Research Problem in Distributed Data Warehouse Environment

Karima Tekaya, ISI ARIANA, Tunisia; E-mail: Karima.Tekaya@isi.rnu.tn

We can't ignore the advantages realized during the past in the area of distribution and support for data localization in a geographically dispersed corporate structure. Therefore, many researches to date investigate building distributed data warehouses with particular emphasis placed on distribution design for data warehouse environment. This article presents the state of the art concerning many process of data fragmentation in data warehouse environment and proposes a number of technical issues that we believe are suitable topics for exploratory research.

1. INTRODUCTION

The users of Data Warehouses do not cease increasing. They are divided more and more geographically on several sites. In addition, data warehouse are characterized by their large size and by the complexity of decisional query requiring several operations of joins and aggregation. Consequently, centralized Data Warehouses are not adapted more to this kind of companies and generate an elevated cost of query execution. To answer to this new need, several techniques of fragmentation have been proposed for the decentralization of a data warehouse, and the optimization of the cost of query execution. The following section, presents some works treating the distributed architecture for a data warehouse. In section 3, some data warehouse fragmentation techniques which exist in the state of the art are evoked and in the section 4, a number of technical issues that we believe are suitable topics for exploratory research will be presented.

2. ARCHITECTURE OF DISTRIBUTED DATA WAREHOUSE

In [Noaman, Barker, 1997] and in [Noaman, Barker, 1999], authors proposed an architecture for distributed data warehouse. It is based on the ANSI/SPARC architecture that has three levels of schemas: internal, conceptual, and external. This work is based on TOP/DOWN approach and presents two fundamental issues: fragmentation and allocation of the fragment to various sites. Authors proposed a horizontal fragmentation algorithm for a fact table of a data warehouse. In [TEKAYA, ABDELLATIF, 2004], we have proposed a methodology for relational distributed data warehouse design. For this purpose, we develop a set of matrix: 'Matrix of data partitioning', 'matrix of data allocation' and a 'matrix of data source' and in [TEKAYA, ABDELLATIF, 2005] we adapt the same methodology to the multidimensional environment. In [Noaman, Barker, 1997], [Noaman, Barker, 1999], [TEKAYA, ABDELLATIF, 2004] and in [TEKAYA, ABDELLATIF, 2005], a basic architecture of a distributed data warehouse has been proposed; we suggest the implementation of these works in a distributed data warehouse environment.

3. TECHNIQUES OF FRAGMENTATION IN DATA WAREHOUSES

Several works shows the importance of fragmentation in a context data warehouse, it represents today a more challenging stake that in a relational or objects database context. In addition, several commercial products showed the utility of fragmentation in the process of queries optimization: In [TEKAYA, ABDELLATIF, MCSEAI'04], we proposed a matrix of fragmentation, 'Horizontal Matrix of fragmentation ', making abstraction to the approach of modelling, this matrix permits from a logical table to generate a set of fragments and it has as input: queries and their frequencies of use. Thereafter, we proposed a matrix of allocation permitting to allocate every fragment to the most adequate site. One data can be a table or a fragment of table, this as while taking account of frequencies of data utilization and the priority of sites. In [TEKAYA, ABDELLATIF, ISDM'05] the same solution was adapted in a dimensional environment; we have experiment the solution through an example. In [TEKAYA, ABDELLATIF, AICCSA'05] the problematic of data warehouse fragmentation was presented and some arguments are showed to prove the importance of fragmentation in distributed data warehouses environment. Otherwise, several works of research and the commercial products showed the utility of fragmentation techniques in the process of queries optimization [Sanjay and al., 2004]. Horizontal fragmentation in data warehouses is more challenging compared to that in relational and object databases. This challenge is due to the several choices of partitioning schemas that can be found: [Bellatrech and Boukhalfa, 2005]

1. *Partition only the dimension tables using simple predicates defined on this table, t*his choice is not suitable for OLAP queries for the following reasons: Any partitioning that does not take into account the fact table is discarded. [Bellatrech and Boukhalfa, 2005].
2. *Partition only the fact table using simple predicates defined on this table, t*his choice has been adopted in [Noaman and Barker, 1999]. The proposed work is essentially based on:
   a. the proposition of a distributed architecture of a data warehouse,
   b. the formal definition of the relational modelling of a data warehouse by the application of normalization rules,
   c. the replication of dimensions on the different sites of the enterprise,
   d. regrouping of selection predicates on facts tables. (Criteria of fragmentation),
   e. the application of the horizontal fragmentation algorithm presented in [Özsu and Valduriez, 1991] and [S. Ceri and G. Pelagatti, 1984].

