A Comprehensive Survey on Privacy-preserving Distributed Data Mining

S.Shunmugam [(1)], Dr. R.K. Selvakumar [(2)], P.Kavitha [(3)]

[(1)]*Research Scholar, Center for Information Technology, and Engineering, Manonmaniam Sundaranar, University, Tirunelveli,Tamilnadu,India*

[(2)]*Professor, Dept. of Computer Science& Engineering, CVR College of Engineering, Hyderabad, India*

[(3)]*Assoicate Professor, Dept. of CSE, Sri Sairam Institute of Tech, Chennai, Tamilnadu, India.*

Abstract- In the modern computer era, there are abundance of data available in every organization. From the available data, valuable hidden information can be extracted by the process called data mining. The field of data mining has emerged as a prominent area in the last two decades and even though fruitful information can be extracted from the data, the possibility that leakage of sensible private information of the individuals is a threat. Also there are laws enacted against compromising information leakage about individual .Recently the databases are not in a central location and mostly distributed even geographically. Every organization is interested to know the outcome of mining but reluctant to share the data freely. Hence preserving privacy in distributed data mining is of significance and lot of techniques have been evolved .This paper presents an extensive survey on the recent research works in privacy preservation of distributed data mining, the challenges, the limitations and upcoming trends.

Keywords- *Distributed data mining, privacy preservation, association rule, clustering, classification, secure multiparty computation, trusted third party*

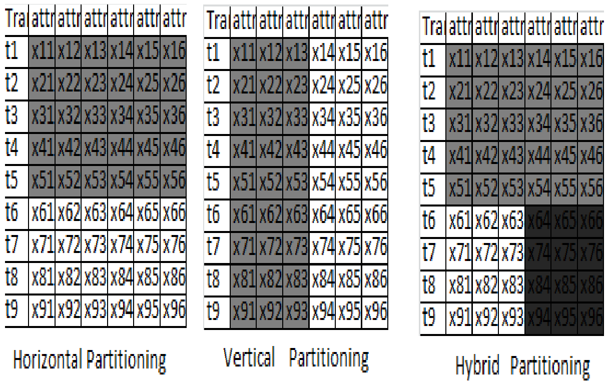
I. INTRODUCTION

With the advent of computer and internet large amount of data are available with every organization. Though maintaining large amount of data is a cumbersome one, most of the entities have data warehouses. Data mining is a process which is used to discover useful hidden patterns, trends, correlation and prediction from large amount of data available in a data warehouse model and used for the progress of the organization [2][6]. Different techniques are used to explore different tasks. Association rule mining, cluster analysis, classification and regression etc. are some of the data mining techniques.

With the increase in competition in businesses, it has also become essential to know how the competitors are performing. The primary concern in such a scenario is that each of the competitors does not want to disclose their individual data. Hence, privacy preservation is an important concern wherein collaborative distributed data mining needs to be undertaken [33]. Privacy preservation in distributed data mining (PPDDM) is a significant secure multiparty computation (SMC) problem among other SMC problems [34–36]. SMC helps in knowing how the competitors are performing without compromising on either party's privacy. The issue of SMC is such that only the data mining results of each of the sites that satisfy a certain function are known in the cumulative data. The confidential data of the collaborating parties remains private.

II.DISTRIBUTED DATA MINING

Data mining techniques were earlier applied in centralized data servers. Now the data servers are spread over at different locations and are belonging to different entities and distributed mining is to be done. Existing Algorithms and techniques were refined to adopt distributed data mining to fetch the desired results efficiently. Data could be distributed and partitioned as follows: [4] [14] [22]



Tra	attr	attr	attr	attr	attr	attr
t1	x11	x12	x13	x14	x15	x16
t2	x21	x22	x23	x24	x25	x26
t3	x31	x32	x33	x34	x35	x36
t4	x41	x42	x43	x44	x45	x46
t5	x51	x52	x53	x54	x55	x56
t6	x61	x62	x63	x64	x65	x66
t7	x71	x72	x73	x74	x75	x76
t8	x81	x82	x83	x84	x85	x86
t9	x91	x92	x93	x94	x95	x96

