8-16-1996

Advances in Database Technology

Sasa M. Dekleva
School of Accountancy, DePaul University, sdekleva@condor.depaul.edu

Advances in Database Technology

Sasa M. Dekleva
DePaul University
School of Accountancy
1 E. Jackson Blvd.
Chicago, IL 60604-2287
USA
sdekleva@condor.depaul.edu

It caught me by surprise that the relational data model, not long ago considered the most advanced, was referred to as a "conventional" model in a recently published database book. What is superseding it, if anything? What are the new database technologies that should be integrated in our database curricula? This tutorial will present an overview of a number of newer database technologies, including object data management, extended relational data model, data warehousing, on-line analytical processing, data visualization, etc., with references to important sources for further study. It will also identify concepts and technologies that are particularly relevant for graduates of IS programs and should thus be integrated into the curricula.

Technological change, facilitated with the continuous advancements in information technology, is mainly driven by two forces. Database technology is responding to changes in (1) the amount and (2) the format of data stored and manipulated in the industry. Business databases that used to occupy only so many megabytes are now measured in giga-, tera-, or even peta-bytes. At the same time, the use of IT is spreading into new business and technical areas with different database needs from those of traditional computing. These new applications include computer-aided design, computer-aided manufacturing, computer-aided software engineering, text and hypertext, multimedia, office automation, Web and various other applications with complex data formats including knowledge bases.

Relational data model and database technology succeeded in providing easy database access to even occasional users. In other words, the reason for its popularity has been its simplicity. However, conventional relational database management systems are not at all appropriate for many of the new database applications. Several new data models and database architectures have emerged to remedy the shortcomings of the relational model. For example, data warehousing, on-line analytical processing (OLAP), and data visualization have been used to efficiently manipulate the much increased amount of data and to improve its interpretation. At the same time, several technologies have evolved to accommodate new data formats. For example, the relational data model has been extended to support procedures, objects, versions, and other new capabilities. Object-oriented models provide for the definition of user-defined data types and support different levels of inheritance and encapsulation. Functional data model provides a data access and manipulation language based on mathematical function notation and is an elegant way to describe databases. Semantic data models also contributed a number of useful concepts, many of which have been incorporated into other data models.

Technologies for Data Analysis

A number of related technologies have been developed to facilitate decision making. They include data warehousing, data mining, OLAP, and data visualization. These products are rapidly becoming the mainstream of business information processing and should be integrated into our database courses.

Data Warehouse: Most of the discussion on data warehouse was extracted from (Inmon, 1995). A data warehouse is a business subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision making process. The data warehouse is always a physically separate store loaded with the data from the operational environment. It differs from the operational data because the data

warehouse is designed around the major subjects of the enterprise, is integrated, excludes data that will not be used for DSS processing, and spans its data over a spectrum of time. The focus is on data modeling and database design exclusively, which contrasts with the more classical process/functional orientation. Perhaps the most important aspect of the data warehouse is that data are always integrated. Integration is achieved in several ways - in consistent naming, in consistent measurement of variables, in consistent encoding structures, in consistent physical data attributes, and so forth. While the operational applications and databases have a short time horizon and usually present the current state in a company, the data in the warehouse represents the progression over a long time horizon, normally from five to ten years. Every key structure in the data warehouse thus contains an element of time, such as day, week, month, etc. This requirement is addressed in research on the so-called temporal databases. An interesting consequence of this is that data do not need to be updated (they may only need to be corrected), which allows us to store data redundantly without the danger of data becoming inconsistent.

The data warehouse contains data at different levels of summarization and a data dictionary or meta data. The different components of the data warehouse are:

- meta data
- current detail data
- older detail data
- lightly summarized data
- highly summarized data

The source of nearly all data warehouse data is the operational environment, but there is little redundancy of data between the operational environment and the data warehouse environment. A data warehouse is implemented in an evolutionary process. The tasks of extracting, cleaning, and loading information into a data warehouse take an unexpectedly large amount of time. It has been estimated that, on average, 80% of the efforts in building a data warehouse go into these efforts. Experienced warehouse architects believe that the meta data is the most important component of the warehouse. It is used as:

- a directory to help the DSS analyst locate the data warehouse content
- a guide to the mapping of data as the data is transformed from the operational environment to the data warehouse environment
- a guide to the algorithms used for summarization between the current detailed data and the lightly summarized data and the lightly summarized data and the highly summarized data, etc.

On-line Analytical Processing (OLAP): Most of the discussion on OLAP was extracted from (Codd *et al*., 1993). Increased global competition drives knowledge workers to improve their strategic and tactical decisions based on corporate information. OLAP systems allow knowledge workers to intuitively, quickly, and flexibly manipulate data using familiar business terms in search of an analytical insight. In general, the OLAP systems support the complex analysis requirements of decision-makers, analyze the data from a number of different business dimensions, and support complex analyses against large input data sets.

