Encrypted Big Data with Data Deduplication in Cloud

Priyanka G. Masal
PG Department
MBES's College of Engineering
Ambajogai, India, 431 517

B. M. Patil
PG Department
MBES's College of Engineering
Ambajogai, India, 431 517

ABSTRACT
Cloud computing is a service by re-arranging resources in the Internet. Cloud service is popular for data storage. The data holder's privacy is the data stored in cloud and in the encrypted form. The cloud data deduplication is new challenges by the encrypted data, which is for processing in cloud and big data storage. The deduplication is not working on encrypted data. Data has wide applications in zones like keeping money, investigative exploration, prescription and government offices. Order is one of the ordinarily utilized assignments as a part of information mining applications. For back as decade, due to the ascent of different protection issues. The numerous hypothetical and commonsense answers for the order issue have been proposed under diverse security models. The late fame of distributed computing with notwithstanding, clients now has the chance to outsource their information, the information mining assignments and in encoded structure to the cloud. The information on the cloud is in existing security protecting characterization methods and encoded structures are not appropriate. In this paper, the characterization issue over encoded information in system concentrates on fathoming. System proposes a safe k-NN classifier over scrambled information in the cloud. The index is created with the help of Vector base cosine similarity (VCS) multiple strings matching algorithm which matches the pre-defined set of keywords with information in the data files to index them and store relevant data. The classification of information, the information access designs and convention ensures security of client's data inquiry is proposed in this system. To the best of their learning, there work is the first to add to a safe k-NN classifier over scrambled information under the semi-legitimate model.

Keywords
Cloud Computing, Access Control, Big Data, Data Deduplication, Encryption data, k-NN classifier, Security , outsourced databases.

1. INTRODUCTION
Lately, the cloud computing model is changing the landscape of the organizations way of working their information especially in the way they save access and process data. As a growing processing model, cloud processing draws many organizations to think about seriously concerning cloud potential with regards to its cost-efficiency, versatility, and offload of management expense. Most often, organizations assign their computational functions in improvement to their information to the cloud. Regardless of remarkable benefits that the cloud offers, security and comfort issues in the reasoning are avoiding companies to utilize those benefits. The information need to be encoded before freelancing to the cloud the information is extremely delicate. Nevertheless, when information are secured, regardless of the actual security plan, executing any information mining tasks turns into very complicated without ever decrypting the information.

There are other privacy worries, confirmed by the following example. Assume an insurance provider contracted its secured clients database and relevant data mining task to a cloud. When a representative from the company needs to figure out the threat stage of a potential new client, the representative can use a classification method to figure out the threat stage of the client. Initial, the representative requires generating a details history q for the client containing certain private details of the client, e.g., credit rating, age, marriage status, etc. Then this history can be sent to the cloud, and the cloud will estimate the class label for q. However, since q contains vulnerable details, to secure the customer's privacy, q should be encoded before delivering it to the cloud. The above example reveals that data mining over encoded information on a cloud also requires securing a user's history when the history is a part of a data mining procedure. The delicate information about the real information products by monitoring the information accessibility styles and cloud can obtain helpful even if the information is encoded.

Cloud computing is a service by re-arranging resources (e.g. storage, computing) and providing them to the users demands. The cloud computing is internet based computing. It provides shared file and computer resources. The big network resources provided by the cloud computing. The secure and confidentiality is important on cloud computing. Therefore authorized de-duplication only gives effective result. Cloud computing has desirable properties, such as elasticity, fault-tolerance, scalability and per-use. Thus, it has become service platform. The client transfers their information to the server form of cloud service provider (CSP) and allows keeping up this information. CSP can't be completely trusted by cloud clients for reasonable to expect. The data can encode the information to cloud for information security and client protection.

