

Data Warehousing Technology

A White Paper by Ken Orr

©Copyright 1996 by The Ken Orr Institute; revised edition, 2000

1.1) "Data in Jail" - the Data Access Crisis

If there is a single key to survival in the 1990s and beyond, it is being able to analyze, plan and react to changing business conditions in a much more rapid fashion. To do this, top managers, analysts and knowledge workers in our enterprises need more and better information.

Information technology itself has made possible revolutions in the way that organizations today operate throughout the world. But the sad truth is that in many organizations despite the availability of more and more powerful computers on everyone's desks and communication networks that span the globe, large numbers of executives and decision makers can't get their hands on critical information that already exists in the organization.

Every day organizations large and small create billions of bytes of data about all aspects of their business, millions of individual facts about their customers, products, operations and people. But for the most part, this data is locked up in a myriad of computer systems and is exceedingly difficult to get at. This phenomenon has been described as "data in jail".

Experts have estimated that only a small fraction of the data that is captured, processed and stored in the enterprise is actually available to executives and decision makers. While technologies for the manipulation and presentation of data have literally exploded, it is only recently that those involved in developing IT strategies for large enterprises have concluded that large segments of the enterprise are "data poor."

1.2) Data Warehousing - Providing Data Access to the Enterprise

Recently, a set of significant new concepts and tools have evolved into a new technology that makes it possible to attack the problem of providing all the key people in the enterprise with access to whatever level of information needed for the enterprise to survive and prosper in an increasingly competitive world.

The term that has come to characterize this new technology is "data warehousing." Data Warehousing has grown out of the repeated attempts on the part of various researchers and organizations to provide their organizations flexible, effective and efficient means of getting at the sets of data that have come to represent one of the organization's most critical and valuable assets.

Data Warehousing is a field that has grown out of the integration of a number of different technologies and experiences over the last two decades. These experiences have allowed the IT industry to identify the key problems that have to be solved.

1.3) Operational vs. Informational Systems

Perhaps the most important concept that has come out of the Data Warehouse movement is the recognition that there are two fundamentally different types of information systems in all organizations: operational systems and informational systems.

"Operational systems" are just what their name implies; they are the systems that help us run the enterprise operation day-to-day. These are the backbone systems of any enterprise, our "order entry", "inventory", "manufacturing", "payroll" and "accounting" systems. Because of their importance to the organization, operational systems were almost always the first parts of the enterprise to be computerized. Over the years, these operational systems have been extended and rewritten, enhanced and maintained to the point that they are completely integrated into the organization. Indeed, most large organizations around the world today couldn't operate without their operational systems and the data that these systems maintain.

On the other hand, there are other functions that go on within the enterprise that have to do with planning, forecasting and managing the organization. These functions are also critical to the survival of the organization, especially in our current fast-paced world. Functions like "marketing planning", "engineering planning" and "financial analysis" also require information systems to support them. But these functions are different from operational ones, and the types of systems and information required are also different. The knowledge-based functions are informational systems.

"Informational systems" have to do with analyzing data and making decisions, often major decisions, about how the enterprise will

operate, now and in the future. And not only do informational systems have a different focus from operational ones, they often have a different scope. Where operational data needs are normally focused upon a single area, informational data needs often span a number of different areas and need large amounts of related operational data.

In the last few years, Data Warehousing has grown rapidly from a set of related ideas into an architecture for data delivery for enterprise end-user computing.

2) Understanding the Framework of the Data Warehouse

One of the reasons that data warehousing has taken such a long time to develop is that it is actually a very comprehensive technology. In fact, data warehousing can be best represented as an enterprise-wide framework for managing informational data within the organization. In order to understand how all the components involved in a data warehousing strategy are related, it is essential to have a Data Warehouse Architecture.

2.1) A Data Warehouse Architecture

A Data Warehouse Architecture (DWA) is a way of representing the overall structure of data, communication, processing and presentation that exists for end-user computing within the enterprise. The architecture is made up of a number of interconnected parts:

- Operational Database / External Database Layer
- Information Access Layer
- Data Access Layer
- Data Directory (Metadata) Layer
- Process Management Layer
- Application Messaging Layer
- Data Warehouse Layer
- Data Staging Layer

