Data fragmentation using the technique NFA to DFA

Charu Chauhan, E. Poovammal
Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India.
chauhancharu08@gmail.com; poovamme@srmist.edu.in

Abstract

Today's business environment needs to manage large quantities of data which has enabled a need for a distributed database, as it provides efficient data processing. Data security is a major aspect as the benchmark for confidentiality, integrity and availability of data remains relatively high. In this paper we propose a technique for data security of a file where the file is fragmented, clustered together and store on the server using the non-deterministic finite automata To reconstruct a file the deterministic finite automata technique is used to fetch the file fragments, reorganize and download the whole file.

Keywords: *Fragmentation, data security, distributed database.*

1. Introduction

Many organizations have important data that is stored daily. Many entities like educational and financial institutes, governmental organizations have important information that needs to be stored and secured.
The security of these entities type of database is the main concern today. Data security is to maintain the confidentiality of the stored data and secure database from any of the external attackers or any of the illegal access to the data. Database security's main concern is to permit and prohibit the user's actions to be done on it.
Sensitive data that are stored on the different servers have to be protected from the external attacks
Sensitive data can be the bank records or the debit or credit card details
There are many issues which are challenging because of

a) lack in the security mechanism
b) limitation in the domain
c) data manipulation by the client
d) system domain restrictions

Data stored on the cloud is cost-effective but also has high-security challenges. There is the possibility to manipulate or hack the data when the communication or transfer of the data happens, also can be attacked by the external attackers. Therefore, the user can be concern about their data privacy.

Data security in the cloud is provided by the encryption algorithm where the data is encrypted by the keys. Different types of keys like the public and private keys are used to secure the data. Any of the users can access the data by providing the keys. Encryption algorithm has more additional storage because of the more extra bandwidth which increases the costs. Which can be one of the reasons that all the cloud are not encrypted.
This paper is proposing the method of fragmentation of the data. Used in the distributed database which is the NoSQL (query optimization is not required)

Distributed database

A distributed database is that which is spread on different networks of computers and also that do not share the physical components of the system. It is the division of the large and single database [21][into the smaller parts and distributed over the different computer connected in the network this helps in reducing the communication costs and increase the performance in terms of the fast retrieval time. There [16] are some of the drawbacks with the distributed database which is the security of the data, maintenance of the database is high. Data integrity in the disturbed database is difficult as it is shared among the various networks and query optimization is also the issue of it.

Types of the distributed database

Homogenous Database

In the homogenous database [17] all the database stored at different sites are identical to each other i.e.) the data structure, operating system all are identical at all the network sites. It is easy to manage

Heterogenous Database

The different network sites may have different software or schema which makes the query processing slow.[17] The different computer present at different network sites may have different operating system, different data structures so for the communication different translation is required.

Distributed Databases have replaced the centralized database in every information processing sector, for example, the educational institutes, health care centres, etc. In the distributed database the data is divided into all the network sites and not stored in the single server.

Applications with large volumes of data can effectively improve the efficiency of distributed processing of the data on database management systems (DBMS). In the distributed database system, the database is available on the various sites but not physically connected. The data available are logically connected have some relationship which is according to the structural query processing and also the transparency is maintained, as connected in a large network.

The two major techniques of distributed databases are Fragmentation and Allocation.

Fragmentation

Data fragmentation was found late in the 70s.

Fragmentation is the concept where the huge data file is broken into many of the smaller files and saved. Fragmentation provides data security and also allows multiple access to the data at the same time. The communication cost is low in the data fragmentation as it opts for the concept of parallel computing. Fragmentation also leads to the Reduction in the disk access as irrelevant data access also reduces.

Fragmentation is classified as horizontal fragmentation, vertical fragmentation and hybrid fragmentation.

Horizontal fragmentation is the division of a relation of the table into tuples (rows) subsets. Each fragment is a subset of tuples of a relation. And also, each fragment has the location to store at the different nodes each fragment is present with the unique row.

Vertical fragmentation is the division of the relation of the table into attribute (column) subsets. The different column has different and unique fragmentation. The partition has done to get the small fragments of the relation.