Their aim is to maximize space savings and minimize redundant data in cloud storage. The cross-user deduplication is widely adopted in a technique. The duplicate data store (either files or blocks) only once in cloud. The user wants to upload a file (block) in the cloud which is already stored. The provider is adding the user to the owner list of that file (block). It achieved cost savings and high space in the Deduplication. The any Big Data storage providers are adopting it. The deduplication can up to 68% in standard file systems and reduce storage needs by up to 90-95% for backup applications. The savings can be passed back directly or indirectly to the cloud users for significant to the economics of business cloud. The practical issue is to manage encrypted data storage with deduplication in an efficient ways. The current industrial solutions for deduplication cannot handle encrypted data. Existing solutions suffer for deduplication on brute-force attacks. The data revocation and access control at

the same time they cannot support. The reliability, security and privacy cannot ensure in most existing solutions.

The data holders allow to managing deduplication is too hard due to a number of reasons.

- First, they could cause storage delay because of the data holders may not be always online or available for management.
- Second, in the deduplication process becomes too complicated for communications and computations to involve data holders.
- Third, how to issue data access rights or deduplication keys to a user in some situations a data holder may have no idea. When it does not know other data holders due to data super-distribution.
- Forth, the process of discovering duplicated data it may intrude the privacy of data holders.

Therefore, the data storage deduplication on the CSP cannot cooperate with data holders in many situations. In this paper, propose a scheme based on Proxy Re-Encryption (PRE) and data ownership challenge to manage deduplication with encrypted data storage. Our aim is to solve the issue of deduplication in the situation. It is difficult to get involved the data holder is not available. The data deduplication is not influenced by the size of data. It is applicable for big data. Specifically, the contributions of this paper can be summarized as below:

- Their aim is to preserve the privacy of data holders and to save cloud storage to manage encrypted data storage with deduplication. The data holder is offline in our scheme can support data sharing deduplication even. The data holder's doses not intrude the privacy.
- The effective approaches to verify check duplicate storage and data ownership with secure challenge and big data support.
- The cloud data deduplication integrates with data access control, thus recounting data deduplication and encryption.
- The security and assess the performance of the proposed scheme with the analysis.

2. LITERATURE SURVEY

An Efficient Privacy-Preserving Ranked Keyword Search Method is proposed by Chen et. al. [1] Proposed system discussed hierarchical approach clusters the documents based on the minimum relevance threshold, and then partitions the resulting clusters into sub-clusters until the constraint on the maximum size of cluster is reached. A linear computational complexity against an exponential size increase of document collection this approach can reach. In order to verify the authenticity of search results, a structure called minimum hash sub-tree is designed in this approach. System also investigated ciphertext search in the scenario of cloud storage. System explore the problem of maintaining the semantic relationship between different plain documents over the related encrypted documents and give the design method to enhance the performance of the semantic search.

Chunhua et al. [2] proposed Analysis and Improvement of Privacy-Preserving Frequent Item Protocol for Accountable Computation. System proposed a protocol of finding frequent item in accountable computing (AC) framework which enables two parties to conduct collaborative computation on their transactional databases to find out the common frequent items without disclosing their private data to the other party. Their scheme was proposed in a secure two-party computation model against malicious adversaries. System also analyzes the implementation details of AC-framework and identifies some security weaknesses in their scheme. Furthermore, system clarifies the security requirements for the AC-framework and presents an augmented solution to enhance security. System also analyzes the search efficiency and security under two popular threat models.

Huang and Rongxing et al. [3] proposed EFPA: Efficient and Flexible Privacy-Preserving Mining of Association Rule in Cloud. System proposed an efficient and flexible protocol, called EFPA, for privacy-preserving association rule mining in cloud. With the protocol, plenty of participants can provide their data and mine the association rules in cloud together without privacy leakage. Detailed security analysis shows that the proposed EFPA protocol can achieve privacy-preserving mining of association rules in cloud. It also present an efficient and flexible privacy preserving association rule mining protocol, called EFPA. Unlike most existing works, EFPA can support distributed data providers to collaboratively achieve association rule mining without exposing any privacy of data providers or mining results, i.e., the providers' data and mining results cannot be revealed by cloud.

