ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

PRIVACY PRESERVING IN BIGDATA USING HASHING TECHNIQUE WITH MD5 AND DES

M.Indu Maheswari*[1], Dr S.Revathy[1]
[1]Sathyabama university,Chennai,India
Email: pop.indu@gmail.com

Received on 12-05-2016 *Accepted on 05-06-2016*

Abstract

A basic security requirement of big data storage is to guarantee the confidentiality of the data. Due to rapid growth in the multiuser communication on the cloud infrastructure, the reliability of the data sets is growing rapidly. Several significant data are subjected to achieve the privacy and security. By using the proposed three planning mechanism, the security can be acheived, they are anonymity, multiple receiver and conditional sharing. Using anonymity can hide the sender and receiver information details. In multiple receiver, by giving category the receiver can access the sender data before that admin check the condition to share the data. If the authorized receiver means the admin can allow the receiver to access the data. The data encryption standard (DES) and message digest (MD5) algorithm for encrypting the data and also used advanced encryption standard (AES) for Re-encrypt the data for high security. The re-encrypted data will be placed in hadoop server, whenever the user wants cipher text will be shared based on conditional sharing mechanism. Finally the data will be decrypted by the legitimate user. By implementing these algorithms the data from HDFS can be retrieved in fastest manner and hence, the time required for the proposed system to fetch the data is reduced compared to the existing system.

Keywords: Hadoop, bigdata, privacy,security,encryption.

1. Introduction

Big data is the collection of large data sets and its difficult to process the data. The dataset which exceeds storage capacity of terabyte are considered to be big data. Some of challenges are capture, storage, search, analysis, sharing etc. The sources of big data are social media, marketing, sensor data, mobile GPS data, scientific data etc. Big data offers, Increase in storage capabilities. Increase in processing power, Availability of data. Security in big data addresses the challenges presented by big data. Application software security: use the secure version of open source software. Account monitoring and control: manage the account for big data users. Secure configuration for hardware and software: build servers based on security for all systems in organizations big data architecture.

The organizations choose to upload their data to Hadoop since the hadoop supports considerable data storage service but also efficient data processing capability. It falls under the concept of cloud. Cloud computing describes a type of outsourcing of computer services. The idea behind cloud computing is similar: The user can simply use storage, computing power, or specially crafted development environments, without having to worry how these work internally. Cloud computing is usually Internet-based computing. The cloud is a metaphor for the Internet based on how the internet is described in computer network diagrams; which means it is an abstraction hiding the complex infrastructure of the internet. It is a style of computing in which IT-related capabilities are provided “as a service”, allowing users to access technology-enabled services from the Internet ("in the cloud")without knowledge of, or control over the technologies behind these servers.

Cloud Computing is a paradigm in which information is permanently stored in servers on the Internet and cached temporarily on clients that include computers, laptops, handhelds, sensors, etc." Cloud computing is a general concept that utilizes software as a service (SaaS), such as Web 2.0 and other technology trends, all of which depend on the Internet for satisfying users' needs.

For example, Google Apps provides common business applications online that are accessed from a web browser, while the software and data are stored on the Internet servers.

Some of related papers are referred such as,

The main idea of this paper, is to place as little trust and reveal as little information to the proxy as necessary to allow it to perform its translations. Proxy re-encryption (PRE) allows a proxy to convert a cipher text to be encrypted under a key which will be provided by sender.[1]

An application called atomic proxy re-encryption, in which a semi-trusted proxy converts a cipher text for Alice into a cipher text for Bob Without seeing the underlying plaintext.The present new re-encryption schemes that realize a stronger notion of security, the usefulness of proxy re-encryption as a method of adding access control to a secure file system.[2] The sender will be provided with unique key whenever he/she want to send the data to the receiver.The id which is given to sender is used for encrypt the data.[3]

The Alice have to send message to bob.In which the selective id, user attributes and identities are used for generating keys, such keys are used for encrypting the data.The generated key will be in constant size.[4]The algorithm called

