International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 05 | May -2017 www.irjet.net p-ISSN: 2395-0072

Two fish Algorithm Implementation for lab to provide data security with predictive analysis

Mrs. Anjana Devi[1], Mrs. Ramya B S[2]

[1]*Student (MTech.), Dept. of IS Engineering, SCEM, Mangalore*
[2]*Asst. Professor, Dept. of IS Engineering, SCEM, Mangalore*

Abstract - *An The increasing demand for clinical laboratories to provide an efficient management system for the patients test report and data is a formidable part of the healthcare industry. This has led the laboratories to devise a faster and efficient system that complies with the logistical and technical methods to operate faster and with much more accuracy and security providing immunity against cryptic attacks. Since even a medium scale laboratory requires large database to store the data of patients, the organization opts for cloud storage to make data storage and retrieval easier, simple and more secure. The proposed model stores patients report in a cloud database by implementing two fish algorithm, a block cipher for storing data by encryption and retrieving the same by decryption. The symmetric encryption algorithm Two-Fish is a 128-bit block cipher with a key length of 128, 192 or 256 bits just as the AES algorithm. The stored database is used for predictive analysis of test reports of the patients with possible diagnosis of the abnormal values of blood count. A part of complete blood count (CBC) parameters are considered as test data with ID3 decision tree analysis algorithm.*

Key Words: **Two fish, CBC, ID3, block cipher.**

1. INTRODUCTION

The increasing demand for clinical laboratories to provide an efficient test results of reports along with efficient management of the patients results and data[2] is a formidable part of the health system. This has led to the laboratories to devise a faster and efficient method which complies with the logistical and technical methods to operate faster and with much more accuracy and security providing immunity against cryptic attacks. Since even a medium scale laboratory requires large database to store data of the patients, the organization opts for cloud storage to make data storage and retrieval easier, simple and more secure. The proposed model stores patients report in a cloud database by implementing two fish algorithm for storing data by encrypting and retrieving the same by decryption.

Cloud computing is a technology that provides access to information and computing resources from anywhere that a network is available. Hence, there is a need to secure the patients data stored on the cloud. The proposed paper uses two fish encryption algorithm to protect integrity of the patients' reports against unauthorized attacks. However, for all cloud computing applications, performance and cost of implementation are also major concerns. And two fish algorithm provides the required security and performance so the encryption algorithm is balanced.

The ID3 decision tree algorithm is used on the decrypted data to carry prediction analysis of the blood count to give possible diagnosis list of the abnormal blood count parameters. The White Blood Cells (WBC), haemoglobin and Platelets are considered for analysis in the training dataset for prediction analysis.

2. TWO FISH ALGORITHM

A. Encryption: Two fish

The Two fish was first published in 1998 by the American cryptographer Bruce Schneier. The two fish algorithm is a type of block cipher that makes use of a key size of 128, 192 or 256 bits and a plaintext of 128 bits. Compared to Rijndael, Two fish is quite complex, but makes use of many similar functions. The basic process of Two fish is given as follows (depicted in figure 1).

i) The plaintext is broken up into four 32 bit words and each is XORd with a 32 bit expanded key(The first word is XORd with $KK0$, the second word with $KK1$ and so on).

ii) The first word is broken up into 4 bytes, each of which is applied to a substitution box (or S-box, like the lookup table mentioned in AES). The second word is first rotated left by 8 bits and then is also applied to the same set of S-boxes.

iii) From here both the first and second words are applied to an MDS matrix (Maximum Distance Separable) which serves

to diffuse the newly substituted data of the 32 bit word amongst its 4bytes.

iv) After the MDS matrix multiplication the first word is applied to a pseudo-Hadamard Transform:

$aa' = aa + bbmmmmmm\ 232$

where a is the first word, b is the second word and a' is the new first word.

Using the 'new' first word as input, the second word is applied to the same transform, which can equivalently be represented as:

$bb' = aa + 2bbmmmmmm\ 232$

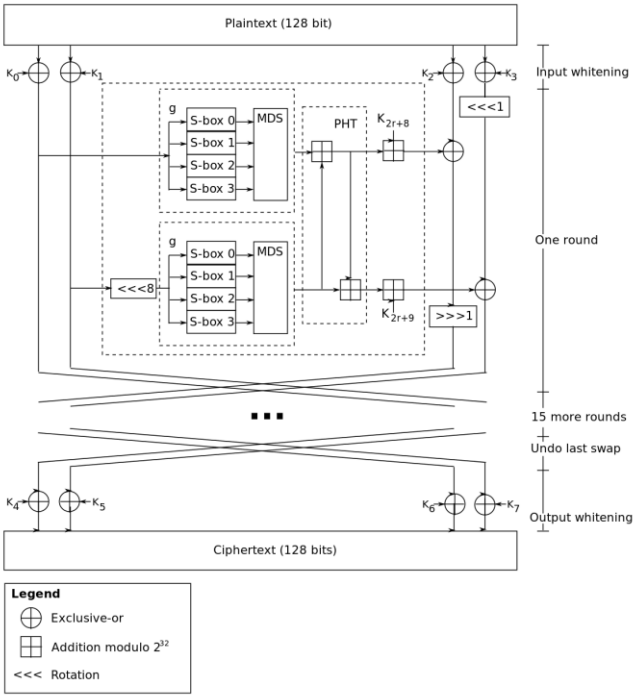

Fig. 1. Two fish algorithm implementation

This operation serves to diffuse the two words amongst each other.

