Performing Data Mining in (SRMS) Through Vertical Approach with Association Rules

Ambarish S. Durani[1], Vinay Kapse[2]

Abstract: ***This system technique is used for efficient data mining in SRMS (Student Records Management System) through vertical approach with association rules in distributed databases. The current leading technique is that of Kantarcioglu and Clifton[1]. In this system I deal with two challenges or issues, one that computes the union of private subsets that each of the interacting users hold, and another that tests the inclusion of an element held by one user in a subset held by another. The existing system uses different techniques for data mining purpose like Apriori algorithm. The Fast Distributed Mining (FDM) algorithm of Cheung et al. [2], which is an unsecured distributed version of the Apriori algorithm. Proposed system offers enhanced privacy and data mining with respect to the Encryption techniques and Association rule with Fp-Growth Algorithm in private cloud (system contains different files of subjects with respect to their branches). Due to this above techniques the expected effect on this system is that, it is simpler and more efficient in terms of communication cost and combinational cost. Due to these techniques it will affect the parameter like time consumption for execution, length of the code is decrease, find the data fast, extracting hidden predictive information from large databases and the efficiency of this system is increased by the 20%.***

Keywords: Data Mining; Vertical Approach; Association Rules

1. Introduction

In recent years the sizes of databases has increased rapidly. This has led to a growing interest in the development of tools capable in the automatic extraction of knowledge from data. The term Data Mining, or Knowledge Discovery in Databases, has been adopted for a field of research dealing with the automatic discovery of implicit information or knowledge within databases [2]. In Previous System, here the problem of data mining of association rules in partitioned databases. In that setting, there are several departments (or users) that hold homogeneous databases, i.e., databases that share the same schema but hold information on different entities. The goal is to find all association rules with support at least *s* and confidence at least *c*, for some given minimal support size *s* and confidence level *c*, that hold in the unified database, while minimizing the information disclosed about the private databases held by those users[1]. Consider the application SRMS in which the different databases are situated in each department, where the all data is stored semester wise. The main aim behind this system that the data should be stored in optimal (minimize) format with some secure techniques. In SRMS the data is used in large scale so, this propose system provide some technique for data mining with encryption /decryption techniques in private cloud. How SRMS worked?

2. Process of Execution

- First SRMS is one kind of web application which is used by a particular organization. Where the modules are Used by all staffs and admin.
- Second, store data in separate file respect to branch and Semester on cloud using encryption algorithm like AES.
- Third, collect data for ex. Admin wants data of semester 3 of CSE branch then collect the data through Association rule then id and name is taken from static Database and only marks are collect from different file which is present on cloud.
- Finally, shown the combine record of related semester and branch.

3. Review Literature

The existing system uses different techniques for data mining purpose like Apriori algorithm. The Fast Distributed Mining (FDM) algorithm of Cheung et al. [2], which is an unsecured distributed version of the Apriori algorithm.

Data mining is not particularly new statisticians have used similar manual approaches to review data and provide business evolutions for many years. Changes and updation in data mining techniques, however, have enabled organizations to collect, monitor, analyze, and access data in new ways. The first change occurred in the area of basic data collection. Before companies uses the transition from paper-based records to computer-based systems, managers had to wait for staff to give records of pieces together to know how well the business was performing or how current performance compared with previous. As all companies started collecting and saving basic data in computers, they were able to start quick answering detailed easily.

Now a day's a third party is exist to provide a service for commercial company. The users could submitted to him their inputs and he would perform the function evaluation and send to them the resulting output. In the absence of such a trusted third party, it is needed to devise techniques that the users can run on their own in order to arrive at the required output *y*. The next aim is to secure the inputs of each user. if the both are combined together(data mining and Secure) the third party involvement is avoided Yao was the first to propose a generic solution for this problem in the case of two players. [3].

In our problem, the inputs are the partial databases, and the required output is the list of association rules that hold in the cache memory with support and confidence *s* and *c*, respectively. They can be applied only to small inputs and functions which are realizable by simple circuits. In more complex settings, such as ours, other methods are required for carrying out this computation.

For example[11], the rule found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy burger. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics. Three parallel algorithms for mining association rules [12], an important data mining problem is formulated in this paper. These algorithms have been designed to investigate and understand the performance implications of a spectrum of trade-offs between computation, communication, memory usage, synchronization, and the use of problem-specific information in parallel data mining [13]. Fast Distributed Mining of association rules, which generates a small number of candidate sets and substantially reduces the number of messages to be passed at mining association rules [14].

For mining of data and encryption/decryption different techniques are available. Like for data mining K-means algorithms, Apriori Algorithm. Fast Distributed Mining and for encryption/Decryption RSA, DES etc [10]. This paper proposes the Fp growth mining with AES algorithm to provide mining and encryption. Figure shows the architecture of scheme for SRMS.