K.Samanthula et al. [4] k-Nearest Neighbor Classification on Semantically Secure Encrypted Relational. The encrypted data proposed k-NN protocol protects the confidentiality. The data, user's input query, and data access patterns. In this work is the first to develop a secure k-NN classifier over encrypted data under the model. To proposed to classification of encrypted data. In encrypted data in the cloud secure k-NN classifier. This system can protect the confidentiality of data.

Lichun Li et al. [5]. In this work Privacy-Preserving-outsourced Association of Mining on vertically partitioned in databases. This system focus on cloud-aided frequent itemset mining solution, which is used to build an association rule mining solution. Here outsourced databases that allow multiple data owners to efficiently share their data securely without compromising on data privacy and leak less information about the raw data than most existing solutions. In comparison to the only known solution achieving a similar privacy level as this proposed solutions, the performance of this proposed solutions is three to five orders of magnitude higher. Based on this experiment findings using different parameters and data sets, system demonstrate that the run time in each of this solution is only one order higher than that in the best non-privacy-preserving data mining algorithms. Since both data and computing work are outsourced to the cloud servers, the resource consumption at the data owner end is very low. It also privacy-preserving outsourced frequent itemset mining solution for vertically partitioned databases. This allows the data owners to outsource mining task on their joint data in a privacy-preserving manner. Based on this solution, system built a privacy preserving outsourced association rule partitioned databases. Compared with most existing solutions, this solutions leak less information about the data owners' raw data.

In this paper a structure for rules from dealings made up of specific items. The data has been randomized to ensure that individual dealings. To restore data ensure utilizing an uncomplicated "uniform" randomization. To assess the qualities of security describe and ruptures of randomization for suppliers. It is limiting the breaks more proficient than

steady randomization. To permit us to restore set encourages from randomized datasets. Show these equations into investigation techniques. In conclusion for implement so as to results on genuine datasets. In this paper issues address of privacy preserving data mining. In which consider two parties owning confidential databases. Run the data mining algorithm. There work to propose need to both protect privileged information. They enable its use for research.

Cloud storage service providers such as Dropbox et al. [6], Google Drive et al. [7], Mozy et al. [8], and others perform only storing one copy of each file uploaded performs data duplication. If users encrypt their data to storage savings by deduplication are failed because of the encrypted data are saved as different contents by applying different encryption keys. The solutions fail in encrypted data deduplication on existing industries. DeDu et al.[9] is an efficient deduplication system, but it cannot handle encrypted data. Jin Li and Yan et al. [10] New de-duplication constructions supporting authorized duplicate check in hybrid cloud architecture in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Proposed system includes proof of data owner so it will help to implement better security issues in cloud computing. File-level de-duplication: This technique is also called as single-instance storage. In this de-duplication compares a file that has to checking all its attributes in the index. In this index updated and stored the file is unique, if the file is not unique then only a pointer to the existing file that is stored references. Only the data of file is unique then saved in the result and relevant copies [11].

3. PROPOSED SYSYTEM ARCHITECTURE

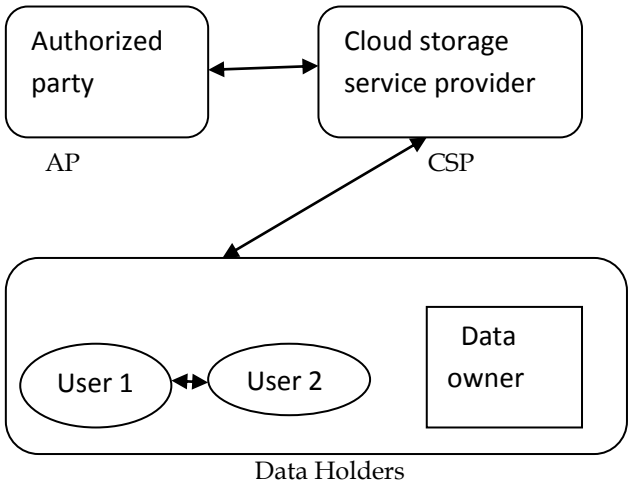

Fig 1: System architecture

The figure shows in Fig. 1, the system contains three types of entities:

