VOL. 13, NO. 8, APRIL 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

UML BASED MODEL FOR DISPLAYING PRIVACY PRESERVING DATA MINING SYSTEMS

Anurag[1], Deepak Arora[2] and Upendra Kumar[3]
[1]Amity Institute of Information Technology Amity University Uttar Pradesh Lucknow Campus, India
[2]Department of Computer Science and Engineering Amity University Uttar Pradesh Lucknow Campus, India
[3]Department of Computer Science and Engineering Birla Institute of Technology, Patna Campus Bihar, India
Email: anurag.smit@gmail.com

ABSTRACT

Physical deployment of privacy preserving data mining system is a great challenge for organizations nowadays. What are different hardware and software files needed for deployment? How different software files across the system should interlinked together for functioning correctly. What execution environment should be provided for different platforms on the same hardware? What are various protocol needed for hardware communication. Its prior knowledge will assist the developers for successful implementation of the entire framework. This paper deals with UML pictorial model for enabling the developers in successful installation of Privacy Preserving Data Mining Systems during run time environment. This would enable for the engineers in developing the software projects within optimum time, within budget, reduce the chances of errors and in turn minimizes the development and maintenance effort.

Keywords: unified modeling language, privacy preserving data mining systems, component diagrams, deployment diagrams, centralize PPDM systems, distributed PPDM systems, object oriented software engineering, query language.

1. INTRODUCTION

Software programs are often too big and complex. It is a great challenge for programmers to develop the error free software. It often involves different stakeholders to coordinate together for writing codes in order to function correctly. Object oriented approach has been evolved with the aim to simulate real world applications. It deals with breaking of programs into objects. One can models the real world and translates the created objects from analysis into design. The first step in Object Oriented analysis is create a precise, concise, understandable and correct model of real worlds. The software components being develops should exhibit the properties of Correctness, reusability, extensibility, compatibility, portability and friendliness [16] [17]. It has many advantages such as protecting the other resources in the operating system, support encapsulating properties and Abstract data types. It abstraction property helps in making the design flexible. Class inheritance and object composition supports reusing functionality, which facilitates newer requirements to be compatible with existing requirements and leads the system to evolve and makes the system robust to the particular types of change. New functionality could be added without altering the existing functions [16]. UML is used to model the full range of practical system needed to be built. Model is very essential for the software development process. It is the visual description of the necessary details of project and guides the software developers built the error free software applications as any flaw in the design could produce the catastrophic result, and disturb the overall budget [18]. Computations mainly involves three ingredients-processors (or thread of control), actions (or functions) and data (also called objects).A system architecture could be obtained from function or from objects. It mainly concern with what the object does, rather than concerning what the system do[19].Object Oriented Paradigm solve all the problems of classical paradigm and is the best approach available today. It makes the task of development and maintenance easier. It exhibits the property of encapsulation which implies object's interdependence amongst each other [20]. Object-Oriented Design is the process of design encompassing object-oriented decomposition and a notation of representing both the logical and physical, as well as static and dynamic model of the system under design [21]. UML is standard diagrammatic notation for specifying, visualizing, constructing and documenting the artifacts of software system, as well as non-software system. It is the de-facto standard for Object Oriented Modeling and its 13 different diagrams are helpful for visualizing the complex software easier before actual implementation [22]. It provides the software professionals a stable and common design language that could be used to built complex applications for creating and discriminating design plans [23].

Concealing sensitive data from the outside world during mining process and simultaneously preserving the underlying data patterns so as to retain data utility which in turn could be exploited for gaining trade benefits has been major concern in the field of healthcare [2], business, web-usage mining [24], market basket analysis [25] and biometric [26] to name a few. Organizations mainly apply data mining for gaining trade benefits for extracting useful patterns while hiding personal information from competitive organizations. In the previous paper, authors have discuss the static modeling of various approach of Privacy Preserving Data Mining Systems by various UML diagrams such as Use Case diagram, Class diagram, Activity diagram and sequence diagram to name a few. Authors in this paper will discuss components and deployment diagrams for its physical deployment during run time. To the best of our knowledge, this is the first work which concern with UML Modeling of generalized PPDM system.

