Designing an Object–Relational Data Warehousing System: Project ORDAWA*
(Extended Abstract)

Johann Eder[1], Heinz Frank[1],
Tadeusz Morzy[2], Robert Wrembel[2], Maciej Zakrzewicz[2]

[1] Institut für Informatik–Systeme, Universität Klagenfurt
Universitätsstr. 65, A-9020 Klagenfurt, Austria
{heinz, eder}@ifi.uni-klu.ac.at
[2] Poznan University of Technology, Institute of Computing Science
Piotrowo 3A, 60-965 Poznań, Poland
morzy@put.poznan.pl
{Robert.Wrembel, Maciej.Zakrzewicz}@cs.put.poznan.pl

Abstract. In this paper we present a research project aiming at the design and development of an Object–Relational Data Warehousing System (ORDAWA). The project is conducted in co–operation of Institute for Informatics–Systems, Klagenfurt University, and Institute of Computing Science, Poznań University of Technology. Important goals of the project are to develop techniques for the integration and consolidation of different external data sources in an object/relational data warehouse, the construction and maintenance of materialized relational as well as object–oriented views, index structures, query transformations and optimizations, and techniques of data mining.

1 Introduction

Database management systems are widely used by organizations for maintaining and processing data that document their day-to-day operations. In applications that update such operational data, transactions typically automate clerical data processing tasks such as order entry, adding a reservation, or depositing a check. These tasks are structured and repetitive, and consist of short, atomic and isolated transactions. The transactions require detailed, up-to-date data, and read or update few records. Operational databases tend to be hundreds of megabytes to gigabytes in size. Consistency and recoverability of the database are critical tasks, and maximizing transaction throughput is the key performance metric. Consequently, the database is

* This work is supported by the grant No. 8T11C04315 from the State Committee for Scientific Research (KBN), Poland, and by grant No. 1/2000 from the Austrian Office for Academic Exchange.

designed to reflect the operational semantics of the intended applications. Such On-Line Transaction Processing (OLTP) applications have driven the growth of the DBMS industry in the past three decades, and will doubtless continue to be important.

Recently, however, organizations have increasingly emphasized applications in which current and historical data are comprehensively analyzed and explored, identifying useful trends and creating summaries of the data in order to support high-level decision making. Such applications are referred to as decision support systems.

2 The Data Warehouse Approach

Organizational decision support systems require a comprehensive view of all aspects of an enterprise. Information collected in an enterprise is often of different data format and complexity (e.g. relational, object–relational, and object–oriented databases, on-line multimedia data stores, Web pages, spreadsheets, flat files etc.). Therefore, one of the important issues is to provide an integrated access to all these heterogeneous data sources.

There are two basic approaches to data integration: the *mediated* approach and the *data warehousing* approach [15]. In the first one, a global query is decomposed and transformed into queries for each of the data sources. The results of each of the queries are then translated, filtered, and merged to form a global result.

Whereas in the *data warehousing* approach, data of interests coming from heterogeneous sources are extracted, filtered, merged, and stored in an integrating database, called *data warehouse*. The data are also enriched by historical and summary information. Then, queries are issued for this data warehouse.

The advantages of the data warehousing approach to data integration are as follows: (1) queries operate on a local centralized data repository, that reduces access time to data, (2) queries need not be decomposed into different formats and their results need not be integrated, (3) a data warehouse is independent of other data sources, that may be temporary unavailable. However, a data warehouse has to be kept up to date with respect to source data, by periodically refreshing it.

Usually, data warehousing is a collection of technologies aimed at enabling the managers and analysts to make better decision. Data warehousing technologies have been successfully deployed in many industries: manufacturing (for order shipment and customer support), retail (for inventory management), financial services (credit card analysis, fraud detection, risk analysis, etc.), healthcare (for outcomes analysis), etc.

