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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Hybrid Big Data Approach For Secure Authorized DE-duplication

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

B.E. Student, Dept. of Computer Engineering, Bharati Vidyapeeth's College of Engineering, Lavale, Pune, India

Assistant Professor, Dept. of Computer Engineering, Bharati Vidyapeeth's College of Engineering, Lavale, Pune, India

ABSTRACT: The use of cloud storage is increasing rapidly over a decade and hence maintaining efficiency of these cloud storage becomes extremely important issue. The main reason that hinders the cloud efficiency is the data redundancy. Hence eliminating data redundancy is very necessary. There are many methodologies which could tackle the issue but most effective technique is to deploy an effective de-duplication scheme over the cloud data which could eliminate duplication amongst the files stored on the cloud, hence only unique data can be stored on cloud thereby improving its space complexity. There are many methodologies and theories which highlight on use of De-duplication in cloud like- Data De-duplication over unencrypted data, Application aware data De-duplication schemes, etc. Most of these systems for De-duplication have some performance issues that can lead to lower accuracy of the technique. This paper proposes a novel De-duplication scheme over the data in which every unique file on the cloud will generate a unique hash key which will be maintained by the mechanism called bloom filter. The De-duplication will be done on basis of the hash key generated and later the data will be encrypted and stored on the cloud.

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

I. INTRODUCTION

The use of mobile computing devices like laptops, cell phones, tablets, etc. is rapidly increasing main issue with these devices is its limited storage capacity. To catalyse this issue use of the cloud storage becomes a necessity. To enhance the performance of cloud it becomes essential to eliminate data redundancy in the storage. For this purpose many models are in use. Most effective one is the data De-duplication of the files stored on the cloud [1].Hashing is the technique which is used to create the string of fixed length to represent set of many strings. Using hashing the set of strings can be identified or indexed with the fixed length hash key generated by the hashing algorithm. As a hashing technique this system uses MD5 hash key generation algorithm [2].The hash key generated by the algorithm is of 32bit value. The MD5 Hash key generation algorithm is performed into two steps known as Padding and Compression consisting three main operations viz. Bitwise Boolean Operation, Modular Addition, Cycle shift Operation. In Padding the set of strings is converted into the blocks of 512-bit (sixteen 32-bit words). Then the entire string is padded to convert the length of the string into multiple of 512. The algorithm then operates on the 32-bit state, divided into four 8-bit words. The processing of each block is done individually. MD5 hash key algorithm gives fixed size hash value as an output regarding the size of the input string.The bloom filter is the data structure introduced in 1970 by Burton Bloom. Initially it was used in data base applications, last few years they have been considered in many networking applications like overlay and peer to peer networking systems [3].The BF is an effective data structure which can be used for hash key management as its working is based on hash functions, thereby allowing false positive; the memory saving capability of BF is very impressive and thus it outweigh its minute drawbacks making its use very tempting. More accurately, a Bloom Filter represents a set X of t elements from a universe U. It uses an array of n bits, denoted by B[1],...,B[n], which are initially all set to 0. A number of r independent hash functions h1,...,h(r) are used, with log2(n) bits long output; the hash functions separately map each element of the universe to a random number which is uniformly distributed over the range. For each element s in X, the bits B [hi(s)] are set to 1, for $1 \leq i \leq r$ (a bit can be set to 1 many times). To answer a question of the form "Is y in X?" one checks whether all bits B[hi(y)] are set to 1. If not, y is not a member of X, while if all B [hi(y)] are set to 1, it is considered as y is in X. As mentioned above, there is a probability of false positives f that can be properly tuned by varying the values n and r. It is a known result [3] that the

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

optimal value of r is r =(n/a)log2; in this configuration ,all bits B[1],...,B[n] are set with probability p =1 /2 (thus, roughly, the same number of ones and zeros occur in the BF) and f is minimized.

