Journal of Advances and Scholarly Researches in Allied Education
Vol. XIV, Issue No. 2, January-2018, ISSN 2230-7540

Secure Cloud based Data Mining with Encryption and Digital Signature

Dr. Kamlendu Kumar Pandey[1]* Dr. Devendra Pandey[2]

[1,2] Assistant Professor, Department of Information and Communication Technology, Veer Narmad South Gujarat University, Surat, India

Abstract – As clouds are now containering most of world data and guarantee its safety with a service level agreement, more and more organizations are shifting their data on cloud. The technological infrastructure provided by the client allows the application to maintain the historical data and pour continuously streaming transactional data. This data after some time joins the data ware house after the suitable treatment. The gross data piled up over time need to be analyzed or various purposes. This is where Big Data technologies come in the picture. Organization uses various mining techniques on this Big Data to get relevant findings. The question here is how secure your big data is and how it safeguards against an adversary which access the data either from the data store or on fly.. In this paper the confidentiality of data is protected by k-means approach off data mining using digital signature and homomorphic encryption for the data which is distributed across various containers.

Keywords— Containerized Database, Security, Data Mining , Encryption

----------X----------

1. INTRODUCTION

The cloud has come up with a hue business value for all kind of digital resources, environments and processing. Almost all kind of resources are virtualized and customized for the needs of the user. The usage and pricing policy is different for different clouds. The clouds can store large amount of data and allow the user to use the data through variety of interfaces. This is possible by using service based technologies like IAAS (Infrastructure as a Service) and PAAS. (Platform as a service) . These services handle all the system lebvel and technology based constraints and allow the users to deploy their data seamlessly. It give them all kind of tools using which they can perform numerous operations on the data. All processing and storage is done by cloud. The cloud also show exceptional redundancy to replicate the data on various region based storage and can handle the failover without any problems to the users. Although everything seems to be in order but the users always have apprehensions to the security off data. The clouds claim to keep the data secure using current internet standards but organisation still want to play safe on data particularly when it is to be accessed for mining. The reason is a security hole in this access can make the data and its findings transparent to competitors and adversaries. The clouds are generally public and many masquereders are just in search of the data stores which can be compromised. The competitirs are generally intrested in the data of rival company so as to make new strategies. The clouds generally have multi fold security to protect the data but dat becomes vulnerable if some network operations like aggregate queries or relationship operations are carried on them. As the data in that travels on network encountering various routers, if not protected can be heard and captured by the eaves droppers or hackers. In this paper we propose only encryprion on the data using the homomorphic encryption which is present in the literature.

2. RELATED WORK

The data mining is a serious operation to be performed by the users who are part of the responsibility chain in the organisation. The situation here is diffrent than the conventional mining. Here we need to extract the information from data stores which are deployed on cloud in various patterns[1]. The storage of the data can be at one place or distributed, that nmeans several chunks will be placed at different location. When we perform mining the data is extracted from all these sources[3][4]. Several researchers have been working in this area have proposed various methods to deal with it. Researchers generally classify the datastores as container based, vitualized and cloud based clusters from which the data is to be mined. Non only the content of data but the privacy of the user involved is also important. All the datamining application must follow adequate authentication and authorization mechanisms to deal with the users[22]. The attack on datastores and onfly queries can be categorized

