Report on the 5th International Workshop on the Design and Management of Data Warehouses (DMDW'03)

Hans-J. Lenz[1], Panos Vassiliadis[2], Manfred Jeusfeld[3], Martin Staudt[4]

[1] Freie Univ. Berlin, Germany, hjlenz@wiwiss.fu-berlin.de

[2] Univ. of Ioannina, Greece, pvassil@cs.uoi.gr

[3] Tilburg Univ., The Netherlands, Manfred.Jeusfeld@kub.nl

[4] Univ. of Applied Sciences Solothurn, Switzerland, martin.staudt@fhso.ch

1 Introduction

This article reports on the 5th International Workshop on Design and Management of Data Warehouses (DMDW) held in conjunction with VLDB'03 in Berlin, Germany, on September 8th, 2003. A short summary on the papers and the discussions held during the workshop is given.

Data Warehousing embraces technology and industrial practice to systematically integrate data from multiple distributed data sources and to use that data in annotated and aggregated form to support business decision-making and enterprise management. Although many database techniques have been revisited or newly developed in the context of data warehouses, such as view maintenance and OLAP, little attention has been paid to the design, management and high quality service of the management of a given enterprise. Little attention is also paid to aspects that are intrinsic to the functionality and usage of data warehouses, either in the back-stage (like data cleansing or data extraction and loading) or in the front-end (like similar and uncertain queries, or what-if analysis). The DMDW workshop is intended as a forum to fill this gap.

This year's main theme for DMDW was "Business Intelligence for Data Warehouses: Functionality, Usability and Quality of Service". In response to the call for papers, 21 papers were submitted. The Program Committee selected 9 papers based on their scientific contribution, novelty and relevance to the workshop. In a popular research topic, DMDW included three papers on the problem of data warehouse design. A second area of interest for this year's workshop included novel technologies and best practices, possibly showing a movement of data warehousing research towards novel technological challenges, like for example, XML, grid computing and spatial data warehouses. Finally, following a tradition of keeping DMDW close to practitioner problems, three experience reports were also accepted for this year's workshop.

Apart from the DBLP server, the electronic edition of these proceedings is also available from the CEUR series:

http://CEUR-WS.org/Vol-77

2 Data Warehouse Design

Data warehouse design is a reoccurring theme to DMDW and appears to be one of the most popular fields for data warehouse research so far. Having matured from the era where data warehouse design was considered simply a selection of materialized views, the topics of this year's DMDW on data warehouse design involve methodologies, correctness criteria and semi-automated support to the DW designer.

The paper by S. Luján-Mora and J. Trujillo, entitled "*A Comprehensive Method for Data Warehouse Design*" presents a UML-based framework for the full span of DW design. The different phases and artifacts of a DW design process are identified and supported by the usage of UML. To this end, the authors start by identifying different kinds of important schemata for data warehouse design. Then, for each phase of the DW design the authors present different stereotypes for different uses. Packages are used to abstract schemata and their details are revealed through zoom out operations.

In his paper entitled "*Well-formed data warehouse structures*", M. Schneider models a DW as a graph with nodes being fact and dimension tables and edges representing the dependency relationships among the different nodes of the graph. Once the architecture of the graph has been established, constraints for the well-formedness of the graph are given. Furthermore, a mapping to relational structures is also given, both for star and snowflake schemata.

Finally, the paper "*Using Design Guidelines to*

Improve Data Warehouse Logical Design" by V. Peralta and R. Ruggia discusses how to automate the design process, taking into account performance requirements, in the context of mapping from the conceptual to logical structures. The goal of the paper is to formalize implementation-related information and then use this information in the form of guidelines for the production of the logical DW schema. Three types of guidelines are discussed in this paper: vertical fragmentation of dimensions, horizontal fragmentation of facts and aggregate materialization.

3 Best Practices and Novel Techniques

The two papers on novel approaches in data warehousing discuss the usage of XML in the integration of data to support OLAP queries and the issues around the architecture and usage of a spatial DW system. A discussion of best practices around the design, usage and deployment of a DW was also included in this session.

The paper "*Applying Grid Technologies to XML Based OLAP Cube Construction*" by T. Niemi, M. Niinimäki, J. Nummenmaa and P. Thanisch discusses handling problems due to the huge volume of OLAP databases and the updates of data. The authors have developed a system for OLAP data collection from distributed XML sources. Whereas the user defines the information that he/she needs over a global schema, a "collection server" acts as a mediator and assigns local servers to do the job. XML is employed for describing the global schema fragmentation and representing the OLAP data, practically, as a selection – group by query. Grid techniques are applied for the security layer and the distribution of Java code.

