ISSN 2319 – 7595
Volume 8, No.2, March - April 2019
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs32822019.pdf
https://doi.org/10.30534/ijiscs/2019/32822019

SECURE CONFIDENTIALITY OF BIG DATA STREAMS USING SELECTIVE ENCRYPTION METHOD AND REGRESSION ALGORITHM

Soniya Joy[1], Neena Joseph[2]
[1] APJ Abdul Kalam Technological University, India, soniyajoy15395@gmail.com
[2]APJ Abdul Kalam Technological University, India, neena.joseph@mangalam.in

ABSTRACT

Security has become an important issue when the applications of big data are dramatically growing in cloud computing. The advantages of the implementation for these technologies have improved the service models and improve application performances in various perspectives. To secure the confidentiality of collected data, there is a need to prevent sensitive information from reaching the wrong people by ensuring that the right people are getting it. Sensed data are always associated with different sensitivity levels of confidentiality based on the sensitivity of the sensed data types. The confidentiality of collected data is improved by using a Selective Encryption (SEEN) method and a Regression algorithm. The Selective Encryption method is based on two key concepts common shared keys and a seamless key refreshment. This method can significantly improve the efficiency and buffer usage at DSM without compromising the confidentiality and integrity of the data streams.

Key words— Selective Encryption, Regression Algorithm, Confidentiality, Machine Learning

1.INTRODUCTION

Many applications such as smart health, smart cities such devices sense the deployed environment and generate a variety of data and send them to the server for analysis as data streams. A Data Stream Manager (DSM) collects the data streams (big data) to perform real time analysis and decision-making for these applications. An attacker may access the data in transit. The challenging task in such applications is to secure the confidentiality of the collected data. Securing the data trustworthiness requires that the system satisfy two key security properties: confidentiality and integrity. To secure the confidentiality of data here we apply machine-learning techniques. Secure the data and information at the time of transmission over the network using some cryptographic techniques. There are two methods are used to secure the data confidentiality, Selective Encryption (SEEN) and Regression algorithms. The Regression algorithm is used to predict the outcome of an event based on the relationship between variables obtained from the data set. Linear regression is one type regression used in Machine Learning. The problem describe here is when a sender sends a data to receiver there is a chance for disclosure of data, in which case confidentiality of data is lost which may lead to unauthorized access of data, misuse of data, modification of data, lost or theft of data. We focus on protecting the sensitive data from harmful intruders.

Machine Learning (ML) is a subset of artificial intelligence. Machine Learning algorithm is trained using a training data set to create a model. When new input data is introduced to the ML algorithm, it makes a prediction on the basis of the model. The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning algorithm is deployed. If the accuracy is not acceptable, the Machine Learning algorithm is trained again and again with an augmented training data set. The machine learning process is described in figure 1.

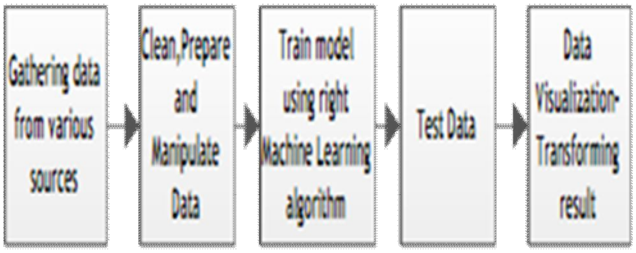

Gathering data from various sources
Clean,Prepare and Manipulate Data
Train model using right Machine Learning algorithm
Test Data
Data Visualization-Transforming result

Figure 1. Machine Learning Process

1. Gathering data from various sources - The very first and the most important step is to collect relevant data corresponding to our problem statement. Accurate data collection is essential to maintaining the integrity of our machine learning.
2. Clean, Prepare and Manipulate data - The data gathered from the previous step most probably is not fit to be used by our machine learning algorithm yet, as this data might be incomplete, inconsistent and is likely to contain many errors and missing values. After taking care of all the inconsistencies, errors and missing data in our dataset we move on to feature engineering. A feature is an attribute or property shared by all of the independent units on which analysis or prediction is to be done. Feature engineering is the process of using domain

knowledge of the data to create relevant features that make machine learning algorithms perform well.

3. Train Model - Model training is the process by which a machine learning algorithm takes insights from the training dataset and learns specific parameters over the training period that will minimize the loss or how bad it performs on the training dataset.
4. Test Data - After the model is trained, it is then evaluated, using some evaluation metric, on the test dataset. Hence, the model tries to perform on the test dataset using only the knowledge gained from the training dataset. Some of the most common evaluation metrics are Accuracy score, F1 Score, Mean Absolute Error (MAE) and Mean Squared Error (MSE).
5. Data Visualization - The performance of the model can further be improved on both the training and testing datasets using various techniques like, cross-validation, hyper-parameter tuning or by trying out multiple machine learning algorithms and using the one which performs the best or even better, by using methods which combine the results from multiple algorithms.

