ISSN 2395-1621

Data Mining Using Secure Homomorphic Encryption

[#1]Raunak Joshi, [#2]Bharat Gutal, [#3]Rajkumar Ghode, [#4]Manoj Suryawanshi,
[#5]Prof U.H. Wanaskar

[1]raunakjoshi007@gmail.com
[2]b.gutal55@gmail.com
[3]rajghode2151@gmail.com
[4]manojsurya045@gmail.com
[5]ujwalaw.267@gmail.com

[#6]PVPIT Pune Maharashtra India
[#5]Student Assistant Professor Computer Enginering Department, Pune University
PVPIT Pune Maharashtra India

ABSTRACT

Data mining is the process that extracts, classifies and analyzes valid and useful information from large volumes of data provided by multiple sources. The data mining has been widely applied into various areas, one of which is to investigate potential security threats. Data mining is a computational tool widely used today which aims to extract useful information from various databases. Nowadays with the large volume of information that is produced, stored in a remote database concerns about confidentiality and privacy of information are arising due to lack of guaranteed security by storage service and the mining algorithm. A new approach of modern cryptography, defined as the homomorphic encryption , allows for the encrypted data to be arbitrarily computed which is a solution that aims to preserve the security, confidentiality and data privacy.

ARTICLE INFO

Article History

Received :4th April 2016

Received in revised form :

6th April 2016

Accepted : 8th April 2016

Published online :

9th March 2016

Keywords— Data Mining, Security, K-means, Encryption

I. INTRODUCTION

Data mining helps in extracting useful knowledge from large data sets, but the process of data collection and data dissemination may, however, result in an inherent risk of threats to confidentiality and data privacy. Some personal information about individuals, companies and organizations must be deleted before it is shared or published, unless such information is encoded. Thus, preserving privacy in data has become a very important issue. With the advancement in technology, industry, and research a large amount of data is being generated which is increasing at an exponential rate. Traditional Data Storage systems are not able to handle Data and also analyzing the Data becomes a challenge and thus it cannot be handled by traditional analytic tools. Data mining is the process of extracting valid and useful information from large quantities of data, analyzing the information and discovering useful patterns with different techniques. It has been applied into many different applications such as medical, health care, marketing, finance, privacy, security and so on.

II. LITERATURE SURVEY

Paper name: Performance of Ring Based Fully Homomorphic Encryption for securing data. .In this paper we have studied homomorphic encryption in database. Paper name: Homomorphic Encryption-based Secure SIFT for Privacy-Preserving Feature Extraction.In this paper we have studied Privacy Preservancy in Data Mining. Paper name: Survey on Recent Algorithms for Privacy Preserving Data mining In this paper we have studied we have studied privacy preserving recent algorithms in data mining. Paper name: Data Storage Security Using Partially Homomorphic Encryption.In this paper we have studied data security using homomorphic encryption Paper name: The Use of Fully Homomorphic Encryption in Data Mining with Privacy Preserving In this Paper we have studied use of homomorphic encryption in data mining.

III. PROPOSED SYSTEM

In this we use a novel methology a secure k-means algorithm in data mining approach assuming the data to be

distributed among different hosts preserving the privacy of the data. The approach is able to maintain the correctness and validity of the existing k-means generate the final results even in the distributed environment. A new approach of modern cryptography, defined as the Homomorphic Encryption allows for the encrypted data to be arbitrarily computed which is a solution that aims to preserve the security, confidentiality and data privacy. This system proposes methods that ensure the confidentiality and privacy in the mining of databases based on fully homomorphic encryption

The system follows a certain flow. This client uploads the data or submits the data. It then gets stored in the database. Then after performing data mining by applying k-means algorithm it gets distributed among different hosts and after applying MPC that is Multi Party Computation it gets encrypted and it gets stored again in database. If the user is the authorized user then he will be able to decrypt the data using the privacy key. Secure multi-party computation (also known as secure computation or multi-party computation/MPC) is a subfield of cryptography with the goal to create methods for parties to jointly compute a function over their inputs, and keeping these inputs private.

