A practical guide to getting started with Data Warehousing
The following terms are trademarks or registered trademarks of the IBM Corporation in the United States, other countries or both:

AIX  
AS/400  
Data Guide  
DB2 DataJoiner  
DataPropagator  
DB2  
DB2 OLAP Server  
DB2 Universal Database  
IBM  
IMS  
Information Warehouse  
Intelligent Miner  
NetFinity  
OS/2  
OS/400  
OS/390  
RISC System/6000  
SP2  
Visual Warehouse

UNIX is a registered trademark of The Open Group.

Windows NT is trademark of Microsoft Corporation in the United States, other countries or both.

Microsoft, Windows, and Windows NT are trademarks of Microsoft Corporation in the United States, other countries or both.

Other company, product, and service names, which might be denoted by a double asterisk (**), might be trademarks or service marks of others.

© International Business Machines Corporation 1999  
10-99  
All Rights Reserved  
This report was prepared for IBM by the Applied Technologies Group. The text is the property of the Applied Technologies Group. IBM has been granted unrestricted rights to produce the text in this format. For additional information, contact the Applied Technologies Group, One Apple Hill, Suite 216, Natick, MA 01760, Tel. (508) 651-1171.
A practical guide to getting started with Data Warehousing
Visual Warehouse:  
The IBM data warehouse & datamart solution

Visual Warehouse is an integrated package of business software that shortens the distance between mountains of complex data and insightful business decisions that deliver a competitive edge. Based on innovative, proven technology, Visual Warehouse can help companies both large and small uncover entirely new insights about their businesses, their markets, their competitors, and their customers. These insights can improve the bottom line by providing a more comprehensive and focused view of the market realities that affect day-to-day business operations.

“We needed to build a very large data warehouse on a massively parallel processing platform. Very few companies play well in that world, and most are proprietary. The open architecture, massively parallel processing capability, and scalability of DB2 were the driving forces behind our decision.”

—Jane Landon, Systems Executive, Prudential Life Insurance of America

“IBM has an extensive range of database products that are very well complemented by a nice set of supporting business intelligence products such as Visual Warehouse.”

—Ray Ziegler, Assistant Vice President, Information Technology, CIGNA Reinsurance

“We need to be able to slice and dice our data by payer, physician, or geography, for strategic planning, budgeting, and negotiating contracts with the insurance community. Visual Warehouse equips us with the capabilities to do this and provides timely answers to high-level questions… We believe this data warehousing solution is going to be a journey over many years and will open us to new opportunities—this is just the starting point.”

—Ron Johnson, Chief Information Officer, Shore Memorial Hospital
Note to the reader: This guide is oriented to both business managers exploring the potential of data warehousing for their organizations, and technical implementers who are seeking practical advice on implementing data warehouse systems. To those unfamiliar with data warehousing terms we would suggest beginning with Appendix A, which reviews the major components and nomenclature. Business managers may wish to focus on the Criteria section (specifically the Business and Process Criteria) and the first few sections of Implementation Tactics. Readers looking for technical information may wish to focus on Technology Criteria and the Implementation section. All readers may wish to look at the case studies that highlight IBM solutions for data warehousing.
Considerations for a successful warehousing project

Overview
Enterprises today, both nationally and globally, are perpetually seeking competitive advantages. It has become an incontrovertible axiom that information is the key to determining how to gain such a competitive advantage. The problem today is how to deal with the mountains of raw data which our ever more efficient information systems are collecting, massaging, processing, deriving, and disseminating. We literally are in the midst of a volcanic eruption of data. Somewhere hidden in this explosion of data is the clues management needs to define their strategic positioning in the market so as to maximize their competitive stance.

Into this picture, technology has inserted the concept of Data Warehousing as one alternative to coping with the well-known information overload described above. Data Warehouses exist to help management and decision makers transform raw data into information and to help management identify key trends. It helps the enterprise foresee predictable events and act in anticipation of those events and helps management understand the entire picture of what has already happened. It allows them to develop a good systemic understanding of events, thus allowing focused reactions to those events, such as redefining and reengineering business processes to take advantage of that understanding.

A clear prerequisite to enabling management to accomplish all of the above is the fact that the data to be analyzed has to be accessible and flexible, and it has to be available in a format that is usable by the requester. To date, too much emphasis has been placed on the raw technology which embodies the concept of Data Warehousing, and not enough on the underlying and concomitant strategy, planning, business processes, and services which develop, maintain, and use the Data Warehouse technology.

Experience has shown us that in projects where the technology has been perceived as a failure, the problem does not usually lie with the technology itself, but rather with the way in which the technology was applied. It is often applied to the wrong business problem, at the wrong scale, with insufficient training, and planning, and with little or no thought to how users need to access the data, etc. Some analysts’ statistics show that over 50% of warehouse projects fail to meet their stated objectives. In order to mitigate the risks associated in these projects, the project must work in close concert with the business community that will benefit from the warehouse and also need to have a solid grounding from a financial return perspective.

Data Warehousing technology can benefit enterprises at different levels and scales of implementation. This applies from departmental systems running on commodity platforms such as Microsoft NT, and database servers such as DB2 Universal Database” for Windows NT, on up to the enterprise level systems running on enhanced parallel MPP architectures such as IBM SP2 with DB2 Universal Database. Database Management Systems play key roles in the long-term viability of Data Warehouses. Issues such as easy access to operational data, scalability, and management of meta data (information about data in the warehouse) stand out. Therefore, thought needs to be given to what the initial strategy needs to be to ensure that an organization truly benefits from the technology. Moreover, we all know we are living in a very dynamic and fast changing business environment, where the only certainty is uncertainty and change. Therefore, the strategy must develop Warehouses such that they can grow and adapt to the changes we know are coming, even though we don’t today know what those changes are. This calls for prudent planning in developing the strategy to select architectures, which can react flexibly to changes in market dynamics, organizational restructuring, economic fluctuations, etc.

