

Emerging Data De-duplication Approach using Clonal Selection in Cloud Data Storage

P. Selvi,

Part-Time Research Scholar,
Department of Computer Science,
Karpagam Academy of Higher Education,
Coimbatore – 641021.

Dr. D. Shanmuga Priyaa,

Professor,
Department of CS, CA & IT,
Karpagam Academy of Higher Education,
Coimbatore – 641021.

Abstract

In current internet era Cloud computing plays a vital role in present information trending technology. The cloud service provider provides the storage to the Clients according to their requirements. In cloud data duplication is the main issue. Data duplication defined as storing same file more than one time in cloud, such that lot of storage is wasted. Due to the dynamic nature of the cloud storage, there is a statistically varying usage of cloud beyond regular use and some block of information chunks might be read in all times frequently but they will now not be used in but yet over length. Certain datasets can be typically retrieved or up to date thru numerous customers whilst, others may also want the unnecessary degree of redundancy for consistency requirement. For this reason, it is vital to aid this Clonal selection based De-duplication on hashed file values that has to be stored in cloud. This paper proposed a clonal selection based De-duplication system which does not store the duplicate data. In this method before uploading the file it is split into blocks and each block is hashed using SHA – 3 message digest algorithm to generate hash value for each block. After generating hash value the duplicate hash values is searched and eliminated using clonal selection algorithm, and the duplicate copies are referred using the corresponding blocks reference id. and along with that this work provided more security to the particular data by validating secret key and authentication details. The experimental result shows that the proposed method performs better De-duplication process in cloud storage than the other two techniques.

Keywords

Cloud storage, De-duplication, redundancy, Clonal selection, Hash value.

1. Introduction

The use of cloud for storing data by companies for backup and common people for sharing information among friends has increased drastically over the past few years. This has created a challenge to the cloud service providers to maintain all this massive data and to offer these services at lower price to the customers. In reality most of the data stored in the servers is often repeated. For example, a service may contain several instances of same data file, storing all these instances would require a large amount of storage space. This problem can be solved by using Data De-duplication technique. Data De-duplication stores solitary single distinctive illustration of the data type either on the tape or disk. In this approach redundant record or dataset is substituted using a pointer which holds the address of the unique record or data copy. This reduces the hardware used to store data and the bandwidth costs required for transmitting and receiving purposes. De-duplication methods on bit and block level cab capable to accomplish firmness ratios under various appropriate situations,

There are different types of De-duplication techniques like,

- 1) Location based De-duplication
- 2) Time based De-duplication
- 3) Chunk based De-duplication

In location based De-duplication there are two types one is client side or server side. In client side data is De-duplicated and send to the server side. In server side the data is send to the server first and then De-duplicated at the server side. In time based there are two inline and post fix. In chunk based the data is divided into the different chunks and data is De-duplicated.

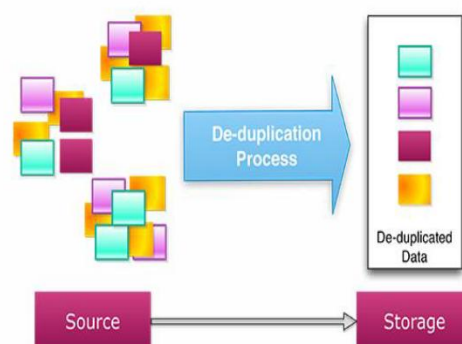


Figure 1. Data De-duplication Process in Cloud Storage

Data De-duplication is termed as a plan applicable to cloud storage providers in order to eradicate the duplicate data by maintaining a single unique copy of data which also consumes valid cloud storage space [fig 1]. Instead of maintaining multiple copies of same data the most useful and auspicious solutions which diminishes cost of storage and rise the accessibility of users by consuming network usage and minimizing the time of backup. Besides the advantage of using Data Deduplication it has the issues relevant to high cost for new security limitations and the related to challenges in privacy preserving. De-duplication [12, 14] shall be applied on the block or file level. In block level the de-deduplication process eliminates the presence of duplicate block and maintains only the single unique copies of the blocks. For the file de-duplication it eliminates the duplicate files from the storage and maintains the single copy of it.

