A Secure AES Based Data De-Duplication technique using Inverted Index in Cloud Computing

Mr.Liladhar M. Kuwar[1], Prof.Kapil Vyas[2]
[1, 2] Department of Computer Science & Engineering
[1, 2] BM College Of Technology, Indore (M.P.)

Abstract- *Cloud computing comes throughout focus development of grid computing, virtualization as well as web technologies. Cloud Computing technology that interrelates between applicant and businesses to use web services without an installation. All the web services uses by business and applicant can access the information and files at any computer system having an internet connection. Cloud computing utilizes both central remote servers and internet to manage the data and applications with use of internet technology. With a lot of benefit of cloud such as scalability, accessibility, cost saving world user tend to shift their data to cloud storage.*

In this paper we are removing duplicate data to save storage space and increase storage speed of network. Hear we applied inverted index technique and tf-idf to identify duplicate data in cloud environment. Once de duplication is achieved system design for secure data transformation in network. Cryptography is common approach to protect the information in Cloud. Encryption algorithm plays a main role in information security system. Security is achieved throw encryption and decryption on data. In this paper we examine secure de duplication technique.. After removal of duplicate data markle hash tree applied.

Keywords- Cloud computing .Data De-duplication, Inverted – Index,tf-idf, markle hash fuction, AES Security Algorithm.

I. INTRODUCTION

Cloud computing comes throughout focus development of grid computing, virtualization as well as web technologies. Cloud computing is usually the world wide web based computing That presents infrastructure as service (IaaS), platform as service (PaaS), software as Service (SaaS). Throughout SaaS, software application form is usually created shown through the cloud provider. PaaS a good application development platform for the developer to Create a internet based application.[1] within IaaS computing . Infrastructure can be sent to be a help towards the requester. In your current application form associated with Virtual Machine (VM).These model usually are developed viewable from a good subscription basis utilizing cost Equally you-use model to be able to customers, regardless regarding their location. Cloud Computing still under inside their development stage and also has quite a few issue in addition to challenges out of a several questions in cloud scheduling plays very important role inside determining your current effective execution.

Digital application are growing fast and use of cloud in internet has increased rapidly. Cloud provide several benefits in term of cost and on demand services. Real time communication like computer adopted various computing ideas form cloud computing. Now day’s maximum amount of data is stored in cloud environment due to storage and networking environment. Data disk are unable to recognize duplicate data appear on disk. Duplicate data can affect storage space of disk. More of duplicate data affect the performance and uses of disk, space, speed and so more performance parameter. data de duplication technology overcome the problem of duplicate data in disk and increase the performance of computer. Duplicate data appear when a common technique is used to store and solve the data. Detection of duplication data is time consuming.

II. RELATED WORK

Now day data duplication is rapidly growing technique use in data backup storage without redundancy. It is very important in unique.

cloud computing data security is moreover a very approachable matter. This paper pays much awareness to the security issues of Cloud computing [2]. In this papers we help to sharing content in media using attribute. this content are secured by the security method .and the load balancing technique is used.[3] We design an interactive protocol and an extirpation based key derivation , which is combined with lazy revocation multi tree structure and symmetric encryption to form a privacy preserving efficient framework for cloud storage.[4] We analyzed the data to determine the relative efficacy of data de duplication, particularly considering whole-file versus block-level elimination of redundancy and also studied file fragmentation, finding that it is not prevalent, and updated prior file system metadata studies, finding that the distribution of file sizes continues to skew toward very large unstructured files [5] Security in data de-duplication can be provided with the use of convergent encryption technique which encrypts the data before uploading it to public system. The the limitations of convergent encryption drives researchers towards building more sophisticated data de-duplication techniques which can fulfil current organizational needs.[9] . As a proof of concept, the work implement a prototype of proposed authorized duplicate check scheme and conduct tested experiments using the prototype. The work shows that the proposed authorized duplicate check scheme incurs minimal overhead compared to normal operations.

Cloud computing is efficient and scalable but maintaining the stability of processing so many jobs in the cloud computing environment The cloud system faces the issues of replication and the data duplication according to scenarios. In this context need to solve the problem of both, to enhance the cloud performance in terms of storage overhead and availability required to manage the entire data in such manner by which the search ability, and the indexing of data can be achieved both. Therefore the following suggestions are made to enhance the existing cloud performance.

