Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1157 - 1166
Received 20 January 2021; Accepted 08 February 2021.

Sensitive Data Identification and Protection in a Structured and Unstructured Data in Cloud Based Storage

M.Rajkamal[1], M. Sumathi[2], N. Vijayaraj[3], S. Prabu[4], G Uganya[5]
[1]Application Developer, IBM, Bangalore.
[2]Assistant Professor, Department of Computer Science and Engineering, K.Ramakrishnan College of Engineering, Trichy,
[3]Assistant Professor, Department of Computer Science and Engineering, VelTech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai.
[4] Research Scholar, SASTRA Deemed University, Thanjavur
[5]Assistant Professor, Department of Electronics and Communication Engineering, Saveetha School of Engineering, SIMATS, Chennai.
[1]rajkamalmurugasean@gmail.com, [2]sumathishanjai.nitt@gmail.com, [3]vijaiphdraj@gmail.com, [4]prabu@cse.sastra.ac.in, [5]uganyaece@gmail.com

ABSTRACT

Nowadays user information is large in size and is stored in a cloud storage location in different forms like structured and unstructured. In these storage representations maximal sizeof the information is common to all users and some of the information is differ from one user to other. The common information does not require security and differed information required security. The conventional entire data encryption technique leads to higher computational complexity and reduce data usability to authorized users. To reduce computational complexity and increase data usability, the selected data encryption technique is proposed in this work. If the sensitive information is present in the form of structured data, the security required attribute is partitioned from other attributes and encrypted instead of entire data. Similarly, if the data present in an unstructured form, the information extraction technique is used to extract the security required sensitive information. Afterwards the extracted attributes are encrypted by Attribute Based Encryption (ABE) technique. When compare to conventional encryption techniques, the proposed selected sensitive attribute technique provides better security to sensitive attributes with lesser computational time and complexity. Similarly the data usability to non-encrypted non-sensitive attributes is increased.

Keywords:
Sensitive Attribute, Information Extraction, Information Protection, Security and Usability.

1.Introduction

Cloud computing provides everything as a service to requestor with a minimal cost and management efforts. The major benefits of cloud computing are on-demand self services, broad network access, rapid elasticity, resource pooling and measured services. Due to these characteristics from individual users tolarge organizations are moving their personalized and official data from their personalized storage devices to cloud storage locations. Every technology having its own merits and demerits. Similarly, cloud computing is also having their own merits and demerits. The major issues of cloud computing are resource allocation, deduplication, load balancing and security issues. When comparing these issues, security issues create critical issues for handling highly sensitive data like financial and medical data. Hence, security issues handling plays a vital role in cloud based sensitive data storage[1]. Conventionally, user data is encrypted as a whole by a data handling organization and stored into the cloud storage location. This entire document encryption reduces data usability of authorized users and increases computational complexities. Similarly, in cloud storage, user data is managed and maintained by third party Cloud Service Providers (CSP). Sometimes, these CSP's are also act as adversaries. To handle these issues an alternate protection technique is required.

In present scenario data owners (D_O) are willing to protect their selected data from the outside

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1157 - 1166
Received 20 January 2021; Accepted 08 February 2021.

world and also willing to publish certain data to authorized users. This partial data protection technique provides trade-off between security and usability of user data with the knowledge of the D_O. Hence, the $D_O$preferred sensitive data (S_D) protection technique is required for secure data storage in cloud storage [2].Applying security technique to entire data leads to lesser data usability to authorized users and security to S_D. Hence, the security preferred S_D is separated from other data and security mechanisms are applied to the S_D instead of entire data. This selective data encryption provides a higher level of protection to S_D. To separate S_D from the remaining data plays a major role in a selective data encryption. Nowadays, user data are maintained in different form like a structured and unstructured. When a data is maintained in a structured form, the row and column based separation technique is used for partition. If the data is maintained in an unstructured form, the information extraction technique and natural language processing based keyword or pattern matching techniques are used for separation of S_D[3]. After segregating the S_D from other data, the S_D is encrypted with an encryption algorithm and stored in cloud storage.