Aligning technology with business objectives

Data Warehousing is a rapidly maturing technology changing with every product release. There is no Rosetta Stone that will tell any one organization what works and what doesn’t. Consequently, experience and evolution are the best overall planning principles, which can be deployed by management today. There are a number of criteria, which should be kept in mind as strategies are developed. The criteria can be grouped into three categories: Business Criteria, Process Criteria, and Technology Criteria.

Business criteria
The first set of criteria have to do with the business problems at hand and what demands are made on the project by the business dynamics encountered:

Critical success factors: Why are you doing this?
No technology project will ever succeed if it is not properly aligned with the business Mission, Vision, and Goals. Therefore, it is vitally important to understand such elements of the strategy such as:

Fig. 1
• What is the problem at hand? Is it a problem related to cycle time, customer satisfaction, more cost-effective decision making, better business intelligence, or a general lack of information with which to make decisions on any of the above?

• Which of the departmental or enterprise goals and responsibilities are directly related to the problem at hand?

• What are the Critical Success Factors—those things that must be done well to solve the problem?

• Which of the organizational components of the enterprise is best positioned to solve the problem?

• Which audience within this organization will best use this technology and how: executives, financial analysts, scientists, engineers, clerical and administrative users, line managers, others and why do they need this (who will it benefit)?

• How can this technology be used to solve the problem? This is where the alignment of the technology to the problem will occur.

Quantifying benefits

Once the benefits of the technology have been aligned to the business objectives, they should then be quantified. The reason for this is that management must be able to answer the question: How will we know if this project is successful? The answer must fit the form: “This project will be successful if it allows us to achieve the following goals...”

In many organizations, quantifying benefits takes the form of a financial analysis, specifically a Return on Investment (ROI) analysis. A study by International Data Corporation (IDC), co-sponsored by IBM, showed that Data Warehousing could provide significant and impressive ROI numbers. The study, which included 62 participants, demonstrated that the overall ROI on warehouse projects was 401% with payback periods of two to three years.

What was interesting about the study, however, was that the smaller, departmental implementations, sometimes known as Data Marts, had a 533% ROI, while the larger, enterprise efforts showed an impressive 322% ROI.

The IDC study identified three kinds of benefits in the use of Data Warehouses.

• Cost avoidance benefits. These benefits were the ability not to spend money that is presently spent on generating endless reports to end-users. This included the resources expended by IT to generate answers on ad hoc queries. In many ways, Data Warehousing represents a liberation for IT by providing users with the tools they have needed over the years to generate their own reports. By allowing the users to do so, one eliminates the often endless loops of the user requesting a report from IT, IT delivering the report, the user either not approving the report due to some miscommunication or changing their minds after seeing what they asked for, etc.

• Efficiency gains from increased productivity among end-user professionals who gather and analyze data. The analyst who must stop an analysis to get information and who has to ask someone else to get the information loses efficiency in two ways. The first is in the loop described above, where there might be several iterations of request and response between the analyst and IT before the analyst is satisfied with the results of the request. The second is in the interruption of the analysis, and the inefficiencies associated with recovering the thought processes that were underway when the analysis was suspended.

• Warehouse dependent savings due to decisions based on analysis that could only come from data in a warehouse. This is a quality-of-decision issue that comes from the fact that certain data associations may not exist in any one operational system and therefore are not available to the analyst. By building those requisite associations and relations in the Warehouse, a situation where the whole is greater than the sum of its parts occurs and those relationships allow the analyst to do their work better, and become more effective. This is not a case of people making bad decisions prior to the advent of the system; this is about giving people better tools to empower them to do better work.

Not all benefit quantifiers are in terms of ROI. A recent article described how firms often eschew formal ROI analysis because they consider the data warehouse a strategic investment. In this case, these organizations were convinced prima facie that the benefits would be worth the costs. A cautionary word, however: make sure that there is an understanding of the projected costs before starting the project if there is no formal financial measurement technique such as ROI.

Regardless of the culture of the organization, whether it accepts soft benefits in its approval process or not, be sure to quantify the anticipated benefits in business terms. Without this, the organization will have no metrics to determine whether or not the project is successful.
The IDC finding would lead one to believe that since the data marts delivered the highest ROI, a viable strategy could be to allow each department to implement their own data marts independently. The problem with this thinking is that it ignores the issue of cross-functional analysis. Let us examine what happens when Finance and Marketing each develop their own data marts independently of each other. Let us further assume that each keeps track of sales, but defines them differently. For example, marketing defines a sale as a booking while finance defines it as a payment. What are some of the potential ramifications of this situation? A primary consequence is that analyses of the same business dynamic—sales—may result in two departments having significantly different views and analyses. More importantly, suppose finance needs something from the marketing Data Mart, or vice versa. Without a unifying design or data standard, the analysis of data that spans two or more organizational Data Marts will not be possible using this strategy. It is therefore important in developing a Warehousing strategy to understand the implications of an independent Data Mart strategy and factor these risks into its formulation. A more detailed discussion of independent vs. dependent Data Marts is presented in the Process Criteria.

“The heat is on…to get warehouses up and running fast” (but) “The biggest benefit comes down the road, when you can support 20 different decision support applications with the same architecture” —GartnerGroup.

This quote from the GartnerGroup highlights the need to develop an overall strategy and architecture for the organization prior to or at least in parallel with the development of the first Warehouse project. At the beginning of these projects, everyone is in a hurry to reap the immediate benefits of the systems and becomes impatient with the planning process, which is sometimes seen to slow things down.

However, experience tells us the payoffs will be that much greater, and subsequently improvements will come that much sooner if a sufficient amount of time is invested at the front end of the project in defining an overall enterprise architecture into which the Data Marts can connect.