2. Problem Definition

In general the users whom shoots queries doesn't have control on the databases to retrieve the results and there is no guarantee on the resultant output will not have duplicates. The problem of record de-duplication while retrieving the records through the queries passed by the user may also occur while using search engines which uses multiple of web databases. Here the focus relies on the same domain databases which mean the same type of records provided by the databases in response to the queries passed by the end users. For instances let consider the availability of p records which consists of group of attributes are also known as fields in a data source A. Another data source B consists of number of records. Here each of the q records can have duplicate of each of the p records in the A data source. The objective of this de-duplication is to identify the matching status of these p records with q records in data source B

**Table 1. Sample Dataset Showing Records from Cora Dataset
S. No, Author Id, Name, Title & Date**

S. No.	Author Id	Name	Title	Date
1	199	M.Ahlskog	J Appl.Phys	1994
2	101	InganasM.R.Anderson	J Appl.Phys	1994
3	101	Inganas and M.R	Journal of Applied	1994
4	199	Ahlskog	Journal of Applied	1994

Here the issue is determination of duplication among the records and display only the unique records. It is possible only when an appropriate matching algorithm is applied or

otherwise it may wrongly determine that all the records in table 1 are different and they are unique. The main motive of this work is to design and develop an approach which determines duplicate pair of records from the resultant query that denote the identical real world database using the help of unsupervised algorithms

In case of web database depending on the queries the relevant records are results are fetched and displayed and it is done only using the online queries and they are highly depend on the content of the query. While using conventional database the data which has to be retrieved are known in advance and based on that prior training on those database are applied and the system was well trained to identify the duplicates and maintain the unique copies of record. But such things are not encouraged during online query processing in the web database where multiple sources are involved to produce a single database like illusion and they don't have the pre-determined training facility on the database.

The major cause to design and develop a different solution for the online database lies on the disadvantages of developing coding manually or learning approaches in online. The foremost disadvantage is the offline approaches are highly relevant to the training dataset and which is not suited for online databases. Because to train the dataset all the possible combination of queries in order to classify the record as either duplicate or as a unique copy. The queries contain maximum possible coverage of queries to precisely determine the unique records by eliminating the duplicate records. It is done with the help of expertise in the concern domain. But in case of web based databases the records are collected from various databases dynamically so that it is not possible for applying such pre-determined dataset. The secondary disadvantage is that even though the dataset used for training represent the complete data but if it is applied for the partial dataset will never produce the optimal result in elimination of the duplicate records.

To overwhelm this issues an unsupervised algorithm for record de-duplication approach to be applied on a specific record matching problem using the hashing technique in order to determine the replicas between records in query results of various databases.

3. Related Work

Various similarity metrics was proposed to find the identical strings using matching algorithms [5]. Due to the typographical errors the mismatches are happen in database.

Most of the mismatches occur in database due to the typographical errors and variations of string. The persistent of inconsistency happened in typing can be determined using some specific similarity metrics approaches. These kinds of metrics contain similarity based on tokens, characters, phonetic and numeric similarity measures.

Faritha and chandrasekar [1] in their research work devised a modified bat algorithm based deduplication method for removing duplicate records. The main intention of this work is to produce a reliable and efficient approach which adopts data mining algorithms. With the help of evolutionary based computation approaches the similarities on data mining record duplication are greatly eliminated. Genetic algorithm this scheme which inhibits random solution and explodes the Optima of bat based recorded application by updating its finest characteristics. In [2] HMM model was proposed to find different fields called states and applying similarity Matrix among the parrot records which are called transition. In this approach the records in the database cleaned, standardized that is normalized through hidden Markov models. The performance of the proposed system two different data sets are used namely restaurant data set and core bibliographic data set. The implementation of hidden Markov model results in classification of records as duplicates and not duplicate. The simulation analysis shows the positive level of the proposed system than the existing approaches

In the work [3] the authors devised genetic approach based deduplication of records that integrates various pieces of evidences which was extracted from the content of data to determine ID duplicate and function using its fitness value that identifies whether the selected to pass of records in a database are duplicate or not. Whilst this genetic programming approach has the capability to automatically adapt the fitness function to find replica boundaries. In this work they used blood bank database as an illustration and using genetic approach for removing duplicates exist in the database.

Fellegi and Sunter [4] proposed a more elaborated statistical approach to deal with this problem. Their method depends on the definition of two boundary values that are used to classify a pair of records as being duplicates or not. It is implemented with Bayes's rule and Naive based classification.

The adaptive approach presented in [5] consists of using examples for training a learning algorithm to evaluate the similarity between two given names, i.e., strings representing identifiers. We use the term attribute to generally refer to table attributes, record fields, data items, etc. This approach is applied to both clustering and pair-wise matching. Koza [6] in his work proposed a genetic programming approach along with hash-based similarity i.e., with MD5 and SHA-1 algorithm.

