Development of a Hybrid Query Processing Algorithm in a Distributed Database Management System.

Dawodu, Adekunle .A., Badmus, Yahyah . A., Bello, Abass M. & Dawodu, Ayokunle Mayowa
[123]Department of Computer Science, D.S Adegbenro ICT Polytechnic, Itori, Ewekoro, Ogun State, Nigeria
[4]Computer Engineering Department, Olabisi Onabanjo University, Ago - Iwoye, Ogun State, Nigeria
E-mails: [1]alandawodu@gmail.com; [2]badmusyahyah@gmail.com; [2]abassholatunde142@gmail.com
Phone Nos: +2348055141696/ +2348132327110; +2347056778672; +2348067287113; +2349152640016

ABSTRACT

A database refers to a collection of logically related information organized so that it can be easily accessible, managed and updated. Database are generally accused electronically from a computer system and are usually controlled by a database management system. Query processing on a distributed database system requires transmission of data between computers on a communication network, then minimizing the amount of data transmission is one of the fundamental principles to reduce the query processing cost and to prevent network congestion. In a distributed database environment a user can access and modify data from several database scattered across numerous sites by means of a distributed database management system. Typically, distributed database operate on two or more interconnected servers on a computers. In conclusion, distributed and parallel database systems have become important tools for data intensive applications and with rapidly growing prominence of these database as a result of organizational and technical reasons, while the arrangement of data transmissions and local data processing is known as a distribution strategy for a query. The inability to obtain an effective query strategy with an efficient accuracy and minimum response time or cost to execute the given query is one of the major key issues of the query optimization in distributed database systems, further inefficient database compression methods, inefficient query processing missing indexes, inexact statistics and deadlocks are further defects.

Keywords: Query Processing, Database Management Systems (DBMS), Distributed Database Systems (DDMS), Hybrid Algorithm, Data Transmission.

1. BACKGROUND OF THE STUDY

Distributed database systems have become important tools for data intensive applications and with rapidly growing prominence of these database as a result of organizational and technical reasons. They are many inherent problems in centralized architectures, consequently distributed database have become a solution to those complications (Uk essays, 2018). Database are generally accessed electronically from a computer system and are usually controlled by a database management system (DBSM) (Sebastian, 2023). A distributed database (DDB) is a group of different, conceptually connected databases that are dispersed across computer network (Waterlloo, 2021). The software that administers the DDB and offe4rs an access method that makes this distribution transparent to users is known as a distributed database management system.

A DDBMS governor the storage and processing of logically related data over interconnected computer systems. One of the advantages of a DDBMS is that the data can be located near the site with the least demand, another one is growth facilitation. According to BOSKO (2021), a distributed database represents multiple interconnected database spread out across several sites connected by a network utilize multiples nodes, scale horizontally and develop a distributed system. Since the database are all connected, they appear as a single database to the users. Therefore, more nodes in the system provide more computing power, offer greater availability, and resolve the single point of failure issue.

1.1 Distributed Database Features

Some general features of distributed database are (BOSKO, 2021)

i.   Location Independency: Data is physically stored at multiple sites and managed by an independent DDBMS.

ii.   Distributed query processing: Distributed database answer queries in a distributed environment that manage data at multiple sites. High – level queries are transformed into a query execution plan for simple management.

iii. Distributed transaction management – provides a consistent distributed database though commit protocols, distributed concurrency control techniques, and distributed recovery method in case of many transactions and failures.

iv.   Seamless integration: Database in a collection usually represent a single logical database and they are interconnected.

v.   Network linking: All database in a collection are linked by a network and communicate with each other.

vi.   Transaction Processing: Distributed database incorporate transaction processing. Which is a program including a collection of one or more database operations. Transaction processing is an atomic process that is either entirely executed or not at all.

1.2 Types Of Distributed Database

There are two distinct types of distributed databases:

i.   Homogenous Distributed Database

This is a network of identical database stored on multiple sites. The sites have the same operating system, DDBMS and data structure, making them easily manageable.

