Secure Deduplication of Encrypted Data in Cloud

Mr.K.Sundeep Saradhi[1], Nikhil Kolla[2],Sarada Mulukuri[3], N R S RamTeja[4], Venkatesh Lankalapalli[5]
[1]*Asst.Professor,* [2,3,4,5]*B.Tech Students*
Dept. of CSE, LBRCE, Mylavaram

Abstract - Cloud computing has become one of the most significant field since it provides flexibility, reliability and scalability; thereby decreasing the operational and support cost. The services ranges from simple backup services to cloud storage infrastructures. In this context, the security of data across cloud is a major concern . In order to protect the data, it is stored in an encrypted format. However, this encrypted data introduces new challenges for cloud deduplication. The standard Attribute Based Encryption (ABE) system does not support secure deduplication, which is crucial for eliminating duplicate copies of identical data in order to save storage space and network bandwidth. In this paper, we present an attribute based storage system with secure deduplication in a hybrid cloud setting, where a private cloud is responsible for deduplication detection and a public cloud manages the storage. The main advantage in this system is to overcome the previous data deduplication systems as it can be used to confidentially share data by specifying access policies rather than sharing decryption keys.

Keywords - Cloud computing, Attribute Based Encryption, secure deduplication, hybrid cloud, private cloud, public cloud

I. INTRODUCTION

There is a huge increment in the measure of information created every day and in 2020 it is normal 44 zettabytes of information will be delivered. The capacity and the board of these expansive volumes of information is turning into the most testing activity today. By re-masterminding different assets over the web distributed computing offers another method for administration arrangement. Among the administrations gave distributed storage administration is the most imperative and well known one. Making the information the board versatile in distributed computing deduplication system has pulled in increasingly more consideration as of late. At the point when similar information is being re-appropriated to the distributed storage by various clients deduplication is best. Information deduplication is a one of the information pressure procedures. This procedure keeps just a single physical duplicate and kills numerous information duplicates with a similar substance and connections other repetitive information to that duplicate. Many distributed storage administrations utilize a deduplication strategy decreasing asset utilization in this manner sparing plate space and system data transfer capacity. Cloud clients transfer private and their own information to the server farm of the cloud specialist organization. It is judgmatic to accept that cloud specialist organizations can't be completely trusted by cloud clients as the clients delicate information is helpless to both inside and pariah assaults. Despite the fact that information deduplication guarantees parcel of advantages, security and protection turns into a difficult issue because of the fast advancement in information mining and investigation methods. So as a divine being practice the client need to encode the information to be put away on cloud so as to guarantee information security and client protection. Deduplication have demonstrated mind-boggling expense investment funds , i.e., it diminishes up to 68 percent in standard record frameworks and 90-95 percent of capacity needs if there should be an occurrence of reinforcement applications .

II. DATA DEDUPLICATION PROCESS

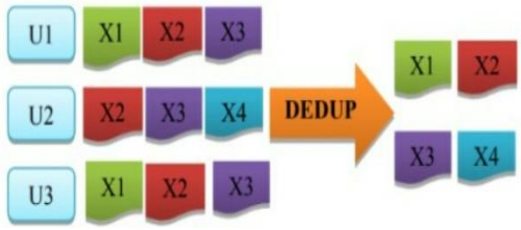

Figure 1. Information Deduplication Process

Information deduplication, likewise called as Intelligent Compression is the methods for lessening the measure of information that should be put away. The procedure of information deduplication works by wiping out the rehashed information and putting away just the primary novel case of any information. On the off chance that the client endeavours to store similar information again just a pointer is made to the initially put away information instead of putting away the repetitive information. For each document or lump (in square dimension) an interesting hash number is made utilizing the hash calculations, for example, MD5 or SHA1. The made hash number is contrasted and existing hash numbers in the file. On the off chance that it exists, at that point the information isn't put away else the new hash number and information is put away. At times the hash calculation may deliver a similar hash number for various pieces of information which is named as hash impact. Staying away from hash impact turns into a need to avoid information misfortune. Figure 1 clarifies the deduplication procedure including three clients. The clients transfer their records to the capacity server. The records X1, X2 and X3 are rehashed and henceforth deduplicated amid the procedure. Deduplication not just spares the extra room and system transfer speed yet in addition accelerates remote reinforcement and disaster recovery process.

III. CLASSIFICATION OF DATA DEDUPLICATION

Data deduplication process can be classified based on Data unit, location and Disk placement which are explained below.

