Enhancing Data Security with De-duplication Using Key Exchange Protocol

[1]N. Pavan Venkata Ramana , [2]S. Vasundra,

[1]PG Scholar, [2]Professor, [1,2]JNTUACEA, Ananthapuramu, A.P. India.

[1]*pavanvenkat07@gmail.com*, [2]*vasundras.cse@jntua.ac.in*

Abstract: Data accumulated in the virtual servers is notably increasing with timely manner. Security of the stored statistics leads to the serious concern. As corporate generate more data in day to day, the same data may be stored into the servers in multiple times which may occupy large storage space. Data storing multiple times in servers may lead to enormous storage of data. The retrieval of data also gets difficult. This may even affect the workload and latency of the servers. With the security as the main concern, data generally uploaded in encryption form. The retrieval of data is also in an encrypted format until the key is produced to decrypt the data to plain text. As the statistics stored in the cloud is in an encrypted format, the files that are uploading to servers are not identified for whether the file consists of similar data or the one of kind information. Hence statistics de-duplication is vital for identifying the redundant records and storing the facts in the cloud. Thus Improved aware position similarity (IPAS) is employed to alleviate the data similarity in storing the data on the server. In IPAS the data blocks which are stored in encrypted format will diagnose for data similarity which enhances the storage space, saving power and number of computations in Central Processing Unit cycles.

Keywords — Deduplication, Position aware similarity, Key exchange protocols, Security, Block sizing, Sample blocks.

I. INTRODUCTION

Data security is a term which deals with protecting data. The data might have been stored in any of the ways in the physical or manual method. The data usually stored on hard drives, servers or cloud and storages attached to networks [1][2]. The secure data is achieved in mentioned ways like data encryption, layered or tired storage security architecture. Duplication is a process of storing a file or information more than one instances which leads to wastage of space in the storage devices. De-duplication is a compression technique which is specialized in eradicating the redundant or repeated data by examining the similar files which are already stored in the database. The circumstance of de-duplication is to avail capacity of storage [3] which is a good measure for the organizations.

Key exchange is a process in cryptology where cryptographic keys are exchanging among two parties, but no another party else have the copy of a key. In the servers with the excess magnification of data, the risk of data controlling leads expensive. According to computer science corporation (CSC) file [4], the general cloud workloads are elevated to boom at a compound annual boom charge (CAGR) of 24% from 2013. By 2018, sixty nine percent of the cloud workloads are foreseen to be non-public cloud, with a purpose to develop at CAGR of 21%. Data de-duplication has risen in its primitive form since 1970's. It is initiated with the idea of corporations that wanted to keep massive amount of touch statistics without using of big quantity of storage area [5]. Sometimes, to de-duplicate the data, it takes large time even months which occurred on hard copies. The difficulty came along when widespread of computer use in an office environment. With the wide use of internet and computers, the data also timely exploded. There, backup's includes tapes, magnetic discs and other alternative hardware was created and used to keep the information. With more explosions of internet and computers, these backups were filled. Here, companies moved to alternative storages or infrastructure like cloud storage which is a virtual environment.

Though the information is held within the cloud, by storing the information once more, ends up in multiple storages. Majority of the massive mass of data that organization generates in-house or its workspace is duplicate data. From the vista of storage, retrieving and so on will cost to the organization. To promote a significant set of distinctive records for streamlining their master knowledge management activites, companies need to diagnose the duplicate data across diverse sources and get rid of it through effective data deduplication process. Professionals worked for the algorithms which can bring down the redundant statistics in the storage. The multiple times of storing data not only affect Information Technology (IT) resources but also occupy expensive bandwidth [6].

The algorithms like message digest (MD5) and secure hash algorithm (SHA-1) provides an eccentric hash value to the data segments for deduplicating the data. Then, these

 DOI : 10.18231/2454-9150.2018.0261

eccentric hash values are corelated with other segment values. If same values found, the data is not stored again. Even though, data de-duplication has a bottleneck [7] in comparing the unique values because of massive generation of data volume uploaded into servers. Thus makes increasing inthe systems latency. Data similarity detection algorithms such as traits, shingle and simhash are basically used. But while performing the deduplication, these algorithms require lots of central processor units (CPU) cycles and memory space. Thus latency increases with the datasets uploaded.