The conventional encryption algorithms[15] work well for conventional personalized storage system. When a data is stored in cloud storage, these conventional techniques are not providing high end security to user S_D. Hence, an efficient security technique is required to handle S_D in cloud storage. At present, Attribute Based Encryption (ABE) technique is a more preferable protection technique in cloud based storage. The security strength of the ABE depends on a number of attributes are involved in an encryption technique [4]. To overcome the issues of ABE technique, group key based encryption technique is required to protect each user data in a separate manner. [14] The group key based encryption technique provides a higher level of security than the existing protection techniques. The security strength of group key depends on encryption algorithm and key management [5]. Thus leads to the motivation of sensitive data identification and protection of S_D in a structured and unstructured data in cloud storage.
The remaining part of the paper is organized as follows: in section 2, the works that are related to S_D identification and protection in a structured and unstructured data is analyzed with the merits and demerits. In section 3 the proposed technique is discussed with the corresponding algorithms and section 4 the security analysis part is analyzed. The experimental results are discussed in section 5 and section 6 concludes with the proposed work and future enhancement.

Table 1. Notations Used in Proposed System

1	CSP	Cloud Service Provider
2	D_O	Data Owner
3	S_D	Sensitive Data
4	ABE	Attribute Based Encryption
5	NS_D	Non-Sensitive Data

2. Related Work

The existing works are related to S_D separation in a structured/unstructured data and protection techniques are going to be discussed deeply in this section.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1157 - 1166
Received 20 January 2021; Accepted 08 February 2021.

2.1 **Sensitive attribute separation**

Cecil Eng et al. proposed the instance-based attribute identification in structured database integration[13]. The author considered schema and summary instance information for processing the attribute instances along with group of attributes. The attribute domain class hierarchy, attribute classification and formation of attribute groups etc [6]. Yong Yi et al. proposed Privacy Protection Method for multiple sensitive attribute based strong rule technique.[12] The association rule based sensitive attribute identification is discussed for identifying multiple sensitive attributes in a structured data. The identified attributes are clustered and applied privacy preserving technique to that data [7]. Cedric du Mouza et al. proposed the automatic detection of sensitive information in a structured database. The semantic rule based attribute detection with linguistic values is discussed for attribute identification. This semantic modeling based attribute identification provides betters result to smaller sized data but not working well for large sized data [8][11].

2.2 **Sensitive attribute protection**

The works that are related to sensitive attribute protection is going to be discussed in this section. Kalyan Nagaraj et al. proposed an encrypting and preserving sensitive attributes in customer churn data using novel dragonfly based pseudonymizer approach. The dual protection technique is proposed in this work such as dragonfly and pseudonymizer algorithms are discussed in this work. This dual protection technique provides better security but takes higher computational time [9]. Razaullah Khan proposed a privacy preserving for multiple sensitive attribute against fingerprint correlation attack satisfying c-diversity. The sensitive attributes are bucketed through fingerprint correlation technique. The one-one to correspondence is identified between the terms for sensitive attribute bucketing. It provides easy privacy preserving to sensitive information but applicable to domain related dataset not for all types of data [10].
Limitations of existing techniques:

- The accuracy of existing S_D technique depends on number of rules and linguistic values.
- The dual protection technique increases computational complexities and not suitable to all types of data.
- The accuracy of finger print correlation technique is depended number of finger prints.

Hence, the generic technique which is suitable to all type and domain related technique is required in a current scenario.

3. Proposed Sensitive Data Identification and Protection Technique in Cloud Storage

The sensitive data identification and protection technique in a structured and unstructured data in cloud based storage provides high protection to sensitive data with a tradeoff between security and processing time. Figure 1.shows the flow diagram of the proposed system. The input data is given as a unstructured (pdf) document and read through python code based PDF reader. The identified S_D is stored in the output text file and passes to protection. Similarly, the input data is in a structured data, the specific attribute name which is given as S_D by the D_O is segregated from other attributes. Now, the segregated S_D in structured and unstructured document is encrypted by an Attribute Based Encryption (ABE) technique. Now, the encrypted S_D is stored in a cloud storage location. Hybrid cloud is used for proposed work. Here, the private cloud is used for store

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1157 - 1166
Received 20 January 2021; Accepted 08 February 2021.

the encrypted data and public cloud is used for storage of Non-Sensitive Data (NS_D). When compared to public cloud, private cloud provides high end security to $S_D$but, the cost of data maintenance in private cloud is high. Hence, private cloud is used for encrypted S_D storage and public cloud is used to store NS_D. This storage scheme provides tradeoff between storage cost and security to S_D.

