

## Secure De-Duplication over Cloud Using Attribute Based Access Policies

**Jeet Mirani<sup>1</sup>, Sanyogita Naik<sup>2</sup>, Ajit Wale<sup>3</sup>, Prof. Archana Patil<sup>4</sup>**

[jeetmirani24@gmail.com](mailto:jeetmirani24@gmail.com), [naiksan22@gmail.com](mailto:naiksan22@gmail.com), [ajitwalepatil@gmail.com](mailto:ajitwalepatil@gmail.com), [patilarchana63@gmail.com](mailto:patilarchana63@gmail.com)

*<sup>1,2,3,4</sup> Pune University, Keystone School of Engineering, Near Handewadi Chowk, Urali Devachi, Shewalewadi, Pune, Maharashtra 412308*

**Abstract:** As the cloud computing services are rapidly being used in recent days for storage and other purposes, cloud deduplication is also such service which has to be focused on. Data deduplication is a technique for reducing the amount of storage space an organization needs to save its data. In most institutes, the storage systems contain duplicate copies of many pieces of data. Deduplication eliminates these extra copies by saving just one copy of the data and replacing the other copies with pointers that lead back to the original copy. Most organizations and companies frequently use deduplication in backup and disaster recovery applications, but it can be used to free up space in primary storage as well. To avoid this duplication of data and to maintain the confidentiality in the cloud Hybrid cloud concept is used. To secure the confidentiality of sensitive data while supporting deduplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, this paper makes the first attempt to formally address the problem of authorized data deduplication using content level duplication checking.

**Keywords:** *Authorization, data security, privilege, deduplication, credential, hybrid cloud.*

### Introduction

Hybrid cloud is a composition of two or more clouds (private, community or public) that remain unique entities but are bound together, offering the benefits of multiple deployment models. Use of Cloud computing has rapid and wider scope now a days. Cloud storage services provide huge amount of virtual Infrastructure environment abstracting the underlying platform and other technical details from the user. Cloud users need to pay as per the resource usage allocated to him/her. Distributed computing has massive variety of degree in data sharing in current period. Distributed computing is give accurate measure of virtual environment concealing the stage and working

framework of the client. Client use the assets for exchanging information. It may, client need to pay by the process of utilization of assets of cloud. Cloud admin distributors are putting forth cloud administrations with ease furthermore with large dependability. Client can transfer the vast sum data on cloud and exchanged information to a large number of clients. Cloud suppliers are offer diverse administrations, for e.g., framework as an administration, stage as an administration, and so forth. Client not has to buy the assets. As the data is get exchanged by the client might be it is basic notification to deal with this regularly expanding information on the

cloud. To make well information administration in the distributed computing. We use duplication technique, which is the best technique in cloud. This technique is turning out to be more moderation for information DE duplication. This system is send the information over the system required little measure of information. This technique has application in information administration and organizing. Information duplication is the procedure of decreasing copy file Also it is the best pressure system for the information DE duplication. This system has application in information administration and organizing. Rather than keeping excess duplicate file of the same information DE duplication just keep unique duplicate and give just references of the first duplicate to the repetitive information. The process of checking the duplication process is two; one is document level duplication check and second is piece content level duplication check. In the document level duplication technique check is expel the same name record from the capacity and square level DE duplication are evacuated the copy pieces. DE duplication techniques need of the some security system. In the conventional system client need to encode his own particular information.

To manage a security from the unapproved information DE duplication focalized information DE duplication is proposed to uphold the data privacy while checking the information duplication. The cloud giving various administrations as attended in the above figure, for example, stage, administrations, base as an administration, and database as an administration.

In this part we are utilizing as a part of distributed storage as an administration.

We are utilizing client accreditations to check the confirmation of the client. In that cases cloud is available two sort of cloud such private cloud and open cloud. In private cloud store the client accreditation and in the open cloud client information present out. In the figure 2. Cloud take focal points of both open cloud and private cloud. Open cloud and private cloud are available in the half and half cloud structural engineering. When any client forward solicitation to people in general cloud to get to the data he have to present his data to the private cloud then private cloud will give a record token and client can get the notifications to the document lives on the general population cloud.

