International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

Deduplication on Encrypted Big Data in HDFS

Saif Ahmed Salim[1], Prof. Latika R. Desai [2]

[1]*Department of Computer Engineering, Dr. D.Y. Patil Institute of Technology, Pune University, Pune, India*
[2]*Department of Computer Engineering, Dr. D.Y. Patil Institute of Technology, Pune University, Pune, India*

Abstract*—Data de-duplication is single of essential data compression systems for rejecting duplicate replicas of repeating data, and has been generally used in cloud storage to decrease the total of storage space and save bandwidth. To make sure the privacy been proposed to ascent the information already outsourcing. To well confirm information security, this paper makes the primary endeavor to formally address the issue of approved information de-duplication. Not the same as usual de-duplication frameworks, the degree of difference assistances of clients are further considered in copy check other than the data itself. We additionally present a limited new de-duplication changes supportive approved copy check in a limit cloud design. Security study demonstrates that our system is protected in expressions of the definitions definite in the planned safety model. As a impervious of thought, we execute a model of our future authorized duplicate check system and conduct test bed experiment with our prototype. We display that our future authorized duplicate verify scheme incurs nominal above compared to normal processes.*

***Key words*—Access control, Big data, HDFS, data-deduplication.**

1.INTRODUCTION

Our aim is to minimize repetitive information and augment space funds. A strategy which has been generally embraced is cross-client deduplication. The fundamental idea behind deduplication is to store duplicate data (either records or pieces) just once. Appropriately, if a customer needs to exchange a record (piece) which is currently secured, the cloud provider will add the customer to the proprietor once-over of that report (square). Deduplication has demonstrated to accomplish high space and cost reserve funds and numerous Huge Information stockpiling suppliers are as of now receiving it. Deduplication can diminish capacity needs by up to 90-95% for reinforcement applications and up to 68% in standard document frameworks. Distributed computing gives apparently boundless "virtualized" assets to clients as administrations over the entire Web, while concealing stage and usage subtle elements. The present cloud advantage providers offer both exceedingly available limit and massively parallel figuring resources at reasonably low costs. As disseminated figuring gets the opportunity to be overwhelming, a growing measure of data is being secured in the cloud and conferred by customers to decided advantages, which describe the get to benefits of the set away data. One fundamental trial of appropriated stockpiling organizations is the organization of the consistently growing volume of data. To make data organization flexible in dispersed registering, de-duplication has been a remarkable technique and has pulled in more thought starting late. Data de-duplication is a particular data weight framework for wiping out duplicate copies of repeating data away. The system is used to upgrade stockpiling use and can similarly be associated with arranging data trades to lessen a number of bytes that must be sent. Instead of keeping various data copies with comparable substance, de-duplication discards dull data by keeping emerge physical copy and implying different overabundance data to that copy. De-duplication can happen at either the report level or the piece level. For record level de-duplication, it discards duplicate copies of the comparable archive. De-duplication can in like manner occur at the piece level, which takes out duplicate squares of data that occur in non-indistinct reports. Conveyed processing is a rising organization show that gives estimation and limit resources on the Web. One engaging convenience that circulated registering can offer is appropriated capacity. Individuals and endeavours are routinely required to remotely record their data to remain from any information mishap if there are any gear/programming frustrations or unexpected disasters. As opposed to purchasing the required stockpiling media to keep data fortifications, individuals and endeavours can essentially outsource their data support organizations to the cloud banquet providers, which give the principal stockpiling advantages for have the data fortifications. While disseminated capacity is engaging, how to give security confirmations to outsourced data transforms into a rising concern. One vital security test is to give the property of ensured cancelation, i.e., data records are forever blocked stores of deletion. Keeping data fortifications forever is undesirable, as fragile information may be revealed later on in perspective of data break or wrong organization of cloud managers. Along these lines, to avoid liabilities, attempts and government associations typically keep their fortifications for a predetermined number of years and request to eradicate (or squash) the fortifications a brief time frame later. For example, the US Congress is figuring the Web Information Maintenance establishment in moving toward ISPs to hold data for quite a while, while in the Joined Kingdom, associations are required to hold wages and pay records for quite a while.

