ABSTRACT

During the last three decades, distributed database technology has emerged as one of the most significant development in the field of database systems. Distributed database technology has become an integral part of most of the business organizations due to its decentralized nature. Distributed databases have eliminated many of the shortcomings of the centralized databases and it fits more naturally in the decentralized structures of many organizations. All the major vendors of database systems nowadays support distributed database technology. So, the design of distribution database is an important area of research.

The design of centralized database has two main issues: designing the conceptual schema and designing the physical database. But the design of distributed databases adds two more issues: designing the fragmentation of global database and allocation of fragments over network. Distributed database design is an optimization problem which needs to address the three key issues: fragmentation of the global database, fragments allocation and replication of the fragments. All these issues complicate the design of distributed database. The fragmentation of the database is a complicated issue in itself. Different techniques have been proposed for fragmenting the database by different researchers. The present study concentrates only on fragments allocation problem assuming that database is already fragmented. The fragmentation and query optimization strategies are not within the scope of the present research work.

The data allocation problem involves finding the optimal placement of the fragments to the sites of the communication network. The optimality can be defined with respect to two different measures: minimal cost and performance. The cost of allocation consists of the cost of storage of data and the data transmission cost i.e. retrieval cost and update cost. The allocation problem attempts to find a fragments allocation schema that minimizes the combined cost function. Minimum response time and maximum system throughput at each site are two different ways to check the performance of distributed database system. The
allocation problem attempts to find an allocation strategy that maintains the performance criteria.

The objective of the thesis is to generate data allocation framework for distributed database design so that the total cost of processing a query can be minimize and the performance of the distributed database system can be increased. Data in distributed database system is allocated according to two different types of access patterns: static and dynamic. In a static environment, the access probabilities of application running on different site to fragments never change but in a dynamic environment these probabilities change over time. The present research work is divided into two parts: Static allocation of data and Dynamic allocation of data.

Biogeography-Based Optimization and Simplified Biogeography-Based Optimization techniques have been used for both non-replicated and replicated static allocation of data. To show the performance of the proposed algorithms, results of the proposed algorithms for static data allocation are compared with the genetic algorithm. Both the new proposed algorithms are giving quality solutions within a shorter period of time as compared to genetic algorithm.

A new heuristic algorithm named Threshold and Time Constraint Algorithm has been proposed for non-replicated dynamic allocation of data. The proposed algorithm re-allocates data with respect to the changing data access patterns with time constraint. The proposed TTCA for non-replicated dynamic allocation of data in distributed database system is an extension of existing two approaches: Optimal algorithm and Threshold algorithm. This new algorithm decreases the movement of data over the network and also improves the overall performance of distributed database system.