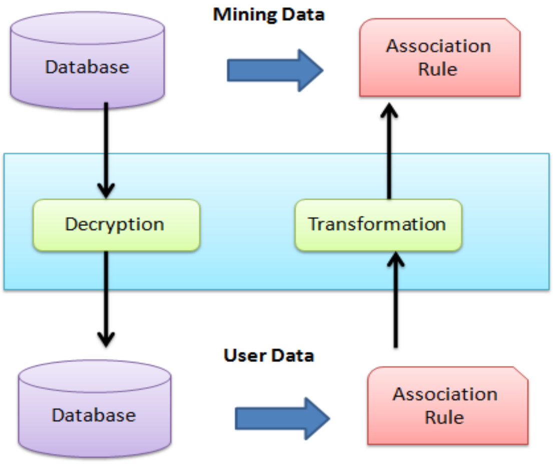


Figure 3.1: Architecture Scheme

4. Methodology

A. Mining Algorithm

Previously Apriori algorithm is used. It uses a generate and test approach generates candidate item sets and tests if they are frequent [4]

- Generation of candidate item sets is expensive (in both space and time)
- Support counting is expensive
- Subset checking (computationally expensive)
- Multiple Database scans (I/O)

For SRMS Apriori algorithm is not beneficial, one disadvantage is overcome by FP Growth algorithm. FP-Growth allows frequent itemset discovery without candidate itemset generation. Two step approach:
Step 1: Build a compact data structure called the FP-tree
Step 2: Extracts frequent itemsets directly from the FP-tree.

Mining is preferably used for a large amount of data [8, 9] and related algorithms often require large data sets to create quality models [7]**.** The relationship between data mining and cloud is worth to discuss. Cloud providers use data mining to provide clients a better service [6].

B. AES algorithm

AES is asymmetric which is encrypted by different keys. Here in this paper the AES is used with different key length, different iteration and perform operation of different file size. The encryption is in fact not difficult to break if a dictionary of words with their expected frequencies is available [5] This will covered in Result analysis.