Horizontal Partitioning

Tra	attr	attr	attr	attr	attr	attr
t1	x11	x12	x13	x14	x15	x16
t2	x21	x22	x23	x24	x25	x26
t3	x31	x32	x33	x34	x35	x36
t4	x41	x42	x43	x44	x45	x46
t5	x51	x52	x53	x54	x55	x56
t6	x61	x62	x63	x64	x65	x66
t7	x71	x72	x73	x74	x75	x76
t8	x81	x82	x83	x84	x85	x86
t9	x91	x92	x93	x94	x95	x96

Vertical Partitioning

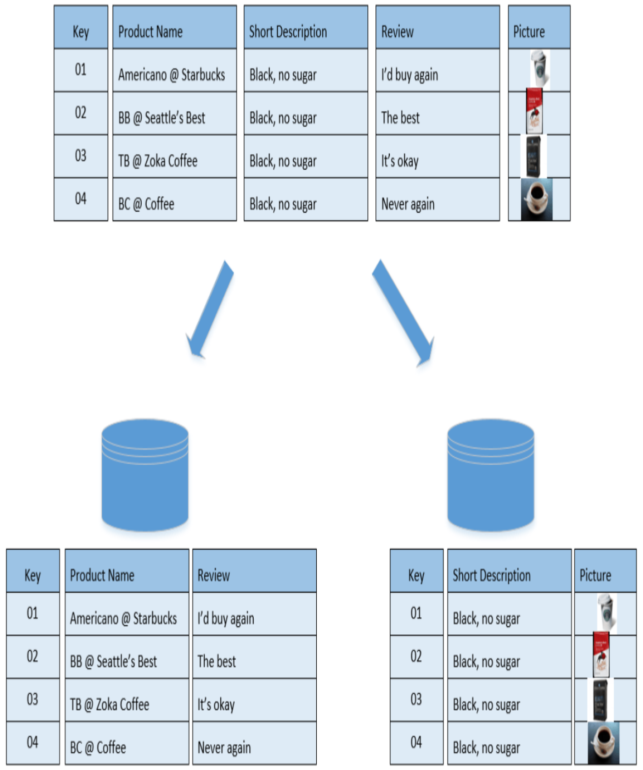
Tra	attr	attr	attr	attr	attr	attr
t1	x11	x12	x13	x14	x15	x16
t2	x21	x22	x23	x24	x25	x26
t3	x31	x32	x33	x34	x35	x36
t4	x41	x42	x43	x44	x45	x46
t5	x51	x52	x53	x54	x55	x56
t6	x61	x62	x63	x64	x65	x66
t7	x71	x72	x73	x74	x75	x76
t8	x81	x82	x83	x84	x85	x86
t9	x91	x92	x93	x94	x95	x96

Hybrid Partitioning

Figure 1. Types of Partitioning

Vertical partitioning

This type of partition divides the table vertically, which means that the structure of the main table changes in the new ones. An ideal scenario for this type of partition is when all the information about the customer in the query is not needed. If only need few details from the current year, the database could split it into two databases, one would hold customer information and current purchases, and the other would hold data about purchases from previous years.



Key	Product Name	Short Description	Review	Picture
01	Americano @ Starbucks	Black, no sugar	I'd buy again	
02	BB @ Seattle's Best	Black, no sugar	The best	
03	TB @ Zoka Coffee	Black, no sugar	It's okay	
04	BC @ Coffee	Black, no sugar	Never again	

Key	Product Name	Review
01	Americano @ Starbucks	I'd buy again
02	BB @ Seattle's Best	The best
03	TB @ Zoka Coffee	It's okay
04	BC @ Coffee	Never again

