Removing Duplicate Data in Cloud Environment using Secure Inverted Index Method

Ajahar Ismailkha Pathan[1], Liladhar M. Kuwar [2], Rijavan A. Shaikh [3], Dheeraj Basant Shukla[4]

[1,2,3,4]*Department of Computer Engineering, PSGVPM's D. N. Patel College of Engineering, Shahada, Dist-Nandurbar Maharashtra 425409*

---***---

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.

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

1. INTRODUCTION

Cloud computing uses number of techniques in which PaaS a good application development platform for the developer to create 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).

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.

2. 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 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 [13].

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.

3. PROPOSED METHODOLOGY

TF –IDF. It is popular and effective scheme in information retrieval. We apply tf-idf technique to make inverted index of term. If 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).

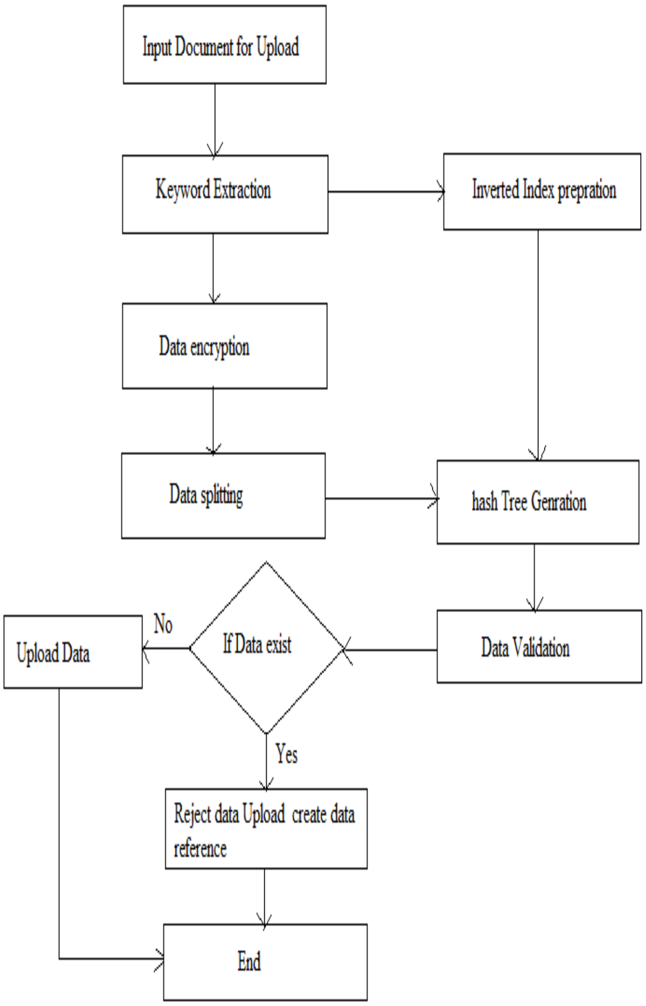


Fig -1: System flow diagram

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 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.

Table -1: Data Encryption Process.