1.1 Related Work

Cloud specialist co-ops offer profoundly accessible storage room and hugely parallel figuring assets at generally low expenses. The coming of Cloud Storage inspires ventures and associations to outsource information stockpiling to outsider cloud suppliers. An expanding measure of information is being put away in the cloud and shared by clients with indicated benefits, which characterize the get to privileges of the put away information. Gmail is a case of distributed storage which is utilized by the greater part of us consistently. One of the significant issues of distributed storage administrations is the administration of the perpetually expanding volume of information. To make information administration versatile in distributed computing, deduplication is a method and has pulled in more consideration as of late. Information deduplication is a specific information pressure system for wiping out copy duplicates of rehashed information away. Information deduplication is otherwise called single instancing or clever pressure system [1]. This system is utilized to enhance stockpiling use. Rather than keeping various information duplicates with a similar substance on the cloud, deduplication disposes of repetitive information by keeping just a single physical duplicate and alluding other access information to that duplicate copy. Deduplication can occur at either the record level or the square level [2]. For document-level deduplication, it takes out copy duplicates of a similar record. Microsoft's Single Instance Server (SIS) and EMC's Centera utilize a record level deduplication [3]. For piece level deduplication, it dispenses with copy squares of information that happen in no indistinguishable records. Dropbox distributed storage utilizes an extensive settled size (4MB) piece level deduplication [3]. Deduplication can happen at Inline, Post-prepare, Client-side, and Target-based [4]. In Inline deduplication, it happens before information put away on cloud i.e. it is performed at the season of putting away information on the capacity framework. It diminishes the plate space required in the framework [4]. In Post-prepare deduplication, it happens in the wake of putting away information on cloud i.e. it alludes to the kind of framework where programming forms, channels the excess information from an informational collection simply after it has as of now been exchanged to an information put away area. In Client-side deduplication, it happens at Owner/User side, in that copy information is first just recognized before it must be sent over the system. This will make trouble on the CPU yet in the meantime decreases the heap on the system. It is proposed to limit transmission capacity and space expected to transfer and store copied information. Kim et al. [3] given that many driving cloud-based capacity administrations including Dropbox, Wuala, Memopal, JustCloud, and Mozy utilize information deduplication methods at a source i.e. at a customer side to save network bandwidth from a user to cloud servers, which in turn increases the speed of data upload as well as storage space. In Target-based de-duplication, it occurs at storage service provider side. The Target deduplication 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 deduplication, the Target deduplication 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 room required [4].

Information deduplication brings a lot of advantages, security and protection concerns emerge as clients' touchy information are defenseless to both insider and untouchable assaults. Conventional encryption, while giving information classification, is incongruent with information de-duplication. Conventional encryption requires distinctive clients to scramble their information with their own keys by which indistinguishable information duplicates of various clients will prompt diverse figure writings, making de-duplication inconceivable [5]. The answer for adjusting privacy and effectiveness in deduplication was portrayed by M. Bellare et al [6] called united encryption. It has been proposed to uphold information classification while making deduplication. It scrambles/unscrambles an information duplicate with a joined key, which is inferred by registering the cryptographic hash estimation of the substance of the information duplicate itself [7]. To forestall unapproved, get to, a protected evidence of proprietorship convention [8] is additionally expected to give the verification that the client undoubtedly possesses a similar document when a copy is found. After the confirmation, ensuing clients with a similar record will be given a pointer from the server without expecting to transfer a similar document.

Nonetheless, past deduplication frameworks can't bolster differential approval copy check [8]. In an approved deduplication framework, every client is issued an arrangement of benefits amid framework instatement. Each record transferred to the cloud is additionally limited by an arrangement of benefits to determine which sort of clients is permitted to play out the copy check and get to the documents. Before presenting his copy check ask for some document, the client needs to take this record and his/her benefits as data sources. The client can locate a copy for this record if and just if there is a duplicate of this document and a coordinated benefit put away in cloud.

2. PROPOSED WORK AND METHODOLOGY

From the above literature survey we have concluded that an existing data de-duplication system, the private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges. Such architecture is practical and has attracted much attention from researchers. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud.

In the proposed research work to design and implement a system which will provide the parallel processing to detect the data de-duplication problem in big data environment. The system also provides benefit access control of data management and proxy revocation of system.

2.1 System Overview

Proposed scheme contain following main aspects

Encrypted Data Upload:

If data duplication check is negative, the data holder encrypts its data utilizing an arbitrarily culled symmetric key DEK in order to ascertain the security and privacy of data, and stores the encrypted data at database together with the token utilized for data duplication check. The data holder encrypts DEK with pkAP and passes the encrypted key to database.

Data owner:

First data owner can upload the text file from at the same system can take all files from data nodes and check the duplication with given file. If the VCS score is greater than threshold system can denote this file as duplicate. If the file is not duplicate, then job manager first check each server load and find the trustworthy of them base on CPU as well as memory load. Then the encrypted data can distribute into HDFS, and store the file tokens and other details into hash table.