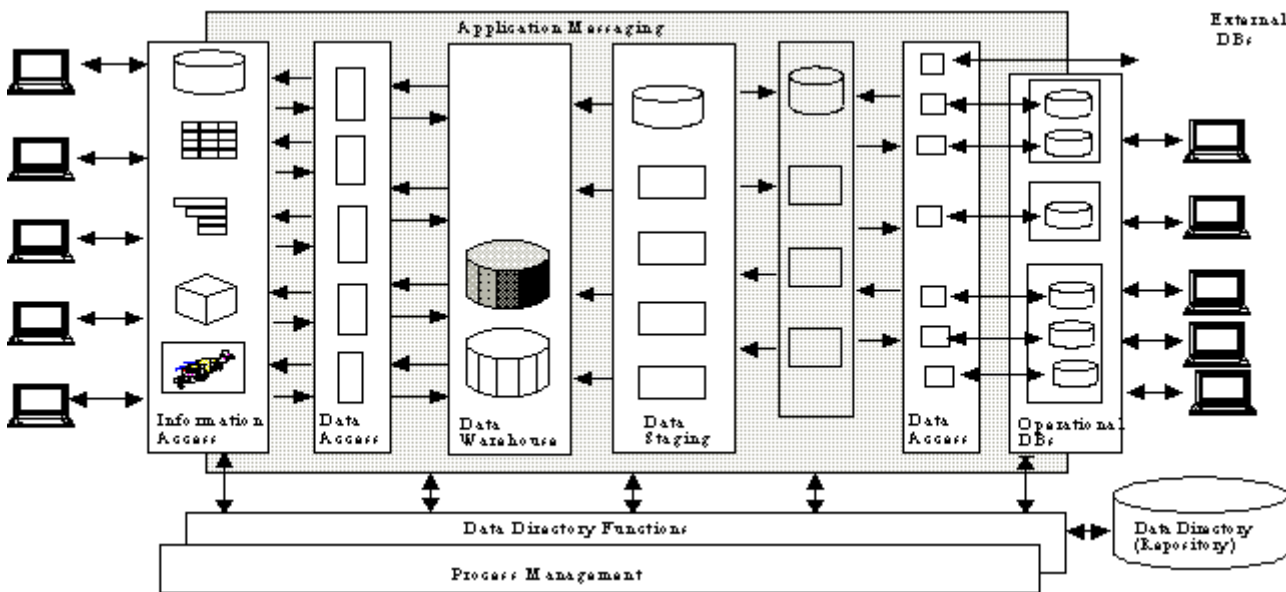


Figure 1 - Data Warehouse Architecture

2.2) Operational Database / External Database Layer

Operational systems process data to support critical operational needs. In order to do that, operational databases have been historically created to provide an efficient processing structure for a relatively small number of well-defined business transactions. However, because of the limited focus of operational systems, the databases designed to support operational systems have difficulty accessing the data for other management or informational purposes. This difficulty in accessing operational data is amplified by the fact that many operational systems are often 10 to 15 years old. The age of some of these systems means that the data access technology available to obtain operational data is itself dated.

Clearly, the goal of data warehousing is to free the information that is locked up in the operational databases and to mix it with information from other, often external, sources of data. Increasingly, large organizations are acquiring additional data from outside databases. This information includes demographic, econometric, competitive and purchasing trends. The so-called "information superhighway" is providing access to more data resources every day.

2.3) Information Access Layer

The Information Access layer of the Data Warehouse Architecture is the layer that the end-user deals with directly. In particular, it represents the tools that the end-user normally uses day to day, e.g., Excel, Lotus 1-2-3, Focus, Access, SAS, etc. This layer also includes the hardware and software involved in displaying and printing reports, spreadsheets, graphs and charts for analysis and presentation. Over the past two decades, the Information Access layer has expanded enormously, especially as end-users have moved to PCs and PC/LANs.

Today, more and more sophisticated tools exist on the desktop for manipulating, analyzing and presenting data; however, there are significant problems in making the raw data contained in operational systems available easily and seamlessly to end-user tools. One of the keys to this is to find a common data language that can be used throughout the enterprise.

2.4) Data Access Layer

The Data Access Layer of the Data Warehouse Architecture is involved with allowing the Information Access Layer to talk to the Operational Layer. In the network world today, the common data language that has emerged is SQL. Originally, SQL was developed by IBM as a query language, but over the last twenty years has become the de facto standard for data interchange.

One of the key breakthroughs of the last few years has been the development of a series of data access "filters" such as EDA/SQL that make it possible for SQL to access nearly all DBMSs and data file systems, relational or nonrelational. These filters make it possible for state-of-the-art Information Access tools to access data stored on database management systems that are twenty years old.

The Data Access Layer not only spans different DBMSs and file systems on the same hardware, it spans manufacturers and network protocols as well. One of the keys to a Data Warehousing strategy is to provide end-users with "universal data access". Universal data access means that, theoretically at least, end-users, regardless of location or Information Access tool, should be able to access any or all of the data in the enterprise that is necessary for them to do their job.

