Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 04, April -2018

IMPLEMENTATION OF PROTECTION OF BIG DATA USING THIRD PARTY AUDITING AND DATA DE-DUPLICATION

Vidya Jagtap[1], Ravi Mohite[2], Chaitali Shinde[3], Khushal Khule[4]

[1] HOD, Department of Computer Engineering, G.H.RAISONI Engineering College, Chas, Ahmednagar, India
[2] U.G. Student, Department of Computer Engineering, G.H.RAISONI Engineering College, Chas, Ahmednagar, India
[3] U.G. Student, Department of Computer Engineering, G.H.RAISONI Engineering College, Chas, Ahmednagar, India
[4] U.G. Student, Department of Computer Engineering, G.H.RAISONI Engineering College, Chas, Ahmednagar, India

Abstract — *In recent years, huge information became a hot analysis topic. The increasing quantity of huge knowledge additionally will increase the prospect of cracking the privacy of people. Since huge private information need high procedure power and massive storage, distributed systems are used. As multiple parties are concerned in these systems, the danger of privacy contravention is accumulated. There are different types of privacy-preserving mechanisms developed for privacy protection at totally different stages (e.g., knowledge generation, knowledge storage, and knowledge processing) of a giant knowledge life cycle. The main goal of this paper is to provide a comprehensive summary of the privacy preservation mechanisms in huge knowledge and gift the challenges for existing mechanisms. Above all, during this paper, we tend to explain the infrastructure of massive knowledge and therefore the onward privacy-preserving mechanisms in every stage of the massive knowledge life cycle. What is more, we tend to discuss the challenges and future analysis directions associated with privacy preservation in huge knowledge.*

Keywords- *Cloud Computing; Big Data; protection of data; Data Security; De-duplication of Data*

I. INTRODUCTION

Cloud service providers manage an enterprise-class infrastructure that offers a scalable, secure and reliable environment for users, at a much lower marginal cost due to the sharing nature of resources. It is routine for users to use cloud storage services to share data with others in a team, as data sharing becomes a standard feature in most cloud storage offerings, including Drop box and Google Docs. Many mechanisms have been proposed to allow not only a data owner itself but also a public verifier to efficiently perform integrity checking without downloading the entire data from the cloud, which is referred to as public auditing.

With evolution of computers the life of people became more and more easily. They were able to keep their data on their devices, and started finding ways to make them accessible to others, for example say by using floppy, writable disks, which was followed by portable hard-disk, all these where expensive in their own way during their time. The data was very much private on personal devices like PC, laptops, mobile phones etc., therefore sharing data with others was considered to be expensive. As the world of computing got more advanced the ways for sharing data started becoming cheaper and cheaper. In recent years a new term has evolved call "Cloud" which is provided by different provides, and which is nothing but facility or service of different resources or components like hardware, platform, storage's, software etc., and it is gaining importance because it frees the user from maintenance perspective on an investment of some money for the use of these services provided by cloud service providers.

II. RELATED WORK

While Big Data gradually become a hot topic of research and business and has been everywhere used in many industries, Big Data security and privacy has been increasingly concerned. However, there is an obvious contradiction between Big Data security and privacy and the widespread use of Big Data. In this paper, we firstly reviewed the enormous benefits and challenges of security and privacy in Big Data. Then, we present some possible methods and techniques to ensure Big Data security and privacy.[2]

III. LITERATURE SURVEY

Present a literature survey and system tutorial for big data analytics platforms, aiming to provide an overall picture for non-expert readers and install a do-it-yourself spirit for advanced audiences to customize their own big-data solutions.

Introduces the Big data technology along with its importance in the modern world and existing projects which are effective and important in changing the concept of science into big science and society too. The various challenges and issues in adapting and accepting Big data technology, its tools are also discussed in detail along with the problems Hadoop is facing. We introduce View the privacy issues related to data mining from a wider perspective and investigate various approaches that can help to protect sensitive information. In particular, we identify four different type of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker.
We introduce a model for provable data possession (PDP) that can be used for remote data checking: A client that has stored data at an untrusted server can verify that the server possesses the original data without retrieving it. The model generates probabilistic proofs of possession by sampling random sets of blocks from the server, which drastically reduces I/O costs. The client maintains a constant amount of metadata to verify the proof. The challenge/response protocol transmits a small, constant amount of data, which minimizes network communication. Thus, the PDP model for remote data checking is lightweight and supports large data sets in distributed storage systems. The model is also robust in that it incorporates mechanisms for mitigating arbitrary amounts of data corruption.

