Optimal Design of Distributed Databases

L.P. Daswin Pasantha De Silva and Manjula Dissanayake
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka

Abstract— **The physical expansion of enterprises with the current trends in globalization has had a positive influence on the technologies of distributed systems. At the forefront of this technological revolution are distributed database systems. For a distributed database to be at optimal performance and thus provide an efficient service it needs to be designed appropriately. The significance of the perfect design is only emphasized by the multiple dimensions required in generating a design. The purpose of this paper is to suggest an approach to generate optimal designs for such distributed database systems and to develop a prototype to demonstrate the said approach. The approach emphasizes on the accuracy of inputs as it largely determines the quality of the final solution. Hence the extraction of network information, a key dimension, is automated to ensure precision. The global schema is fragmented considering data requirements as well as connectivity of each site. Allocation of fragments is treated as a combinatorial optimization problem and assigned to a memetic algorithm. An estimation of distribution algorithm complements the search effort of this memetic algorithm. Site options for replication server environments are investigated based on a shortest path algorithm. Usability of the system in an object oriented development environment, through conditional object-relational mapping, is also explored. The prototype was developed using an evolutionary prototyping approach. It was evaluated by several experts in the relevant fields of application. The results of which, confirmed the practicality of the suggested approach.**

Keywords—**Distributed Database Design, Memetic Agorithms, Object-Relational Mapping, Bandwidth Measurement**

I. INTRODUCTION

Data are ubiquitous just as they are indispensable. All information systems regardless of the application area, user skill level, language or country of use require data to successfully complete its purpose. The creation of the relational model made way for the database paradigm as a suitable storage and access mechanism for data. Expansion of enterprises on the grounds of globalization prompted the hitherto centralized database technology to be geographically distributed along with the disseminated enterprise operations, thus leading to the distributed database technology.

A distributed database is a collection of multiple, logically interrelated databases distributed over a computer network [1]. As accentuated by the definition it is the fusion of two conceptually different technologies, database systems which realize centralization and networking which realize decentralization. The product of this combination in essence can be stated as the best of both worlds. The complication that is to harness the benefits of both the parent technologies mentioned above ensures the importance of the design stage for any successful distributed database system. Thus, a good design although intricate, is mandatory. Various factors, both technical and non technical, need to be considered in the production of a good design. Ranging from the global schema to the disk capacity of a server, much information needs to be extracted from the distributed environment.

The key concerns when devising a distribution solution are fragmentation, replication and allocation. Fragmentation arises with the need to access different data of a single relation, in a global schema, by more than one site in the distributed environment. In such instances the relation is broken into several disjoint sets. Replication is the storage of copies of relations at different sites as per the access frequencies. Replication can be extended to fragments as well as complete schemas. Allocation is the task of identifying the best locations for replicas, fragments and relations so that the performance, of the environment as a whole, is maximized and the costs incurred are minimized.

Documented academic research on this problem domain although available, is individually inadequate. Identification of only a subset of the problem domain, unrealistic assumptions on the behavior of the distributed environment, homogeneous application and over simplification of certain constraints are some notable limitations. The focus of this paper is a complete and integrated solution for optimal design of distributed databases.

An overview of the sections to follow; Section 2 further explores distributed database design along with existing research and techniques. Section 3 goes into detail on the suggested methodology while section 4 describes the implementation of the prototype and the features offered. Future enhancements are discussed in Section 5.

II. EXISTING RESEARCH AND TECHNIQUES

As accentuated in [1], there exist two key strategies for fragmentation. Horizontal fragmentation; where division is based on tuples and vertical; where division is based on