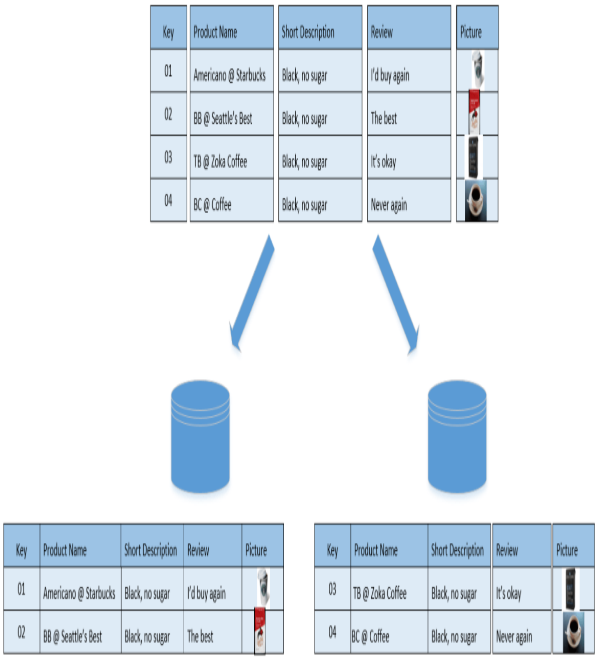
Key	Short Description	Picture
01	Black, no sugar	
02	Black, no sugar	
03	Black, no sugar	
04	Black, no sugar	

Figure 2: Example of Vertical Partitioning

Horizontal partitioning

In this partitioning the database is partitioned as rows. For example, if there is a large database of customers, that could be divided into four new tables: When partitioning the data, need to assess the number of rows in the new tables, so each table has the same number of customers and will grow by a similar number of new customers in the future. This might also partition the dataset based on the recent customers, for example, the clients that are not being active at your store are stored in one DB. In addition, the active customer database might be split into more tables, to get the results faster.

The structure of the original table stays the same in the new tables, i.e., we have the same number of columns. See Figure 3 for visual representation of the partition.



Key	Product Name	Short Description	Review	Picture
01	Americano @ Starbucks	Black, no sugar	I'd buy again	
02	BB @ Seattle's Best	Black, no sugar	The best	
03	TB @ Zoka Coffee	Black, no sugar	It's okay	
04	BC @ Coffee	Black, no sugar	Never again	

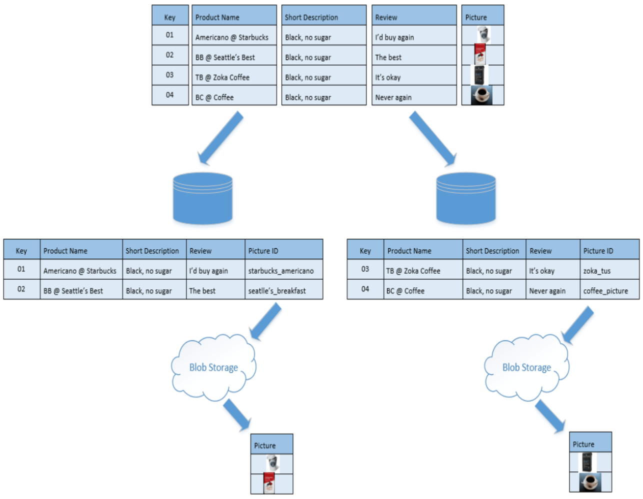
Key	Product Name	Short Description	Review	Picture
01	Americano @ Starbucks	Black, no sugar	I'd buy again	
02	BB @ Seattle's Best	Black, no sugar	The best	

Key	Product Name	Short Description	Review	Picture
03	TB @ Zoka Coffee	Black, no sugar	It's okay	
04	BC @ Coffee	Black, no sugar	Never again	

Figure 3: Example of Horizontal Partitioning

Hybrid Partitioning

This division combines vertical and horizontal partitioning. If there is a large dataset where you keep different types of data, that could horizontally partition the customer information and vertically divide the database into string values based on the criteria in a SQL DB, and pictures could be stored in Blob storage. See Figure 4 for visual representation of the partition.



Key	Product Name	Short Description	Review	Picture
01	Americano @ Starbucks	Black, no sugar	I'd buy again	
02	BB @ Seattle's Best	Black, no sugar	The best	
03	TB @ Zoka Coffee	Black, no sugar	It's okay	
04	BC @ Coffee	Black, no sugar	Never again	

Key	Product Name	Short Description	Review	Picture ID
01	Americano @ Starbucks	Black, no sugar	I'd buy again	starbucks_americano
02	BB @ Seattle's Best	Black, no sugar	The best	seatlle's_breakfast

Key	Product Name	Short Description	Review	Picture ID
03	TB @ Zoka Coffee	Black, no sugar	It's okay	zoka_tus
04	BC @ Coffee	Black, no sugar	Never again	coffee_picture

Figure 4: Example of Hybrid Partitioning

III. PRIVACY PRESERVING DISTRIBUTED DATA MINING (PPDDM)

In distributed data sets, even though the parties show interest in extracting useful mining results for them but hesitant to reveal the data due to risk of exposing the privacy to others.