IV. PROPOSED SYSTEM

We will give a comprehensive overview of the state-of-the-art technologies to preserve privacy of big data at each stage of big data life cycle. Moreover, we will discuss privacy issues related to big data when they are stored and processed on cloud, as cloud computing plays very important role in the application of big data. Furthermore, we will discuss about potential research directions. In this system we are providing attribute Based encryption as well as third party auditing in which we figure out the collusion attack in the exiting scheme and provide an efficient public integrity auditing scheme with secure group user revocation based on vector commitment and verifier-local revocation group signature. We design a concrete scheme based on our scheme definition. We are providing concept of deduplication, in simplified terms, data deduplication compares objects (usually files or blocks) and removes objects (copies) that already exist in the data set. The deduplication process removes blocks that are not unique. A public verifier is able to correctly verify shared data integrity.

A public verifier cannot distinguish the identity of the signer on each block in shared data during the process of auditing. The ring signatures generated for not only able to preserve identity privacy but also able to support block-less verifiability.

V. THE OPERATION OF THE SYSTEM

5.1. Project task set

5.1.1. File upload

File level De-duplication system:

If a file duplicate is found, the user will run the PoW protocol POWF with each S-CSP to prove the file ownership. for the j-th server with identity idj, the user first computes $\phi F;idj$= TagGen′(F, idj) and runs the PoW proof algorithm with respect to ϕF, idj. If the proof is passed, the user will be provided a pointer for the piece of file stored at j-th S-CSP. Otherwise, if no duplicate is found, the user will proceed as follows:

First divides F into a set of fragments $\{Bi\}$ (where i = 1, 2, · · ·).

For each fragment Bi, the user will perform a block-level duplicate check.

Block Level deduplication:

If there is a duplicate in S-CSP, the user runs PoWBon input: $\phi Bi;j$= Tag Gen′(Bi, idj) with the server to prove that he owns the block Bi. If it is passed, the server simply returns a block pointer of Bi to the user. The user then keeps the block pointer of Bi and does not need to upload Bi.

5.1.2. Privacy preserving in data processing

AES Algorithm:

The Advanced Encryption Standard (AES), also referenced as Rijndael (its original name), is a specification for the encryption of electronic data established by the U.S. National Institute of Standards and Technology (NIST) in 2001.The

key size used for an AES cipher specifies the number of repetitions of transformation rounds that convert the input, called the plaintext, into the final output, called the cipher text. The number of cycles of repetition is as follows:
10 cycles of repetition for 128-bit keys.
12 cycles of repetition for 192-bit keys.
14 cycles of repetition for 256-bit keys.
Each round consists of several processing steps, each containing four similar but different stages, including one that depends on the encryption key itself. A set of reverse rounds are applied to transform cipher text back into the original plaintext using the same encryption key.

High-level description of the algorithm

1. KeyExpansions: round keys are derived from the cipher key using Rijndael's key schedule. AES requires a separate 128-bit round key block for each round plus one more.
2. Initial Round:
   1. AddRoundKey: each byte of the state is combined with a block of the round key using bitwise xor.
3. Rounds:
   1. Sub Bytes: a non-linear substitution step where each byte is replaced with another according to a lookup table.
   2. Shift Rows: a transposition step where the last three rows of the state are shifted cyclically a certain number of steps.
   3. Mix Columns: a mixing operation which operates on the columns of the state, combining the four bytes in each column.
   4. AddRoundKey

4.Final Round (no Mix Columns)
   1. Sub Bytes
   2. Shift Rows
   3. AddRoundKey

MD5 Algorithm:
MD5 algorithm was developed by Professor Ronald L. Rivest in 1991. According to RFC 1321, “MD5 message-digest algorithm takes as input a message of arbitrary length and produces as output a 128-bit "fingerprint" or "message digest" of the input …The MD5 algorithm is intended for digital signature applications, where a large file must be "compressed" in a secure manner before being encrypted with a private (secret) key under a public-key cryptosystem such as RSA.”

5.1.3. Third party auditing

User store data on cloud and the intention of Third Party Auditing is to verify the integrity of data. The TPA who can verify the data from cloud server on data owners benefit needs to be authorized by data owner. There is also security risk if the Third Party can ask for dubious integrity proof over inescapable dataset. This step only required when client want some third party to verify data.

5.1.4. Data De-duplication

Concept of De-duplication, in simplified term, data de-duplication compares objects (usually file or blocks) and remove objects (copies) that already exists in the dataset. The de-duplication removes blocks that are not unique.

5.1.5. File Download

To download a file *F*, the user first downloads the secret shares *{cij,mfj}* of the file from *k*out of *n* storage servers. Specifically, the user sends all the pointers for *F* to *k* out of *n* servers. After gathering all the shares, the user reconstructs file *F*, *macF*by using the algorithm of Recover (*{·}*). Then, he verifies the correctness of these tags to check the integrity of the file stored in S-CSPs.

