A Secure Data Evaluation and Publishing Technique for Big Data

Pratiksha Patil
Computer Science & Engineering Department
Sushila Devi Bansal College of Technology, Indore
Madhya Pradesh, India

ABSTRACT

The number of applications is become large now in these days which are dealing with thousands of users in a second. Therefore, the data is such application is also collected and processed in large quantity. To deal with such data the big data technology is used that is combination of software and hardware for efficient data processing. The aim of the proposed work to address the different privacy and content sensitivity issues in big data environment. In addition, of that the effort is made for improving the content to prevent the data leakage during the content publishing in public domain. Therefore, the proposed work is contributed for designing an attribute key encryption technique that works on random attribute selection policy. The selected attribute is used for common key generation, which is used for the shared files. To generate the key for encryption the MD5 algorithm is used. Additionally for the encryption the efficient algorithm namely the AES algorithm is used. Secondly for identifying the sensitive content on the user's text the NLP (natural language processing) based technique is applied. That technique is used to extract the part of speech information from the text and to identify the noun from the text. That are the target data which is need to be encrypted. After encryption of the target text, the data is again reformed for publishing. To implement the entire scenario the web application is used which is usage the Hadoop storage for preserving the data. After implementation, the performance of the system measured in terms of time and space complexity. According to the results, the performance of system found acceptable.

Keywords

Privacy preserving, big data, data publishing, data leakage, NLP, POS

1. INTRODUCTION

Big data is a term, which is used to indicate a significant amount huge collection of data. That kind of data cannot deal with the normal traditional techniques of information processing or normal human efforts. Therefore a combined infrastructure which is based on computational hardware and software is designed that is big data. Additionally the processes or techniques that are supporting analysis of large amount of data are termed as big data processing or analytics. As per their overview when the huge amount of data is combined and processed in a same place there are some chances or mistakes can also be possible. But the mistakes and small problems in such critical environment where the sensitive and private data is processes can damaged some person's privacy and security either financially or socially. In this presented work the proposed work is focused on the privacy and sensitivity management in big data.

Now in these days a number of organizations where the rapidly data is generated in terms of seconds are usages the services of big data. The big data is offers quick processing of information, extraction of decisional points and other opportunities. Among them the large e-commerce web applications and social media, applications are primary consumers of the big data services. Both kind of online service providers collect a huge amount of data in fraction of seconds and distribute all the relevant information to other web application users. But both kinds of applications are includes various critical and private information such as mobile number, credit card information and other kinds of social comments. Therefore a common sand secure framework is required that understand the needs of data flow and involved processes and safeguard the private and sensitive information from become public. In order to demonstrate the required solution for security and privacy a cryptographic model is proposed that first identify the sensitive and private information and preserved the information to disclose to others.

2. PROPOSED WORK

The main aim of the proposed work is to involve the security and privacy in big data analysis when the data is aggregated and distributed to the other clients. In this context a model is designed and described in this chapter.

2.1 System Overview

Now days the traditional computing techniques are not much functional due to large amount of data generation and their analysis for making future decisions and other applications. In this context new technology is used in various applications for supporting the need of new generation requirements that technology is known as big data and big data analytics. This technique usages the concept of machine learning, parallel computing and distributed computing to analyze large amount of data with less resource utilization. Therefore highly crowded applications are required to use this infrastructure such social media, email servers and the e-commerce web applications. In these applications a significant amount of users are live and frequently exchanging or generating data for other one.

In such environment a significant amount of data is collected from a number of users and numbers of sources additionally it is also distributed for different other purposes. In this case a small or little mistake or leakage of data can create a complex issue. In this context to prevent this case of mistake or data leakage issue a new cryptographic security technique is proposed for design and implement. In this context a social media site is considered for providing the key issue and proposed solution. The social media data is private and sensitive in different context additionally the publication of data can create various issues. In addition of that a significant amount of traffic is concentrated over this source of information. Therefore, the social media is frequently supporting the use of big data. Therefore, the social media site is basically used for demonstrating both the aspects of proposed system development. This chapter introduced the overview of the proposed system design and the next section describes the key issues and challenges to resolve in this work.