3.1Sensitive data identification in an unstructured document

Algorithm 1 is used for the separation of S_D from an unstructured document such as medical document, financial document, registration document, criminal records etc. These documents contain maximum sized common information and minimal sized unique information. Instead of providing security technique to entire document, the unique information is identified from the document and security techniques are applied on it. This technique reduces computational complexities like encryption and decryption time.

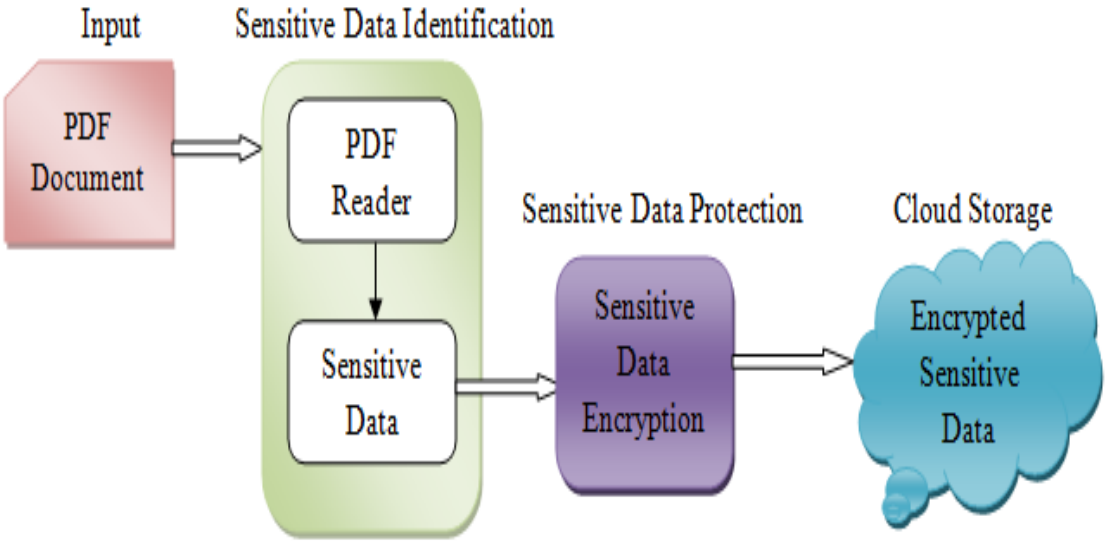

Figure 1. System architecture - Sensitive data identification and protection

---

Algorithm 1: Separation of Sensitive Data

Input: Unstructured Document
Output: Sensitive Data
Algorithm:

```
for all document
        fromfpdf import FPDF
        # traverse text
        for i in range(len(text)):
        char = text[i]
        pdf = pdfplumber.open('Medical_Rec-1.pdf')
        page = pdf.pages[0]
```

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1157 - 1166
Received 20 January 2021; Accepted 08 February 2021.

```
        text = page.extract_text()
        #print(text)
        name_search=re.search(r'(?<=Name    : ).*?(?=\s)', text)
        found_name=name_search.group()
returnfound_name
```

From the unstructured document, the specified list of words is identified through patterns which are exactly matched to the given terms by a keyword extraction process. The predefined list of words is identified through D_O willingness. Eg. The names are identified by the pattern
name_search=re.search(r'(?<=Name:).*?(?=\s)', text)
i.e. the terms which are followed by Name: is identified as S_D. Similarly, the other S_D is identified in the document. Afterwards, the identified list of S_D is sent to protection part.

3.2 Sensitive attribute identification in a structured data

Sometimes user data is maintained in a structured format like a table. In this case, the specific list of attributes is identified as S_D by an attribute partition technique. The fuzzy rule based classification technique is used for attribute partition process. When compared to association rule fuzzy rule provides higher accuracy rate. Hence, if-then rule based fuzzy rules are used for attribute partition in a structured data. The categorical value of an attribute is converted to ordinal values for fuzzy rule formation. Three different threshold values are fixed for the classification such as higher (>75), middle (<25 and >75), lower (<25). Here the higher and middle level values are identified as S_D. Now, the identified S_D is passed to the protection stage.