as sniffing, masquerading, trojanic, Denial of Service or man in the middle attacks which adversary will be using. The security concerns can be classified as that on the level of application , in network and vartualized docker based containers[23] . Companises like IBM and Oracle have implemented Software as Service to handle such attacks. In [24] the auther discusses the security concers which can occur at the application level. The schemes like k-anonymity [10][11] has been proposed for multi cloud data mining [25]. The new technologies have come with other facilities which can avoid costly virtualization. The container based technology allows the databases and ware houses to run in a docker based container. The mining candidates which are software application or python libraries too can be deployed in separate containers. This is a much safer mechanism to be deployed on cloud . All big players have the facility to do so. The AWS has Elastic Container System to deploy container images while Elastic Kubernetes Engine which is container orchestrator can be maintain the container clusters . The EKEs are having a key based mechanism to protect the data in the inter container travel. The Kubernetes create a own virtual network and containers join this network in secure manner. The paper propose to have a encryption module as a part of service in the cloud environment being execcuted under kbernetes engine. This paper deals with the container based datastores where the data is present in more tahn one docker container and the both the containes are deployed in different cloud based Kubernates engine. The datamining application is using KNN based also called k-nearest neighbor . It also uses Support vector mechine as well as k-means to achieve the desired encryption to the query calls and processing calls . A homomorphic encryption scheme is used in the whole process. The secure data mining is a state of art and efforts are done by researchers to have a commin strategy for protection against attacks. Digital signatures a major role in establishing privacy of the user as it comprises of the public – private key pair. The operation in that case can be encrypted as well as authenticated and authorized. There is a very stiff competition amongst client service provider. Id a cloud give an assurance of security as in case of data mining through multiple containers then it will have an edge.

3. THE METHODOLOGY

In this approach we consider entire data warehouse whose data is separated on two different container on the cloud. This can be written as

Let W= { w1 .. wn}

where W be a clustered warehouse, where n is number of properties representing the multi-dimenssional data.

The warehouse is is deployed in two docker containers as X and Y .i.e. container X and container Y. X has Hx= {w_1^x … w_n^x} and container Y has h= { w_2^x .. w_2^y}.

The mining has to be performed on the containers spanning these two data containers in a secure way. The service will send the input value and fetch final results . All values fetched in between will not be known to the attackers

1. *Encryption Formulas*

The encryption formulae must confirm to the privacy of the data store and all the communications and operations doen on the encoded data must give the same result as it is done in the case of plain text operations. This approach of security is called homomorphic encryprion.

We are using Pallier cryptosystem which suits our requirement for secure retrieval of data . The popular equation S(m).S(n)=S(m+n) and $S(m)^n$=S(m*p) in this , where S is the secure way of encryption required for the purpose.

2. *Notations to represent data*

A NoSQL based database having multi-dimenssional properties in form of JSON is shown as

W= {w_1, w_2,….. w_n} in which container X has H_x={w_1^x,……, w $_1^x$} and container Y D $_B$={w_1^y,……, w_1^y}.

It is assumed to have a multidimenssional database like MongoDB, in which each data chunk is represented by a vector set =$x_{i,1}$,……, $x_{i,m}$.

let container H has several containrised clusters $H1^x$, $H2^x$,……., HK^x Here, k represents no. of containers or clusters. while conhainer Y has H $_1^B$, H_2^y,……., H_K^y and (C_1,C_2,……,C_k)={ H_1^A + H_1^y ,……, H_k^x + H_k^y } as the combined container cluster centers[9].

3. *Algorithm conceptualized for the current scenario*

Assumption : Z_i is the no of nodes which are result which is the sum of container X and container Y i.e. H^X and H^Y respectively where Z_i= H^x+ H^y.

Given : 1) Warehouse W_x and W_y belonging to container X and container Y

2) The total no of clusters represented by k

Outcome: k is the clusters which is the aggregation of W_x and $\mathbf{W_y}$

a) The data is first normalized and scanned for the purpose of data mining

b) container X and container Y choose their own k container centers H_1^x, H_2^x,……., H_K^x

and H_1, H_2^y,……., H_K^Y in a purely random manner. $(Z_1,Z_2,.....,Z_k)=\{ H_1^X + H_1^Y ,……, H_k^X + H_k^Y \}$

c) Perform calculations fpr local k-means for container X and container Y.

d) Retain and store the container centres H_j^{Xi}, $H_{jY}^{,i}$.

e) Save $H_j^{X,i+1}$, $H_j^{Y,i+1}$.

On observation of the value of container center if the difference of value between the current and precious container center is less than or equal to minimum limit of threshold than stop the loop else continue the new value as initial value container X, Y generate a public private key pair as Pubk and Pvtk

4. THE WAY TO IMPLEMENT DIGITAL SIGNATURES

The digital signature is created by using a public key private key pair generated by keygen on SSL using PKCS standard.