The technique used for tagging the multiple users to the same file for de-duplication of the data is the Subset Vector Creation. This technique can be used for allowing the access of the same file to the users who have uploaded it[4]. In Subset Vector Creation each file is identified by its unique Hash Key and for that hash key the Subset Vector is created. The each entity in the Subset Vector contains two values, the first one is username (the user who uploaded the file) and the second one is filename (the name of the file he uploaded). The first entry in the Subset Vector is always the first user who uploads the file. Then the multiple users and their respective file name is added to the same set if the hash key of the files they are uploading is same based on the content of the file[5]. The example would be if the user 'ABC' uploads the file '123.txt' with the unique hash key 'KEY' then the subset vector for that unique hash key is created, then if user 'XYZ' uploads the file '987.txt' with the same hash key generated based on the content then the Subset Vector for that hash key would be as follows:

{'KEY'} = {(ABC, 123.txt), (XYZ, 987.txt)}

Where, KEY = unique Hash Key for that particular file

The advantage of this Subset Vector Creation technique is that if even multiple users upload the same file with different file name still those users can be tagged to the previously uploaded file because it creates the vector based on the hash key. This technique as a file tagging tool is proven effective due to the use of data structure called vector. Though the understanding of vectors is slightly complicated its use is application oriented in the proposed system.

II. RELATED WORK

Looking at system internal architecture of existing process de-duplication for cloud storage, Figure shows that how system flow works, server authenticate the client side services it could be done with user account verification process. After the verification and validation process user permission granted to access the services. Here we introduces bloom filter it should do very major task with the MD5 [7] algorithm, bloom filter don't able to store elements itself. If the elements are present then you don't use bloom filter for testing purpose [8] you can use it if undoubtedly not present.MD5 algorithm process the data contents and produces Hash Tag which is being stored in hash table, bloom filter check availability of the tag and return true or false [9] internal architecture shows that process. If the hash tag is available it can be notify to user and create pointer and link with that user file or folder. If it's not available then reverse circle cipher key encryption can done with the data contents [10].At the end of procedure data can be moved to storage area.

III. PROPOSED METHODOLOGY

Below Section Details complete Methodology used in Data de-duplication.

Phase 1:This Phase Data is Be uploaded by authenticated user.De-duplication System applies Reverse circle Chipper Encryption Algorithm and Data is been sent to Next Phase.

Phase 2: All Encrypted Content is been Hashed Using MD5 Algorithmic procedure. And File to Hash List is been maintained.

Phase 3: once the document hash is created it experienced through the procedure of De-duplication. In this procedure sprout channel accumulated all the hash estimation of past records. In this arrangement of past qualities, new hash esteem is analysed. When match is discovered, this hash esteem s bolstered to the similarity index relationship calculation for location of % connection. In the event that the esteem returned by similarity index is 1 then the record is copied and it won't spare to the cloud.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

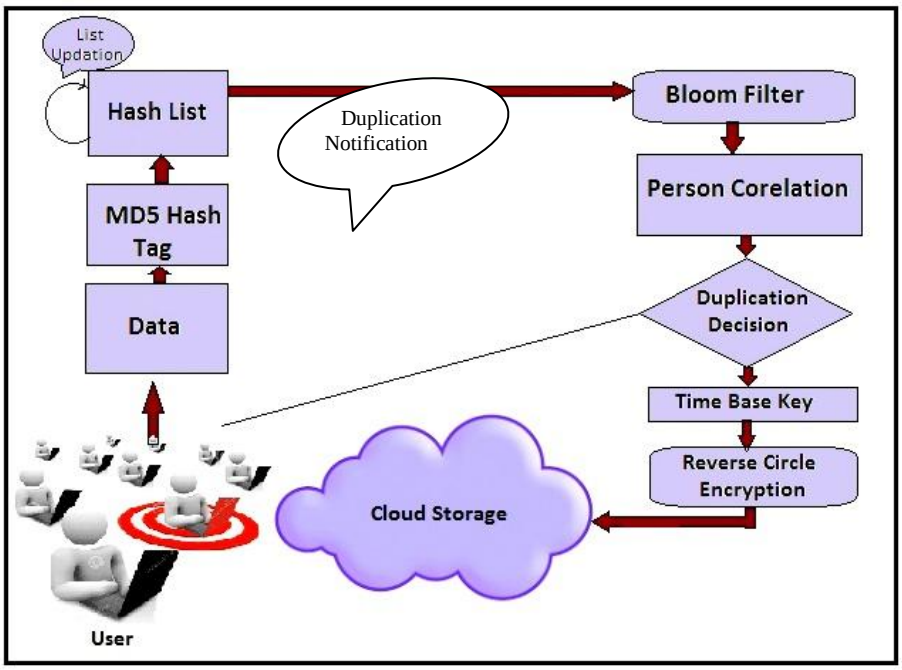

Fig.1.Internal System Architecture

Phase 4: when duplication is recognized, every one of the references of this record with past documents is kept up forthe future utilize.