MD5 ,DES and AES is used to encrypt the patient records,so patients details will be kept safe by this the security and privacy of patient data will be acheived.[5] The scheme of proxy re-encryption with deterministic finite automata is used.The data will be encrypted based on DFA of index string.The proxy cannot view plain text in encryption and decryption mechanism.The plain text is unknown to proxy.[6] The technique called multihop-identity based conditional proxy re-encryption is used to secure the data against ciphertext attacks.[7]

The importance of algorithm for calculating the accuracy while handling and clustering the data.[8] The representation of data based on clustering techniques and methodologies,shows significance management of bigdata.[9]The account of validating data while handling dataset by means clustering technique is done,shows validating of data is plays a vital role in bigdata.[10]The process of extracting the optimal cluster from the given dataset,shows how to extract the relevant data.[11]The process of providing immediate results to the patient based on the user question and disease.[12]

2.Proposed System Architecture:

In the Proposed system the novel notion called password based encryption with hashing techniques are used.This technique helps to preserve the anonymity for ciphertext sender/receiver, conditional data sharing and recipient-update. The ciphertext will be re-encrypted with the help of hash value.With the help of hashing technique, the generated key length will be more, so its hard to intruder to hack the data. The fig 1 shows that how the user given plain text is processed and placed in the hadoop server after encryption process and how the user receives their original result based on the conditional sharing mechanism.

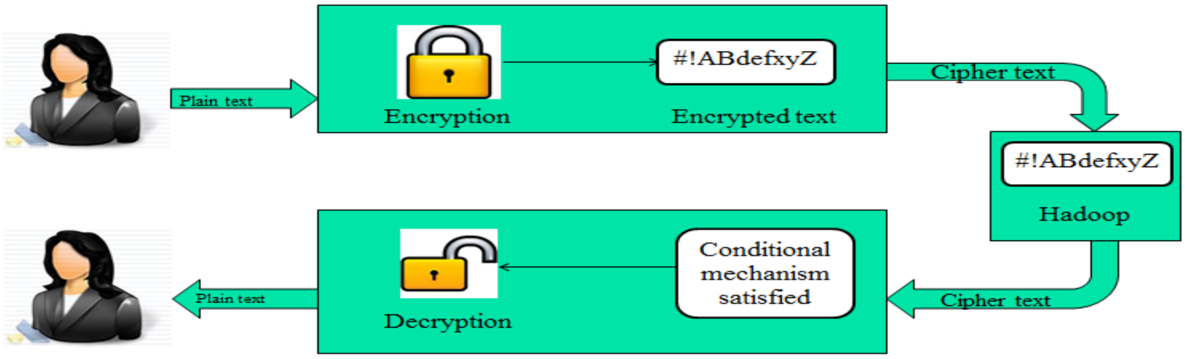


Fig: Proposed System Architecture.

All these process can be implemented using MD5,DES and AES algorithms. The proposed system architecture includes,

- Collect Data & Data Encryption
- Multiple receiver update

- Conditional sharing
- Data Access and Evaluation

2.1 Collect Data & Data Encryption:

The process starts with collecting the patient dataset.The dataset will be collected using following links,

http://worldwidescience.org/topicpages/p/patau+syndrome+trisomy.html

http://biotext.berkeley.edu/data/dis_treat_data/labeled_sentences_for_each_relation/only_dis

In that Patient dataset, it contains some attributes such as name, age, gender, month, location and symptoms of the patient's disease.Once the dataset has been chosen, it should be tokenized. The tokenization process includes removing unwanted symbols.

Algorithmfor pre-processing dataset:

Input: Dataset D

Output: Tokenisation(D)

```
Tokenize(d, delimit(,));
For(int i=0;i<=n; i++)  //where n is the number of records in the dataset
{
If(i="null" or i="unwanted symbols")
Remove (i);
}
```

For encryption algorithm called MD5 and DES is used.First To generate a salt (random): 8 bytes. Then Key Generation Starts here. Need to append the salt to the password. Further hash the data using MD5 as a result 16 byte hash will be generated. In that first 8 bytes is the key and next is the initialization vector. Then Pad the input string with1-8 bytes. By using that here one key is generated. To the next encrypt the input string using DES and encode using Base 64 Encode.