The Data Access Layer then is responsible for interfacing between Information Access tools and Operational Databases. In some cases, this is all that certain end-users need. However, in general, organizations are developing a much more sophisticated scheme to support Data Warehousing.

2.5) Data Directory (Metadata) Layer

In order to provide for universal data access, it is absolutely necessary to maintain some form of data directory or repository of meta-data information. Meta-data is the data about data within the enterprise. Record descriptions in a COBOL program are meta-data. So are DIMENSION statements in a FORTRAN program, or SQL Create statements. The information in an ERA diagram is also meta-data.

In order to have a fully functional warehouse, it is necessary to have a variety of meta-data available, data about the end-user views of data and data about the operational databases. Ideally, end-users should be able to access data from the data warehouse (or from the operational databases) without having to know where that data resides or the form in which it is stored.

2.6) Process Management Layer

The Process Management Layer is involved in scheduling the various tasks that must be accomplished to build and maintain the data warehouse and data directory information. The Process Management Layer can be thought of as the scheduler or the high-level job control for the many processes (procedures) that must occur to keep the Data Warehouse up-to-date.

2.7) Application Messaging Layer

The Application Message Layer has to do with transporting information around the enterprise computing network. Application Messaging is also referred to as "middleware", but it can involve more than just networking protocols. Application Messaging for example can be used to isolate applications, operational or informational, from the exact data format on either end. Application Messaging can also be used to collect transactions or messages and deliver them to a certain location at a certain time. Application Messaging in the transport system underlying the Data Warehouse.

2.8) Data Warehouse (Physical) Layer

The (core) Data Warehouse is where the actual data used primarily for informational uses occurs. In some cases, one can think of the Data Warehouse simply as a logical or virtual view of data. In many instances, the data warehouse may not actually involve storing data.

In a Physical Data Warehouse, copies, in some cases many copies, of operational and or external data are actually stored in a form that is easy to access and is highly flexible. Increasingly, Data Warehouses are stored on client/server platforms, but they are often

stored on main frames as well.

2.9) Data Staging Layer

The final component of the Data Warehouse Architecture is Data Staging. Data Staging is also called copy management or replication management, but in fact, it includes all of the processes necessary to select, edit, summarize, combine and load data warehouse and information access data from operational and/or external databases.

Data Staging often involves complex programming, but increasingly data warehousing tools are being created that help in this process. Data Staging may also involve data quality analysis programs and filters that identify patterns and data structures within existing operational data.

3) Data Warehouse Options

There are perhaps as many ways to develop data warehouses as there are organizations. Moreover, there are a number of different dimensions that need to be considered:

- Scope of the data warehouse
- Data redundancy
- Type of end-user

Figure 2 shows a two-dimensional grid for analyzing the basic options, with the horizontal dimension indicating the scope of the warehouse, and the vertical dimension showing the amount of redundant data that must be stored and maintained.