2. RELATED WORKS

Resource constrained sensing device are widely to build and deploy self organizing wireless sensor network for a variety of critical applications such as smart cities, smart health etc. A malicious adversary may access or tamper with the data in transit. One of the challenging tasks in such application is to assure the trustworthiness collected data so that any decision are made on the processing of corrected data [5]. The management of privacy and security in the context of data stream management systems (DSMS) remains largely an unaddressed problem to date. Unlike in traditional DBMSs where access control policies are persistently stored on the server and tend to remain stable, in streaming applications the contexts and with them the access control policies on the real-time data may rapidly change [6]. Cloud computing and big data analysis are gaining lots of interest across a range of applications including disaster management. These two technologies together provide the capability of real-time data analysis not only to detect emergencies in disaster areas, but also to rescue the affected people. This framework supports emergency event detection and alert generation by analysing the data stream, which includes efficient data collection, data aggregation and alert dissemination. One of the goals for such a framework is to support an end-to-end security architecture to protect the data stream from unauthorized manipulation as well as leakage of sensitive information [2]. We explore techniques to improve the performance of symmetric key cipher algorithms. Eight popular strong encryption algorithms are examined in detail. Analysis reveals the algorithms are computationally complex and contain little parallelism. Cryptographic processors would have to deliver orders of magnitude more performance to meet the bandwidth demands of secure servers and virtual private network (VPN) routers [10]. Present a mechanism for data aggregation in WSN that enforces both confidentiality and integrity of the aggregated data. The proposed mechanism is based on a novel application of peer monitoring and on a delayed aggregation of sensed data. The security of our scheme relies on the concept of additive homomorphic encryption and on a light weight key distribution technique [8].

Selective encryption technique is one of the most promising solutions to increase the speed of encryption as compared to the full encryption. Selective image after encryption becomes more secure against the attacks. Selective encryption is advantageous for the multimedia content like images, video and audio. Selective encryption is faster as compared to the full encryption of the data [3]. The typically large dimension of the input space, noisy signals, and the irregular training make context awareness a challenging problem to learning algorithms. The Kohonen Self-Organizing Map can be very unstable, especially in the initial untangling stage when the parameters that control the flexibility are high. To avoid overwriting, KSOM was extended with a second layer of labelled on-line k-means sub-clusters. The k-means clustering assures rapid vector quantization and a better safekeeping of memorized prototypes, while the KSOM gradually creates its topological map [9]. Here proposed a shared key synchronization method to ensure an end-to-end security in big data stream processing system consisting of distributed sensors and cloud-hosted stream processing engines (DSM). Proposed method synchronizes the shared key without communication between sensing devices and DSM, where sensing devices obtain the shared key reinitialization properties from its neighbours [7]. the proposed probabilistic selective encryption algorithm can be one of the most promising solutions in the area of information security which reduces the cost of data protection in wireless networks. The proposed approach is most suitable and gives better results when compared to other approaches. Thus, our solution delivers a realistic solution for protected wireless communication [4].

3. PROPOSED SYSTEM

The system can be introducing the hospital scenario using machine learning in hospital the data of each patient, blood donation detail, etc can be stored. An attacker may access the data stored in hospital is the important issue comes in this system. To avoid this type of attack, secure the confidentiality of data stored in the hospital. The system introduces the confidentiality of data Selective Encryption method are used. Then classify the result Regression Algorithm are used. The Selective Encryption (SEEN) method can be designed based on a symmetric key block cipher and multiple shared keys use for encryption. The cryptographic function with selective encryption, the Data Stream Manager (DSM) efficiently rekeys without retransmissions [1]. Selective Encryption can save the computational complexity. Apply different keys to encrypt the data packets for different data sensitivity levels. The SEEN method has some features they are:

- Efficient key broadcasting without retransmission
- Ability to recover the lost keys with a proper detection
- Seamless key refreshment without interrupting the data streams
- Maintain the data confidentiality based on the data sensitivity level

This method can be designed based on common shared keys which is initialized and updated by a DSM without requiring retransmission. And performs seamless refreshing of the shared key without disrupting ongoing data encryption or decryption. Then classify the results using a Regression algorithm, Linear Regression is used. Regression algorithms predict the output values based on input features from the data fed in the system. The algorithm builds a model on the features of training data and using the model to predict value for new data. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

