Data Extraction and Transformation for the Data Warehouse

A Presentation by
Cass Squire, Channel Director of Professional Services
Prism Solutions, Inc.

Corporations worldwide are finding that understanding and managing rapidly growing, enterprise-wide data is critical for making timely decisions and responding to changing business conditions. To manage and use business information competitively, many companies are establishing decision support systems built around a data warehouse of subject-oriented, integrated, historical information.

In order to understand why the data warehouse must replace old legacy applications for effective information processing, it is necessary to understand the root causes of the difficulty in getting information in the first place. The first difficulty in getting information from the base of old applications is that those old applications were shaped around business requirements that were relevant as much as twenty-five years ago. These applications that were shaped yesterday do not reflect today's business.

The second reason why older applications are so hard to use as a basis for information is that those applications were shaped around the clerical needs of the corporation. A clerically focused application of necessity does not have the historical foundation required to support a long-term view.

Another reason why the clerical perspective of applications does not support management's need for information is that the clerical community focuses on detailed data. While detailed data is fine for the day-to-day clerical needs of the organization, management needs to see summary data in order to identify trends, challenges and opportunities.

Yet another reason why the clerical perspective of applications does not suffice for management's need for information is that the clerically-oriented applications were built an application at a time, and there was little or no integration from one application to the next. The result is that the old legacy applications cannot easily or reliably be combined to produce a unified perspective of data.

For these basic reasons, the older foundation of applications will not suffice as a basis for the important informational processing that organizations need to do in order to become efficient, competitive corporations. Nothing short of an entire change in architecture and a fundamental restructuring of the applications foundation will suffice.

Fortunately there is an alternative architecture, which consists of a separation of processing into two broad categories—operational processing and decision-support processing. At the heart of decision-support (DSS) processing is the structure known as the **data warehouse**.

The data warehouse contains data which has been gathered and integrated from the legacy systems environment. There are different levels of data within the data warehouse. Some data is very detailed. Other data is summarized. Other older detailed data is placed in secondary storage. In addition there is a component of the data warehouse known as "meta data." Meta data, or information about data, is a directory as to what the contents of the data warehouse are and where the contents came from.

Six major steps are involved in implementing a data warehouse: 1) building the data warehouse data model, 2) defining the system of record, or "best data" for the warehouse, 3) designing the physical data warehouse, 4) creating the transformation programs, 5) loading and maintaining the data warehouse, and 6) building and maintaining directory of meta data.

Building a data warehouse requires extraction of data from legacy systems, operational applications and external sources. As data passes from the system of record into the data warehouse, it passes through a set of programs called integration and transformation programs. These programs take application-oriented data and turn the data into corporate data. The integration and transformation programs perform a wide variety of functions, such as—

- reformatting data,
- recalculating data,
- modifying key structures of data,
- adding an element of time to data warehouse data,
- identifying default values of data,
- supplying logic to choose between multiple sources of data,
- summarizing data,
- tallying data,
- merging data from multiple sources, and so forth.

Among the challenges involved in data extraction and transformation include the fact that source data exists in heterogeneous mainframe and UNIX-based environments. The navigational paths of these legacy systems and operational applications are complex. What's more, inconsistencies between naming conventions, business rules and definitions used must be resolved. In addition, data integrity and quality must be verified and maintained throughout the transformation process.

The integration and transformation programs are very susceptible to maintenance as they need to be modified each time the operational environment changes or each time the data warehouse environment itself changes.

There are several advantages to automating the development and maintenance processes. Automated tools can reduce implementation time and cost substantially by eliminating the need for manual programming. Structured code generation increases productivity, promotes consistency across programs and allows quick response to change. Data integrity is maintained by performing data extraction and transformation automatically rather than manually. Finally, automated tools actively capture and maintain meta data related to source files, output tables, transformations and mappings, providing a record of changes and enhancements to the data warehouse over time.

One reason why meta data management is mandatory for the data warehouse environment is due to the size of the warehouse. Meta data serves as the "card catalog," helping users navigate the data warehouse and find relevant information for analysis. Another reason why meta data is so important is because of the horizon of time that is managed in the data warehouse. It is typical for 5 to 10 years of data to be stored in the data warehouse. One of the implications of managing such a lengthy time period of data is that the structure of data will change over time. Meta data stores the context of this historical information.

Meta data exists at two levels in the data warehouse—at the developer level and at the end user level. Developer meta data is used by the developer to manage and control the development and maintenance process. End user meta data resides on the data warehouse platform itself and is available to the end user as a regular part of the access and analysis of the warehouse.