Understanding product integration issues
The current maturity level of the Data Warehousing industry leads to a proliferation of many disconnected offerings in the marketplace. Many small vendors have joined the fray with products targeted at one or two elements of the Data Warehousing architecture. As a result, there are very few offerings in the market, which answer all of the needs of a potential end user. This leads to a risk regarding how well different packages will integrate, which common platforms are supported, and so on. A trend is emerging where a number of large vendors form integrated product teams with smaller vendors that hold a solid solution in an important warehousing niche. For example, if a large vendor had a great DBMS solution, they might team with product vendors with extract technology, data cleansing technology or specialized OLAP tools. One benefit of these vendor consortiums is to agree on common approaches to sharing meta data between tools and databases from different vendors. Meta data, with its global significance, is a key to product integration in contemporary data warehousing solutions. The “investment” of these vendors integrating products from different vendors benefits end users by resolving the tricky and risky issues associated with integrating products from different vendors.

Analytical applications
One of the fastest growing trends in the Data Warehousing market is the emergence of packaged analytical applications. The ability to script data analysis applications to derive particular end results is paramount to the evolution of effective data warehousing solutions. In this regard, early data warehousing solutions were all one-off customized applications built to deal with particular business challenges. This continues to be an important part of data warehousing activities. But in addition, as applications and business needs become more generalized and as custom applications begin to be widely deployed, there is a viable place for sets of off-the-shelf analysis applications that can be deployed quickly and effectively to meet many common business problems. These off-the-shelf applications also include the capacity for semi-customization and are the trend of the future. The costs of these applications should continue to drop, while the real value will continue to grow.

Dealing with cultural issues
Data Warehousing is about pooling resources (data), which implies sharing, which in turn can imply a loss of control, a concept sometimes inimical to many data owners. This kind of organizational provincialism can sometimes throw up impediments to a Warehousing project and must be dealt with in the early phases of the project.

Business unit and process considerations
Technology is only useful to the extent that it supports our ability to carry out our assignments and achieve our corporate goals. Therefore the introduction of any new technology must be aligned with the business units and processes which it is intended to support. The IT organization may or may not have all the requisite technical skills, but IT will not successfully implement a warehouse project without the involvement and commitment of the business unit. Too often, technology is developed independently of any business process considerations, many times with catastrophic results.

Choosing technical services
Very few, if any, IT organizations have the requisite combination of skills and resources required to perform all of the technology, planning and implementation tasks required for successful Data Warehouse projects. This is not intended as an affront to IT organizations, but rather a simple observation that it will take a broad spectrum of talent which crosses many disciplines to make this work. The chances that a single IT organization will have all of these talents available for this project at the same time are small. Therefore, at some point in time, many organizations will need to locate a partner to consult in the technical planning for the Warehouse and then eventually assist in the implementation of the system. The partner selected must be able to operate within the constraints enumerated in the organization's Warehousing strategy, including the methodology chosen.
implementation style chosen (see Process Criteria - Methods), and so on. Choosing the wrong partner, for example one who has Data Warehousing experience but not Data Mart experience, or one who does not have experience across the entire spectrum of products and services, can increase the risks associated with these systems. For example, IBM offers specific services in conjunction with its Visual Warehouse solution as well as custom services for any scope of warehouse implementation.

The state of standards

OMG committee for common warehouse meta data

IBM, in conjunction with Oracle and Unisys, is sponsoring an OMG (Open Management Group) subcommittee for the standardization of Common Warehouse Meta Data. The objectives of this committee are to establish an industry standard for common warehouse meta data interchange and to provide a generic mechanism that can be used to transfer a wide variety of warehouse meta data. The intent is to define a rich set of warehouse models to facilitate the sharing of meta data, to adopt open APIs (Java and Corba) for direct tool access to meta data repositories, and to adopt XML as the standard mechanism for exchanging meta data between tools. The subcommittee, chaired by IBM, is in the process of accepting vendor proposals for the above objectives.

OMG committee for XML/XMI

A related OMG subcommittee has been formed to standardize XML Meta Data Interchange (XMI). IBM, Unisys and other industry leaders are also involved in this work. IBM and Unisys have submitted a proposal co-submitted by Oracle, DSTC, and Platinum Technology and supported by numerous other vendors. The proposal for an XML Meta Data Interchange Format specifies an open information interchange model that is intended to give developers working with object technology the ability to easily interchange meta data between modeling tools and between tools and meta data repositories. In a data-warehousing context, the proposal defines a stream-based interchange format for exchanging instances of UML models.

One of the inherent risks of a new technology is the lack of standards, and Data Warehousing is no exception. There are countless examples of competing technologies that resolved themselves into one standard, and most of the time the resolution creates winners and losers. Eight-track tape owners were losers in the technology battle with cassettes, Betamax owners lost against VHS in video recording, CPM lost to MS/DOS, and the list goes on. Obviously, one risk mitigation strategy in this arena is to align the project with a big player in the industry—one that will have an influence in determining the winning standards. Another innovative mechanism which was used in a major bank IT shop was to include in the cost/benefit analysis an estimated cost to bury the existing system and replace it with a new one in case the wrong decision was made. If the project still made sense after including this cost, then the bank would go ahead with it in spite of the standards risk.

Lack of attention to training

For many years, exposure to the world of the database was limited to the inhabitants of the glass house—the IT department. Consequently, many end users are not familiar with the concepts behind navigating a data schema or unraveling the mysteries of joining tables via keys to get queries answered. A liberation of sorts will ensue from allowing users access to their own data. However, the users must be prepared for life in this liberation and must be ready to accept the responsibilities of runaway queries, etc. Therefore, it is incumbent on management to make sure that adequate training is provided to users to allow them to use the system effectively. They have to use the system without getting so frustrated that they give up, or worse, poison the project by maligning it to others. They cannot bring the system to its knees by constructing queries, which run forever, or base an analysis on faulty data because the user was not familiar enough with the system to understand the information for which he or she was asking.

Process criteria

A number of elements of the Warehousing strategy have to do with processes: processes by which the strategy is implemented, and processes which are supported by the overall strategy.

Scope of effort: How big should it be?