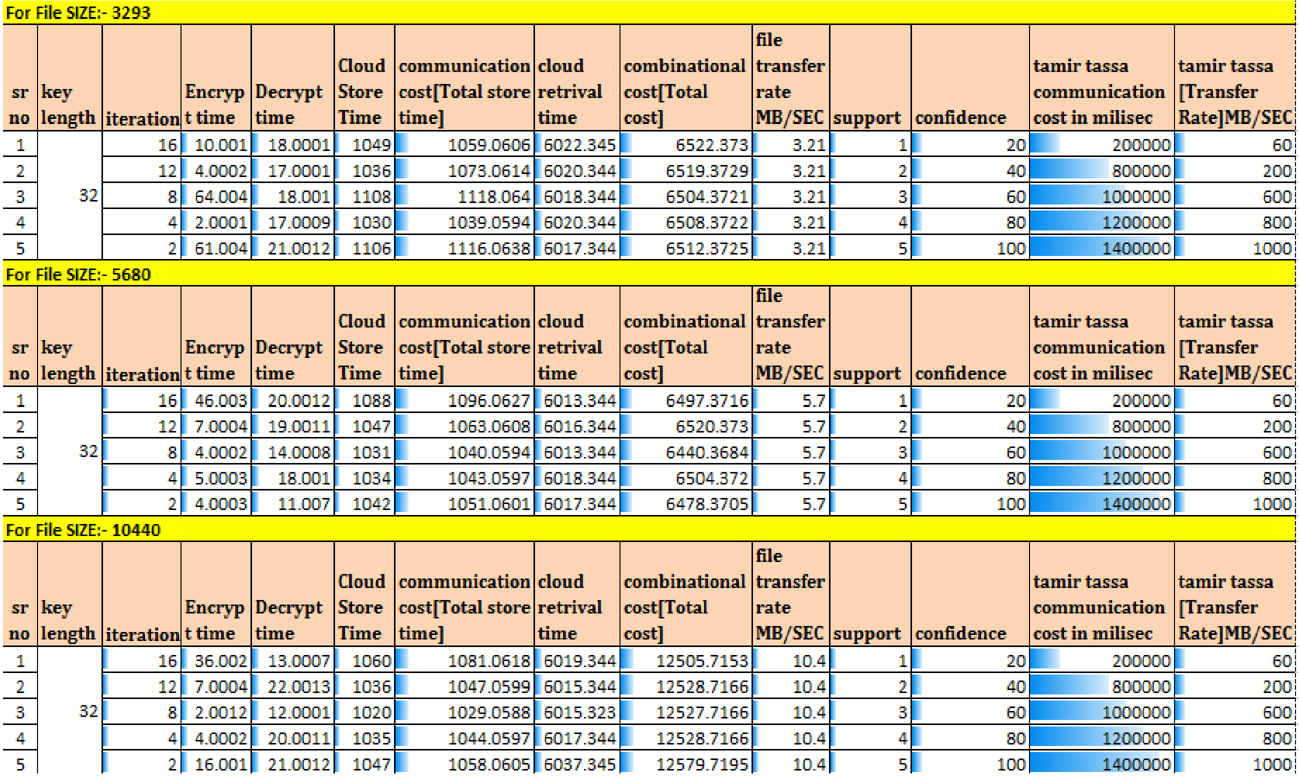
C. Cloud Interface

Cloud is used for only storage purpose, now a day there are two options available for storing data, first is server and second is cloud. for better security and larger space cloud is a better option. It is quite difficult to locate path of cloud where actual data is stored. In SAMS the files are encrypted using AES algorithm, The files are stored in “SAMS” domain in cloud. The union of record is performed by FP algorithm and data is return to the admin module. Cloud is also shows the message that how many space is used by user and show remaining space.

5. Result Analysis

The result analysis is performed by the calculating the communication cost and combinational cost, with the encryption and decryption time. The readings are calculated by using different key length, file size and iterations. All experiments were implemented in C# (.net 4) and were executed on an Intel(R) Core(TM) i3 personal computer with a 1.66GHz CPU, 8 GB of RAM, and the 32-bit operating system Windows 7 ultimate. Table 5.1 Shows the computational and combinational cost for 32 key length and different file size.

Table 5.1: Computational cost and Combinational cost with encryption and decryption time.



For File SIZE:- 3293

sr no	key length	iteration	Encrypt time	Decrypt time	Cloud Store Time	communication cost[Total store time]	cloud retrival time	combinational cost[Total cost]	file transfer rate MB/SEC	support	confidence	tamir tassa communication cost in milisec	tamir tassa [Transfer Rate]MB/SEC
1	32	16	10.001	18.0001	1049	1059.0606	6022.345	6522.373	3.21	1	20	200000	60
2		12	4.0002	17.0001	1036	1073.0614	6020.344	6519.3729	3.21	2	40	800000	200
3		8	64.004	18.001	1108	1118.064	6018.344	6504.3721	3.21	3	60	1000000	600
4		4	2.0001	17.0009	1030	1039.0594	6020.344	6508.3722	3.21	4	80	1200000	800
5		2	61.004	21.0012	1106	1116.0638	6017.344	6512.3725	3.21	5	100	1400000	1000

For File SIZE:- 5680

sr no	key length	iteration	Encrypt time	Decrypt time	Cloud Store Time	communication cost[Total store time]	cloud retrival time	combinational cost[Total cost]	file transfer rate MB/SEC	support	confidence	tamir tassa communication cost in milisec	tamir tassa [Transfer Rate]MB/SEC
1	32	16	46.003	20.0012	1088	1096.0627	6013.344	6497.3716	5.7	1	20	200000	60
2		12	7.0004	19.0011	1047	1063.0608	6016.344	6520.373	5.7	2	40	800000	200
3		8	4.0002	14.0008	1031	1040.0594	6013.344	6440.3684	5.7	3	60	1000000	600
4		4	5.0003	18.001	1034	1043.0597	6018.344	6504.372	5.7	4	80	1200000	800
5		2	4.0003	11.007	1042	1051.0601	6017.344	6478.3705	5.7	5	100	1400000	1000

For File SIZE:- 10440

sr no	key length	iteration	Encrypt time	Decrypt time	Cloud Store Time	communication cost[Total store time]	cloud retrival time	combinational cost[Total cost]	file transfer rate MB/SEC	support	confidence	tamir tassa communication cost in milisec	tamir tassa [Transfer Rate]MB/SEC
1	32	16	36.002	13.0007	1060	1081.0618	6019.344	12505.7153	10.4	1	20	200000	60
2		12	7.0004	22.0013	1036	1047.0599	6015.344	12528.7166	10.4	2	40	800000	200
3		8	2.0012	12.0001	1020	1029.0588	6015.323	12527.7166	10.4	3	60	1000000	600
4		4	4.0002	20.0011	1035	1044.0597	6017.344	12528.7166	10.4	4	80	1200000	800
5		2	16.001	21.0012	1047	1058.0605	6037.345	12579.7195	10.4	5	100	1400000	1000

