Contents Based Hash Method for elimination of data Duplication in Storage Clouds

[1]Mr. B.V. Satish Babu,Asst.Professor, [2]D.Vamsi Krishna, [3]K.Dilip Kumar, [4]M.Eswar Rao

[1,2,3,4]*Dept. of Computer Science of Engineering*

[1,2,3,4]*Andhra Loyola Institute of Engineering and Technology, Vijayawada*

Abstract— *As Cloud Computing has been emerging from the past decade, gathering information to benefit cloud for progressing towards becoming an fascinating design, which is an advantage in getting tremendous information support and administration. On the whole, as the distributed storage is not dependent ,security issues raise which is a worry to the most efficient method to know information again being duplicated in cloud while performing inspection correctly .Here ,we mainly focus on the operation of correct examination of providing security against de-duplication on the information stored in cloud.*

Keywords*—cloud; de-duplication; hashing; convergent encryption;*

I. INTRODUCTION

Storage in cloud is a method of arranging warehouse of information in virtualized tools of capacity which most of them are provided by third party entities. Distributed storage technologies will give the clients with the benefits like ease of administration, reducing capital expenditure(CapEX) and operational expenditure(OpEx).These outstanding provisions will increase number of clients to avail distributed storage: as per the report, the amount of information in cloud is going to increase to 45 trillion GB in 2025.

In general the fact in cloud storage structured model has been received , that it neglects to obey some vital rising needs, for e.g., the storage capacities of evaluating respectability of cloud by cloud consumers and recognizing duplicate documents by cloud storage servers. We illustrate both problems below.

Here the main issue is uprightness reviewing. The cloud server can highlight its customers from the overwhelming importance of capacity management and support. The most different part of distributed storage from customer point of view is that the information is handshake and transferred by the help of Internet and put separate from dubious space, which not in the regime of the customers by any manner, which definitely raises customers' incredible worries against trustworthiness of owned information.

One more issue is securable de-duplication. The quick appropriation of storage cloud administrations is joined by expanding capacities of information put away at the place of remote cloud servers.

All among these clouds put away records, the vast majority of them are copied: as per a current review by EMC [2], 75 percent of late advanced information is copied duplicates. This reality arises a new innovation specifically de-duplication, where the storage cloud servers might want to de-duplicate by keeping just a solitary duplicate for each record (or piece) and make a connection to the document (or square) for each customer who claims or makes a request to store a similar document (or piece). Lamentably, this activity of de-duplication would prompt various dangers possibly influencing the capacity framework [2], [3].

II. LITERATURE SURVEY

Information protection is more important for most organizations which are outsourcing finance and organizations utilizing outer E-mail administrations to hold delicate data, security is a major concern where frequently referred to problem to distributed computing; experts and doubtful organizations ask "who might believe their fundamental information out there some place?" There are likewise prerequisites for review capacity, according to feeling of Sarbanes-Oxley and HIPAA controls that must be accommodated company information to be migrated to storage cloud.

Storage cloud clients confront security dangers both from internal and external to the cloud. Huge numbers of the security problems required in protecting mists from external dangers like confronting vast storage server farms. In the cloud, this duty is separated among many possibilities, includes the client of cloud , the seller of cloud, and an merchants from outside that clients depend on security-sensitive programming or setups.

The client of cloud controls utilization level security. The supplier of cloud controls physical security, and likely to implement outer firewall strategies. protection for halfway layers is shared between the client, administrator; in the low level of deliberation given to client, the greater pact runs with it.

Amazon EC2 clients will have much specialized agreement for security than do Azure technology who have a greater amount of duties when compared to App-Engine . This client agreement can be given to others who will provide better security administrations. The properties line homogeneous and institutional models of stages for e.g. EC2 make it better for an company to offer,like model design administration

While distributed computing may make outside confronting security less demanding, it poses the new issue of inside confronting security. Cloud suppliers must prepare for burglary or forswearing of-administration assaults by clients. Clients should be shielded from each other.

Trustworthiness examining provable information ownership ensures that objective records are possessed by the cloud server without downloading or recovering the information totally. It was created by ateniese genuine. PDP request the confirmation from server side to demonstrate that the server precisely claims these pieces. PDP on element situation proposed an element PDP pattern yet without addition it is enhanced by presenting validated flip table idea. These are affected by the computational for label creation at the customer side.

To conquer this issue wang et.al presented PDP out in the open cloud. Verification of retrievability does not ensures that the cloud server claim's information, but rather ensures full recuperation of document wang et.al enhanced this by adjusting the merkel hash tree for the square label validation. Xu and Chang enhance the POR composition with polynomial responsibility for diminishing the correspondence cost. Le et al considered another distributed storage design with two autonomous cloud server for trustworthiness inspecting to diminish the customer side load operations.