Hybrid fragmentation is the combination of both the technique vertical and horizontal. first the horizontal fragments then the vertical fragments are generated from the single or more horizontal fragments the same is done in the case of vertical fragments. There is an allocation process also after the file in the database is fragmented then it is assigned to the different server(nodes) in the distributed database. For the backup it depends on the user to replicate the fragment and store on the different servers on maintaining it as the single copy. Benefits of replication of file fragments is that efficiency is increased and also the lost data fragment can be retrieved increase the available data and the performance of the system

2. Literature review

Katarzyan [1] The paper deals with the various data security challenges that are exposed. The author proposed here the architecture that deals with data confidentiality when the information is been shared with others. Architecture introduced the two storage of the data and a communication channel. The data stored is fragmented and stored on different cloud storage. This increases the security of the data; the fragments are also of public fragments and private fragments the public fragments will be stored on the different clouds whereas the private fragments will be secured. While downloading the fragments the public fragment is sent and when the user request it the private fragment is sent through the separate communication channel which is secured through which receiver will be able to collect all the fragments and download the file.

Asma H. AL-Sanhani [2] the author did the comparative study of the data fragments in the distributed database. Fragments are the division of the relation into the many and stored while to restore the file it can be done without the loss of any kind of the information. Some of the advantages of the fragment is to increase the efficiency, security can also be maintained. A horizontal fragment is that where it contains all the information of the query site. Vertical fragments that contain the all information of the particular column and hybrid are the mixture of both the technique. comparison was done of the fragments which was that the condition of a tuple is true for vertical fragment and false for the horizontal and mixed fragment.

Shahidul Islam Khan [3] the author proposed the algorithm where the fragmentation technique is done at the initial stage and a later stage in the distributed database. At the initial stage, the horizontal fragmentation is done by considering the attribute locality precedence which informs the important attribute in the distributed database the value of it is considered from the CRUD matrix which is the create, read, update and delete it makes it easier to locate the data in the table which can indicate the row and the column of the entities.

The author presented in the paper the proper fragmentation at the initial stage by using the method where the query execution is not happening. The allocation process is also synchronized with fragments which leads to no complexity in the allocation process in the distributed database. Performance increased as there is no remote access and also the transfer of the data.

Amjad Alsirhani [4] the author proposed the method to increase the data security and confidentiality of the data by using the encryption algorithm and the fragmentation in the distributed database system. The different types of the encryption algorithm are used as per the queries also initially the database is encrypted and it is stored in the mater cloud and any of the keys are not disclosed. To get the required tuple only one of the columns is kept as the open (unencrypted) so that
desire relation can be fetched by querying it.

P.Ravi Kumar [5] the author proposed the different cloud computing model that deals with the data security of the cloud. The cloud models are cloud service consumer, cloud service provider these are the cloud models where the cloud service consumer is that where the service done by the cloud. Some of the data security issues authentication which is not properly and issues related to the confidentiality, integration, and authentication of the data where it is on the cloud.

To overcome these issues the method was introduced software-defined networking which deals with the design of the software and the network of the abstract application which protects the data and its privacy.

Chun-Hung Cheng [6] the author proposed the different clustering techniques for the data partition. The partition is done by formulating the travel salesman problem. Which has shown the results of the for both the technique of the fragmentation.

Himel Dev [7] the data security of the cloud is the main concern of every cloud user. The flexibility of the services on the cloud has a risk. The author proposed the architecture in the distributed system which can eliminate the risk. The architecture allows that all the data is not stored at the same cloud the data is broken or split and stored at the multiple nodes. It is made difficult for the hacker to access all the cloud where the data is stored, and also the exact information can not retrieve. The author used the concept of the RAID (redundant array of independent disks) which help in maintaining the data and also it is lower at cost. The author used the technique to increase the security of the data that is fragmentation, categorization and distribution which also maintains the privacy of the data. The mining of the attack is done to increase the performance of the system which also leads to the frequently asked of the data which can be stored separately so access time is less comparatively.