The trend towards data warehousing is complemented by an increased emphasis on powerful analysis tools. There are three basic classes of analysis tools that are emerging. First, DBMSs that are designed to support complex and aggregated queries. Such systems can be regarded as systems applicable for decision support systems. Second, there are systems that support queries involving group-by and other aggregation operators. Applications dominated by such queries are called On-Line Analytical Processing (OLAP). Finally, there are tools for data mining in which users look for interesting previously unknown patterns in the data. Evaluating OLAP or data

mining queries over data distributed globally within the scope of the distributed organization is likely to be extremely slow and inefficient. The natural solution is to create a centralized repository of all the data. Such a repository is called data warehouse. The availability of a data warehouse facilitates the application of OLAP and data mining tools, and conversely, the need to apply such analysis tools is a strong motivation for building a data warehouse.

3 The New Research Challenges

Data warehouses contain consolidated data from many databases and other external data sources, spanning long time periods, and augmented with summary information. There are many new challenges in designing, creating and maintaining large data warehouses. There are several new promising research issues some of which are related to problems that the database community has worked on for several years, but others are only just beginning to be addressed.

In recent years special attention has been paid to relational data warehouse systems consolidating data from relational database systems. Several commercial data warehousing systems have been introduced on the market, e.g. *Oracle Express Server*, *DB2*, *Sybase IQ*, *OnLine Dynamic Server*, *OnLine Extended Parallel Server*, *Red Brick Warehouse*, *Teradata*. These systems offer limited functionality and support solutions only for a narrow set of problems concerning relational data warehousing. However, recently, contemporary database systems are used more and more frequently to store and process large amounts of data of complex structure and behavior. Moreover, important business data are stored also on Web pages, in on-line data stores, etc. To process such data efficiently, object–oriented and object–relational database systems are required. The next step in information processing is the integration and analysis of complex data coming from different sources, both relational, object-oriented as well as semi-structured. For this purpose, object–oriented or object–relational data warehousing systems are most promising. This is a new and challenging field of research that is only beginning to be worked on.

Since data warehouses contain huge volumes of data and complex analytical queries are addressed to them, the system's efficiency is very important. In order to speed up the evaluation of OLAP queries different techniques are being developed, i.e. new data structures, new techniques of query optimization and data mining, new tertiary storage organizations.

For modern (i.e. object–relational or object–oriented) data warehousing systems the mechanisms of increasing their efficiency have to be more sophisticated, as data to be processed are of arbitrary complex structure and behavior. Firstly, the optimization of object–oriented queries may be performed by means of inverse methods [6], materialized methods [1, 9, 11]. Secondly, different kinds of index structures, defined on complex attributes and methods have to be applied [2], and are still under the development. Next, data are processed in parallel.

Special attention is being paid to object–oriented views that are important mechanisms assuring logical data independence, providing mechanism for data hiding

and security, simplification of a database schema, and shorthand for queries. Moreover, views are applied to the integration of data coming from different distributed sources as well as for the materialization of data. However, in this field we still face open research issues concerning: (1) the design of efficient and scalable object–oriented views, (2) techniques for materialization of object–oriented views, and (3) the maintenance algorithms of such materialized views.

With recent developments of data mining technology, it is expected that data mining tools will be introduced for sophisticated data analysis in data warehouses. The primary goal of data mining is to discover frequently occurring, previously unknown, and interesting patterns from very large databases. The discovered patterns are usually represented in the form of association rules or sequential patterns. The results of data mining are mostly used to support decisions, observe trends, and plan marketing strategies. Data warehousing imposes new requirements for data mining, e.g. data mining query optimization, materializing data mining results, multidimensional data mining algorithms.

Other just emerging research areas within the context of data warehousing concern: (1) the maintenance of summary data for periodically changing dimension data as well as (2) the maintenance of a data warehouse schema and data in the presence of changes made to the schemas of source databases. For both research areas it seems to be reasonable to use methods of object-oriented or object-relational data warehouses to describe dependencies between different schema versions and to efficiently migrate data between different versions of schemas and dimension-date to allow multi period comparisons and computations of trends.

Furthermore, other challenging research issues, still remaining open problems for this new type of data warehouses, concern efficient data storage techniques and tertiary storage systems.

4 The Goals of Project ORDAWA

The aim of this project is to develop techniques and algorithms for object-oriented and object-relational data warehousing. To check the correctness and efficiency of the developed techniques and algorithms an experimental prototype of an object-relational data warehousing system will be developed.