Another prominent adaptive approach presented in [7] has a concept of genetic programming which has the capability to determine identical entries of records in a repository. It has the rule of evidence to find and fix boundaries of Replica. In [8] the authors did a study on deduplication by applying many existing methods and highlighting the merits and demerits in various cases. In paper [9] a novel strategy was proposed to determine similar set of data stored on a big data environment. Indus the data duplicates are removed from the media of storage and freeze up the storage space by increasing the system performance in terms of operational speed and reducing the time complexity with the help of deduplication process.

In the work [10] the authors proposed a Framework on record deduplication for the data stored in cloud. The application Storage for cloud data owners and providing high security for data is done by designing secure Cloud Computing Framework which comprises of 4 different segments namely intrusion detection, identity management system prevention, data deduplication and Secure cloud storage. Mandagere, et al., [11] proposed a deduplication approach which replicates the fold think of Deduplicated storage in terms of reconstruction of bandwidth overhead of metadata and resource usage usefulness.

The authors in [12], remember the outcomes of De-duplication at the reliability of the archival device. They proposed a method to improve reliability by using growing a method to weigh and degree the significance of each chew through inspecting the variety of information files that percentage the chew, and use this weight to pick out the level of redundancy required for the bite to guarantee QoS. Hence this proposed work overcomes the existing problem of data De-duplication in cloud storage by introducing the Clonal selection based searching method and hash the blocks using SHA -3 algorithm.

4. Proposed Methodology

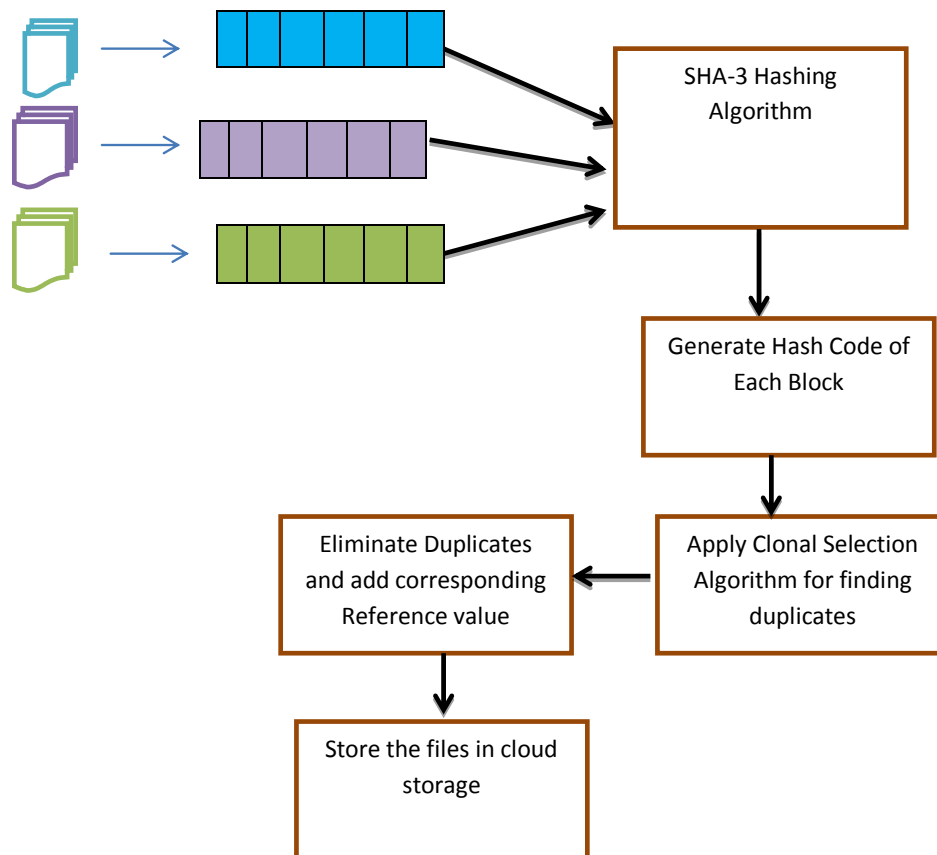


Figure 2. Overall Framework of Proposed Methodology

In this work the De-duplication is done for the files or blocks stored in the cloud storage are considered. As the storage cost is very high the presence of duplicate files or blocks should be identified and eliminated for overcoming the problem of memory consumption and cost. Initially the files are split into chunks known as blocks. Then each block is hashed using the SHA 3 algorithm. Once the whole blocks of all the files are hashed then the De-duplication process is started. The Clonal selection algorithm plays an important role in finding the duplicate hash values and maintains the first copy and the corresponding redundant occurrence of the blocks are replaced with the corresponding reference id of the matched blocks.