- **CSP:** The CSP is allowing to data owner for storage services of data. It cannot be fully trusted. Therefore it is curious about the contents of stored data. It should perform honestly on data storage in order to gain profit.
- **Data Holder:** The data holder can uploads and saves their data and files in the CSP. In this system is possible to number of data holders could save their files in encrypted raw data in the CSP. The file is regarded as data owner by the data holder that produces or creates the file. The data holder is in normal form than the higher priority of owner.
- **AP:** an authorized party (AP) that is fully trusted by the data holders. The data holders to verify data ownership and handle data deduplication. It does not collude with CSP. In this case, CSP should not know the plain user data in its storage. AP cannot know the data stored in CSP.

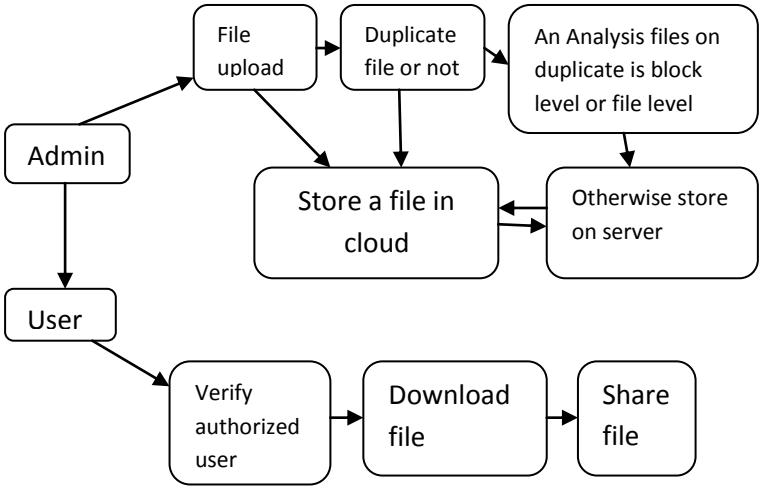

Fig 2: Block diagram of proposed system

The above Fig 2 shows the block diagram of the proposed system architecture. The user want to upload the file in the server then the system shows to upload their file is successfully otherwise the data is same them shows the data duplication in the server. The content of data is not same then store the data or file in cloud. The new user wants to upload their files in this cloud they want to register in the server and get the user ID and password. Login to this ID and password and upload the data for storage. The authorized user is after login view to access the files. Only authorized user can access the files and this user can download the files. The user can upload files, download files, share files and revoke files. The user share files with encrypted key to the other user. The authorized user can download this share files with the encrypted key. In this proposed scheme using VCS algorithm shows the duplication of big data. Analysis files on duplication in data server.

The secure search of system in the cloud the typical participants involve the cloud server the data user and the data owner, as shown in Fig. 3 The data owner outsources the encrypted dataset and the corresponding secure indexes to the cloud server, where data can be encrypted using any secure encryption technique, some particular search-enabled encryption techniques while the secure index is generated.