Again for re-encryption AES is used. The key must be generated. Get the key value by converting the key into bytes. By using key specification, get the key value and apply AES. Then convert into cipher by using the cipher instance of AES. Initialization vector is processed. Encryption will be done using the key.Then byte Data To Encrypt. In this input string is convert into byte data. That byte data has been ciphered. After BASE64Encoder() is used for encode the plaintext. As a result Encrypted data will be obtained. For decryption, the encrypted data form the hadoop server will undergone process with decryption of MD5 and DES, it includesBase64 decode the input message. Extract the salt (first 8 bytes).

The rest is the encoded text.Use derived key generation as in Encrypt above to get the key Decrypt the encoded text using key. Finally by remove padding the decrypted result will be obtained. Again decrypted result will be re-decrypted with AES. Process includes Get the key and key value. Convert value into bytes using Secret KeySpec then Get the cipher instance using Cipher.getInstance. The initialization vector is processing using Cipher.DECRYPT_MODE.For get the decoded value apply base 64 encoder.Finally the decoded plaintext of cipher text will be obtained.

Anonymity

Once the encryption is completed,the encrypted data will be stored in HDFS. The stored data is then shared among the multiple receiver. In this process the details of sender and receiver who communicates will be hided to ensure the security.

2.2 Multiple Receivers Communication

In this process multiple number of receivers are there to share their details such as profiles, id, password, and category to the senders.

Each and every receivers will be provided with unique id and password inorder to maintain the security.Category defines the data in which receiver is going to receive information,so with the help of unique id and key receivers can receive their information.

Algorithmfor multiple receivers communication

```
For (int  i=0; i<=n; i++)
{
Register (cloud);
Update (profile);
Anonymize (sensitive Attribute);
Generate (unique ID);
    Generate random number=unique id;
Update Profile;
Send to Owner;
```

```
    }
```

2.3 Conditional Sharing

Once the Multiple receiver updation is completed, conditional sharing mechanism will be processed. In this approach conditional sharing takes place based on the user's category, received data category and their unique key.If all the conditions are satisfied then then the user can able to receive the data in the encrypted form.This mechanism helps to ensures upto the maximum level security.

Algorithmfor conditional sharing

```
For (int i=0; i<=n; i++)
{
Receive (Multiple receiver details);
Share data ();
Share data ()
{
If (receiver category satisfies condition or matched with patient category)
{
Receive cipher text from owner;
}
}
```

2.4 Data Access and Evaluation

Receiver can access the data only if they are authenticated.The retrieved data will be in encrypted format.Once the receiver get the key from sender and decrypt the data and then can view their original data what they need.Evaluation of performance is based on encryption ad decryption time of existing and proposed system.

Algorithmfor data access and evaluation:

```
    For (int i=0; i<=n; i++)
    {
    Get keys from owner;
    Decrypt();
    Access data;
```

}

2.5 Privacy Preserving Security Algorithams

A. DES Algorithm

The DES (Data Encryption Standard) is an encryption algorithm to encrypt the data.

DES works on bits, or binary numbers in the form of 0s and 1s.In this paper, the patient data is encrypted in the form of 0s and 1s.

B. Message digest5 Algorithm

In this paper, we are using Message Digest5 (MD5) for encrypting the patient data. The encrypted data of patients are in the form of 0s and 1s.

C. AES Algorithm

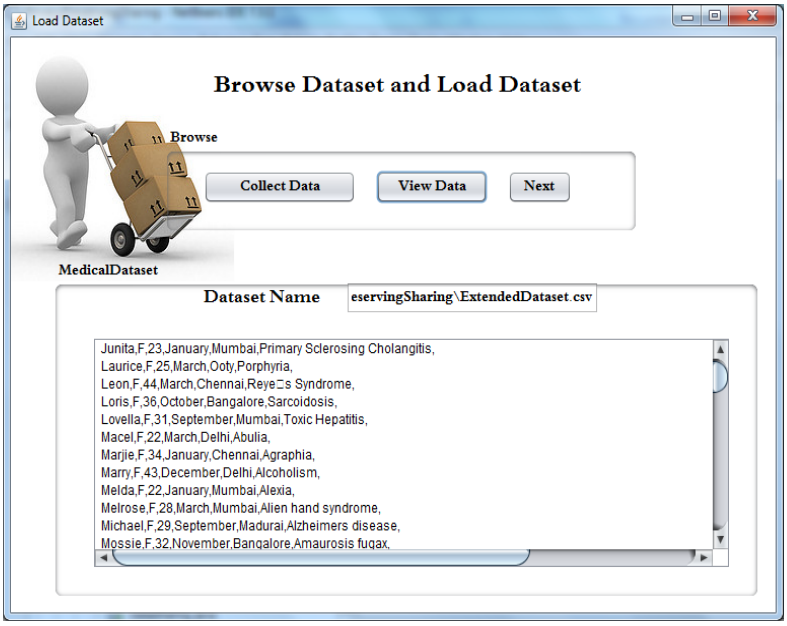
In this paper, AES algorithm is used for re-encryption and decryption. In Re-encryption the encrypted data will be encrypted again. In decryption, the encrypted data will be changed to the original format.