4.1 Splitting files into chunks

In this research work, the deduplication of data is implemented among the records of the files. It removes the redundant data copies using the clonal selection method. Before performing the searching process the file is fitted into chunks, which is known as blocks. After dividing the files into blocks each of them is hatched using the hashing function for generating unique identifier. Once unique ID is generated for whole blocks of the files then comparison on objects either files or blocks is done and the existing duplicate copies are removed from the data set. The data deduplication process eliminates the blocks of files which are not unique. The following four steps describes about the process of recording duplication

1. The given input data is divided into chunks or blocks
2. The selected block of data is hashed to provide hash value of the concern selected block or chunk
3. This hash values are used to find if there is any other block which contains same hash value of the given storage
4. Those identified duplicate data are replaced with their references in terms of pointers to the already existing copy of object in the database

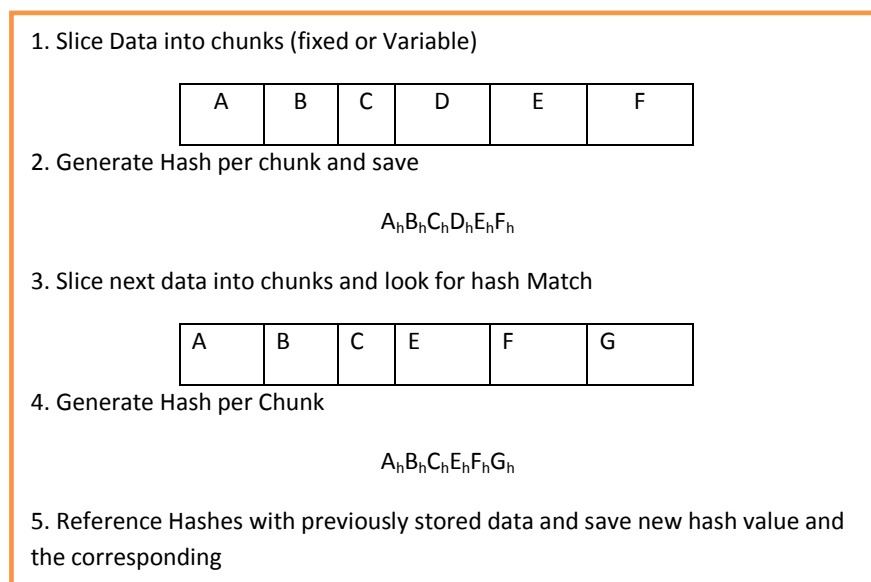


Figure 3. Process of Splitting Files into Chunks and Applying Hashing Value

4.2 SHA 3 Hashing Algorithm

The expansion of SHA 3 is secure hash algorithm 3 which is the latest version of secure hash algorithm family which is released by NIST [13, 14]. electing n bit string as input a function known as padding is applied then a permutation function f is computed on bits of blocks whose width is b , under the rate of R content length of output D which has the capacity $c = b-r$ and overall z value is calculated as

$$Z = \text{sponge}[f, \text{pad}, r]$$

- The selected N input string is padded using pad function of SHA 3 algorithm and resultant padded bit string contains p bits with the length which is divisible by r
- The split P bits into n consecutive r bit pieces $P_0, \dots, P_{n-1}$
- Assign the State S to a string of 0 b bits.
- Now grip the input string into the state :
 - For each block P_i ,
 - prolong P_i at the termination by a series of c 0 bits, resilient one of length b ,
 - XOR that with S and
 - put on the block permutation f to the consequence, acquiescent a new state S
- prepare Z to be the unfilled string
- while the length of Z is less than d :
 - attach the first r bits of S to Z
 - if Z is still less than d bits long, apply f to S , yielding a new state S .
- trim Z to d bit

4.3 Clonal Selection Algorithm

Through the study of immunity response in human body its inspiration derived an artificial immune algorithm. in brief the system builds antibodies of the immune system by adaptive learning of the features of the including antigens and Learns to act upon it. In this proposed work clonal algorithm which was originally device by Decastro and Van Zuben [15] starts designing algorithm by proposing a function called as $f(x)$ which has to be optimized. Few possible solutions of candidates are produced and in the purpose function these antibodies are used to compute their affinity and which will be used to identify the ones which will be cloned for the next step or iteration. In this the values of cloned antibodies IC change Unmute waited with some predefined ratio and again the affinity values or recomputed and saved.