The private key is stored in a secure key store while public key is used to create a sinature pattern encripted by the private key. The signature is finally exported to a SSL certificate which automatically exports public key on an SSL call to the containers where data chunks are stored

5. SETUP FOR IMPLEMENTATION

1. Ingress Controller : It is API gateway to access the data mining services deployed on cloud and containers

2. Docker – This a conatiner in which the database chunks can be deployed

3. Kubernetes – It is a container orchestrator which monitors the containers and looks into their scaling and health aspects

4. variables used for testing the application (k) used as no of containers. (x) – no of execution cycles . Md – mesaure of the distance Dc as delta convergence . Threshold value is kept at 0.4

5. Fedora 28 Gnome 3.28 64-bit

6. RESULTS AND DISCUSSION

The experiments focussed on the various parametes to be checked for the following

1. Genuinity

How genuine and correct are the results is often a question for every researcher. The results obtained on one platform may significantly differ on the other. Every cloud is having its own way to handle the data. So there has to be threshold based on which genuinity can be verified with a permitted degree of relaxation. The containers in the cloud may be down , destroyed or instantiated on various situations.

2. Protection against attacks

There must be a mechanism to generate various attack condition and ensure that the algorith is capable to withstand thouse attacks. One need to generate a log which gives correct account off the attack and the actions done to protect against attack. The container data has to be protected with a strong encryption and hashing algorithm. The assymetric way of identity establishment and encryption makes the whole affair very strong. This must be tested ahainst all atacks specially the DDOS attacks.

3. Output

The solution which is proposed here applies k-mean clustering on the containers where datachunks are stored in a Kubernetes environments. The queries from the containers are handled by mining service and the encrypted intermediate query results are processed by the service and is stored int other container. The results are also obtained by a single cluster container and multiple cluster container, the results are almost in the agreement in both the cases which justify the genuineness of the method.

7. CONCLUSION

The paper presents the secure way of performing database operations in for a warehouse distributed in two container chunks and the challenge of querying the data for the purpose of mining is dealt with so that the adversary cannot get the view of the intermediate processed results a homomorphic encryption is used in this along with digital signatures. This will stop the adversary to have a stealth mode while attacking the system. The approach only with two locations it can further be enahanced for n number of location and encryptions and signature checks to be performed as a integral part of data mining.

REFERENCES

1. S. Owen, A. Robin, T. Dunning, and E. Friedman (2012). Mahout in Action. Manning Publications, 2012.

2. http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linuxsingle-node-cluster

3. C. Tai, J. Huang, and M. Chung (2013) "Privacy Preserving Frequent Pattern Mining on Multi-cloud Environment." 2013 International Symposium on Biometrics and Security Technologies (ISBAST), IEEE, pp. 235-240.

4. R. Bhadauria, R. Borgohain, A. Biswas and S. Sanyal (2012). "Secure Authentication of Cloud Data Mining API" arXiv preprint arXiv:1204.0764.

5. K. Beaty, A. Kundu, V. Naik, and A. Acharya (2013). "Network-level Access Control Management for the Cloud." 2013 IEEE International Conference on Cloud Engineering (IC2E), IEEE, pp. 98-107.

6. http://archive. ics.uci.edu/ml/databases.

7. H. Dev, T. Sen, M. Basak, and M. E. Ali (2012). "An Approach to Protect the Privacy of Cloud Data from Data Mining Based Attacks" In High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:, pp. 1106-1115. IEEE, 2012.

8. P. Paillier (1999). "Public-key cryptosystems based on composite degree residuosity classes." In Advances in cryptology-EUROCRYPT'99, pp. 223-238. Springer Berlin Heidelberg.

9. Shobha Rajak (2012). Ashok Verma "Secure Data Storage in the Cloud using Digital Signature Mechanism" International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 4.

10. Wojciech Kinastowski (2013). Digital Signature as a Cloud-based Service" The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization, CLOUD COMPUTING.

Corresponding Author

Dr. Kamlendu Kumar Pandey*

Assistant Professor, Department of Information and Communication Technology, Veer Narmad South Gujarat University, Surat, India

kspandey@vnsgu.ac.in