International Journal of Advanced Research in Science Management and Technology
ISSN(Online)-2454-4159 Volume 3, Issue 3, March 2017

"Survey of Various Data Deduplication Method and Security Algorithm Issue In Cloud Computing"

Gauravkumar Pralhad Patel[1], Prof. Kailash Patidar [2]
PG Scholar, Department of Computer Science Engineering, SSSIST, Sehore, M.P., India[1]
Professor, Department of Computer Science Engineering, SSSIST, Sehore, M.P., India[2]

Abstract— Cloud term is meant thusly sort arrange where data stow away, for example, web access supplier system .and registering term is indicated, for example, information preparing and processing to take care of the issue in PC. Information preparing apply anyplace in PC or any figuring gadget such as single PC or any server device. In arrange number of server accessibility and distinctive class separates is reliant on site information. We Computers Internet administrations and correspondence fringe gadgets application are deploying quick and utilization of distributed computing has expanded quickly in this day and age. Lessening expense is real idea of web and utilizing single server cost is so high so we utilize shred facilitating ,virtual private server and devoted server to decrease cost of web availability .Content assume a noteworthy part in web a ton of substance will associated more number of individuals and builds administrations heap of web and to enhance content we acquaint free web access with permitted client for substance composing by utilizing email, web journals, Wikis, Forums , Chat, Social Networking, web based shopping.
We utilized server farm and this server farm is consolidated thousands of PC server combined as a cluster. Data focuses joined with each other is called framework of server farms and this group and lattice mix of system present distributed computing. Web fills in as cloud and all server farms additionally act as a cloud in web. So web and all related enormous server farms organization make assertion for every huge data focuses to give condition as administrations work as business perspective administration act as distributed computing.

Keywords— Cloud term, Internet administrations, Lessening expense, virtual private server and devoted server.

I. INTRODUCTION

Cloud show give on request organize administrations to servers, stockpiling, applications, and administrations. Today situation practically information is put away in advanced frame because of improvement in systems administration and capacity innovation. to be use information we put away in the cloud space since plate space is exorbitant to put away and keep up.

Versatility: Website engineer or client make there on disjoin and pay a ton of many to manintain the facality of there claim server . To utilize their own one of a kind processing structure, client or designer must make an interest in equipment and programming bundle. In the event that necessity on their frameworks later increment, have need to give extra assets and including them with their current structure. Besides, if stack in the long run decays, clients are staying with relinquished limit. With distributed computing , customer speak with server supplier notwithstanding, clients ought to interface this buy just these assets yet they wish in little additions and may effortlessly alter the assets put resources into them in answer to changes effective.

Supply, dependability, notwithstanding worldwide accessibility: Cloud figuring imarging as a wide supply managar Because cloud suppliers are in the from more noteworthy organizations of size than the users.In cloud a considerable measure of clients utilize business of highlight and utilize registering assets to a ton clients they regularly have more noteworthy abilities in overseeing frameworks. Therefore, their frameworks will have higher accessibility and consistency than frameworks clients could field freely. cloud suppliers' servers permits clients to utilize the information from anyplace without expecting to run their own dependably on server with a universally reachable IP address.

Practicality and accommodation: Cloud organizations assets and server farms are accessible in gigantic sum for client to give abstracting without end data of these hidden segments, and once in a while, the product, cloud organizations

free shoppers from maintaining individual's assets. Particularly, programming as-an administration shoppers regularly can utilize an application without expecting to expressly mount any product, essentially by method for exploring to Website.

Distributed computing is the most recent pattern in registering administration arrangement. This sort of administration has seen this innovative notwithstanding social modify of figuring administration supply from being given in your group to being given remotely by outsider organizations . Usefulness, for example, equipment, handling and furthermore other usefulness has gotten to be offered on request, as an administration notwithstanding both unreservedly and from cost. The customer has viably lost control over how their information is being put away, shared and utilized, in addition to over this security used to ensure the information.

