An Analysis of the Publication: "An Overview of Data Warehousing and OLAP Technology"

CSci 8701: Overview of Database Research
University of Minnesota

Michael Goshey

Tuesday, September 19, 2006
# Contents

INTRODUCTION....................................................................................................................................................... 3  
Assignment summary........................................................................................................................................... 3  

PROBLEM ADDRESSED ......................................................................................................................................... 3  
Problem statement............................................................................................................................................... 3  
Significance......................................................................................................................................................... 3  

MAJOR CONTRIBUTIONS..................................................................................................................................... 3  
Bridging the gap between industry and academia............................................................................................... 3  
OLTP vs. OLAP: clarifying the differences......................................................................................................... 3  
Concise survey of relevant issues, architectures and tools.................................................................................. 4  
Concrete list of data warehouse design and build steps...................................................................................... 4  

KEY CONCEPTS....................................................................................................................................................... 4  
OLTP and OLAP ................................................................................................................................................. 4  
Relational and dimensional data models............................................................................................................. 4  
Data warehouses and data marts........................................................................................................................ 4  
ETL ...................................................................................................................................................................... 5  
Metadata.............................................................................................................................................................. 5  
Managed query vs. ad hoc query and analysis environments ............................................................................. 5  
ROLAP and MOLAP ........................................................................................................................................... 5  
Materialized views............................................................................................................................................... 5  
Bitmap indices ..................................................................................................................................................... 5  
SQL extensions .................................................................................................................................................... 6  

VALIDATION METHODOLOGY........................................................................................................................... 6  
ASSUMPTIONS.......................................................................................................................................................... 6  

2006 REWRITE .......................................................................................................................................................... 7
Introduction

Assignment summary

The assignment required each team to select a database research paper scheduled to be covered during the course and to provide a written analysis of the chosen paper. I selected the following publication:


Problem Addressed

Problem statement

The purpose of this survey paper was to provide an overview and brief assessment of the burgeoning database field of data warehousing and on-line analytical processing (OLAP) systems which had rapidly grown in prominence and importance during the years immediately preceding the paper's publication. The authors indicate particular interest in identifying how data warehouse and OLAP requirements differ from those of the more traditional transactionally-based (or OLTP for on-line transaction processing) systems.

Significance

The significance of the paper lies in the fact that data warehousing has become a major avenue of growth and key area of concentration in applied database technology. At the time the paper was published many of its key concepts were new enough to benefit greatly from an academic work that would serve as a focal point to marshal consensus on an approximate taxonomy of terms, architectures, issues and technologies.

Major Contributions

The paper provides a number of key contributions summarized below.

Bridging the gap between industry and academia

As stated in the introduction to the chapter from the course text¹, data warehousing (and by extension OLAP technologies) "largely originated in industrial practice." The authors assert that because the academic research came later the field was "...largely misunderstood." This paper is an academic work by researchers affiliated with industry (Microsoft Research and the HP Palo Alto Lab) and draws successfully from both worlds, establishing a common interest in the field.

OLTP vs. OLAP: clarifying the differences

A major technical contribution of the paper is that the authors articulate differences between systems designed for OLTP and those built to emphasize OLAP (summarized in the "Key Concepts" section). The differences are straightforward but the field benefits from this clear articulation in a reference-able publication.

Concise survey of relevant issues, architectures and tools

A good survey article provides an overview of the technical landscape and its relevant features in a particular area of concentration. The authors succeed in this task by providing a meaningful, concise treatment of the issues, architectures and tools most germane to the field of data warehousing at the time the paper was published.

Concrete list of data warehouse design and build steps

Another noteworthy contribution was the author’s inclusion of a detailed list of activities involved in designing, building deploying and maintaining a typical data warehouse environment. While the list doesn’t break new technical ground per se its inclusion in the paper helps ensure that academic researchers and industry practitioners alike are speaking from the same conceptual understanding of what comprises a data warehouse environment.

Key Concepts

A number of technical concepts are addressed in the paper; those I have deemed to be of particular importance are listed and defined below.