Leetam used the key-scatter paradigm in order to settle the solve problem of a noteworthy number of joined keys in concurrent encryption Secure de-duplication is a technology in which just a single duplicate of the document can be spared at server side with the end goal of circle space of storage cloud servers and additionally organize data transmission is spared. De-duplication at customer side prompts spillage of side channel data.

To maintain a strategic distance from this issue Halevi et al. presented the verification of possession convention that lets a customer productively demonstrate to a server that customer abstractly contained with this record. A few method of proprietorship conventions in view of the Merkle's hash tree is introduced to empower protective customer side de-duplication.

Pietro and Sorniotti proposed a proficient evidence of possession plan by picking the projection of a record onto some haphazardly chose bit-positions as the document verification. A different profession for secure deduplication concentrates on the classification of de-duplicated information and considers to make deduplication on scrambled information.

Ng et al. firstly presented the private information deduplication method as a basis of open information deduplication conventions of Halevi et al. introduced an encryption that is promising cryptographic basic method for guaranteeing information protection in method of de-duplication. Bellare et al formal the primitive as message based encryption, and one more applications is investigated in space - productive secure out sourced stockpiling.

Abadi et al. enhanced the Bellare et al's security definitions by taking plain content dispersions that is rely upon people in general parameters of the patterns. With respect to useful usage of joined encryption for better de-duplication, Keelveedhi et al. outlined the DupLESS framework in which customers scramble under document dependent keys resulted from a key generating server by means of a negligent pseudorandom work convention.

Every one of the works characterized above considered either deduplication or trustworthiness evaluating, while in this paper creators have endeavored to tackle both issues at the same time, furthermore, it is advantageous taking note of that our work is additionally recognized with which reviews the information in the cloud with deduplication since they likewise consider A) outsource the calculation of label era. B) review and de-duplicate scrambled information in the conventions

III. PROPOSED METHOD

Here we determine that proposed framework has accomplished both trustworthiness inspecting and document based de-duplication. In any case, it not possible to keep the cloud storage servers from supervising the content of records that have been consider away.

As it were, the properties of trustworthiness inspecting and secure de-duplication are just enforced on plain text documents. In this arena, we introduce a framework, which uprightness reviewing and de-duplication on encoded records. Cloud storage clients has vast content records to be keep away and based on the cloud for data upkeep and computations. They can be single purchasers or business related associations.

Cloud Servers virtualized the important entities as per necessities of clients and to discover them as storage capacity pools. Ordinarily, the cloud consumers will purchase or rent stockpiling from storage cloud servers, and keep their individual data

in these purchased or leased spaces for future usage. Evaluator helps customers transfer and analyze its external information that implements a MapReduce cloud and works like an authentication specialist.

This decision assumes that the important point is related with a few of open and private related keys. Its open key is made and it is accessible to other elements in the framework

The outline objective is of recording classification requires to keep the cloud servers from getting to the substance of documents. Extraordinarily, we need that the objective of document secrecy should be impervious to "word reference assault". That is, even the enemies have pre-information of the "word reference" which incorporates all the conceivable records, despite everything they can't recoup the objective document.

The Below Figure portrays that: The User will tries to transfer a record and that document will goes to approval prepare for approving the data. In approval handle they experience into two process sends reference and transfer. In the event that the transferring is now existed then it goes under sends reference and if transferring information is not existed before in the database it will checks and compute the hash estimation and produces some hash value of an information and stores in the database and record will straightforwardly transferred into Storage clouds.

The validator Checks whether the database contents the hash value if the hash value is already existing in the database it will sends reference of file where the user will be trying to upload and the file will be referred and checked and download that file if he wants that or simply visits that file and compare to his older file and make use in cloud storage.

If the validator approves that the data which is not present in the database, then it will undergo into a process and checks the hash based value of the data present in the file.

The new hash based value which is calculated while the file uploads into the database is calculated and it will store in a table in our database and the file will directly upload into the cloud storage.

