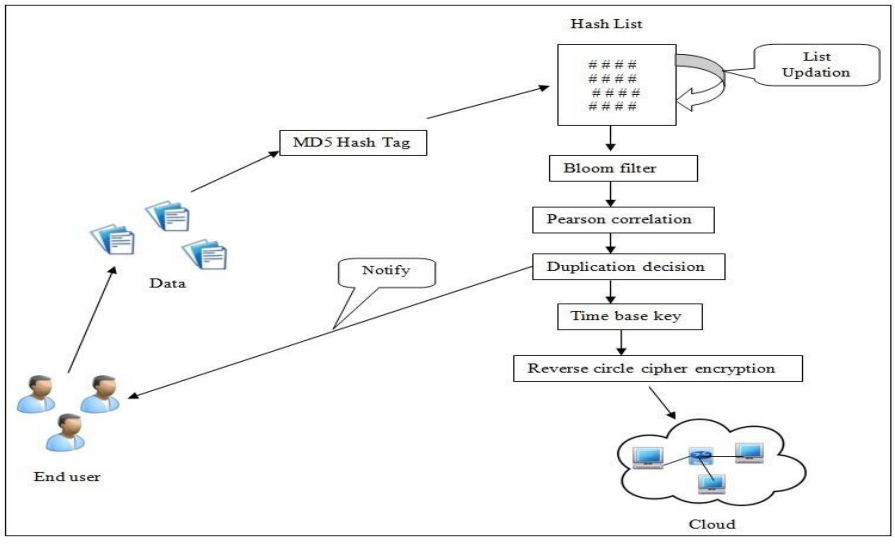

Fig.1.System Structure

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Same convergent key and hence the same ciphertext. To prevent unauthorized access, a secure proof of ownership protocol is also needed to provide the proof that the user indeed owns the same file when a duplicate is found. After the proof, subsequent users with the same file will be provided a pointer from the server without needing to upload the same file

A user can download the encrypted file with the pointer from the server, which can only be decrypted by the corresponding data owners with their convergent keys. Thus, convergent encryption allows the Big Data to perform DE-duplication on the cipher texts and the proof of ownership prevents the unauthorized user to access the file. However, previous deduplication systems cannot support differential authorization duplicate check, which is important in many applications. In such an authorized deduplication system, each user is issued a set of privileges during system initialization each file uploaded to Big Data is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files.

IV. EXISTING SYSTEM: & EXISTING METHODOLOGY

To make data management scalable in Big Data computing, deduplication has been a well-known technique and has attracted more and more attention recently. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Deduplication can take place at either the file level or the block level. For file level DE-duplication, it eliminates duplicate copies of the same file. Deduplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files.

V. DATA DE-DUPLICATION TYPES

1. File-level de-duplication

It is commonly known as single-instance storage, file-level data de-duplication compares a file that has to be archivedor backup that has already been stored by checking all its attributes against the index.
The index is updated and stored only if the file is unique, if not than only a pointer to the existing file that is stored references. Only the single instance of file is saved in the result and relevant copies are replaced by "stub" which points to the original file.

2. Variable block level de-duplication

Compares varying sizes of data blocks that can reduce the chances of collision, stated Data link's Orlandini.

3. Block-level de-duplication

Block-level data de-duplication operates on the basis of sub-file level. As the name implies, that the file is being brokeninto segments blocks or chunks that will be examined for previously stored information vs. redundancy.The popular approach to determine redundant data is by assigning identifier to chunk of data, by using hash algorithmfor example – it generates a unique ID to that particular block. The particular unique Id will be compared with thecentral index. In case the ID is already present, then it represents that before only the data is processed and storedbefore .Therefore only a pointer reference is saved to the previously stored data. If the ID is new and does not exist,then that block is unique. The unique chunk is stored and the unique ID is updated in the Index.The size of the chunk which needs to be checked varies from vendor to vendor. Some will have fixed block sizes, whilesome others use variable block sizes likewise few may also change the size of fixed block size for sake of confusing.Block sizes of fixed size may vary from 8KB to 64KB but the main difference with it is the smaller the chunk, than itwill be likely to have opportunity to identify it as the duplicate data. If less data is stored than it obviously meansgreater reductions in the data that is stored. The only major issue by using fixed size blocks is that in case if the file is modified and the de-duplication result uses the same previously inspected result than there will be chance of notidentifying the same redundant data segment, as the blocks in the file would be moved or changed, than they will shiftdownstream from change, by offsetting the rest of comparisons.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

VI. DEDUPLICATION SITUATIONS OCCURRENCE'S

A. CLIENT-SIDE DE-DUPLICATION

Client-side de-duplication is different from all other forms of de-duplication in that duplicate data is first only identifiedbefore it has to be sent over the network. This will definitely create burden on the CPU but at the same time reduces the load on the network. Leveraging client side de-duplication gives us lot of advantages, because of the high level of duplicate information in a virtual environment and also the fact that data is sent across a highly congested IP network.