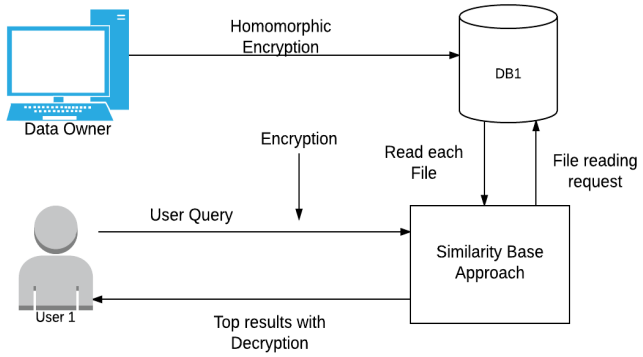

Fig 3: Proposed system architecture

4. SCHEMES

- **Encrypted Data Upload:** The data holder want to store the data then it encrypts its data using a randomly selected symmetric key DEK to ensure the security and privacy of data. To stores the encrypted data at CSP with used for data duplication check. The data holder encrypts DEK with pkAP and passes the encrypted key to CSP.
- **Data Deduplication:** Data holder tries to store the same data that has been stored already at CSP when Data duplication occurs at the time. The comparison is checked by CSP through VCS algorithm. If the comparison of data is same, CSP contacts AP for deduplication and the data holder's PRE public key. The AP challenge is checks the eligibility of the data holder data ownership and then issues a re-encryption key. It can convert the encrypted DEK to a form that can only be decrypted by the eligible data holder.
- **Data Deletion:** When the data holder deletes data from CSP. It can firstly manage the records of duplicated data holders it removing the duplication record of this user. If the rest records the CSP will not delete the stored encrypted data. The block data access from the holder that requests data deletion.
- **Data Owner Management:** The real data owner uploads the data later than the data holder. The CSP can manage to save the data encrypted by the data owner. In the cloud with the owner generated DEK and after, AP supports re-encryption at CSP for eligible data holders.
- **Encrypted Data Update:** The data owner wants to update data new encrypted raw data is provided and replace old storage for the reason of security. CSP issues the new re-encrypted to all data holders with the support of AP.

5. ALGORITHM

The algorithms used in deduplication technique and the implementation process details are described in this chapter.

5.1.SHA -256

- SHA-256 a cryptographic hash function with digest length of 256 bits.
- SHA-256 operates in the manner of MD4, MD5, and SHA-1: The message to be hashed is first

(1) The result is a multiple of 512 bits long padded its length.

(2) Message blocks $M^{(1)};M^{(2)}\ldots. M^{(N)}$ parsed into 512-bit.

The message blocks are processed one at a time: Beginning with a fixed initial hash value $H^{(0)}$, sequentially compute

$H^{(i)} = H^{(i-1)} + C_M^{(i)}(H^{(i-1)});$

The SHA-256 compression function in C and + means word-wise mod 232 addition. $H^{(N)}$ is the hash of M.

- **Vector base cosine similarity (VCS)**

Input Query Q, Threshold t

Output duplicate if returns 1 else unique

Here they have to find similarity of two vectors: $\vec{a} = (a_1, a_2, a_3, \ldots)$ and $\vec{b} = (b_1, b_2, b_3, \ldots)$, where a_n and b_n are the components of the vector (features of the document, or values for each word of the comment) and the n is the dimension of the vectors:

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

------- (1)

Step 1: Read each row R from dataset D

Step 2: for each (Column c from R)

Step 3: score= Formula1(R,Q)

If (score > t)

Break;

Early stop;

Else step 2 continue

End for

Output: duplicate if returns 1 else unique

6. RESULT AND DISCUSSION

For the system performance evaluation, calculate the matrices for accuracy. The system is implemented on java 3-tier MVC architecture framework with INTEL 3.0 GHz i5 processor and 8 GB RAM with public cloud Amazon EC2 consol. System also evaluated the computation. The graph shows the system performance with different experiments that has been classified in graphs.

The below graph Fig.4 shows the file uploading time i.e. the time required for secrete key generation and upload the file. In this graph they can consider the file uploading time is equals to file uploading ending time minus file uploading starting time. Here time required to upload five files is considered and generated graph showing result for this. Here x axis represents file length (kilo byte) and y axis represents file uploading time. If the file contains more data then time required to upload a file is more. In below figure nature of the graph completely depends on the file length and system i.e. graph changes with the varying size of the file. The file uploading time is depends on the system also the system is proper then file upload in few milliseconds than the system slower.