2.2 Problem Domain

The social media is one of the platforms which are frequently used for sharing contents to their friends, public exposure and as a message to someone confidentially. In technical terms the data in social media either shared between users in one to one and one to many manners. Additionally the nature of this data is heterogeneous in format or multiple in formats which is kept on a same place for distribution. Therefore the following issue can occurred during the distribution of content by minor mistake:

1. ***Access control:*** Access control on data which is being published in the social network need to be regularized for unauthorized users. For example, a post which is needed to be share between only the close friends is published as a public post.
2. ***Data owner privacy:*** Most of post is published as text post the post with a name or some confidential information can also create issue for the data owner.
3. ***Multiple key management*****:** In a small amount of time a number of users are creating different number of post for each post if a new key is managed then a significant amount of time and resources are need to maintain for only key generation and their management.

2.3 Proposed Methodology

The initial concept of the proposed system can be understood using the diagram 2.1 where first end of users are continuously generating contents from their devices and the application collect all the data into a single place. On the other hand the overall data is reflected to different users according to their groups, communities or messages. According to the proposed problem domain all the data has the issue of privacy and security during data distribution and publication. Therefore most of the systems are consumes the techniques of cryptography.

These cryptographic techniques are secure and low cost for implementation for any security applications. There are two major concerns are first the control on data and the content confidentiality. In addition of that the implementation of cryptographic technique also needs additional effort for managing the keys and key generation techniques to manage the frequently developed data.

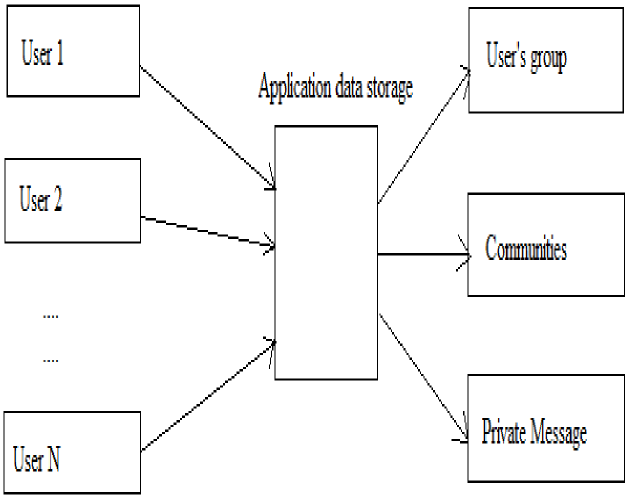

Figure 2.1 Social Media Interactions

Here two terms control on data and confidentiality of content are the primary issue of the work. Therefore first a cryptographic technique is developed using the hybridization of ASE and MD5 encryption algorithm. In addition of that for improving the security the ABE (attribute based encryption) technique is used for reducing the overhead of key management technique. Now first consider the diagram 2.2 where the initial cryptographic technique is provided.

The given cryptographic system is simple inn design and implementation. That is composition of MD5 algorithm and AES algorithm for encrypting the target data. The use of MD5 algorithm is performed for finding the fixed 128 bit hash code for algorithm to utilize as the encryption for the input text data. on the other hand system usages any length of plane text to encrypt the information from others. The system usage the 128 bit MD5 generated hash key and plain text to generate the cipher text using AES algorithm. This text is send to the application storage for utilizing in future.