Homogenous DDBs allow users seamless access to data from each of the databases. AN examples of a homogenous DDB is Shown in figure 1.1a
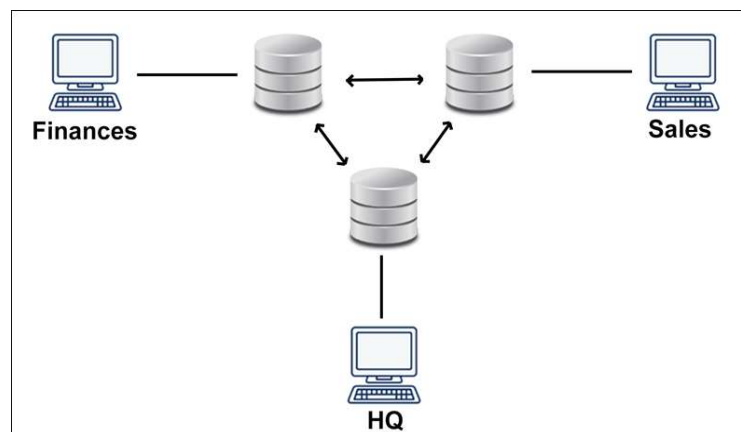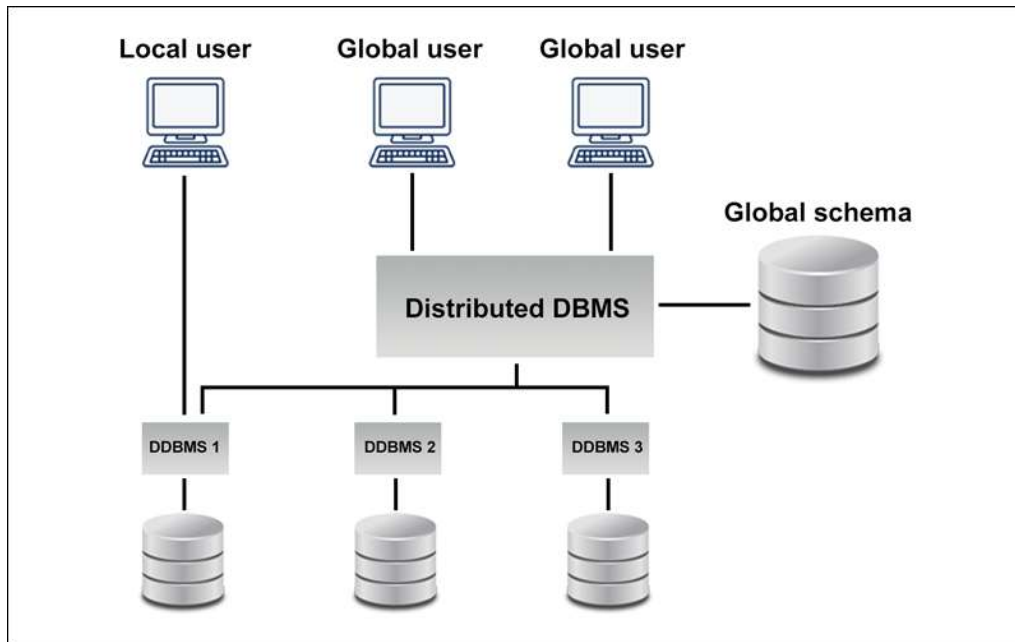

Figure 1.1a: Schematic Representation of Homogeneous Distributed Database
(Source: Bosko, 2021)

ii.   **Heterogeneous Distributed Database**

A heterogeneous distributed database uses different squamous operating system, DDBMS and different data models in which a particular site can be completely unaware of other sites causing limited cooperation in processing user request. The limitation in why transactions are required to established communication between sites. An example of heterogeneous database is shown in figure 1.1b