K-means Algorithm for data mining AES Algorithm for homomorphic encryption

The approach is able to maintain the correctness and validity of the existing k-means to generate the final results even in the distributed environment.

Let us consider a set S

where, S= {U, R, SER, D, N, C, K-means ()}

Here, S: System which includes: U: Set of Users Where U={U1, U2, U3 …, Un}

SER: Server.

R: Set of Request.

Where R= {R1, R2, R3……., Rn}.

D: Database with horizontal partitions(p1,p2).

N: Number of Cluster. (i.e. 2)

C: Set of Centroid.

Where C={C1, C2,..Cn}.

K-means (N): It is the algorithmic part of the system.

Where N is number of cluster i.e. 2.

IV. RESULT ANALYSIS

Correctness refers to the validity of the final result obtained or the outcome of the experiments performed using the proposed approach, on the same hardware and software platform as compared to the original or base approach. The correctness is checked by comparing the deviation of the results from the anticipated results. Security this parameter evaluates the proposed algorithm in terms of security i.e. the capability of the algorithm to prevent the attackers, with malicious intent, to gain access to the confidential user data and valuable information inferred from the data.Actual results display the proposed approach performs k-means clustering on a dataset which is horizontally partitioned and stored on two different locations. The approach first run locally then performs a joint computation on encrypted intermediate results so as to obtain complete result. It was observed that running secure k It was observed that running secure k means on the partitioned data with same parameters and computation environment as the original single party k-means, produced the same end results and same inference, thus, validating the correctness of the proposed approach.