2. RELATED WORK

Let us review some of the basic works done in Privacy Preserving Data Mining Systems. Fiasco Gionatti *et al*. [1] proposed privacy preserving association rule mining in which data is outsourced to the cloud server and each item is made indistinguishable from at least k-1 transformed items. This approach is useful for the prevention of background knowledge attack. Ximeng Liu *et al.* [2] apply additive homomorphic encryption on patient sensitive data for training naïve bayes classifier which helps the upcoming patient to predict disease risks. Lin Zhang *et al*. [3] proposes privacy preserving decision tree mining for perturbing data by introducing two differential noises-Laplacian and exponential. Jun Zhou *et al*. [4] proposes cloud- assisted e healthcare system by medical text mining and image feature extraction. Privacy Preserving ID3 decision tree classification on medical datasets have been proposed in by Ye Li *et al* [5]. Arshveer Kaur [6] proposes Privacy preserving based on data anonimization technique. Awalia and Lakshamiwati [7] proposed hybrid privacy preserving technique by partitioning data based on entropy and then combining distorted data. Peter Shaojui Wang *et al.* [8] proposed an algorithm for the prevention of inside attackers in distributed karnel based data mining. Surbhi Sharma *et al.* [9] propose cryptographic privacy preserving data mining for organizational data based on C4.5 decision tree classification. Vadlana Baby *et al.* [10] proposes privacy preserving k- means clustering based on secret sharing technique for distributed datasets. P. Usha *et al.* [11] proposes the privacy preserving mining technique in which sensitive attributes of the heterogeneous datasets are anonimized and non-sensitive attributes are published. Vikas G. Ashok *et al*. [12] propose association rule mining on vertically partitioned datasets without compromising the efficiency during global data mining tasks. Brinal Colaco[13] proposed privacy preserving data mining on social network using fuzzy inference technique. Hare Ram Shah *et al.* [14] propose privacy preserving on image datasets by using visual steganography technique. The image is sliced and the query strings are searched to find the required image. L. Sumalatha *et al.* [15] presents the random decision tree using fuzzy logic for encrypting data in a cryptographic manner.

3. UML MODELING OF THE PROPOSED FRAMEWORK

Authors in this paper have discussed both centralize and distributed framework for achieving privacy preserving data mining tasks. Let us first discuss the centralize approach.