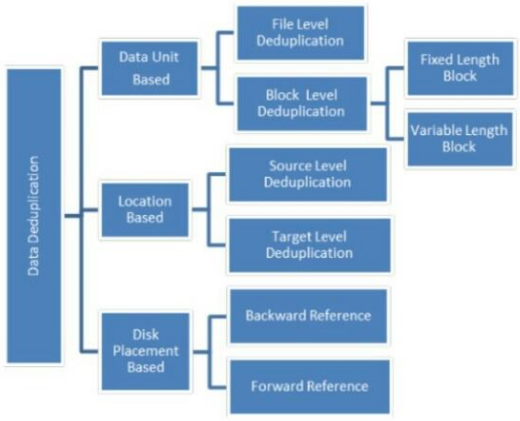

Figure2.Classification of data duplication

A. Information unit based deduplicaton -

Information deduplication can work at the document level or square dimension. In record level deduplication two documents are contrasted and their novel hash esteems. On the off chance that the qualities are same, at that point the documents are expected to have comparable substance and in this way just a single duplicate is spared and the pointers are made for different duplicates. In block level deduplication lumps are framed by part up the record substance. The lumps framed might be of fixed length or of variable length. As the name suggests fixed size piecing isolates the document into same measured lumps. It is quicker among other piecing calculations however it experiences "Limit Shift" issue when there is any change in the information. Variable length squares accomplish great information deduplication throughput.

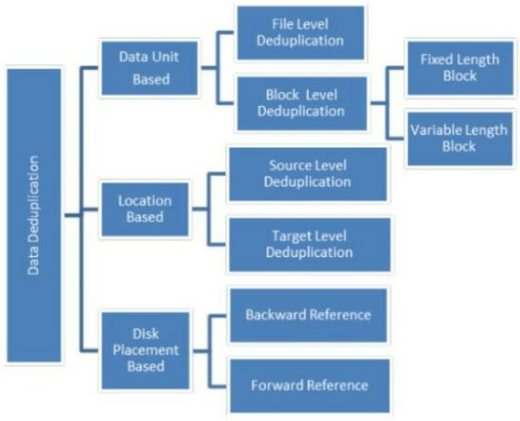

Figure 3. Grouping of Data Deduplication

B. Area based deduplication -

Area based deduplication is additionally isolated into two kinds source based and target based deduplication. Source based deduplication is performed at the customer side. Before the information is being transmitted to the capacity server the deduplication process is completed. This outcomes in sparing system transfer speed just as extra room. Target based deduplication is done at the server side and the customer is uninformed of the procedure. There is no overhead on the customer. Target based deduplication spares extra room however neglects to spare system transmission capacity.

C. Circle Placement based Deduplication -

Circle position deduplicationt depends on how the information is put away in circle. Two procedures are utilized in particular forward reference and in reverse reference. Forward reference keeps up new information lumps while making pointers to old information lumps. In Backward reference past information lumps are divided exceptionally.

IV. EXECUTION EVALUATOR OF A DATA DEDUPLICATION SYSTEM

The execution of the any deduplication framework is estimated by two imperative calculation. Dedupe proportion and Throughput.

1. Dedupe ratio=size of genuine information/size of information after deduplication.
2. Throughput= Megabytes of information deuplication/second.

V. TECHNIQUES USED IN DATA DEDUPLICATION

Coming up next are the safe natives utilized in deduplication.

A. Symmetric Encryption -

Symmetric Encryption uses a typical mystery key k for both encryption and decoding. Symmetric encryption can be characterized by three essential capacities.

- $KeyGen_{SE}$ $(1\lambda) \rightarrow k$ - is the key age calculation that produces k utilizing security parameter 1λ .
- Enc_{SE} $(k, M) \rightarrow C$ - is the symmetric encryption calculation that takes the mystery key k and message M as info and yields the ciphertext C.
- Dec_{SE} $(C, k) \rightarrow M$ - is the symmetric decoding calculation that takes the mystery key k and figure content C as info and yields the message M.

B. Focalized Encryption -

Focalized Encryption guarantees information mystery in deduplication. For each message M the client determines a focalized key and encodes the message with that focalized key. Also a tag is likewise determined for message M which is utilized to distinguish copies. In the event that two messages are same, at that point the labels are likewise the same. United encryption can be characterized by four essential capacities.

