Report on the ACM Fourth International Workshop on Data Warehousing and OLAP (DOLAP 2001)

Joachim Hammer
Department of Computer & Information Science & Engineering
University of Florida
Gainesville, FL 32611-6120, U.S.A.
jhammer@cise.ufl.edu

Introduction

The Fourth Annual ACM International Workshop on Data Warehousing and Online Analytical Processing (DOLAP 2001) was held in Atlanta, GA, USA, in November 2001, in conjunction with the Tenth International Conference on Information and Knowledge Management (CIKM 2001). Although this was only the fourth annual meeting, DOLAP has already become an important and broadly accepted forum for researchers and practitioners to share their findings in theoretical foundations, current methodologies, practical experiences, and new research directions in the areas of data warehousing and online analytical processing (OLAP). Despite the fact that conference attendance has been down since the horrific events of September 11, DOLAP 2001 attracted researchers from Europe, Asia and the Americas.

The DOLAP 2001 program, which occupied a full day immediately following CIKM 2001, included a keynote talk, technical presentations, and a final discussion involving all workshop participants. As in previous years, the quality of the submitted papers was high and the program committee had a difficult time deciding which of the 31 submissions from 16 different countries should be accepted for presentation. The submitted research papers covered the state-of-the-art in data warehousing and related fields including data warehouse architecture, design and evolution, multi-dimensional modeling, query optimization, indexing, view materialization and maintenance, data warehouse quality, XML- and object-based warehouses, and data warehousing and the Web. In addition, the committee received a number of industrial submissions describing ongoing data warehousing projects and novel applications for warehouses. After careful review, 12 research and industrial papers were selected for presentation at DOLAP 2001. The proceedings are published by ACM Press, and are also available online at www.informatik.uni-trier.de/~ley/db/conf/dolap/dolap2001.html.

In addition, the authors of the three highest-ranked papers have been invited to submit an updated, more detailed version to the Special Issue on Advances in Data Warehousing of the Data and Knowledge Engineering Journal to be published in early 2003.

Keynote Address

An important goal of DOLAP is to bring together researchers and practitioners of data warehouse technology from academia as well as industry. This is particularly important since the work in industrial laboratories has often gone unnoticed by the academic research community. In order to provide workshop participants with an industrial perspective, the DOLAP program committee invited Rick Cole, Senior Technical Staff Member in the Business Intelligence Development group at IBM's Silicon Valley Laboratory to talk about his R&D experience and how Red Brick's products have impacted the field of data warehousing.

Rick Cole is a 7-year veteran of Red Brick Systems, Informix Software and recently IBM's Silicon Valley Laboratory, where he has held principle engineering, architecture, and senior management positions. Rick Cole has responsibility for DB2 and Red Brick product integration and technology transfer.

During the hour-long presentation titled "The Red Brick Road," Cole took the workshop attendants on a tour of his various career stops, outlining his early experiences in data warehousing applications, star schema data modeling, and star schema query processing. The tour started at the IBM Advanced Business Systems Division in Rochester, Minnesota where Cole played a lead role in the design and development of SQL query processing in DB2 for the IBM AS/400. For example, he described how query-processing problems in the document and control database of IBM's System/38 were due to its use of a conceptual schema reminiscent of the now familiar star schema. Subsequent work on overcoming the query problem lead to the first implementation of the star join algorithm.

At IBM's Santa Teresa lab, Cole gained first hand experience of the benefits of close collaboration among various database product groups and one of the research divisions at IBM. This collaboration resulted in better understanding of the strengths and weaknesses of several products, faster technology transfers from research to production, and eventually in significant product improvements such as faster sorting for SQL/MVS, hash joins for AS/400 and the use of partial indexes for SQL/400.

Next stop on the tour was the Volcano project at the University of Colorado at Boulder, whose breakthroughs in query optimization laid some of the foundations for modern query processing in database systems including dynamic query processing in the Red Brick data warehouse.