II. LITERATURE SURVEY

In [8], Y. Deng et al. anticipated knowledge de-duplication had been wide used in knowledge backup system because of the considerably reduced needs of storage capability and network information measure. However, performance of data de-duplication step by step decreases with the expansion of de-duplicated data. This is because the unique values grow notably with the increase of backup, and more portions of unique values have to store on drives. Every time, lot of accesses arise to locate precise values and disturbs the activities of de-duplication. Rather to it, unique values which belong to existed file may be stored with drives discretely. This produce minute or tiny disc accesses and results load performance which is degraded.

Along with it, an eccentric values appears only once in the backup activity will result in a low cache rate due to lacking locality. Author J.Xie proposed to induce file similar activity to increase unique value fetching which increases cache rate. Later unique values are made sequential to stream data through backup to sustain locality. The final result is that novel improvement can degrade the number of unique values accessing the drives and alleviating the bottleneck of the disk in de-duplication.

In [9] Y. Hua planned for data similarity aware computation. The cloud is rising for ascendable and economical cloud services. To handle huge data and decrease data shifting, the activity of virtual infrastructure needs better management for the cache. The author proposed an efficient multi-layered cache technique called MERCURY. To capture similarities, he influenced a low complexity sensitive hashing.

With the drawbacks of inefficiency and homogeneous data placement, Author proposed a multi-core sensitive hashing for similarities among data

In [10] Y. Zhou et al. proposed that processing same files having same unique values is dealing with data management. Segmentation of files is an easy approach to identify similarities. For that technique like traditional approach is defined. But, a small modification in traditional approach cause ineffective operation and the whole operations in it may lead inaccurate performance. Containing the activities performed by traditional approach, also, more activity performing technique is raised namely positional awareness algorithm. So, Position awareness sampling scheme is defined which performs efficientactivities in duplicate data.

In [11] L. P. Cox et al. proposed online backup garage is companion clean possibility for absolutely everyone to shop digital information, documents, and exclusive transmission files. This makes the storage servers loaded similarly as additional disk garage is wanted to save plenty of an oversized quantity of identical records. This disadvantage changed into a triumph over with a mechanism known as information de-duplication. This approach is employed for disposing of reproduction facts and to lessen redundancy at the server node.

For the duration of this paper, author studied and updated paintings on de-duplication and proposed an answer that could be a Parallel design for inline records de-duplication that makes use of the relaxed hash set of rules 256 for interest information de-duplication task with the intention to beat the problems of it slow and to scale back hash collision. All through this layout write and delete operations square measure accomplished for potency and time analysis. This structure is useful for garage servers anyplace an great amount is maintained every day and computer code industries continuously look for new tendencies so they may hold their storage systems up to this point and unfastened for not pricey utilization of the server nodes.

III. IMPROVED POSITION-AWARE SIMILARITY ALGORITHM [IPAS]

The algorithms such as traits and shingle exist; even they need to split files into blocks by Rabin hash algorithm. But, this ignores the content of the file. Also, slight modification at the file head leads failure of duplicate file detection. Algorithms like traditional sampling and position similarity exist, but a small modification in the file may result in the failure detection. Thus improved enhanced position aware similarity algorithm is proposed which keeps the same steps of traditional and position similarities, but with the improved steps for similarity identification. The improved enhanced position aware similarity algorithm begins with calculating the total file size samples the data blocks from both file head and tail end. Later block sizing is performed with eigen values allocating to those segments. Finally, the similarity is identified by using the eigen factors and deletes the unnecessary chunks from the file, and the original file is stored securely.

Algorithm :

Step 1: Begin

Step 2: Calculate the file size

File size =(File size/S) // S→sampling position

 DOI : 10.18231/2454-9150.2018.0261

Step 3: Sampling data blocks from file head and tail end

For (i←1 to N/2) // 'i' is a variable, N→number of sampling blocks

Step 4: Block sizing

Step 5: Fingerprint allocation

Step 6: Similarity identification

Step 7: End

The framework of improved position similarity algorithm

An improved position similarity mainly consists of the following steps: Sampling data blocks, block sizing, fingerprint allocation, similarity identification.

- **Sampling data blocks:** A file is sliced into N number of portions for ease identification of the similar data with the previous file to be compared with.
- **Block sizing:** Each block is then sized with LENC to a hash function which is a fixed arbitrary size to a data.
- **Fingerprint allocation:** The N-fingerprint values are collected as a set of sigA(N; LENC). In the same way, the set sigB(N; LENC) of another file is collected for comparison.
- **Similarity identification:** If any similarities are found then the de-duplication is done, and the file is updated to an old file which does not repeat the redundancy.