Nelson Santos [8] the author used the different techniques of the fragmentation which is applied to the cloud which restricts the unauthorized access of the user. In the implementation process the pattern fragmentation is done where the file is split and it is stored with the index number and the length of the split file is calculated so that the reconstruction process of the file happens.to mask the length of the file chunk the header bytes were also padded to it and then uploaded to the cloud. The benefit of doing this is that the attacker would not know that which is the header file and which is the chunk of the file as both are of the same length. In this paper for the encryption process the AES (Advanced Encryption Standard) algorithm is used for the encryption the original file is encrypted and the cipher text is divided in the split file and it is arranged according to the random pattern fragmentation and the index value is noted.at the time of the downloading the file all the split file chunk are first downloaded and read in sequence till the time all the files are extracted and later the cipher text is recreated to match and maintain the data integrity and security.

Yves Deswarte [9] the author used the technique of the fragmentation, redundancy, and also of scattering by using these techniques the data which is destroyed or lost due to the accidental fault could be recovered. This technique helps to maintain the data security of the data also the integrity maintained across the distributed database system.

Intrusion detection is the lager number of attackers attacking the particular file of data. Intrusion target can be the confidentiality, integrity, and availability of the data, so to control this the different technique is used firstly the sensitive data is fragmented and scattered to the different sites in the distributed system. The backup or the replication of the fragment file is done to recover it when the accidental fault happens. This process makes it complicated for the header to hack and collect the whole information of the file.

Sandhiya Hari Kumar [10] the author proposed the technique of the fragmentation which is the hybridization which is done by applying the algorithm known as the subspace clustering. Hybrid fragmentation is the combination of both the vertical and horizontal technique where the partition is done combining the attributes as well as the tuples. Clustering of fragments is done when the dimensions are found in the subspace algorithm of all the fragments the relevant dimension fragments are clustered together according to the relevant set of tuples and set of attributes. The method is considered to be better with a very large database as relevant fragments are cluster together.

The distance between clustered is calculated using the Manhattana distance where all the records are nearby or fall in the same distance and the average of it is taken and see if no cluster is overlapping to other by this the subspace dimension can be found by the algorithm to store the fragments.

3. Methodology

In this paper we aim to increase the security of the data stored on different databases by the concepts of data fragmentation and NFA to DFA technique. This is the alternative method for data security and the process is faster. Here the large data file is broken or fragmented into the multiple smaller files and these file fragments are grouped but no file in the consecutive order is grouped together. These files are stored on the different server (nodes) of the database. This technique. The attacker would not be able to reconstruct the whole file if in any case the fragment is attacked still would not able to access the whole file. The proposed method allowed to store at the different grouped fragments to the various servers which increases the security of the data

A. Fragmentation

The fragmentation is the technique that splits the original file into small fragments as determined by the user. As mention in the figure1.The user has an option from the interface where a user can fragment the file according to the need. The fragments are given the ID special fileID is the name which is then stored. In this paper we have used the horizontal technique for the fragmentation. The horizontal technique is that which takes the records from the database table and storing it at a different server. The fragments that are stored on the different servers are of the same size which eliminates the adding of the padding bits to the fragmented file. The proposed method determines the correct order of the fragmented file and stored on the different nodes for the increase in the performance of the machine.

B. Clustering

In this paper clustering means as the grouping together. When the file is fragmented they are grouped together and stored on the different servers. The grouping of the file fragments are done in such a way that no consecutive file fragment is stored together by storing such a way security of the file increases. By any chance if the folder (which is grouped together) is hacked still the information which is been hacked will be of no use as it is not interlinked with each other fragmented file which is grouped. Once the file is fragmented and grouped together then the folder is uploaded and saved on the various servers seen in figure 2.