Rick Cole became directly involved with data warehousing when he joined Red Brick Systems. Although the company has undergone several changes and transitions over the past years, its data warehousing product consistently remained at the forefront of the technology curve. Among Red Brick's contributions to the field of data warehousing are several star schema join algorithms, analytic extensions to SQL (RISQL), dynamic query plans, parallel loading, various intelligent scanning mechanisms, integrated data mining, and integrated materialized views. The talk concluded with Cole's observation that IBM's acquisition of Red Brick has allowed him to come back (home) to the place where his career started many years ago (hence the title of the talk).

During the subsequent discussion immediately after the talk, Cole also offered his opinion on what he thinks the important and worthwhile research problems in data warehousing are. His list included the need for researchers to consider warehouse schemas containing multiple fact tables having a subset of common dimensions as well as fact tables which are connected via hierarchical relationships; parallel data warehousing (cube) algorithms; efficient handling of changing dimensions, which includes (1) deleting old data from dimension tables while ensuring referential integrity between the fact and dimensions, and (2) updating of dimensional attributes while maintaining an accurate history of changes; the need for technologies to help ease system configuration and tuning; and more generally, data warehouse support for new applications including customer relationship management, e-commerce, etc.

Paper Presentations

Originally, data warehousing and online analytical processing have emerged as key technologies for enterprises wishing to streamline the management of their operational and business data and to improve decision support. However, judging by the number and diversity of this year's submissions, it is apparent that data warehousing continues to take on broader roles, for example, in the context of integrating semistructured data sources, metadata management, and internet-based e-commerce. This broadening was also reflected in the DOLAP 2001 program.

In order to provide a framework for the different presentation topics, the accepted papers were divided into three topic areas: Multi-dimensional Modeling, Query Processing and Optimization, and Data Warehouse Operation.

Multi-dimensional Modeling

The usability of a decision support system depends on the implementation of the multi-

dimensional data cube and the ease with which the cube can be designed to fit the analysts' needs. Important inputs to the design process are the expected decision-support queries and a description of contents and schema of the sources or the integrated data warehouse if one exists. The four papers in this first session describe various approaches to simplifying the design of OLAP systems, using different assumptions about the inputs and desired output.

The paper by Niemi, Nummenmaa and Thanisch (University of Tampere, University of Edinburgh) presented a new technique to automate cube design given the data warehouse, functional dependency information, and sample OLAP queries. Their method constructs complete but minimal cubes with low risks related to sparsity and incorrect aggregations. The design process is iterative and improves over time, as more information about the expected queries is known.

Abelló, Samos, and Saltor (U. Politècnica de Catalunya, U. de Granada) provided a theoretical foundation for understanding multi-dimensional data and how it should be modeled. By providing a sound and complete algebra for manipulating cubes (using an object-oriented approach), the authors argue that designing and hence querying cubes can become more intuitive, and storing cubes can be made more efficient.

The papers by Pokorný (Charles University) and by Golfarelli, Rizzi and Vrdoljak (University of Bologna, University of Zagreb) recognized the fact that increasing amounts of interesting data is stored in the XML data format; both presentations described tools to help integrate XML data in data warehousing environments. Pokorný showed how the existing star schema approach to multi-dimensional modeling could be adapted to an XML-based data warehouse, where the underlying sources are also XML-based. Golfarelli, Rizzi and Vrdoljak described how the design of a data mart could be carried out starting directly from an XML source (rather than converting it into an equivalent relational representation first). To do so, the authors proposed a semi-automatic approach for building the conceptual schema, leaving the choice of how to implement the data mart to the user.

Query Processing and Optimization

The four papers in this topic area considered new approaches to improving the performance of OLAP queries against a multi-dimensional data warehouse. Although ultimately concerned with improving the execution speed, two of the papers were equally concerned with improving the users' ability to formulate meaningful and useful queries.