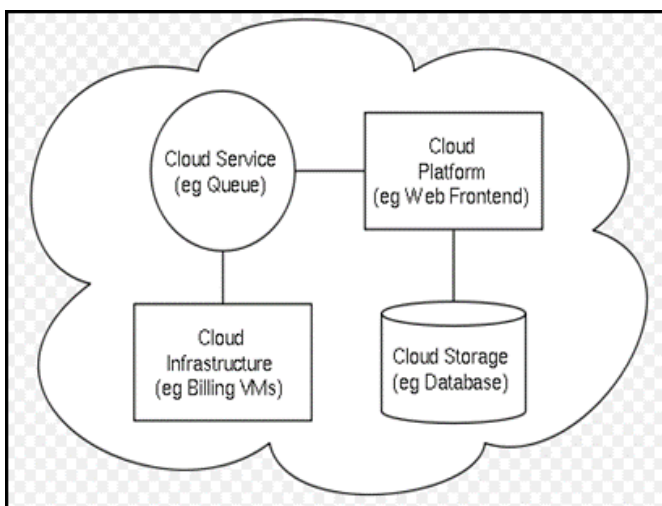
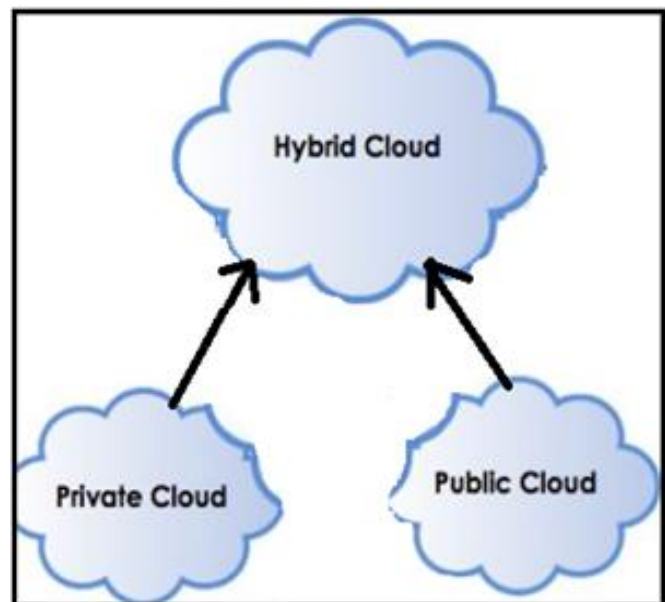


Figure1: Cloud architecture and services.



*Figure2: Hybrid Cloud Architecture.*

We have utilized a half and half cloud construction modeling as a part of proposed. We have to need to mind the file name in record information duplication and information DE duplication is checked at the square level. On the other hand, client needs to recover his information or download the information record he have to download both of the document from the cloud server this will prompts perform the operation on the same record this abuses the security of the distributed storage.

### **Literature Survey:**

There are so many researchers have been completed to secure duplication check of data on cloud. The cloud storage and data DE duplication are two methods present in existing system. First method of the data Deduplication is perform as post processing method [1] In this which data is first store on the storage device and then duplication check is applied on the data. The use of this method is there are no need to wait for calculating the hash function and the speed of storage not get downgrade. The main drawback with this system is that if storage capacity of the device is very low then the file storage may get full. Some issue of this the post processing method is not useful at all because it checks the file after storing it on the cloud server. Second method of duplication check is the inline duplication check. It is checked when new entries are to be added to the database the duplication of the file. It will checks for the block level duplication of the file before adding new data or the new entry to the database. This method have some drawback such as each time need to calculate the hash function which may lead to slower throughput of the storage device. But the few of the vendors have proof that data duplication check have same output in the inline and post processing method. Another method of duplication check is source duplication check in which

file duplicate contents are checks for duplication before storing it on the cloud server. Third method of Deduplication is source data Deduplication in which data duplication is completed at the side of the source. The file duplication is check before it get uploaded on the cloud server. The duplication is checked at the target level in which file get scanned periodically and hash get generated for the software can check for the hash value if both value get new matched with the existing hash value then the new file not get uploaded on the server of the cloud, only link to that data is to be provide to the file user. If new file is to be added to the cloud server and it get match the hash function of the old file then it only remove the new file and provide hard link to the old file resides on the cloud server.