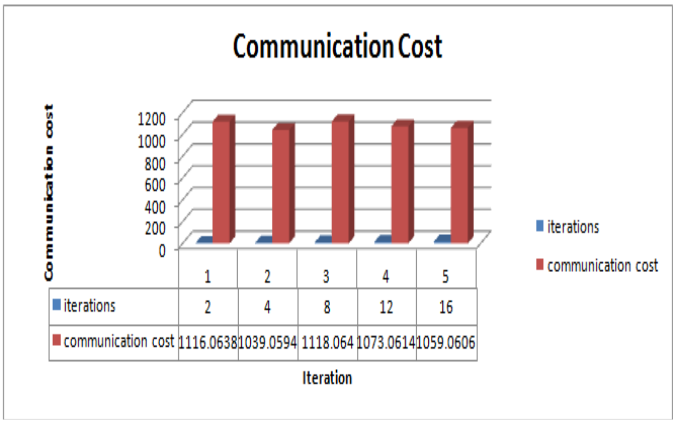


Figure 5.1: Fig.5.1. Shows communication cost

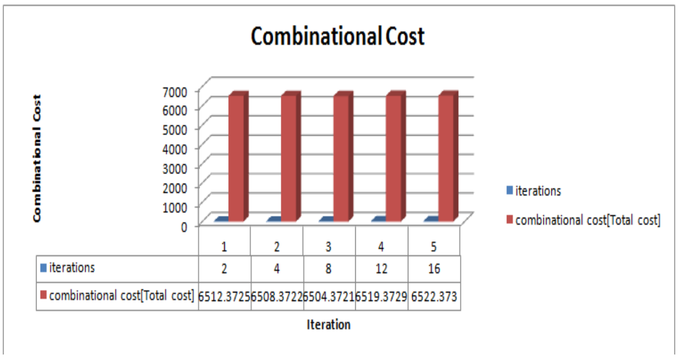


Figure 5.2: Shows combinational cost

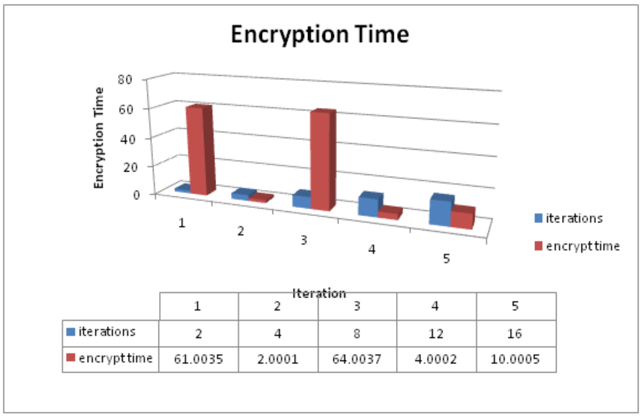


Figure 5.3: Shows Encryption Time

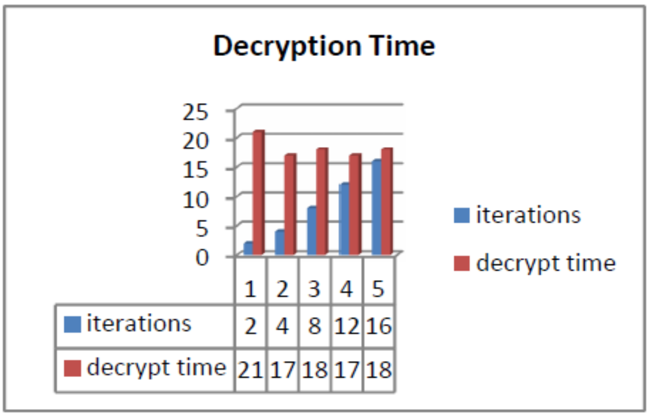


Figure 5.4: Shows Decryption Time

Example 2:

1) Fixed key length 16 key length , File Size 3293,5680,10440 KB, Iterations 2,4,8,12,16.

Table 5.2: shows the Communication, Combinational cost with encryption and decryption time

For File SIZE:- 3293

sr no	key length	iteration	Encryp t time	Decrypt time	Cloud Store Time	communication cost[Total store time]	cloud retrival time	combinational cost[Total cost]	file transfer rate MB/SEC	support	confidence	tamir tassa communication cost in milisec	tamir tassa [Transfer Rate]MB/SEC
1	16	16	4.003	18.001	1034	1055.0603	6019.344	6517.3728	3.3	1	20	200000	60
2		12	12.001	44.0026	1041	1048.0599	6027.345	6500.3718	3.3	2	40	800000	200
3		8	4.0002	65.0037	1031	1056.0604	6031.345	6539.374	3.3	3	60	1000000	600
4		4	17.001	44.0025	1045	1053.0603	6062.347	6572.376	3.3	4	80	1200000	800
5		2	4.0002	52.003	1031	1044.0597	6029.345	6511.3724	3.3	5	100	1400000	1000

For File SIZE:- 5680

sr no	key length	iteration	Encryp t time	Decrypt time	Cloud Store Time	communication cost[Total store time]	cloud retrival time	combinational cost[Total cost]	file transfer rate MB/SEC	support	confidence	tamir tassa communication cost in milisec	tamir tassa [Transfer Rate]MB/SEC
1	16	16	26.002	17.001	1059	1074.0614	6017.344	6503.372	5.7	1	20	200000	60
2		12	4.0002	20.0011	1033	1044.0597	6025.345	6530.3735	5.7	2	40	800000	200
3		8	4.0003	18.001	1034	1046.0599	6016.344	6530.3735	5.7	3	60	1000000	600
4		4	42.002	51.0029	1083	1093.0626	6015.344	6542.3742	5.7	4	80	1200000	800
5		2	11.001	16.0009	1051	1061.0606	6020.344	6520.373	5.7	5	100	1400000	1000