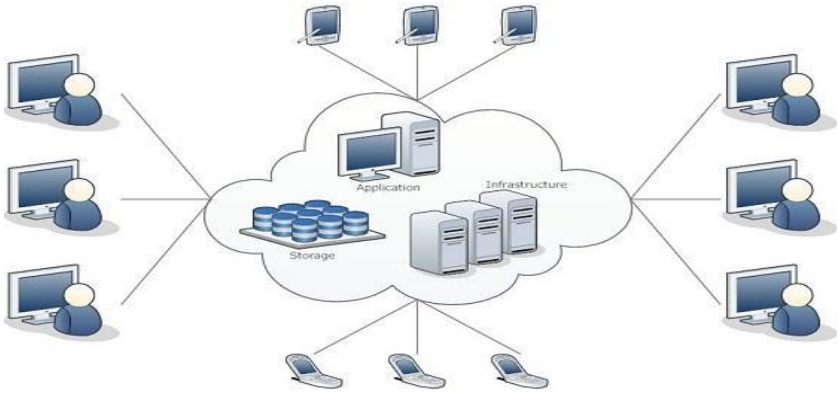

Fig.1 Cloud Computing Model

II. LITERATURE SURVEY

[1]. Cloud provide great achievement, it also introduces a many of security threats to the information and data which is now being ported from on-premises to off-premises. This paper discusses the various service models such as IaaS, PaaS, SaaS. Some products offer Internet-based services—such as storage, middleware, collaboration, and database capabilities—directly to users such as Iaas,Paas,Saas.Most cloud computing systems in operation today are proprietary, rely upon infrastructure that is invisible to the research community, or are not explicitly designed to be instrumented and modified by systems researchers. This paper also explores the types of cloud computing environment. Paper also discuss the cloud database including deployment model, characteristics. Also it identifies some technological and legal issues in cloud computing.

[2]. Compute clouds are enormous server farms packed with computing power and storage space accessible through the Internet. Instead of having to manage one's own infrastructure to run applications, server time and storage space can be bought from an external service provider. From the customers' point of view the benefit behind this idea is to be able to dynamically adjust computing power up or down to meet the demand for that power at a particular moment. This kind of flexibility not only ensures that no costs are incurred by excess processing capacity, but also enables hardware infrastructure to scale up with business growth. But security is very big concern in the data sharing in the network. Same as in cloud computing data security is also a very sensitive matter. This paper discusses the concept of Cloud computing to achieve a complete definition of what a Cloud is and its security issues, More than 20 definitions have been studied allowing for the extraction of a consensus definition as well as a minimum definition containing the essential characteristics. This paper pays much attention to the security issues of Cloud computing.

[3] The Attribute based accessto extensible media in cloud to help sharing the contents to the network.The data encryption standard cipher text policy algorithm is use. It's a important to protect the data from unauthorized access. Media contents such as images, audio, video, and text etc..These contents are protected by this security method. . All the storage is done on cloud so the security of data is threatened. So there is need to secure the data on the cloudThe public cloud services are sent to the client by internet. The load balancing technic is use in cloud computing.Cloud Load

Balancing simply means doing load balancing, i.e. allocate workload across different servers, network channel, CPU, disk drives, on the cloud, i.e. inside the Internet itself. This method is use helps to obtain , as other load balancing methods, helps to achieve optimal system utilization, maximum throughput, minimum response time, and prevent overload of system.

[4] Privacy security is a key issue for cloud storage. To solve this problem, the paper1 proposes a privacy-preserving cloud storage framework, which includes the design of data organization structure, the generation and management of keys, the treatment of change of users' access right and dynamic operations of data, and the interaction between participants. We design an interactive protocol and an extirpation-based key derivation algorithm, which are combined with lazy revocation, multi-tree structure and symmetric encryption to form a privacy-preserving, efficient framework for cloud storage. A system is realized which is based on the framework. The paper analyzes the effectiveness of extirpation-based key derivation algorithm, the overhead of the system and the privacy security of the framework. Finally, we summarize our work and introduce our future research directions.