3. Results and Discussion

This section provides calculation for the experiments performed in this paper for better understanding of the underlying validation scheme.

Input: Patient dataset ,Output: Filtered data

The fig 3.1 shows the experimental analysis of how the dataset has been collected and displayed.

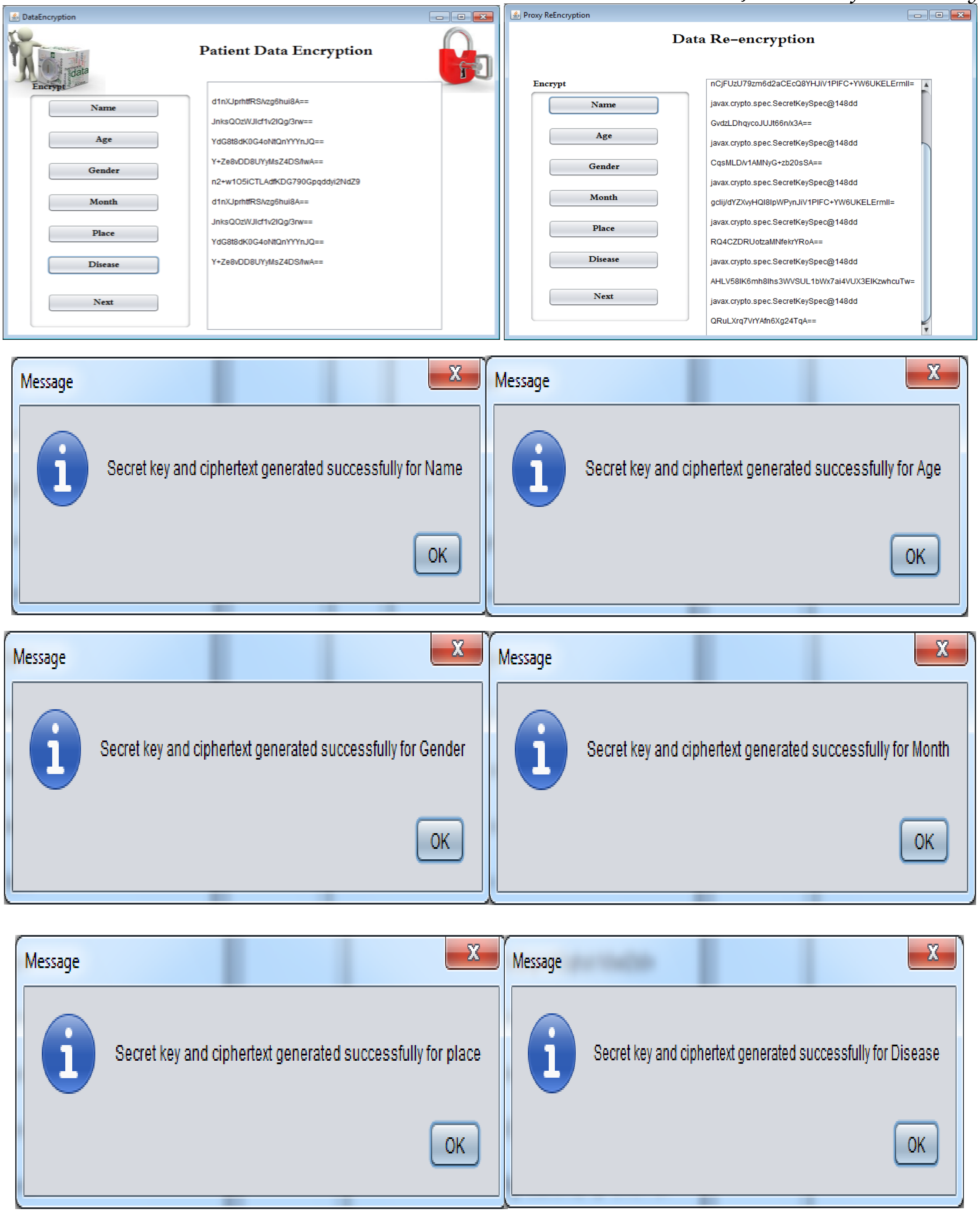


Dataset Name: eservingSharing\ExtendedDataset.csv

```
Junita,F,23,January,Mumbai,Primary Sclerosing Cholangitis,
Laurice,F,25,March,Ooty,Porphyria,
Leon,F,44,March,Chennai,Reye□s Syndrome,
Loris,F,36,October,Bangalore,Sarcoidosis,
Lovella,F,31,September,Mumbai,Toxic Hepatitis,
Macel,F,22,March,Delhi,Abulia,
Marjie,F,34,January,Chennai,Agraphia,
Marry,F,43,December,Delhi,Alcoholism,
Melda,F,22,January,Mumbai,Alexia,
Melrose,F,28,March,Mumbai,Alien hand syndrome,
Michael,F,29,September,Madurai,Alzheimers disease,
Mossie,F,32,November,Bangalore,Amaurosis fugax,
```

Fig 3.1 Experimental result for data collection.

The fig 3.2 shows the experimental analysis of how data has been encrypted using MD5 and DES and re-encrypt using AES algorithm for all the attributes