For File SIZE:- 10440

sr no	key length	iteration	Encryp t time	Decrypt time	Cloud Store Time	communication cost[Total store time]	cloud retrival time	combinational cost[Total cost]	file transfer rate MB/SEC	support	confidence	tamir tassa communication cost in milisec	tamir tassa [Transfer Rate]MB/SEC
1	16	16	17.001	11.0006	1054	1073.0614	6015.344	6447.3688	10.4	1	20	200000	60
2		12	5.0003	15.0009	1045	1065.0609	6020.344	6518.3728	10.4	2	40	800000	200
3		8	41.002	11.0007	1073	1083.062	6016.344	6505.3721	10.4	3	60	1000000	600
4		4	4.002	14.0008	1036	1107.0633	6017.344	6507.3722	10.4	4	80	1200000	800
5		2	21.001	12.0007	1048	1058.0606	6020.344	6520.3729	10.4	5	100	1400000	1000

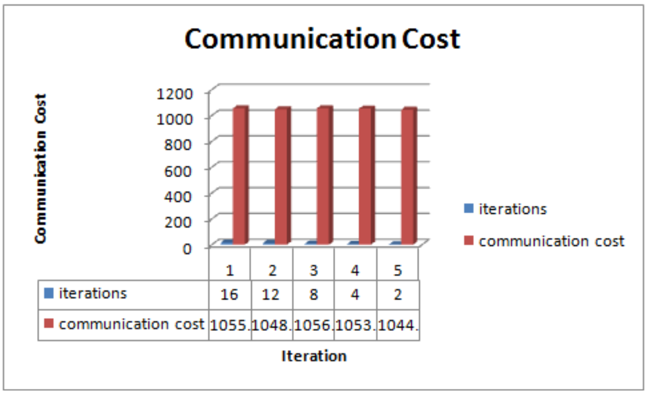


Figure 5.5: Shows communication cost

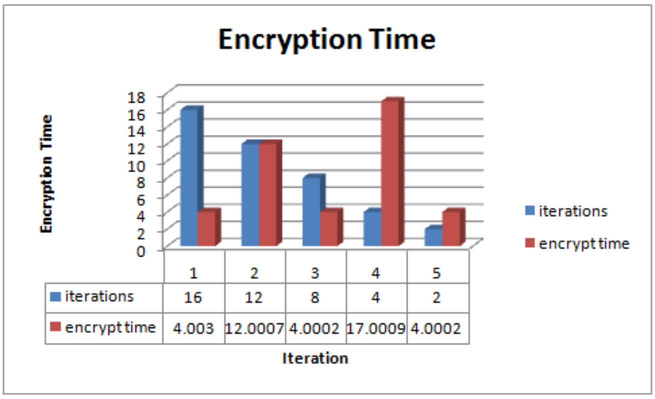


Figure 5.7: Shows Encryption Time

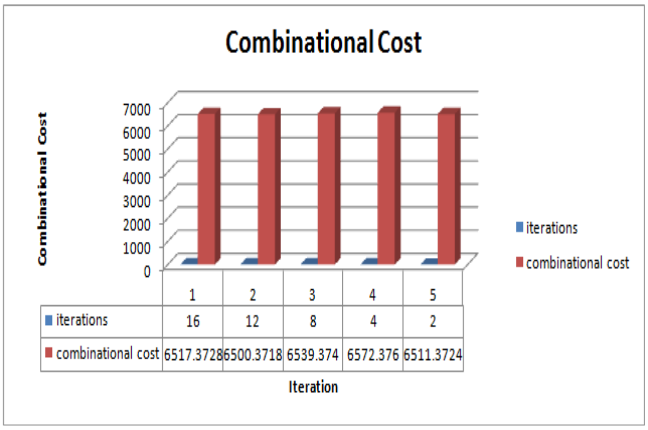


Figure 5.6: Shows combinational cost

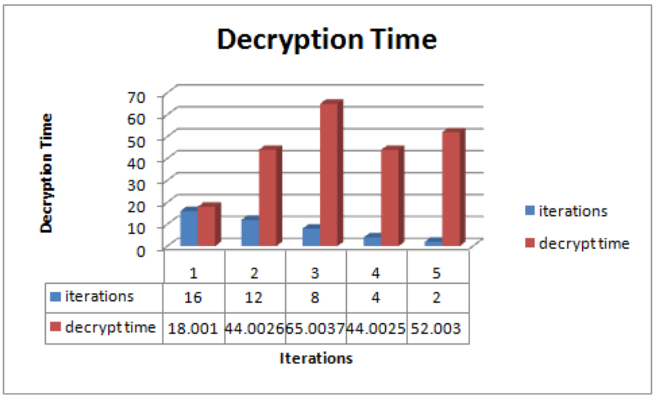