[5] Cloud Computing is the recent trend in IT industry that changes the way the business is done by delivering computing as a service on demand to the user by pay per use model. Major advantages such as reduced cost, flexibility ensure cloud computing to be much sorted technology in the computing industry. Cloud computing moves the data to the large data centers which is not trustworthy as there is chance for the cloud service provider to misuse the data without the knowledge of the users. We introduce a novel model which reduces the fear of loss of data confidentiality in the cloud service provider side. This model has third party security vendor which takes care of encryption and decryption of the data based on the preference of the user. So data can reside safely on the cloud service provider side.

[6] We collected file system content data from 857 desktop computers at Microsoft over a span of 4 weeks. We analyzed the data to determine the relative efficacy of data deduplication, particularly considering whole-file versus block-level elimination of redundancy. We found that whole-file deduplication achieves about three quarters of the space savings of the most aggressive block-level deduplication for storage of live file systems, and 87% of the savings for backup images. We 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. Categories and Subject Descriptors: D.4.3.

III. PROPOSED APPROACH

In our proposed framework different information and records are put away by information proprietor. Records where put away in cloud server. CSP cloud server supplier where recognize copy information utilizing altered list procedure. When information de duplication is checked then markle hash tree create hash capacity to produce secure key and encryption and decoding perform at client end. Our proposed framework contain taking after term portray underneath

1. Reversed Index: We proposed a rearranged file based information de duplication method. In this information de duplication plot defeat the issue of seeking copy information in informational collection. Our plan bolster conjunctive multi key world looking. Upset list is a key information structure hidden present day Information Retrieval. For each term t , we should store a rundown of all archives that contain term – distinguish each reports by a docID , a records serial number .We utilized settled size cluster for this.

So we require variable size postings is typical and best .on circle , a persistent keep running of postings is ordinary and best . in memory can utilize connected rundown or variable length exhibit .

Altered record development :-

Records, tokenizer (TOKEN STREEM), strategic modules(MODIFIED TOKEN) indexer (INVERTED INDEX)

Starting phase of content preparing:-

Tokenization:- slice character grouping into world tokens .

Standardization:- outline and question term to same frame ex you need U.S.A and USA

Stemming:- we may wish to various types of a root to match ex approve approval

Stop world : we may preclude extremely normal words (or,not) ex the, a,to of .

Building a modified record for de duplication any sort of seeking framework requires to work various errand while parsing the page or any reports.

1 getting the records.

2 evacuating the stop words like "is ", "a", "we","the" and so on.

3 stem to the root word. Recover may get to be "retrive" ,watchman stemmer.

4 Record reports id. We need to remembered reports as ID retrieve==> docID30 . On the off chance that I get same word in different archives , I can compose retriev == > docID30&docID40. I can enhance by term show up in report for give rank . so I will compose as retrieve== > docID30 |3| &docID40 |1|.

5. Union and store the terms.

2.TF-IDF: Tf-idf is best known weighting plan in data recovery plot. In our proposed calculation we are applying TF-IDF strategy in Inverted record to discover number of times of a world shows up in the reports .TF-IDF remains for term recurrence opposite archive recurrence

Regularly, the tf-idf weight is formed by two terms: the primary processes the standardized Term Frequency (TF), otherwise known as. the quantity of times a word shows up in a report, isolated by the aggregate number of words in that archive; the second term is the Inverse Document Frequency (IDF), registered as the logarithm of the quantity of the records in the corpus partitioned by the quantity of archives where the particular term shows up.

• **TF:** Term Frequency, which measures how as often as possible a term happens in a report. Since each record is distinctive long, it is conceivable that a term would seem significantly more circumstances in long reports than shorter ones. Therefore, the term recurrence is regularly separated by the report length (otherwise known as. the aggregate number of terms in the archive) as a method for standardization:

TF(t) = (Number of times term t shows up in a report)/(Total number of terms in the record).