OLTP and OLAP

On-line transaction processing (OLTP) systems feature databases specifically modeled to support speedy and efficient handling of insert, update and delete transactions such as those found in airline reservation systems, insurance claims processing and e-commerce sites to name a few examples. This is in contrast to on-line analytical processing (OLAP) systems which are optimized for data retrieval, reporting, aggregation and ad hoc query and analysis.

Relational and dimensional data models

Relational data models typically feature highly normalized data that minimizes redundant storage and supports the goals of OLTP systems. Conversely dimensional models accept the trade-off of redundant data storage in favor of optimizing response time of anticipated queries. A dimensional model consists of de-normalized dimensions (tables that model entities of interest in the data, for example, auto manufacturer, customer and model) surrounding a fact table that consists of numeric measures of importance for the given set of dimensions (such as unit sales, cost of goods sold, profit margin, etc.) combined with foreign key fields that directly tie the fact table to each of the dimensions, providing a robust ability for users to analyze the measures by filtering and aggregating based on any of the attribute columns in any of the dimensions. This multidimensional construct of dimensions in a logical constellation around a fact table is referred to as a star schema and is the predominant model in data warehousing and OLAP.

A lesser-used variation on this theme is to re-normalize certain dimensions to more easily exploit their natural hierarchies. This technique is referred to as snowflaking a dimension.

Data warehouses and data marts

While data warehousing usually refers to the set of related decision support technologies, a data warehouse more narrowly means the data store itself which emphasizes aggregated and historical data over the detailed, individual records usually found in transactional systems. The inclusion of history combined with operational data from across the enterprise often results in data warehouses that exceed the size of operational databases by orders of magnitude. Given the complexity and scope data warehouses can take years to build. The authors’ view of a data mart is as a departmental or subject-
specific mini-warehouse that is easier to implement and requires less political capital to promote consensus across an entire enterprise.

**ETL**

A critical component of data warehousing systems are the processes that 1) extract data from multiple source OLTP systems, 2) transform the data to ensure consistent naming and typing conventions, impose conformance to business rules and provide aggregations to optimize query response times, and 3) load the data into de-normalized structures that support the goals of OLAP. This collection of ETL processes provides a data processing bus to support the warehouse environment.

**Metadata**

Data warehouses thrive on “data about data” (metadata) in a couple of key ways. Since data warehouses support fast, efficient reporting and query and analysis, much importance is placed on making the data easier for users to understand. The cryptic field names that exist in operational systems are translated into more verbose representations and are supplemented by detailed information about the source systems from which they were extracted or business rules and calculations that have been applied. A second type of metadata monitors the state of the system itself, allowing the warehouse administrator(s) to answer questions such as “what are the most/least frequently used queries?” or “who are the most/least active users?”

**Managed query vs. ad hoc query and analysis environments**

The authors point out that data warehouses are often designed to support multiple user communities with distinct needs. A managed query environment provides an extensive set of pre-written queries or parameterized stored procedures that feed a specific set of canned reports consumed by a (typically) less-sophisticated audience of data consumers. Technology savvy users may prefer to use ad hoc query and analysis tools that allow them the freedom to construct their own queries, write their own reports and perform “what-if” analysis on the data.

**ROLAP and MOLAP**

The authors identify a couple of architectural alternatives used in OLAP systems. While both approaches are based on a multi-dimensional view of data the implementations are distinct. Multi-dimensional OLAP (MOLAP) systems utilize specialized cube-like data structures that are built and maintained separately from traditional relational tables. Relational OLAP (ROLAP) systems rather use standard relational tables in multi-dimensional ways such as building star-schemas and creating fact tables aggregated to various levels of anticipated query usage.

**Materialized views**

Materialized views are typically used to maintain local copies of the results of queries as well as to provide pre-stored results from aggregations. Since extracting and combining data from remote operational systems as well as providing robust aggregations are common features of the data warehouse, materialized views are a well-used tool in the warehouse developer’s toolkit.