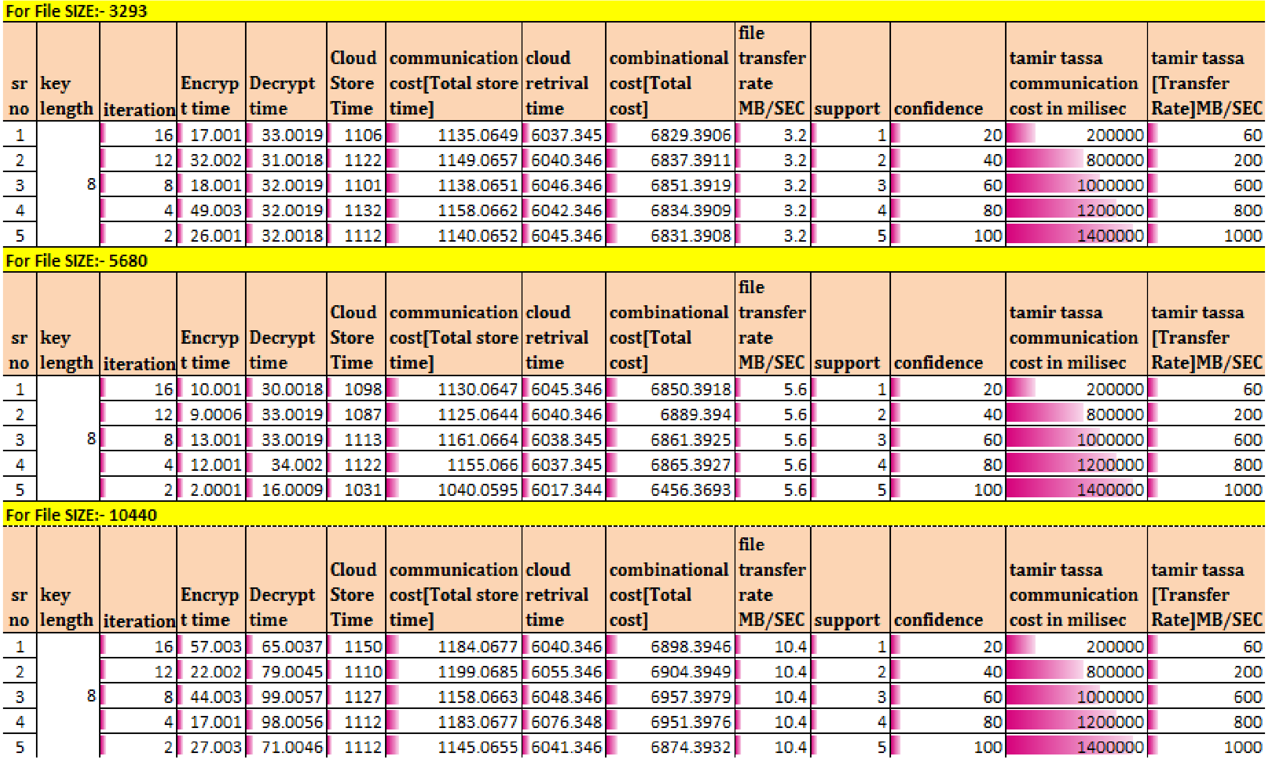


Figure 5.8: Shows Decryption Time

Example 3:

1) Fixed key length 8 key length ,File Size3293,5680,10440 KB, Iterations 2,4,8,12,16

Table 5.3: Shows the Communication, Combinational cost with encryption and decryption time



For File SIZE:- 3293

sr no	key length	iteration	Encrypt time	Decrypt time	Cloud Store Time	communication cost[Total store time]	cloud retrival time	combinational cost[Total cost]	file transfer rate MB/SEC	support	confidence	tamir tassa communication cost in milisec	tamir tassa [Transfer Rate]MB/SEC
1	8	16	17.001	33.0019	1106	1135.0649	6037.345	6829.3906	3.2	1	20	200000	60
2		12	32.002	31.0018	1122	1149.0657	6040.346	6837.3911	3.2	2	40	800000	200
3		8	18.001	32.0019	1101	1138.0651	6046.346	6851.3919	3.2	3	60	1000000	600
4		4	49.003	32.0019	1132	1158.0662	6042.346	6834.3909	3.2	4	80	1200000	800
5		2	26.001	32.0018	1112	1140.0652	6045.346	6831.3908	3.2	5	100	1400000	1000