After specific iteration of affinity, the affinity which holds the smallest value is considered to be closer to the solution of the specified problem.

4.3.1 Procedure for clonal selection algorithm is as follows:

Step 1: Producing a random number of set of antibodies which are the present candidates solution of a specified problem

Step 2: Compute the affinity values of every possible candidate solutions

Step 3: The obtained antibodies or started sorting from the ascending order of its value. The antibodies with lowest affinity is considered as a better match among antibody and antigen

Step 4: Clone the best matching antibodies with predetermined ratio

Step 5: With the predetermined ratio used mutate some of the antibodies. The ratio of mutation depends on how the clones are matching strongly or weakly. The weakly matching clones are mutated much more in order to reach the optimal solution then mutating the strong matching clones

Step 6: Compute every antibodies new affinity values

Step 7: Repeat Steps 3 through 6 while the lowest error criterion is not met.

4.3.2 Procedure for Record De-duplication using Clonal Selection:

1. Process of Record Antigen Selection:

From the given hash value of each records a single antigen is applied for the random record set where each antigen is associated with a single record.

2. Exposure of antigens to the system:

In this whole database is exposed to the selected *hash value of* record antigen. Affinity value which is a measure of similarity among the records is calculated using hamming distance measure for each record in the dataset with the selected antigen record.

3. Process of Selecting Similar records: A set of *n hash value* records are selected from the entire database that has the maximum affinity for the selected record antigen.

4. *Cloning*: The set of *hash value* selected records are then cloned in proportion to their degree of affinity.
5. *Affinity Maturation (mutation)*: The set of *duplicate hash value records* are then subjected to the process of mutation to find the same clone of records in remaining part of the database..
6. *Clone Exposure*: The clone records are then exposed to the record antigen, and again they measure the affinity of the record chosen.
7. *Candidature*: The records with highest affinity value are better finds the duplicate records and they are eliminated and the corresponding already existing copies reference id is added. The records which are unique to be consider as the candidate for further process of record De-duplicaiton. The duplicated records are removed from the database.
8. *Replacement*: Finally, the database contains only the unique records after eliminating the Clonal copies of the records.

5. Elimination of Duplicate Hashed blocks in Cloud Storage

For big data storage like cloud in order to improvise the performance of deduplication technique on such cloud storage data the encryption model greatly helps in this process for easy and precise result. Data deduplication technique is also termed as a specialized data compression method which eliminates duplicates of repeated data set a record in cloud storage space. Related and the relevant terms are intelligent compression or data compression of single instances data storage. This method is used to improvise the utilization of cloud storage very efficiently and it can be applied for transferring packets with the reduced number of bytes to consume the time complexity.

With respect to the data granularity the Strategies for deduplication can be classified into two main categories they are file level deduplication and block level deduplication these are nobody is becoming the most common Strategies for maintaining unique set of data in cloud storage. The science of the block can be either pixel or variable the block level deduplication data can be suitable for both type of approaches. The next categorization is depending upon the location in which such kind of data deduplication has to be performed whether the data has to be duplicated at the client side which is called as source based deduplication or else it is target based. The source based data deduplication first hashes each data segment which has to be uploaded and

since these has results to the storage provider to find whether such kind of data or already exist or not so that only the an duplicated are unique data segments will be actually uploaded to the cloud storage by the user. Adopting data deduplication on the client side will receive a band with conception but unfortunately it may vulnerable attack the system in which the attackers can immediately Discover whether a certain record or data has been stored or not. On the other side buy deduplication of data at the space of storage provider the system is greatly protected against the side channel attacks but such kind of solution increases the communication overhead.