III. PROPOSED WORK

A. TF –IDF

It is popular and effective scheme in information retrieval . we apply tf-idf technique to make inverted index of term . tf is term frequency of any world that appear in document divided by the total number of term in the documents.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF is Inverse documents frequency calculate term important in documents. IDF(t) = log_e(Total number of documents / Number of documents with term t in it.

B. Inverted index

Inverted index scheme overcome the problem of searching duplicate data in data set. It is key data structure for modern information retrieval system. We store Doc ID that contains term. Inverted index work on the read documents create token streem and modified token and then converted in inverted index first it cut character sequence in the world token map text and streaming.

It is very important data management De-duplication technique is used for technique of keeping multiple data copies with the same content using tf-idef technique using Inverted index eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy .

C. Hash Function

Our proposed system work with Merkle Hash Tree is a well-studied authentication structure [7] when unique term identified and documents are arrange with the inverted index then hash code is generated for documents to identified duplicate data over storage . The hash code is generated by the client and is stored at both the client and the server side. The tree has four leaf nodes viz. m1, m2, m3 and m4. Initially, we apply hash on each documents to obtain h(m1), h(m2), h(m3) and h(m4). Then, h(m1) and h(m2) are hashed and combined together to get ha. Similar process happens with message m3 and m4 and here, we get hb. Here, h is a secure hash function. This can be expressed as ha = h(h(m1)|| h(m2)) and hb = h(h(m3)|| h(m4)) Further, ha and hb are combined and rehashed to obtain the root as hr. This can be expressed as hr= h(ha|| hb)It is used to efficiently prove that a set of elements are undamaged and unaltered. It helps greatly in reduction of server time [9]. computing cryptographic hash function over data and use of hash value to determine similar data. Identification of duplicate data then but pointer to file ownership is updated thus saving storage and bandwidth and new data is not uploaded. When it comes to client side de-duplication, hash values of data are computed at client and send for duplicate check. An attacker, who gains access to hash value of a data which not authorized to him/her, may claim de-duplication of file and thereby gaining access to the file It is used by cryptographic methods to authenticate the file blocks.

D. AES

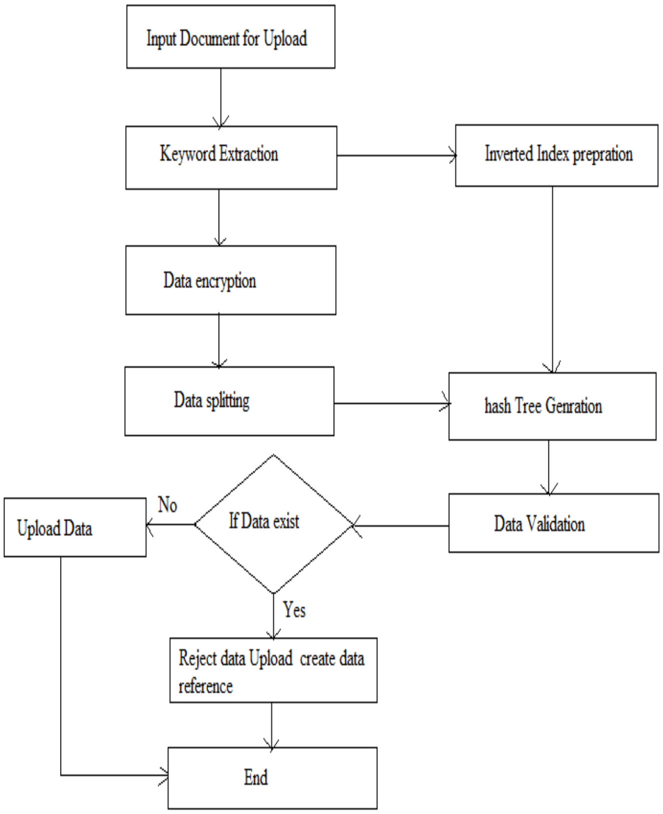
Once the duplication detected by hash function over data and use of hash value to determine similar data. Identification of duplicate data then but pointer to file ownership is updated thus saving storage and bandwidth and new data is not uploaded. An attacker, who gains access to hash value of a data which not authorized to him/her, may claim de-duplication of file and thereby gaining access to the file. To defend such an attack, method. AES works as an interactive algorithm between two parties . AES is a block cipher. The algorithm supports a variety of key sizes as 128,192 or 256. The default size is 256 bits. The encryption of data blocks is done in 10, 12 and 14 rounds depending on the size of the key used. It provides fast and flexible encryption and can be easily implemented on various platforms. In this paper, AES-128 is used and so encryption is done in 10 rounds. This algorithm is used for both encryption and decryption. For encryption, it takes data blocks and the secret key as the input and outputs the encrypted data blocks. For decryption, encrypted data blocks and key are given as inputs and original file blocks are the output.