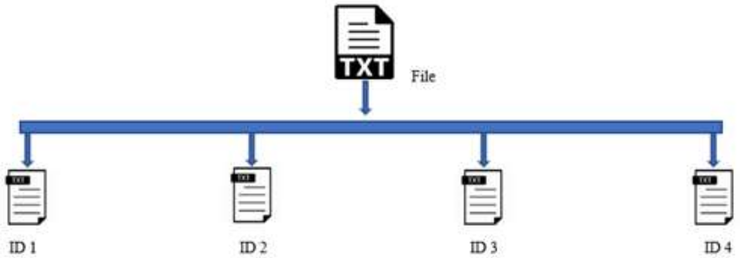

FIGURE 1: Fragmented file

We also proposed the interface for this method.

Where the user can be allowed to give the unique file ID for every original file and the fragment size so that all the fragments are of the same size and after that have an option to upload the file to the different server and the also have an option to download the whole file.

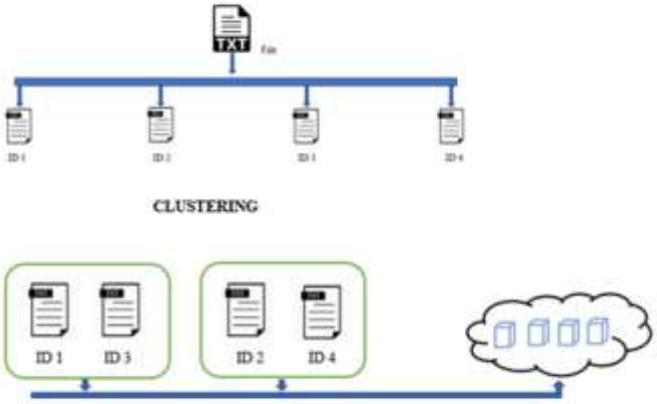

FIGURE 2: Clustering

During the process of download, a file the client has an option in the interface to download the file. The fragments can be downloaded on the client machine where the fragments are aligned and organized aging back according to the unique file Id number which is stored on the lookup table.

Once the fragments are is correct order the original file can be downloaded on the client machine device. Due to the unique file Id number store on the lookup table the process is faster to de-fragment it by which the performance of the machine is increased.

C. NFA and DFA

Finite automata are the simplest machine which accepts the regular expression or language which is used to compute in the many model machine. In this paper we have used the deterministic finite automata and non-deterministic finite automata which are the regular language. Non deterministic finite automata is the 5 tuple abstract machine which consists of (*i)* finite number of state (ii) finite number of input alphabets or symbols (iii) next state is the transition state which is the power state of both finite number and symbol. (iv) initial state (v) final state with the acceptance. An NFA can reach to the multiple state transition at the same time and using the same string of the transition. If there is a string from a1..a9 and there a transitions from the initial state to the final state still there will be transition on the path that contain the label from the a1 to a9. At the same time many transitions can happen time required by NFA is less. Whereas in the deterministic finite automata there will be only one path of the transition of the input string.

Deterministic finite automata refer to the uniqueness in the computation. If the alphabet in the string is present the transition is accepted or else it is rejected. The deterministic finite automata are faster as compared to the non-deterministic finite automata but it requires a large amount of memory.

Lookup table its constructed to store all the transition of the DFA. A state label is used as the to mark as the address point of the transition. The table is used to store the pattern of the location pointed by the common output state.

In the paper we proposed that the fileID which is stored for each of the file fragments is used as the input to the string-matching pattern of the non-deterministic automata which is stored to the different servers and the entry of each fragment is stored in the lookup table for the reference. The file fragments are stored on the server and the backup of the fragments is also stored on the server. It is that on the half server of the total the clustered file is stored, and half is for the backup so, if any case there is an accidental fault also will be able to download and re-organize the whole file.

4. Result

The interface for the client is created by which the file can be uploaded and downloaded with not using the no- sql database. The client should give the unique fileID and indicate the size of each fragment to be stored and can upload the file. To reconstruct the file download option is available all the fragments of the file can be downloaded and reorganize.

In DFA only one active state at a time as it can be match or do one transition at the time of input alphabet. NFA is constructed as it can match many alphabets of the string at a time.