The system model consists of three entities: Data owner, User, Cloud service provider.

a. Data Owner

The owner acts as an administrator. Owner registration is accepted by the "cloud service provider (csp)". Then the owner will upload files, and the owner will check for user requests, if any user sends a request to download the file, which was uploaded by owner, then the owner will generate an encryption key and send to CSP. This shows the Deduplication files and giving a response to those files.

b. User

For accessing the files, the userwants to register first. The user views the files in the cloud,then to download the file user need CSP and owner permission, so he will send a request to the owner, after getting a response from both owner and CSP.

c. Cloud Service Provider(CSP)

CSP can view the list of users and owners, view the uploaded files and verify them. After receiving the encryption key from the owner, CSP will send a re-encryption key to the user again to download the file.

The following screenshots shows the working of similarity detection:

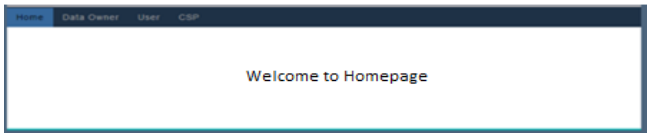

Figure 1: Home page

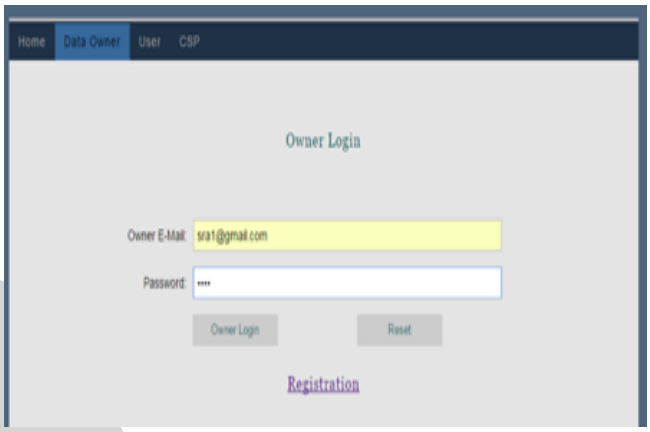

Figure 2: Owner login page

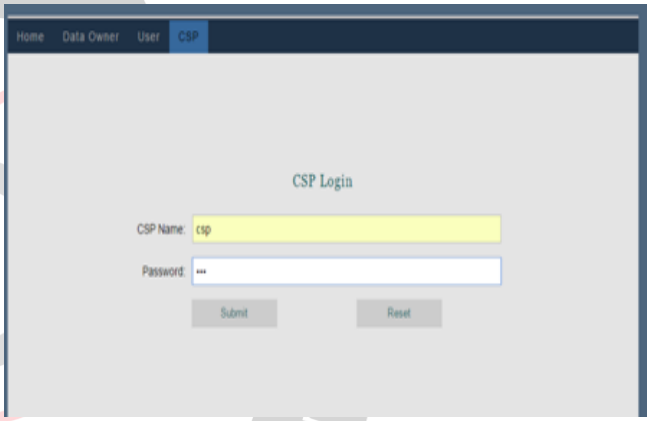

Figure 3: CSP login page

Figure 4: CSP process

Figure 5: Owner uploading the file

 DOI : 10.18231/2454-9150.2018.0261

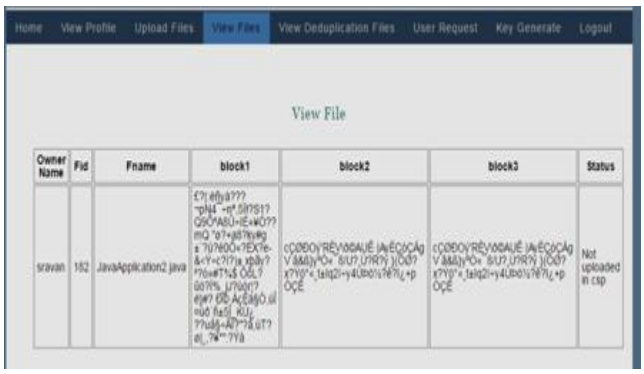

Figure 6: Owner viewing file in encrypted format

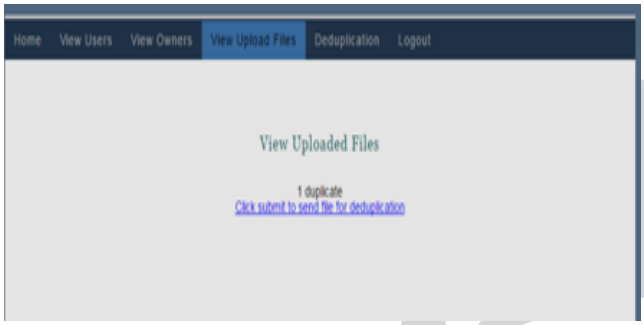

Figure 7: CSP verifying uploaded files

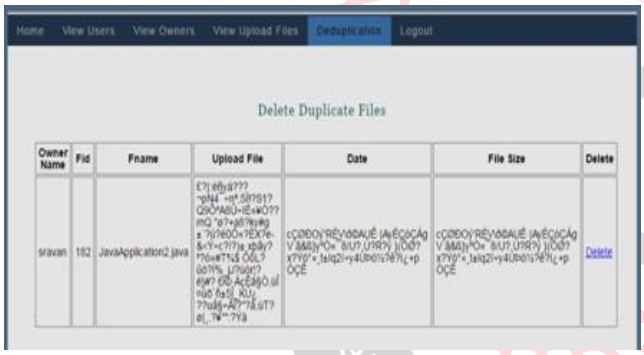

Figure 8: CSP deleting files with owner acceptance

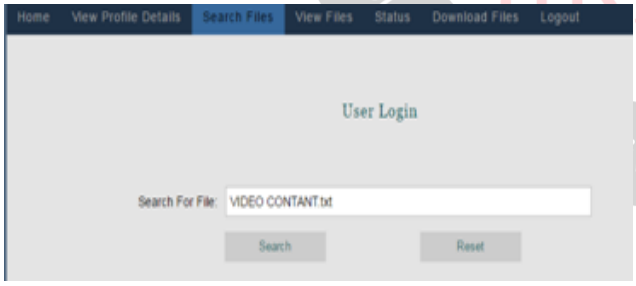