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
7. End for
```

4. IMPLEMENTATION & RESULTS

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.

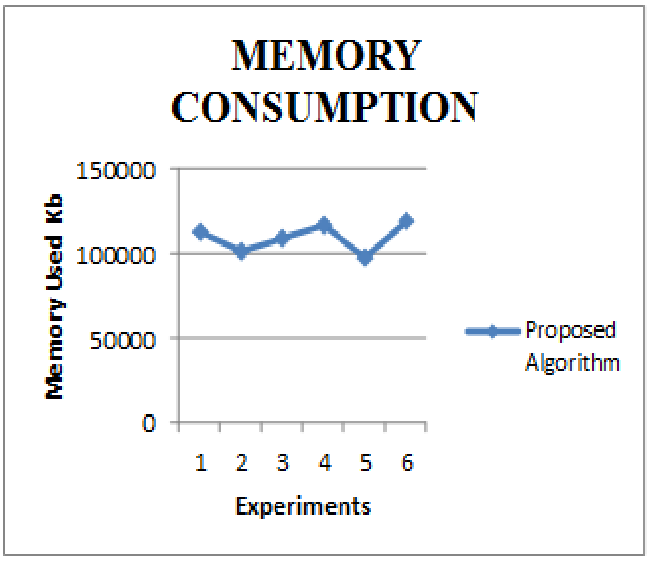
4.1. Memory Consumption

The total memory consumption of the algorithm is computed using the following formula.

$$Consumed\ Memory = Total\ Memory - Free\ Memory$$

The table 2 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.

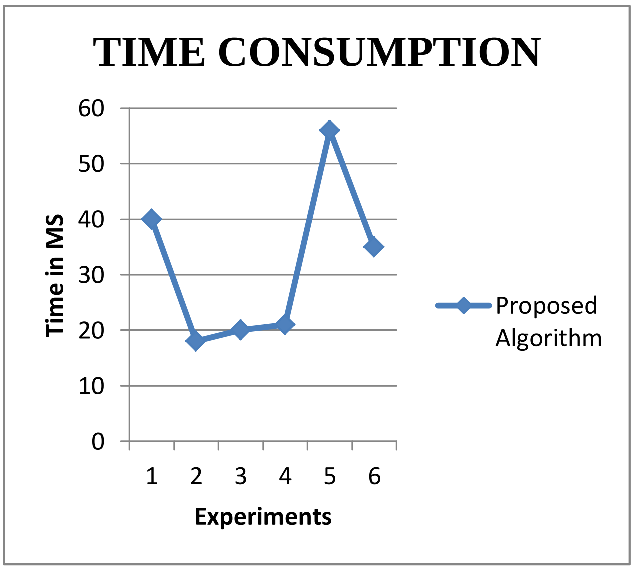
Table -2: Memory Consumption.



4.2. Uploading Time Consumption

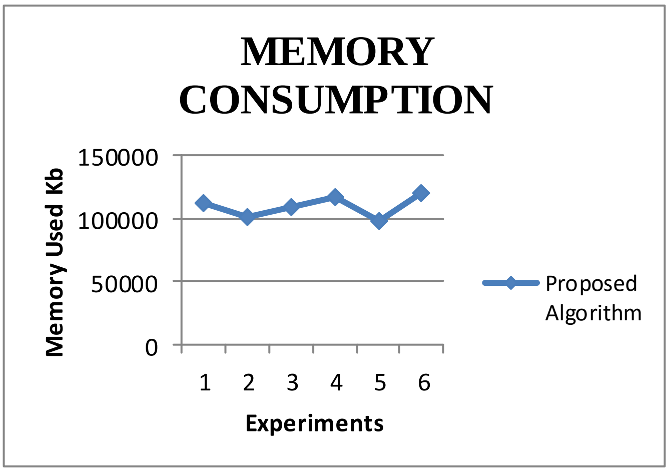
The amount of time required to develop the upload a data file on the server for cryptographic model is termed as the time complexity of the algorithm or time consumption of system. Table 3 shows the total time required to upload data.

Table -3: Time Consumption.



4.3. Downloading Memory Consumption

Table -4: Download Memory Consumption.



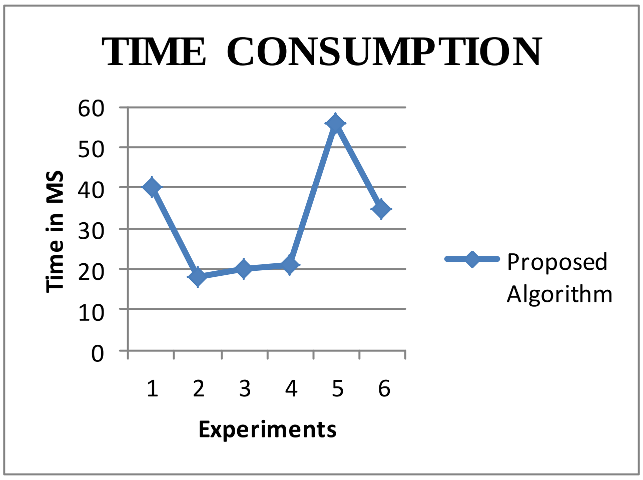
The algorithms need a significant amount of main memory to store the data for processing. This storage requirement is termed as the memory consumption or the space complexity of the system. Here the downloading based memory consumption is computed.

Table 4 shows the total memory required to download data.

4.4. 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.

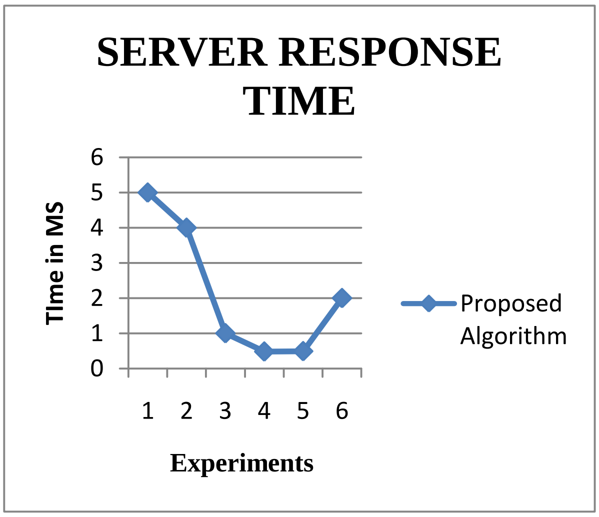
Table -5: Download Time Consumption.



4.5. 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 -6: Server Response Time.



The computed response time of the proposed technique for cloud based secure communication is demonstrated using the table 6. 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.

5. CONCLUSION

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. A simple approach that makes de-duplication compatible with encrypted data. A strategy needs to study for data duplication and secure transmission over cloud computing environment. We work for a new security approach for secure data transmission 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, 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] Ajahar Ismailkha Pathan, Amit Sinhal, “Encode Decode Linux based Partitions to Hide and Explore File System”, International Journal of Computer Applications, Vol. 75(12), pp. 40-45, Aug 2013.