• **IDF:** Inverse Document Frequency, which measures how critical a term is. While registering TF, all terms are considered similarly vital. In any case it is realized that specific terms, for example, "is", "of", and "that", may show up a considerable measure of times however have little significance. In this manner we have to overload the incessant terms while scale up the uncommon ones by registering the accompanying:

IDF(t) = log_e(Total number of records/Number of archives with term t in it).

• Consider a record containing 100 words wherein the word run shows up 3 times. The term recurrence (i.e., tf) for run is then (3/100) = 0.03. Presently, accept we have 10 million archives and the word feline shows up in one thousand of these. At that point, the opposite archive recurrence (i.e., idf) is figured as log(10,000,000/1,000) = 4. Accordingly, the Tf-idf weight is the result of these amounts: 0.03 * 4 = 0.12.

The tf idf weight of a term is the result of its tf weight and idf weight.

W td = log(1+tf td) * log 10 (N/df t) Figure tf-idf score is weight of tf-idf .

Figure the estimation of archives recurrence for the term that show up in each records .so reports recurrence is depand upon the term show up in the reports.

So Tf score rely on upon both the terms and records and the idf score depands on the term. Wen we join both of things so the general weight depands on both the records and terms (this was made a cound grid shape yet in any reports vector portrayal does not consider the requesting of words in an archives Ex johan is quick then wed and mary is quick then

johan have the same vectore this is known as the sack of worda model. It could be said , this is a stage back the position file could recognize these two archives . we will take a gander at recouping positional infoemtion later .

The log recurrence weight of term t in d is { 1+log 10 tf id if tf id >0 W tf= { 0 generally

3. Merkle hash tree : Our proposed framework work with Merkle Hash Tree is an all around examined verification structure [7]. It is utilized to proficiently demonstrate that an arrangement of components are undamaged and unaltered. It helps enormously in decrease of server time [9]. It is utilized by cryptographic techniques to confirm the record squares. The leaf hubs of the MHT are the hash estimations of the first document squares. The thought behind producing MHT is to break the document into various pieces. Apply hashes to the credible information values i.e. the first document squares and join iteratively. Presently, reiterate the outcome hash hubs and consolidate in a tree-like design and rehash this method till we get a tree with a solitary root. The MHT is produced by the customer and is put away at both the customer and the server side. Fig 2 portrays a case of MHT. The tree has four leaf hubs viz. m1, m2, m3 and m4. At first, we apply hash on each of these record pieces and get h(m1), h(m2), h(m3) and h(m4). At that point, h(m1) and h(m2) are hashed and joined together to get ha. Comparative process occurs with pieces m3 and m4 and here, we get hb. Here, h is a protected hash work. This can be communicated as ha = h(h(m1)|| h(m2)) and hb = h(h(m3)|| h(m4)) Further, ha and hb are joined and repeated to acquire the root as hr. This can be communicated as hr= h(ha|| hb)

4. AES: AES is a square figure. The calculation underpins an assortment of key sizes as 128,192 or 256. The default size is 256 bits. The encryption of information squares is done in 10, 12 and 14 rounds relying upon the extent of the key utilized. It gives quick and adaptable encryption and can be effortlessly executed on different stages. In this paper, AES-128 is utilized thus encryption is done in 10 rounds. This calculation is utilized for both encryption and decoding. For encryption, it takes information squares and the mystery key as the info and yields the scrambled information pieces. For unscrambling, scrambled information pieces and key are given as sources of info and unique record squares are the yield. Why AES? AES has rapid key setup time and a decent key

• Dexterity. It is appropriate for confined space situations as the memory necessity for its execution is less. It makes proficient utilization of assets due to its

• Characteristic parallelism which brings about a decent programming execution. It doesn't have any genuine frail keys.

