Exploration of File Fragmentation Techniques in Cloud Data Center

[1]Charu Chauhan, [2*]Poovammal. E

[1,2]Department of Computer Science and Engineering
SRM Institute of Science and Technology, Kattankulathur, India – 603203
[1]chauhancharu08@gmail.com, [2]poovamme@srmist.edu.in

Article Info
Volume 83
Page Number: 18893 - 18898
Publication Issue:
March - April 2020

Article History
***Article Received**: 24 July 2019*
***Revised**: 12 September 2019*
***Accepted**: 15 February 2020*
***Publication**: 30 April 2020*

Abstract

information storage was a difficult task in the history of human mankind and before the modernized computer system exits. Then the data was stored in the different file system in the separate location which was later combined together into one unit called as the database. The different type of the database are available like relational database, object oriented database. Client server architecture was introduced where the single system was centralized. Distribution database is that where the database is divided into the smaller parts and divided among all the different network sites. The fragmented framework plays the important role in providing the security to the data present on the various databases. In this paper different types of fragmentation techniques such as horizontal fragmentation, vertical fragmentation and hybrid fragmentation were analyzed considering important aspects. Also, this paper recommends the cost effective best model.

Keyword: *distributed database, fragmentation, centralized database*

1. Introduction

Information storage was a difficult task in the history of humankind and before the modernized computer systems exists [4]. Computer technology has a speedy growth from past three decennium. In order to manage huge quantities of data, the numerous techniques are available. In client server architecture the data is primarily collected and handled by centralized servers [8].

The centralized server has the benefits of safety, flexibility and has control over substitute also it has the benefits of safe recovery of the data. Anyhow, [15] the high communication costs is one of the major drawback of client server architecture, when the customer is far off and frequent communications are made. The second drawback of client server architecture is that in case of system failure or bottleneck situation on a single server, there is lack of availability.

Distributed Databases have replaced the centralized database in every information processing sectors for [17] example the educational institutes, health care centres etc. In the distributed database the data is divided to 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, 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 large network.

The two major techniques of distributed database are Fragmentation and Allocation. [15] Fragmentation is dividing the large file into smaller set of files which is known as the fragments. Allocation is the process where each of the fragments are allocated in the memory space. When there is an immense growth of the data in the database various fragmentation techniques [17] makes the number of the available server insufficient. With enormous database volumes and constant updates in place, researchers are forcing to evolve now a days, one cannot predict the manner in which information is transmitted during development time. Periodic reorganization of data is required to reduce the costs of interaction and to recover information efficiently and quickly. So, cloud computing is used.

Cloud Computing [19] is becoming more relevant, and in the science and industry it has been receiving growing attention. In a study by Gartner cloud computing has been seen as the first among 10 leading technologies and by companies and organizations with a better prospect over the next few years.

Cloud computing offers universal, efficient, on-demand of network access to configure the computing resources that can be accessed and released with minimum operational or service interactions [22] (e.g. network, databases, storage systems, applications and services). Cloud Computing appears for both a computing and a distributed architecture.

The data fragmentation framework, in a cloud storage system collectively addresses the security and performance of recovery time [22]. Even the latest technologies can sometimes fail to provide adequate confidentiality and privacy, as service suppliers are likely to take advantage of customers’ trust and benefits of transferring the personal data to other parties [15].

The transition of sensitive data to Cloud Service providers often ensures that these services have to rely on data to be adequately protected from outside attacks. Information is stored on several complex virtual servers throughout the Cloud in cloud bases. The data provide is fragmented before it is distributed to multiple servers for the security of the data from the various attacks.

2. Related Terms and Research Works

The paper discusses about fragmentation methods which are in practice and relevant advantages and issues of each method. Also, we addressed few of those issues. Hence, introducing various related terms and research works attempted and successfully implemented, in the next few sections.

Database

There has been a lot of development of databases in the 20th century to meet the needs of information storage and recovery.[21] It began with simple file systems, stored at separate locations individually. Nonetheless, huge amounts of data were not very successful in storing due to replication problems, separation and isolation of data, data dependence and incompatible file formats.