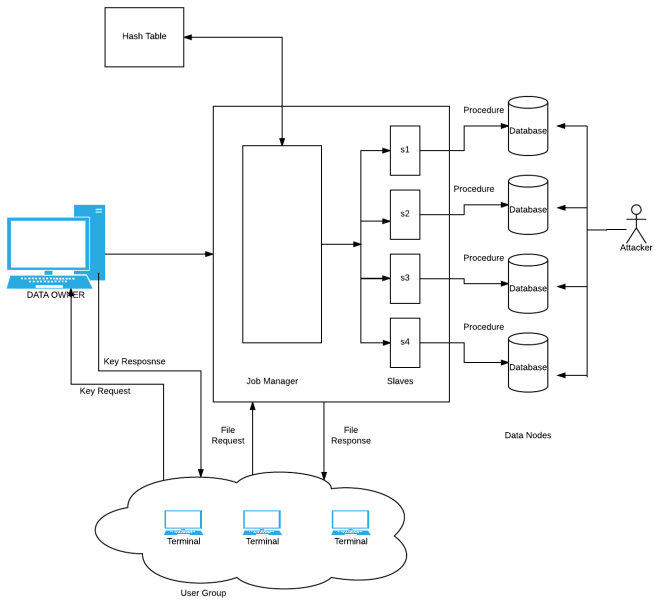

Figure-1: Proposed System architecture

User:

In case that an authentic data owner uploads the data later than the data holder, the job manager can manage to preserve the data encrypted by the authentic data owner at the HDFS. But at the same time manager can use the hash table efficient data retrieval, that can be reduce the time and cost also.

2.2 ALGORITHM

Document retrieval Algorithm

Input: Users query as Q , Network Connection N;

Output: result from relevancy calculation top k pages' base on Q.

Step 1: User provide the Q to system.

Step 2: if (N!=Null)

Process

Read each attribute A from ith Row in D

Res[i]=Calcsim(Q,A)

Else No connection

Step 3: For each (k to Res)

Step 4: Array list Objarray to bind Q to Res[i] or k

Step 5: Return to users Objarray

Step 6: Display Objarray

Weight Calculation Algorithm

Input: Query generated from user Q, each retrieved list L from webpage.

Output: Each list with weight.

Here system have to find similarity of two lists:

$\vec{a} = (a_1, a_2, a_3, \ldots)$ and $\vec{b} = (b_1, b_2, b_3, \ldots)$

where a_n and b_n are the components of the vector (features of the document, or values for each word of the comment) and the is the dimension of the vectors:

Step 1: Read each row R from Data List L

Step 2: for each (Column c from R)

Step 3: Apply formula (1) on c and Q

Step 4: Score=Calc(c,Q)

Step 5: calculate relevancy score for attribute list.

Step 6: assign each Row to current weight

Step 7: Categorize all instances

Step 8: end for end procedure

2.3 MATHEMETICAL MODEL

S= {s, e, F, X, Y,}

Where,

s = Start of the program.

1. Log in with webpage.

2. Input Query.

e = End of the program.

Retrieve the similar features.

F=Function Using Algorithm

1) Searching algorithm

2) Find similar clusters base on query approach

3) Select the results similar to query

4) Similarity Function (vector base cosine similarity) and return the result list

X = Input of the program. Input should be query.

Y = Output of the program.

First query submitted into server then server load datasets then divide into subspaces, again apply constraint propagation on subspaces then clustering after that server will do updated these layer repeatedly and final output will have generated into ensemble clusters.

X, Y U

Let U be the Set of System.

U= {Client, D, C, C1, N, E }

Where Client, F, S, T, M, D are the elements of the set.

Client= User, Server

D= Divide data into subspace

C= Apply constraint propagation on subspaces

C1= Clustering Solutions

N= update layers/nodes

E= Result clusters

Here system proposes to find similarity of two vectors: $\vec{a} = (a_1, a_2, a_3, \ldots)$ and $\vec{b} = (b_1, b_2, b_3, \ldots)$,

where a_n and b_n are the components of the vector (features of the document, or values for each word of the comment) and the is the dimension of the vectors:

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

Success Condition

if(Query != Null)

Failure Condition

if (Query==Null || Db==Null || Connection==Null)

3. EXPERIMENTAL RESULT

The final results of the designed system are given below. From those results we get the detailed information to Check de-duplication and upload the files, Fetching the Signs using Hashing Algorithm, checking for Duplication, file uploading, file downloading and attacker trying to attack(block) on data node. Detailed procedure of the proposed system is given. Based on this we confirm that securely authorized de-duplication is successfully achieved with hybrid cloud approach.

We also evaluated the computation costs of system for varying values of k, l and K. Throughout this sub-section, we fix m = 6 and n = 2000. However, we observed that the running time of grows almost linearly with n and m. The below tables 1 shows current system evaluation outcome.

Table-1: current system evaluation outcome

Approach	Data Records	Times in Seconds
Proposed	2000	35
	4000	68
	6000	102
	8000	132
	10000	171

For the results and comparative analysis, we compare the system with some existing approaches like cloud base de-duplication, KNN base duplication, the below graphs show the time required for retrieve the data with propose as well as existing.