4. Sensitive data protection

In the proposed system, the S_D is encrypted instead of entire data. This technique reduces processing time and computational overhead. Similarly, this technique provides tradeoff between security and time management system. Algorithm 2 shows the encryption part of the proposed work.
The ABE encryption technique is used for protection of S_D. The 128-bit block size and 256-bit key size is taken for encryption and decryption process. When compare to conventional techniques the proposed technique provides perfect security to security preferred attributes. Hence, ABE technique is preferred in this work. Algorithm 2 clearly shows the encryption process of the proposed work.

Algorithm 2: Protection of Sensitive Data

Input: Sensitive Data
Output: Encrypted Sensitive Data
Algorithm:

```
        def encrypt(text,s):
        # Encrypt uppercase characters
                if (char.isupper()):
                result += chr((ord(char) + s-65) % 26 + 65)
        # Encrypt lowercase characters
                else:
```

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1157 - 1166
Received 20 January 2021; Accepted 08 February 2021.

```
result += chr((ord(char) + s - 97) % 26 + 97)
return result
encrypted_name=encrypt(found_name,5)
encrypted_text = text.replace(found_name, encrypted_name)
print(encrypted_text)
```

Now the encrypted S_D is stored in a private cloud and non-encrypted NS_D is stored in a public cloud. This storage technique reduces processing cost without sacrificing the security of S_D. The non-encrypted NS_D provides better usability to authorized users without delay and computational complexities than the existing techniques.

5. Security analysis

The strength of the security technique depends on the key confidentiality. In ABEtechnique the security strength is higher than the conventional encryption techniques like DES, AES and Blowfish algorithms. The proposed technique is secure against the conventional attacks like man-in-the middle attack, brute-force attack, known plain text attack and known- cipher text attacks.

- **Man-in-the-Middle Attack** – In a proposed technique, the S_D is encrypted by an ABE technique. The attributes which are used for the encryption is known by a D_O only. Hence, adversaries are unable to find the key for an encryption and decryption process.
- **Brute-force Attack** – The S_D are encrypted by different attribute values. In a brute-force attack, the adversaries are tried number times for finding the key value. In a proposed technique, 256 bit key size is used for encryption. If the number of SD group is 'n', then the adversaries works 2^n times to find the entire list of key values. This 2^n finding increase key identification time. Practically, this key identification is an impractical task.
- **Known Plaintext Attack**- In a known plaintext attack, the adversaries are tried to identify the unknown attributes through known attributes. In a proposed technique, the related attributes are grouped together and unrelated attributes are grouped separately. Hence, adversaries are unable to predict the other group attributes. Thus, the known plaintext attack is impossible in a proposed technique.
- **Known Ciphertext Attack**- The known ciphertext attack also impossible in a proposed technique because of the different grouping of S_D.

Based on the analysis, the proposed technique provides better security than the conventional techniques.

6. Experimental results

The proposed technique is implemented in python language with our own synthetic dataset. The financial dataset with 1000 users containing 30 attributes. From this 30 attributes 12 attributes are selected as sensitive. Hence, these 12 attributes are encrypted instead of entire 30 attributes. This reduced attribute encryption takes lesser encryption time than the entire data encryption. Figure 2 shows the comparison graph for entire attribute and selected S_D size. When compared to entire document the S_D size is nearly 50% is lesser. These lesser size S_D takes lesser encryption time than entire data encryption. Figure 3 shows the comparison of entire data and S_D encryption time.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1157 - 1166
Received 20 January 2021; Accepted 08 February 2021.