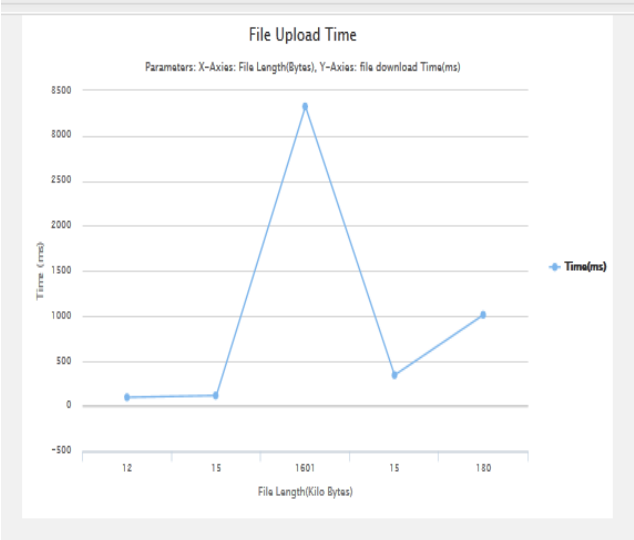

Fig 4: File uploading time

Second graph Fig 5 shows graph for file downloading time i.e. time required by the user for verification and file download time. Here also time required to download files is considered and generated a graph showing result for this. Here x axis represents file length (kilo bytes) and y axis represents file downloading time. In below figure nature of graph depends on the key verification time required for that file i.e. if file requires more time for key verification then more time is required to download that file. In the figure shown below nature of graph varies with the key verification time of files.

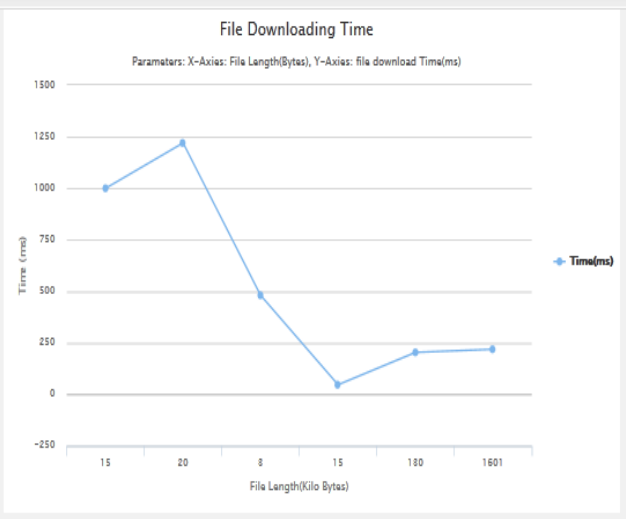

Fig 5: File downloading time

Next graph Fig 6 shows comparison graph for existing system and proposed system. Here X- axis represents existing system and proposed system as parameters and Y- axis represents values of deduplication to which access is restricted to the files. Suppose any user finds out that one of the authorized user in the server with whom he has shared one file. The uploaded files access is the data of the file is same then the file shows the deduplication of the data. In the below figure first bar is lower than the second bar because in first bar access of all the previously uploaded files is deduplication of the data. So proposed system is better because while using the proposed system the time for deduplication. Proposed system gives less time for duplication than the existing system.

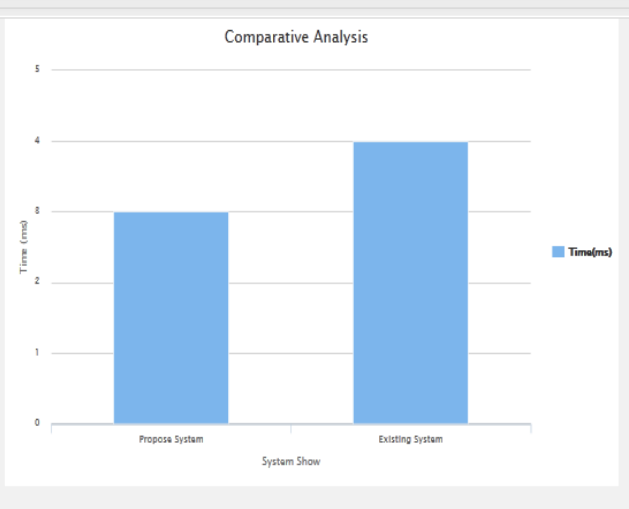

Fig 6: Comparison graph for proposed system and existing system

Analysis graph in Fig 7 shows comparison graph for time complexity with existing system and proposed system. Here X-axis represents existing system and proposed system as parameters and Y- axis represents values of number of documents to which access is restricted to the files. In the first experiment system first compare the time complexity with different existing algorithms. The below graph show the how much time required in milliseconds calculate the relevancy with no. of documents.

