EFFICIENT KNOWLEDGE DISCOVERY IN CLOUD USING K-MEANS CLUSTERING AND HOMOMORPHIC ENCRYPTION

[1]Mr.Abishek M John, [2]Subin Omanakuttan
[1]PG Scholar, [2]Assitant Professor
[1]Department of Computer Science,
[1]College Of Applied Science (IHRD), Mavelikara, Kerala, India

Abstract: In this modern world, storing of data is a major issue. Terabytes and terabytes of data are produced every day. This rapidly increasing data is called big data. These extremely large or complex data sets cannot be handled by a traditional data processing application. One of the best solutions for this problem is to use a cloud system. Cloud Computing offers a number of benefits and services to its customers who pay for the use of hardware and software resources (servers hosted in data centers, applications, software...) on demand which they can access via internet without the need of expensive computers or a large storage system capacity and without paying any equipment maintenance fees. But there is a main issue of data security and privacy while storing the big data on cloud. A major issue in Data Mining based attacks, is the entry of an unauthorized user to extract valuable information by analysing the raw data. This paper proposes an efficient technique for securely mining the data by means of k-means clustering and using a Homomorphic encryption for extra security of data in a cloud system. 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 the correctness and confidentiality of final result.

IndexTerms **- Cloud Computing, Security, K-Means, Data Mining, Encryption**

I. INTRODUCTION

Cloud computing is a computing paradigm, where a large pool of systems are connected in private or public networks, to provide dynamically scalable infrastructure for application, data and file storage. With the advent of this technology, the cost of computation, application hosting, content storage and delivery is reduced significantly. Cloud computing is a practical approach to experience direct cost benefits and it has the potential to transform a data center from a capital-intensive set up to a variable priced environment. 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).Cloud computing presents many unique security issues and challenges. In the cloud, data is stored with a third-party provider and accessed over the internet. This means visibility and control over that data is limited. It also raises the question of how it can be properly secured. It is imperative everyone understands their respective role and the security issues inherent in cloud computing. Despite the numerous benefits of cloud computing, only 33% of companies have a "full steam ahead" attitude toward adopting the cloud. That's according to a survey of over 200 IT and IT security leaders by the Cloud Security Alliance (CSA), which identified issues holding back cloud projects. Chief among them, companies are worried about how secure their data is once it leaves the company's firewall. These days, there are news headlines about data breaches and software vulnerabilities every day.

Cloud service providers treat cloud security risks as a shared responsibility. In this model, the cloud service provider covers security of the cloud itself, and the customer covers security of what they put in it. In every cloud service—from software-as-a-service (SaaS) like Microsoft Office 365 to infrastructure-as-a-service (IaaS) like Amazon Web Services (AWS)—the cloud computing customer is always responsible for protecting their data from security threats and controlling access to it

.

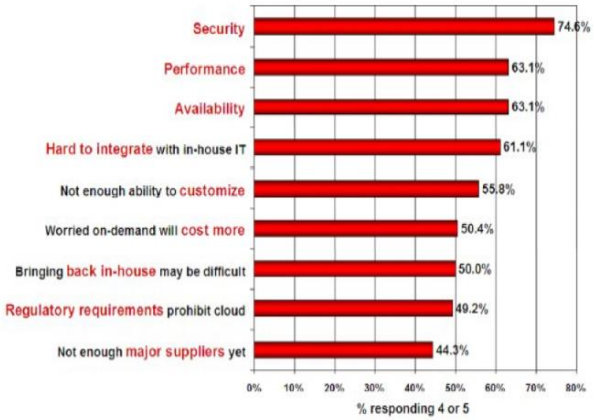

Fig 1: issues in cloud computing

This paper presents an approach to mine the data securely using k-means algorithm from the cloud even in the presence of these issues. This approach assumes that the data is not stored in a single location but is distributed to various hosts. This proposed approach prevents any intermediate data leakage in the process of computation while maintaining the correctness and validity of the data mining process and the end results. For extra security the clustered data is encrypted using Homomorphic encryption.

