

An Overview of Data Warehousing Technology

By Jim Luisi

Scope and Purpose

The goal of this paper is to objectively position *DataPulse*, a data warehousing product from DataPulse, as an entirely new DBMS paradigm for providing decision support capabilities for multiple tera-bytes of customer and transaction system data. Although targeted for Fortune 100 companies, it can also accommodate the pocketbook of Fortune 2000 companies. It positions companies to incorporate all categories of production data with external prospect lists and customer demographic overlays.

We will identify the ideal set of requirements for data warehousing, as well as, DataMarts, and then focus on the existing applied technologies to identify their limitations and challenges.

The background for this includes examination of the following elements:

- Development - Building Your Data Warehouse
- Available Technologies – Capabilities and Limitations
- Business Use - Analysis and Decision Making

Data Warehousing Environment

The Data Warehousing industry is attempting to house and manage vast quantities of production data involving customer information and transaction history into large repositories utilizing database software designed for transaction systems. As a result, the largest and most powerful hardware solutions have been required in order to perform the most basic decision support requirements.

Due to the limitations of technology, large data warehouses have difficulty incorporating all customer and transaction history, let alone prospective customer data, and demographic enhancements, as well as, economic and environmental data.

Data warehouse subsets, called DataMarts, and other types of smaller databases, have been constructed to address the unwieldy nature of large data warehouses. With smaller amounts of data, organizations have been able to address some of the more advanced decision support capabilities, such as multi-dimensional analysis and statistical modeling. As a result, however, DataMarts are usually found to be lacking in their breadth and volume of useful data.

Development – Building Your Data Warehouse

The path to building a data warehouse is an iterative one. Due to the redundancies in the potential sources of production data, one must decide which sources provide the best quality of data. The results of this analysis help to identify opportunities where production systems can be fortified to improve future data capture and facilitate data clean up.

Data scrubbing reports can assist technologists and management in identifying the areas of production data that require the most attention. Additionally, name and address standardization software can reduce the replication of customer data by improving the match rate for customers during production and data warehouse load procedures.

Available Technologies – Capabilities and Limitations

Although recent developments in database technology for transaction systems have provided new thresholds in transactions processed per second, these do not compare to the levels of data manipulation required by decision support applications. New hardware technology, including, faster CPUs, larger memory devices and faster parallel I/O substructures do not come close to bridging the gap from 'transactions per second' involving a few rows per table, to 'transactions per minute' involving millions of rows for many tables.

As a result, implementations based upon transaction system database technology cannot begin to support real-time decision support applications, regardless of how much money a corporation is willing to spend.

Business Use – Analysis and Decision Making

The ability to analyze the demographics of existing and prospective customers, business transactions and customer contact data, quickly and easily, is essential to making informed decisions for direct marketing, product development, and customer retention. Only a unified repository of data can support questions that pertain to many combinations of attributes.

The three most useful areas of functionality, which require an advanced real-time decision support capability, involve cascade counts, descriptive statistics, and comparative analysis.

Cascade Counts – Is the ability to choose a business object, such as customers or business transactions, perform an analysis based upon various selection criteria and define the most useful organization to provide cascade counts.

Descriptive Statistics – Is the ability to generate descriptive statistics, such as mean, minimum, maximum, frequency distribution, and standard deviation for various business objects.

Comparative Analysis – More robust than multi-dimensional analysis technologies, it is the ability to correlate the count or statistic of any selection criteria real-time, without generating a predetermined set of counts or statistics.

Marketing requires the identification of the demographic characteristics of the most valuable customers, which can be used to determine the most valuable prospects.

The needs of management also include the ability to identify emerging markets, demographic trends, identification of possible expansions and consolidations, competitive pressures, mergers and acquisitions, and enhancements to internal policies and procedures.

The Challenges

What are the challenges of MIS and EIS reporting?