v) At this point the first two words are XORd with a round key each.

vi) Following this step the third word is XORd with the output of the first word (the new first word or $WW0'$) and then rotated right by one bit producing what will be the new third word ($WW2'$). At the same time the fourth word is rotated left by one bit and then XORd with the output of the operations on the second word ($WW1'$) producing what will be treated as the new fourth word($WW3'$).

vii) The next round begins (starting at step 2) with the first and second words ($WW0$ and $WW1$) at the beginning of the previous round becoming this rounds third and fourth words while the new first and second words are the output from the previous round ($WW2'$ and $WW3'$respectively).

viii) Steps 2 through 7 are repeated for a total (including the first) of 16 rounds.

ix) The first and second words are swapped with the third and fourth words, effectively undoing the seventh step of the final round.10. The words are XORd with another set of round keys ($KK4 - KK7$) producing the cipher text.

An important note to make about Two fish however is that the S boxes used in the rounds (while the same 4 S-boxes for each round) are key dependant (dependant on the original source key, not the round keys)[3]. This adds an extra layer of security in that the S-boxes are now an unknown quantity to a would be attacker, making it much more difficult to crack a single round without knowledge of the key, since for two unique keys the primary substitution scheme will be different[6].

B. Cloud storage

Security is a major concern in cloud computing as our data is stored in a cloud and it becomes very difficult to perform operations on the encrypted data, hence we can use symmetric two fish encryption to secure our data and also perform operations on it[7]. With symmetric two fish encryption, a laboratory can encrypt its entire database of reports and upload it to a cloud. Then it could use the cloud-stored data as desired—for example, to search the database to understand how its workers collaborate. The results would be downloaded and decrypted without ever exposing the details of a single patient report.

The encrypted file is stored in cloud database and secure from external unintended modifications to the file. The encrypted file can later be retrieved as and when required maintaining the integrity of the data by a user with proper login credentials. The users can access data based on the permissions set by the administrator.

C. ID3 Decision tree algorithm

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983). The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric---information gain[9]. To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked (i.e. minimizing the depth of the tree). Thus, we need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function.

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree learning algorithms generate decision trees from training data to approximate solutions to classification or regression problems.

The decision tree in the proposed paper is applied for the blood count parameters to predict the possible diagnosis of the abnormal parameters whose values lie outside the normal range of cell count as predefined by the standards followed by the laboratories. The blood count parameters such as White Blood Cells(WBC), Platelets, Hemoglobin count are considered to give analysis report of the possible diagnosis for each individual abnormal values of the cells.

Pseudo code:

This pseudo code assumes that the attributes are discrete and that the classifications are either yes or no. It deals with inconsistent training data by choosing the most popular classification label whenever a possible conflict arises.

```
def id3(counts,classification_attribute,attributes):
        create a root node for the tree
        if all counts are normal/yes:
                return root node with normal/yes label
        else if all counts are out_of_range/no:
                return root node with out_of_range /no label
        else if there are no attributes left:
                return root node with most popular
   classification_attribute  label
        else:
                best_attribute=attribute from attributes that
        best classifies counts
                assign best_attribute to root node
                for each value in best_attribute:
                add branch below root node for the value
                  branch_counts=[counts that have that
                                  value for best_attribute]
                if branch_counts is empty:
                        add leaf node with most popular
                    classification_attribute label
                    else:
                    addsubtreeid3(branch_examples,
                  classification_attribute,
                     attributes-best_attribute)
```

If there's an attribute for the data to be split on, the algorithm calls itself recursively, with the original set of examples being split into groups based on the value of the best attribute and the set of available attributes to split on having the best attribute removed from it. Because this algorithm is a recursive one, the base cases: all examples having the same classification, no attributes being left, or no examples remaining, are tested first.

3. IMPLEMENTATION

The proposed paper is laboratory software for storing patients data in Mysql database which is stored in a cloud. The person authorized to view and edit data permissions are set by the administrator. The administrator sets the respective read write permissions for accessing database securely thus preventing unnecessary threat to the patient database. The large dataset of the patient requires cloud storage whose services can be availed based on the laboratory requirement whether small, medium or large scale laboratory instead of investing in data storage equipments which would further require manpower for maintenance and updation.