Privacy Preservation in distributed data mining literally means processing of data using algorithms even without visualizing it and the data of the parties to remain private even after mining [26].Also to ensure confidentiality of data without compromising on data mining techniques[26].

PPDDM involves either trusted third party (TTP) [11] or Secure Multiparty computation (SMC) [10] [16] [29][30]. The parties involved in distributed data mining are classified into three categories. They are

1. Honest party
2. Semi honest party
3. Corrupted or malicious party

Honest party is one who genuinely shares the required data and not involving in unethical activity during the process. Semi honest party is one who follows the protocol and having intention to know about other party's information [31].Corrupted or malicious party is one who not follows the protocol or doing unethical activity during the process [31].The PPDDM has to take care of honest parties from getting the desired results even if the corrupted party foul play during the process and has to assume the presence of at least the semi honest parties. Also assume that intruders presence due to usage of network between the parties.

IV.CRYPTOGRAPHY IN PPDDM

Cryptography is a technique used for data security which hides the raw information such that others could not understand using encryption and may be decrypted by the intended party to know the information. It may use public key and private key for encryption or decryption [3].In PPDDM cryptographic methods are employed to prevent the intruder from knowing the information shared in the network connecting the parties and to prevent parties involved in mining from knowing the other party's data [3][7].

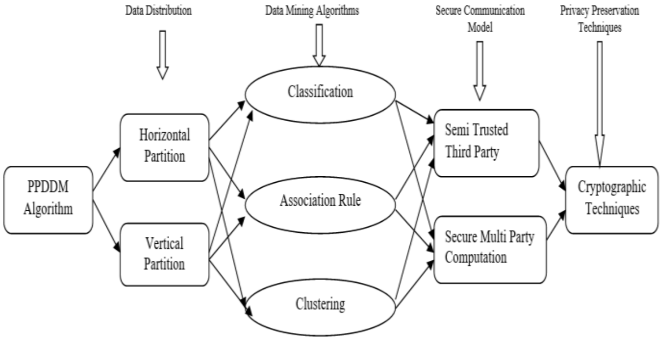


Figure 5.Generic PPDDM Structure

There are different kinds of cryptography protocols available .Though it prevents leakage of private information, hard to use where large data bases and many parties involved [18]. The generic model of PPDDM is shown in Figure 2. [4] [22]

V. Comparison of different Approached on PPDDM

a. Clustering Approach

Ahmed M. Elmisery et al [1] proposed a novel clustering approach over vertically partitioned data of different healthcare datasets for privacy preservation secure computation uses homomorphic encryption .This protocol protects from collusion attacks but time and communication complexity is high for large datasets.

Jinfei Liu et al [13] proposed a density based clustering algorithm DBSCAN clustering for horizontally, vertically and arbitrarily partitioned data. Privacy preserving was achieved by using multiplication protocol which enables the parties cooperate to obtain the clustering results without revealing the data at the individual parties. Time and communication is effective in this work.

Privacy preserving distributed K-means clustering was proposed by Sankita Patel et al [27].Shamir's secret sharing scheme was used over horizontal partitioned data. Better communication complexity was observed than using homomorphic encryption protocol and suitable for large datasets.

Zhenmin Lin [32] designed an efficient privacy preserving protocol using Expectation Maximization (EM) clustering over arbitrarily partitioned datasets. The two-party privacy preserving EM clustering protocol [5] disclosed only the number of iterations and the computation cost was high.

Meera Treesa Mathews et al [20] proposed Apriori algorithm for mining frequent item sets and Extended distributed Rk-secure sum protocol for privacy preserving. In this protocol all parties are arranged in Bus network and hence could not know the data of other parties. Even the number of parties increased the number of rounds remain the same and therefore computation and communication complexity are low.