Management and executive information systems face the following reporting challenges:

- **The inability to evaluate transaction history with information on millions of customers including their demographic characteristics from a unified repository.**
- **The continual need to evaluate more information.**
- **Timely and accurate information.**

Requirements

The ideal tool set for the data warehousing community is capable of performing real-time iterative search and discovery with large volumes of data. The data warehouse should include transaction history and demographically enhanced customer data, as well as, any additional economic or environmental data desired.

Costs

The average Fortune 500 company spends anywhere between twenty and eighty million dollars to create a data warehouse and approximately ten million dollars each year to support the infrastructure of hardware, software and personnel.

Resource expenditures include:

- **Vast amounts of hardware are applied to compensate for the inefficient use of transaction database management systems for decision support applications.**
- **Human resources are diverted to support large hardware, software and operational infrastructures.**
- **Business decisions based on unasked and unanswered questions due to a lack of quick and easy information.**
- **DataMarts, which are subsets of data, resulting in redundant, stale and incomplete information.**

New Technology

DataPulse is a new approach to data warehousing. It redefines the data warehousing business by supporting three major categories of real-time analysis for millions of customers, and billions of business transactions. It represents one of a new generation of products that has been built on DataPulse's **DP DBMS** the first DBMS designed specifically for decision support applications.

The **DP DBMS** (version 8) is a database management system designed to support the requirements of decision support technology. It is comprised of advanced data representations with compact organizations and unique access strategies. This decision support DBMS uniquely addresses the massive I/O demands that are generated by all types of decision support applications, and with the advent of version 6, the **DP DBMS** has the additional capability of handling time series data.

DataPulse makes it quick and easy to incorporate prospects and customers, demographic enhancements, transaction history, as well as economic and environmental data, into a unified repository supporting MIS/EIS reporting. With the ability to analyze the data warehouse through cascade counts, descriptive statistics and comparative analysis you will maintain a pulse on your data.

The specialized graphical user interface on the User Workstations is comprised of specially developed C++ Controls, while the Database Server uses C++ in a variety of hardware configurations ranging from a single Intel processor, to Beowulf-type clusters of multiple Tera-byte machines.

All products based on the **DP DBMS** provide a wide range of unique capabilities facilitating the population, update, and retrieval of large amounts of data, typical of decision support applications.

For example, *DataPulse* can load in hours, what would normally take weeks, and can perform thirty table joins in minutes, which is faster than other products performing five table joins on any platform.

Summary of the *DataPulse* Data Warehouse

DataPulse's new DBMS was designed for use by decision support applications. It can be used on new, lower cost, Intel PC platforms which enhance its affordability. For the first time, many tera-bytes of information can be analyzed to answer questions about customers, prospects, and transaction history. Fortune 100 firms can enhance their data warehousing capabilities and smaller firms can now afford to utilize data warehousing applications.

Its advantages include:

- **Easy to Use**

Cascade counts, descriptive statistics and comparative analysis functions are intuitive for all business users.

- **Fast**

Performance will remain unsurpassed by any competing product that is dependent upon a DBMS designed for transaction system usage.

- **Handles large volumes of data**

Performs real-time analysis from among an excess of 100 million customers, hundreds of millions of current contracts, and billions of transactions.

- **Readily interfaces to in-house systems**

The *DataPulse* data warehouse repository can be readily fed hundreds of gigabytes of data from any transaction system. It can also export data to any transactional DBMS, in the necessary format, in order to support any existing applications and reporting.

- **Economically Feasible**

The *DataPulse* solution makes use of less expensive hardware platforms and software environments providing the lowest possible infrastructure cost.

- **Complete Installation and Training Support**

All products built on the *DP DBMS* include worldwide installation and training support.

- **Complete Help Desk Support**

All products built on the *DP DBMS* include twenty-four by seven help desk support.

- **Full investment is retained**

The client owns the solution.

About the Author

Jim Luisi has provided advanced database technology consulting to Fortune 500 companies for over twenty years. Having researched direct and mass marketing solutions for large direct marketing and advertising firms, he has been able to identify new technologies to meet the challenges of decision support systems.