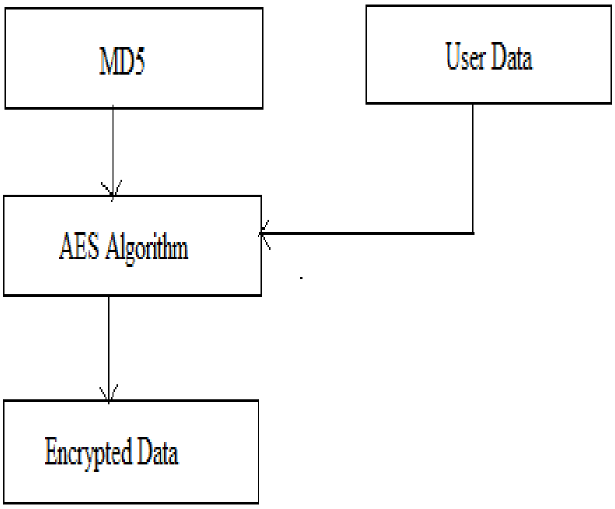

Figure 2.2 Cryptographic Technique

But the current model needs a key to encrypt the data additionally if when a new text or data appeared then we need to manage one more key. Therefore in order to reduce this effort the proposed technique includes the attribute based concept for securing data. The technique utilize on user attribute decided by the cryptographic system designed which is provided by the user to system. But the selection of attribute is performed once and attribute selection is performed on random manner. Now when the data is only keep in social network for self use then only selected attribute is used as the input to the MD5 algorithm for generating the encryption key. But there are two aspects remain to introduce:

1. When data is shared between number of friends
2. Data is communicated to a target user

The data which is stored in a cloud is encrypted using a symmetric key encryption technique namely AES algorithm. Additionally the secrete key which is randomly selected by the designer is used for securing data. in this similar manner the data shared between two person is secured by selected the target user's attribute and the user's own attributes and for generation of dynamic key both the attributes are processed using the MD5 algorithm which always provide a similar length of key say 128 bit. In this similar manner when a post is publicly that means the post is secured by a password which is generated according to the all the friend's attributes therefore only the friends in the user's list can able to recover the data accurately in the given social network.

Therefore using attribute based data encryption technique can prevent the data to access un-authorize manner and but the how to prevent the sensitive data to be public. Therefore the proposed technique includes the sensitivity scanning technique for performing this task. The sensitivity of data is evacuated using the NLP (natural language processing) technique therefore a POS (part of speech) tagger is used for evaluating the text find the noun words from the post data. That word is replaced with the encrypted text for security and privacy concern.

2.4 System Architecture

The proposed system model is defined using the figure 2.3. In this model the entire working of the proposed privacy preserving model is tried to demonstrate. The system initiate work when the user provide input text to publish on the social media friends feed or wall. That text is shared among all the friends who are associated with the particular user. Therefore in first the key generation is performed by the system. That process is taken place once and stored with the user account.

When a new friend is included in the user's friend list then the key is again updated. For generation of the encryption key the system select a random attribute from the entire friend list.

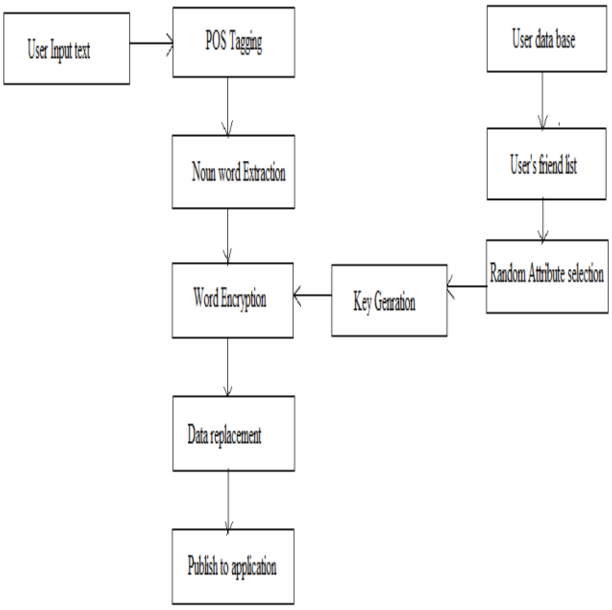

Figure 2.3 Proposed Model

And the selected attributes are processed using the MD5 algorithm to generate a fixed 128 bit string. On the other hand the user's input text or post is used with the POS tagger which is a NLP (natural language processing) library. That tagger identifies all the text part of speech. After that form the text noun words are targeted and encrypted using the generated attribute based key. The encrypted data is replacing the noun word for sensitivity removal from text. Finally the post is ready to publish. The following advantage is observed after performing this task.

1. If the post is publically posted among all the social media then nobody can see the encrypted words
2. Only the friends can able to see the data

2.5 Proposed Algorithm

This section includes the step process followed by the system to reduce the privacy issue of the shared text in big data environment.