Homomorphic encryption is the conversion of data into cipher text that can be analysed and worked, as if it were still in its original form. Homomorphic encryptions allow complex mathematical operations to be performed on encrypted data without compromising the encryption. In mathematics, homomorphic describes the transformation of one data set into another while preserving relationships between elements in both sets.

II. Related Work

For preserving privacy of data mining, the researchers implement different types of techniques. They develop different types of data mining algorithms which help to keep the privacy and security of the data in big data.
While maintaining the correctness of the algorithm, security of two party k-means is need to be improved. Some PPDM (Privacy Preserving Data Mining) methods are k-anonymity, noise transformation and multiplicative transformation. Compared to PPDM the proposed system is more efficient and effective way for preserving the privacy of data mining.
Different types of modified cryptography techniques and trusted computing can be used for privacy. Data mining attacks in cloud can be classified into 3 levels: network level, application level and virtualization level .To solve the network level attacks IBM SCE developed a new technique which is known as "security as a service". It can protect high level security attacks. Cryptographic method is a best way for preserving privacy but cryptography cannot alone handle the security mechanism. Fragmentation technique or partitioning of database onto chunks is another best method for privacy. Here these nodes will prevent the intruder from accessing the data completely.
K-anonymity in a multi cloud environment can be used to perform frequent pattern mining. In this method the distributed data or a multi cloud environment prevents the attacker from accessing the complete data. Another method is one time pass key mechanism.it also protects the privacy of user and the service provided

III. Proposed System

This thesis proposes a secure data mining 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 an additional access security by verifying the user who tries to access the stored data by means of an OTP number. Thus preventing unauthorized access of data stored in 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 using a Homomorphic encryption system in 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 requirement .RSA is a partially homomorphic crypto system. It involves four steps: key generation, key distribution, encryption and decryption. RSA involves a public key and private key. The public key can be known by everyone, and it is used for encrypting messages. RSA is one of the first practical public key crypto systems and is widely used for secure data transmission.

Let n=pq where p and q are primes. Pick a and b such that
$ab \equiv 1 \pmod{\phi(n)}$
n and b are public while p and q are private.
$e_k(x)=x^b \bmod n$
$d_k(y)=y^a \bmod n$
the Homomorphism: Suppose X_1 and X_2 are plain texts. Then,
$e_k(x_1)\, e_k(x_2) = x_1^b x_2^b \bmod n = (x_1\, x_2)^b \bmod n = e_k\,(x_1\, x_2)$

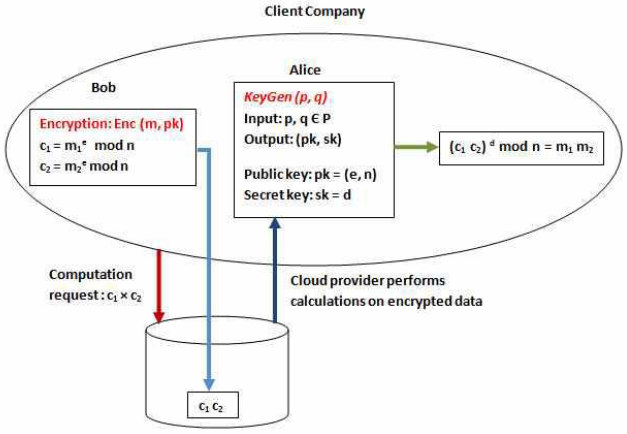

Fig 2: Multiplicative Homomorphic Encryption Applied to Cloud Computing

IV. Proposed Algorithm

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 $C_i = H^A + H^B$.
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 $H1^A$, $H2^A$,……..,Hk^A and $H1^B$,$H2^B$,……..,Hk^B(locally) randomly.
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 Hj,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