Figure 1.1b: Schematic Representation of Heterogeneous Distributed Database
(Source: Bosko, 2021)

1.3 Architecture Of Distributed DBMS.

In a distributed database environments, a user can access and modify data from several database scattered across numerous sites by means of a distributed DBMS. Instead of being arranged as general purpose computers, distributed DB system. Architecture uses dedicated servers for each site. In a distributed system, companies employ various servers such as applications servers and database servers, for various purposes (UK essays, 2018). For example, a bank implement database system on different computers in the network of the bank with computer systems located at the different branches providing communication between the systems using the network link of the bank, in this scenario, local DBMS is allowed to access single site while DDBMS have access to several sites. The major components of a DDBMS are: Network software and hardware, computer workstations, communication media, transaction processor, data manager.

The main steps involved in distributed query processing are decomposition, localizations and optimization. In decomposition step. It generates query tree for given structured query language. These relations are replaced by fragments on localization step. The process of reducing cost of a tree is done in optimization steps, the steps are as follows.
   i.   Localization
   ii.   Distributed Query operators
   iii. Cost based optimization

1.4 Statement of the Problem

Researchers in the field have attempted to find solutions towards reduction of complicity in query processing in DDBMS and results have so encouraging as result of ever increasing volume of data in today's ICT world and the need to process data in a distributed system environment.

It is in the light of this that – a hybrid query processing algorithm in a distributed database management system (DDBMS) is being proposed.

1.5 Aim and Objectives

The aim of this research is to develop a hybrid query processing algorithm in a distributed database management system (DDBMS) and the specific objectives are:

  i.   Carry out an extensive survey of the existing technique and algorithm existing query processing algorithm using the localization approach.
  ii. Design the algorithm based on the localization techniques from above.
  iii. Develop the algorithm using the design in (ii) above and software tools.
  iv. Test the developed algorithm using standard performance metrics.

2. LITERATURE REVIEW

Query optimization in distributed database management systems is a process of producing a plan for the processing of a query to a distributed database system. The plan is called a query execution plan, Bressan (2009) according to Shyarn et al., (20015), the query optimization problem in large-scale distributed database is NP-hard in nature and difficult to solve because the complexity of the optimizes increase as the number of relations and numbers of joins in a query increase. Research and projections have continually been carried out to find an appropriate algorithm to seek an optimal solution especially when the size of the database increase. Azhir et al., (2022) proposed the relations needed by a query be kept in several locations which significantly increases potential corresponding query execution Plans (QEP's) for a user query which does not sound logical to investigate all potential query plans in a high setting like this. They concluded the best query plan must be generated for processing a given query in a distributed database systems.

In a multiple client/single server (DBMS) a database is accessed by more than one client and may possibly lead to instances of looks. In multiple-client /multiple-server, database (DDBMS) access is distributed across many servers. So in order to process a user queries, servers should communicate with each other according to the request made by the user, peer to peer is the advanced architecture which requires each host to behave as client and server (Ukessays, 2018). Distributed query processing is the procedure of answering queries (which means mainly read operations on large data sets) in a distributed environment where data is managed at multiple sites in a computer network (Satler, 2009). It was further asserted that query processing involves the transformation of a high-level query (e.g, formulated in SQL) into a lower-level query operators as well as the execution of the result. The goal of the transformation is to produce a plan which is equivalent to the original query (returning the same result) and efficient, ie, to minimize resources consumption like total costs or response time (Satler, 2009).

Query processing involves processing user queries and retrieving data from the distributed database system which is a complex task as data may be stored across multiple sites and may need to be combined to answer user queries (Akashkumal 17, 2023). According to (Tarunsingwap 7 2023) query processing in a distributed database management system requires the transmission of data between the computers in a network using a query distribution strategy of ordering data transmissions and local data processing in a database system.

In a distributed database system, schema and queries refer to logical units of data. The process of query optimization in distributed systems involves optimal utilization of resources in the distributed system. The current distributed database management system are subject to some problems, such as the complexity of management and control.