Chunk level duplication checker is another method of the duplication calculation. In this part for each chunk identification is get assigned generated by the software. For the pre-processing file checking we have to make some assumption that identification is same then data is also same but this is not right in all the cases due to the pigeonhole principal. It will produce wrong result that if for two blocks of the data same identification number is get generated it simply remove the one block of the data.

### **Proposed System:**

In the proposed system we are doing duplication check in authenticated way. For the file duplication check proof of ownership is also set at the time of file upload the proof is added with the files this proof will decide the access privilege to the file. It is decide who can perform duplication check of the file. User is needed to submit his/her file and proof of ownership of the file before sending the request to for the duplicate check Request to the cloud. When there is file on the cloud and also privileges of the user only that time to approved the duplicate check request.

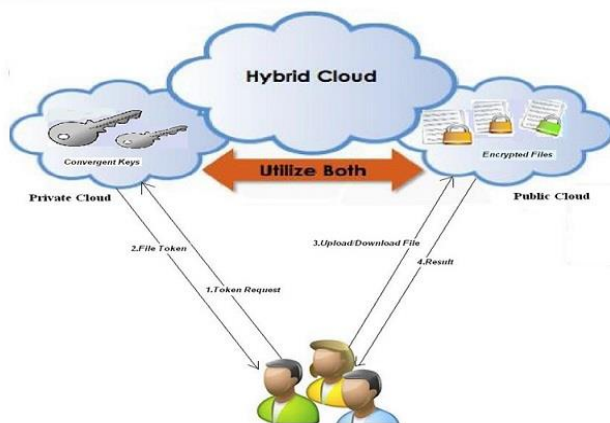


Figure3: Overview of the system.

Above fig.3 shows the proposed system architecture which comprises of public cloud, private cloud and user.

Proposed system architecture contains only one public cloud and one is private cloud. All data of user is contains in public cloud such as files. And private cloud consists of user credentials. User for every transaction with the public cloud need to take token from the private cloud. If the user's credentials stored at the public cloud and private cloud are get matched then user can have access for the duplicate check. Following operations are need to be done in the authenticate duplicate check.

#### A. Encryption of File:

We are using secrete key resides at the private cloud to encrypt the user data and this key is used to convert plain text to cipher text and again for the decryption of the user data. To encrypt and decrypt we have used three basic functions as follows:

- 1.) Key GenSE: It is generate the secrete file by using security parameter. In this k is the key generation algorithm.
- 2.) EncSE (k, M): In this we have generated a cipher text using formulae M is the text message and k is the secrete key.

3.) DecSE (k, C): In this we have to generate plain text using C is the cipher text and k is the encryption key.

#### B. Confidential Encryption of data:

This ensures the data confidentiality in the duplication. User derives a convergent key from each original data and encrypt the data copy with the generated convergent key. User also add the tag for the data so that the tag will helps to find the duplicate data. By using converged key generation algorithm to encrypt the user data. This will ensures the security, authority and ownership of the data.

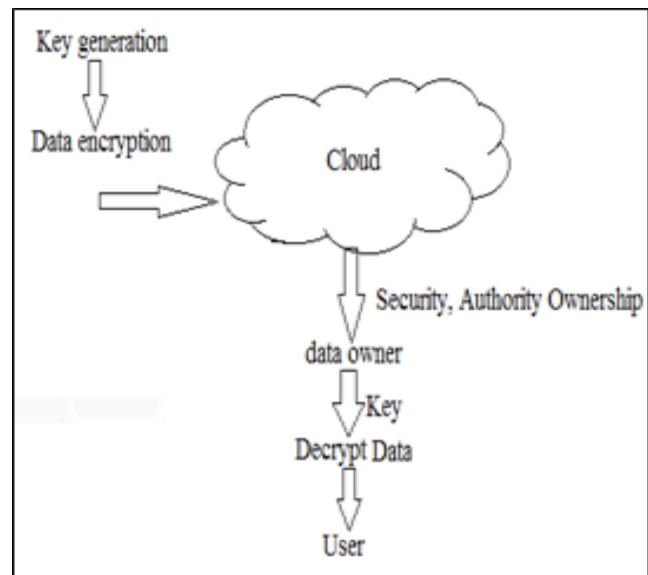


Figure4: Confidential data encryption.