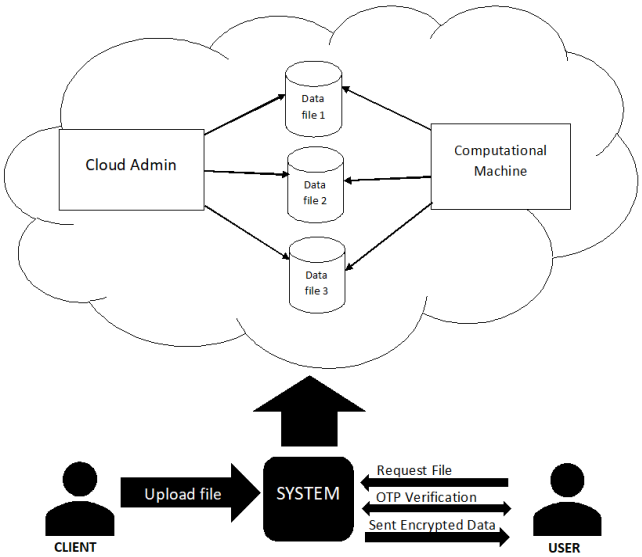

Fig 3: proposed system architecture

V. NORMALIZATION OF PRIVATE DATA

As we implement a multivariate database in place of standard XML document, the value of the variable is obtained as the sum of different attributes. Thus for large amount of data the value of variable will be high, which can dominate the entire matric. So a normalization method is used to standardize the multi attribute data. For this private mean computation is used.

Let Host A has $d_A = \sum_{i=1}^{n} d_i^A$ with *n* data entries

And Host B has $d_B = \sum_{i=1}^{m} d_i^B$ with *m* data entries

Then mean $M = \frac{d^A + d^B}{n+m}$

This mean is generated using RSA Homomorphic crypto system. Thus making the data unable to intercept by the attacker.

VI. DISTANCE MEASURING AND UPDATION OF CLUSTERS.

A local k mean is performed by all hosts on their respective data set after the standardization. Then initializes the cluster centers for each attribute. Each cluster center is assigned with data objects nearer to it. This distance between cluster centers and data objects are determined using Euclidean or Manhattan distance methods. These methods are chosen according to the application or database.

- ***Cluster updation***

  For every data object's values in the j^{th} attribute in i^{th} cluster, calculate sum as

$S_j^A = C_{ij} * n_j$ where, n_j is number of data objects for j^{th} cluster

$S_j^B = C_{ij} * m_j$ where, m_j is number of data objects for j^{th} cluster

Now, new i^{th} cluster center for j^{th} attribute is

$$C_{i,j} = {S_j^A + S_j^B}/{n_{j+} m_j}$$

Iteration stopping criteria

K means is iterative in nature, so they have some criteria to stop the iteration if the output is obtained. This criteria is that the Euclidian distance between two consecutive clusters must be less than ϵ threshold value. ie,

$$\text{Dist}\ (C_j, C_{j+1}) = \text{Dist}\ (H_j^{A,i+1} + H_j^{B,i+1}, H_j^{A,i} + H_j^{B,i}) < \epsilon \text{ or } (H_j^{A,i} + H_j^{B,i}) - (H_j^{A,i+1} + H_j^{B,i+1}) < \epsilon$$

VII. Analysis and Evaluation

Evaluation of the proposed system can be done basis of these two parameters

1. correctness of the algorithm

Here the accuracy of the output of the proposed algorithm is checked by performing different tests in the system. Then the output is compared with the current working system .the differences in the output are analysed.

2. Security

Here the security ability of the proposed algorithm is analyzed. How much security is provided by the algorithm against attacks on the system? Here the confidentiality, privacy and access rights of the stored data are analyzed.

VIII. Results

The proposed approach is based on k means clustering which is horizontally portioned and store the data in different location. First the data is clustered locally and then perform join computation on encrypted intermediate results and obtain complete results. The secure k means partitioned data with same parameters and produce same end result and same inference and also validating the correctness.

