Chapter VI
Selecting and Allocating Cubes in Multi–Node OLAP Systems: An Evolutionary Approach

Jorge Loureiro
Instituto Politécnico de Viseu, Portugal

Orlando Belo
Universidade do Minho, Portugal

ABSTRACT

OLAP queries are characterized by short answering times. Materialized cube views, a pre-aggregation and storage of group-by values, are one of the possible answers to that condition. However, if all possible views were computed and stored, the amount of necessary materializing time and storage space would be huge. Selecting the most beneficial set, based on the profile of the queries and observing some constraints as materializing space and maintenance time, a problem denoted as cube views selection problem, is the condition for an effective OLAP system, with a variety of solutions for centralized approaches. When a distributed OLAP architecture is considered, the problem gets bigger, as we must deal with another dimension—space. Besides the problem of the selection of multidimensional structures, there’s now a node allocation one; both are a condition for performance. This chapter focuses on distributed OLAP systems, recently introduced, proposing evolutionary algorithms for the selection and allocation of the distributed OLAP Cube, using a distributed linear cost model. This model uses an extended aggregation lattice as framework to capture the distributed semantics, and introduces processing nodes’ power and real communication costs parameters, allowing the estimation of query and maintenance costs in time units. Moreover, as we have an OLAP environment, whit several nodes, we will have parallel processing and then, the evaluation of the fitness of evolutionary solutions is based on cost estimation algorithms that simulate the execution of parallel tasks, using time units as cost metric.
INTRODUCTION

The revolution operated at business environment and technology level motivated Data Warehousing (DWing). Globalization has generated highly competitive business environments, where proper and timely decision making is critical for the success or even the survival of organizations. Decision makers see their business on a multidimensional perspective and, mainly, need information of aggregated nature. These concomitant factors impose a new class of applications coined as On-Line Analytical Processing (OLAP).

The success of the DWing and OLAP concept brings to them an increasing number of new users and more and more business information areas. Its enlargement is a natural consequence: the stored data becomes huge, as well as the number of users. This reality imposes a high stress on the hardware platform, as OLAP query answers might be given in seconds. Some solutions were proposed and implemented, being two of them the most relevant: the pre-aggregation and materializing of queries and the distribution of data structures.

The former is an extension of the DWing concept, as an eager approach (Widom, 1995): why waiting for a query to scan and compute the answer? The aggregation of possible huge detailed data to answer to an aggregated query may take a long time (possibly some hours or days) and then, the pre-computing and materializing of aggregated queries’ answers, denoted as materialized OLAP cubes or OLAP structures, are, certainly, a sine qua non performance condition in OLAP systems.

The second solution is, naturally, another view of the old maxim “divide ut imperes”: as OLAP users increase and structures get huge, we may distribute them by several hardware platforms, trying to gain the known advantages of database distribution: a sustained growth of processing capacity (easy scalability) without an exponential increase of costs and an increased availability of the system, as it eliminates the dependence from a single source. And this distribution may be achieved in different ways: 1) Creating different cubes, each one inhabiting in a different hardware platform: that’s the solution coined as data mart approach; 2) distributing the OLAP cube by several nodes, inhabiting in close or remote sites, interconnected by communication links: that’s a multi-node OLAP approach (M-OLAP); 3) using, as base distribution element, not the subcube, but only a part of it, a component called subcube fragment. Those solutions may be conjunctly applied, building, on its largest creation, the Enterprise Data Warehouse or the Federated Data Marts. But those creations, a materialization of the referred advantages, don’t come for free. Many problems have to be solved, being the most relevant the so called “cube view selection problem”, an optimizing issue derived from the former referred solution: materialization of subcubes.

As the number of these structures (and especially size and refresh cost) is huge, we have to select only the most beneficial ones, based on the maxim that “an unused subcube is almost useless”. Periodically, or almost in real-time, we may decide which of the subcubes are the most beneficial and provide for its materialization and update (possibly adding or discarding some of them). Also, when the distributed OLAP approach gets on the stage, other disadvantages may be pointed: the increased complexity of DW administration and a huge dependency on the proper selection and allocation (into the several nodes) of the data cube, a question which is partially answered by the proposals of this paper: a new algorithm for the selection and allocation of distributed OLAP cubes, based on evolutionary approaches, which uses a linear cost model that introduces explicit communication costs and node processing power, allowing the use of time as the reference cost unit.

Related Content

Data Mining for Obtaining Secure E-Mail Communications
www.igi-global.com/chapter/data-mining-obtaining-secure-mail/10858?camid=4v1a

Data Mining in Diabetes Diagnosis and Detection
www.igi-global.com/chapter/data-mining-diabetes-diagnosis-detection/10603?camid=4v1a

Data Provenance
www.igi-global.com/chapter/data-provenance/10873?camid=4v1a

Data Management in Three-Dimensional Structures
www.igi-global.com/chapter/data-management-three-dimensional-structures/10598?camid=4v1a