5.1.6. Output:

User can upload, download, recover, share files on cloud server and provide data deduplication and reliability.

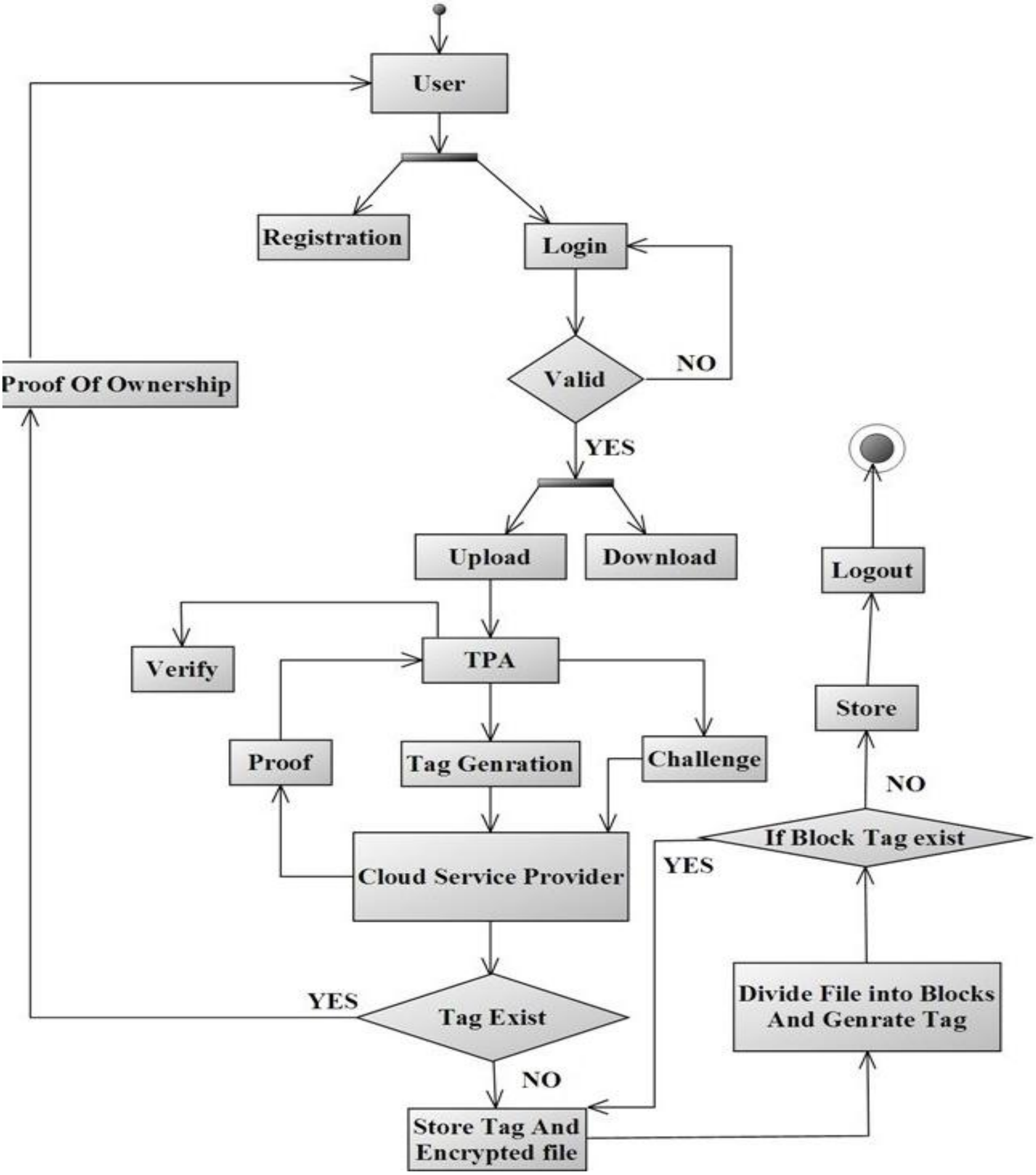


Fig.1: Activity Diagram

VI. Architecture Design

In this paper, we will give a comprehensive overview of the state-of-the-art technologies to preserve privacy of big data at each stage of big data life cycle. Moreover, we will discuss privacy issues related to big data when they are stored and processed on cloud, as cloud computing plays very important role in the application of big data. Furthermore, we will discuss about potential research directions. The remainder of this paper is organized as follows. The infrastructure of big data and issues related to privacy of big data because of the underlying structure of cloud computing. Data owners could perform integrity verification by themselves or delegate the task to trusted third parties. The basic framework of any integrity verification scheme consists of three participating parties: client, cloud storage server (CSS) and third party auditor (TPA). The client stores the data on cloud and the objective of TPA is to verify the integrity of data.