For File SIZE:- 5680

sr no	key length	iteration	Encrypt time	Decrypt time	Cloud Store Time	communication cost[Total store time]	cloud retrival time	combinational cost[Total cost]	file transfer rate MB/SEC	support	confidence	tamir tassa communication cost in milisec	tamir tassa [Transfer Rate]MB/SEC
1	8	16	10.001	30.0018	1098	1130.0647	6045.346	6850.3918	5.6	1	20	200000	60
2		12	9.0006	33.0019	1087	1125.0644	6040.346	6889.394	5.6	2	40	800000	200
3		8	13.001	33.0019	1113	1161.0664	6038.345	6861.3925	5.6	3	60	1000000	600
4		4	12.001	34.002	1122	1155.066	6037.345	6865.3927	5.6	4	80	1200000	800
5		2	2.0001	16.0009	1031	1040.0595	6017.344	6456.3693	5.6	5	100	1400000	1000

For File SIZE:- 10440

sr no	key length	iteration	Encrypt time	Decrypt time	Cloud Store Time	communication cost[Total store time]	cloud retrival time	combinational cost[Total cost]	file transfer rate MB/SEC	support	confidence	tamir tassa communication cost in milisec	tamir tassa [Transfer Rate]MB/SEC
1	8	16	57.003	65.0037	1150	1184.0677	6040.346	6898.3946	10.4	1	20	200000	60
2		12	22.002	79.0045	1110	1199.0685	6055.346	6904.3949	10.4	2	40	800000	200
3		8	44.003	99.0057	1127	1158.0663	6048.346	6957.3979	10.4	3	60	1000000	600
4		4	17.001	98.0056	1112	1183.0677	6076.348	6951.3976	10.4	4	80	1200000	800
5		2	27.003	71.0046	1112	1145.0655	6041.346	6874.3932	10.4	5	100	1400000	1000

