International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Secure Data Mining in Cloud using Homomorphic Encryption

Arya R

Student, Dept. of Dual Degree Computer Applications, Sree Narayana Guru Institute of Science and Technology N Paravur, Kerala, India

---

Abstract - *With the advancement in technology, industry, ecommerce and research an outsized amount of complex and pervasive digital data is being generated which is increasing at an exponential rate and sometimes termed as big data. Traditional Data Storage systems aren't ready to handle Big Data and also analyzing the large Data becomes a challenge and thus it can't be handled by traditional analytic tools. Cloud Computing can resolve the matter of handling, storage and analyzing the large Data because it distributes the large data within the cloudlets. No doubt, Cloud Computing is that the best answer available to the matter of massive Data storage and its analyses but having said that, there's always a possible risk to the safety of massive Data storage in Cloud Computing, which must be addressed. Data Privacy is during a ll|one amongst|one in every of"> one among the main issues while storing the large Data in a Cloud environment. Data processing based attacks, a serious threat to the info, allows an adversary or an unauthorized user to infer valuable and sensitive information by analyzing the results generated from computation performed on the data. This thesis proposes a secure k-means data processing approach assuming the info to be distributed among different hosts preserving the privacy of the info. The approach is in a position to take care of the correctness and validity of the prevailing k-means to get the ultimate results even within the distributed environment.*

***Key Words*: Cloud computing, Security, k-means, data mining, encryption.**

1. INTRODUCTION

Cloud computing refers to the web-based computing, providing users or devices with shared pool of resources, information or software on demand and pay per-use basis. It frees a user from the concerns about the expertise within the technological infrastructure of the service. It allows user and little companies to form use of varied computational resources like storage, software and processing capabilities provided by other companies. The cloud services are often divided into three categories: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS)[2]. Amazon, Microsoft, Google are a number of the main cloud service providers. Google App Engine (GAE) may be a sort of PaaS provided by Google which allows web application hosting. Windows Azure, SQL Azure is a few of the services offered by Microsoft providing processing and storage capabilities for giant datasets [3]. Amazon Web Services (AWS) including Simple Storage Service (S3), SQS, EC2 are cloud services provided by the Amazon [1]. This paper presents an approach to mine the info securely using k-means algorithm from the cloud even within the presence of adversaries. This approach assumes that the info isn't stored during a centralized location but is distributed to varied hosts. This proposed approach prevents any intermediate data leakage within the process of computation while maintaining the correctness and validity of the info mining process and therefore the end results.

2. EXISTING SYSTEM

Of all the problems that bug the fashionable cloud computing the difficulty of security or data privacy is one among the most important concerns of the cloud providers also because the client. Maintaining the client privacy isn't only important to take care of the confidentiality of the sensitive and valuable data of user but also to take care of the reputation among customers for the cloud provider. The info stored within the cloud must be secured from inside also as outside the cloud attacker. On the within a client may face attacks from the opposite user like theft or Denial-of Service attacks. The within attacks are often prevented by using authentication at the time of access of the info. Virtualization is one other technique which will be used to stop the users from each other but this still cannot solve the matter completely as all the resources can't be virtualized and also the virtualized environments aren't completely error-free. Any error within the network virtualization environment may cause incorrect transfers or leakage of sensitive information within the communication process or maybe the irrecoverable loss of valuable data. On the within a cloud provider can also pose as a threat to users data if he/she misuses the info because the provider

is present at rock bottom layer of the safety and should skills to bypass the applied security techniques. Various encryption techniques are proposed within the literature to secure the info albeit the authorization fails and therefore the attacker line up of the info.