• Any square size and key sizes are bolstered by

•AES that are products of 32 (more noteworthy than 128-bits) No direct and differential cryptanalysis attacks.

•Have yet been demonstrated on AES.

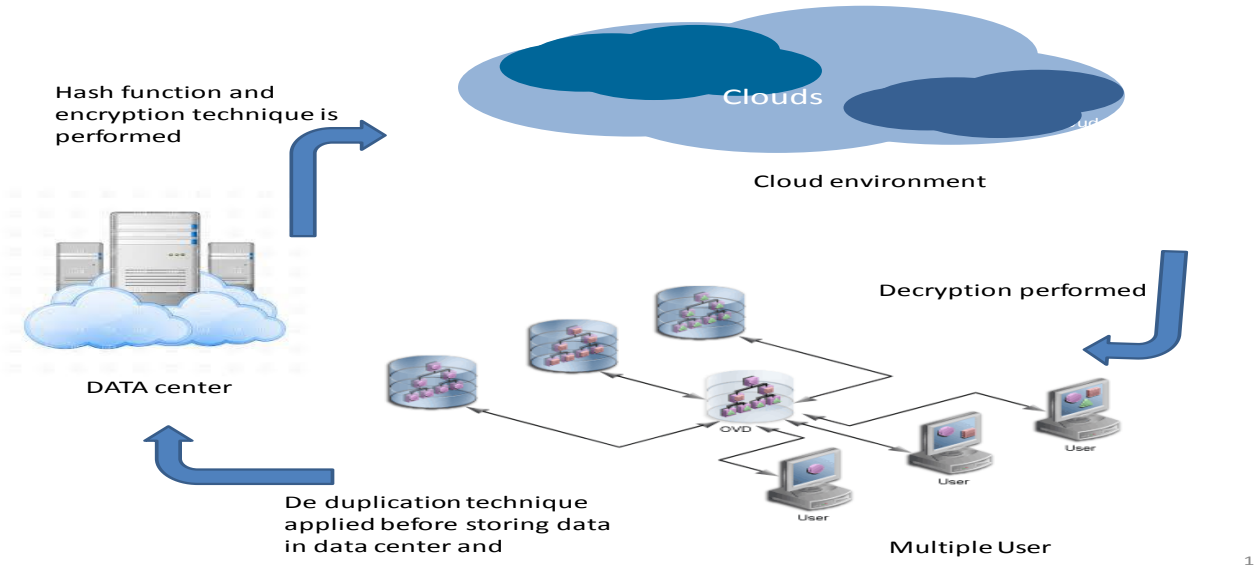

Fig.2 Architecture diagram

IV. RESULT ANALYSIS

1. Uploading Memory Consumption-

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.

$$\text{Consumed Memory} = \text{Total Memory} - \text{Free Memory}$$

Table 1. Memory Consumption

Number of Runs	Proposed Technique
1	112712
2	101772
3	109178
4	117089
5	97833
6	12257