In the process of deduplication unique chunks of data or dye patterns where identified and saved during an analysis phase. As the continuation of analysis other channels are also compared simultaneously to find the store copy and whenever a match occurs the relevant chunks replaced with a pointer which holds the address of already stored chunk position. This methodology even if same type of big patterns may occur hundred or even thousands of times the amount of data that has to be stored or transferred will be greatly reduced by referring the pointers to the unique data copy

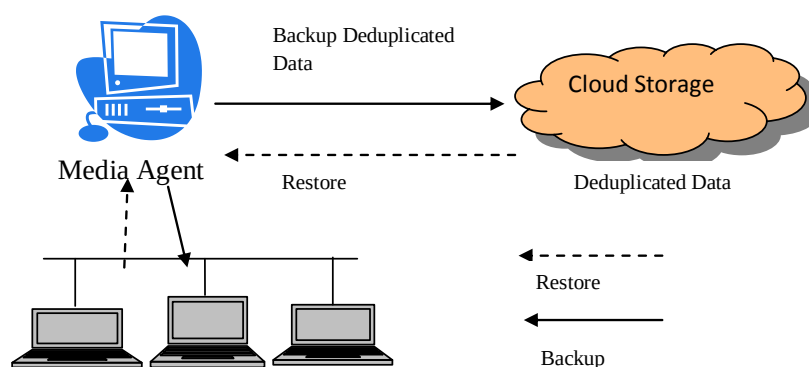


Figure 4. Direct De-duplication using Hash Based Clonal Selection Algorithm in Cloud Storage

6. Experimental Result

The proposed De-duplication technique using clonal selection was implemented using MATLAB Code. A set of sample files (with and without duplicate copies) are taken for De-duplication. If a new file comes, it is saved on the cloud storage. If a duplicated file comes, it is

not saved, only index file is updated, the storage space is saved. Testing was done for different combinations of files and results were tabulated

6.1 Performance Analysis

Table 2: TEST 1: Different Text Files with and Without De-duplication

S. No.	File Name	File Size (KB)	No. Of Users	Before De-duplication (KB)	After De-duplication (KB)
1	File1.txt	516	125	59920	516
2	File2.txt	347	45	27120	347
3	File3.txt	512	112	55400	512
4	File4.txt	340	50	27600	340
5	File5.txt	425	110	55600	425
6	File6.txt	475	70	27280	475

Total Size of files Before De-duplication = $59920 + 27120 + 55400 + 27600 + 55600 + 27280 = 252920\text{KB}$.

Total Size of files After De-duplication = $516 + 347 + 512 + 340 + 425 + 475 = 2615\text{KB}$.

Total Size of the files Reduced = $252920 - 2615 = 250305\text{ KB}$.

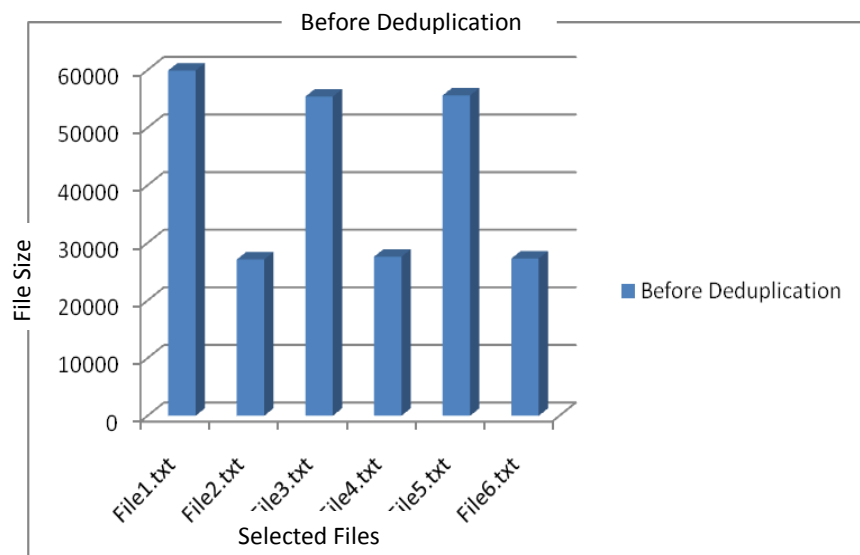


Figure 5. File Size Comparisons Before Deduplication

From the table 2 it shown that the comparison is done on the six different files with various sizes. Each file was accessed by many users and that also listed in it. While managing this file in cloud storage management without performing any De-duplication process then the storage size of these files will be really high. Whereas while applying the De-duplication process using Clonal selection algorithm it can be clearly understand from the result that the file size is not increased even how many users access the files. This is because the hashed value of each file blocks are searched for duplication by the Clonal selection algorithm if it those files are duplicates only there reference will be stored for the concern file otherwise if it is a new file then its original content will be stored in cloud.