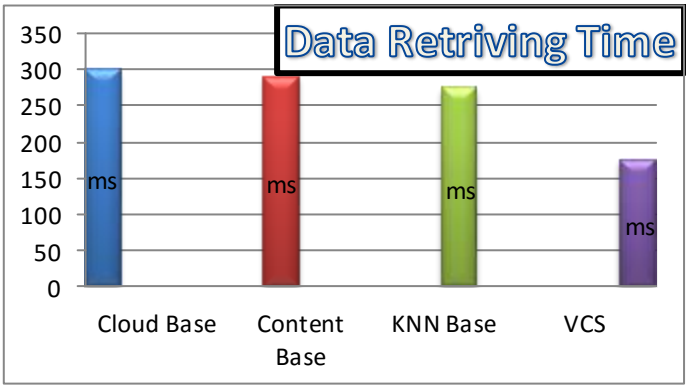

Figure-2: Proposed system comparison with others

After the complete implementation of system evaluate with different experiments. For the second experiment system focuses on time complexity of cryptography algorithm. The system takes use different time for data encryption as well as data decryption purpose. The below figure 3 shows the encryption and decryption time complexity.

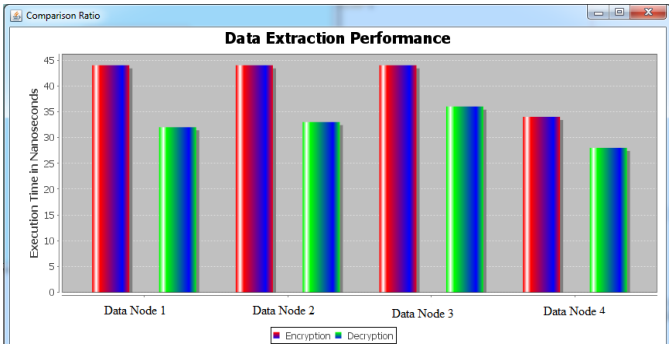

Figure-3: Data encryption and decryption performance with different approaches

4. CONCLUSIONS

Managing encrypted data with deduplication is consequential and consequential in practice for achieving a prosperous cloud storage accommodation, especially for astronomically immense data storage. In this paper, we proposed a practical scheme to manage the encrypted sizably voluminous data in cloud with deduplication predicated on ownership challenge and PRE. Our scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only sanctioned data holders can obtain the symmetric keys utilized for data decryption. Extensive performance analysis and test showed that our scheme is secure and efficient under the described security model and very opportune for sizably voluminous data deduplication. The results of our computer simulations further showed the practicability of our scheme.

Future work includes optimizing our design and implementation for practical deployment and studying verifiable computation to ascertain that SP deports as expected in deduplication management.

REFERENCES

[1] Deepak Mishra, Dr. Sanjeev Sharma, "Comprehensive study of data de-duplication", International Conference on Cloud, Big Data and Trust, Nov 2013.

[2] Gaurav Kakariya, Prof. Sonali Rangdale, "A Hybrid Cloud Approach for Secure Authorized Deduplication", International Journal of Computer Engineering and Applications, Volume VIII, Issue I, October 2014.

[3] Daehee Kim, Sejun Song, Baek-Young Choi, "SAFE: Structure-Aware File and Email Deduplication for Cloud-based Storage Systems".

[4] Pooja S Dodamani, Pradeep Nazareth, "A Survey on Hybrid Cloud with De-Duplication", International Journal of Innovative Research in Computer and Communication Engineering, December 2014.

[5] Boga Venkatesh, Anamika Sharma, Gaurav Desai, Dadaram Jadhav, "Secure Authorised Deduplication by Using Hybrid Cloud Approach", November 2014.

[6] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication", in Proc. IACR Cryptology ePrint Archive, 2012

[7] Jin. Li, Xiaofeng Chen, M. Li, J. Li, P. Lee, and W. Lou., "Secure Deduplication with Efficient and Reliable Convergent Key Management", In IEEE Transactions on Parallel and Distributed Systems,June- 2014.

[8] Jin. Li, Yan Kit Li, Xiaofeng, P. Lee, and W. Lou., "A Hybrid Cloud Approach for Secure Deduplication", In IEEE Transactions on Parallel and Distributed Systems, 2014.

[9] Jan Stanek, Alessandro Sorniotti, Elli Androulakiy, Lukas Kencl, "A Secure Data Deduplication Scheme for Cloud Storage".

[10] Jaehong Min, Daeyoung Yoon, and Youjip Won, "Efficient Deduplication Techniques for Modern Backup Operation", IEEE Transactions on Computers, Vol. 60, No. 6, June 2011

[11] Mihir Bellare,Sriram Keelveedhi and Thomas Ristenpart, "Message-Locked Encryption and Secure Deduplication", Proceedings of Eurocrypt, Vol. 6,March 2013.