Chapter VII

Swarm Quant’ Intelligence for Optimizing Multi-Node OLAP Systems

Jorge Loureiro
Instituto Politécnico de Viseu, Portugal

Orlando Belo
Universidade do Minho, Portugal

ABSTRACT

Globalization and market deregulation has increased business competition, which imposed OLAP data and technologies as one of the great enterprise’s assets. Its growing use and size stressed underlying servers and forced new solutions. The distribution of multidimensional data through a number of servers allows the increasing of storage and processing power without an exponential increase of financial costs. However, this solution adds another dimension to the problem: space. Even in centralized OLAP, cube selection efficiency is complex, but now, we must also know where to materialize subcubes. We have to select and also allocate the most beneficial subcubes, attending an expected (changing) user profile and constraints. We now have to deal with materializing space, processing power distribution, and communication costs. This chapter proposes new distributed cube selection algorithms based on discrete particle swarm optimizers; algorithms that solve the distributed OLAP selection problem considering a query profile under space constraints, using discrete particle swarm optimization in its normal (Di-PSO), cooperative (Di-CPSO), multi-phase (Di-MPSO), and applying hybrid genetic operators.
INTRODUCTION

Nowadays, economy, with globalization (market opening and unrulement) shows a growing dynamic and volatile environment. Decision makers submerged into uncertainty are eager for something to guide them, in order to make timely, coherent and adjusted decisions. Within this context a new Grail was born: information as condition for competition. Data Warehouses (DW) emerged, naturally, as a core component in the constitution of organization’s informational infrastructure. Their unified, subject oriented, non-volatile and temporal variability preserving vision, allowed them to become the main source of information concerning business activities.

The growing interest on DW’s information by knowledge workers has motivated a fast enlargement of the business enclosed areas. Also the adoption of Data Warehousing (DWing) by most of Fortune 400’s enterprises has helped to make the huge size of today’s DW (hundreds of GB or even tenths of TB). A query addressed to such a database has necessarily a long run-time, but they must be desirably short, given the on-line appanage characteristic of OLAP systems. This emphasis on speed is dictated by two orders of reasons: 1) OLAP users’ need to take business decisions in a few minutes, in order to accompany the fast change of markets, operated in short time intervals; 2) the strong dependence of the productivity of CEO’s, managers and all knowledge workers and decision makers of enterprises in general, on the quickness of the answers to their business questions.

However, this constant need for speed seems to be blocked by the huge amount of DW data: a query like “show me the sales by product family and month of this year related to last year” may force a scanning and aggregation of a significant portion of the fact table in the DW. This is something that could last for hours or days, even disposing, hypothetically, of powerful hardware and suitable indexes.

The adoption of a DWing “eagger” (Widom, 1995) approach allowed to solve this problem through the generation and timely updating of the so called materialized views, summary tables or subcubes (mainly used from now on). In essence, they are Group By previously calculated and stored by any kind of dimensions/hierarchies’ combinations. These subcubes need space and especially time, enlarging the size of the DW even more, perhaps one hundred times bigger, since the number of subcubes may be very large, causing the well-known “data explosion”. So, it is crucial to restrict the number of subcubes and select those that prove to be the most useful, due to their ratio utilization/occupied space. This is, in the essence, the views selection problem: selecting the right set of subcubes to materialize, in order to minimize query costs, characteristically NP-hard (Harinarayan, Rajaraman, and Ullman, 1996).

Two constraints may be applied to the optimization process: the space that is available to cube materializing and the time disposable to the refreshing process. But multidimensional data continues growing and the number of OLAP users too. These concomitant factors impose a great stress over OLAP underlying platform: a new powerful server was needed or simply empower the architecture with the aggregated power of several small (general purpose) servers, distributing multidimensional structures and OLAP queries through the available nodes. That’s what we called Multi-Node OLAP architecture, shown in Figure 1.

A number of OLAP Server Nodes (OSN) with predefined storage and processing power, connected through a network, using real characteristics of communication inter-node links, which may freely share data or issue aggregation queries to other nodes participating in a distributed scenario, constitutes the M-OLAP component, where inhabit the multidimensional structures. This system serves a distributed knowledge worker community, which puts a set of queries on their daily routine. This brings to