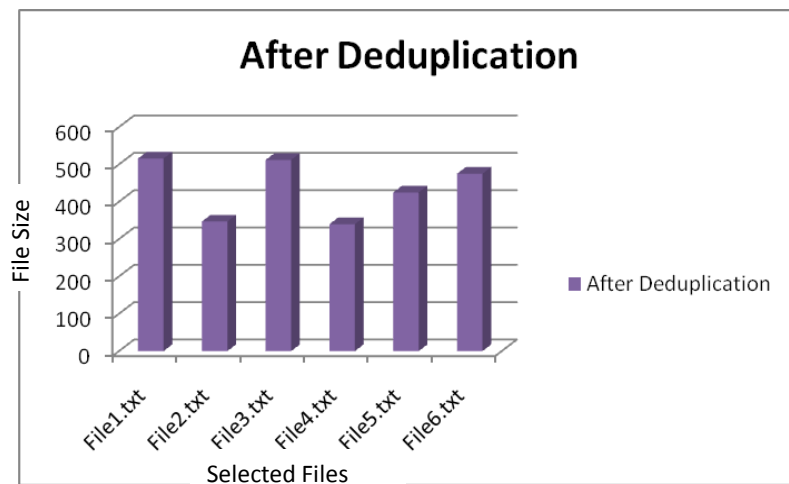


Figure 6. File Size Comparisons After Deduplication

The figure 6 shows the size of each file after performing the De-duplication and while the same files are accessed for manipulation by more number of users.

Table 3: TEST 2 Updating the Existing Text Files with and without De-Duplication

S. No.	File Name	Original File Size With De-duplication (KB)	Updated file Size (KB)	Updated file before De-duplication (KB)	Updated file after De-duplication (KB)
1	File1.txt	420	436	436	16

The table 3 shows the performance of the proposed De-duplication process while updating the existing file only the updating part alone is modified while keeping the remaining block of the file. Thus it considerably reduces the cloud storage space and the time complexity for searching the particular file or uploading them is also reduced in an effective manner.

Table 4. Performance Comparison of Clonal Selection Based De-duplication Method with Bat Algorithm and Genetic Algorithm

De-duplication Methods	Precision	Recall	Accuracy
Clonal Selection Algorithm	94.50%	92.67%	95.80%

Bat Algorithm	90.68%	89.65%	92.30%
Genetic Algorithm	86.50%	82.40%	88.50%

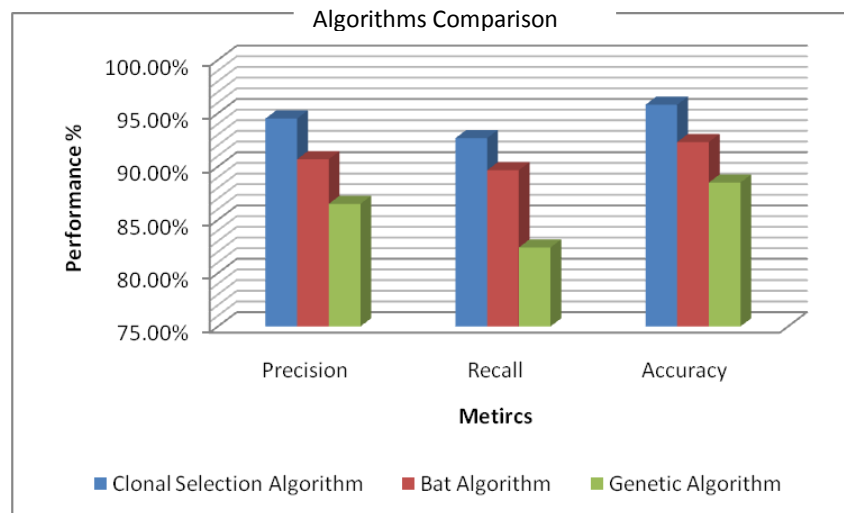


Figure 7. Accuracy of a De-duplication Technique is Obtained by Finding the Ratio of the Total Number of Correctly Retrieved Duplicate Records with the Total Number of Records.

$$\text{Accuracy} = \frac{\text{Total Number of Correctly Retrieved Duplicate Records}}{\text{Total Number of Records}}$$

Precision of a De-duplication technique is determined by finding the ratio of the number of correctly identified duplicate record pairs with number of identified duplicate record pairs.