The following research issues will be pursued while developing the system:

- integration and consolidation of different external data sources within a data warehouse,
- index structures for data warehouses,
- relational and object-oriented views for data warehousing [3, 4, 5, 7, 14, 16, 17],
- algorithms for incremental relational and object-oriented view maintenance [7, 18],
- query transformation and optimization techniques in data warehousing environment [6],
- parallel processing in data warehouse environment,

- techniques and algorithms for data mining in data warehousing environment [10, 12, 13].

References

1. Bertino E.: Method precomputation in object–oriented databases. *Proceedings of ACM–SIGOIS and IEEE–TC–OA International Conference on Organizational Computing Systems*, 1991
2. Bertino E., Catania B., Chiesa L.: Definition and Analysis of Index Organizations for Object-Oriented Database Systems. Information Systems, Vol.23, No.2, pp. 65-108, 1998
3 Dobrovnik M., Eder J.: Adding view support to ODMG-93. Proceedings of the International Workshop on Advances in Databases and Information Systems, 1994
4. Dobrovnik M., Eder J.: Logical data independence and modularity through views in OODBMS. Proceedings of the Engineering Systems Design and Analysis Conference, Vol. 2, 1996, pp. 13-20
5. Dobrovnik M., Eder J.: Partial Replication of Object–Oriented Databases. Proceedings of the Second East-European Conference on Advances in Databases and Information Systems – ADBIS'98. Poland, 1998, LNCS No. 1475, pp. 260-271
6. Eder J., Frank H., Liebhart W.: Optimization of Object–Oriented Queries by Inverse Methods. Proceedings of East/West Database Workshop, Austria, 1994
7. Gupta A., Mumick I.S. (eds.): Materialized Views: Techniques, Implementations, and Applications. The MIT Press, 1999
8. Hammer J., Garcia–Molina H., Widom J., Labio W., Zhuge Y.: The Stanford Data Warehousing Project. Data Engineering Bulletin, Vol. 18, No. 2, June, 1995
9. Kemper A., Kilger C., Moerkotte G.: Function Materialization in Object Bases: Design, Realization, and Evaluation. IEEE Transactions on Knowledge and Data Engineering, Vol. 6, No. 4, 1994
10. Morzy T, Wojciechowski M., Zakrzewicz M.: Pattern-Oriented Hierarchical Clustering. Proceedings of the third East-European Symposium on Advances in Databases and Information Systems – ADBIS'99, Slovenia, 1999, LNCS 1691, pp. 179-190
11. Morzy T., Wrembel R., Koszlajda T.: Hierarchical Materialisation of Method Results in Object–Oriented Views. Proceedings of 2000 ADBIS-DASFAA Symposium on Advances in Databases and Information Systems, Czech Republic, 2000
12. Morzy T., Zakrzewicz M.: Group Bitmap Indexing for Association Rules Retrieval. Proceedings of the Int. Conf. on Knowledge Discovery in Databases – KDD'98, New York, 1998, AAAI/MIT Press, Menlo Park, CA
13. Morzy T., Zakrzewicz M.: SQL-like language for database mining. Proceedings of the first East-European Symposium on Advances in Databases and Information Systems – ADBIS'97, Russia, 1997
14. Roussopoulos N.: Materialized Views and Data Warehouses. SIGMOD Record, Vol. 27, No. 1, 1998, pp. 21-26
15. Widom J.: Research Problems in Data Warehousing. Proceedings of 4th Int. Conference on Information and Knowledge Management (CIKM), 1995
16. Wrembel R.: Deriving consistent view schemas in an object-oriented database. Proceedings of the fourteenth International Symposium on Computer and Information Sciences – ISCIS'99, Turkey, 1999, pp. 803-810

17. Wrembel R.: On a formal model of an object–oriented database with views supporting data materialisation. Proceedings (of short papers) of the third East-European Conference on Advances in Databases and Information Systems - ADBIS'99, Slovenia, 1999, pp. 109-116
18. Wrembel R.: On Materialising Object–Oriented Views. Proceedings of the fourth IEEE Baltic Workshop on DB&IS, Lithuania, 2000