- $KeyGen_{CE}$ $(M) \rightarrow K$ - is the key age calculation that creates the key K and maps the message M to united key K .

- Enc_{CE} (K, M) → C - is the encryption calculation that takes the key K and message M as information and yields the ciphertext C.
- Dec_{CE} (C, K) → M - is the unscrambling calculation that takes the key K and figure content C as information and yields the message M.
- TagGen(M) → T(M) - is the tag creating calculation that maps the label T with message M.

Focalized encryption scrambles/unscrambles with a concurrent key that is acquired by processing the cryptographic hash estimation of the substance of the message. Indistinguishable information from diverse clients produce a similar figure content which makes deduplication doable alongside information privacy.

C. Evidence of Ownership -

The idea of verification of possession (Pow) allows the client to demonstrate the responsibility for duplicate M to the capacity supplier. Pow is actualized as an intuitive calculation by the client also, the capacity server. The capacity server determines φ(M) for the information duplicate M. The client sends φ' to the capacity server to demonstrate the proprietorship. On the off chance that φ'=φ(M) at that point the client is acknowledged as the information proprietor of the information duplicate M by the capacity server.

D. Recognizable proof Protocol -

The recognizable proof convention has two stages Proof and Verify. In the confirmation stage the client can demonstrate his character to the verifier by showing the conspicuous evidence. In the check stage the verifier checks the recognizable proof confirmation put together by the client furthermore, yields the acknowledge or reject message as indicated by the evidence submitted.

VI. DEDUPLICATION STORAGE SYSTEMS

A powerful deduplication framework is characterized by the help it gives as far as three associating contending objectives.

1. **Deduplication productivity**: This is the essential pressure objective which alludes how productively the framework recognizes the copy information units. Capacity cost is decreased by great deduplication productivity.
2. **Adaptability:** It is the capacity of the framework to help tremendous measure of crude stockpiling with stable execution. A decent versatility helps in lessening the generally speaking expense by diminishing the absolute number of hubs where every hub can deal with more information.
3. **Throughput:** Throughput alludes to the information exchange rate all through the framework. High throughput results in quick reinforcements. A deduplication framework shares information among documents naturally which is contradictory to a conventional reinforcement framework. So there emerges a requirement for solid reference the board which keeps track of portion use and guarantee back the liberated space. A couple deduplication stockpiling frameworks are examined underneath where each one is favored for various capacity purposes.

Venti is square dimension arrange capacity framework, expected for documented information. As the squares are tended to by the unique finger impression (a one of a kind hash created by an impact safe hash work) of their substance the alteration on a square is impossible without changing its location. This property actualizes writeonce arrangement which recognizes Venti from other capacity frameworks. In spite of the fact that Venti is viewed as the structure square of numerous capacity applications it can't effectively manage vast measure of information and experiences versatility.

HYDRAstor is a versatile, auxiliary stockpiling arrangement. The front end is the customary record interface where the back end is a framework of capacity hubs. The usage of variable-sized square, inline and hash-checked worldwide copy end on capacity hubs makes the framework exceptionally versatile. Extraordinary Binning is a versatile and parallel deduplication framework for piece based record reinforcement. It utilizes record likeness instead of area therefore expanding the throughput of the framework. The plan of comparable information documents into receptacles makes deduplication simpler by expelling copied lumps from each container. Extraordinary Binning is ground-breaking making information the board errands strong with low overhead.

MAD2 is an exact deduplication arrange reinforcement administration which takes a shot at both document level and square dimension. The procedures which help the framework in quickening the deduplication process are sorting out fingerprints into Hash Bucket Matrix, Blossom Filter Array to rapidly recognize approaching non copy object, double reserve and Load Balance method.

Duplicate Data Elimination (DDE) features are address-by block, just work as a foundation procedure, square dimension content hashing (160 piece SHA1), lethargic update and duplicate on write that ensures consistency among information and information hash. DDE can ceaselessly improve capacity proficiency as the information set develops.

VII. CONCLUSION

Information deduplication is a rising pattern and secure deduplication is a standout amongst the most imperative worries for clients. The paper centers around nuts and bolts of deduplication including the procedure of deduplication, order, techniques utilized in deduplication and couple of information deduplication frameworks. Different strategies accessible for secure information deduplication are too talked about alongside their favourable circumstances and disservices. In future secure information deduplication frameworks ought to be work with models giving elite proportion and throughput along with client protection.