Homomorphic Encryption for Secure Data Mining in Cloud
Prof. Vikas Maral¹, Sagar Kale², Ketan Balharpure³, Sourabh Bhakkad⁴, Pranav Hendre⁵
Professor¹, Student², ³, ⁴, ⁵
Department of Computer Engineering
Savitribai Phule Pune University
kalesagar1711@gmail.com², ketanbalharpure@gmail.com³, sourabh123459@gmail.com⁴, pranavhendre@gmail.com⁵

Abstract:
Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Big data is difficult to handle process and analyze using traditional approach. Using cloud services, we can resolve problems like resource sharing, storage capacity, and data transfer etc. But there is a main issue of data security and privacy while storing the big data on cloud. A major threat in Data Mining based attacks, allows an adversary or an unauthorized user to extract valuable and sensitive information by analyzing the results generated from computation performed on the raw data. In order to provide privacy, security for cloud user as well as cloud provider. We proposed a system for secure data mining using well known techniques like homomorphic encryption system, k means clustering, AES algorithm. In this process flow, cloud server is unaware of data uploaded by the user. And the client only gets the computational results. Through an experimental evaluation, we can maintain correctness and confidentiality of final result.

Keywords: cloud computing, Security, k-means, data mining, encryption.

I. INTRODUCTION
Cloud computing is a kind of Internet-based computing that uses the internet and central remote servers to maintain data and applications. Cloud computing allows consumers and businesses to use applications without installation and access their personal files at any computer with internet access. It frees a user from the concerns about the expertise in the technological infrastructure of the service. It allows end user and small companies to make use of various computational resources like storage, software and processing capabilities provided by other companies.

The cloud services can be divided into three categories: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Amazon, Microsoft, Google are some of the major cloud service providers. Google App Engine (GAE) is a type of PaaS provided by Google which allows web application hosting. Windows Azure, SQL Azure is some of the services offered by Microsoft providing processing and storage capabilities for large datasets. Amazon Web Services (AWS) including Simple Storage Service (S3), SQS, EC2 are cloud services provided by the Amazon. Thus convenience, on demand measured access, shared easily configurable computational resources, and self-service are some of the major characteristics of a cloud environment.

Problem statement:- The main goal of this article is to provide the security of the data, to take the backup of the data retrieve of the data.

II. EXISTING SYSTEM
Data mining can be a serious threat to the cloud security. Specially, to the organizations dealing with the financial, Governmental, education or legal issues of people. To maintenance of client privacy along with data privacy in cloud is a major area of concern for the cloud provider as well as cloud user.

III. PROPOSED SYSTEM
We are proposing a system in data mining on the given data using k-means clustering approach while maintaining the privacy
of the content at both the host and also preventing the intermediate values to be leaked to the adversary.

It is desired that the hosts know their inputs, the final outputs and no intermediate values and increased performance as compared to existing system.

In proposed system we are focusing on security of data on cloud server. For performance improvement we are using K-means as well as AES algorithms.

We are adding extra constrain along with cloud system i.e. Homomorphic Encryption. We can also proposed public key and private key.

The proposed approach uses the public key cryptosystems where M is the message or the plain text which is to be encrypted. The system can be divided into 3 parts (K,E,D):

- A pair of public and private key (l_k,p_k) is generated.
- A ciphertext or encrypted message c=Elk(m,r) is obtained where m € M and r is a random value.
- Decryption Dpk(c)=m is used to obtain plain text again.

[A] Private Data Normalization
A standard XML document is used to submit the data so that a data standard is maintained. But as we are dealing with multivariate database, i.e. a multi-attribute database, the value of variable is obtained as a sum of different attributes. Thus, the probability of some variables having large values is high, which can dominate the entire metric. Thus, a normalization method is used to standardize the multi-attribute data, using private mean computation of the data objects.

Let Host A has \( d_{A_i} \) with n data entries

And Host B has \( d_{B_j} \) with m data entries

Then mean

\[
M = \frac{\sum_{i=1}^{n} d_{A_i}^A + \sum_{j=1}^{m} d_{B_j}^B}{n+m}
\]

This mean is generated using Pallier Homomorphic cryptosystems so it also cannot be intercepted by the adversary. Now, the data is standardized locally using the above mean value as \( X_i = X - M \) for all data objects \( X_i \)

[B] Distance measuring and updation of clusters
After the standardization of the data a local k-means is performed by all host on their respective datasets and initializes the cluster center for each attribute and assign data objects to the nearest cluster center using Euclidean or Manhattan distance which can be chosen according to the application or database, i.e. \( h_{1A}, \ldots, h_{kA} \) for Host A and \( h_{1B}, \ldots, h_{kB} \) for Host B. As these cluster centers are calculated locally there is no need of any security protocol but in the next step of updating the cluster centers, joint centers are to be found which needs to be calculated privately.

Cluster Updation: for every data object’s values in the \( j^{th} \) attribute in \( i^{th} \) cluster, calculate sum as \( S_{i} = C_{i} + n_{j} \) where, \( n_{j} \) is number of data objects for \( j^{th} \) cluster \( S_{i} = C_{i} + m_{i} \) where, \( m_{i} \) is number of data objects for \( j^{th} \) cluster. Now, new \( i^{th} \) cluster center for \( j^{th} \) attribute is \( C_{i} = S_{i}^A + S_{i}^B / n_{j} + m_{j} \)

Pallier Homomorphic cryptosystem is used to do the above computations privately as: Host A, B and Third Party randomly generates a pair of public/private keys \( (l_k, p_k) \). Host A and B encrypts their sum value with Third Party’s public key and send it to Third Party along with their Public keys. As at the end of an iteration the local cluster centers are combined to get a global cluster center which is used by next local iteration, the correctness of algorithm stands true even in the distributed environment.