3. PROPOSED SYSTEM

This project proposes a secure data processing for a cloud based system using k-means clustering without losing data integrity and prevents the intermediate values from being leaked. This method also proposes a further access security by verifying the user who tries to access the stored data by means of an OTP number. Thus preventing unauthorized access of knowledge stored within the cloud. The given data is clustered by using k-means clustering approach and stored in multiple locations in cloud. These host locations must know their inputs, final output and no intermediate values. These clustered data must be encrypted. Here we are employing a Homomorphic encryption system during which if any specific operation is performed on encrypted data or cipher text, the results generated matches the operation performed on the plain text when decrypted. For this purpose we are using RSA (Rivest-Shamir-Adleman) encryption which satisfies the need .RSA may be a partially homomorphic crypto system.

It involves four steps: key generation, key distribution, encryption and decryption.

RSA involves a public key and personal key. The general public key are often known by everyone, and it's used for encrypting messages.

RSA is one among the primary practical public key crypto systems and is widely used for secure data transmission.

4. ENCRYPTION FORMULA

To preserve the privacy of the info of every host and therefore the intermediate results which are communicated to and fro we'd like an encryption system during 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 technique of encryption is understood as Homomorphic encryption system. For this purpose we use the Pallier cryptosystem [16] which satisfies the necessity of the approach. We use E(a).E(b)=E(a+b) and E(a)b=E(a*c) during this approach, where E is that the required encryption scheme.

4.1 ASSUMPTIONS

A semi-honest model of adversary is assumed by the proposed approach during which a number can reveal other host's data, if not secured, while maintaining the privacy of its own. • This approach assumes that the info input by client is stored as chunks [12] at different locations rather than storing whole of the info centrally, as, the centrally stored data is more susceptible to the attacker. Thus the client's data is stored during a decentralized manner by partitioning the database horizontally. Horizontal partitioning is mentioned the partitioning scheme where each site has different records which contain same or equal set of attributes

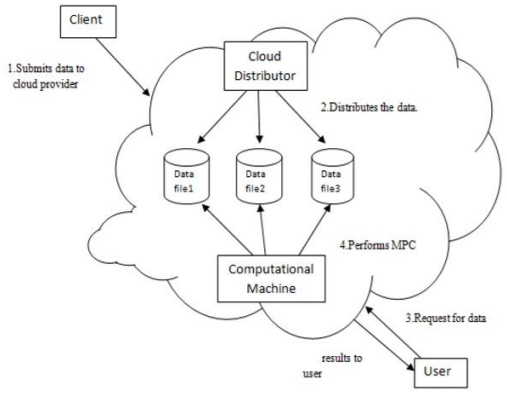


Fig 1: System architecture

5. ALGORITHAM

Notations: C,i represents the combined clustering centers which is the sum of Host A and Host B's share i.e. HA and HB respectively where Ci= HA+ HB.

Input: 1) Database DA and DB 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 DA and DB or D.
1) Each party performs Data Normalization on local data.

2) Host A and Host B select their respective k cluster centers H1A, H2A,……., HKA and H1B, H2B,……..,

HKB(locally) randomly. (C1,C2,……,Ck)={ H1A + H1B ,……, HkA + HkB }.

3) Calculate or perform local k-means for Host A and Host B.

4) Save the cluster centers HjA,i, HjB,i .

5) Perform the secure cluster updation and reassign the data objects to their closest clusters locally

6) Save HjA,i+1, HjB,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 else repeat step 4 onwards.

6. DETAILED APPROACH

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 (lk,pk) 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.

7. IMPLEMENTATION

Implementation means converting a new design into operation. During implementation there has to be a strong interaction between the developer and the users. This is the phase where the new system is given full chance to prove its worth and to minimize the reluctance to change. The proposed system may be entirely new, replacing an existing one or it may be modifications to the existing system. In either case, proper implementation is necessary to provide a reliable system to meet organizational requirements

7.1 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 dA=∑= n i 1 diA with n data entries And Host B has dB=∑= m i 1 diB with m data entries Then mean 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-M for all data objects.

7.2 UPDATAION 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. ,……, for Host A and , 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.