The software has been developed in PHP with algorithm implementation in Java. The database stored is encrypted using two fish algorithm and retrieved using decryption two fish algorithm which uses same key known for both sender and receiver. The encryption of the file is done as shown in fig 2 using 128 bit encryption. The encrypted file is stored in cloud database and secure from external unintended modifications to the file. The encrypted file can later be retrieved as and when required maintaining the integrity of the data by a user with proper login credentials. The decrypted file from cloud can be then used by researchers for study and analysis of blood count data in predictive analysis for various health sector research studies. This contributes to tremendous help by researchers to the community resulted by the data analytics.

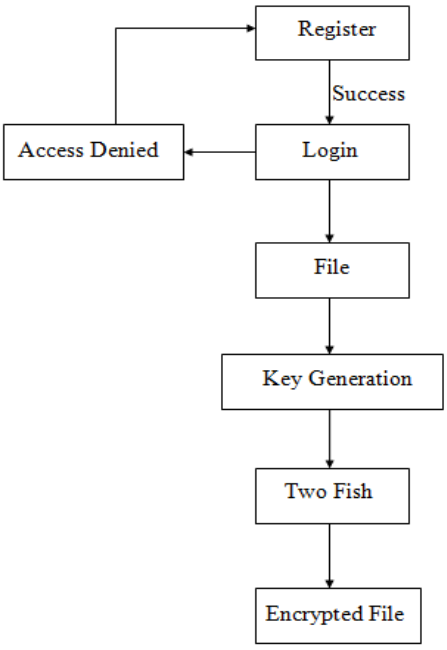

Fig 2. Two fish algorithm implementation to the patients' database

The generalized implementation is as shown in fig 3 with cloud storage making the system more secure as it reduces the frequent crashes of the local database and can be retrieved at any point of time.

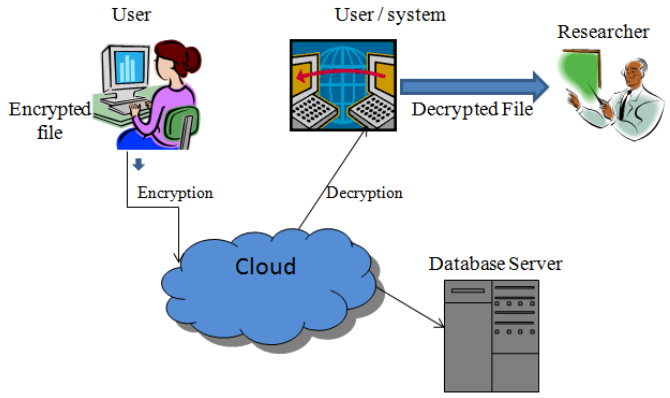

Fig 3. Implementing two fish algorithm in a lab system along with research analysis

4. CONCLUSION

Data Security is the major concern in laboratory software of any healthcare institution. To achieve security, various cryptographic algorithms are used to encrypt and decrypt the data. The chosen algorithm should provide best performance in securing database which the two fish algorithm implementation provides for patient database in the proposed paper. As analyzed, Two fish will give better performance than blowfish. Because as compared to Blowfish, Two fish is a 128-bit block cipher and uses at most 128-bit key. The prediction analysis of the dataset can be further carried out on complete blood count(CBC) so that all the parameters are considered in giving possible diagnosis whose study would be a boon to the society.

This work can also be extended to determine the performance of cloud in terms of throughput, power consumption and memory consumption. This work can also be extended to the future work in encryption and decryption of large size of text files, images, audio files and video files.

REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[3] R.anderson and E. Bihan, "Two practical and provably secure block cipher BEAR and LION", fast software encryption, third international workshop proceedings, springer – verlag, 1996, pp. 113-120

[4] U. Blumenthal and S. Bellovin, "A Better Key Schedule for DES like ciphers" pragocrypt "96 proceedings, 1996, pp.42-54.

[5] Tom M. Mitchell, (1997),Machine Learning, Singapore, McGraw-Hill.

[6] Tingyuan Nie and Teng Zhang , "A study of DES and Blowfish encryption algorithm", IEEE Region 10 Conference, 2010.

[7] K. Hwang and D. Li, "Trusted Cloud Computing with Secure Resources and Data Coloring", IEEE Internet Computing, Vol. 14, No. 5, pp. 14-22, 2010.

[8] Paul E. Utgoff and Carla E. Brodley, (1990), 'An Incremental Method for Finding Multivariate Splits for Decision Trees', Machine Learning:Proceedings of the Seventh International Conference, (pp.58),Palo Alto, CA: Morgan Kaufmann.

[9] Wei Peng, Juhua Chen and Haiping Zhou, of ID3,' An Implementation Decision Tree Learning Algorithm',University of New South Wales, School of Computer Science & Engineering,Sydney, NSW 2032, Australia. Quinlan, J.R. 1986, Induction of Decision trees, Machine Learning.