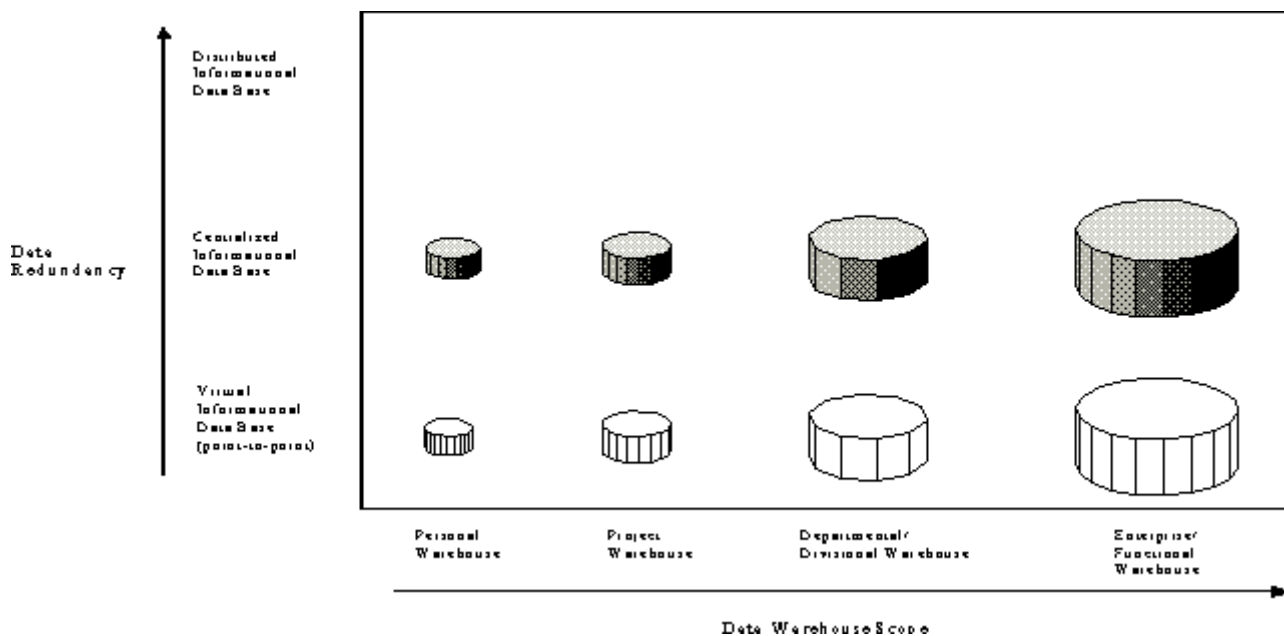


Figure 2 - Data Warehouse Options

3.1) Data Warehouse Scope

The scope of a data warehouse may be as broad as all the informational data for the entire enterprise from the beginning of time, or it may be as narrow as a personal data warehouse for a single manager for a single year. There is nothing that makes one of these more of a data warehouse than another.

In practice, the broader the scope, the more value the data warehouse is to the enterprise and the more expensive and time consuming it is to create and maintain. As a consequence, most organizations seem to start out with functional, departmental or divisional data warehouses and then expand them as users provide feedback.

3.2) Data Redundancy

There are essentially three levels of data redundancy that enterprises should think about when considering their data warehouse options:

- "Virtual" or "Point-to-Point" Data Warehouses
- Central Data Warehouses
- Distributed Data Warehouses

There is no one best approach. Each option fits a specific set of requirements, and a data warehousing strategy may ultimately include all three options

3.2.1) "Virtual" or "Point-to-Point" Data Warehouses

A virtual or point-to-point data warehousing strategy means that end-users are allowed to get at operational databases directly using whatever tools are enabled to the "data access network". This approach provides the ultimate in flexibility as well as the minimum amount of redundant data that must be loaded and maintained. This approach can also put the largest unplanned query load on operational systems.

As we will see, virtual warehousing is often an initial strategy in organizations where there is a broad but largely undefined need to get at operational data from a relatively large class of end-users and where the likely frequency of requests is low. Virtual data warehouses often provide a starting point for organizations to learn what end-users are really looking for. Figure 3 below shows a Virtual Data Warehouse within the Data Warehouse Architecture.

3.2.2) Central Data Warehouses

Central Data Warehouses are what most people think of when they first are introduced to the concept of data warehouse. The central data warehouse is a single physical database that contains all of the data for a specific functional area, department, division, or enterprise. Central Data Warehouses are often selected where there is a common need for informational data and there are large numbers of end-users already connected to a central computer or network. A Central Data Warehouse may contain data for any specific period of time. Usually, Central Data Warehouses contain data from multiple operational systems.

Central Data Warehouses are real. The data stored in the data warehouse is accessible from one place and must be loaded and maintained on a regular basis. Normally, data warehouses are built around advanced RDBMs or some form of multi-dimensional informational database server.

3.2.3) Distributed Data Warehouses

Distributed Data Warehouses are just what their name implies. They are data warehouses in which the certain components of the data warehouse are distributed across a number of different physical databases. Increasingly, large organizations are pushing decision-making down to lower and lower levels of the organization and in turn pushing the data needed for decision making down (or out) to the LAN or local computer serving the local decision-maker.

Distributed Data Warehouses usually involve the most redundant data and, as a consequence, most complex loading and updating processes.

3.3) Type of End-user

In the same sense that there are lots of different ways to organize a data warehouse, it is important to note that there are an increasingly wide range of end-users as well. In general we tend to think in terms of three broad categories of end-users:

- Executives and managers
- "Power" users (business and financial analysts, engineers, etc.)
- Support users (clerical, administrative, etc.)

Each of these different categories of user has its own set of requirements for data, access, flexibility and ease of use.

4) Developing Data Warehouses

Developing a good data warehouse is no different from any other IT project; it requires careful planning, requirements definition, design, prototyping and implementation. The first and most important element is a planning process that determines what kind of data warehouse strategy the organization is going to start with.

4.1) Developing a Data Warehouse Strategy

Before developing a data warehouse, it is critical to develop a balanced data warehousing strategy that is appropriate for its needs and its user population. Who is the audience? What is the scope? What type of data warehouse should we build?