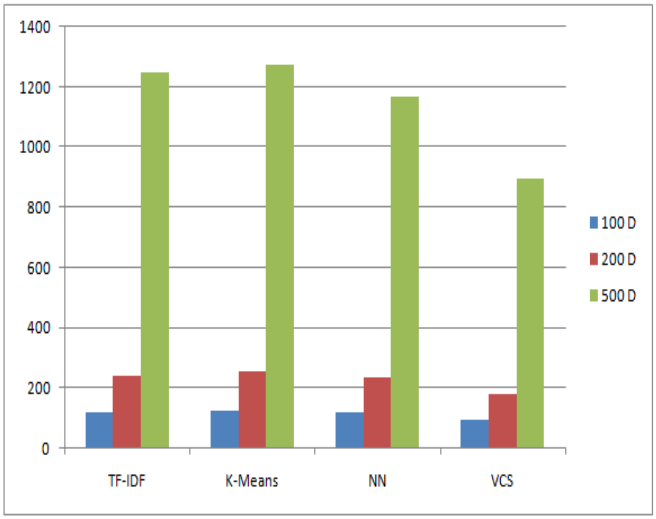

Fig 7: Comparison graph for time complexity with existing system and proposed system

7. CONCLUSION

Discuss about to managing encrypted data with deduplication is for achieving a successful cloud storage service, for big data storage. In this paper, proposed a scheme to manage the encrypted big data in cloud with deduplication based on PRE and ownership challenge. This scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Only authorized data holders can use the symmetric keys used for data decryption by encrypted data can be securely accessed. The performance analysis and graphs showed that there scheme is secure under the described security model for big data deduplication. The results of the computer simulations further showed in the variation of graph. Future work is to include optimizing there design and implementation for big data storage in cloud. To ensure that CSP behaves as expected in deduplication management Studying verifiable computation. The part of cloud computing has brought many researchers from different fields; yet, much effort remains to reach use and the broad acceptance of cloud computing technology. In the future scope propose a big data deduplication. For the future environment system can focus on personalize search on user feedback sessions as well as recommendation base on user point of interest with database security is the interesting part of system.

8. REFERENCES

[1] Chi Chen at. Al. proposed An Efficient Privacy-Preserving Ranked Keyword Search Method IEEE 2016.

[2] Chunhua Su at. al. proposed Analysis and Improvement of Privacy-Preserving Frequent Item Protocol for Accountable Computation Framework IEEE 2012.

[3] Cheng Huang and Rongxing Lu proposed EFPA: Efficient and Flexible Privacy-Preserving Mining of Association Rule in Cloud in IEEE 2015.

[4] Bharath K. Samanthula at. Al. k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data MAY 2015.

[5] Lichun Li at. al. Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases in AUGUST 2016.

[6] M. Bellare, S. Keelveedhi, and T. Ristenpart, "DupLESS: Serveraided encryption for deduplicated storage," in Proc. 22nd USENIXConf. Secur., 2013, pp. 179–194.

[7] Dropbox, A file-storage and sharing service. (2016). [Online]. Available: http://www.dropbox.com

[8] Google Drive. (2016). [Online]. Available: http://drive.google.com

[9] Mozy, Mozy: A File-storage and Sharing Service. (2016). [Online]. Available: http://mozy.com/

[10] Z. Yan, W. X. Ding, and H. Q. Zhu, "A scheme to manage encrypted data storage with deduplication in cloud," in Proc. ICA3PP2015, Zhangjiajie, China, Nov. 2015, pp. 547–561.

[11] Z. Sun, J. Shen, and J. M. Yong, "DeDu: Building a deduplication storage system over cloud computing," in Proc. IEEE Int. Conf. Comput. Supported Cooperative WorkDes.,2011,pp.348.355,doi:10.1109/CSCWD.2011.5960097.