Data De-duplication gives benefits but security and data confidentiality is still major issues. So, usual way to provide security is encryption. But there is confliction between data de duplication and encryption. Because it is possible that same plaintexts may lead to different cipher texts.

Enhance data de-duplication process and security. In order to protect the user's information from reveal, Siani Pearson [10]. put forward design principles in design process of cloud computing services to ensure that user's message and business information would not leaked out.

Calculations: Setup, Key Generation, encryption and decoding.

Setup: The setup algorithmic project takes no info barring the understood security parameter.



Input: Web page documents is provided as a input to read Document is denotes as term D. data is store in data center and all the web pages is available in cloud storage so for all documents we required Cloud storage S .and Inverted map IM as a input used to calculate inverted index which is designed to allow very fast full-text searches. An inverted index consists of a list of all the unique words that appear in any document, and for each word, a list of the documents in which it appear

Read R= read Document(D) from the documents which is available in cloud storage .

Extract Text Features(R)= E

Tf-Idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

Encrypt data (R)= E AES encryption is performed over data.

Generate has function and check the value of has function to performed this function till end loop,in these step massage is encrypted and secure for communication .

```
1  $Sp[] = E.splitFile(E)$
2  $for(i = 0; i \leq Sp.length; i++)$
      a. $H = genrateHash(Sp[i])$
      b. $if(H != HashTree.node)$
            i. $HashTree.createNode(H)$
      c. else
            i. Remove H
      d. End if
3  End for
```

Step :

Input: document D, cloud storage S, Inverted Map IM
Output: de-duplicated storage D_S

Process:

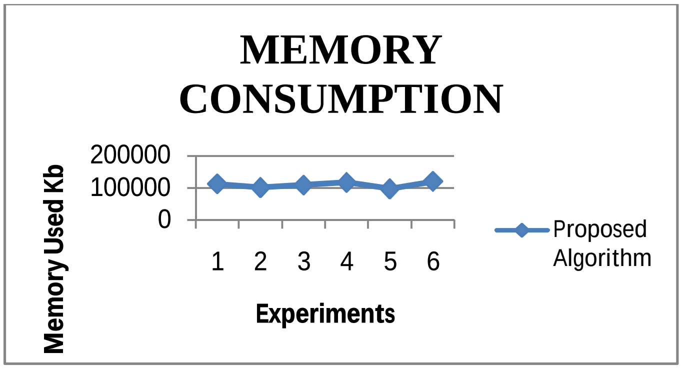
```
1. $R = readDocument(D)$
2. $E = extractTextFeaters(R)$
3. $IM.createEntry(E)$
4. $E = EncryptData(R)$
5. $Sp[\,] = E.splitFile(E)$
6. $for(i = 0; i \leq Sp.length; i++)$
      a. $H = genrateHash(Sp[i])$
      b. $if(H != HashTree.node)$
            i. $HashTree.createNode(H)$
      c. Else
            i. Remove H
      d. End if
End for
```

IV. RESULT

The amount of main memory required to execute the algorithm with the input amount of data is known as the memory consumption or space complexity. The total memory consumption of the algorithm is computed using the following formula.