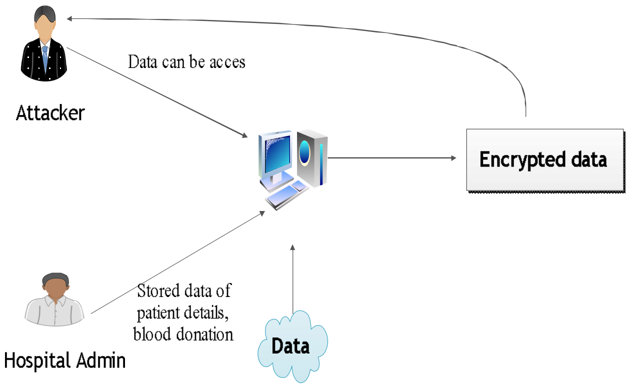


Figure 2. Overview of Proposed System

In figure 2 shows the overview of proposed system. The hospital admin can be stored the details of patient, blood donation details etc and the other data from cloud can be stored in a system. An attacker can be accessing these data stored in the system. In the system the data can be store very confidential. The encrypted data are stored in the system, so an attacker can be accessing the data the encrypted data format will give. So, it will secure the data stored in the system. Different types of attacks are coming under this system like blackhole attack, confidentiality attack etc.

4. EXPIREMENTAL RESULTS

In this section describe the comparison of some encryption method with the proposed encryption method called Selective Encryption (SEEN). In figure 3 shows the comparison of some encryption schemes. Compare the performance of SEEN security with advanced encryption standard (AES-128, AES-192), LSec and DPBSV. The results of the SEEN security method are about 81% is better than AES-128, AES-192 and LSec algorithms with different data packets. The performance of SEEN shows that it is little more efficient and littler faster than AES protocols. To secure the confidentiality are designed to prevent sensitive information from reaching the wrong people, while making sure that the right people can get it. Confidentiality in hospital is the right of an individual to have personal, identifiable medical information kept private. Such information should be available only to the physician of record and other health care and insurance personnel as necessary

In figure 4 shows the encrypted format of the data stored in the system. All data are stored in the system are encrypted format, so attacker does not give any data from the system. The data can be stored in this format in system an attacker can access the data from system attacker does not attack the data and does not destroy the data stored in the system. So, secure the confidentiality of data by using some encryption method and classify the output using some machine learning algorithm are used. Regression algorithm is used in Regression Linear Regression is used.

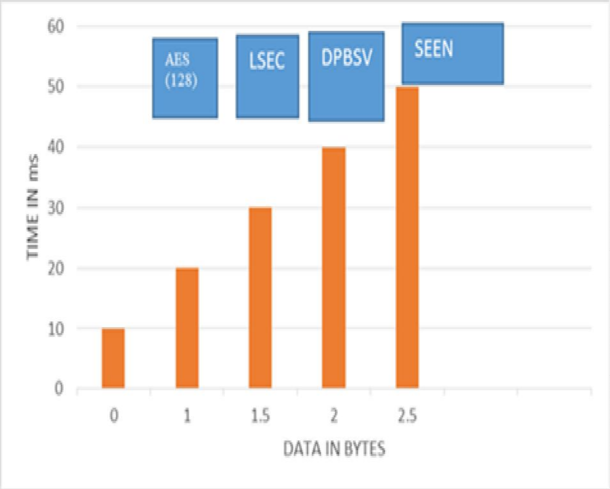
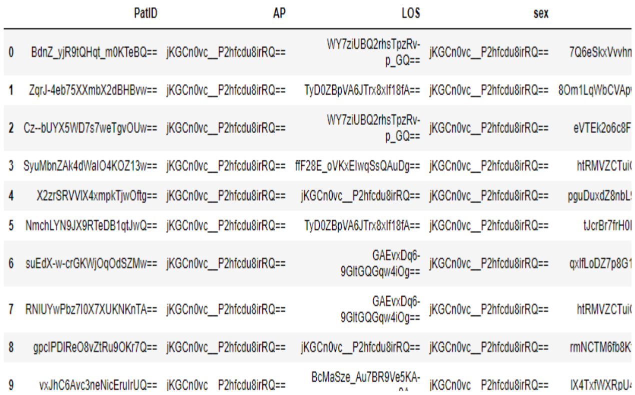