Figure 9: User login and search for file

Figure 10: User found file and sending request

Figure 11: File downloaded by user

IV. RESULT ANALYSIS

With the aid of an Improved Position Aware Similarity algorithm, the total size of the file that occupied in the servers before and after applying the proposed work is compared and shown below:

Table 1: Comparison of file size in existing and proposed systems

Size in Kb's	Existing system			Proposed system		
	File 1	File 2	Total	File 1	File 2	Total
Sample 1	25	35	60	25	35	35
Sample 2	250	350	600	250	350	350
Sample 3	1200	1450	2650	1200	1450	1450

In table 1, file1 has some information, and file 2 has the same information as in file 1 and extended information. In the existing and proposed system, the file 2 contains the same data that contains in file 1. But, the file 2 contains the updated data later which increased the size of the file 2. In the existing system, same files with different names are stored in the database which is said as multiple storages of data resulting in wastage of storage. The representation of multiple storage and resulting of it is shown in figure 12.

In the proposed system, same files with same data, but different names are identified, and duplicates are ignored with IPAS. In this system, the updated data is added to the existing statistics by deleting the duplicate file, and the entire file is stored in the server which results in avoiding the excess occupancy in storage devices.

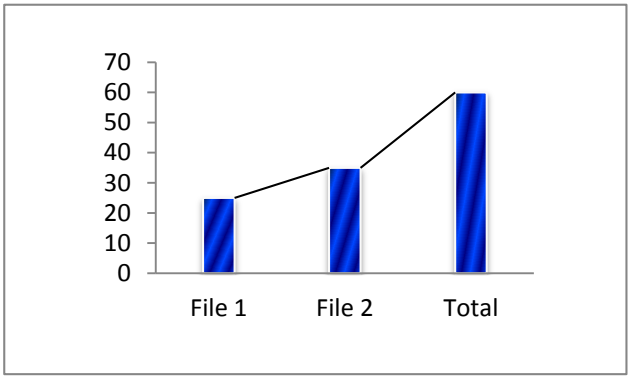

Figure 12: File size in Existing System

 DOI : 10.18231/2454-9150.2018.0261

In figure 12, file 1 and file2 which is of size 25 kb and 35 kb results in total occupancy of 60 kb.