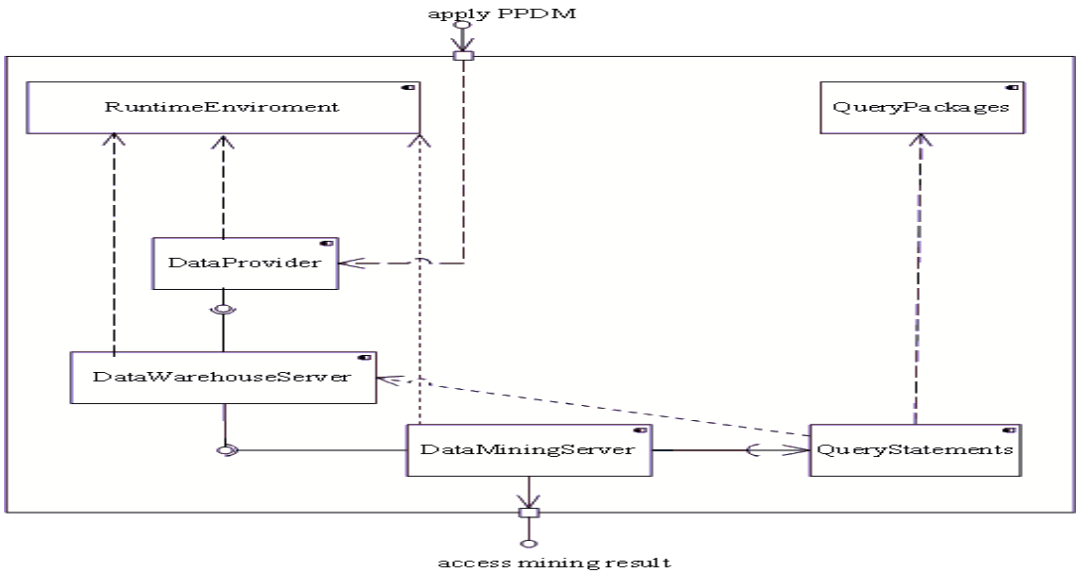

Figure-1. Component diagram of centralized PPDM system.

Runtime Environment, a software component containing source classes, object classes and all the necessary packages providing environment suitable for java programs execution, is installed in each sets of the hardware components. Figure below represents how each of these package and object files interact together for achieving the desired operations. Each Data Provider's information is stored in the tables and java object class fetch data from the table and applies privacy preserving operations on it. The data are outsourced to the Data Warehouse Server. The clean and integrated data from multiple Data Provider's datasets are stored in Data Warehouse database, which also include other information such as login credentials. As the operation on Data Mining Server involves any Object Oriented Program in the front-end and Query Language in backend, both Object Oriented Programming and Query Package environment are needed to installed in the Data Mining Server. Queries are sent to the Data Warehouse Server. Username and Password are needed which are stored in login credential database, and connection took place after successful authentication. Data is fetch from the Data warehouse database and privacy preserving operations are applied on it. Figure below represents the entire operations