Many large warehouse projects have failed because of an inability of the organization to handle the size and scope of the project. It is tempting to think of a single repository where all of the enterprise’s data problems can be solved in one fell swoop. And, if the organization can indeed come up with an integrated data model and solve all of the issues associated with such architecture, the benefits are indeed significant. However, this is sometimes not realistic and not necessary for the problem at hand. Industry studies have estimated the average size of a large warehouse project to be nearly two million dollars1 with a time to completion measured in years. No doubt, some business problems require the integration of data from many systems and will require a global strategy. Many others will not, and a simpler strategy, one of Data Marts, will likely be less risky.

This question of project scope should be readily answered from the exercise described earlier on aligning the technology to the corporate Vision and Mission. Once the questions relating to who needs the technology and which problems are being addressed are answered, the scope should be relatively straightforward to determine, which should allow management to allocate appropriate resources to manage it.

A strategy, which entails an enterprise wide scope, has certain implications associated with it, which should be understood. First and foremost it requires integration and cooperation among multiple organizational elements. Many issues will arise regarding different definitions of similar or identical terms, competing objectives and agendas, data parochialism and an unwillingness to give up control. This is a situation, which can be difficult to manage and successfully navigate. Oftentimes, change management is necessarily intertwined with this exercise, since the organization will have to wrestle with inter-departmental issues as
described above. Management must assess whether or not the organization is ready to deal with these kinds of issues, or whether they might not better wait until a more appropriate time.

A Departmental, or Data Mart approach is by definition smaller in scope, more focused in its outcome, quicker to achieve, less costly. However, there is a risk to developing Data Marts in a vacuum, as described in more detail in the method section below. Ideally there is a need to think globally about future integration with other departmental applications and Data Marts to avoid developing Data Islands.

Data Marts have their place and present a strong business case for starting with such an implementation, but if an organization determines that an enterprise warehouse is the appropriate strategy then there are many successful models to emulate.

**Implementation approach options**

Deciding whether a Data Warehouse or Data Mart is right for the enterprise is an appropriate beginning. It must, however, be followed by a decision on whether to buy an integrated package from a single vendor, engage a systems integrator to bring together a collection of best of breed products, or have the enterprise’s IT department create a best of breed solution.

The obvious advantage to dealing with an organization which can offer a complete solution is, of course, faster realization of business goals, usually at a reduced cost, and with a good degree of certainty that the ultimate solution will work (lower technological risk). These benefits do come at a price, though, and that price is the potential compromises that have to be made in functionality and performance by accepting the full suite of products from a single vendor. Further, the enterprise may have to adapt business processes to fit the specific characteristics of technology sourced from a single vendor.

The best of breed concept is certainly not new. It has as its foundation the tenet that the enterprise will be better off if it can somehow bring together the best extraction tools, databases, SMP/MPP hardware, disk drives, analysis tools, network, etc. and get them to function as a unified system. Aside from the difficulty and religious wars that accompany the attempts to define best, the price paid for the anticipated exceptional performance is primarily the pain associated with integrating the disparate components. Different vendors value different architectures and functionality characteristics, and the best extraction tools, for example, may not integrate well with the best meta data repository, and so on. Many implementations of components which “should” integrate, are advertised as “compatible”, but require extensive work to get all details to work for an organization’s application. This kind of scenario is especially significant in technologies which are young, and in which standards are not yet entrenched. Data Warehousing meta data standards are still evolving, although IBM, in conjunction with Oracle and Unisys, is sponsoring an OMG (Open Management Group) subcommittee for the standardization of Common Warehouse Meta Data. This will result in a rich set of warehouse models to facilitate the sharing of meta data. Another OMG subcommittee has also been formed to standardize XML Meta Data Interchange (XMI). These standards initiatives will go a long way to resolving these important issues. Most integrators find that all projects require compromises to achieve integration, although some level of custom fit with the enterprise is generally achieved.

There is a second, subtler price associated with this custom fit, and that is what to do in the future about upgrading to later releases. Assume, for example, that an integrator combines extraction tool A with meta data repository B, database C and analysis tools D. Assume further that all of the products, (A, B, C, and D) are at release 1.0. The integrator finishes the job, and the customer is satisfied, and some time later extraction tool A moves to release 2.0, three months from then the analysis tool moves to release 2.0, and so on, each release bringing with it features and functions the organization would like to incorporate. The challenge is: how to do that? Are these independent release updates compatible with each other? Will the resultant system be backward compatible with the original system? If the organization continues to rely on the integrator to maintain a system at the latest best-of-breed status, this is tantamount to a full employment act for the integrator, and certainly of dubious value to the customer. On the other hand, a customer who has purchased an integrated package from a single vendor can put the onus of compatibility and upgrading on the shoulders of the vendor, and upgrade the warehouse to the next release by buying a single upgrade.

The third option, that of using the enterprise’s own IT shop to integrate the package is probably not a viable option except for very large and sophisticated organizations. Few IT shops have the expertise to integrate “best of breed” components, and fewer still have the available resources. Remember that one of the reasons to build a Warehouse is because the IT
Top-down and bottom-up approaches

How should an organization approach the design and construction of the Warehouse? Should it go for a top-down technique, setting up an enterprise-wide architecture and then constructing Warehouses or Data Marts that conform to that architecture? Or perhaps take a bottom-up philosophy, starting right in with highly focused and targeted Data Mart projects aimed at specific critical areas of the business? The issues here have to do with deciding between addressing short-term tactical requirements to help individual departments and long term strategic planning issues regarding data architectures that have to cut across political and organizational boundaries.

The Top-Down approach can yield the best long-term results, but it also invokes the most angst within an organization. It is very difficult, expensive, and time consuming to achieve consensus on a single, consistent, accepted and valid view of the business, the data it needs, etc. In the prior example of the Marketing and Finance departments, two different operational definitions for “sales” were developed. Which of the definitions will be used in the Warehouse, or will there be two terms that must now be defined? If so, they cannot both be called sales, and the meta data had better be clear as to what the final determination became. It is clear that many more change management issues will have to be dealt with in this approach, as organizations grapple with the sins of the past in not having promulgated an accepted data standardization program, etc.