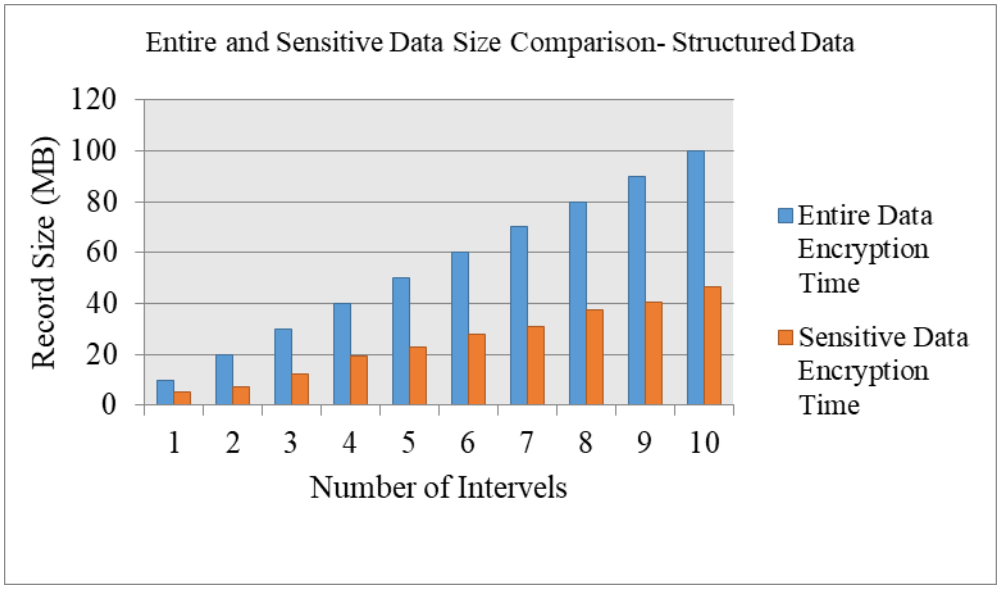

Figure 2. Comparison of entire attribute and sensitive attribute encryption - structured data

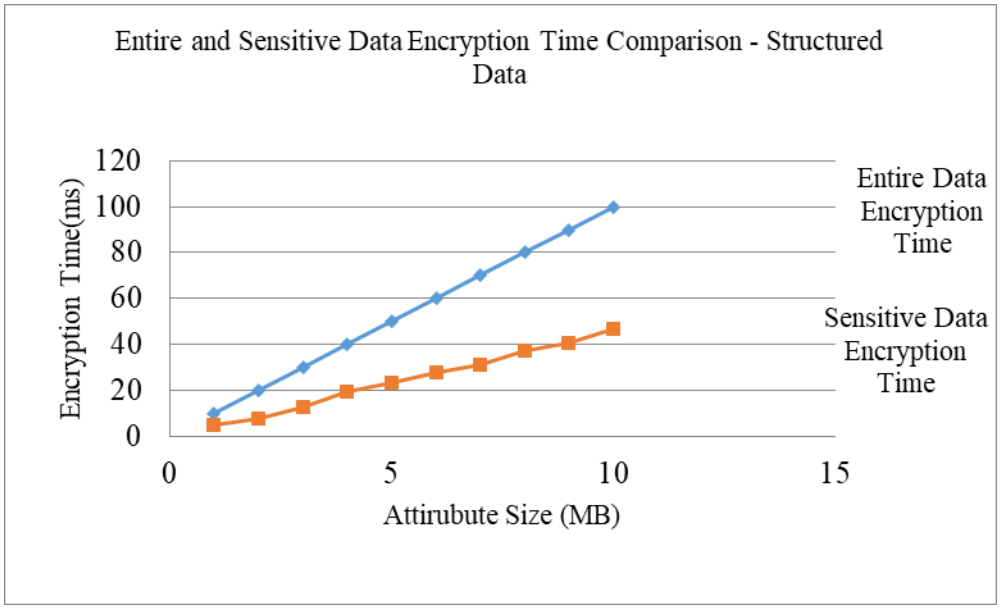

Figure 3. Entire and sensitive data encryption time comparison

Similarly the unstructured medical document is taken for sensitive attribute identification. The extracted sensitive information is encrypted instead of entire document. When compared to structured data, the common information in unstructured data is high. Such that, nearly 82% of information is common for all users and 18% of information is differ from user to user. Hence, 18% of data isencrypted instead of 100% data. Figure 4 shows the entire document size and S_D size of medical document. It clearly shows that the S_D occupies lesser than 200 document size in 1000 document size. This size reduction reduces processing time also. Figure 5 shows the encryption time of the entire document and S_D for unstructured document.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1157 - 1166
Received 20 January 2021; Accepted 08 February 2021.

