Secure Data Mining in Cloud Using Homomorphic Encryption

Prof. Madhuri Patil[1] Srushti Deshmukh[2] Savita Nagvanshi[3] Sai Naik[4] Hitendra Gupta[5]
[1,2,3,4,5]Department of Information Technology
[1,2,3,4,5]MGM's College of Engineering & Technology, Kamothe- Mumbai University, India

Abstract— In advanced technology, industry, ecommerce and research a huge amount of complex and pervasive digital data is being generated which is increasing at an exponential rate and often termed as big data. Data mining techniques while allowing to individuals to extract knowledge on one hand introduce number of privacy threats on other hand. In this paper we are going to discuss some of the privacy issues in the data mining and cloud computing. Also the detailed discussion on different data mining techniques and the applications of data mining to provide the security.

Key words: Data Mining, Cloud Computing, Security, Homomorphic Encryption

I. INTRODUCTION

Cloud computing known as 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 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)[2]. 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 [3]. Amazon Web Services (AWS) including Simple Storage Service (S3), SQS, EC2 are cloud services provided by the Amazon [1].

II. OBJECTIVE

The main objective of this post to understand how to provide the security to the data.when very large amount of data is generated then how to secure that data by using the various data mining algorithm is also the objective of this paper.

III. EXISTING SYSTEM

To store the privacy of the data mining algorithm has been a concern of researchers for long and a number of algorithms have been proposed for the same. [7] Focuses on improving the security of two-party k-means while maintaining the correctness of algorithm. K-anonymity [10], noise transformation and multiplicative transformation [9][17] are some PPDM(privacy preserving data mining) methods. Compared to PPDM secure cloud mining is a relatively newer field..

IV. PROPOSED SYSTEM

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 have to perform data mining on the given data using k-means clustering approach while keeping 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

V. EXPLANATION

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. Each party performs Data Normalization on local data.
3) 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 }
4) Calculate or perform local k-means for Host A and Host B.
5) Save the cluster centers HjA,i, HjB,i .
6) Perform the secure cluster updation and reassign the data objects to their closest clusters locally
7) 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.

VI. SYSTEM ARCHITECTURE

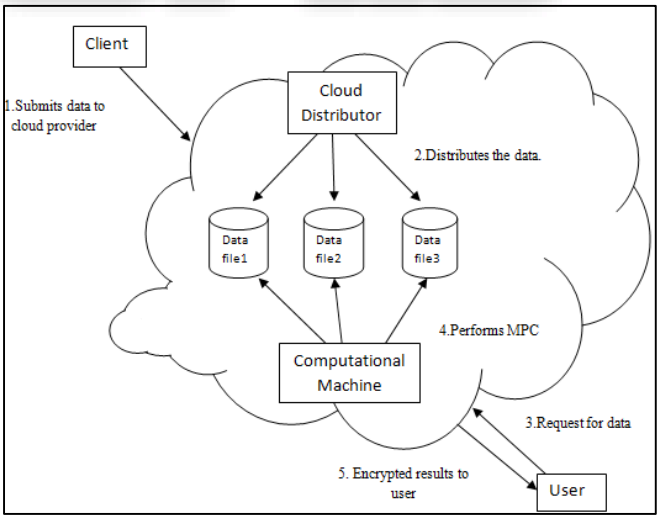

Fig.1: System Architecture

VII. RESULT AND ANALYSIS

A. *Evaluation Parameters*

1) Correctness

Correctness refers to the validity of the final results 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.

2) Security

This parameter evaluates the proposed algorithm in terms of security i.e the capability of the algorithm to prevent the attackers with malicious attacks to get access to the confidential user data and available valuable information extract from raw data.

VIII. 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. This paper explains a method to resolve 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 Pallier 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

REFERENCES

[1] M. Brantner, D. Florescu, D. Graf, D. Kossmann, and T. Kraska, "Building a database on S3." In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 251-264. ACM, 2008.
[2] J. Carolan , S. Gaede, J. Baty, G. Brunette, A. Licht, J. Remmell, L. Tucker, and J. Weise, "Introduction to cloud computing architecture." White Paper**,** 1st edn. Sun Micro Systems Inc (2009).
[3] Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica, "Above the clouds: A Berkeley view of cloud computing." Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS 28 (2009): 13.
[4] R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina, "Controlling data in the cloud: outsourcing computation without outsourcing control." In Proceedings of the 2009 ACM workshop on Cloud computing security, pp. 85-90. ACM, 2009.
[5] D. J. Solove, "I've got nothing to hide and other misunderstandings of privacy," San Diego L. Rev. 44 (2007): 745.
[6] P. K. Rexer, "Data miner survey highlights the views of 735 dataminers" 2010.
[7] Su, F. Bao, J. Zhou, T. Takagi, and K. Sakurai, "Privacy-preserving two-party k-means clustering via secure approximation." InAdvanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on, vol. 1, pp. 385-391. IEEE, 2007.
[8] Md. Riyazuddin , Dr.V.V.S.S.S.Balaram , Md.Afroze , Md.JaffarSadiq , M.D.Zuber. "An Empirical Study on Privacy Preserving Data Mining". International Journal of Engineering Trends and Technology (IJETT). V3(6):687-693 Nov-Dec 2012. ISSN:2231-5381
[9] K. Che, and L. Liu, "A random rotation perturbation approach to privacy preserving data classification." (2005).
[10] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification." In Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on, pp. 429-440. IEEE, 2009.