```
d1nXJprhtfRS/vzg6hui8A==
JnksQOzWJlcf1v2lQg/3rw==
YdG8t8dK0G4oNtQnYYYnJQ==
Y+Ze8vDD8UYyMsZ4DS/lwA==
n2+w1O5iCTLAdfKDG790Gpqddyi2NdZ9
d1nXJprhtfRS/vzg6hui8A==
JnksQOzWJlcf1v2lQg/3rw==
YdG8t8dK0G4oNtQnYYYnJQ==
Y+Ze8vDD8UYyMsZ4DS/lwA==
```

```
nCjFUzU79zm6d2aCEcQ8YHJiV1PIFC+YW6UKELErmlI=
javax.crypto.spec.SecretKeySpec@148dd
GvdzLDhqycoJUJt66n/x3A==
javax.crypto.spec.SecretKeySpec@148dd
CqsMLD/v1AMNyG+zb20sSA==
javax.crypto.spec.SecretKeySpec@148dd
gclij/dYZXvyHQI8IpWPynJiV1PIFC+YW6UKELErmlI=
javax.crypto.spec.SecretKeySpec@148dd
RQ4CZDRUotzaMNfekrYRoA==
javax.crypto.spec.SecretKeySpec@148dd
AHLV58IK6mh8Ihs3WVSUL1bWx7ai4VUX3EIKzwhcuTw=
javax.crypto.spec.SecretKeySpec@148dd
QRuLXrq7VrYAfn6Xg24TqA==
```

Secret key and ciphertext generated successfully for Name

Secret key and ciphertext generated successfully for Age

Secret key and ciphertext generated successfully for Gender

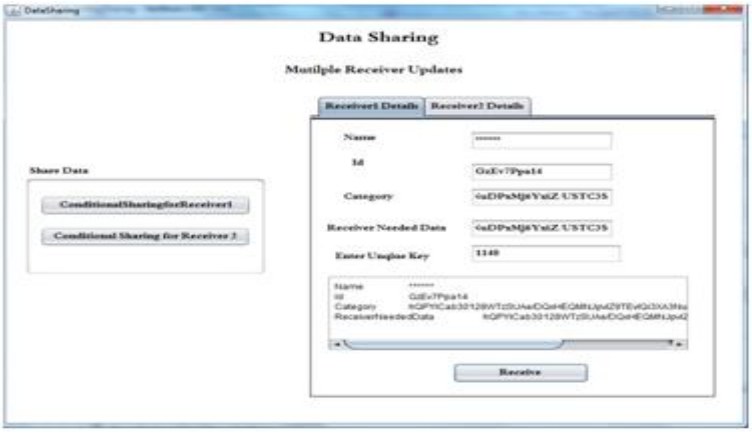
Secret key and ciphertext generated successfully for Month