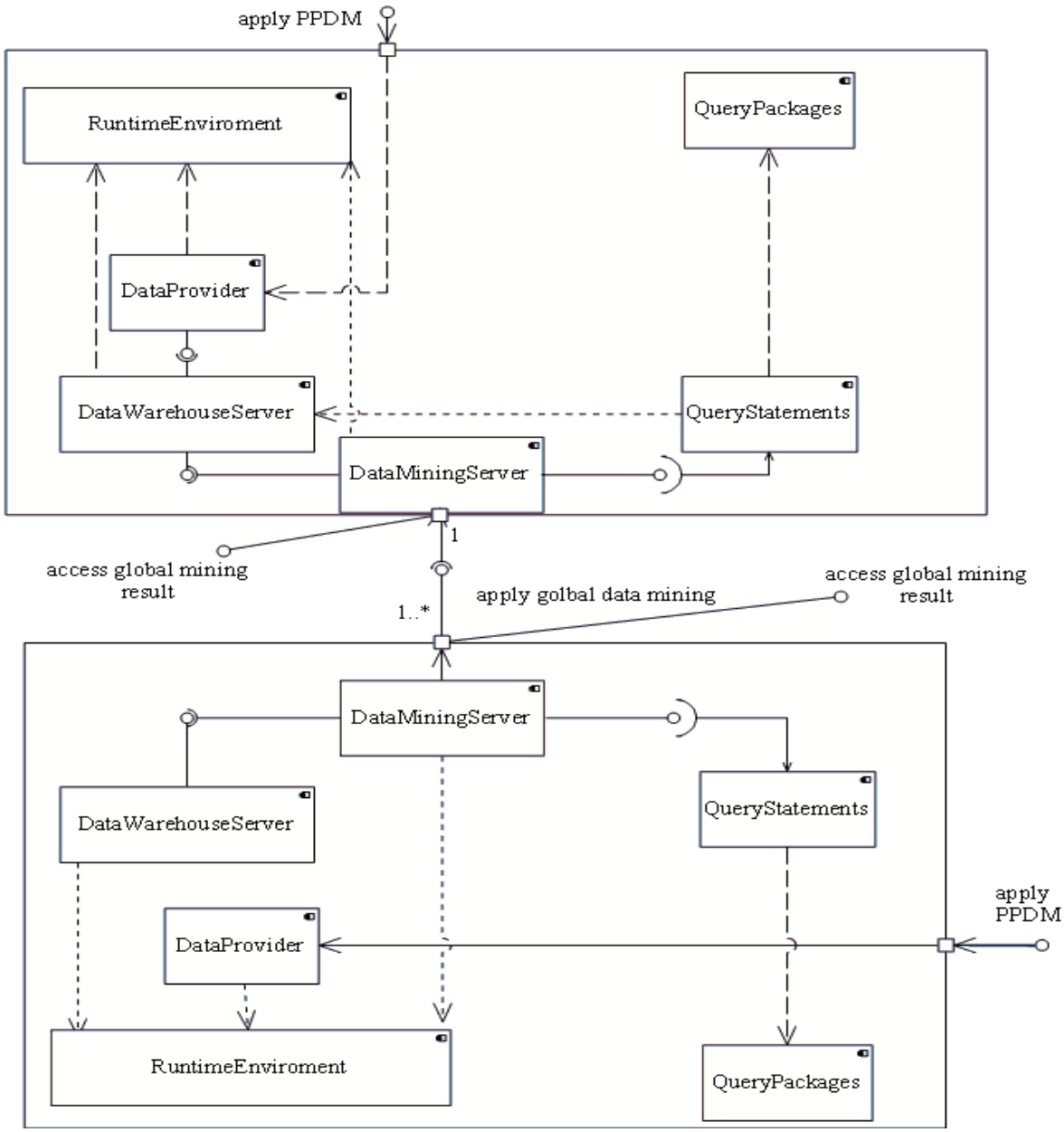

Figure-2. Component diagram of distributed PPDM system.

Figure-2 represents the component diagram of distributed Privacy Preserving Data Mining Systems. Different PPDM frameworks are connected via TCP/IP protocol suite and shares local data mining results to perform global data mining.

Figure-3 represents the deployment diagram of centralize Privacy Preserving Data Mining Systems. Various hardware components (or nodes) - Data Provider, Data Warehouse Server and Data Mining Server are installed in the systems. Data Provider and Data Warehouse Server interact together by TCP/IP protocol, while communication between Data Mining Server and Data Warehouse Server takes place connectivity. Data Provider gives input for performing the entire operations while Data Miners access the data mining output after PPDM operations.