   The algorithm generates a set of horizontal fragment based on the definite applications on the dimension tables in link with the table of facts. Note that a fact relation stores foreign keys and raw data which are usually never contain descriptive (textual) attributes because it is designed to perform arithmetic operations. On the other hand, in a relational data warehouse, most of OLAP queries access dimension tables first and then the fact table. This choice is also discarded.
3. *Partition some/all dimension tables using their predicates, and then partition the fact table based on the fragmentation schemas of dimension tables, t*his choice has been adopted in [Bellatreche, Karlapalem and Mohenia,2000] and in [Bellatreche, Karlapalem and Mohania, 2002]. This work aims essentially to:
   a. Propose an algorithm for partitioning dimension tables and the fact table of a star schema, (b) Fragmenting the fact table based on all predicates given in OLAP queries might be prohibitive.

Therefore, authors showed that dimension tables play a very important role in fragmenting the fact table. They develop a greedy algorithm for selecting the best dimension tables for partitioning the fact table. (c) Develop a cost model for executing the most frequent OLAP queries on partitioned and unpartitioned star schemas, finally to evaluate the partitioning algorithm with some experiments study and show the tradeoffs partioned and unpartitioned data warehouse. This approach is best in applying partitioning in data warehouses. Because it takes into consideration star join queries requirements (these queries impose restrictions on the dimension values that are used for selecting specific facts; these facts are further grouped and aggregated according to the user demands. The major bottleneck in evaluating such queries has been the join of a large fact table with the surrounding dimension tables).

An evolution of the same work has been proposed in [Bellatrech and Boukhalfa, 2005]. The proposed work is essentially based on:

1. The Formalization the problem of the horizontal fragmentation in data warehouse environment using star join diagram,
2. The Developpement of a methodology of fragmentation for the fact table, which is based on the genetic algorithms,
3. The Description a coding process of fragmentation diagrams. In order to measure the quality of the chosen solution and a model of cost (selective function) have been developed,
4. The implementation of the proposed solution, by the development of a genetic motor in Visual C and to validate the achieved survey, they used a Benchmark APB-1 IIS releases [Council, 1998].

Authors proposed the development or the planning of the genetic algorithm to take account of the query evolution (structures and frequencies). They also aimed an auto-evolutionary data warehouse.

4. PROBLEMS AND RESEARCH ISSUES

Few works quantified contributions of fragmentation in a data warehouse context. Several axes of researches are opened and several ideas remain to explore. The main objective is to bring kindness of techniques used in distributed databases to the domain of data warehouses and benefit the implication of distribution especially on the process of queries optimization. Applications only work on subsets of relations. It is therefore preferable to distribute these subsets.

We can distribute the complete relations but it would generate a lot of traffic, either a replication of data with all problems that it causes (problems of updating and problems of storage). The small fragment utilization permits to make turn more process simultaneously, what drags a better utilization of the computer network.

Therefore, a problem puts itself: how a good degree of fragmentation to define? In fact, the objective aimed by researches essentially consists in formalizing the problem of the horizontal fragmentation in a data warehouse environment and to propose an algorithm permitting to solve it. This solution has been developed through the application of a genetic algorithm in [Bellatrech and Boukhalfa, 2005]. The solution admits as input a list of dimension tables, fact tables and a list of queries including a set of aggregations functions; as result the algorithm generate an optimized fragmentation diagram. Some authors appraise that the set of queries can changes (in the level of structure or frequency) and therefore it is interesting to develop an algorithm witch take in account the evolution of queries. It becomes especially very important for the evolutionary information systems like a data warehouses. Queries evolves according to the informational needs of the enterprise, this evolution has two types: frequency and structure. One request can become less frequent or same unused during the time, another one becomes more important but change of structure, (change of attribute or condition of restriction). Aggregations functions are very frequent and changes according to needs. Of this fact, we must answer to new requests (utilization of news tables dimension or of facts) witch can be very important for the information system and answers must be imminent for the decision making. Therefore, the static fragmentation diagram cannot answer to all these new constraints. Consequently, works may be oriented to develop techniques that permit to take in consideration the evolution of queries of a data warehouse and generate dynamic fragmentation diagrams.