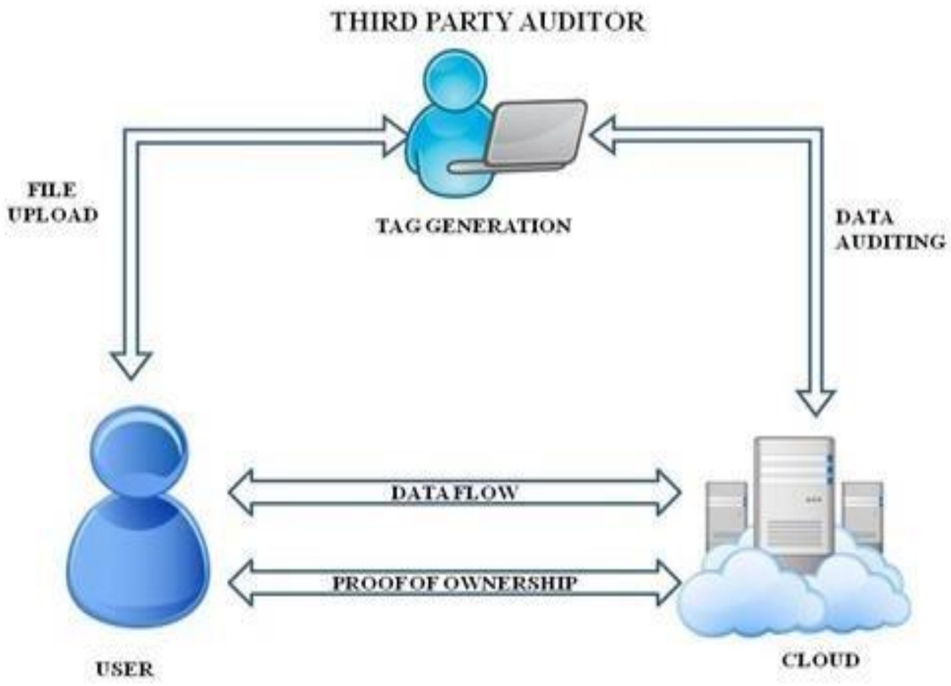


Fig.2: Architecture Diagram

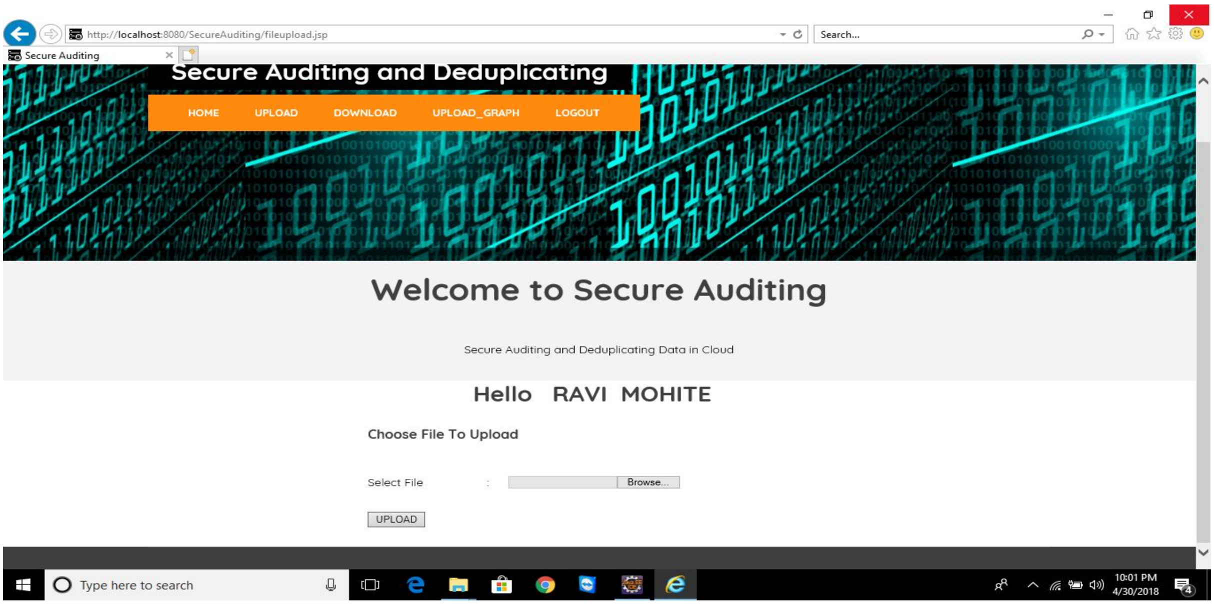
VII. Conclusion

We have seen how delegation of responsibility trusted 3rd party which provides security services secures user data. It reliefs the client from maintaining any kind of key information and allowing the client for using any browser enabled device to access the cloud services. It allows the client to verify the integrity of the data stored on download or retrieval of its own stored data in cloud. The client can share the data securely with specific band of people without any overhead of key distribution. We are providing concept of Deduplication which operates at the file or block level. File Deduplication eliminates duplicate files, but is not an efficient means of Deduplication.