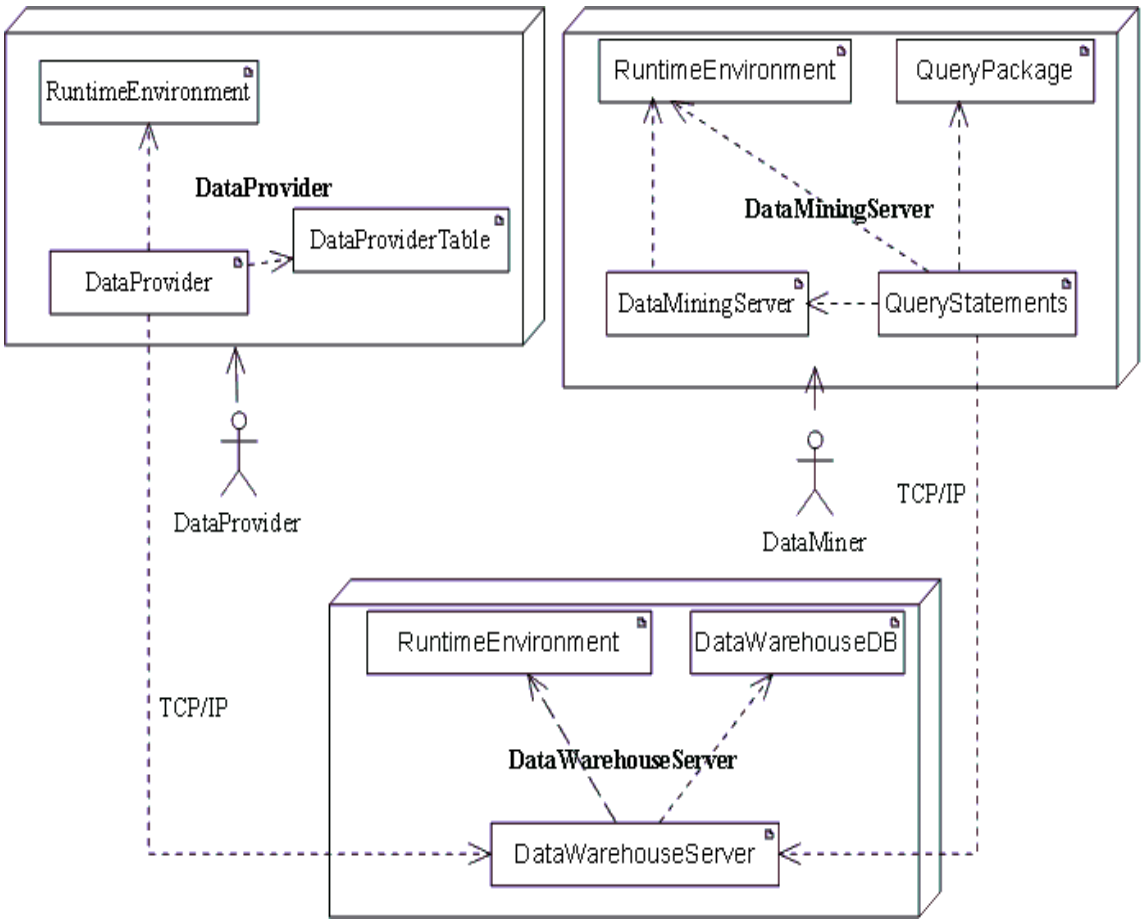

Figure-3. Deployment diagram of centralize PPDM systems.

Figure-4 depicts the deployment diagram of Distributed PPDM systems. Each PPDM frameworks are connected via TCP/IP Protocol suite. Each Data Mining Server across different frameworks share the local mining result with other framework by TCP/IP protocol suite to perform global data mining operations. Data Provider input the data and mining output is accessed by the each data miners of the corresponding framework.