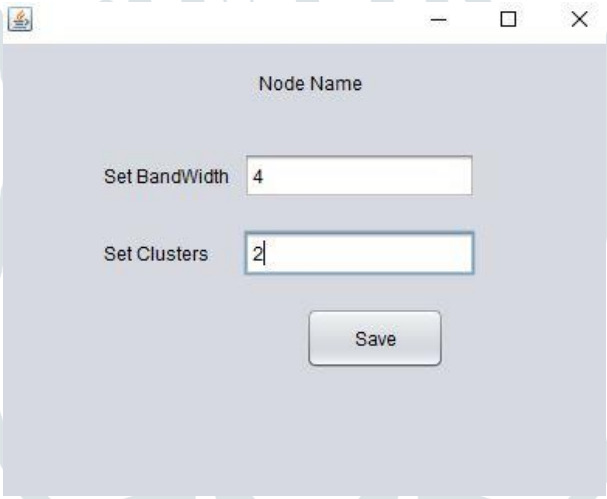

Fig 4: setting parameters

Fig 5: upload file by client

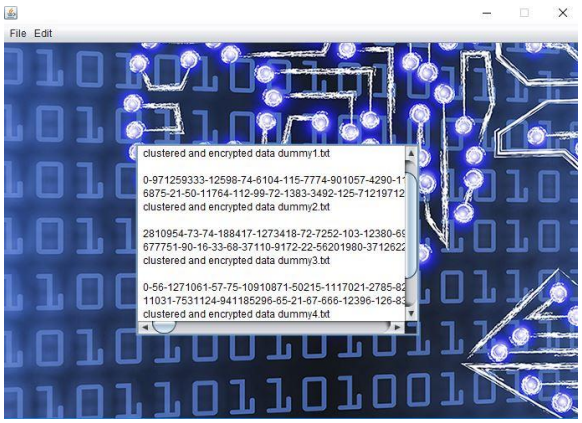

Fig 6: Resulting clusters

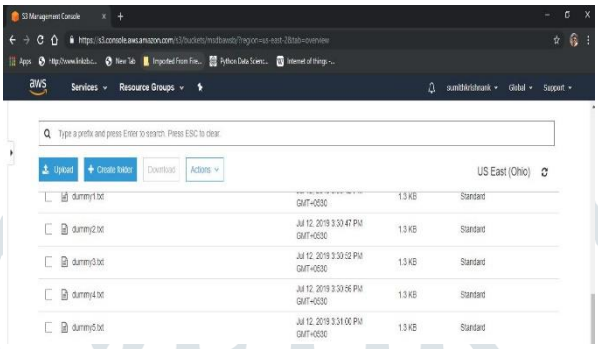

Fig 7: clusters residing in cloud

The figure shows the final clusters. We can see that the data is clustered and encrypted according to their proposed system. These clusters can be merged to obtain the final data. This proves the correctness and validity of the system. This proposed system can be applied to all single party k means situations.

To solve the security issue the fragmentation technique is used, that is the data set is fragmented horizontally and stored in different hosts. So if an intruder accesses the data set, he couldn't get the correct information from that data set. The second method is semi honest adversary that is the participant try to leak the data of one another while maintaining their privacy. So if a third party is trying to access the data, he needs to decrypt it and in the last approach the data goes to the third party encrypted with the key. Here if an intruder tries to pick the data in the transition he will not be able to decipher the encrypted data. This prevents sniffing attack on the data-in transit.

IX. Conclusion

The usage of cloud services in our day today life is increasing rapidly. Confidentiality and privacy of our data is the first priority. It assumes that the user data is distributed on two hosts and performs a combined k-means clustering using the RSA Homomorphic encryption system, for security purpose, so as to prevent any interpretation of intermediate results by an attacker.
The Security of Cloud Computing based on Homomorphic Encryption is a new concept of security which enables us to provide the results of calculations on encrypted data without knowing the raw entries on which the calculation was carried out respecting the confidentiality of data. Apart from this, with the usage of OTP verification of user we authorize each user who tries to access the data. 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 hosts.

X. Acknowledgment

We are thankful to our guide Prof. Subin Omanakuttan Assistant Professor in Department of Computer Science for his valuable support. Also we are thankful to the Department for the technical support.

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 Engineering Vol. 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

[5] Maha TEBAA , Said EL HAJII "*Secure Cloud Computing through Homomorphic Encryption"* IJACT Volume5, Number16, December 2013

[6] Prof.Vikas Maral, Sagar Kale, Ketan Balharpure, Sourabh Bhakkad, Pranav Hendre "*Homomorphic Encryption for Secure Data Mining in Cloud* " IJESC 2016

[7] Rinkal Patel "*Review on Data Security on Cloud using Homomorphic Encryption over Big Data*" IRJET April 2017