In another line of research, H. Bâazaoui Zghal, S. Faïz and H. Ben Ghézala authored the paper entitled "*CASME: A CASE Tool for Spatial Data Marts Design and Generation*". The paper is focused on spatial decision support. A metamodel is presented, classifying data in spatial objects (location of a road) and non-spatial objects (name of a road). Moreover, the paper describes user applications, the architecture of a tool that has been implemented and provides a set of steps for the design of a spatial DW. During the presentation of the paper, the validation of the tool through road accident databases was discussed. Most important, the usage of spatial DW's was also highlighted: spatial data mining and spatial OLAP are employed to support queries which deviate from the traditional OLAP queries in the context of alphanumeric information (e.g., instead of a count(*) query the user can start with a query of the form "show me the accidents on the map").

Finally, in a paper entitled "*Best Practice for Implementing a Data Warehouse: A Review for Strategic Alignment*", R. Weir, T. Peng and J. Kerridge make a three-way effort to shed light on elements of best practice in implementing a data warehouse. To this end they (a) provide a review of existing literature, (b) discuss a successful case study and (c) provide a framework to rationalize on the success of this particular case study. In terms of the first part, elements of best practice are discussed and compared as pre-2000 vs. post-2000. It appears that management buy-in, training and politics are more significant in the literature now than before; on the other hand, end-user buy-in and project goals seem to appear less important. Then, a case study of a company that was in a problematic situation is discussed. The company implemented a data warehouse and from $60M losses turned to $211M profit in 2 years. The paper discusses the management framework under which the data warehouse was originated and exploited for this success. Yet, during the discussion that followed, it appears that the strategy of the organization plays the major role in this kind of success and the IT is still just one of the means to support it.

4 Experience Reports

DMDW nurtures a tradition of bringing academia close to practical problems. This year's DMDW, in a very interesting session, hosted a performance tuning paper from IBM, a paper from Oracle discussing the incorporation of MOLAP functionality in the ROLAP server of Oracle and an experience report from Telecom Italia discussing their back-end ETL system.

In a paper by M. Nicola and H. Rizvi, entitled "*Storage Layout and I/O Performance in Data Warehouses*" the problem of data placement on disks in DW's is discussed and a comparison of storage layouts for data warehouses is performed. The authors, both coming from IBM, try to confront the DBA's problem of optimizing I/O performance. The problem is typically translated to decisions such as how to spread data on disks or how to perform striping. Current solutions involve tricks like placing tables and indexes/logs/temporary space on different disks. Still, no practical systematical solution is offered, mainly due to the fact that research starts with the unrealistic assumption that a workload on which to place the design can be identified. Unfortunately, practical experience suggests that workloads are not always available, and if they are, they change too often. The paper reports two experiments: (a) on 1 TB TPC-H data over a cluster of 32 machines, where all data were uniformly distributed across the nodes and (b) experiments on a 50GB star Schema, over an SMP system with 22 disks. The findings of the paper suggest that (a) high I/O throughput is more important than seek times or I/O contention; (b) striping everything across all disks with a large to medium stripe size and a significant number of disks dedicated to fact tables is a performance winner (better than separating indexes from tables) as well as the easiest way to administer the DW.

While the previous paper tackled the DW physical design and administration process, the support of the querying process from another vendor is discussed in the paper by M. Bastien, entitled "*Accessing multidimensional Data Types in Oracle 9i Release 2*". Oracle has incorporated the MOLAP Express Server in its relational engine and the paper discusses how an object-relational interface can be created, so that SQL queries can be dispatched to the MOLAP engine for processing. Two types of ADT's are required to handle queries: one for cube rows and another to form tables out of these rows. The multidimensional structures are wrapped through a 'table' function and potentially a view (to make the interface pure relational) before they are accessed through SQL queries. An extra feature of the discussed architecture is that the pre-aggregation of all possible cubes is not required in this environment.