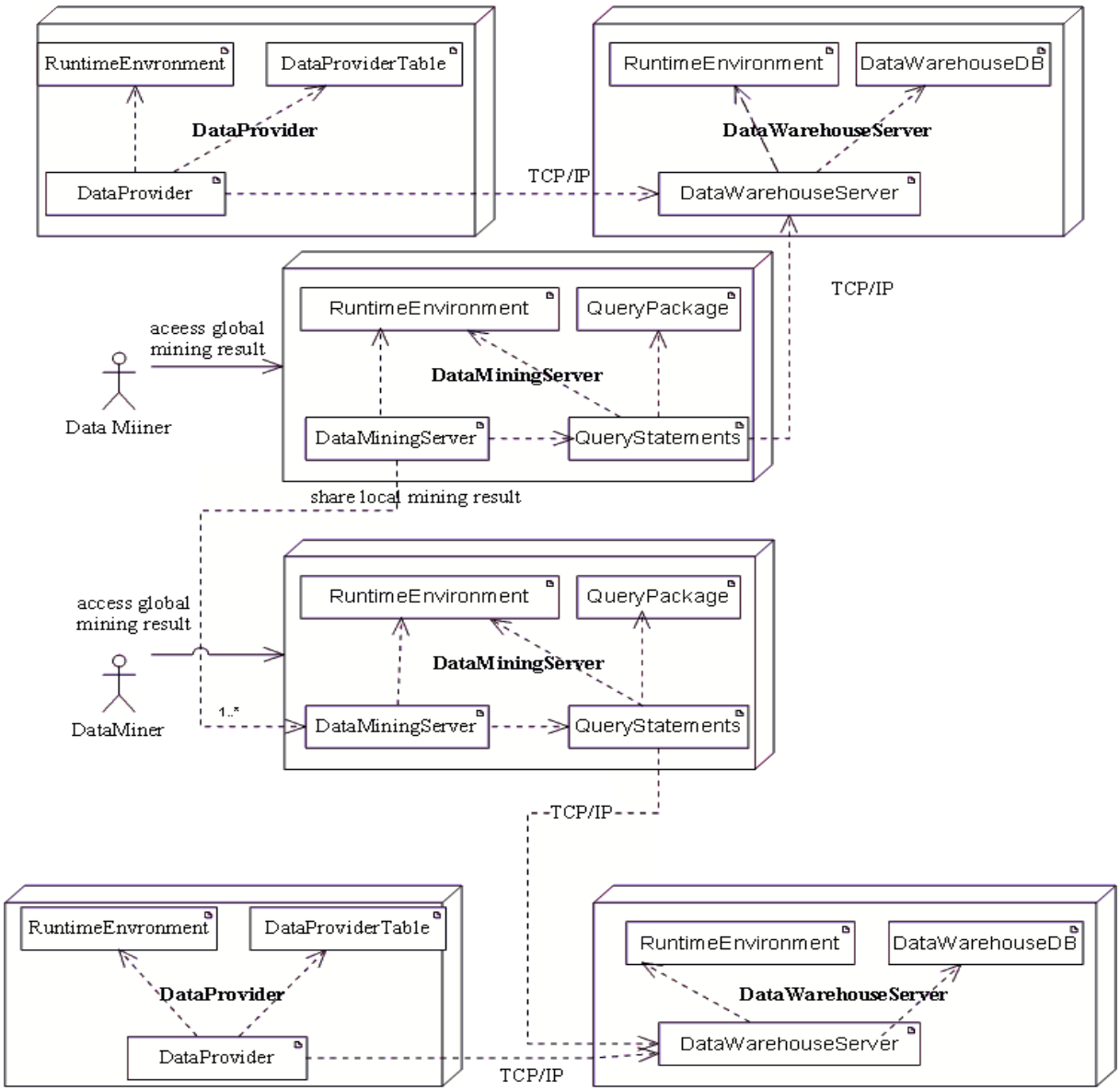

Figure-4. Deployment diagram of distributed PPDM systems.

CONCLUSIONS

Complexity is inherent in the development of software codes. The building of complex software often results in late, over budget and deficient in needed requirements, often called as software crisis. So, it is often required to decompose the software into objects. Deployment of Privacy Preserving systems is one of the biggest challenges nowadays. This could only be successfully achieved if developers have prior knowledge of complete hardware and software artifacts and its mutual interaction during deployments. Authors have explained step by step pictorial representation of various diagrams for both centralize and distributed data mining systems. This enables the software technicians to develop error free and robust application which in turn reduce further development and maintenance effort.

REFERENCES

[1] Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, Hui Wang. 2013. Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases. IEEE Systems Journal. 7(3): 385-385.

[2] Ximeng Li, Rongxing Lu, Jianfeng Ma and Baodong Qin. 2016. Privacy-Preserving Patient-Centric Clinical Decision Support System on Na¨ıve Bayesian Classification. IEEE Journal of Biomedical and health informatics. 20(2): 655-668.

[3] Lin Zhang; Yan Liu; Ruchuan Wang; Xiong Fu; Qiaomin Lin. 2017. Efficient privacy-preserving classification construction model with differential

privacy technology. Journal of Systems Engineering and Electronics, IEEE. 28(1): 170-178.

[4] Jun Zhou, Zhenfu Cao, Xiaolei Dong, Xiaodong Lin. 2015. PPDM: A Privacy-Preserving Protocol for Cloud-Assisted e-Healthcare System. IEEE Journal of Selected Topics in Signal Processing. 9(7): 1332-1344.