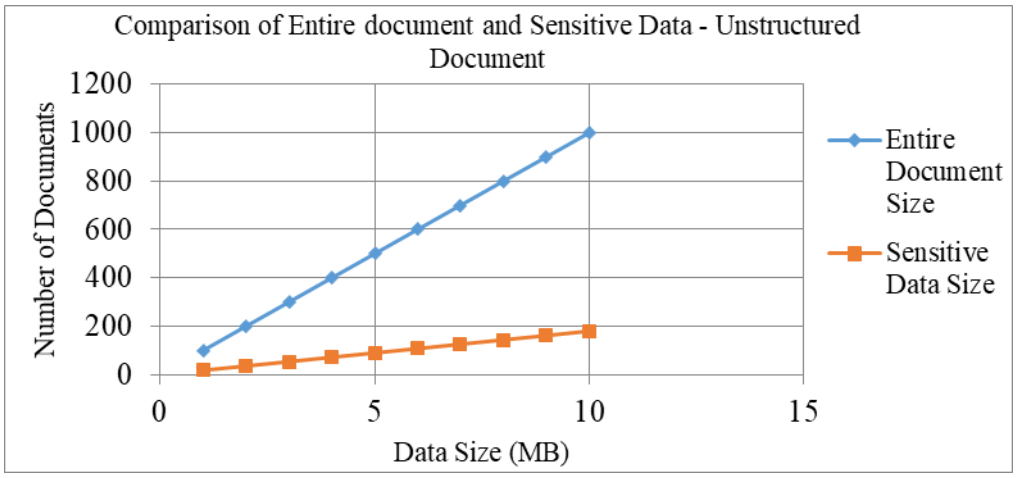

Figure 4 Comparison of entire document and sensitive data size - Unstructured document

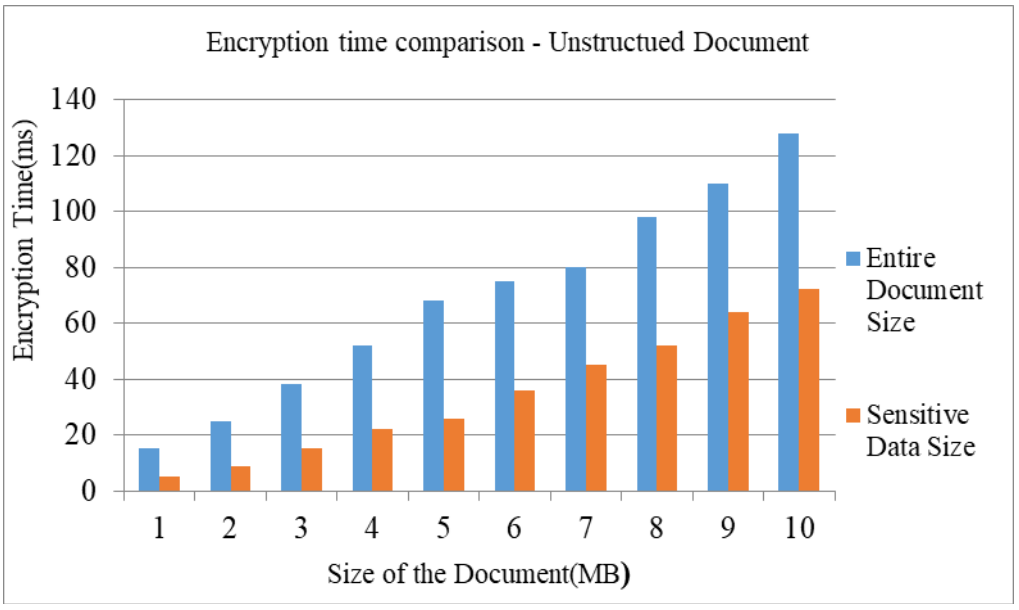

Figure 5 Comparison of Entire Document and Sensitive data encryption time - Unstructured Data

7. Conclusion

The sensitive data identification and protection technique in a structured and unstructured data in cloud based storage provides high protection to sensitive data with a tradeoff between security and processing time. The experimental results show that the proposed technique lesser processing time and encrypt only limited data. This technique provides better security to sensitive data and usability to non-sensitive data. Hence, this technique is applicable to large size data like big data applications. In our future work the same technique is going to be implemented in a big data related applications

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1157 - 1166
Received 20 January 2021; Accepted 08 February 2021.