Favourable circumstances related with the database were quicker and shared access, information uprightness and information consistency and so forth. Later database structures were introduced by IBM’s [15] IMS (Information Management System) in 1966 was the main business various levelled database framework. It gave the highlights of simple information refreshing, quick recovery, different affiliation and effortlessness of structure. And also provide the characteristics like updating of the data was easy and fast, structure was simple. Also, it has the restriction of the data replication. Large amount of computer storage made the scope of hierarchical databases limited.

Then the graph structured objects had come [18] which was supported by the database which is object-oriented database. This database had many features like encapsulation, identity of the object, inheritance. The problem with these databases was that mixing of the conceptual relationship with storage

Later the relational database was introduced [15-16] where tables from which the data can be retrieved and reassembled in many ways without the knowledge of the tables in the database and the application interface of the relational database is the structured query language.

Centralized Database

A centralized database is that where all the connections of the systems are made to the single server and the processed information is also stored at that single server. The single server is known as the centralized server.[18] Any modification on the data, which can be data retrieval or storing of the data should be done on the centralized server only. The data modification will automatically be done to all the system connected to it.

Centralized database has the advantages [21] that data can be easily accessed, it is portable as all the information is stored at single location, maintenance of the database is easy and required less power, data redundancy in minimum.

The drawbacks of the centralized database, [19] if the access of the data is more may create a bottleneck situation. Simultaneous access of the data by the multiple user can lower the efficiency of the system. The failure in the centralized system may lead to destroy of the data present in the database measure should be taken to back up the data
in the database

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 different computer connected in the network this helps in reducing the communication costs and increase the performance in terms the fast retrieval time.

There [16] are some of the drawbacks with the distributed database which is 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 issues of it.

Distributed Database may be homogenous or heterogeneous by nature. In the homogenous database [17], the database stored at the different sites are identical to each other. That is, the data structure, operating system all are identical in all the network sites and hence it is easy to manage. In the heterogeneous database, the different network sites may have the different software or schema which makes the query processing slow.[17] The different computer present at different network sites may have the different operating system, different data structures, so for the communication different translation / transformation is required.

To start with the research works presented in this domain**,** Aniqa Rehman [1] proposed the architecture which deals with the payroll processing system which is a continuous processing functions and on state transition diagram. The adjustments can be made in payroll processing system which is the effective way of business approach as it provide the flexibility and enhances the productivity of the business.

The work done in this payroll system is different because the author here presented the unified modelling language and also the non-deterministic finite based models, and also provided formal specifications by the method of VDM-SL which provides the reliability and also the method is used for the correctness of a system.

Sandhya Harikumar et al [2], the authors have proposed in their research work, the strategy for allocating the fragments which are grouped together. For the query answering it focuses on the duplication of the fragments However, major work here focuses on the, how the attributes and the tuples are used in the queries. To produce the fragment from the similar pattern of the data is easy.

The author proposed the subspace algorithm which is used for fragmenting the data for the effective query processing. Algorithm also improves the processing time duration of query in the distributed database system. Also improving the way the fragmentation is done and as foundation of the allocation of those fragments. The main work done by the author are, on the basis of the projected clustering a hybridized fragmentation is done of the global relation and the proper merging of the fragments for allocation purpose.

Vlastimil Kosar et al [3] the author proposed hardware architectures which is presented on the Deterministic Finite Automata (DFA). In his work, the size of the memory and also the speed is limited as the determinization may lead in the growth in the number of the states and also in the transition tables. Hence, it is based upon the Nondeterministic Finite Automata (NFA) because it is limited by the size of the field programmable gate array chips. The transition table is mapped on the field programmable chip logic present growth of the attacks, worms and viruses. Network Intrusion Detection Systems matches the regular expression which means field programmable logic increases due to the regular expression and also due to the network links.

The author Lisbeth Rodríguez et al [4] proposed the architecture for the dynamic vertical partitioning for the database. The system of database considered was distributed in nature and the statistics are collected. After analysing the statistical information, which is collected by monitoring the queries, decision is made to see whether the new portioning is required or not. The author proposed the algorithm for the vertical partitioning and re-fragmentation of database without intervention of a database administrator.