3. METHODOLOGY
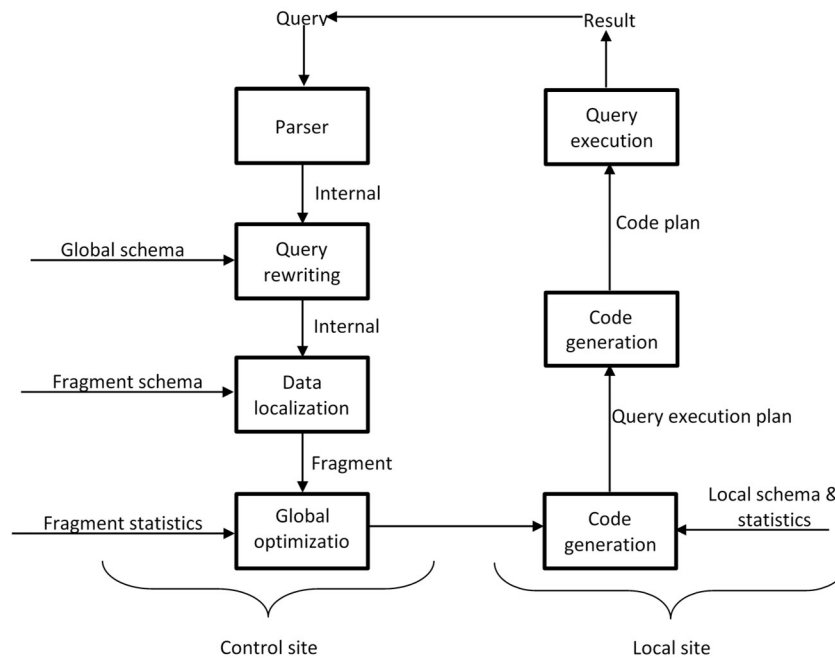
Fig. 3.1a; Distributed Query Processing
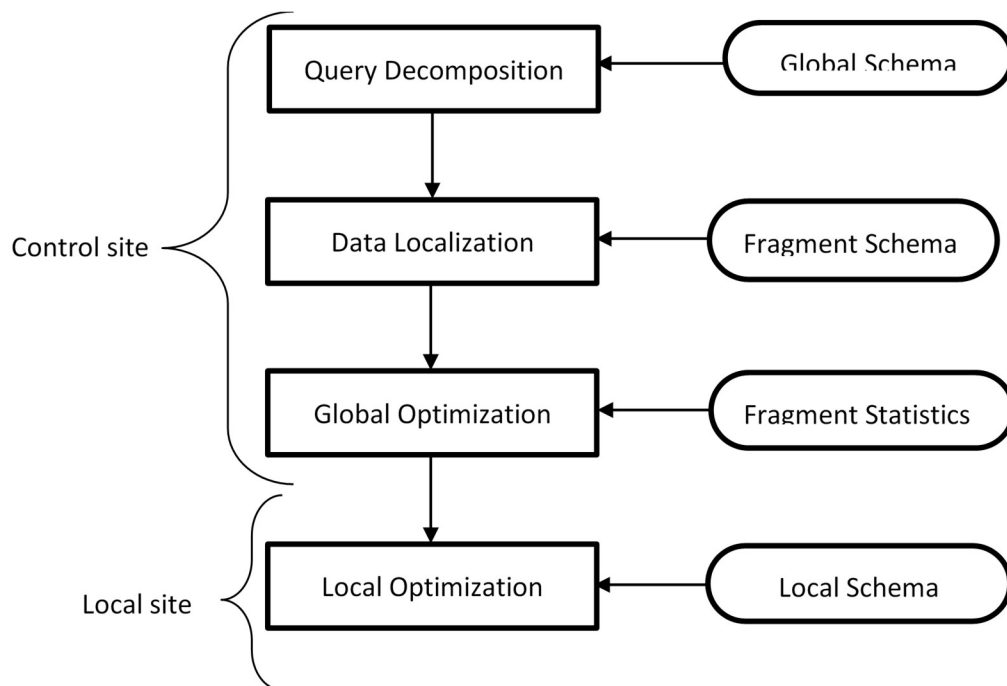
Fig. 3.1b; Generic Layering Scheme for distributed query processing

The interest of the diagrams above is to find how each algorithm performs in a general query environment. The advantage obtained from hybrid algorithm is not taken away by the reprocessing the required to formulate the schedule.