$$\text{Precision} = \frac{\text{Number of Correctly Identified Duplicate Record Pairs}}{\text{Number of Identified Duplicate Record Pairs}}$$

Recall of a De-duplication technique is calculated using number of correctly identified duplicate record pairs with number of true duplicate records.

$$\text{Recall} = \frac{\text{Number of Correctly Identified Duplicate Record Pairs}}{\text{Number of True Duplicate Record Pairs}}$$

The table 4 and the figure 6 show the performance comparison of the proposed clonal selection based De-duplication in cloud storage by the precision, recall and f-measure. The

Result shows that the intelligent effect of antigens has greatly identified the duplicates copies more efficiently than the bat algorithm and the genetic algorithm by producing 94.5% of precision, 90.68% of recall and 95.8% of accuracy which produces optimal result than the other two existing methods.

7. Conclusion

The usage of cloud storage in cloud computing often leads to increasing in reputation of the files. It offers on demand for virtualized storage space and consumers mode emolument for the opening they openly spent. As the increasing need for and records keep inside the cloud, records De-duplication is more essential to optimally utilize the cloud storage. This paper utilizes two highly inspired techniques SHA-3 and Clonal Selection. The SHA-3 is used for producing the hash value of each blocks of the files and from the resultant each block Clonal selection maintains the single unique copy of the hash value of corresponding blocks and if it determines any duplicates they are not stored again instead the reference value of the original hashed block is referred thus it reduces the redundancy avails in the existing approaches. From the simulation result it is highly proved that the proposed method contributes more in process of De-duplication than the other two techniques namely genetic algorithm and bat algorithm.

References

1. FarithaBanu .A, Chandrasekar .C, An Optimized Approach of Modified BAT Algorithm to Record De-duplication, International Journal of Computer Applications (0975 – 8887) Volume 62– No.1, January 2013
2. A Novel Approach for Record De-duplication using Hidden Markov Model
1 R.Parimaladevi & 2 DR.V.Thigarasu R.Parimaladevi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 8070-8073
3. S Kavitha, K Arunadevi, S V Sivaranjani and V Karthika, Record De-duplication Using Genetic Programming Approach, Vol. 2, No. 2, May 2013 Int. J. Engg. Res. & Sci. & Tech.
4. I.P. Fellegi and A.B. Sunter, “A Theory for Record Linkage,” J. Am. Statistical Assoc., vol. 66, no. 1, pp. 1183-1210, 1969.

5. I. Bhattacharya and L. Getoor, "Iterative Record Linkage for Cleaning and Integration," Proc. Ninth ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery, pp. 11-18, 2004.
6. J.R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, 1992.
7. J. R. Waykole, S. M. Shinde, A Survey Paper on De-duplication by Using Genetic Algorithm Alongwith Hash-Based Algorithm, *Int. Journal of Engineering Research and Applications* www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 1(Version 1), January 2014, pp.343-346
8. Shrivashital Shankar Gujar, Avinash , A Survey of Record De-duplication, *International Journal of Engineering Science and Innovative Technology (IJESIT)*, Volume 3, Issue 2, March 2014
9. Shamsher Singh, Ravinder Singh, Next Level Approach of Data De-duplication in the Era of Big Data , *International Journal of Advanced Research in Computer Science*, Volume 8, No. 4, May 2017 (Special Issue)
10. R. Shobana, K. ShanthaShalini, S. Leelavathy V. Sridevidei, Duplication of Data in Cloud, *Int. J. Chem. Sci.*: 14(4), 2016.
11. N. Mandagere, P. Zhou, M. A. Smith, and S. Uttamchandani, "Demystifying data De-duplication," presented at the Proceedings of the ACM/IFIP/USENIX Middleware '08 Conference Companion, Leuven, Belgium, 2008.
12. D. Bhagwat, K. Pollack, D. D. E. Long, T. Schwarz, E. L. Miller, and J. F. Paris, "Providing High Reliability in a Minimum Redundancy Archival Storage System," in *Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, 2006. MASCOTS 2006. 14th IEEE International Symposium on, 2006, pp. 413-421.
13. https://www.nist.gov/itl/csd/201508_sha3.cfm
14. https://www.nist.gov/manuscript-publication-search.cfm?pub_id=919061
15. De Castro, L.N., Von Zuben, F.J.: The Clonal Selection Algorithm with Engineering Applications. In: *GECCO 2002 - Workshop Proceedings*, pp. 36–37 (2000)