The model proposed by N V Muthu Lakshmi et al [19] involves a sign based secure sum cryptographic technique for finding the global association mining over horizontally partitioned data using trusted party who prepares the merged list. In this model each site calculates the partial as well as the total support for all the item sets of the merged list provided by the trusted party. Using the sign based [23] cryptographic technique and global frequent item sets were generated by the trusted and informed to all sites. The communication complexity is low.

Jyotirmayee Rautaray et al [17] in their paper suggested an approach which is a combination of FP Tree algorithm and hybrid secure sum protocol over horizontally partitioned data.FP Tree algorithm is a top down approach has two parts namely FP Tree building and mining the FP Tree. In the hybrid secure sum protocol each party fragment the data into number of segments (less than 3) and assign their own random number. In [15] they proposed a method which involves Data Encryption Standards (DES).

Ehsan Molaei et al [9] suggested a method using Distributed algorithm for privacy preserving data mining based on ID3 and improved secure sum.In ID3distributed algorithm [8], instead of gathering data on a server, will calculate parameters like gain, entropy etc. by getting help from local data servers.

Author	**Algorithm/ Approach**	**Partitioning**	**No. of sites**	**SMC/Secured Trusted Third party**	**Advantages/ Limitations**
N V Muthu lakshmi et al 2012	Association Rule	Horizontal	>2	Trusted Third party/Sign based secure sum cryptography	Communication cost is low
Jinfei Liu et al 2013	DBSCAN	Horizontal Vertical Arbitrary	>=2	Yao's Millionaires' problem protocol and Multiplication protocol	Communication cost is low
Jyotirmayee Rautaray et al 2013	FP Tree Rule	Horizontal	>2	Hybrid secure sum protocol	Efficient communication cost
Jyotirmayee Rautaray et al 2013	FP Tree Rule	Horizontal	>2	SMC Data Encryption Standards(DES)	Communication cost is low
Jyotirmayee Rautaray et al 2013		All	>2	SMC Modified RK secure sum protocol	Communication cost is low/zero leakage
Nirali R Nanavati et al 2014	Association Rule	Horizontal	>2	No Third party/Shamir's secure sum and symmetric key based scheme	Better communication and computation cost
Meera Treesa Mathews et al 2014	Association Rule	Horizontal	>2	Extended Distributed RK secure sum protocol	Efficient computation and communication cost
Ehsan Molaei et al 2014	ID3 Distributed Algorithm	Horizontal	>2	AES cryptography and RSA asymmetric cryptography	Communication cost is moderate
Chirag N Modi et al 2015	Association Rule	Horizontal	>2	Elliptic curve cryptography and Onion Routing	Communication and computation cost is optimal for small data sets
Mayur B Tank et al 2015	Association Rule	Horizontal	5	Coordinator and using random number/no third party	Limited communication
T Nusrat Jabeen et al 2016	FP Growth Algorithm		>2	Associative third party/Elliptical curve cryptography	Computation and communication cost are low
Bhawani Singh Rathore et al 2016	Association Rule	Horizontal	4	RSA public key& Homomorphic Encryption	Computation and communication cost are low

Table 1.Survey of PPDDM

VI. CONCLUSION AND FUTURE WORK

An extensive study has been conducted in the Privacy Preservation of Distributed Data Mining in this paper. Most of the works carried out recently have taken care of the efficiency, time and communication complexity in privacy preserving in distributed data mining. Some models engaging Trusted Third party as intermediary in privacy preservation may reduce the computation complexity but in the real world scenario not encouraged. Some models where SMC are employed the computation complexity is high.   In case of association Rule mining much emphasis was made on finding positive association rules based on frequent item sets and the negative rules have been ignored. We propose to work on both positive and negative association rule mining in distributed databases in future while considering the weighted items/attributes.

REFERENCES

[1] Ahmed M. Elmisery and Huaiguo Fu," Privacy Preserving Distributed Learning Clustering Of HealthCare Data Using Cryptographic Protocols", 2010 34th Annual IEEE Computer Software and Applications Conference Workshops, 2010.