Table 2.1 Proposed Algorithm

```
Input: user friend list $L_n$, text to post T
Output: improved post I

Process:
    1. $for(i = 1; i \leq n; i++)$
        a. $A_i = getRandomAttribute(L_i)$
    2. $end\ for$
    3. $K = MD5.GenrateHas\square(A_n)$
    4. $TagList = POS.ParseText(T)$
    5. $for(j = 1; j \leq TagList.lengt\square; j++)$
        a. $if(TagList_i == Noun)$
            i. $E = AES.Encrypt(word_i, K)$
        b. End if
    6. End for
    7. $I = reformText(T, E)$
    8. Return I
```

3. RESULT ANALYSIS

This chapter provides the discussion about the obtained results and the parameters on which the performance evaluation performed. The provided work is based on the cryptography and web application thus the following parameters are taken for consideration.

3.1 Encryption Time

The time required to encrypt the data to be published using the proposed encryption technique is termed as the encryption time of the algorithm. To calculate the encryption time the following formula is used.

$$Encryption\ Time = Encryption\ end\ time - Start\ time$$

Table 3.1 Encryption Time

Experiments	Time in MS (millisecond)
1	32
2	27
3	40
4	36
5	31
6	34
7	29

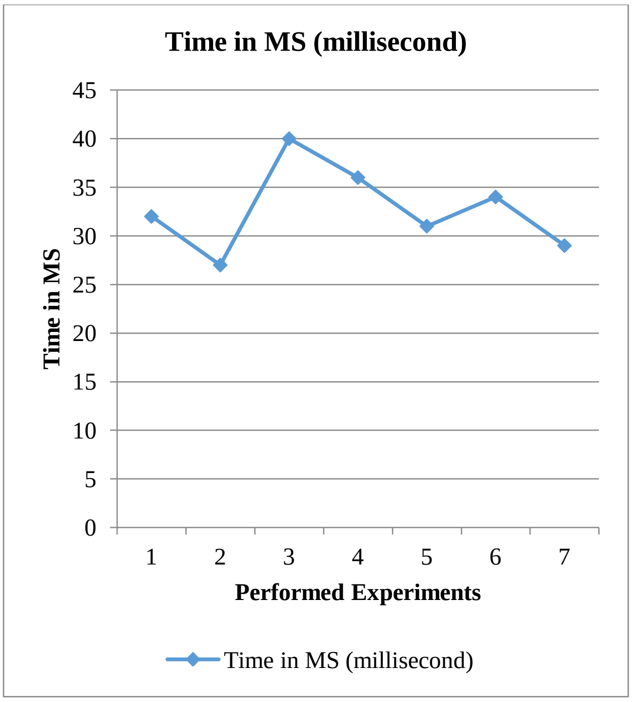

Figure 3.1 Encryption Time

The encryption time of the proposed big data security model is demonstrated using figure 3.1 and the table 3.1. In this diagram the X axis contains the different experiments performed with the system and the Y axis contains the time consumed with the concerned experiment. The time reported in this diagram and table is measured in terms of milliseconds. According to the given results the encryption time is varied in all the experiments because the amount of text for encryption is provided different length. But overall results are acceptable and not consume enough time for encryption process.

3.2 Decryption Time

To decode the encrypted cipher text the consumed amount of time is termed as the decryption time of cryptographic algorithm. According to the definition the time difference between the decryption process initialization and complete the decoding process of target text is defined as the decryption time. That is computed using the following formula:

$$Decryption\ Time = Decoding\ end\ time - Decoding\ start\ time$$

Table 3.2 Decryption Time

Experiments	Time in MS (millisecond)
1	30
2	24
3	38
4	33
5	29
6	32
7	28

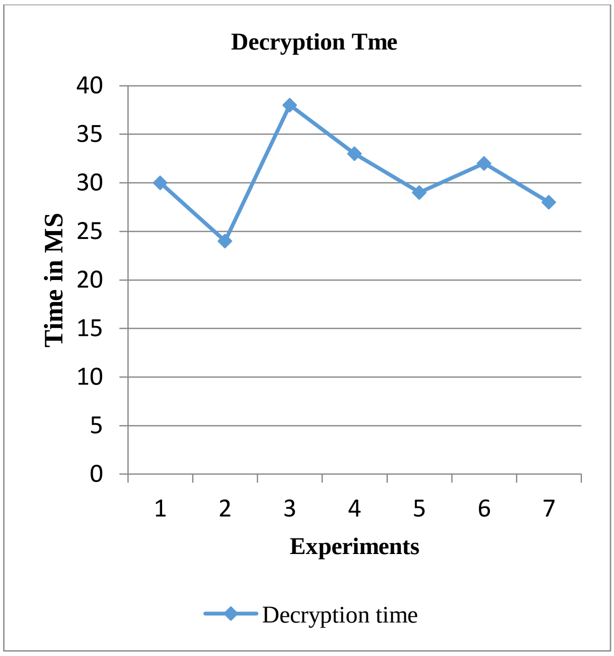

Figure 3.2 Decryption Time

The computed decryption time of the proposed security technique is given in figure 3.2 and table 3.2. In this diagram the Y axis includes the time required to decrypt data and the X axis records the number of experiments. The reported time in this diagram is defined in milliseconds. According to the graph and table the time consumption of encryption and decryption is approximately similar, but decryption process reflects the less time requirements as compared to the encryption process. According to the observations the decryption process needs acceptable amount of time for decoding the text.

3.3 Encryption Memory

The main memory is required to place data and instructions for encryption process are notified as the encryption memory. The memory consume of the system is computed using the following formula:

$$Memory\ usage = Total\ alloted\ memory - free\ memory$$

Table 3.3 Encryption Memory

Experiments	Memory in KB (Kilobytes)
1	23662
2	22741
3	22558
4	23056
5	23591
6	22746
7	22510

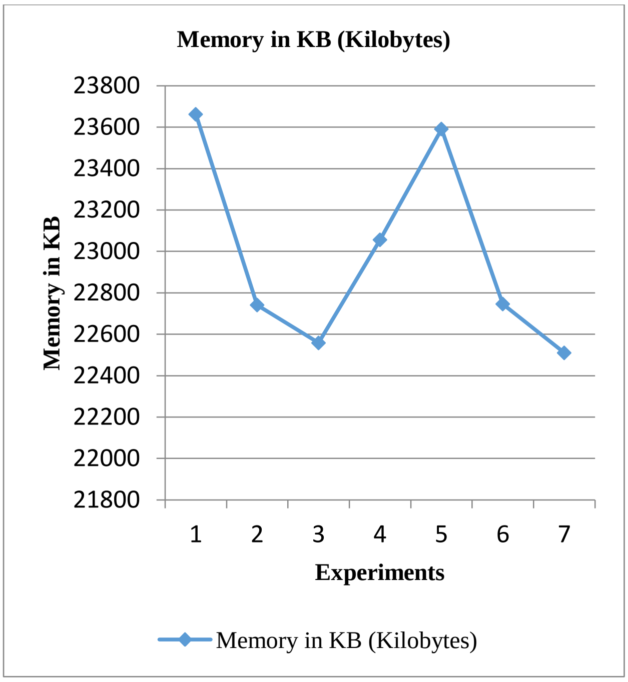

Figure 3.3 Encryption Memory

The memory utilization of the proposed system during the encryption of the data is reported using figure 3.3 and table 3.3. The diagram shows the memory usages in Y axis and the X axis contains the different observations made during the experiments. The results show the memory usages varies between 22000-23800KB (kilobytes). But not exceeding in any experiments. Therefore the encryption algorithm is acceptable for offering security of the system and efficiently performed required operations.

3.4 Decryption Memory

The amount of memory required to hold the decrypted data and to execute the decoding process is required memory is known as the decryption memory usages. As the encryption memory computed in the similar manner the decryption memory is computed using the formula, the formula for computing memory usage is given as:

$$Memory\ Usage = total\ alloted\ memory - free\ memory$$

Table 3.4 and figure 3.4 represents the memory usages of the system during the decryption of encrypted text. To report the performance of the algorithm for memory usage the X axis includes the experiments conducted with the system and Y axis reports the corresponding memory usages of the algorithm. According to the calculated results the memory consumption of algorithm is moderate in both the conditions i.e. encryption and decryption. Therefore the performance of the encryption is acceptable for both the parameters namely time and space complexity.