Phase 5: In above case if File is not duplicate system flow comes to this step.

Phase 6: All Information of File is been saved using Inverted Index. Clouds plug-in have been used to deduplicate cloud data.

Proposed framework makes utilization of switch circle figure an encryption calculation for forcing the solid security approach. Turn around circle figure is secured contrasted with other in light of the fact that it makes utilization of private key for encryption reason. Once the info string is acquired it is partitioned into pieces of 10 characters. At that point these individual pieces are turned by their particular file and after that nourished to the encryption module. Encryption module acknowledges the pivoted string and in light of the ASCII estimation of each of the character encryption is performed. Detail usage technique for turnaround circle figure calculation is clarified in beneath calculation.

Phase7*:*De-duplication System provides feature to download file which was upload by him.

Phase8*:* This phase user is able to share his file with trusted users and one who are authorized.

IV. EXPERIMENTAL RESULT

Proposed system of De-duplicationis been deployed as a web application using Apache Tomcat and developed using J-creator IDE.Performance is evaluated based on the precision and recall parameters. So precision can be defined as the ratio of the number of relevant images and text files are identified as duplicates to the total number of irrelevant and relevant images and text files are identified as duplicates.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

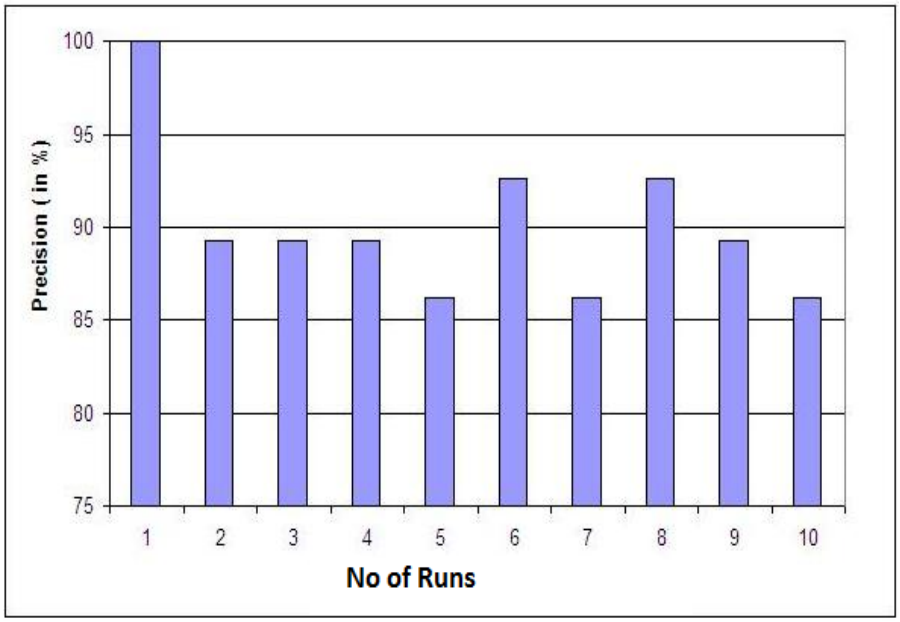

Fig 2: Average precision of the proposed approach

- P = The number of relevant images and text files are identified as duplicates
- R=The number of irrelevant images and text files are identified as duplicates
- So, Precision = (P/ (P+ R))*100

It is been observe that the tendency of average precision for the images and text files are identified as duplicates is more than the average of the other de-duplication techniques.

Recall is the ratio of the number of relevant images and text files is identified as duplicates to the total numbers of relevant images and text files are not identified as duplicates and it is usually expressed as a percentage.