The Bottom-Up approach favors the use of smaller, more focused applications of Warehouses that can avoid the pitfalls of the Top-Down approach by simply limiting the extent of the implementation. This approach also exhibits simpler data archaeology problems: there are usually limited data sets, limited user views, a good understanding of the data needs and how they relate to the business problem. In its purest form, this approach trades the near term pain of dealing with data standardization issues for the longer term inability to operate cross-organizationally.

In this approach, each department is responsible for extracting whatever data they need, defining their own meta data and using their own private Warehouses for decision support at the departmental level. Three obvious problems arise: (1) this architecture is difficult to scale up to an enterprise view; and (2) the lack of standardization prevents analysts in one organization from accessing information from another organization’s Warehouse which might be of use to them in their analysis. And (3) they may derive different answers to the same question—e.g., what were sales last month. This is demonstrated by the brick walls inserted between departmental systems in Fig. 3 preventing interactions. The primary cause of this is the Warehouse Tower of Babel syndrome inherent in the underlying philosophy. Each mart can create confusing, overlapping and contradicting views of the business, like the proverbial six blind men trying to describe an elephant, each only being able to relate to the elephant according to the portion of the animal they were feeling. What is a customer? a product? a sale?

This approach works if the organization has a business problem with a single focus and the data to solve that problem exists in only a few places, with no political ownership issues.

Some organizations use a hybrid approach to gain the speed and cost advantages of the highly focused departmental approach, yet at the same time making sure that the implementation is consistent with the overall goals of the organization.

This approach uses the principles developed in Rapid Application Development (RAD) methodologies and intentionally delivers iterations of the departmental Warehouse, attempting at each iteration to come closer to an overall enterprise data model and data architecture.

This allows the organization to take advantage of the speed and cost savings of the smaller approach while at the same time mitigating and eventually overcoming the lack of integration and islands of automation problems inherent in this approach. Other organizations are using the smaller implementation as a proof of concept and prototype/pilot installation. This assists in proving the benefits of the technology on a smaller scale, and smaller risk, and then scaling the solution into a more
Technology criteria

The technology dimension will of course play a major role in the enterprise strategy. Different strategies will require different technological characteristics and features. Just as the technology must align itself with the business mission, the strategy must also consider the technology.

Scalability

Scalability refers to the ability of a system to increase in capacity as users demand more, as data stores grow, as more users are added to the system and as more applications are developed against the Warehouse.

One of the challenges associated with introducing new technologies and new applications is that of determining the actual system load after implementation. JAD (Joint Application Development) sessions attempt to mitigate the risk by attempting to produce a picture of user requirements, but the truth is that users who have never had the opportunity to work with a new technology don't really know what to ask for. Further, they don't know ahead of time what kinds of demands they will make on the system until they actually sit down and start working with it. Therefore, as the users become familiar with the query capabilities and the navigational issues, their own success will cause them to demand more from the system as word spreads and more users exploit the features of the system. In time, users will become more sophisticated and begin exploring with ideas of data mining and visualization as their analyses become more complex. All of these are factors that demand scalability in a system. Companies acquiring warehousing technology need to be assured that scalability is "built-in" to the architecture.

Manageability

Data Warehouses require the development and implementation of new processes, tools, and work systems to manage the extraction/transformation of operational data, the administration of users (adding, deleting, changing access rights) and so forth. In the course of deploying a Warehouse, a number of operational management decisions must be made and supported by the product solution set:

- What is the relationship and meaning of the data being loaded to the intended business use? How frequently should data loads and transformations be made? Should they be daily, weekly, monthly, quarterly?
- How will the Warehouse reporting adapt to the business processes as they change over time? How will the system model and track the business?
- How much data needs to come in on each load, and how long will it take for the system to recompute all the indexes, meta data updates, and other administrative details that must be undertaken? How will the system monitor database operations?
- How will the system deal with backups/ restores and what should the process be to administer a disaster recovery plan?
- If the system is to be paid for by usage chargebacks, is there a mechanism for the systems administrator to keep track of usage by account number or password, and are there reports available to facilitate this feature?

Performance

How well the system performs will be the ultimate arbiter in the success or failure of the project. The intent of the Warehouse is to help people do their jobs more effectively and efficiently. If the response time is not adequate, users will not use the system, the enterprise will not derive any benefit from the expenditure, and the project may wind up with a negative ROI. Therefore the technology dimensions of performance must be thoroughly considered in developing a strategy.

Over the years, there have been a number of studies to determine the limits of human patience in dealing with computer response times. In general, users want to see something back in a timeframe that does not interrupt their thought processes. Some have said that two to five seconds is a good target response time. But when we are dealing with such gargantuan database sizes, it is difficult to conceive of doing a table scan on a multi-million row table in that timeframe, which leads us to the second human factors point. That is the fact that the amount of time a user will wait is proportional to the perceived difficulty of the procedure requested. Therefore, if a user knows they have entered a particularly nasty query, they will be more tolerant of delay. The best advice on this subject is to work with the user community in establishing meaningful metrics and working cooperatively to set and meet expectations on both sides.

There are several areas that directly impact the performance of a system: the hardware architecture and the database architecture. The hardware dimension is simply the use of Symmetric Multi Processing (SMP) and Massively Parallel Processing (MPP) architectures. The top end of the MPP capacity does outstrip the top end capabilities of SMP. However, the applications where this kind of performance is necessary are few and far between. For this reason, many industry analysts are predicting that SMP will be the hands down winner in the Warehousing market. This does not mean that every warehouse implementation should be on a parallel technology of one sort or another. Many applications can perform satisfactorily on non-parallel systems.