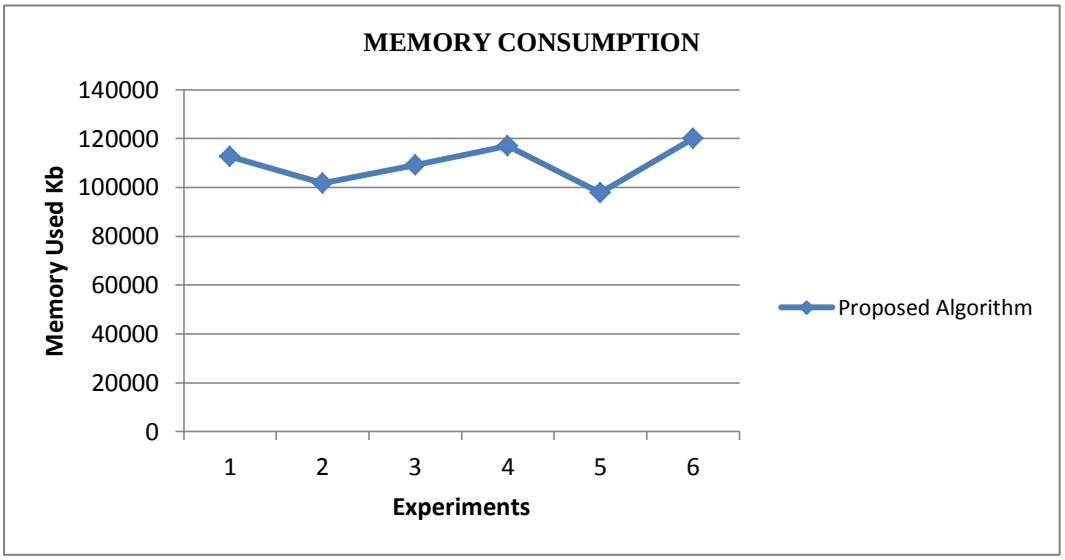

Fig.3 Memory Used

The figure 1 and the table 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.

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. The time consumption is given using following formula:

$$\text{Time consumption} = \text{End time of file put on Server} - \text{Start time of File put on data model}$$

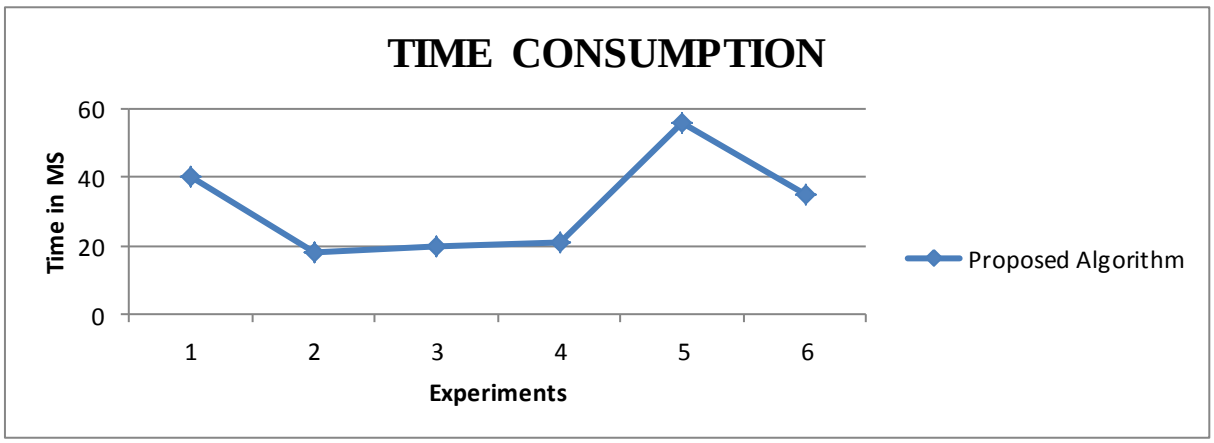

Fig.4 Uploading Time Consumption

Table 2. Uploading Time Consumption

Number of Runs	Proposed technique
1	40
2	18
3	20
4	21
5	56
6	35

The figure 3 and table 4 shows the amount of time consumed during uploading a file on server. In this graph the blue line shows the performance of proposed multiple replica cryptographic approach. The X axis show the different number experiments to analysis the model performance with respective to choose different data files and Y axis shows the amount of time consumed during the encryption process.