The paper by Theodoratos and Tsois (New Jersey Institute of Technology, National Technical University of Athens) presented an architecture called CBS star that uses one-dimensional hierarchical clustering and encoding techniques to organize the dimension tables and multi-dimensional access methods to organize the fact table. User queries against a traditional star schema are rewritten by the query processor to run over the corresponding CBS star schema instead. The authors showed that this re-writing in conjunction with the optimized CBS star improves the performance of a large class of OLAP queries containing expensive star-join operations.

Like Thedoratos and Tsois, Choong, Laurent and Marcel (Université F. Rabelais), have developed a new design for multi-dimensional data sets; however, their main objective is to improve navigation rather than processing performance. To this end, the authors have defined a measurement of the quality of a representation of multi-dimensional data and presented a framework for computing the best possible representation for a given usage scenario. An important contribution of the work was their analysis of the difficulties of computing such representations.

Deschler and Rundensteiner (Worcester Polytechnic Institute) examined the problem of how to balance query performance against warehouse update performance to meet the maintenance window in large data warehouses. Specifically, the authors have introduced the RB+ (Red-Black+) tree as a practical replacement for the popular B+ tree. The RB+

tree uses persistent red-black binary trees instead of sorted records for leaf pages. The paper shows that this organization improves memory performance up to 3,000% for updates and provides query performance comparable to a B+ tree, making it practical for large, frequently updated warehouses.

The paper by Espil and Vaisman (Pontificia Universidad Católica Argentina, Universidad de Buenos Aires) presented a language called IRAH (Intensional Redefinition for Aggregation Hierarchies) for specifying the maintenance of dimension hierarchies in a multi-dimensional database. The authors showed how IRAH supports the definitions of exceptions (e.g., caused by imprecise knowledge or unpredictability of the underlying data) that override the extensions of rollup functions implied in the hierarchies. Moreover, IRAH supports materialization of exception paths, which extends the usefulness of the multi-dimensional model and enhances query processing.

Data Warehouse Operation

The final session included three papers, which were concerned with issues relating to data warehouse operations including data warehouse replication, metadata management and OLAP query support in commercial database servers. Unfortunately, due to the events of September 11, 2001, the authors of the fourth paper, Drs. Bębel and Wrembel (Poznań University of Technology), were unable to attend the workshop.

The paper by Schlesinger, Bauer, Lehner, Ediberidze and Gutzmann (University of Erlangen-Nürnberg) focused on efficient synchronization of multi-dimensional data in a client-server environment, which allows users to store and maintain parts of the warehouse on personal computers closer to where the data is needed. The main contribution of the paper is an algorithm to detect changes to the schema data, and to efficiently synchronize between client and server while exploiting the special needs and requirements of data warehousing.

Vaduva, Kietz and Zücker (University of Zürich, Swiss Life) reported on a new metamodel called M4 (MetaModel of Mining Mart), to support metadata management in a variety of environments including data preprocessing for data mining and data warehousing. To illustrate the merits of M4, the authors discussed the benefits of metedata management in integration systems such as increased support for consistency, uniform and easy access to all information, and data lineage tracing.

The final paper of the workshop by Andreas Weininger (IBM) outlined specific features that a database server should have to efficiently process queries on a database with a star schema model. The basis for his investigation was an analysis of the features provided by the IBM Extended Parallel Server (XPS). His results concluded that special star join methods like the Push-Down Hash Semi Join, new access methods like Generalized Key (GK) indices, and specific index usages like multi-index scans are essential for the efficient processing of queries that arise frequently in data warehousing contexts.

Discussion and Closing Remarks

The workshop ended with a lively discussion of the next-generation data warehouse problems offered by the keynote speaker earlier in the day. The workshop participants concluded that data warehousing and OLAP remain interesting and fruitful research areas with many unsolved challenges.

Acknowledgements

DOLAP 2001 was sponsored by ACM's Special Interest Groups on Information Retrieval (SIGIR) and Management Information Systems (SIGMIS). We also express our thanks to Professor Il-Yeol Song, the CIKM 2001 Workshop Chair, to Professor E.K. Park, treasurer and registration chair, and to the extremely competent members of the DOLAP Program Committee for helping us put together a strong workshop program.