Finally, J. Adzic and V. Fiore in their paper "*Data Warehouse Population Platform*" discuss an ETL platform and a case study for Telecom Italia. The DW Population Platform (DWPP) is a general platform, comprising a set of libraries for common ETL tasks. The architecture of the system is discussed in the paper, with particular interest in (a) a "shared data area", which is an in-memory area for data transformations, with a specialized area for rapid access to lookup tables and (b) the pipelining of the ETL processes in DWPP. A case study for mobile network traffic data is also discussed, involving around 30 data flows, 10 sources, and around 2TB of data, with 3 billion rows. The throughput of the population system is 80M rows/hour, 100M rows/day, exploiting low level OCI calls, in order to avoid storing data to files and then loading them through loading tools. With 4 hours of loading window for all this workload, the main issues identified involve (a) performance, (b) recovery, (c) day by day maintenance of ETL activities and (d) adaptable and flexible activities. The discussion revealed two main problems, specifically (a) the choice of partitions as the main design problem and (b) the need for practically real time DW, with 2 hour data freshness, due to user requirements!

5 Keynote

Data warehousing is mature enough an area to be able to host keynote speeches in its dedicated workshops and conferences. DMDW could not be an exception; this year we had the opportunity to enjoy a keynote speech by Prof. Stefano Rizzi from the Univ. of Bologna, with title "*Open problems in data warehousing: eight years later*".

Stefano began by surveying landmarks in the field of data warehousing both as a practical but mostly as a research area (actually, the "eight years later" part of the title refers to the seminal paper of J. Widom in CIKM'95).

In terms of the relationship between research and practice, the following table summarizes Stefano's assessment of the situation as of today.

	Tool implementation	User satisfaction
architectures	😐	☺
conceptual modeling	☹	☹
OLAP	☺	☺
query lang. and processing	😐	😐
optimization and tuning	☹	😐
physical aspects, indexing	☺	☺

Stefano gave a list of open issues for different stakeholders in the data warehouse environment (which are not necessarily research topics, though). In a nutshell, these issues involve:

- Methodologies for DW design and data integration
- Project documentation
- Metadata standards
- Data quality assessment
- DW evolution

Stefano proceeded in presenting some results in the fields of DW evolution and project documentation and concluded with three promising research problems:

- Federations of XML warehouses under P2P environments.
- Optimization and processing for complex nested aggregation queries.
- Maintenance in highly dynamical environments.

Following the talk, a discussion took place, mainly on the future of data warehousing research. Apart from the topics raised by Stefano, other areas of research were also suggested, like data warehousing for multimedia content and real-time data warehousing (where the back-stage cannot afford the luxury of daily refreshment and has to sustain refreshment rates in terms of minutes or hours). The agonizing question on whether data warehousing as a research area has any future at all was also raised and there was a rather wide agreement that without any drastic breakthrough the area will eventually freeze, even if the practical problems are not yet solved.

6 Summary

It would not be unrealistic to say that this year's DMDW closes a cycle of five years of interesting data warehousing research. After five years, DW design is the only topic that consistently remains in the issues of the workshop; on the other hand, traditional topics like materialized view selection, indexing, aggregate query processing, metadata management and conceptual modeling seem to disappear, not only from the final program but also from the submissions too. However, this does not mean that these problems have found satisfactory solutions in practice. The delay between taking over proposed (in part even standardized) solutions by the DW technology vendors seems to be rather increasing.

While it is clear that the field of data warehousing as a research area is going through a crisis, seeking for new topics, we cannot summarize without a tone of optimism: indeed, novel research areas can be identified, including:

- the handling of non-traditional content, like spatial or multimedia information;
- the need of addressing real-time requirements, i.e., moving the focus from pure DW's for decision support to general purpose (integrated) Operational Data Stores, where many aspects are similar but additional technology must be provided;
- advanced data warehouse architectures like P2P data warehousing, and
- the management of the dynamically evolving nature of DW's.

Of course, traditional areas like data integration and query processing and optimization will possibly continue to offer research topics from the viewpoint of data warehousing. Naturally, it is up to the research community, whether to embark on these challenges or follow different directions.

7 Acknowledgments

On behalf of the Program Committee we would like to thank all the authors of submitted papers and all the referees for their careful work. The support of Freie Universität Berlin, Oracle Deutschland GmbH and Sunsite Central Europe (CEUR) is also gratefully acknowledged. We would like to express our gratitude to the members of the Organizing Committee of VLDB'03 for their support in organizing this workshop. Finally, we would also like to thank Avigdor Gal for detailed comments over a previous version of this report.