Data consolidation is the process of synthesizing pieces of information into single blocks of essential knowledge. The highest level in a data consolidation path is referred to as that data's dimension. Until recently, the end-user products that had been developed as front-ends to the relational DBMS provided very simplistic functionality. The query and report writers and spread-sheets have been limited in the ways in which they can be aggregated, summarized, consolidated, summed, viewed, and analyzed. In fact, relational DBMS were never intended to provide the very powerful functions for data synthesis, analysis, and consolidation that are being defined as multi-dimensional data analysis. These types of functions were always intended to be provided by separate, end user tools that were outside and complementary to the relational DBMS products. The functions include:

- calculations and modeling applied across dimensions, through hierarchies and/or across members
- trend analysis over sequential time periods

- slicing subsets for on-screen viewing
- drill-down to deeper levels of consolidation
- reach-through to underlying detail data
- rotation to new dimensional comparisons in the viewing area

There are two principal architectures for OLAP systems: multidimensional and relational. The first utilizes a multidimensional database to provide analyses. The relational architecture, on the other hand, provides data access directly from the relational databases.

Data Mining and Visualization: Data mining is also a decision support process where we search for patterns of information in a database. Data mining tools are specially designed to identify significant relationships among variables in very large databases. Such tools can evaluate many possible relationships. For instance, a consumer goods company may track 200 variables about each consumer. This search may be initiated by the user or by a program that automatically searches the database and finds information by itself. In this case, the program looks for interesting and significant patterns with no hypotheses posed by the user. In a mixed mode, the user and the system interact, with the system providing valuable assistance. The user can look at the data from several viewpoints, getting appropriate prompts from the system. Once the information is found, it is presented in a suitable form, with graphs, reports, text, hypertext, etc. Two other data mining techniques are sometimes supported: predictive modeling, where patterns discovered in the database are used to predict the future, and forensic analysis, where the extracted patterns are used to find anomalous, or unusual data elements.

The aim of data visualization tools is to let the user easily and quickly view graphical displays of information from different perspectives. For example, the tools let the user quickly see graphs of different subsets of data or different summaries of data.

The concepts common to data warehouse, OLAP, data mining, and data visualization are large amounts of data, decision support, different levels of data summarization, different information dimensions, ease of use, and rapid response. Data can be stored in a data warehouse to be analyzed by the tools classified into the overlapping categories called OLAP, data mining, and data visualization.

Technologies for Complex Data Formats

The division in two technologies is not perfect. Some data mining techniques, for example, are used for analysis of graphical information, such as satellite images. The techniques discussed above have to a large extent been used for analysis of data stored in traditional computer formats. Object data management addresses complex data formats such as drawings, text and hypertext, images, multimedia, and compound documents.

Object Data Management Systems (ODMS): A number of different approaches to object data management have been used in the development of prototypes and products. This diversity is in contrast with the consistency across implementation of the relational model. Existing products are being constantly revised and enhanced and new solutions are being introduced. A trend toward compliance with the results of standardization efforts can also be noticed. Until this research area stabilizes, it will be necessary to carefully evaluate available products and their vendors to match them with the user's application and IT strategic architecture.

There is no general agreement on what are the necessary ODMS features. Even the strict definition of encapsulation provided by abstract data types, where procedures are public and data are private, is often too restrictive. Such requirement would make ad hoc querying virtually impossible. Some authors therefore consider database systems to be object-oriented even without the requirement for encapsulation.

Research and development follows two dominant directions. One includes a variety of attempts to extend existing database systems to provide additional functionality through new or extended database query languages that incorporate procedures and other ODMS features. Such systems present two separate environments: the application programming language and the extended database query language. The two have different data types and run-time execution environments.

The second direction is to extend existing object-oriented (OO) programming languages, in most cases C++. Added features provide persistence (permanent data storage), concurrency control, and other capabilities. These systems may also provide a query language which is compatible with the programming language and executes in the application program environment and shares the same data types.

Most implementations of new data models, such as object data model, have been developed for the client/server platform. They will coexist with, rather than substitute, the existing operational databases. With the increasing penetration of object database management systems and other new database technology, our database curriculums need to be enhanced to present the advantages and weaknesses of these technologies and to describe the methodologies for their implementation.

Several other related developments and concepts deserve our attention, such as parallel processing, manipulation of text and multimedia documents, large-scale digital libraries, and integration of all these with the Web technology and resources.

Standardization: Standardization is essential to the acceptance of this new database technology. The SQL standard, regardless of the criticism surrounding it, enabled a high degree of portability and interoperability among systems. Database systems vendors, reacting to market forces favoring open solutions, have contributed to various formal and informal standardization initiatives. National Institute of Standards & Technology (NIST) and ISO are also cooperating on the development of an extended relational database architecture. The most significant and comprehensive effort is being undertaken by the NIST X3H2 Working Group, which is enhancing SQL into a computationally complete language for the definition and management of persistent, complex objects. The draft standard, known as SQL3, is already available. It includes the specification of abstract data types, object identifiers, methods, inheritance, polymorphism, encapsulation, and all of the other facilities normally associated with object data management. The expected year for completion of SQL3 is currently 1998.

The Object Database Management Group (ODMG) is a consortium of object-oriented database management system vendors. Interested parties are working on standards to allow portability of customer software across object-oriented DBMS products. It has published the ODMG-93 specification, which includes an object model, an object definition language, an object query language, bindings to C++ and Smalltalk, and several other components. Continuing work is planned for later releases

that will address added functionality.

References are available on request and on the Web at http://condor.depaul.edu/~sdekleva/dbms/biblio.html