Dalia Nashat et al [5] proposed the anti-instantiation. The instance is used where the similar data can be found in the distributed database. Author has used the clustering algorithm to fragment the data in the database. The query mechanism used supports the replication of the data but it limits the overgeneralization. The problem of data replication can be expressed as a special bin packing problem

The database partitioning is done using the horizontal technique by Rizik M. H. Al-Sayyed et al [6]. The relations are split into different fragments which are disjoint by nature. The generation of the disjoint fragments are done by removing all the common attributes from the two fragments. By this technique, all the disjoint fragments which are newly formed do not overlap. Then technique proposed by the author was the network sites clustering which is clustering the sites for the network that has the common physical property.

Clustering together can increase the performance and also the communication cost. For the fragmentation allocation process first fragments are allocated to all the clusters. If the fragments show the positive value for the cluster then its allocated to the cluster and if the value is not positive then the fragment is allocated to different cluster.

Castro-Medina et al [7] proposed algorithm which can solve the overlapping problem. Overlapping happens in the horizontal technique for the fragmentation of the data and also replication of the data is the problem which is solved in the distributed database. The algorithm is used get the predicate matrix where the rows of the matrix is set of predicates from which the fragmentation is done by this when the fragments are returned it will return the tuples as well

Raju Kumar et al [8] analysed, two way of allocating the fragmentation they are static or the dynamic way here the author proposed the different approach to fragments that are dynamically allocated and also which are not replicated in the database system also the distance factor is introduced. The algorithm for the dynamic fragment allocation used to fine the factor like threshold value, time constraints to access and volume. The author extended the algorithm for the new factor distance which named as TTVDCA (threshold constrains, access time constraints, volume and distance constraints).

The proposed algorithm has two step first is the preparation phase and the active phase. In the preparation phase the distance matrix is there which keeps the information about the distance of fragment that is present on the site to the distance of fragment on the another site the matrix also contains the information about the fragment threshold value for the fragments when it is reallocated and also the access time constraints of data fragments when it is relocated. Fragments are allocated to different sites using the static method then each of the sites where the fragmentation is present the information about it is store in the access log.

In active phase the distance is calculated of the fragments from the initial site, if the distance is more than the reallocation of the fragment happens by relocating the fragments to the nearer sites so, the travel would be less which increase the efficiency also for the allocation of the dynamically allocated fragments.

Raji Ramachandran et al [9] the author proposed the different method for the horizontal fragmentation using the clustering technique. The clustering technique is where the similar kind of the data are grouped together that forms one cluster it is the fast and efficient way to access the similar kind of the data sets, the clustering of the data is done by many algorithm in data mining but the author uses the K- type algorithm to cluster the data.

The cluster are fragmented but there can be chance that the cluster are grater in number than the network sites. If the cluster are greater than the network sites more time is taken to fragment it. So, the similar kind of the cluster are again grouped together this is done by calculating the Jaccard coefficient. By this method the information can be retrieve fast and query processing is faster.

Hassan I. Abdalla et al [10] the author proposed heuristic approach which is used for the horizontal fragmentation which finds the optimal solution by retrieving the attribute and also by updating it for the allocation problem. The heuristic approach was introduced the cost of the fragmentation is evaluated by collecting the retrieval information and update information it is found by the query the higher the cost for the fragmentation that is selected for the allocation .by this method the data transfer cost is reduced and the remote access.

3. Fragmentation

Data fragmentation is the broken of the large data into many small pieces of data which are not similar to each other.

The technique of fragmentation are vertical fragmentation, horizontal fragmentation and hybrid fragmentation.

A. Vertical Fragmentation

The division of the relation of the table into attribute (column) subsets.[15] The different column has different fragmentation which are unique. The partition done to get the small fragments of the relation.