To enhance the security more, a mechanism to secure the keys in security cloud can be an area of research. To reduce the overhead of network traffic can be another area of research.

VIII. Result

Following are the snapshots of complete flow of execution.


http://localhost:8080/SecureAuditing/fileupload.jsp

Secure Auditing

Secure Auditing and Deduplicating

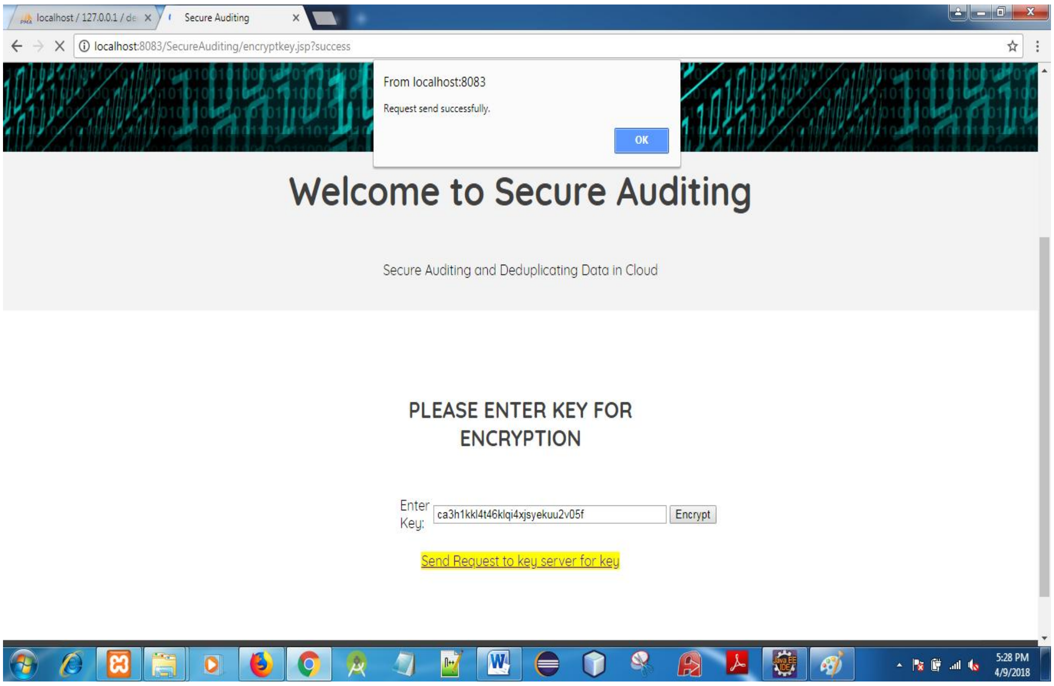
Welcome to Secure Auditing

Secure Auditing and Deduplicating Data in Cloud

Hello RAVI MOHITE

Choose File To Upload

Select File :


localhost / 127.0.0.1 / de
Secure Auditing
localhost:8083/SecureAuditing/encryptkey.jsp?success
From localhost:8083
OK
5:28 PM
4/9/2018

Request send successfully.