Some database vendors have followed the hardware architecture by introducing parallelism into the database functions. For example, IBM has continually enhanced products such as DB2 and it now provides capabilities such as parallel query, loads, joins, scans, and utility processing providing tremendous value.
Flexibility
If there is one thing that distinguishes market conditions today it is the pace of change. In fact, as pointed out in the introduction, one of the drivers leading enterprises to consider Warehousing technology is the need to keep up with that change. Deployment of a Warehouse will not alter the pace of change or the need to keep up with the change, which means the Warehouse technology itself must be flexible, allowing for rapid responses to changing conditions. The Warehouse must be adaptable to changes in the enterprise, such as reorganizations, mergers, and acquisitions. It may be necessary to implement new queries based on responding to a competitors product introduction—queries not envisioned in the original design. If the database needs to be redesigned, reloaded or indexed, this could mean a significant delay in responding to the competition.

Ease/speed of implementation
The development and maintenance tools available with the Warehouse will be key to the success of the project. Are the tool sets tightly integrated, as is usually the case in a one-stop-shop solution, or will the development team have to wrestle with the tools as well as with the disparate products to make them work together? Are the user interfaces graphical or the old command line metaphor? Do the development tools automatically generate the meta data content, or will there be a second step to build the meta data and then a third to reconcile the mistakes made in doing this manually? These considerations apply not only to the initial deployment of the Warehouse, but also speak to the flexibility, since changes to the Warehouse will likely at some point involve the use of these tools again. A trade article in Datamation put it best: “Having tools that are well and deeply integrated with each other is a necessity, not a luxury.”

Tool integration
Integration deals with how well and how smoothly the different architectural components interact. Obviously these are subjective terms, and therefore there are degrees of integration in many different areas such as administrative interfaces, Warehouse management, problem determination, etc.

Tight meta data integration would yield benefits from needing only to input meta data once and, once operational, preventing meta data from becoming out of synchronization. For example, an extraction tool and a meta data repository could interact according to several different models as follows:

- In the first model, each component has its own structures and syntax for the meta data and a third component is interposed between them to effect information transfer between them.

In Fig. 5, a translation module changes the formats of the information from one component to the other. Therefore, in this model there is no direct integration between components.

- A second option (see Fig. 6) is to have each module maintain different internal structures and syntax, but to have a common defined interface, such as a meta data standard. In this case, the accepted transfer mechanism and translation is built into each module. Compromises must still be made because the internal structural differences may yet interfere with some functions, but overall the interaction between the components is facilitated by the acceptance of a common mechanism for information interchange. This situation is considered to be moderately well integrated, and in fact do have an accepted architecture that each component follows to effect this integration.

- Finally, on the other end of the spectrum are tools that use the same structures and syntax internally as well as present a common interface to each other (see Fig. 7). These components are completely integrated. The more vendors involved in the solution, the more challenges exist integrating them—some single vendor solutions make this option relatively achievable.

The degree of integration of the Warehouse is an important consideration in developing an architecture and an implementation strategy. First of all, the Warehouse must integrate with the existing architecture and infrastructure of the organization. Reverting to the three models discussed earlier, if the Warehouse can only integrate with the existing architecture by means of extensive interposition of custom code, the project will be very expensive, lengthy and complex. Questions must be asked about the degree to which the proposed Warehouse must conform to existing supported operating systems, networks, data standards, existing databases, existing application development environments, and more.

If the Warehouse supports common, open architectures, such as in the second model, the likelihood of being able to add analysis functionality later, such as data mining for example, will be higher. Open architectures also make customization of products easier due to the more regular and predictable nature of the interaction among architectural components.
With regard to the existing operational applications from which the Warehouse will extract its information, the tighter the integration with all data sources, the better. Remember that some of the systems in the enterprise will have flat files or older hierarchical or network database architectures which can present a challenge. Which of the three models of integration are required for the enterprise?

Data can be extracted from the databases in discrete batches or continuously, on a transactional basis. Depending on the Warehouse application, either of these may be preferable or both may be required. Do extraction tools exist for both cases, or will there need to be some special code written? Finally, can the existing Warehouse automatically extract portions of the meta data from the existing operational systems, or will there need to be special software or processes employed to do so?

How easy will it be to integrate the Warehouse with existing operations, including administrative activities? Data extracts, cleansing, transformation and loading should not be done manually if it can be avoided. Not only does this introduce more cost in labor, but it can also interject too many vulnerabilities for error. Understanding all of the dimensions of integration with your existing environment and what is possible to automate is a must.

The Warehouse should also integrate with enterprise standards, both de jure standards, those promulgated by officially sanctioned bodies such as ANSI, OMG, the Meta Data Council and de facto standards that are practices and products generally accepted in the industry.

Completeness
The completeness of a solution talks to the existence of all the architectural components. This means that all of the parts an organization needs to work are in place and functional, from extraction and transformation to storage and meta data, to the analytical, management, and change processes required.

Many large vendors, particularly IBM, have invested significantly in their core technology and have evaluated the issues surrounding Warehouse technology and implementation and developed partnerships and processes that address the issues discussed in this section. The case studies that follow are good examples that illustrate many of the points just described.

Implementation tactics
Following is a high level guideline for implementing successful Data Warehousing projects.

Planning the project
Many organizations are in such a hurry to install a system that they tend to gloss over this vital step. To paraphrase what Lewis Carroll’s Cheshire Cat told Alice when she told him she didn’t know where she was going, if you don’t know where you’re going any road will get you there. The corollary, of course, is that if one does have a destination certain, then the choice of roads is important, and one had best spend some time setting it up.

Gaining commitment at top
Senior Management Sponsorship is crucial in these projects, and the level of support is commensurate with the scope of the project. If the project is Departmental in scope, then the senior Departmental leadership must be on board. Obviously support even higher will only help, but be careful of bypassing chain-of-command positions which could turn a potentially ally into a snubbed detractor.

Forming a user/IT partnership
In today’s dynamic environment in which data warehousing solutions are becoming key ingredients to business success, users are learning nearly as much about their data requirements as OLTP users. Therefore, it is critical to form a team where both technical and end-user personnel can work together to develop a mutually acceptable and technologically achievable set of specifications and requirements. This should result in continuous collaboration throughout the project as many warehouse projects can be viewed as a discovery process where the business perspective must be weighed alongside the technical issues.