Table 1: vertical fragmentation

Name	Reg no	course	Dept
Fragmentation 1	Fragmentation 2	Fragmentation 3	

B Horizontal Fragement

It division of the a relation of the table into tuples (rows) subsets.[14] Each fragment is a subset of tuples of a relation. And also, each fragment has the location to store at the different node each fragment is present with the unique row

Table 2: Horizontal Fragmentation

Name	Reg no	course	Department
Fragmentation 1			
Fragmentation 2			
Fragmentation 3			

C Hybrid Fragmentation

In this technique [2] both fragmentation technique is used in the combination of horizontal and vertical fragmentation. In this technique first the horizontal fragments then the vertical fragments are generated from the single or more horizontal fragments same is done in the case of vertical fragments

Table 3: Hybrid fragmentation

Name	Reg no	Course	Department
Fragmentation 1	Fragmentation 3	Fragmentation 4	
Fragmentation 2	Fragmentation 5		

Case study on the horizontal fragmentation

The complicated problem can be solved when its divided into smaller subproblem same is the fragmentation technique that is dividing the data [12] review a cost model for the proper horizontal fragmentation and allocation the modal have two condition in first condition all the constrained are normally executed in the second condition all the constraints were imposed which modal identify correct partitioned and assigning of the fragments with the memory allocation is done

Table 4: fragments and allocation

Fragments	S1	S2	S3	S4
F1	0	0	0	1
F2	1	1	1	1
F3	1	0	1	1

Case study on the vertical fragmentation

The same modal is selected [12] here again the fragmentation and the data is done and allocated but in the vertical fragmentation after it is done the similar kind of the data is grouped together. First the fragments are created, recognized and then clustered.

Table 5: cluster of fragments

cluster	S1	S2	S3	S4	S5	S6
C1	1	1	1	0	0	0
C2	0	0	0	1	1	0
C3	0	0	0	0	0	1

Case study on the horizontal fragmentation

The author Harikumar and Ramachandra (2015) proposed the algorithm subspace clustering the algorithm first fragmentation were created when the tuples and attributes are correlated then they were grouped together into the cluster for the allocation purpose.

It was concluded [11] that fragmentation technique only concentrates on the fragmentation of data and allocation is not done properly in the vertical and horizontal technique as it reduces the complexity, but using the hybrid technique the static and the dynamic allocation can be done.

4. Conclusion

We have discussed about the different fragmentation techniques and allocation process. And we were able to arrive at the conclusion that the horizontal fragmentation is suitable when the user wants to retrieve the tuple value from the relation. The vertical fragmentation is suitable when the user wants to retrieve the attribute from the relation. The hybrid fragmentation is to be preferred when user requirement is more on the security of data rather than the response time.

References

[1] Rehman, Aniqa, and Nazir Ahmad Zafar. "NFA based formal verification of automatic payroll processing system." *2016 International Conference on Emerging Technologies (ICET)*, IEEE, 2016

[2] 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.

[3] 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.

[4] 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.

[5] Wiese, Lena. "Clustering-based fragmentation and data replication for flexible query answering in distributed databases." *Journal of Cloud Computing* 3.1 (2014): 18.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Abdalla, Hassan I. "A synchronized design technique for efficient data distribution." *Computers in Human Behavior* 30 (2014): 427-435.

[11] 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

[12] 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.

[13] Kaundal, Gurpreet, Sukhleen Kaur, and ShevetaVashisht. "Review on Fragmentation in Distributed Database Environment." *IOSR*

Journal of Engineering (IOSRJEN) 4.03 (2014): V6.

[14] Marwa, F. F., I. E. Ali, and A. A. Hesham. "A heuristic approach for horizontal fragmentation and alllocation in DOODB." *Proc. INFOS2008* (2008): 9-16.

[15] Verma, Sunil Kumar. "Fragmentation techniques for distribution database: A Review." *International Journal of Innovative Computer Science & Engineering* 3.2 (2016): 47-50.

[16] Özsu, M. Tamer, and Patrick Valduriez. *Principles of distributed database systems*. Vol. 2. Englewood Cliffs: Prentice Hall, 1999

[17] Armbrust, Michael, et al. "A view of cloud computing." *Communications of the ACM* 53.4 (2010): 50-58.

[18] https://www.tutorialride.com/distributeddatabases/distributed-databases-tutorial.html

[19] https://en.wikipedia.org/wiki/ Centralized_database