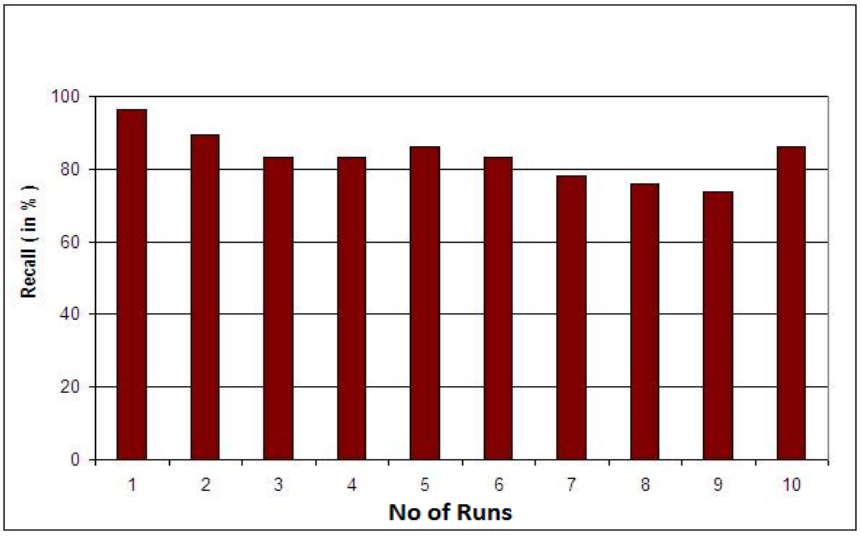

Fig 3: Average Recall of the proposed approach

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

It is been observe that the tendency of average Recall for the images and text files are identified as duplicates is more than the average of the other De-Duplication techniques. So this shows that our proposed system is achieving high accuracy than any other method.

V. CONCLUSION AND FUTURE WORK

The developing necessity for secure cloud storage services and better Encryption Decryption Lead to combine them, thus, defining an innovative solution to data Management and storage in cloud. Numerous De-duplication Schemes exists but fail to provide a complete reliable solution to De-duplication.Above proposed System is scalable and achieves better Scalability as compared to other De-duplication approaches.System can enhance to implement in Internet of things paradigm. Futures enhance to work in all formats of data. Proposed System is Prototype which could be enhancing to take any size input data.

REFERENCES

1. Jingwei Li, Jin Li, DongqingXie and Zhang Cai, "Secure Auditing and De-duplicating Data in Cloud", IEEE Transactions on Computers, 2015
2. PriyankaOra, Dr.P.R.Pal, "Data Security and Integrity in Cloud Computing Based On RSA Partial Homomorphism and MD5 Cryptography", IEEE International Conference on Computer, 2015
3. N. Bonelli, C. Callegari, S. Giordano, and G. Procissi, "A Bloom Filter Bank Based Hash Table for High Speed Packet Processing", IEEE International Conference on High Performance Computing and Communications, 2014
4. VenkateshwarKottapalli, Sunil Khatri, "A Practical Methodology to Validate the Statistical BehaviorofBloomFilters" IEEE, 2016
5. Ebenezer R.H.P. Isaac, Joseph H.R. Isaac and J. Visumathi, "Reverse Circle Cipher for Personal and Network Security", IEEE, 2013
6. Bo Mao, Hong Jiang, Suzhen Wu and Lei Tian, "Leveraging Data De-duplication to Improve the Performance of Primary Storage Systems in the Cloud", IEEE Transactions on Computers, 2015
7. Mike Dutch Data Management Forum Data De-duplication& Space Reduction SIG Co-Chair EMC Senior Technologist/2008 STORAGE NETWORKING INDUSTRY ASSOCIATION.
8. 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 secured authorized de-duplication.
9. A review of Comparative Study of MD5 and SHA Security Algorithm/International Journal of Computer Applications (0975 – 8887) Volume 104 – No.14, October 2014.
10. Bellare, M., Keelveedhi, S., Ristenpart, T.: Message-locked encryption and secure De-duplication. In: Advances in Cryptology{EUROCRYPT 2013. Springer 296{312
11. Harnik, D., Margalit, O., Naor, D., Sotnikov, D., Vernik, G.: Estimation of De-duplication ratios in large data sets. In: IEEE MSST '12. (april 2012) 1 {11}
12. 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
13. Bellare, M., Keelveedhi, S., Ristenpart, T.: Message-locked encryption and secure De-duplication. In: Advances in Cryptology {EUROCRYPT 2013. Springer 296{312

BIOGRAPHY

Gharge Nitin S, Bhor Ganesh G, Karajange Ganesh M,Jagdale Pravin R. Bharati Vidyapeeth's College of Engineering Lavale,Pune.Savitribai Phule Pune University,Studied in last year of Computer Engineering.

Prof.Kadam Yogesh V. IsaAssistant professor in the Department ofComputer Engineering, Bharati Vidyapeeth's College of Engineering Lavale, Pune, Savitribai Phule Pune University.His Research interests areBig Data and Hadoop.