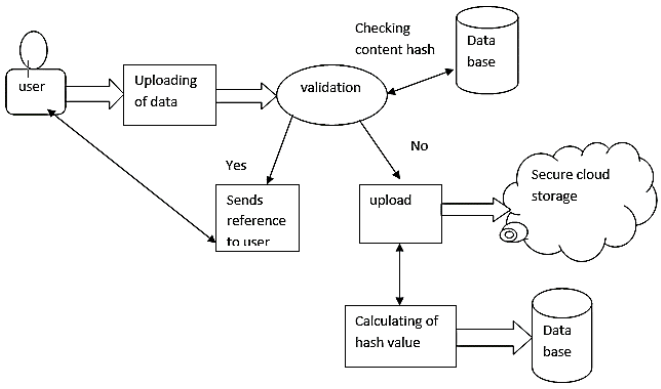


Figure 1: Architecture of Proposed system

IV. RELATED WORK

Installing Dropbox (storage cloud) and Create an account in Dropbox (storage cloud) Login to that account and make an App in dropbox API. it will give one App key, Secret key and Token to permit into dropbox without the entrance of dropbox account.

From that point we are utilizing these App Key, Secret Key, and Token those are permitting client to do all operations from front end like upload, delete, download and view operations.

We made front end by utilizing HTML and CSS and associations through PHP. By utilizing wamp server we are running our PHP scripts and made one database to store values in the table.

Running our **wamp server** in any web program that feels comfort for us and attempting to run our primary application in web server it results to home page where we can upload, view and so on., when we are transferring a record which contains a few information to the cloud, it will check the hash value and checks whether the hash value of this document were at that point show in information base or not.

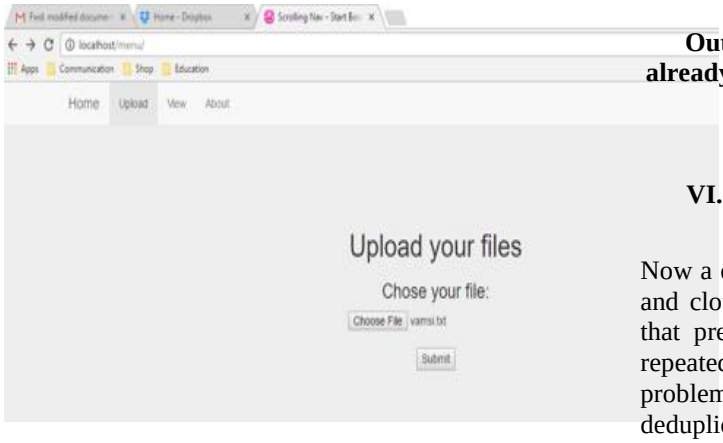
If the hash value is as of now present in the database then the document would not be transferred and make an impression on the client the record is as of now existed and if the hash value was not found in database then the document is transferred into the cloud, then the client gets the notice that the document is transferred to the cloud and the hash value is put away in database.

Now the user will get access to the file and make usage of data that is present in the cloud storage server

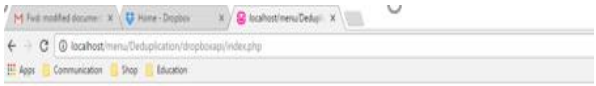
V. RESULTS

The following snapshots will contain the brief description of a project i.e., from where we have to start how to start our project it will contains Home page, Upload page, View page and About Section in our Snapshots

Output 1: Home Page



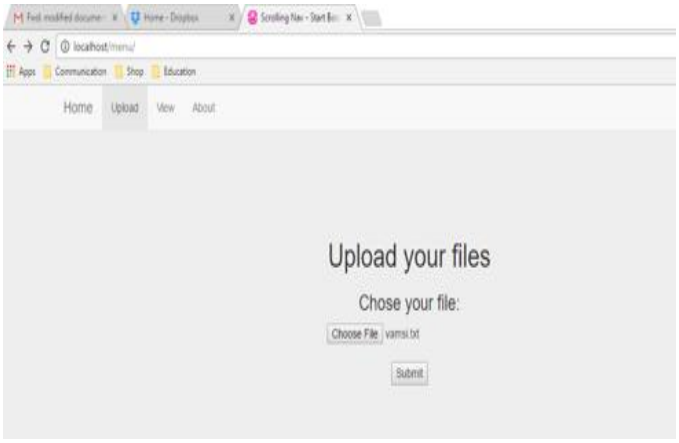
Output 2: Uploading File



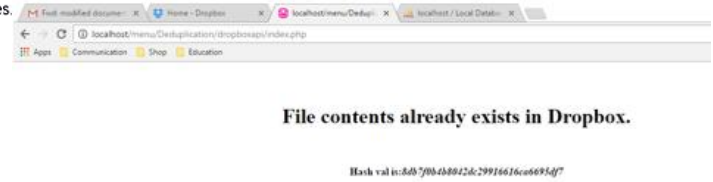
File is uploaded into Dropbox.

Hash val is:*8db7f0b4b8042dc29916616ca6695df7*

Output 3: The file is uploaded into dropbox and hash value for the content of the file



Output 4: Now again trying upload same file which is already present



File contents already exists in Dropbox.