#### C. Proof of Data:

When file upload and download user need to provide proof of the data. User need to submit his/her convergent key which was generated at the time of file upload. To generate the hash value of the data we have used MD5 message digest version 5 algorithm to generate the hash value of the user data. If there is any small change in data occur the hash value of that data get changed.

**Mathematical model:**

A. Set theory:

- $S = \{R, T, P, H, D\}$
- R=Registration of the user with specifying size on cloud.
- T=token generated and forwarded to user through mail for activation.
- P-User Privileges
- H- Hash function calculation.
- D= Matching contents of user uploaded data with existing database
- $R = \{r0, r1\}$
- Where ,
- r0-Provide information to the registration authority.
- r1-Registration authority validate the information.
- r2-user get cloud id and user id.
- $r0 \rightarrow t1$
- $T = \{t1, t2\}$
- t1-token gives to user through mail.
- t2- get privilege to user.
- $D = \{d0, d1, d2, d3\}$
- Where,
- d0- get the data file name and key
- d1- Generate hash function and encrypt file.
- d2-check matching content by entering upload button.
- $t2 \rightarrow d3$
- d3-download/update/upload file by providing token/key or any other details.

**Graphical Analysis:**

**File Size:**

To evaluate the effect of file size to the time spent on different steps, I upload 100 unique files of particular file size and record the time break down. Using the unique files enables us to evaluate the worst-case scenario where I have to upload all file data. The average time of the steps from test sets of different file size are plotted in Figure 5. The time spent on

downloading, encryption, upload increases linearly with the file size, since these operations involve the actual file data and incur file I/O with the whole file. In contrast, other steps such as duplicate check and token generation only use the file metadata for computation and therefore the time spent remains constant. With the file size increasing from 10MB to 120MB.

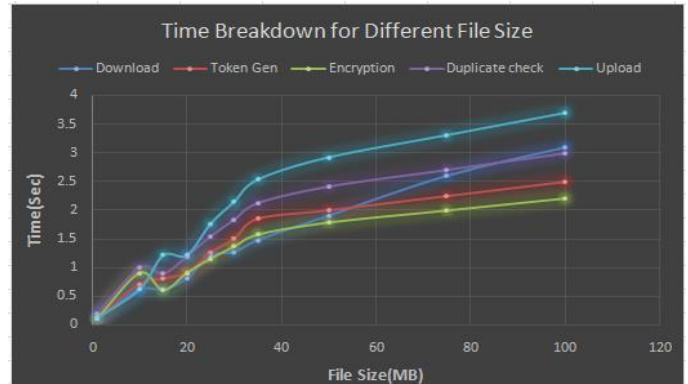


Figure5: Time Breakdown for different File size.

**Number of Stored Files:**

To evaluate the effect of number of stored files in the system, I upload different number of unique size files and record the breakdown for every file upload. From Figure 5, every step remains constant along the time. Token checking is done with a hash table and a linear search would be carried out in case of collision.

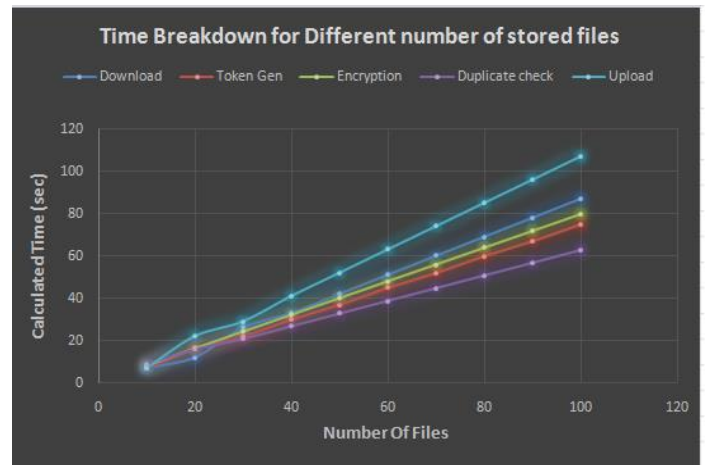




Figure6: Time Breakdown for different number of stored file.