[C] Iteration Stopping Criteria
As it is known that k-means is iterative in nature, so theremust be a criteria which when met stops the iterations. This iteration stopping criteria is reached when output requirement are satisfied. For k-means this criteria is that the Euclidean distance between two consecutive cluster calculations is less then (threshold value).i.e. \( Dist(C_{i+1}, C_{i}) = Dist(\sum_{i=1}^{n_{A}} h_{iA}^A, \sum_{j=1}^{m_{B}} h_{jB}^B) < \) or \( Dist(\sum_{i=1}^{n_{A}} h_{iA}^A, \sum_{j=1}^{m_{B}} h_{jB}^B) < \) . To check this Host A computes Enc \( (\sum_{i=1}^{n_{A}} h_{iA}^A) \) and host B Enc \( (\sum_{j=1}^{m_{B}} h_{jB}^B) \) locally with third party’s public key. Then third party do multiplication of intermediate encrypted values and HOST A and B decrypt with their private key as follows:

\[ T = Dec([Enc((H_{1A}^A - H_{jB}^B)]\cdot Enc(H_{jB}^B - H_{jB}^B)]) \]

If T< , then the desired output is reached and the iterations can be stopped.

3.1 Overall Description
Let D= \{ \} be a multivariate database, where n is the number of attributes, which holds the user’s data. The Database is horizontally partitioned and stored at two locations .i.e. Host A and Host B. Host A has = \{ \} and Host B has = \{ \}. We want to perform data mining on the given data using k-means clustering approach while maintaining the privacy of the content at both the host and also preventing the intermediate values to be leaked to the adversary. It is
desired that the hosts know their inputs, the final outputs and no intermediate values.

[A] Encryption Formulas
To preserve the privacy of the data of each host and the intermediate results which are communicated to and fro we need an encryption system in which if any specific operation is performed on encrypted data or cipher text, the results generated matches the operation performed on plaintext when decrypted. This system of encryption is known as Homomorphic encryption system. For this purpose we use the Pallier cryptosystem which satisfies the need of the approach. We use $E(a).E(b)=E(a+b)$ and $E(a)b=E(a*c)$ in this approach, where $E$ is the required encryption scheme.

[B] Assumptions
A semi-honest model of adversary is assumed by the proposed approach in which a host can reveal other host’s data, if not secured, while maintaining the privacy of its own. This approach assumes that the data input by client is stored as chunks at different locations instead of storing whole of the data centrally, as, the centrally stored data is more vulnerable to the attacker. Thus the client’s data is stored in a decentralized manner by partitioning the database horizontally. Horizontal partitioning is referred to the partitioning scheme where each site has different records which contain same or equal set of attributes.

[C] Data Distribution
A multivariate relational database depicted as $D= \{d_1,d_2, \ldots, d_n\}$ in which Host A has $D_A=\{d_1^A, \ldots, d_1^A\}$ and Host B $D_B=\{d_1^B, \ldots, d_1^B\}$. As the database is multivariate, each data object is denoted by a vector set $d_i=\{x_{i1}, x_{i2}, \ldots, x_{im}\}$ where $m$ is the number of attributes. Now, let Host A have a set of private clustering centers $H_1^A, H_2^A, \ldots, H_k^A$ while Host B has $H_2^B, H_2^B, \ldots, H_k^B$ and $(C_1,C_2, \ldots, C_k) = \{H_1^A + H_1^B, \ldots, H_k^A + H_k^B\}$ as the joint cluster centers. Here, $k$ is the number of clusters.

[D] Proposed Algorithm
Notations:
- $C_i$ represents the combined clustering centers which is the sum of Host A and Host B’s share i.e. $H_A$ and $H_B$ respectively where $C_i = H_i^A + H_i^B$.

Input:
1) Database $D_A$ and $D_B$ belonging to Host A and Host B respectively having $n$ data objects.
2) ‘$k$’ which is the total number of clusters.

Output:
The $k$ cluster which is the combination of $D_A$ and $D_B$ or $D$.
1) Each party performs Data Normalization on local data.
2) Host A and Host B select their respective $k$ cluster centers $H_1^A, H_2^A, \ldots, H_k^A$ and $H_1^B, H_2^B, \ldots, H_k^B$ (locally) randomly.
3) Calculate or perform local k-means for Host A and Host B.
4) Save the cluster centers $H_{A,i}, H_{B,i}$.
5) Perform the secure cluster updation and reassign the data objects to their closest clusters locally
6) Save $H_{A,i+1}, H_{B,i+1}$, if the difference between the previous cluster center and the current one is less than or equal to threshold value then stop the iteration

VI. CONCLUSION
Security and privacy is the major issue concerning the clients as well as the providers of cloud services as a lot of confidential and sensitive data is stored in cloud which can provide valuable information to an attacker. According to author this method solves the privacy issues of the cloud. 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.

VII. ACKNOWLEDGEMENT
We should like to thank Prof. Vikas Maral, Assistant professor Computer Engineering department, in KJ College of Engineering And Management Research, Pune, Maharashtra, India for their useful guidelines.

REFERENCES


BIOGRAPHY
Mr. Sagar M Kale, Mr. Ketan Balharpure, Mr. Sourabh Bhakkad, Mr. Pranav Hendre are students of KJ College of Engineering and Management Research. Currently they are pursuing their B.E. in computer engineering. Their research interests are Computer Networks (Cloud Computing), Algorithms etc.