B. TARGET-BASED DE-DUPLICATION

Target de-duplication will remove the redundancies from a backup transmission as and when it passes through an appliance that is present between the source and the target. Unlike source de-duplication, the target de-duplication does not reduce the total amount of data that need to be transferred across a WAN or LAN during the backup, but it reduces the amount of storage space required.

C. GLOBAL DE-DUPLICATION

Global data de-duplication is a procedure of eliminating redundant data when backing up data to more number of deduplication devices. This situation might require backing up data to more than one target de-duplication system or in the case of source de-duplication it might require backing up to multiple backup nodes which will be they are be backing up multiple clients.

D. INLINE DE-DUPLICATION

Inline de-duplication is the most economic and efficient method of de-duplication. It reduces the raw disk space needed in system, since the full, not still de-duplicated data set would never be written to disk. Inline de-duplication reduces time to disaster recovery readiness because the system does not need to wait to utilize the entire data set and before it begins duplication of data at the remote side, it is de-duplicated.

VII. ADVANTAGES OF DEDUPLICATION

- Effectively increased network bandwidth - In case De-duplication takes place at the source end than no copies of data need to be transmitted via the network.
- Greener‖ environment - Fewer cubic feet of space is required to store the data in both primary and remote locationsand less electricity is needed.
- Buying and maintaining less storage will return us with the faster returns.

VIII. DISADVANTAGES OF EXISTING SYSTEM

- Users' sensitive data are susceptible to both insider and outsider attacks.
- Some times deduplication impossible.

IX. CONCLUSION AND FUTURE WORK

Cloud computing has reached a maturity that leads it into a productive phase. This means that most of the main issues with cloud computing have been addressed to a degree that clouds have become interesting for full commercial exploitation. This however does not mean that all the problems listed above have actually been solved, only that the according risks can be tolerated to a certain degree. Cloud computing is therefore still as much a research topic, as it is a market offering.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

REFERENCES

1. De-duplication of encrypted big data in cloud(IEEE base).
2. Open Security Foundation: Data Loss DB (http://datalossdb.org/).
3. Meister, D., Brinkmann, A.: Multi-level comparison of data deduplication in ī backup scenario. In: SYSTOR '09, New York, NY, USA, ACM (2009) 8:1{8:12
4. Mandagere, N., Zhou, P., Smith, M.A., Uttamchandani, S.: Demystifying data deduplication. In: Middleware '08, New York, NY, USA, ACM (2008) 12{17
5. Aronovich, L., Asher, R., Bachmat, E., Bitner, H., Hirsch, M., Klein, S.T.: The design of a similarity based deduplication system. In: SYSTOR '09. (2009) 6:1{6:14
6. Dutch, M., Freeman, L.: Understanding data de-duplication ratios. SNIA forum (2008) http://www.snia.org/sites/default/files/Understanding_Data_Deduplication_Ratios- 0080718.pdf.
7. Harnik, D., Margalit, O., Naor, D., Sotnikov, D., Vernik, G.: Estimation of deduplication ratios in large data sets. In: IEEE MSST '12. (april 2012) 1 {11
8. Harnik, D., Pinkas, B., Shulman-Peleg, A.: Side channels in Big Data services: Deduplication in Big Data storage. Security Privacy, IEEE 8(6) (nov.-dec. 2010) 40 {47
9. Halevi, S., Harnik, D., Pinkas, B., Shulman-Peleg, A.: Proofs of ownership in remote storage systems. In: CCS '11, New York, NY, USA, ACM (2011) 491{500
10. Di Pietro, R., Sorniotti, A.: Boosting e_ciency and security in proof of ownership for deduplication. In: ASIACCS '12, New York, NY, USA, ACM (2012) 81{82
11. Douceur, J.R., Adya, A., Bolosky, W.J., Simon, D., Theimer, M.: Reclaiming space from duplicate _les in a serverless distributed _le system. In: ICDCS '02, Washington, DC, USA, IEEE Computer Society (2002) 617{632
12. Storer, M.W., Greenan, K., Long, D.D., Miller, E.L.: Secure data deduplication. In: StorageSS '08, New York, NY, USA, ACM (2008) 1{10
13. Bellare, M., Keelveedhi, S., Ristenpart, T.: Message-locked encryption and secure deduplication. In: Advances in Cryptology{EUROCRYPT 2013. Springer 296{312
14. Xu, J., Chang, E.C., Zhou, J.: Weak leakage-resilient client-side deduplication of encrypted data in Big Data storage. In: 8th ACM SIGSAC symposium. 195{206.

BIOGRAPHY

Bhor Ganesh G, Gharge Nitin S, Karajange Ganesh M, Jagdale Pravin R. all are students of Bharati Vidyapeeth Collage of engineering lavale,Pune. Studied in last year of Computer Engineering.

Kadam Yogesh v.is a Assistant professor in the Department of computer engineering, College of of Bharati Vidyapeeth Collage of engineering lavale, Pune, Savitribai phule Pune University.