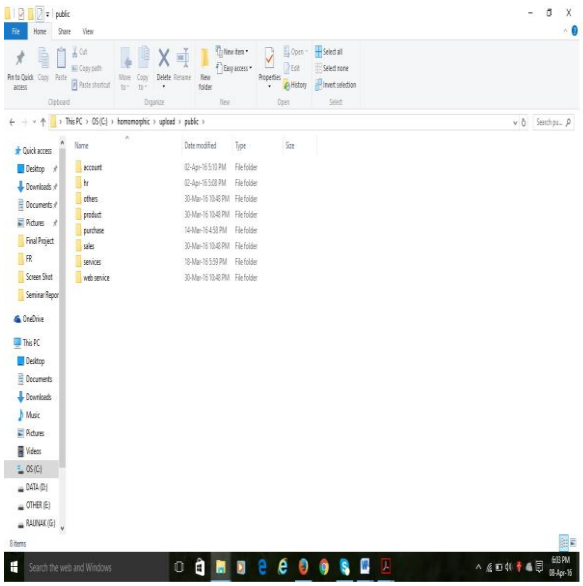

Fig 1: Different file Clusters

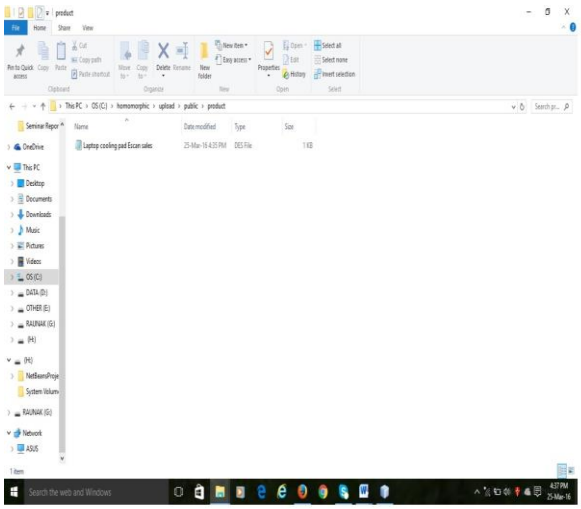

Fig 2: File clustering of data uploaded by user

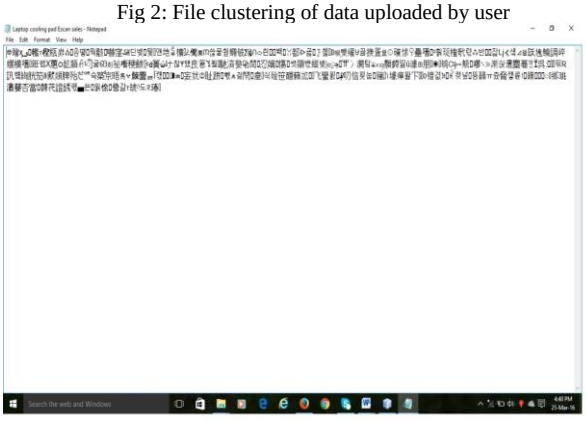

Fig 3: File encryption

Fig 4: File decryption

The above figures show the correctness of the proposed algorithm. It can be seen that the final cluster centers obtained by the merging of the clusters and the clustered points obtained in the final iteration of the two-party k-mean computation is similar to the cluster center obtained by the running of k-means algorithm single time. The correctness can be further seen as both the algorithms were run on same environment and platforms with same hardware and software configuration.

Thus, it is proved that the algorithm maintains the correctness and validity of the final result and thus can be applied to all situations where a single party k-means can be used. Coming to the security we know that the model uses a partitioned approach to store the large dataset. the dataset is fragmented horizontally thus, homomorhpic encryption is the first step towards the security against data mining based attacks as the intruder

which otherwise could, after getting an unauthorized access or entry to the data storage point, easily use the cheap and simple data mining techniques to extract valuable information from the data. But, as the data is encrypted and kept in clusters at different locations getting the correct information from the incomplete data becomes impossible thus fending off the attack by the adversary. Secondly, the assumed model is that of a semi-honest adversary .i.e. users try to leak the data of one another while maintaining their privacy. This approach deals with this threat as the intermediate results of both the party goes to a third party, and that too in an encrypted form, and it performs the computation on the encrypted data and returns the encrypted results to each party. Thus, each party only knows their intermediate values and the final value but not the data of the other party. Lastly, as the data goes to the third party encrypted with a key, if an intruder is able to pick the data in the transition he/she will not be able to decipher the encrypted data to get the original values and to simulate the approach with those values. This prevents Sniffing attack on the data-in transit.

V. CONCLUSION

Security and privacy are the major threats concerning the clients as well as of services as a lot of confidential and sensitive data is stored which can provide valuable information to an attacker. This proposes a method to solve the privacy issues of the database. It assumes that the user data is distributed on two hosts and performs a combined k-means clustering using the Homomorphic encryption system for security purpose so as to prevent any interpretation of intermediate results by an attacker. The proposed approach can further be extended by adding a digital signature or hashing technique to authenticate the third party so as to prevent an adversary from posing as the third party to host's. Also it can be generalized or extended to more number of hosts if required.

V. FUTURE SCOPE

It can be implemented in small scale or mid segment software industries. It can be used in big data and hadoop on a large scale platform. Industries with large database or big data having potential threat can be implemented.Where there is a risk of data privacy and preservancy can be deployed. In real time applications it can be deployed or simulated in cloud platform. The proposed approach can further be extended by adding a digital signature or hashing technique to authenticate the third party so as to prevent an adversary from posing as the third party to host's. Also it can be generalized or extended to more number of hosts if required.

REFRENCES

1. The Use of Fully Homomorphic Encryption in Data Mining with Privacy Preserving July 2014A. Laécio A. Costa1, B. Ruy J. G. B. de Queiroz
2. Security related data mining IEEE 2014 publications on computer and information technology
3. Q. Lu, Y. Xiong, X. Gong, and W. Huang. "Secure collaborative
4. outsourced data mining with multi-owner." 2012
5. IEEE 11th International Conference on Trust, Security and Privacy in
6. Computing and Communications (TrustCom) , IEEE, pp. 100-108, 2012.
7. S. Owen, A. Robin, T. Dunning, and E. Friedman. Mahout in Action.
8. Manning Publications, 2012.
9. R. Bhadauria, R. Borgohain, A. Biswas and S. Sanyal. "Secure
10. Authentication of Data Mining API " arXiv preprint arXiv:1204.0764, 2012.

11. R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina, "Controlling data outsourcing computation without outsourcing control." In Proceedings of the 2009 ACM workshop on security, pp. 85-90. ACM, 2009.