There are a number of strategies by which organizations can get into data warehousing. One way is to establish a "Virtual Data Warehouse" environment. A Virtual Data Warehouse is created by: (1) installing a set of data access, data directory and process management facilities, (2) training the end-users (3) monitoring how the data warehouse facilities are actually used and then (4) based on actual usage, create a physical data warehouse to support the high-frequency requests.

A second strategy is simply to build a copy of the operational data from a single operational system and enable the data warehouse from a series of information access tools. This strategy has the advantage of being both simple and fast. Unfortunately, if the existing data is of poor quality and/or the access to the data has not been thought through, then this approach can create a number of significant problems.

Ultimately, the optimal data warehousing strategy is to select a user population based on value to the enterprise and do an analysis of their issues, questions and data access needs. Based on these needs, prototype data warehouses are built and populated so the end-users can experiment and modify their requirements. Once there is general agreement on the needs, then the data can be acquired from existing operational systems across the enterprise and/or from external data sources and loaded into the data warehouse. If it is required, information access tools can also be enabled to allow end-users to have access to required data using their own favorite tools or to allow for the creation of high-performance multi-dimensional information access systems using the core data warehouse as the basis.

In the final analysis, there is no one approach to building a data warehouse that will fit the needs of every enterprise. Each enterprise's needs are different as is each enterprise's context. In addition, since data warehouse technology is evolving as we learn more about developing data warehouses, it turns out that the only practical approach to data warehousing is an evolutionary one.

4.2) Evolving a Data Warehouse Architecture

The Data Warehouse Architecture in Figure 1 is simply a framework for understanding data warehousing and how the components of data warehousing fit together. Only the most sophisticated organizations will be able to put together such an architecture the first time out. What the Data Warehouse Architecture provides then is a kind of roadmap that can be used to design toward. Coupled with an understanding of the options at hand, the Data Warehouse Architecture provides a useful way of determining if the organization is moving toward a reasonable data warehousing framework.

One of the keys to data warehousing is flexibility. It is critical to keep in mind that the more successful a data warehouse strategy is, the more end-users are going to want to add to it.

4.3) Designing Data Warehouses

Designing data warehouses is very different from designing traditional operational systems. For one thing, data warehouse users typically don't know nearly as much about their wants and needs as operational users. Second, designing a data warehouse often involves thinking in terms of much broader, and more difficult to define, business concepts than does designing an operational system. In this respect, data warehousing is quite close to Business Process Reengineering (BPR). Finally, the ideal design strategy for a data warehouse is often outside-in as opposed to top-down.

But while data warehouse design is different from what we have been used to, it is no less important. The fact that end-users have difficulty defining what they need as a bare minimum is no less necessary. In practice, data warehouse designers find that they have to use every trick in the book to help their users "visualize" their requirements. In this respect, robust working prototypes are essential.

5) Managing Data Warehouses

Data Warehouses are not magic-they take a great deal of very hard work. In many cases data warehouse projects are viewed as a stopgap measure to get users off our backs or to provide something for nothing. But data warehouses require careful management and marketing. A data warehouse is a good investment only if end-users actually can get at vital information faster and cheaper than they can using current technology. As a consequence, management has to think seriously about how they want their warehouses to perform and how they are going to get the word out to the end-user community. And management has to recognize that the maintenance of the data warehouse structure is as critical as the maintenance of any other mission-critical application. In fact, experience has shown that data warehouses quickly become one of the most used systems in any organization.

Management, especially IT management, must also understand that if they embark on a data warehousing program, they are going to create new demands upon their operational systems: demands for better data, demands for consistent data, demands for different kinds of data.

6) Future Developments

Data Warehousing is such a new field that it is difficult to estimate what new developments are likely to most affect it. Clearly, the development of parallel DB servers with improved query engines is likely to be one of the most important. Parallel servers will make it possible to access huge data bases in much less time.

Another new technology is data warehouses that allow for the mixing of traditional numbers, text and multi-media. The availability of improved tools for data visualization (business intelligence) will allow users to see things that could never be seen before.

7) Conclusion

Data Warehousing is not a new phenomenon. All large organizations already have data warehouses, but they are just not managing them. Over the next few years, the growth of data warehousing is going to be enormous with new products and technologies coming out frequently. In order to get the most out of this period, it is going to be important that data warehouse planners and developers have a clear idea of what they are looking for and then choose strategies and methods that will provide them with performance today and flexibility for tomorrow.

[[Home](#)] [[Up](#)] [[home](#)]



The Ken Orr Institute

5883 S.W. 29th St., Suite 101, Topeka, KS 66614

Phone: 785.228.1200 Fax: 785.228.1201

Email: webmaster@kenorrinst.com

kenorr@kenorrinst.com