**Results:**

This system should prevent user from uploading duplicate data on cloud. Data stored on cloud must be in secure encrypted format. Malicious user not able to upload or download data on cloud. The user who has proof of ownership only that user can modify data.

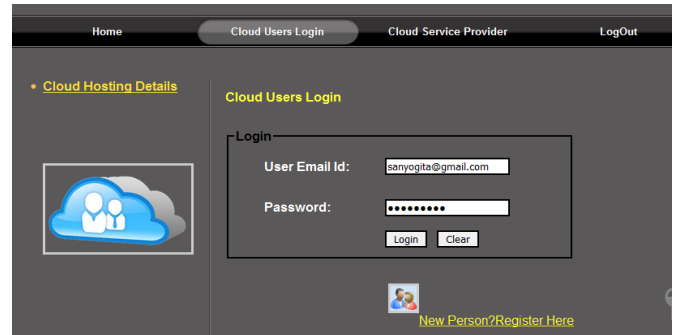


Figure7: User Login Page.

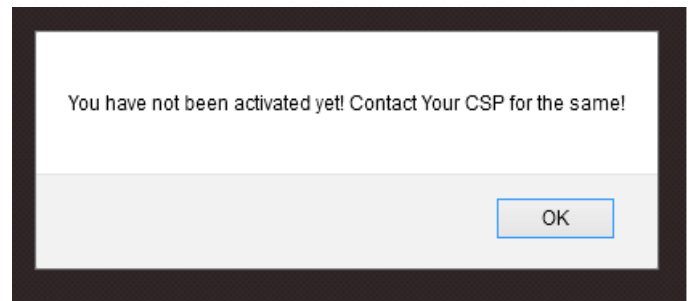


Figure8: New user before activation message.

**Result Snapshots**

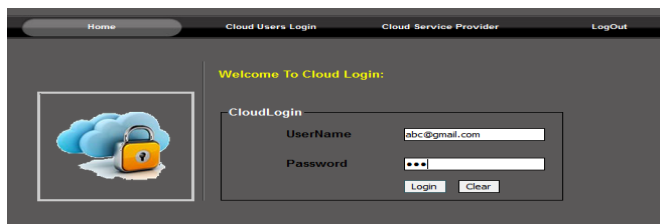


Figure6: CSP Login Page.

User Id	UserName	Email Id	Contact No	Date	Status	Activation
2	sanyogita	sanyogita@gmail.com	7894561230	04/10/2017	A	Deactivate
3	sampleuser	sample@gmail.com	9876543210	04/10/2017	A	Deactivate
4	xyz	xyz@gmail.com	8974561231	04/10/2017	A	Deactivate
5	mno	mno@gmail.com	3216549870	04/10/2017	A	Deactivate

Figure9: New user activated by CSP Admin.

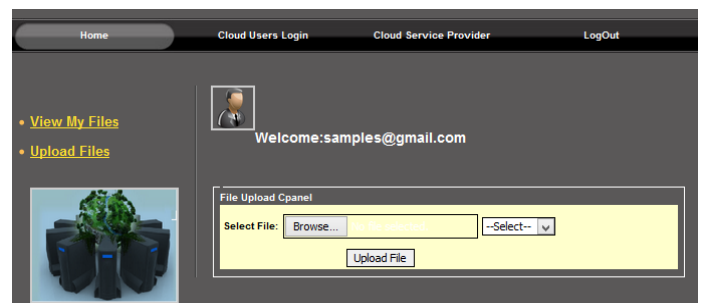


Figure10: File Upload Page.