4. DISCUSSION OF FINDINGS

The shift of paradigm from the centralized to its counterpart Distributed Database Systems have brought enormous opportunities and problems along with it. In this present world of globalization, distributed database system is no more an obsession but a need of almost every organization to remain competitive and more importantly to survive in this competition because of critical performance problem, query processing has always been the centre of attraction. This study also addressed query processing in distributed database systems.

This study is mainly focused on the development of hybrid query processing strategies for distributed database systems which utilized the database statistics to generate the hybrid algorithm program. The study has presented an overview of an existing hybrid processing algorithms and their downfall.

5. CONCLUDING REMARKS

This research work has presented hybrid query processing algorithm for database optimization. The proposed development of a hybrid query processing algorithm in a distributed DBMS techniques gives a methodology to generate an efficient distributed query processing plan which improves the reply time of user queries. These methodologies achieved the distributed query processing plan generation as a single objective for hybrid algorithm problem. At the same time, take advantage of the parallelism of the algorithm itself and distributed database cluster environment to accelerate the algorithm convergence further. Finally, this showed that the results have efficiency in the hybrid algorithm for database processing optimization.

6. SUGGESTIONS FOR FURHER WORKS

This paper is needed to be improving more in the unique of database and query optimization processing systems, such as conducting more deep analysis and evaluation of design entropy based stochastic query optimizer by considering the impact of a variety of selection approaches on development of hybrid query processing algorithm. Furthermore, results of each of approach can be compared with each other results in a very effective and efficient way and to conduct further research on the impact of data allocation of optimize the wage and the process of the system. In conclusion, query processing in World Wide Web (WWW) is probably the most interesting, challenging and form parts of future work in this field.

REFERENCES

1. Sebastian T. (2023). Database. https://corporatefinanceinstitute.com/resources/data-science/database
2. Waterloo, 2021; Distributed Database Management Systems. https://cs.uwaterloo.ca/~tozsu/courses/cs856/F02/lecture-1-ho.pdf
3. Bosko M.. (2021). What Is a Distributed Database? https://phoenixnap.com/kb/distributed-Database
4. UKEssays. (2018). A Distributed And Parallel Database Systems Information Technology Essay. Retrieved fromhttps://www.ukessays.com/essays/information-technology/a-distributed-and-parallel-database-systems-information-technology-essay.php?vref=1
5. Sattler, KU. (2009). Distributed Query Processing. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_704
6. tarunsinghwap7 (2023). Query Processing in Distributed DBMS. https://www.geeksforgeeks.org/functions-of-distributed-database-system
7. Akashkumar17 (2023). Functions of Distributed Database System. https://www.geeksforgeeks.org/functions-of-distributed-database-system
8. Bressan, S. (2009). Distributed Query Optimization. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_708
9. Shyam P., Sushant K., Akhilesh ., Parth K. (2015). Query Optimization Strategies in Distributed Databases.
10. Yadav, Pramod Kumar & Rizvi, Sam. (2018). Query Optimization: Issues and Challenges in Mining of Distributed Data. 10.1007/978-981-10-6620-7_67.
11. Azhir, E., Navimipour, N.J., Hosseinzadeh, M. (2022). Join queries optimization in the distributed databases using a hybrid multi-objective algorithm. Cluster Computing https://doi.org/10.1007/s10586-021-03451-9