7.3 CRITERIA STOPPED

As it is know that k-means is iterative in nature, so there must 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)

8. RESULT AND ANALYSIS

8.1Evaluation Parameters

8.1.1 Correctness

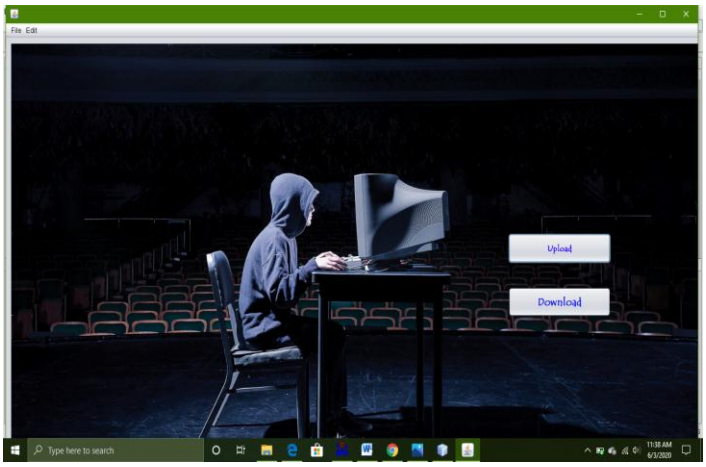
Correctness refers to the validity of the ultimate results obtained or the result of the experiments performed using the proposed approach, on an equivalent hardware and software platform as compared to the first or base approach. The correctness is checked by comparing the deviation of the results from the anticipated results.

8.12. Security

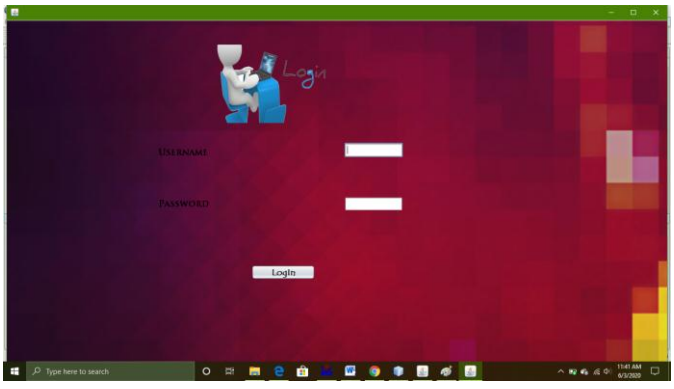
This parameter evaluates the proposed algorithm in terms of security i.e. the potential of the algorithm to stop the attackers, with malicious intent, to realize access to the confidential user data and valuable information inferred from the data .

B. Results

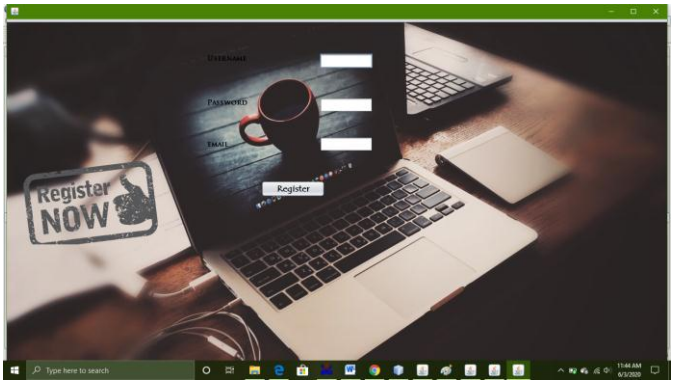
The proposed approach performs k-means clustering on a dataset which is horizontally partitioned and stored on two different locations. The approach first run locally then performs a joint computation on encrypted intermediate results so on obtain complete result. it had been observed that running secure k means on the partitioned data with same parameters and computation environment because the original single party k-means, produced an equivalent end results and same inference, thus, validating the correctness of the proposed approach.