5. CONCLUSION

Distributed data bases system is sufficiently complete to unload users of all competition problems, reliability, and optimization of requests or transaction on data managed by different DBMS on several sites. Of this fact, we studied the contribution of distribution in domains of data warehouses characterized by their big volume of data and by a number of users distributed geographically more and more. We noted that fragmentation plays today a more important role in a context of data warehouse that in a relational or object context, we put some problematic and oriented readers toward several axes of researches. We synthesized techniques of fragmentation that have been achieved in data warehouse context; we noted that can works be more developed. The domain remains opened for possible researches.

6. REFERENCES

[TEKAYA and ABDELLATIF, 2004] Karima Tekaya, Abdellaziz Abdellatif, 'Modélisation de la répartition des données d'un data warehouse'. Eighth Maghrebian Conference on Software Engineering and Artifical Intelligence, (MCSEAI'04). 2004.

[TEKAYA and ABDELLATIF, ISDM'5] Karima Tekaya, Abdellaziz Abdellaziz, 'Modélisation dimensionnelle de la répartition des données d'un data warehouse'. Journal International des Sciences de l'Information et de la Communication (ISDM05)

[TEKAYA, and ABDELLATIF, AICCSA'05], Karima Tekaya, Abdellaziz Abdellaziz, 'Fragmentation and Replication process in data warehouse environment' The 3rd ACS/IEEE International Conference on Computer Systems and Application (AICCSA'05).

[Noaman and Barker, 1999] A.Y. Noaman, and K. Barker, 'A Horizontal Fragmentation Algorithm for the fact relation in a Distributed Data Warehouse'. Proceeding of the Eigth International Conference on Information and Knowledge Management (CIKM'99). November 1999, Pages154-161.

[Noaman and Barker, 1997] A.Y. Noaman and K. Barker. 'Distributed data warehouse architectures'. Journal of Data Warehousing, 2(2):37-50, April 1997.

[Bellatreche, Karlapalem and Mohenia,2000], Ladjel Bellatreche, Kamalakar Karlapalem, Mukesh Mohenia, 'What can partitioning do for your data warehouses and data marts', 2000.

[Bellatreche, Karlapalem and Mohania, 2002] Ladjel Bellatreche, Kamalakar Karlapalem, Mukesh Mohania. 'Some issues in design of data warehousing systems'. Data warehousing and web engineering". IRM Press Hershey, PA, United States. Pages 22 – 76. 2002.

[Bellatrech, Boukhalfa, 2005] Ladjel Bellatreche and Kamel Boukhalfa, 'La fragmentation dans les entrepôts de données: une approche basée sur les algorithmes génétiques', Revue des Nouvelles Technologies de l'Information (EDA'2005), Juin, 2005, pp. 141-160.

[Bellattreche, Lorinquer and Mohenia, 2004] L. Bellattreche, M. Schneider, H. Lorinquer, et M. Mohenia. 'Bringing together partitioning materialized views and indexes to optimize performance of relational data warehouse'. Proceeding of the international conference on Data Warehousing and Knowledge Discovery (DAWAK'2004), pages 15-25, September 2004.

[Bellatreche, Kamalakar and Mukesh, 2002] Ladjel Bellatreche, Kamalakar Karlapalem, Mukesh Mohania. 'Some issues in design of data warehousing systems'. Data warehousing and web engineering. IRM Press Hershey, PA, United States. Pages 22 – 76. 2002.

[Theodoratos, Sellis, 1999] Dimitri Theodoratos, Timos Sellis , ' Dynamic Data Warehouse Design'. Data Warehousing and Knowledge Discovery . 1999.