[2] J. Aruna Santhi* et al.," A Comprehensive Survey on Privacy Preserving Distributed Data Mining With Evolutionary Computing", (IJITR) International Journal of Innovative Technology and Research, Volume No.4, Issue No.5, August – September 2016, 3770 – 3772.

[3] Bhawani Singh Rathore, Anju Singh and Divakar Singh,"Secure Sum based Privacy Preservation Association Rule Mining on Horizontally Partitioned Data", International Journal of Computer Applications (0975 – 8887) Volume 134 – No.14, January 2016.

[4] V. Baby and N. Subhash Chandra," Privacy-Preserving Distributed Data Mining Techniques: A Survey", International Journal of Computer Applications (0975 – 8887), Volume 143 – No.10, June 2016.

[5] Chris Clifton, Murat Kantarcioglu and Xiaodong Lin, Michael Y. Zhu," Tools for Privacy Preserving Distributed Data Mining", ACM New York, NY, USA, ISSN: 1931-0145 EISSN: 1931-0153, Volume 4 Issue 2, December 2002.

[6] Chin-Chen Chang, Jieh-Shan Yeh and Yu-Chiang Li, "Privacy-Preserving Mining of Association Rules on Distributed Databases", IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.11, November 2006.

[7] Chirag N. Modia, Ashwini R. Patilb and Nishant Doshi," An Efficient Approach for Privacy Preserving Distributed Mining of Association Rules in Unsecured Environment",2015 International Conference on Advances in Computing, Communications and Informatics (ICACCl), 10-13 August 2015.

[8] Ehsan Molaei, Mehrdad Jalali and Hossein Vadiatizadeh, ”A Novel Algorithm for Privacy Preserving Distributed Data Mining”, International Journal of Computer Applications (0975 – 8887) Volume 76– No.8, August 2013.

[9] Ehsan Molaei, Hossein Vadiatizadeh et al,” Distributed Algorithm for privacy preserving data mining based on ID3 and improved secure sum”, International Journal of Advanced studies in Computer Science and Engineering IJASCSE, Volume 3, Issue r 1, 2014.

[10] Hemanta Kumar Bhuyan and Narendra Kumar Kamila, “Privacy preserving sub-feature selection in distributed data Mining”, Applied Soft Computing 36 (2015) 552–569, 2015.

[11] Israt Jahan, Nure Naushin Sharmy et al,” Design of a Secure Sum Protocol using Trusted Third Party System for Secure Multi-Party Computations”, 2015 6th International Conference on Information and Communication Systems (ICICS), 2015.

[12] Jayanti Danasana, Raghvendra Kumar and Debadutta Dey, “Mininig Association Rule for Horizontally Partitioned Databases using CK Secure Sum Technique”, International Journal of Distributed and Parallel Systems (IJDPS) Vol.3, No.6, November 2012.

[13] Jinfei Liu, Li Xiong, Jun Luo and Joshua Zhexue Huang,”Privacy Preserving Distributed DBSCAN Clustering”, Transactions on Data Privacy 6 (2013) 69–85, 2013.

[14] Jyotirmayee Rautaray and Raghvendra Kumar,” Distributed RK- Secure Sum Protocol for Privacy Preserving”, IOSR Journal of Computer Engineering (IOSR-JCE), e- ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 1, Jan. - Feb. 2013.

[15] Jyotirmayee Rautaray and Raghvendra Kumar,”Privacy Preserving in Distributed Database Using Data Encryption Standard”, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 2, Issue 3, March 2013.

[16] Jimmy Secretan,” An Architecture for High-Performance Privacy-Preserving and Distributed Data Mining”, Ph.D. Thesis, School of Electrical Engineering and Computer Science, College of Engineering and Computer Science, University of Central Florida, Orlando, Florida, 2009.

[17] Jyotirmayee Rautaray and Raghvendra Kumar ,” FP Tree Algorithm using Hybrid Secure Sum Protocol in Distributed Database”, International Journal of Scientific and Engineering Research Volume 4, Issue3, March-2013.

[18] Md. Golam Kaosar and Xun Yi,” Semi-Trusted Mixer Based Privacy Preserving Distributed Data Mining for Resource Constrained Devices”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 1, April 2010.