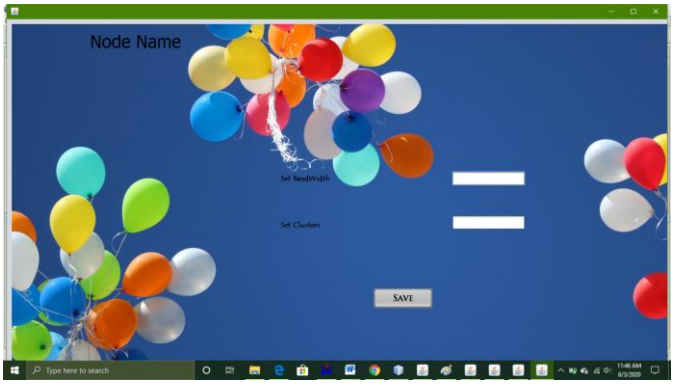
Client/user page



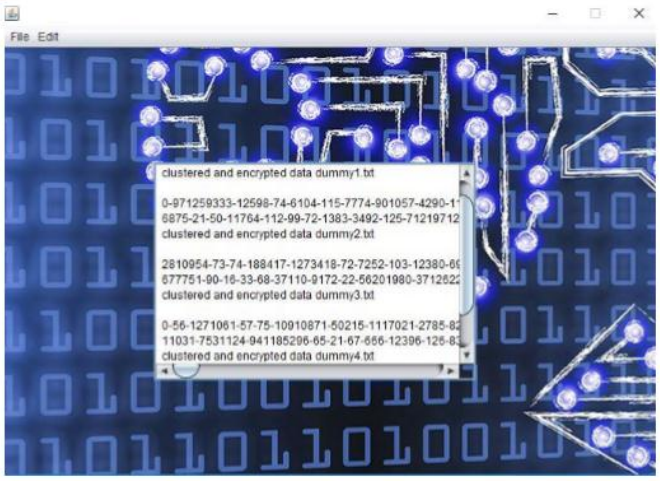
User login page



User registration page



Set parameters



Clusterd data

9. CONCLUSIONS

Security and privacy is that the major issue concerning the clients also because the providers of cloud services as tons of confidential and sensitive data is stored in cloud which may provide valuable information to an attacker. This paper proposes a way to unravel the privacy problems with the cloud. It assumes that the user data is distributed on two

hosts and performs a combined k-means clustering using the Pallier Homomorphic encryption system for security purpose so on 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 on prevent an adversary from posing because the third party to host's. Also it are often generalized or extended to more number of hosts if required.

FUTURE SCOPE

The above approach prevents the data leakage to an adversary if he/she intercepts the data in the middle of communication. But, as a Third Party is used to make communication possible between the two Hosts, an adversary can pose as an imposter posing as a Third Party and can get the data from both the Hosts. To prevent this, a digital signature or hashing technique can be incorporated in the algorithm to prevent an adversary to pose as the third party

REFERENCES

[1] Deepti Mittal, Damandeep Kaur, Ashish Aggarwal."Secure Data Mining in Cloud using Homomorphic Encryption" 2014 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM).

[2] Raunak Joshi, Bharat Gutal, Rajkumar Ghode, Manoj Suryawanshi, Prof U.H. Wanaskar."Data Mining Using Secure Homomorphic Encryption", International Journal of Advanced Research in Computer and Communication EngineeringVol. 4, Issue 10, October 2015

[3] Sneha Sakharkar, Shubhangi Karnuke, Snehal Doifode, Vaishnavi Deshmukh "A Research Homomorphic Encryption Scheme to Secure Data Mining in Cloud Computing for Banking System" 2018 IJSRSET Volume 4

[4] Mohit Marwaha, Rajeev Bedi "Applying Encryption Algorithm for Data Security and Privacy in Cloud Computing" IJCSI 2013 [

BIOGRAPHIES

Arya r
D/O Radhakrishnan
DDMCA student at
SNGIST.N Paravur, Ernakulam,
Kerala, India