Consumed Memory = Total Memory – Free Memory

Table 5.1: Uploading Memory Consumption



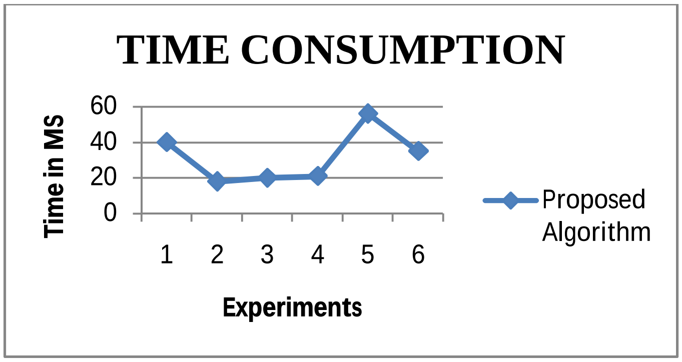
The table 5.1 show the memory or space complexity of proposed cryptographic approach. In this diagram the amount of main memory consumed in terms of kilobytes (KB) is given in Y axis and the number of experiments are reported at X axis. According to the obtained results the proposed algorithm consumes lesser resources and gives better performance of the encrypted and decrypted file.

Uploading Time Consumption

The amount of time required to develop the upload a data file on the server for cryptographic model is termed as the time complexity of the algorithm or time consumption of system. The time consumption is given using following formula:

Time consumption = End time of file put on Server – Start time of File put on data model

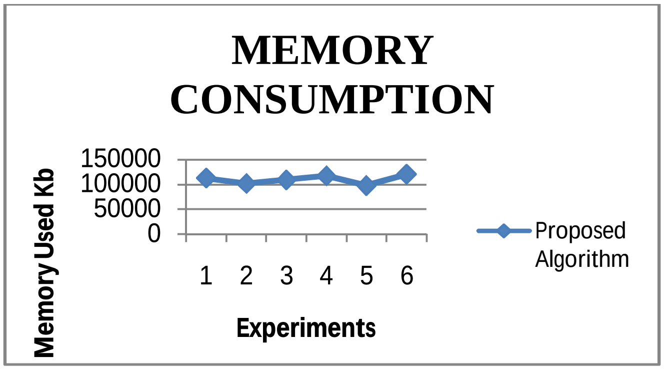
Table 5.2: uploading Time Consumption



Downloading Memory Consumption

The algorithms need a significant amount of main memory to store the data for processing. This storage requirement is termed as the memory consumption or the space complexity of the system. Here the downloading based memory consumption is computed.

Table 5.3: Downloading Memory Consumption

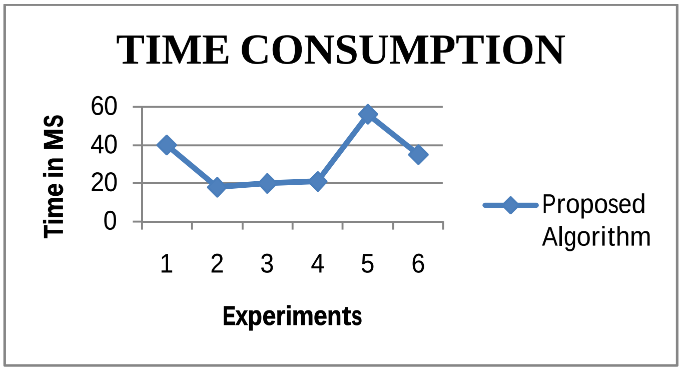


Downloading Time Consumption

The computational algorithms need an amount of time for producing the outcomes. Here downloading time is the time required by the server to do download the data file on the user system. That is computed using the following formula.

Time Consumption = End time of File put on system – Start Time to download file by Server

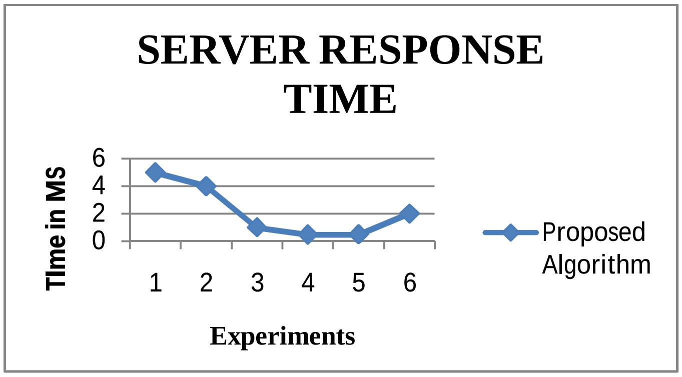
Table 5.4: Downloading Time Consumption



Server Response

The amount of time required to produce the outcome after making the request from the server is termed as the server response time. The response time not included the encryption or decryption activity during these measurements.

