An Objective Function for Evaluation of Fragmentation Schema in Data Warehouse

D

Hacène Derrar
USTHB, Algeria

Omar Boussaid
University of Lyon 2, France

Mohamed Ahmed-Nacer
USTHB, Algeria

INTRODUCTION

Data Warehousing (DW) and Online Analytical Processing (OLAP) are becoming critical components of decision support. Analytical queries usually identify business trends rather than individual values; those are much more complex than transactional ones. Processing these queries may take hours and days. To improve performance, several techniques have been developed. We quote materialized views (Chuan & Xin, 2001), indexes (Chaudhuri, 2004), data fragmentation (Bellatreche, 2005; Boukhalfa, 2009), distributed and parallel processing (Furtado, 2004). In this article, we will concentrate on the technique of data fragmentation (also known as partitioning. So, in this article we use both terms fragmentation and partitioning).

Partitioning tables, indexes and materialized views in fragments stored and accessed separately improve significantly data manageability, accessibility and query execution time. Thus, traditional fragmentation techniques and more particularly horizontal and vertical fragmentation, developed in relational DBMS, were applied to the data warehouse.

These approaches are designed from a statistical analysis of more frequent queries based on both qualitative and quantitative information. So, algorithms used to design an optimal partitioning schema are static algorithms. Their entries are bases on workload gathered from data exploitation. If a change occurs in the inputs of these algorithms, they must be rerun to determine a new optimal fragmentation schema. Moreover, these algorithms are based on the clustering principle which is considered as combinatorial problem and requires for its resolution to use heuristics methods. So, in the case of models evolution and / or changes in workload these algorithms become very complicated, or unworkable.

In the context of relational and object oriented databases and in any environment (centralized, parallel, distributed) much of the literature has addressed this issue. Researchers concentrate their work on data redistribution or fragments reallocation in the event of performance degradation. So, it was considered that the solution lies at the physical level by applying load balancing strategies of treatment and data between nodes. The logical aspect, namely the design of the fragmentation schema, itself, remains adapted because the workload is almost stable.

Conversely, in data warehousing the evolution of data model and workload are dynamic. This is due more particularly to the specific characteristics of OLAP queries. So, an inappropriate and badly conceived fragmentation schema have a considerable influence on the system's performance and more particularly during the execution of the expensive operations such as the joint and the multi-joint which characterize the decisional queries. Xinjian (2005) have clearly demonstrated through theorems and lemma that the choice of partition keys and how to arrange the records in the fact table have a great impact on the OLAP queries response time.

For efficient use of fragmentation technique in data warehouse, it is not only to analyze the data access frequencies to choose an optimal fragmentation schema, but to make that choice dynamic and adapted to changing workload.

DOI: 10.4018/978-1-4666-5888-2.ch188

Also, there are no techniques to know that the implemented schema remain the more optimal and the produced partitions are relevant. We propose in this article an approach based on analysis of the OLAP queries execution to evaluate the effectiveness and the relevance of the data fragmentation schema.

BACKGROUND

There are three fragmentation approaches: vertical fragmentation (Bouakkaz, 2012; Navathe, 1984), horizontal fragmentation (Ceri, 1982) and hybrid fragmentation (Gorla, 2012, pp. 559-576; Ziyati, 2006). Vertical Fragmentation (VF) consists in dividing a relation into partitions of different schema, by projection with duplicating the key. It consists in grouping together attributes that are frequently accessed by queries.

Horizontal Fragmentation (HF) consists in dividing a relation into partitions with the same schema using query predicates. Each partition preserves part of the tuples according to restriction criteria. It reduces query processing costs by minimizing the number of irrelevant accessed instances. Two versions of HF are cited by the researchers : primary HF and derived HF. Primary HF of a relation is performed using predicates that are defined on that relation. On the other hand, derived HF is the partitioning of a relation that results from predicates defined on another relation.

Finally, hybrid fragmentation consists of either horizontal fragments that are subsequently vertically fragmented or vertical fragments that are subsequently horizontally fragmented.

All fragmentation approaches, horizontal, vertical or mix, are based at the time of their design on the analysis of statistical data collected starting from the execution of most frequent queries. So, the adaptation of fragmentation techniques to data warehouse proving more delicate because mainly of nature of the OLAP queries. These queries are long, complex and require sometimes a great number of selection operations and aggregation. They can handle hundreds not to say thousands of tuples. The analytical queries are extremely variable, they are made up generally in an interactive way and can be executed once or many times. This type of queries called also ad hoc queries corresponds to queries seized on line without a long preliminary reflection (Gardarin, 2005). All these characteristics makes, with time, the fragmentation schema, inappropriate since it was conceived starting from unstable statistical data.

The works that has addressed this issue suggest approaches for evaluating the benefits of the use of an algorithm over another according to a cost function, which generally covers the time query execution. We quote more particularly the work of Boukhalfa (2009), which evaluates, at the design phase using a cost model, the relevance of fragmentation schema which will be generated. There is not, to our knowledge, an approach that evaluates the "goodness" of the fragmentation schema already implemented to measure the relevance of fragments.

To profit fully from their advantages, the application of fragmentation techniques in data warehoused must be constantly re-examined and adapted to data warehouse exploitation. In this article we propose a formal approach for evaluating a schema fragmentation implemented, that is to say, the relevance of the fragments, based on the access of the most frequent queries.

Our approach consists to evaluating the effectiveness of the existing fragmentation schema. Nevertheless, to measure the quality of a fragmentation schema which one does not know a priori the used objective function is a task which is not always obvious. Furthermore, the majority of the logical design algorithms of fragmentation are directed by the affinity measurement (i.e., the calculation of the queries access frequency only between one pair of attributes which does not allow, consequently, to measure affinity between all the attributes of a partition).

It proves necessary to define an approach making possible to evaluate and measure the affinity of a fragmentation schema in order to study the possibility of determining another more optimal schema. For this purpose, we describe below an approach based on a new objective function for evaluating a fragmentation schema. This objective function will have to be generic and flexible allowing to take possibly into account of another metric such as: the type of queries, the data placement information, the storage capacity and the transfer cost of the data between sites.

EVALUATING A FRAGMENTATION SCHEMA

When data warehouse is vertically or horizontally divided into data fragments considering the access frequencies of OLAP queries, the data stored in frag-