Figure 3. Comparison of AES, LSEC, DPBSV, SEEN



	PatID	AP	LOS	sex	
0	BdnZ_yjR9tQHqt_m0KTeBQ==	jKGCn0vc__P2hfcdu8irRQ==	WY7ziUBQ2rhsTpzRv-p_GQ==	jKGCn0vc__P2hfcdu8irRQ==	7Q6eSkxVvhn
1	ZqrJ-4eb75XXmbX2dBHBvw==	jKGCn0vc__P2hfcdu8irRQ==	TyD0ZBpVA6JTrx8xlf18fA==	jKGCn0vc__P2hfcdu8irRQ==	8Om1LqWbCVAp
2	Cz--bUYX5WD7s7weTgvOUw==	jKGCn0vc__P2hfcdu8irRQ==	WY7ziUBQ2rhsTpzRv-p_GQ==	jKGCn0vc__P2hfcdu8irRQ==	eVTEk2o6c8F
3	SyuMbnZAk4dWaIO4KOZ13w==	jKGCn0vc__P2hfcdu8irRQ==	ffF28E_oVKxElwqSsQAuDg==	jKGCn0vc__P2hfcdu8irRQ==	htRMVZCTui
4	X2zrSRVVlX4xmpkTjwOftg==	jKGCn0vc__P2hfcdu8irRQ==	jKGCn0vc__P2hfcdu8irRQ==	jKGCn0vc__P2hfcdu8irRQ==	pguDuxdZ8nbL
5	NmchLYN9JX9RTeDB1qtJwQ==	jKGCn0vc__P2hfcdu8irRQ==	TyD0ZBpVA6JTrx8xlf18fA==	jKGCn0vc__P2hfcdu8irRQ==	tJcrBr7frH0
6	suEdX-w-crGKWjOqOdSZMw==	jKGCn0vc__P2hfcdu8irRQ==	GAEvxDq6-9GltGQGqw4iOg==	jKGCn0vc__P2hfcdu8irRQ==	qxlfLoDZ7p8G1
7	RNIUYwPbz7I0X7XUKNKnTA==	jKGCn0vc__P2hfcdu8irRQ==	GAEvxDq6-9GltGQGqw4iOg==	jKGCn0vc__P2hfcdu8irRQ==	htRMVZCTui
8	gpclPDIReO8vZtRu9OKr7Q==	jKGCn0vc__P2hfcdu8irRQ==	jKGCn0vc__P2hfcdu8irRQ==	jKGCn0vc__P2hfcdu8irRQ==	rmNCTM6fb8K
9	vxJhC6Avc3neNicErulrUQ==	jKGCn0vc P2hfcdu8irRQ==	BcMaSze_Au7BR9Ve5KA-	jKGCn0vc P2hfcdu8irRQ==	lX4TxfWXRpU

Figure 4. The encrypted format of the data stored in the system

5. CONCLUSION

In this paper, we proposed a Selective Encryption (SEEN) method and Regression algorithms to maintain the confidentiality of big data streams. This method can be designed based on symmetric key block cipher and multiple shared keys use for encryption Regression Algorithm is Regression is used to predict the output. SEEN supports the source node authentication and shared key recovery without incurring additional overhead. In future, the system will increase the efficiency of symmetric key encryption. will implement deduplication identification mechanism for flexible storage mechanism.

REFERENCES

[1] Improved selective encryption method for IOT-BSN using stream classification adaptive model, Meena J, P. Balamurugan, P. Gayathri Devi.

[2] A Secure Big Data Stream Analytics Framework for Disaster Management on the Cloud, Deepak Puthal, Surya Nepal, Rajiv Ranjan, and Jinjun Chen.

[3] A Study on Different Approaches of Selective Encryption Technique, Saurabh Sharma , Pushpendra Kumar Pateriya.

[4] Implementing AES Algorithm for Selective Encryption in Wireless Networks, Manjula G, Dr. Mohan H.S.

[5] SEEN: A Selective Encryption Method With Time Based Access Control, Blessy Alex, R Sujitha

[6] A Security Punctuation Framework for Enforcing Access Control on Streaming Data, Rimma V. Nehme Elke A. Rundensteiner , Elisa Bertino

[7] A Synchronized Shared Key Generation Method for Maintaining End-to-End Security of Big Data Streams, Deepak Puthal, Surya Nepal, Rajiv Ranjan.

[8] Confidentiality and integrity for data aggregation in WSN using peer monitoring, Roberto Di Pietro1, Pietro Michiardi and Refik Molva2

[9] Combining the Self-Organizing Map and K-Means Clustering for On-Line Classification of Sensor Data, Kristof Van Lacrhoven

[10] Architectural Support for Fast Symmetric-Key Cryptography, Jerome Burke John McDonald Todd Austin
Advanced Computer Architecture Laboratory University of Michigan.