The non-deterministic conversation is required as the flow to fetch the file should be from many to one. Storing the grouped file to many servers using the NFA and fetching only one file fragments to reorganize it. As the non-deterministic can have many flow operations which can lead to the complex situation and time consuming is also more in such case. For the conversion of the NFA to DFA the powerset algorithm is used based on which the transition state for the DFA is automatically constructed.

Conversation from NFA-to-DFA does not lead to growth of the larger state of the machine.

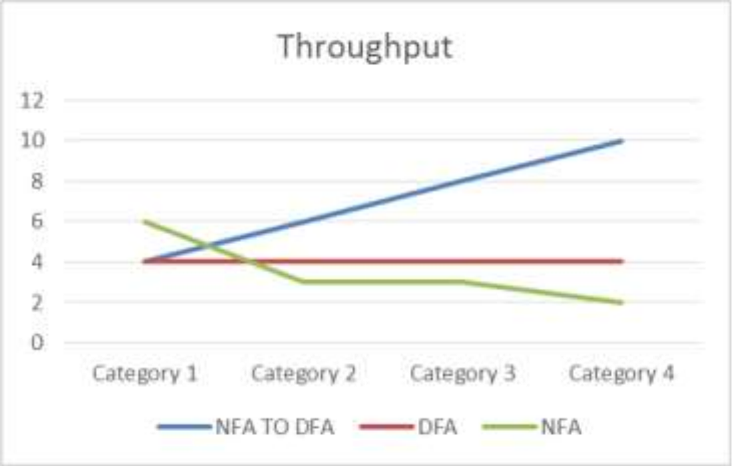

FIGURE 3: Throughput

The throughput of the technique is compared in the figure 3, which tell about the different throughput from the graph it is seen that the conversion of NFA to DFA is more. Which we have used in our research to implement.

Data security is done by many ways. The different type of the encryption algorithm is used by which the keys is been generated and passed to the client after many researches also in almost all the encryption algorithm key is passed to the client. The encryption algorithm provides the security to the file but it increases the complexity and this affect the efficiency of the system. Which may impact drastically the memory usage of the system.

The proposed fragmentation technique is considered to be better in the case of the data protection, recovery. Data recovery is not possible it is been stored at different servers and getting the whole file is a cumbersome task for the user without having the whole information about the file.

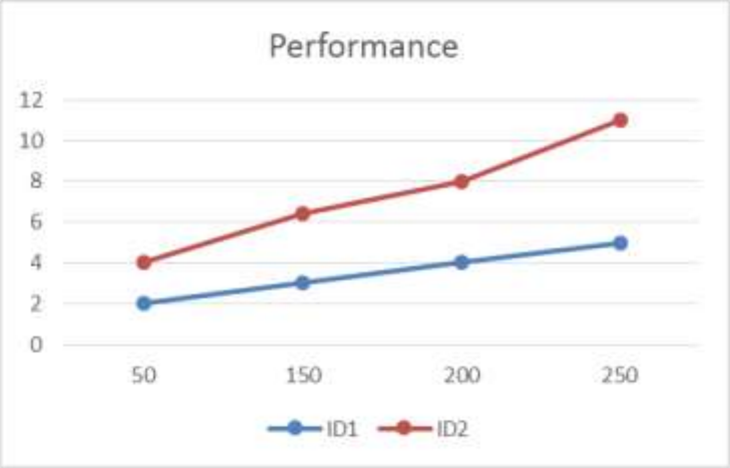

FIGURE 4: performance of file

As seen in the above chart of the two files. The more the number of the fragments done for the file gives the more data protection to it. For the higher high throughput and protection level a fragment number should be more and of the same size.

The proposed method aims to provide the alternative method for the data protection without the much need of the resources and not making the computational complex and making it ideal use for the different environment.