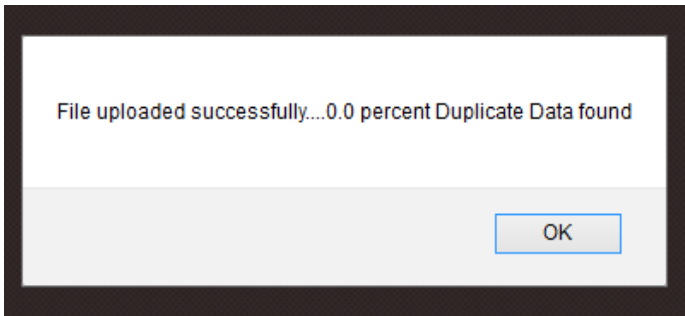


Figure11: File Upload Message.

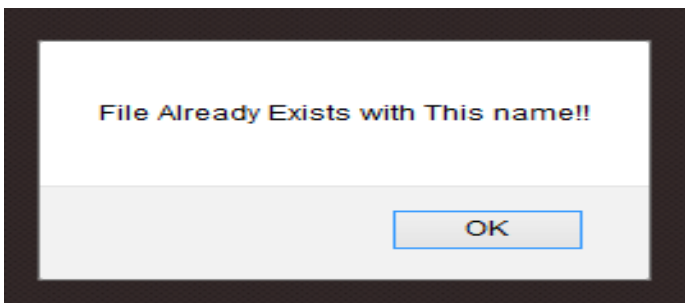


Figure12: If same file name exist then file already exist Message.

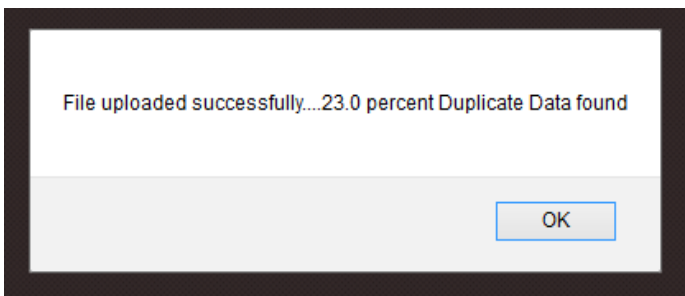


Figure13: Our system checks for duplicate data in file and according to file data system generate chunks.

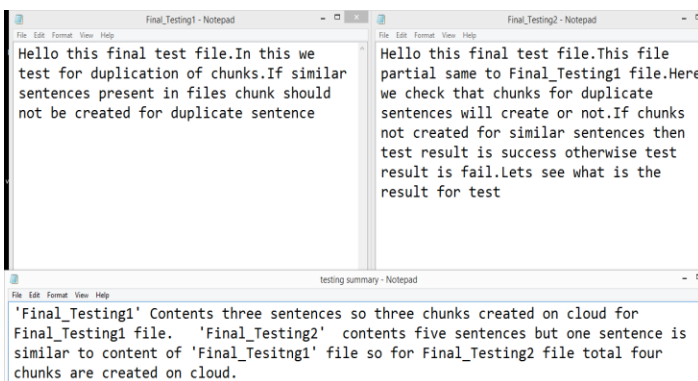
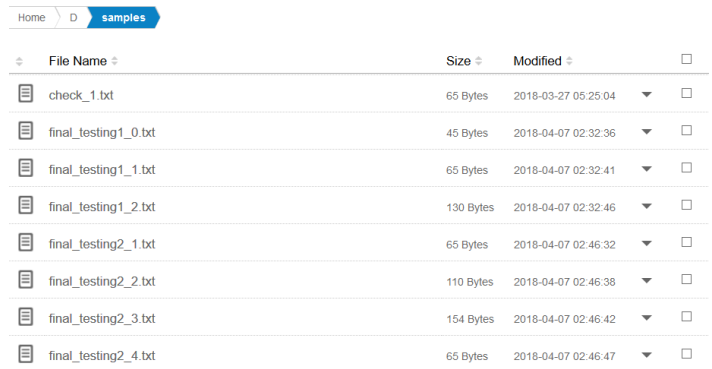
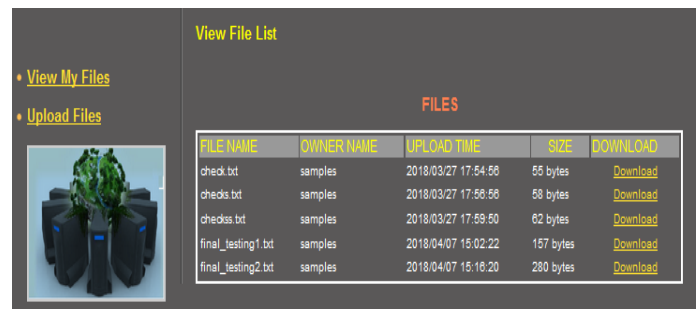


Figure14: Chunk Testing Files.