Table 5.5 Server Response Time



The computed response time of the proposed technique for cloud based secure communication is demonstrated using the table 5.4. The X - axis of this diagram contains the amount of experiments performed using the system and the Y axis shows the amount of time required for generating the response through the server for traverse the hash tree. That can also term as the communication overhead for the system. According to the computed results the response time is not depends on the amount of file size or other parameters. That is directly depends on the amount of work load on the target server where the data is stored or the application is hosted.

IV. CONCLUSION

Cloud Computing is interest of topic for research. De-duplication is a method available in cloud storage for saving bandwidth and storage capacity. As in the starting of paper it is very clear that our proposed mechanism is able to identify duplicate data. This result detection of replica and duplicate data removal by Appling hash function over it .therefore it is a efficient de-duplication method to remove duplicate data to save storage space and increases performance of network , de-duplication is less feasible with encrypted data. Data security issue is known to better secure data. so AES algorithm is applied in this system for secure transformation of data file or massage over network for secure transition of information in cloud computing.

This paper would be helpful to new researcher who wants to research on secure data de-duplication Security methods studied here in future we work to improve performance of our proposed work in security prospect. simple approach that makes de-duplication compatible with encrypted data. A strategy needs to study for data duplication and secure transmition over cloud computing envoirment. we work for a new security approach for secure data transmition and de duplication mechanism using one of the security algorithm .

REFERENCES

[1] Rahul Bhoyar Prof. Nitin Chopde M.E (Scholar) M.E (Computer Engineering) ,Cloud Computing:Service models,Types,Database and ssues , IJACCSEE Volume 3, Issue 3, March 2013.

[2] Neeraj Shrivastava and Rahul Yadav IES, IPS, Academy Indore, MP, INDIA. A Review of Cloud Computing Security Issues, International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 1, July 2013

[3] Tejashri Khandve, Megha Talekar , SheetalDhiwar --Security and Load Balancing In Cloud Computing .International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 10, October 2015

[4] RuWei Huang1,2 Si Yu1 --Design of Privacy-Preserving Cloud Storage Framework. (2010 Ninth International Conference on Grid and Cloud Computing IEEE

[5] M.Thamizhselvan R.Raghuraman, - A novel security model for cloud using trusted third party encryption IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS’15

[6] DUTCH T. MEYER, The University of British Columbia, Microsoft Research WILLIAM J. BOLOSKY, - A Study of Practical Deduplication ACM Transactions on Storage, Vol. 7, No. 4, Article 14, Publication date: January 2012..

[7] Mr. Avinash R. Dhok Ms. Ashwini P. Kolhe, A Survey on Scalable Data Security and Load Balancing in Multi Cloud Environment IJIRST –International Journal for Innovative Research in Science & Technology| Volume 1 | Issue 8 | January 2015 ISSN (online): 2349-6010

[8] Riddhi Movaliya Department of Computer Engineering PIET, Harshal Shah —A Survey of Secure Data Deduplication. International Journal of Computer Applications (0975 – 8887) Volume 138 – No.11, March 2016.

[9] Mr. Yendhe A.1, Ms. Dumbre T.2, Ms. Mahadik S.3, Ms. Gholap A.4, Prof. Gunjal A.5 —survey on secure privileged based data deduplication in cloud using twin cloud . Vol-1 Issue-4 2015 IJARIIE-ISSN(O)-2395-4396

[10] M. Karthigha1* and S. Krishna Anand 2— A Survey on Removal of Duplicate Records in Database. Indian Journal of Science and Technology | Print ISSN: 0974-6846 | Online ISSN: 0974-564 Aprial 2013 IJST.

[11] Akhila Ka*,Amal Ganesha,Sunitha Ca —A Study on Deduplication Techniques over Encrypted Data. Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016 doi: 10.1016/j.procs.2016.05.123.

[12] Zheng Yan Mingjun Wang Athanasios Vasilakos —Encrypted Data Management with Deduplication in Cloud Computing. IEEE Cloud Computing • March 2016 DOI: 10.1109/MCC.2016.29

[13] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee, and Wenjing Lou A— Hybrid Cloud Approach for Secure Authorized Deduplication IEEE transactions on parallel and distributed systems, vol. 26, no. 5, may 2015