5. REFERENCES

1. Kapusta, Katarzyna, Han Qiu, and Gerard Memmi. "Secure Data Sharing by Means of Fragmentation, Encryption, and Dispersion." *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2019.
2. Al-Sanhani, Asma H., et al. "A comparative analysis of data fragmentation in distributed database." *2017 8th International Conference on Information Technology (ICIT)*. IEEE, 2017.
3. Khan, Shahidul Islam, and A. S. M. L. Hoque. "A new technique for database fragmentation in distributed systems." *International Journal of Computer Applications* 5.9 (2010): 20-24.
4. Alsirhani, Amjad, Peter Bodorik, and Srinivas Sampalli. "Improving database security in cloud computing by fragmentation of data." *2017 International Conference on Computer and Applications (ICCA)*. IEEE, 2017.
5. Kumar, P. Ravi, P. Herbert Raj, and P. Jelciana. "Exploring data security issues and solutions in cloud computing." *Procedia Computer Science* 125 (2018): 691-697.
6. Cheng, Chun-Hung, Wing-Kin Lee, and Kam-Fai Wong. "A genetic algorithm-based clustering approach for database partitioning." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 32.3 (2002): 215-230.
7. Dev, Himel, et al. "An approach to protect the privacy of cloud data from data mining-based attacks." *2012 SC Companion: High Performance Computing, Networking Storage and Analysis*. IEEE, 2012.

8. Santos, Nelson, et al. "Performance analysis of data fragmentation techniques on a cloud server." *International Journal of Grid and Utility Computing* 10.4 (2019): 392-401.
9. Deswarte, Yves. "Fragmentation-Redundancy-Scattering: a means to tolerate accidental faults and intrusions in distributed systems." *Proceedings of the ERCIM Workshops, INESC, Lisbonne (Portugal)*. 1991.
10. Harikumar, Sandhya, and Raji Ramachandran. "Hybridized fragmentation of very large databases using clustering." *2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*. IEEE, 2015.
11. Košař, Vlastimil, Martin Žádník, and Jan Kořenek. "NFA reduction for regular expressions matching using FPGA." *2013 International Conference on Field-Programmable Technology (FPT)*. IEEE, 2013.
12. Rodríguez, Lisbeth, and Xiaoou Li. "A dynamic vertical partitioning approach for distributed database system." *2011 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2011.
13. Wiese, Lena. "Clustering-based fragmentation and data replication for flexible query answering in distributed databases." *Journal of Cloud Computing* 3.1 (2014): 18.
14. Al-Sayyed, Rizik MH, et al. "A new approach for database fragmentation and allocation to improve the distributed database management system performance." *Journal of Software Engineering and Applications* 7.11 (2014): 891.
15. Castro-Medina, Felipe, et al. "Design of a Horizontal Data Fragmentation, Allocation and Replication Method in the Cloud." *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2019.
16. Kumar, Raju, and Neena Gupta. "An extended approach to non-replicated dynamic fragment allocation in distributed database systems." *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*. IEEE, 2014.
17. Ramachandran, Raji, Dhiti P. Nair, and J. Jasmi. "A horizontal fragmentation method based on data semantics." *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2016.
18. Abdalla, Hassan I. "A synchronized design technique for efficient data distribution." *Computers in Human Behaviour* 30 (2014): 427-435.
19. Nashat, Dalia, and Ali A. Amer. "A Comprehensive Taxonomy of Fragmentation and Allocation Techniques in Distributed Database Design." *ACM Computing Surveys (CSUR)* 51.1 (2018): 12.
20. Khan, Shahidul Islam, and A. S. M. L. Hoque. "A new technique for database fragmentation in distributed systems." *International Journal of Computer Applications* 5.9 (2010): 20-24.
21. Kaundal, Gurpreet, Sukhleen Kaur, and Sheveta Vashisht. "Review on Fragmentation in Distributed Database Environment." *IOSR Journal of Engineering (IOSRJEN)* 4.03 (2014): V6.
22. Özsu, M. Tamer, and Patrick Valduriez. *Principles of distributed database systems*. Vol. 2. Englewood Cliffs: Prentice-Hall, 1999.
23. https://www.tutorialride.com/distributeddatabases/distributed-databases-tutorial.htm
24. https://en.wikipedia.org/wiki/Centralized_da database