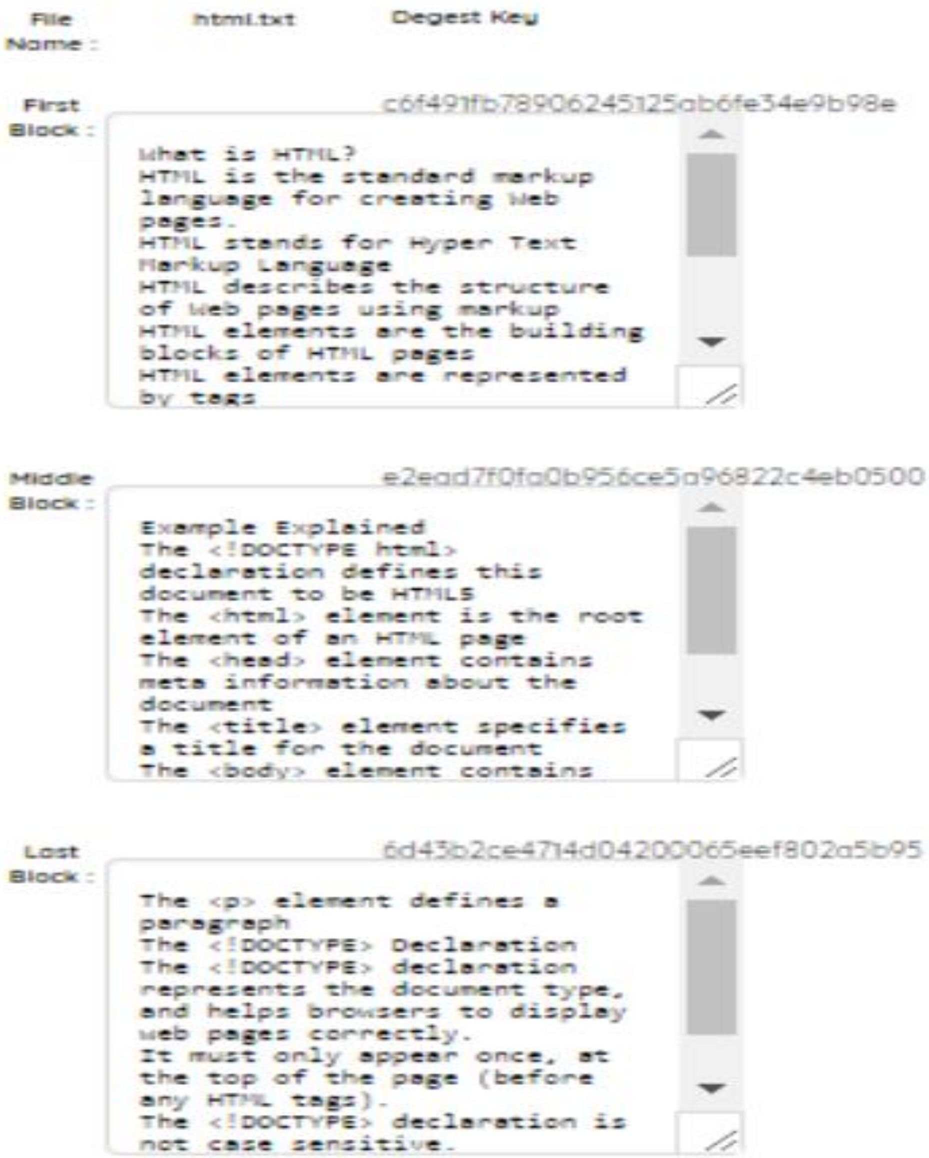
Welcome to Secure Auditing

Secure Auditing and Deduplicating Data in Cloud

PLEASE ENTER KEY FOR ENCRYPTION

Enter Key: ca3h1kkl4t46klqi4xjsyekuu2v05f

Encrypt
Send Request to key server for key



File Name : html.txt

Degest Key

First Block : c6f491fb78906245125ab6fe34e9b98e

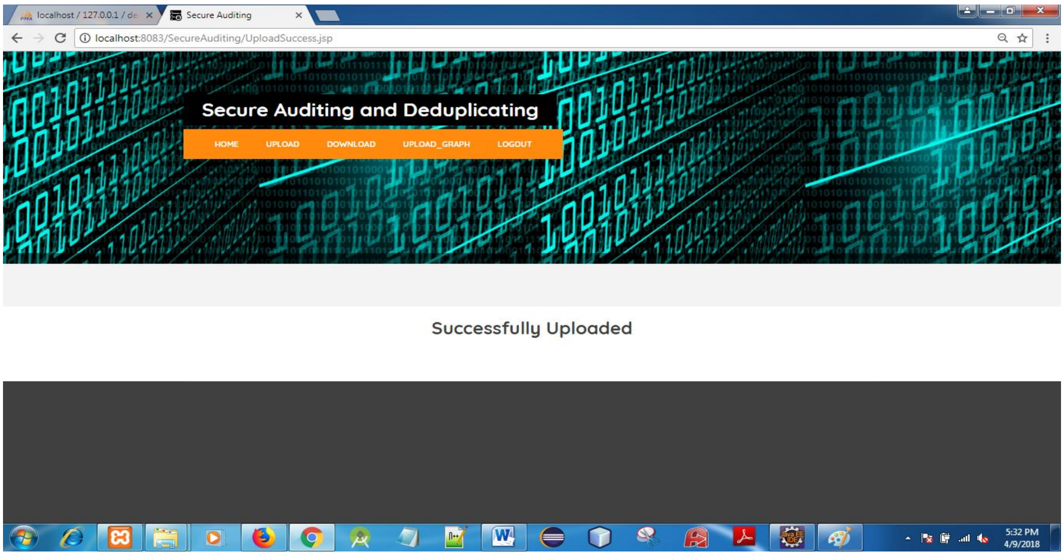
```
What is HTML?
HTML is the standard markup
language for creating Web
pages.
HTML stands for Hyper Text
Markup Language
HTML describes the structure
of Web pages using markup
HTML elements are the building
blocks of HTML pages
HTML elements are represented
by tags
```