**Bitmap indices**

In contrast to B-Tree indices commonly used in OLTP systems the static nature of data warehouses (absence of direct inserts by users, limited updates occurring at less-frequent intervals) allows them to utilize highly compact bitmap indices (storing zeros or ones as bit flags) which can be very fast but require a lot of overhead to maintain. Such indices are especially useful in low cardinality columns (i.e., those with a low number of distinct possible values such as gender, continent, etc.).
**SQL extensions**

The authors briefly describe specialized extensions to the SQL language. These **SQL extensions** have been added to support specific goals of OLAP and data warehousing, such as more robust aggregation functions, reporting-specific trending functions like moving average, column comparisons and others.

**Validation Methodology**

A survey paper is by definition less concerned with making novel claims than with bringing order to a topical area of interest that perhaps lacks clarifying foundational reference works. Therefore the authors do not provide long-winded justifications of their statements but rather cite the work of many others in the field who’s findings they are summarizing.

However, one sense of order that was required in the data warehousing field was to bridge the gulf between an industry-dominated field and academia and toward this aim the authors attempt to validate the legitimacy of their paper by leveraging many sources from both academia and industry. Referencing the work of Inmon and Kimball along side of numerous academic citations is a tactically sound way of marshalling respect from both worlds.

Along the same lines the authors also mention a number of DBMS, ETL, OLAP and data warehousing tools and vendors to demonstrate knowledge in the field and an awareness of the state of the art and the industry at that time.

**Assumptions**

A key assumption of the authors is their firm positioning of data warehouse environments as read-only, static repositories of data. While this claim is usually true in theory it does not always hold up in real-world enterprise data warehouses.

For example, some users may require the ability to write their own materialized views that become part of the warehouse environment, not only for themselves but perhaps to publish and expose to other users. Many of the Business Intelligence (present-day term for Decision Support) vendors now provide such capability as standard.

A data warehouse environment can even become tethered to (limited) transactional requirements. A recent case involved a hotel quality reporting system primarily used to aggregate key numeric performance metrics and provide scorecard-like reporting. The system had the added requirement that the actual report consumers also needed the ability to insert data (the results of their on-site inspections) and have the inserted data be reported back to them as part of the overall scorecard.

Likewise some systems are subject to unanticipated data revisions. As a real-world example customer satisfaction survey data from a surveying system may be one of the operational sources feeding the warehouse. As this data is aggregated and reported back to stakeholders it can be disputed by those involved (“Hey! John Smith is commenting negatively on the behavior of my staff but there are no POS records indicating Mr. Smith visited our location and we don’t have any staff named Bill- he must be referring to another location”, etc.) and occasionally determined to be invalid. Changing history in unplanned ways can result in dramatic consequences such as the need to re-aggregate data.

These are a few examples of how loosening or removing this rigid assumption would likely lead to an interesting exploration of tactics and techniques employed by data warehouse architects to overcome the imperfect circumstances that every organization faces.
2006 Rewrite

Many of the changes I would propose for a 2006 rewrite of this publication stem from the expected changes in terminology, tools and vendors in the data warehouse industry in the 10 years since the paper was conceived. For example fact constellations are now more commonly referred to as conformed dimensions; decision support has become Business Intelligence (BI); none of today’s most major tools and vendors in the BI and ETL space are identified.

Aside from this I would encourage the authors to address trending, slowly changing dimensions and historical snapshots. One of the biggest issues facing data warehouse projects is how to handle history. If I query sales data from three years ago do I want to see it through the lens of today’s organizational hierarchy or do I want it left as it was then? If Mary was the regional director for the West through June and switched to the East in July should her annual rollup numbers include both? Such questions provide daily challenge to those tasked with building data warehouses and would make great additions to the paper.

Lastly I would add more detail about the notion of supporting multiple user constituencies which was only given cursory treatment in the paper. In many cases the requirements of one set of users of a data warehouse (e.g.: must support simplified and streamlined processing of pre-built queries) directly contradict the requirements of other users (e.g.: must support ultimate ad hoc flexibility) and the tensions that are caused by these competing requirements make for enlightening discussion material.