References

[1] M.Sumathi and S.Sangeetha, “Survey on Sensitive Data Handling- Challenges and Solutions in Cloud Storage System”, Advances in Big data and Cloud Computing, PP 1-17, 2019.

[2] M.Sumathi and S.Sangeetha, “Scale based secured sensitive data storage for banking services in cloud”, International journal of Electronic Business, 14 (2), PP 171-188, Inderscience publisher, 2018.

[3] M.Sumathi, S.Sangeetha and Anu Thomas, “Generic cost optimized and secured sensitive attribute storage model for template based text document on cloud”, Computer Communication, Vol.150, PP 569-580, Elsevier publisher, 2020.

[4] M.Sumathi, R.Lekaa, R.Kavirakshana, N.Nishanthini, K.Nirmala, “Improved CiphertextAttribute Based Sensitive document protection and Secure sharing in cloud storage”, International Journal of Advanced Science and Technology, Vol. 29, No.03, PP 8702-8708, 2020.

[5] M.Sumathi and S.Sangeetha, “A group key based sensitive attribute protection in cloud storage using modified random Fibonacci cryptography”, Complex & Intelligent Systems, PP 1 -15, Springer publisher, 2020.

[6] Cecil Eng H. Chua, Roger H.L.Chiang, Ee-Peng Lim, "Instance-based attribute identification in database integration", The VLDB Journal (2003) 12.

[7] Tong Yi and Minyong Shi, "Privacy Protection method for multiple sensitive attributes based on strong rule", Journal of Mathematical Problems in Engineering, Hindawi Publishing Corporation, Vol.2015 PP 1 - 14.

[8] Cedric du Mouza, Elisabeth Metais, NadiraLammari, Jacky Akoka, Tatiana Aubonnet, Isabelle, "Towards an automatic detection of sensitive information in a database", Advances in database knowledge and database applications, 2nd International conference 2010.

[9] Kalyan Nagaraj, Sharvani GS and Amulyashrre Sridhar, "Encrypting and Preserving sensitive attributes in customer churn data using novel dragonfly based pseudonymizer approach", Journal of information, Vol. 2019, 10, 274 PP 1 -21.

[10] Razaullah Khan, Xiaofeng Tao, AdeelAnjum, Haider Sajjad, Saifur Rehman Malik, Abid Khan, and FatemehAmiri, "Privacy Preserving for multiple sensitive attributes against fingerprint correlation attack satisfying c-Diversity", Wireless Communications and Mobile Computing, Vol 2020, PP 1- 18.

[11] T.M.Nithya, J. Ramya, L. Amudha,“Scope Prediction Utilizing Support Vector Machine for Career Opportunities”, International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249- 8958, Volume-8 Issue-5, June 2019, pp.2759-2762.

[12] L. Amudha, Dr.R.PushpaLakshmi, “Scalable and Reliable Deep Learning Model to Handle Real-Time Streaming Data”, International Journal of Engineering and Advanced Technology, ISSN: 2249 – 8958, Volume-9 Issue-3, February, DOI: 10.35940/ijeat.C6272.029320, 2020, Retrieval Number: C6272029320/2020©BEIESP, pp. 3840 – 3844

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1157 - 1166
Received 20 January 2021; Accepted 08 February 2021.

[13] T.M.Nithya, K.S.Guruprakash, L.Amudha. (2020). DEEP LEARNING BASED PREDICTION MODEL FOR COURSE REGISTRATION SYSTEM. International Journal of Advanced Science and Technology, 29(7s), 2178-2184

[14] Nithya, T.M., Chitra, S.. (2020). Soft computing-based semi-automated test case selection using gradient-based techniques. Soft Computing. 24. 12981–12987 (2020)

[15] K.S.Guruprakash, R.Ramesh, Abinaya K, Libereta A, Lisa Evanjiline L, Madhumitha B. (2020). Optimized Workload Assigning System Using Particle Swarm Optimization. International Journal of Advanced Science and Technology, 29(7), 2707-2714.