This is a good area to use RAD (Rapid Application Development) techniques to allow technicians to demonstrate to end-users the potential capabilities of the system and to allow users a tangible feedback mechanism.

Determining Metrics for Success of Project
How will the organization know that the project was successful? If there is an intention of using the original project as a proving ground for an enterprise-wide rollout, how can there be a rollout if there is no yardstick for measuring the results of the pilot? As part of the planning process it is critical that metrics be defined which clearly demonstrate that the results of the project meet and are aligned with the original business goals which drove the project to begin with. It is also imperative that the metrics be objective rather than subjective.

The project plan
This step is anathema to many technologists, but a project plan detailing the objectives, approach, strategy, ownership, timeframe, and resources’ responsibilities is a must if the project is to be managed with any degree of professionalism and efficiency.

Identifying the areas of expertise required
One of the outputs of a good plan is identification of the resources required and an analysis of whether they exist in-house and whether or not they are available. Make sure they are represented on the team. Having senior management support is a good prerequisite for being able to get the people you need, and if they are not available, for being able to get permission ahead of time to go outside (hire service providers) if necessary.

Developing a communications plan
A Warehouse is only useful to the organization if users exploit its abilities. It is foolhardy to spend effort on developing such a project only to spring a surprise onto a department or enterprise which is not ready to take advan-
itage of it because they didn’t know it was coming, or when it would be ready, etc. A communications plan is essential to disseminate information about the project.

**Using benchmarking techniques**

If possible, investigate other organizations that have successfully completed projects similar to the one at hand. This can take the form of literary research or actual trips to view operational systems, interview users, developers and management as to the best practices they have encountered in their project. Remember that not all practices that work in one environment are universally portable, so make sure that the context in which a particular practice is said to work can be extended to the target environment.

**Selecting a methodology**

A methodology that is known and accepted by the organization will go a long way to smoothing out many of the rough spots which any project will hit. A methodology provides three components to a project:

- A logical series of activities to achieve a desired end. This structure will tell the organization what has to be done and when in order to produce a quality product.

- A definition of deliverables associated with each activity and progress reports on the business benefits from each activity. This tells the organization what the output is of each step.

- Roles and responsibilities for actors in the activities.

The Methodology will also help define a project structure and what has to be done to manage the project: when reviews are to be held and what each review will cover. One trap to avoid here is the one of putting someone in charge who is too narrow in his or her scope. For example, a technologist who has no appreciation for or understanding of the business problem, or a functional specialist who has no patience for technological issues. There has to be a blending of business and technology in order for these projects to succeed, and therefore the lead people have
to be sensitive to the various cultural differences of the team members. As the Data Warehousing Institute said in a recent publication: “Data Warehousing is a service business, not a storage business.”

**Using service providers**

Any organization will have areas where there is either a lack of expertise or where there are insufficient skilled resources available for the project. The project manager should look for areas of expertise which are not represented on the team. In many IT shops, the most significant missing element where consulting services might be required is the planning elements. On the business side, organizations will need to adapt to the new opportunities the technology provides. For example, reporting and analysis, prior to Warehouses, was very “flat” file oriented—run a report and look at it on paper—run another report and get it on paper six weeks later. With Warehouse tools such as OLAP, reports become threedimensional and you pivot and drill through them searching for information. Designing these reports with little or no prior experience is very challenging in that you must think differently. Users will need support in learning, specifying and evolving the output of data warehouses.

**Architecting the solution**

Architecture is defined as the definition of the components of a solution and their interaction (see Fig. 8). This guide has delved into several areas where the benefits of a well-defined and integrated architecture have been demonstrated. Defining the components of the solution and how they interact is a critical step in implementing a successful project. The solution should be an end-to-end solution and allow for all the characteristics described earlier, including scalability, extensibility and manageability.

The choice available to IT range from tightly coupled and integrated product “suites” provided by some vendors (e.g., IBM Visual Warehouse”) or individual/groups of building blocks of Warehouse products that IT would take the primarily role in integrating to meet their needs.

Defining a directory such as IBM Visual Warehouse Information Catalog into the architecture will allow the system to draw business meta data from dozens of other products, including DB2, Oracle, Sybase, Hyperion Essbase, CASE tools, and more. Tools such as the Information Catalog can help provide users with the keystone component of consistent, synchronized meta data that helps developers, maintainers and users alike.

![Fig. 8](attachment:chart.png)
Designing the system
Designs should be based on open, well-understood architectures with well-defined interfaces between the components. Use of standards will help the scalability and flexibility of the system over time. Navigational tools for users and user interfaces in general should be designed with the business problem in mind in a cooperative process with end users and managers.

Departmental systems should be designed within the context of an enterprise framework. That is, where possible and practical, identify data elements and concepts that span multiple departments and try to define them in the broadest possible terms so as to be able to include other departments later on.

Developing the system
The development environment should be capable of using RAD techniques to help end users who might not be able to grasp Warehousing and analysis concepts immediately. Use iterative prototyping with time boxing and multiple releases where necessary to gain consensus from the user community.

One of the challenges associated with building a system is the ongoing integration of meta data. The value of integrated meta data is that it reduces the time to implement a Warehouse, as well as providing greater maintenance and management efficiencies. When meta data is updated manually this process can introduce increased error rates for end users (e.g. defining tables incorrectly). One solution is put forward by the Meta Data Council: a corollary to the OMG (Open Management Group) subcommittee for the standardization of Common Warehouse Meta Data and the related OMG subcommittee to standardize XMI. It is a federated approach whereby all stores of meta data type information (data dictionaries, passive repositories, encyclopedias, data base catalogs, etc.) pass through a “metahub” that changes the syntax and other relevant information about the data. This allows any tool that adopts this approach to interoperate and interchange meta data within and around the Warehouse on an ongoing basis. This approach is not yet widely available but holds great promise for resolving one of the major issues in meta management.

Piloting the implementation
Pilot systems, also known as prototype systems or proof of concept systems are among the most misunderstood concepts in the industry. There are three possible reasons for an organization to embark upon a pilot program:

• The technology is foreign to the organization, and there is a need to understand the benefits of the technology to see how it might help in the business problems at hand.