Middle Block : e2ead7f0fa0b956ce5a96822c4eb0500

```
Example Explained
The <!DOCTYPE html>
declaration defines this
document to be HTML5
The <html> element is the root
element of an HTML page
The <head> element contains
meta information about the
document
The <title> element specifies
a title for the document
The <body> element contains
```

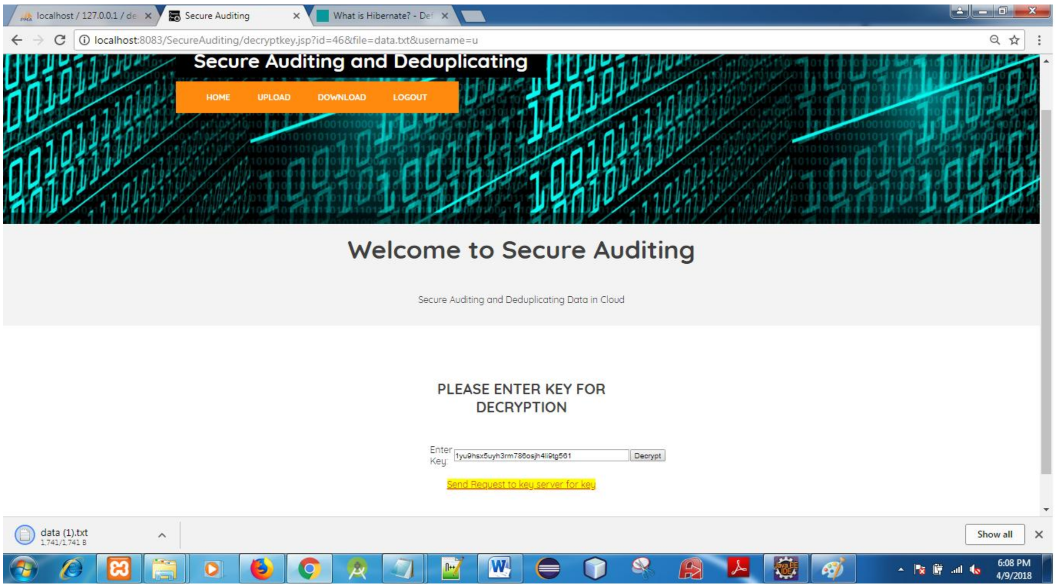
Last Block : 6d43b2ce4714d04200065eef802a5b95

```
The <p> element defines a
paragraph
The <!DOCTYPE> Declaration
The <!DOCTYPE> declaration
represents the document type,
and helps browsers to display
web pages correctly.
It must only appear once, at
the top of the page (before
any HTML tags).
The <!DOCTYPE> declaration is
not case sensitive.
```


localhost / 127.0.0.1 / de
Secure Auditing
localhost:8083/SecureAuditing/UploadSuccess.jsp
HOME
UPLOAD
DOWNLOAD
UPLOAD_GRAPH
LOGOUT
5:32 PM
4/9/2018

Secure Auditing and Deduplicating

Successfully Uploaded


localhost / 127.0.0.1 / de
Secure Auditing
What is Hibernate? - Def
localhost:8083/SecureAuditing/decryptkey.jsp?id=46&file=data.txt&username=u
HOME
UPLOAD
DOWNLOAD
LOGOUT
data (1).txt
Show all
6:08 PM
4/9/2018