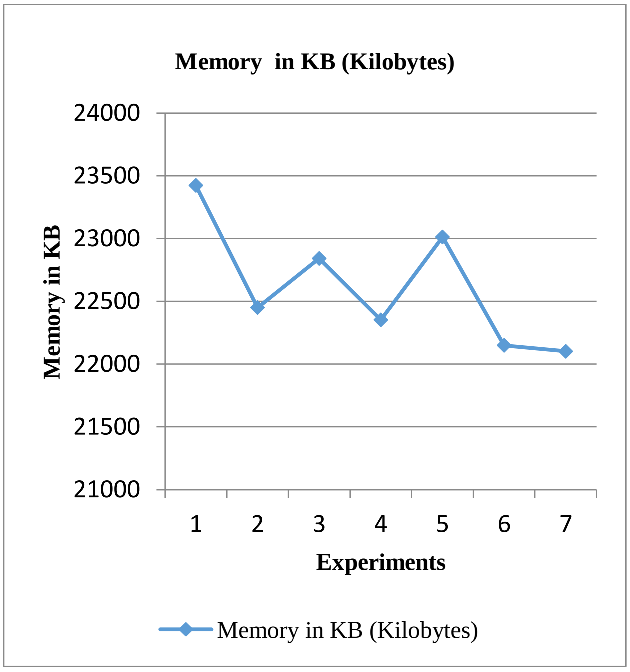

Figure 3.4 Decryption Memory

Table 3.4 Decryption Memory

Experiments	Memory in KB (Kilobytes)
1	23424
2	22451
3	22842
4	22353
5	23013
6	22150
7	22102

3.5 Response Time

The response time of the server is termed the amount of time required to accept the user end user’s requires and provide the return the required resource. Thus the time different between user request and server execution is termed as the server response time. The server response time for the proposed technique is demonstrated in figure 3.5 and table 3.5. In this diagram the X axis shows the experiment performed and Y axis shows the respective time consumed for server response. Basically the server response time depends on the target hosting and the hardware and software configuration therefore it is not an effective parameter for system evaluation. But the server response time defined in these experiments is acceptable and provides the outcomes in less time as compared to encryption and decryption time.

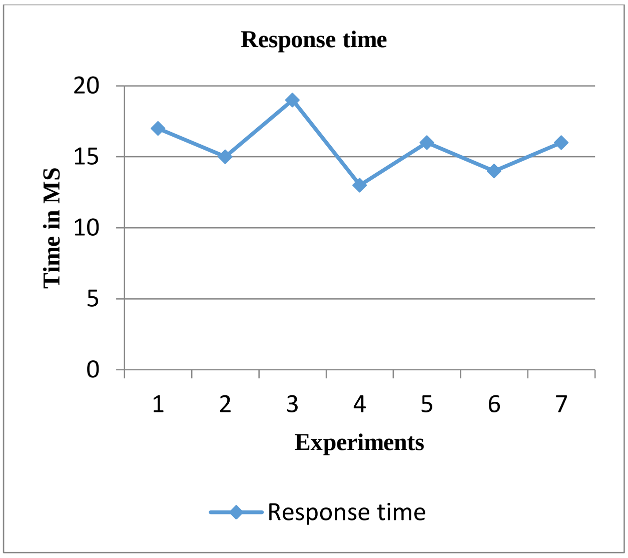

Figure 3.5 Response Time

Table 3.5 Response Time

Experiments	Response ime
1	17
2	15
3	19
4	13
5	16
6	14
7	16

4. CONCLUSION

The proposed work is intended to provide the data security and data owner's privacy in big data environment. This chapter provides the conclusion of the conducted research work and obtained experimental results. Additionally the future extension of the work is also included in this chapter.

4.1 Conclusion

In this era of technology a number of applications are existing that generate and consumes data in bulk amount such as e-commerce, social media applications and others. These applications collect a huge amount of data in fraction of seconds, process it and generate response according to requirements. In order to perform such kind of computations the normal computing is not much effective therefore the big data and their analytic techniques are required. in this work the big data technology and the issue of security and privacy solution investigation is the key area of study. Therefore to demonstrate the need to of security and privacy and the data sensitivity the social media application is used.