3. 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 3. Memory Consumption

Number of Runs	Proposed technique
1	112712
2	101772
3	109178
4	117089
5	97833
6	12257

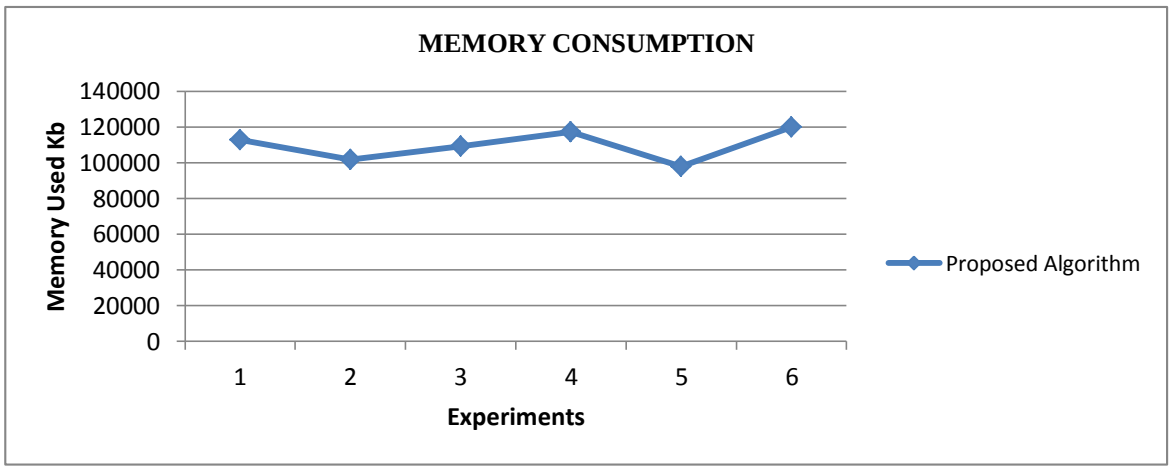

Fig.5 Downloading Memory Used

The memory consumption of the proposed multiple replica cryptography is demonstrated using the figure 5 and table 3. In this diagram the X axis shows the list of experiments and Y axis shows the memory consumption of the implemented system in terms of KB (kilobytes). According to the obtained results the performance of system is depends on the amount of file for process. Thus as the file size increases the required memory is also increases. According to the system performance the proposed technique is adoptable due to less amount of memory consumption.

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. That is computed using the following formula.

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

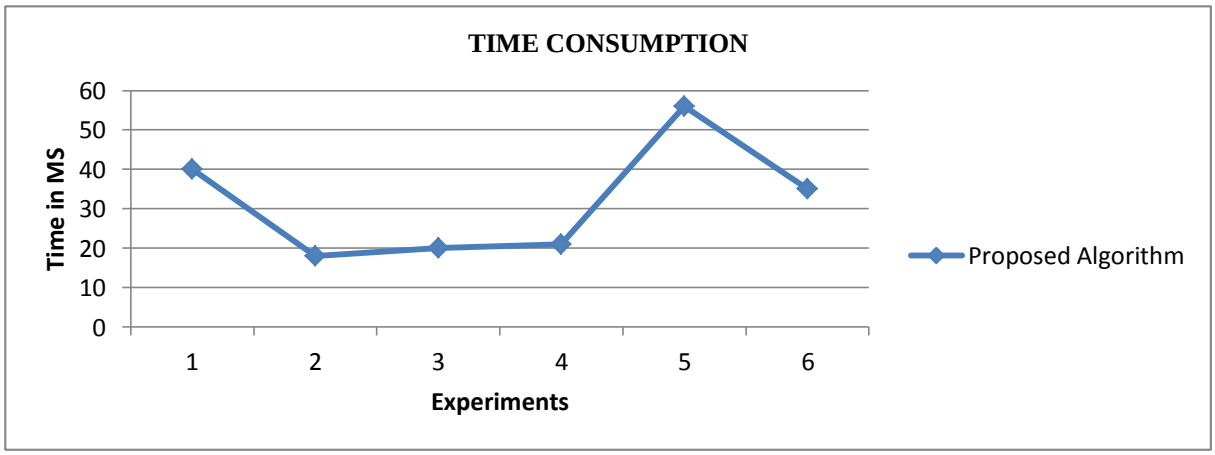

Fig. 6 Downloading Time Consumption

Table 4. Downloading Time Consumption

Number of Runs	Proposed technique
1	40
2	18
3	20
4	21
5	56
6	35

The downloading time complexity of the proposed cryptographic technique is given using figure 6 and table 4. In this diagram the X axis represent the number of experiments and the Y axis shows the amount of time consumed for

downloading the file on system. According to the obtained results the performance of the system is depends on the amount of file size. As the amount of file size is vary the amount of time for downloading is vary.