Secure Auditing and Deduplicating

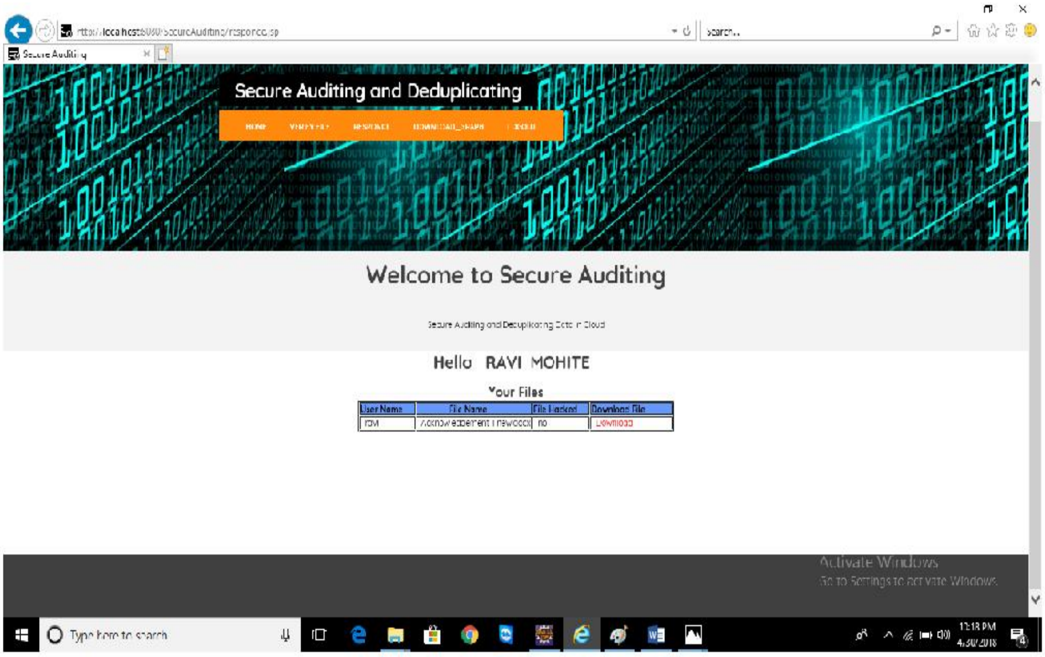
Welcome to Secure Auditing

Secure Auditing and Deduplicating Data in Cloud

PLEASE ENTER KEY FOR DECRYPTION

Enter Key: 1yu9hsx5uyh3rm786osjh4li9tg561 Decrypt

Send Request to key server for key



Secure Auditing and Deduplicating

Welcome to Secure Auditing

Secure Auditing and Deduplicating Data in Cloud

Hello RAVI MOHITE

Your Files

User Name	File Name	File Hacked	Download File
roM	Acknowledgement i newdocx	no	Download

REFERENCES

1. Abid mehmood, iynkaran natgunanathan yong xiang (senior member, ieee), gaug hua (member, ieee), and song guo (senior member, ieee) "protection of big data privacy". Hong-ryeol gil1, joon yoo1 and jong-won lee2 ,'an on-demand energy-efficient routing algorithm for wireless ad hoc networks', proceedings of the 2nd international conference on human. Society and internet hsi'03, pp. 302-311, 2003.
2. B. Matturdi, X. Zhou, S. Li, and F. Lin, ``Big data security and privacy:A review," China Commun., vol. 11, no. 14, pp. 135145, Apr. 2014.
3. J. Gantz and D. Reinsel, ``Extracting value from chaos," in Proc. IDCI View, Jun. 2011, pp. 112.
4. A. Katal, M. Wazid, and R. H. Goudar, ``Big data: Issues, challenges, tools and good practices," in Proc. IEEE Int. Conf. Contemp. Comput, pp. 404409, Aug. 2013.
5. L. Xu, C. Jiang, J.Wang, J. Yuan, and Y. Ren, ``Information security in BigData Privacy and data mining," in IEEE Access, vol. 2, pp. 11491176,Oct. 2014
6. H. Hu, Y. Wen, T.-S. Chua, and X. Li, ``Toward scalable systems for BigData analytics: A technology tutorial," IEEE Access, vol. 2, pp. 652687,Jul. 2014.
7. Z. Xiao and Y. Xiao, ``Security and privacy in cloud computing," IEEECommun. Surveys Tuts., vol. 15, no. 2, pp. 843859, May 2013
8. N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, ``Privacy-preserving multikeywordranked search over encrypted cloud data," IEEE Trans. ParallelDistrib.Syst.,vol. 25, no. 1, pp. 222233, Jan. 2014.
9. C. Hongbing, R. Chunming, H. Kai,W.Weihong, and L. Yanyan, ``Securebig data storage and sharing scheme for cloud tenants," China Commun.,vol. 12, no. 6, pp. 106115, Jun. 2015.
10. O. M. Soundararajan, Y. Jenifer, S. Dhivya, and T. K. P. Rajagopal, ``Data security and privacy in cloud using RC6 and SHA algorithms," Netw. Commun. Eng., vol. 6, no. 5, pp. 202205, Jun. 2014.
11. M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, ''A classification of location privacy attacks and approaches,'' Pers. Ubiquitous Comput., vol. 18, no. 1, pp. 163–175, Jan. 2014.
12. W. Peng, F. Li, X. Zou, and J. Wu, ''A two-stage deanonymization attack against anonymized social networks,'' IEEE Trans. Comput., vol, 63, no. 2, pp. 290–303, 2014.