Secret key and ciphertext generated successfully for place

Secret key and ciphertext generated successfully for Disease

Fig 3.2 Experimental result for encryption and Re-encryption for all attributes.

The fig 3.3 shows the experimental analysis of how the multiple receiver details are received.


DataSharing
Data Sharing
Mutilple Receiver Updates
Share Data
ConditionalSharingforReceiver1
Conditional Sharing for Receiver 2
Receiver1 Details
Receiver2 Details
Receive

Name: *******

Id: GzEv7Ppa14

Category: 6aDPsMj8YaiZ USTC35

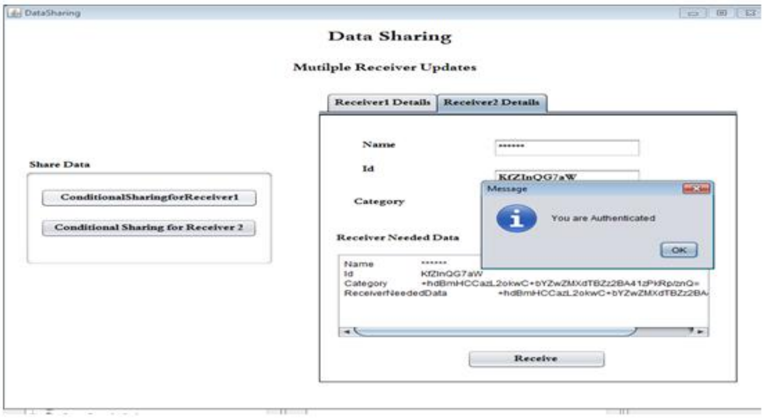
Receiver Needed Data: 6aDPsMj8YaiZ USTC35

Enter Unqiue Key: 1140

```
Name                *******
Id                  GzEv7Ppa14
Category            tQPYiCab30128WTzSUAeDQe4EQMhUpd29TEvtQt3XA3fNa
ReceiverNeededData           tQPYiCab30128WTzSUAeDQe4EQMhUpd2
```

Fig 3.3 Experimental result for multiple receiver update.

The fig 3.4 shows the experimental analysis of conditional sharing mechanism and how the user has been authenticated.


DataSharing
Data Sharing
Mutilple Receiver Updates
Share Data
ConditionalSharingforReceiver1
Conditional Sharing for Receiver 2
Receiver1 Details
Receiver2 Details
Receive
Message
OK

Name ******

Id KfZInQG7aW

Category

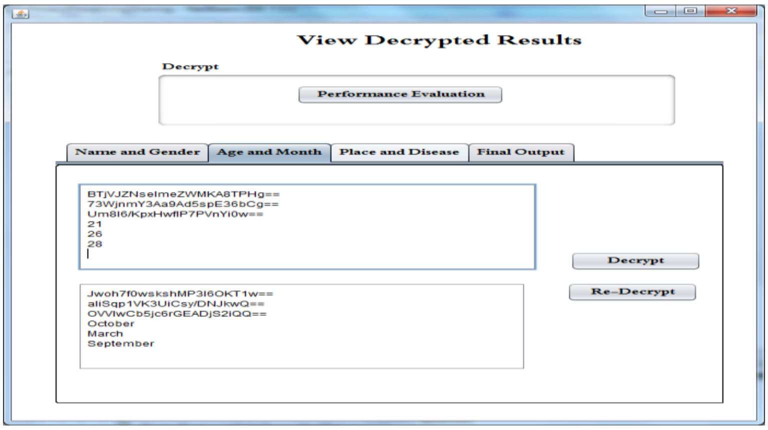
Receiver Needed Data

You are Authenticated

```
Name                ******
Id                  KfZInQG7aW
Category            +hdBmHCCazL2okwC+bYZwZMXdTBZz2BA41zPkRp/znQ=
ReceiverNeededData          +hdBmHCCazL2okwC+bYZwZMXdTBZz2BA
```

Fig 3.4 Experimental result of conditional sharing mechanism.