In the proposed model first a social media application developed which usages the Hadoop storage for preserving data when any user of social media platform write a post. The proposed model first analyzes the text using the POS tagger which provides the part of speech information (i.e. noun, pronoun and others). It is assumed that the sensitive data contains the noun part of the information. The proposed model is responsible for find such kind of noun words form the post and encrypt the target text. To encrypt the words the proposed AES and MD5 based cryptographic algorithm is used.

The MD5 usages the user attribute data which is randomly selected by the application. Therefore the proposed cryptographic algorithm works on attribute based cryptographic scenario. After replacing the sensitive content form the user's post the data is able to publish on social media platform. The proposed cryptographic algorithm is secure even when the system usages the symmetric key encryption because the key generation is depends on the users who are going to access the target post or data. The proposed model is implemented on the JAVA based technology and with the Hadoop. After the implementation of the proposed model the system is evaluated on different parameters the observation of experimentation is reported using table 4.1.

Table 4.1 Performance Summary

S. No.	Parameters	Remark
1	Encryption time	The acceptable amount of time consumption is found for sensitivity scan and the encryption of target text that is varies between 15-38 MS
2	Decryption time	The decryption time less as compared to encryption time
3	Encryption memory	Less amount of memory required for encryption process because short text are processed using these algorithms
4	Decryption memory	The acceptable amount of memory resource required for computed decoded text
5	Server response time	That is depends on the server and server configuration which found acceptable

According to the reported performance parameters and obtained results the proposed technique found efficient and effective securing the data and data sensitivity in a complex environment. Thus the system is acceptable for big data security purpose.

4.2 Future Work

The proposed and implemented system is a concept of privacy preserving technique on big data. That is implemented and it is found that is acceptable for security and privacy concern in huge data communication. In near future the following future extension is included in this work.

1. The proposed work is developed for simulation of the proposed concept therefore the limited amount of data and limited amount of user profiles are used. It is required to test the proposed system for huge amount of data and significant amount of profiles and text size.
2. The proposed model currently developed using the AES and MD5 based cryptographic approach for efficiency point of view. In near future the split key encryption policy is also included for improving the cryptographic security.

5. REFERENCES

[1] Liang, Kaitai, Willy Susilo, and Joseph K. Liu, "Privacy-preserving ciphertext multi-sharing control for big data storage", IEEE transactions on information forensics and security 10.8 (2015): 1578-1589.

[2] “Big Data: What it is and why it matters”, online available at: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[3] “What is Big Data? The Basics – Meaning and Usage”, the windows club, online available at: http://www.thewindowsclub.com/what-is-big-data

[4] Research Trends: Special Issue on Big Data, 30 September 2012

[5] Available online at: http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

[6] Kuchipudi Sravanthi and Tatireddy Subba Reddy, “Applications of Big data in Various Fields”, (IJCSIT) International Journal of Computer Science and Information Technologies, Volume 6 (5) 2015, pp. 4629-4632

[7] Greveler, Ulrich, Benjamin Justus, and Dennis Loehr, "A Privacy Preserving System for Cloud Computing", 2011 IEEE 11th International Conference on Computer and Information Technology (CIT), IEEE, 2011.

[8] Benjamin C. M. Fung, Ke Wang, Rui Chen, Philip S. Yu, "Privacy- preserving data publishing: A survey of recent developments," In: ACM Computing Surveys (CSUR), Vol. 42, pp 1- 53, 2010.

[9] S. Hansell, "AOL removes search data on vast group of web users," New York Times, 2006.

[10] Divyakant Agrawal, Amr El Abbadi, and Shiyuan Wang, "Secure and privacy-preserving data services in the cloud: A data centric view." Proceedings of the VLDB Endowment 5, Number 12 (2012): 2028-2029.

[11] V. Abricksen, “A Survey on Cloud Computing and Cloud Security Issues”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622International Conference on Humming Bird (01st March 2014).

[12] Mohammad Asadullah and R. K. Choudhary, “Data Outsourcing Security Issues and Introduction of DOSaaS in Cloud Computing”, International Journal of Computer Applications (IJCA), PP. 40-45, Volume 85 – No 18, January 2014.