[19] N V Muthu lakshmi and Dr. K Sandhya Rani, “Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques”,(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (1), 2012, 3176 – 3182.

[20] Meera Treesa Mathews and Manju E.V,” Extended Distributed RK- Secure Sum Protocol in Apriori Algorithm For Privacy Preserving”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-3, Issue-4, April 2014.

[21] Masooda Modak and Rizwana Shaikh,” Privacy Preserving Distributed Association Rule Hiding Using Concept Hierarchy”, 7th International Conference on Communication, Computing and Virtualization 2016, Procedia Computer Science 79 (2016) 993 – 1000.

[22] Mayur B Tank and Tushar A Champaneria,” Privacy Preserving IJIRST – International Journal for Innovative Research in Scienc February 2015.

[23] Mayur B Tank and Tushar A Champaneria,” Privacy Preserving Distributed Data Mining”, IJSRD - International Journal for Scientific Research & Development, Vol. 3,Issue 04, 2015.

[24] T. Nusrat Jabeen and M. Chidambaram,” Privacy Preserving Association Rule Mining in Distributed Environments using FP-Growth Algorithm and Elliptic Curve Cryptography”, Indian Journal of Science and Technology, Vol 9(48), December 2016.

[25] Nirali R. Nanavati, Prakash Lalwani, and Devesh C. Jinwala” Analysis and Evaluation of Schemes for Secure Sum in Collaborative Frequent Itemset Mining across Horizontally Partitioned Data”, Hindawi Publishing Corporation, Journal of Engineering, Volume 2014, Article ID 470416, 2014.

[26] Prajna M.S and Sumana M,” Comprehensive Research on Privacy Preserving Emphasizing on Distributed Clustering”, International Journal of Science and Research (IJSR),Volume 5, Issue 4, April 2016.

[27] Sankita Patel, Sweta Garasia, and Devesh Jinwala,” An Efficient Approach for Privacy Preserving Distributed K-Means Clustering Based on Shamir’s Secret Sharing Scheme”, IFIPTM 2012, IFIP AICT 374, pp. 129–141, 2012.

[28] Selva Rathna and Dr.T. Karthikeyan,” Two Phase Secured Multiparty Sum Computation Protocol (2PSMC) for Privacy preserving data mining”, International Journal of Engineering and Computer Science ISSN: 2319-7242 Volume 4 Issue 4 April 2015, Page No. 11453-11456.

[29] V. Thavavel and S,Sivakumar,” A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012.

[30] G vikramaditya Reddy and Sayyad Rasheeduddin,

”Privacy-Preserving Distributed Data Analysis to Learn Data Model”, International Journal of Computer Application and Engineering Technology Volume 3, Issue 4, Oct 2014. Pp. 267-271.

[31] Lu, Yunmei, "Privacy Preserving Data Mining For Horizontally Distributed Medical Data Analysis", Dissertation, Georgia State University, 2016.

[32] Lin, Zhenmin, "Privacy Preserving Distributed Data Mining” (2012).Theses and Dissertations-Computer Science. Paper 9, University of Kentucky, 2012.

[33] Nirali R. Nanavati, Prakash Lalwani, and Devesh C. Jinwala, “Analysis and Evaluation of Schemes for Secure Sum in Collaborative Frequent Itemset Mining across Horizontally Partitioned Data” Hindawi Publishing Corporation Journal of Engineering Volume 2014.

[34] W. Du and M. J. Atallah, “Secure multi-party computation problems and their applications: a review and open problems,” in Proceedings New Security Paradigms Workshop (NSPIN ’01), V. Raskin, S. J. Greenwald, B. Timmerman, and D. M. Kienzle, Eds., pp. 13–22, ACM, September 2001.

[35] O. Goldreich and A. Warning, “Secure multi-party computation,” 2002, http://www.wisdom.weizmann.ac.il/∼oded/pp .html.

[36] P. Bogetoft, D. Christensen, I. Damgard et al., “Secure multiparty computation goes live,” in Financial Cryptography and Data Security, Lecture Notes in Computer Science, pp. 325–343, Springer, Berlin, Germany, 2009.