Hash val is:*8db7f0b4b8042dc29916616ca6695df7*

Output 5: Displays a message file contents already exist in Dropbox that means we achieved de-duplication

VI. CONCLUSION AND FUTURE WORK

Now a days usage of cloud is increasing day by day and cloud becomes more precious to store data. In that precious memory we are uploading our files repeatedly by our mistake, by that deduplication problem occurs for that we are achieving deduplication in our project by using hash value of the file content. If the file content once uploaded it will not uploaded again and again.

In our project we uploaded multimedia files also. While uploading .3gp file it will takes much time to upload into dropbox and file content with more than 10mb of size it will not uploading into dropbox

A Limitation to our venture is the substance is hashed in record, if the document substance is 100% equivalent to another record content then just it won't permit to transfer document. For that, the future work is evading tab spaces and the substance which is close equivalent to the substance display in the other which is attempting to transfer.

VII. References

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A.Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M.Zaharia, "A view of cloud computing," Commun. ACM, vol. 53,no. 4, pp. 50–58, 2010.

[2] J. Yuan and S. Yu, "Secure and constant cost public cloud storage auditing with deduplication," in Proc. IEEE Conf.Commun. Netw.Secur., 2013, pp. 145–153.

[3] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," in Proc. 8th ACM Conf. Comput. Commun. Secur., 2011, pp. 491–500.

[4] S. Keelveedhi, M. Bellare, and T. Ristenpart. (2013). Dupless: Server-aided encryption for deduplicated storage in Proc. 22nd USENIX Conf. Secur., pp. 179–194 [Online]. Available: https://www.usenix.org/conference/usenixsecurity13/technicalsessions/presentation/bellare

[5] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner,
Z. Peterson, and D. Song, "Provable data possession at untrusted storages," in Proc. 14th ACM Conf. Comput. Commun. Secur., 2007, pp. 598–609.

[6] G. Ateniese, R. Burns, R. Curtmola, J. Herring, O. Khan, L. Kissner, Z. Peterson, and D. Song, "Remote data checking using provable data possession," ACM Trans. Inform. Syst. Secur., vol. 14, no. 1, pp. 1–34, 2011.
[7] G. Ateniese, R. Di Pietro, L. V. Mancini, and G. Tsudik, "Scalable and efficient provable data possession," in Proc. 4th Int. Conf. Secur. Privacy Commun. Netow., 2008, pp. 1–10.

[8] C. Erway, A. K€upc¸€u, C. Papamanthou, and R. Tamassia, "Dynamic provable data possession," in Proc. 16th ACM Conf. Comput. Commun. Secur., 2009, pp. 213–222.

[9] A.Faritha Banu, C. Chandrasekar "A Survey On Deduplication Methods"International Journal of Computer Trends and Technology **(IJCTT),**V3(3):343-347 Issue 2012 .ISSN 2231-2803. Published by Seventh Sense Research Group..

[10] Jyoti Malhotra, Jagdish Bakal "FiLeD: File Level Deduplication Approach". **IJCTT** V44(2):74-79, February 2017. ISSN:2231-2803. www.ijcttjournal.org. Published by Seventh Sense Research Group.

[11] Y. Zhu, H. Hu, G.-J. Ahn, and M. Yu, "Cooperative provable data possession for integrity verification in multicloud storage," IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 12, pp. 2231–2244, Dec. 2012.

VIII. Web and video references

[13]https://www.youtube.com/watch?v=FsQZyNpDWv0&t=451s
[14] https://www.youtube.com/watch?v=xFM7_1pdiFE

[15] https://www.youtube.com/watch?v=2cIlcsrk2nA

[16] https://www.youtube.com/watch?v=wfb6h9JyhBY

[17] https://www.youtube.com/watch?v=2puV9yXHiAA

[18] https://www.youtube.com/watch?v=wlB276xVgsw

B.V.Satish,currently working as an assistant professor in "Andhra Loyola institute of Engineering and technology". His areas of interests include cloud computing, Big data analysis, Image processing, IOT, Distributed systems, Network Programming, Computer programming, Server programming.

D. Vamsi Krishna currently pursing B.Tech degree in Computer Science and Engineering at Andhra Loyola Institute of Engineering and Technology (ALIET). His Research area include Cloud Computing

K. Dilip Kumar currently pursing B.Tech degree in Computer Science and Engineering at Andhra Loyola Institute of Engineering and Technology (ALIET). His areas of interest include cloud computing

M. Eswar Rao currently pursing B.Tech degree in Computer Science and Engineering at Andhra Loyola Institute of Engineering (ALIET). His areas of Interest include Cloud Computing