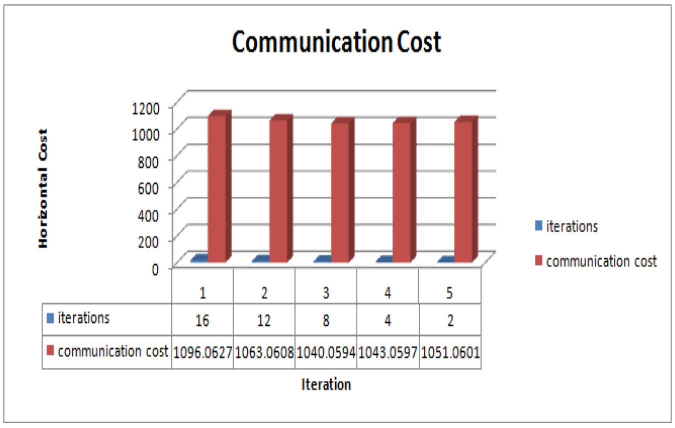


Figure 5.9: Shows communication cost

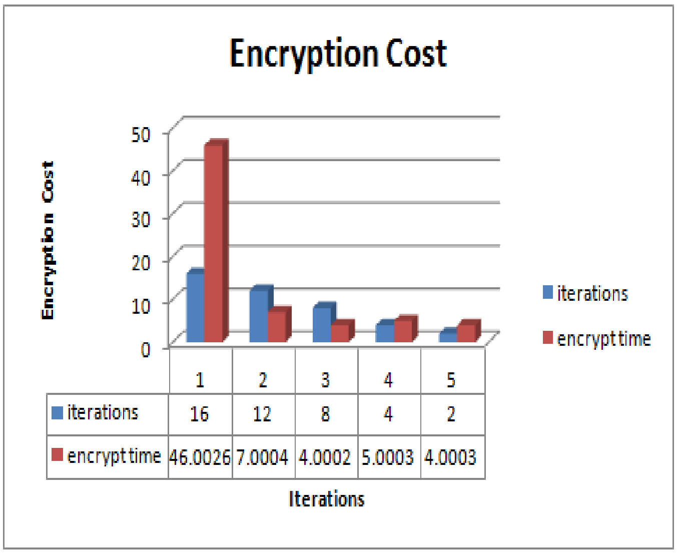


Figure 5.11: Shows Encryption Time

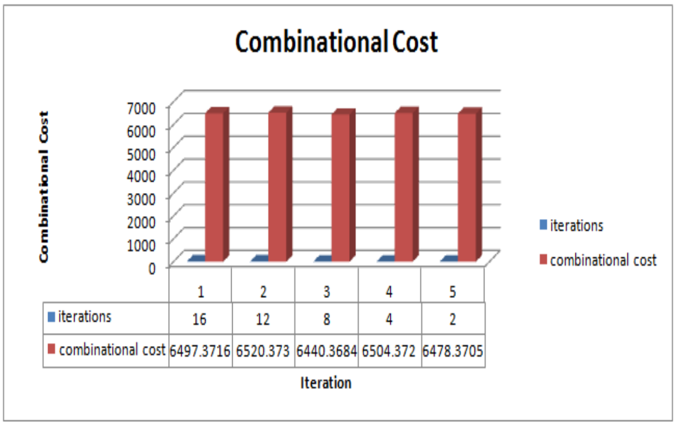


Figure 5.10: Shows combinational cost

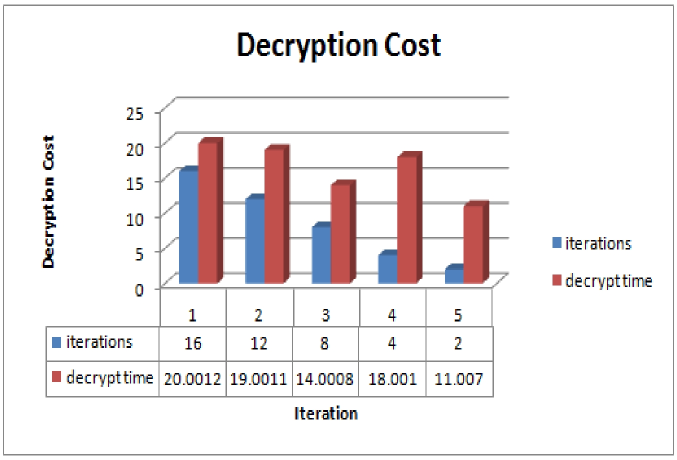


Figure 5.12: Decryption Time

6. Conclusion

From the above experimental setup, the communication cost and combinational cost is optimize in small data entries. To apply these techniques in lager database the Dynamic FP-Growth algorithm is beneficial for mining. It provides security with cloud storage and mining with FP Growth algorithm. This experimental setup is applied for small database file size and is produced a optimize results.

7. Future Work

For lager records, the dynamic FP-growth Algorithm is used for mining which overcome the limitation of Apriori and FP-Growth algorithm.

References

[1] Tamir Tassa.“Secure Mining of Association Rules in Horizontally Distributed Databases,” In *IEEE Transactions On Knowledge And Data Engineering,*

[2] R. Fagin, M. Naor, and P. Winkler.”Comparing Information Without Leaking It,” *Communications of the ACM*, 39:77–85, 1996

[3] Murat Kantarcıoˇglu and Chris Clifton,” Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data,” in *Ieee Transactions On Knowledge And Data Engineering, To Appear, 29 Jan. 2003; revised 11 Jun. 2003, accepted 1 Jul. 2003.*

[4] D. W.-L. Cheung, V. Ng, A. W.-C. Fu, and Y. Fu, “Efficient mining of association rules in distributed databases,” *IEEE Trans. Knowledge Data Eng.*, vol. 8, no. 6, pp. 911–922, Dec. 1996.

[5] D. Agrawal and C. C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms,” in *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. Santa Barbara, California, USA: ACM, May 21-23 2001, pp. 247–255. [Online]. Available: http: //doi.acm.org/10.1145/375551.375602

[6] Y. Lindell and B. Pinkas, “Privacy preserving data mining,” in *Advances in Cryptology – CRYPTO 2000*. Springer-Verlag, Aug. 20-24 2000, pp. 36–54. [Online]. Available: http://link.springer.de/link/service/series/0558/bibs/1880/18800036.htm

[7] C. Yao, “How to generate and exchange secrets,” in *Proceedings of the 27th IEEE Symposium on Foundations of Computer Science*. IEEE, 1986, pp. 162–167.

[8] L. Torgo. Data Mining with R: Learning with Case Studies. Chapman & Hall/CRC, 2010.

[9] L. Van Wel and L. Royakkers. Ethical issues in web datamining. Ethics and Inf. Technol., 6:129–140, 2004.

[10] M. Kantardzic. Data Mining: Concepts, Models, Methods and Algorithms. John Wiley & Sons, Inc., 2002.

[11] Pallavi Dubey, “Association Rule Mining on Distributed Data”, *International Journal of Scientific & Engineering Research*, Volume 3, Issue 1, January-2012 1 ISSN 2229-5518

[12] M. H. Dunham.“Data Mining. Introductory and Advanced Topics,” Prentice Hall, 2003, ISBN 0-13-088892-3.

[13] A. Schuster and R.op Wolff , "Communication-Efficient Distributed Mining of Association Rules," *Proc. ACM SIGMOD Int'l Conf. Management of Data,* ACM Press, 2001,pp. 473-484.

[14] D.W. Cheung, et al., "A Fast Dis *Distributed Information Systems,"* IEEE CS Press, 1996,pp. 31

Author Profile

Mr. Ambarish Durani is MTech pursuing in Computer Science & Engineering From Vidharbha Institute of Technology, Nagpur. Nagpur University (M.H.), India. His area of interest is Data Mining

Prof. Vinay Kapse is M.E. in WCC from G. H. Raisoni, Nagpur. Nagpur University, (M.H.), India and now he is working as Assistant Professor in CSE Department, Vidharbha Institute of Technology, Nagpur. His area of interest is Data Mining and Algorithm