[5] Ye Li,Zoe L. Jiang,Xuan Wang, S.M. Yiu. 2017. Privacy-Preserving ID3 Data Mining over Encrypted Data in Outsourced Environments with Multiple Keys. IEEE International Conference on Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC). pp. 548-555.

[6] Arshveer Kaur. 2017. A hybrid approach of privacy preserving data mining using suppression and perturbation techniques. 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), IEEE. pp. 306-311.

[7] Awalia W. Putri, Laksmiwati Hira. 2016. Hybrid transformation in privacy-preserving data mining. 2016 International Conference on Data and Software Engineering (ICoDSE), IEEE. pp. 1-6.

[8] Peter Shaojui Wang. 2016. Insider Collusion Attack in Privacy Preserving karnel-based Data Mining Systems. Latest Advances and emerging applications of data hiding, IEEE. 4: 2244-2255.

[9] Surbhi Sharma; Deepak Shukla. 2016. Efficient multi-party privacy preserving data mining for vertically partitioned data. International Conference on Inventive Computation Technologies (ICICT). Vol. 2, IEEE.

[10] Vadlana Baby, N. Subhash Chandra. 2016. Distributed threshold k-means clustering for privacy preserving data mining. International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE. pp. 2286-2289.

[11] P. Usha; R. Shriram, S. Sathishkumar. 2015. Sensitive attribute based non-homogeneous anonymization for privacy preserving data mining. International Conference on Information Communication and Embedded Systems (ICICES). pp. 1-5.

[12] Vikas G. Ashok; K. Navuluri; A. Alhafdhi; R. Mukkamala. 2015. Dataless Data Mining: Association Rules-Based Distributed Privacy-Preserving Data Mining. 12th International Conference on Information Technology. pp. 615-620.

[13] Brinal Colaco; Shamsuddin S. Khan. 2014. Privacy preserving data mining for social networks. International Conference on Advances in Communication and Computing Technologies (ICACACT). pp. 1-4.

[14] Hare Ram Sah, G. Gunasekaran. 2015. Privacy preserving data mining using visual steganography and encryption. 10th International Conference on Computer Science & Education (ICCSE), IEEE, 2015Communication and Embedded Systems (ICICES). pp. 154-158.

[15] L. Sumalatha; P. Uma Sankar. 2016. Fuzzy random decision tree (FRDT) framework for privacy preserving data mining. SAI Computing Conference (SAI). pp. 195-202.

[16] Erich Gamma, Richard Helm, Ralph Johnson, John Vlissides. 1994. Design Patterns, Elements of Reusable Object Oriented Software. Addison Wisely.

[17] Jean-Marc Jezequel. 1980. Object Oriented Software Engineering with Eiffel, Addison Wisley.

[18] James Rumbaugh, Ivar Jacobson, Grady Booch. 1999. The Unified Modeling Language reference Manuel. Addison –Wesley.

[19] Bertrand Meyer. 1988. Object Oriented Software Construction. 2nd Edition, Prentice Hall.

[20] Stephen R. Schech. 2001. Object Oriented and classical software engineering. 8th Edition, McGraw Hill.

[21] Grady Booch. 1988. Object-Oriented analysis and design with applications. Addison - Wiesley.

[22] Craig Larmen, Applying UML and patterns: An Introduction to Object-oriented design and the Unified process. 2nd Edition, Adobe framework.

[23] Donald Bell. 2000. UML Basics:An introduction to Unified Modeling Language. IBM Global Solutions.

[24] T. Brijs, G. Swinnen, K. Vanhoof and G. Wets. 1999. Using association rules for product assortment decisions: A case study. in Proc. SIGKDD. pp. 254-260.

[25] B. Mobasher, H. Dai, T. Luo and M. Nakagawa. 2001. Effective personalization based on association rule discovery from Web usage data. in Proc. WIDM. pp. 9-15.

[26] C. Creighton and S. Hanash. 2003. Mining gene expression databases for association rules. Bioinformatics. 19(1): 79-86.