File Name	Size	Modified
check_1.txt	65 Bytes	2018-03-27 05:25:04
final_testing1_0.txt	45 Bytes	2018-04-07 02:32:36
final_testing1_1.txt	65 Bytes	2018-04-07 02:32:41
final_testing1_2.txt	130 Bytes	2018-04-07 02:32:46
final_testing2_1.txt	65 Bytes	2018-04-07 02:46:32
final_testing2_2.txt	110 Bytes	2018-04-07 02:46:38
final_testing2_3.txt	154 Bytes	2018-04-07 02:46:42
final_testing2_4.txt	65 Bytes	2018-04-07 02:46:47

Figure15: Chunks for respective file save to cloud in encrypted format.



FILE NAME	OWNER NAME	UPLOAD TIME	SIZE	DOWNLOAD
check.txt	samples	2018/03/27 17:54:58	55 bytes	<a href="#">Download</a>
checks.txt	samples	2018/03/27 17:58:58	58 bytes	<a href="#">Download</a>
checkss.txt	samples	2018/03/27 17:59:50	62 bytes	<a href="#">Download</a>
final_testing1.txt	samples	2018/04/07 15:02:22	157 bytes	<a href="#">Download</a>
final_testing2.txt	samples	2018/04/07 15:16:20	280 bytes	<a href="#">Download</a>

Figure16: User can able to view or download files list in View File List section.

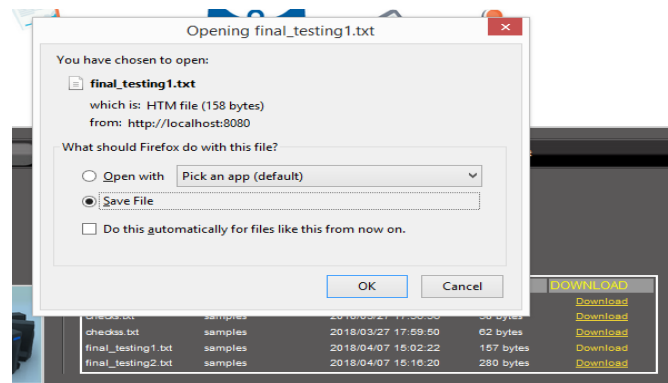


Figure17: File Download By user.

**Test Cases:**

Sr.No	Module	Input	Description	Output
1.	User	User Details	User enter his details and make registration. Then he provide the user name and password for login.	User registration And login successfully
2.	Cloud	Login with Cloud is necessary	Here, Must login with cloud to activate and uploading the data in the cloud.	Login successful
3.	Deduplication	Upload your text data over the cloud	Here, text data uploaded by user is dividing into chunks.	Chunking successful.
4.	Checking Deduplication	Adding text value in the file	Check the chunk value and reference value of duplication data.	Deduplication Successful and provide reference value.

### Conclusions:

Here we provided reason that our proposed framework information DE duplication of record is done approves way and safely. In this we have additionally proposed new duplication check system which produce the token for the private document. The information client needs to present the benefit alongside the united key as a proof of possession. We have settled more basic piece of the cloud information stockpiling which is just endured by diverse systems. Proposed routines guarantee the information duplication safely.

### Acknowledgments:

I would like to take this opportunity to express my profound gratitude and deep regard to my guide Prof. Archana Patil for her exemplary guidance, valuable feedback and constant encouragement throughout the duration of the project. Her valuable suggestions were of huge help throughout my project work. Her perceptive criticism kept me working to make this project in a much better way. Working under her was an extremely knowledgeable experience for me.

### References and footnotes:

- [1]. J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.
- [2]. P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.
- [3]. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.
- [4]. J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.
- [5]. C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013.
- [6]. W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446. ACM, 2012.