The fig 3.5 shows the experimental analysis for how the data is decrypted and accessed in receiver side.



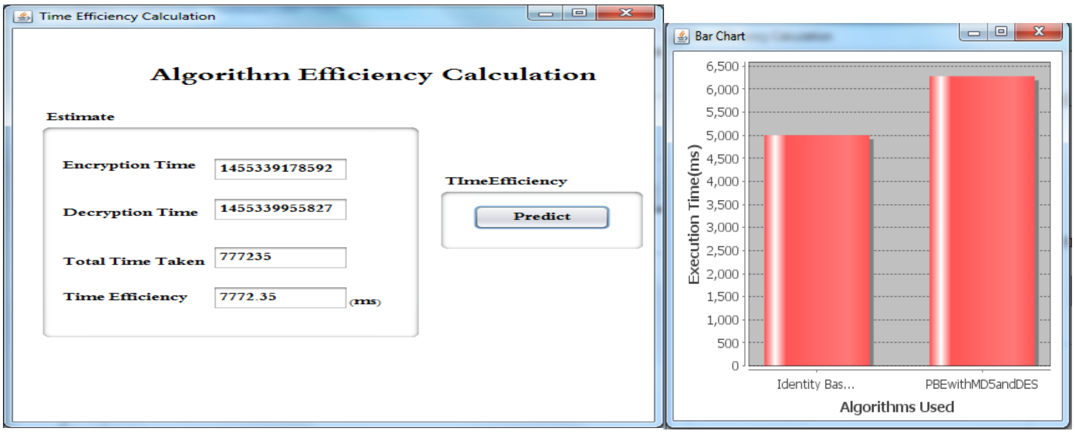
```
BTjVJZNseImeZWMKA8TPHg==
73WjnmY3Aa9Ad5spE36bCg==
Um8I6/KpxHwfIP7PVnYi0w==
21
26
28
```

```
Jwoh7f0wskshMP3I6OKT1w==
aliSqp1VK3UiCsy/DNJkwQ==
OVVIwCb5jc6rGEADjS2iQQ==
October
March
September
```

Fig 3.5 Experimental result for data accessing in receiver side.

4. Performance Analysis

The performance analysis shows that how the proposed system works better when compared to existing system The fig 4.1, fig 4.2 clearly shows that proposed system algorithm reduces the time taken for processing the data for both receivers ,which shown in the graph.



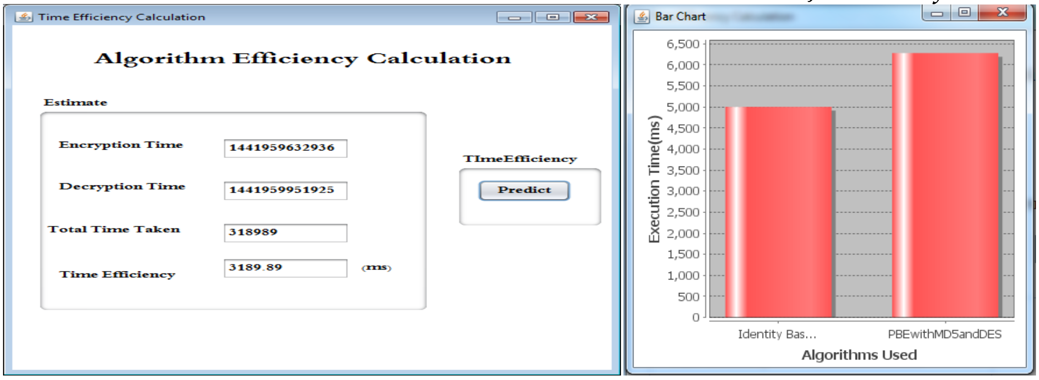
Encryption Time 1455339178592

Decryption Time 1455339955827

Total Time Taken 777235

Time Efficiency 7772.35 (ms)

Fig 4.1 Performance analysis of proposed algorithm based on execution time (receiver 1).