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 5. Server Response Time

Number of Runs	Proposed Technique
1	5
2	4
3	1
4	0.48
5	0.49
6	2

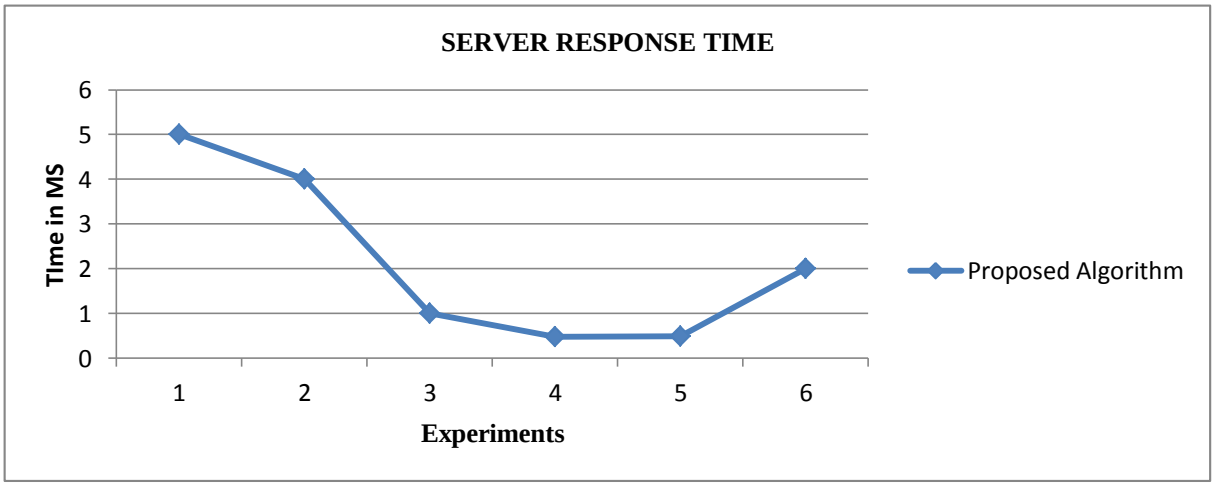

Fig.7 Response Time

The computed response time of the proposed technique for cloud based secure communication is demonstrated using the figure 7 and table 5. 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.

VI. CONCLUSON

Distributed computing improve its execution utilizing this proposed framework. This proposed framework help to enhance execution of information putting away by evacuating information duplication on server. Significant arrangement gave by proposed answer for be discover copy information. Copy information stockpiling is a noteworthy issue for any information base framework. Copy information store numerous duplicate of same data so stockpiling is not use by some other client required for capacity more data into server or any information stockpiling. In our propose framework same duplicate of information is expelled purchase utilizing tf-idf procedure and utilizing rearranged list in our proposed work. We mimic that when we utilize reversed file to avert copy information. Execution of framework builds more as some time recently. Server get to time is increments. Time to make seek with their pertinence regarding accuracy and review and De-duplication time required with the measure of records. In future we work with more duplication system with

secure information transmit particle . we will center to apply de duplication with productive security and concentrate on time proficiency.

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 DeduplicationInternational 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 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] 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] Fatema Rashid, Ali Miri, Isaac Woungang A Secure Data Deduplication Framework for Cloud Environments2012 Tenth Annual International Conference on Privacy, Security and Trust 978-1-4673-2326-0/12/$31.00 ©2012 IEEE 81.