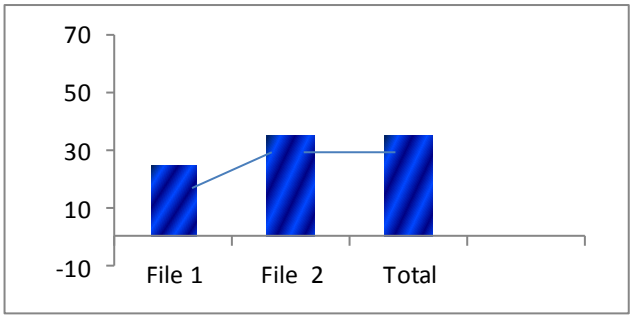

Figure 13: File size in Proposed System

In figure 13, with the IPAS, the redundancy of multiple storages of data in the servers is reduced, and the updated data of 10 kb is attached to the original file and then stored with the occupancy of 35 kb in the database which resulted in saving space in the storages. By these results, we can see that our technique performs better than the existing system.

V. CONCLUSION

In this paper, we present an Improved Position-Aware Sampling set of rules (IPAS) for the cloud surroundings. Comprehensive experiments are executed to choose top of the line parameters for IPAS. Corresponding evaluation and dialogue of the parameter selection are introduced in this paper. The assessment of precision and consider demonstrates that IPAS may be very powerful in detecting document similarity in assessment to Shingle, Simhash, Traits, and PAS. The experimental consequences additionally endorse that the time overhead, CPU and reminiscence occupations of IPAS are lots much less than that of these algorithms. Therefore, we trust that IPAS can be applied to the cloud surroundings to lessen the latency and attain both the efficiency and accuracy.

REFERENCES

[1] J. Gantz and D. Reinsel, "The digital universe decade-are you ready," IDC iView, 2010.

[2] Y. Deng, "What is the future of disk drives, death or rebirth?" ACM Computing Surveys (CSUR), vol. 43, no. 3, p. 23, 2011.

[3] H. Biggar, "Experiencing data de-duplication: Improving efficiency and reducing capacity requirements," The Enterprise StrategyGroup, 2007.

[4] Http://www.cse.com/digital_enterprise.org /wiki/csc

[5] M. Lillibridge K. Eshghi D. Bhagwat "Improving restore speed for backup systems that use inline chunk-based deduplication" Proc. 11th USENIX Conference on File and Storage Technologies pp. 183-197 2013.

[6] A. Muthitacharoen, B. Chen, and D. Mazieres, "A low-bandwidth network file system," in ACM SIGOPS Operating Systems Review, vol. 35, no. 5. ACM, 2001, pp. 174–187.

[7] B. Zhu, K. Li, and R. H. Patterson, "Avoiding the disk bottleneck in the data domain deduplication file system." in Fast, vol. 8, 2008, pp. 1–14.

[8] Y. Zhou, Y. Deng, and J. Xie, "Leverage similarity and locality to enhance fingerprint prefetching of data deduplication," in Proceedingsof The 20th IEEE International Conference on Parallel andDistributed Systems. Springer, 2014.

[9] Y. Hua, X. Liu, and D. Feng, "Data similarity-aware computation infrastructure for the cloud," IEEE Transactions on Computers, p. 1, 2013.

[10] Y. Zhou, Y. Deng, X. Chen, and J. Xie, "Identifying file similarity in large data sets by modulo file length," in Algorithms and Architectures for Parallel Processing. Springer, 2014, pp. 136–149

[11] L. P. Cox, C. D. Murray, and B. D. Noble, "Pastiche: Making backup cheap and easy," ACM SIGOPS Operating Systems Review, vol. 36, no. SI, pp. 285–298, 2002..

Acknowledgements

I thank Prof. S. Vasundra, Dr. Venkatesh and Asst. Prof Gn Vivekananda for the guidance given in processing this work.

Authors Biography

N. Pavan VenkataRamana received his B.E degree in Computer Science and Engineering from Dr.Sri shivakumara maha swamy college of engineering, Bengaluru, in 2016. Currently, he is pursuing his M.Tech in Software Engineering from JNTUA College of Engineering, Ananthapuramu, Andhra Pradesh, India. His areas of interests include Cloud Computing and Network Security.

Dr. S. Vasundra is working as Professor & Head of the Department in Computer Science and Engineering, JNTUA College of Engineering Anantapur, Ananthapuramu, Andhra Pradesh, India. She received her Ph.D. from JNTUA University in the year 2011. Her areas of interests include Mobile Ad hoc Networks, Computer Networks, Data Mining, Cloud Computing and Data Science. She is a professional Body Member of ISTE, IE, IEEE and CSI.

 DOI : 10.18231/2454-9150.2018.0261