[7] Prof. Liladhar M. Kuwar, Prof. Ajahar Ismailkha Pathan, “Implementation of Chaotic Neural Network Using Chaos for Data Encryption”, international journal for science and advance research in technology (IJSART), Volume 4 Issue 5, ISSN [Online]: 2395-1052, Page | 1788-1792, May 2018.

[8] Jangid sheetal Kailash, Vaishnika Balmukund Patil, Neha Vinay Patil, Ajahar Ismailkha Pathan,” Behavioural, Emotional State Based Music Selection & Playlist Generating Player”, International Journal Of Current Engineering And Scientific Research, Vol-4(12), pp. 39-43, Dec 2017.

[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 Anand2, “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] Ajahar Ismailkha Pathan, “Proposed: Tech Learning Community Management”, International Journal for Scientific Research & Development(IJSRD) Vol. 5, Issue 07, 2017 | ISSN (online): 2321-0613, Aug 2017.

[12] Ajahar Ismailkha Pathan, Shezad Habeeb Shaikh, “A Survey on ETS Using Android Phone”, International Journal Of Innovative Research In Technology (IJIRT), Volume 5 Issue 3 | ISSN: 2349-6002, Page-99-104, August 2018.

[13] International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438 Volume 4

Issue 8, August 2015 www.ijsr.net Licensed Under Creative Commons Attribution CC BY A Hybrid Cloud Approach for Secure Authorized Deduplication Jagadish 1, Dr.Suvarna Nandyal2.

[14] 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

AUTHORS

Prof. Ajahar Ismailkha Pathan
Dept. of Computer Engineering
PSGVPM's DNPatel COE, Shahada.

Prof. Liladhar M.Kuwar
Dept. of Computer Engineering
PSGVPM's DNPatel COE, Shahada.

Prof. Rijavan Abddulrahim Shaikh
Dept. of Computer Engineering
PSGVPM's DNPatel COE, Shahada.

Prof. Dheeraj Basant Shukla
Dept. of Computer Engineering
PSGVPM's DNPatel COE, Shahada.