Time Efficiency Calculation

Algorithm Efficiency Calculation

Estimate

Encryption Time: 1441959632936

Decryption Time: 1441959951925

Total Time Taken: 318989

Time Efficiency: 3189.89 (ms)

TImeEfficiency

Predict

Fig 4.2 Performance analysis of proposed algorithm based on execution time (receiver 2).

5. Conclusion

- The Proposed System proposes a privacy-preserving ciphertext multi-sharing mechanism to achieve better improvement and privacy.
- It combines the merits of proxy re-encryption with anonymous technique in which a ciphertext can be securely and conditionally shared multiple times without leaking both the knowledge of underlying message and the identity information of ciphertext senders/recipients.
- Furthermore, this project shows that the new primitive is secure against chosen-ciphertext attacks in the standard model. As conditional sharing mechanism provides more accuracy. As strong encryption mechanisms are used, the security of data is enhanced properly
- It can be applicable to many real-world applications, such as secure email forwarding, electronic encrypted data sharing, where both anonymity and flexible encrypted data sharing are needed.

5.1 Future Enhancement

- In future the method called as threshold proxy re-encryption scheme can be used.
- This scheme integrated with decentralized erasure code which formulated a secure distributed storage system. This system not only supports storing and retrieving in secure manner, but also supports forwarding messages from one storage server to another server without retrieving. The main technique implemented here is, encoding the encrypted messages, forwarding methods as well as decryption. Serial actions of our proposed system is encrypting, encoding, and forwarding. The proposed system provides copy of robustness data in all storage servers and which allow more flexible.

Reference

1. G. Ateniese, K. Benson, and S. Hohenberger,Key-private proxy re-encryption, springer, Year: 2009 PP. 279-294.

2. G. Ateniese, K. Fu, M. Green, and S. Hohenberger.Improved proxy re-encryption schemes with applications to secure distributed storage. ACM transaction on information and system security Year: 2005, PP. 29-43.
3. D. Boneh and X. Boyen ,Efficient selective-ID secure identity-based encryption without random oracles, Springer, Year: 2004
4. D. Boneh, X. Boyen, and E.-J. Goh ,Hierarchical identity based encryption with constant size cipher text, Springer, Year: 2005, pp. 440-456.
5. M.Indu Maheswari,Dr S.Revathy, R.Tamilarasi,Secure data transmission for multisharing in big data storage,Global Congress On Computing And Media,NOV-2015.
6. Kaitai Liang, Man Ho Au, Joseph K. Liu and Duncan S. Wong,A DFA-Based Functional Proxy Re-Encryption Scheme for Secure Public Cloud Data Sharing,Information Forensics and Security, IEEE Transactions, year:2014.
7. Kaitai Liang, Willy Susilo, and Joseph K. Liu “Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage” IEEE transactions on information forensics and security, vol. 10, no. 8, august 2015
8. Revathy S, Parvathavarthini B. “On Rough Fuzzy Cluster Validity Indices”, 2013 Proceedings of Sri Eshwar Engineering College International Conference on Advanced Computing and Communication Systems(ICACCS), Coimbatore, 2013, pp. 1-4.
9. Revathy S, Parvathavarthini B. “Integrating rough clustering with Fuzzy Systems”, Proceedings of MGR University International Conference on Sustainable Energy and Intelligent System[SEISCON 2011], Chennai, 2011, pp. 865-69.
10. Revathy S, Parvathavarthini B. “Decision theoretic Evaulation of Rough Fuzzy Clustering”, Arab Gulf journal of Scientific Research, 2014, 32 (2/3), pp. 161-67.
11. Revathy S, Parvathavarthini B. ”Futuristic validation method for Rough Fuzzy Clustering”,Indian Journal of Science and Technology, 2015, 8 (2), pp. 120-27.
12. R.Tamilarasi, M. Indu Maheswari,**”**Performance Evaluation Of MapReduce Workloads On Medical Dataset**”**, paper will be published in “Research Journal Of Pharmaceutical, Biological and Chemical Sciences(RJP BCS)”ISSN:0975-8585.

Corresponding Author:
M.Indu Maheswari*,
Email: *pop.indu@gmail.com*