• The technology is understood, but finding an appropriate application is not. The organization wants to know if a particular application of this technology to a business problem is appropriate, which is to say, will this technology solve the problem?

• The technology is understood as well as the application, but the cost/benefit equation is not understood. The organization wants to know if the cost of the solution is worth it.

Each of these motivations will result in different pilots, in different places in the organization and with different associated metrics.

Regardless of the motivation, however, a primary question that must be asked is: How will we know if the pilot answered our questions? The only way is to develop quantitative as well as qualitative metrics and in addition develop tools and techniques for capturing and analyzing the results at the end of the pilot program.

Deploying the system
If the strategy calls for replicating successful projects in other departments, then a roll out plan must be developed to identify the order of the rollout as well as any integration efforts that must be dealt with. These might include process reengineering, especially if multiple departments are today involved in a process that is not automated, and one of these departments will receive the automation prior to the others.

Training users
One sure way to snatch defeat from the jaws of victory is to develop a technically outstanding Warehouse and then let users loose on it with no training. That which is intuitive to a technologist steeped in Warehousing concepts is gibberish to an end user whose focus is running a business. Training programs must be developed which target not the technological niceties behind the screens, but rather that address the how’s of using the system from a business viewpoint. Focus on training the user on understanding the meta data and the navigation capabilities as well as how to use those in analyzing a business problem.

Managing warehouse operations
Developing a Warehouse is one thing, keeping it operational is another. It requires managing the timeliness of file transfer and load processes as the system grows in size and complexity. Daily operations create meta data changes, such as the addition of a user, or loading of a new star schema, which means the meta data repository must be managed. In addition, the organization will need feedback reports on the Warehouse operations. Managers will want to know that the quarterly data extraction actually was started on schedule and completed on time with all appropriate indexes regenerated.

Some users, in spite of all best training efforts will construct queries that are capable of bringing a system to its knees. In these situations, organizations will need database management tools that report on how much system resources are being used by a query and allow the interruption and subsequent termination of the query.

Some reporting tools are also able to identify which areas of the database are being hit hardest, or most frequently, or even identify which times of the month/quarter/year those areas are most likely to be accessed. This will allow managers to govern the extraction and indexing processes—change table structures, drop columns, define different aggregations—according to anticipated uses.

New demands by business units enabled by technology or new business theory will always tax a core business system such as a Warehouse, but good architecture will deliver flexibility downstream to deal with these “opportunities” such as today’s web enabled marts or integrated business process marts.
**Summary**

Data Warehousing can provide significant new benefits to an organization. The technology is moving from an early adopter stage to a maturity phase. This has removed much of the technology risk associated with early adopters stage but risk associated with implementation and strategy are still very much factors. By considering the issues described in this guide you should be able to reduce risk associated with implementing data warehouse solutions and deliver higher value solutions in a reduced timeframe.

**Appendix A—Elements of a warehouse**

Fig. 9 represents the basic elements of a Data Warehouse architecture that responds to the needs of enterprises today. This model is slightly different from most of the ones being promulgated throughout the industry in that it combines the technological elements with the process and planning elements mentioned earlier. The human figures in the diagram represent those elements that are primarily of a service nature, such as planning, analysis, processes and management.

In this section we will briefly touch on the individual elements and their contribution to the overall architecture.

**Planning/analysis**

Arguably the most important element of the Data Warehouse architecture, the planning and analysis constituent is the one in which the enterprise determines which problems it needs to solve, what the characteristics of the problem are and how best to solve them. Once the problem is clearly framed, then the team (a joint IT/Business Unit) can move on to identifying and resolving other dependent questions such as:

- What existing business processes are involved in the defined problem?
- How large a warehouse do we need to satisfy our business problem?
- What operational data systems (existing operational data sources) do we need to access to get the data we need?
- What transformations will need to be built to convert the operational data into usable warehoused data?
- What are the characteristics of the Data Model we need to build to answer the anticipated queries?

**Data acquisition**

Data acquisition includes extraction of data from operational systems, cleansing the data (restructuring records, removal of operational data, translating field values to a common data dictionary) and checking data integrity and consistency. It involves transformation (adding time fields, summarization, and derived fields), loading of the clean data into the Warehouse database, updating the Warehouse indexes, etc. It is the component of the architecture that provides the raw data into the warehouse repository that will then be used for analysis. Clearly, it will be important to understand how the data will be used so that transformations, cleanings, etc. render a suitable format with which to accomplish our query goals.

**Data storage**

The Data Storage element represents the database and accompanying structures that are used in the analytical process. Relational and multidimensional databases are primarily used in architectures today, and while there is a fair amount of debate in the industry as to which provides the best results, relational databases offer more flexibility, smaller overall size, and easier access to atomic data for drill down operations. A trend emerging within the industry is for database vendors to increasingly add multi-dimensional functions into relational products.

**Meta data**

Just as any other kind of warehouse needs to keep an inventory of its holdings, a Data Warehouse needs to keep track of what data it is currently holding along with the pedigree of that data. This is the role of the meta data repository, to give users and technicians easy-to-understand information about the data such as where the data came from, which rules were used in creating the data, what the data elements mean, etc. Many systems divide the Business or End-User Directory, so that technical information about the data is kept in a separate directory from that which is required by the user to understand the data from a business perspective. (For example...
IBM Visual Warehouse Manager deals with technical meta data while their Visual Warehouse Information Catalog manages the business meta data.

**Data access**

Here, the rubber meets the road in that the tools presented here to the end user are the ones that will ultimately drive their perception of the utility of the system. Access can be categorized using one of the following descriptors: Standard Query, Data Interpretation, Multidimensional Analysis, Data Mining, and Enterprise Reporting.

Standard query tools allow users to develop an hypothesis and create questions (queries) to test its validity. This is sometimes called a verification-driven approach. Examples include business statistics and optimization (linear programming). Multi-dimensional analysis tools facilitate flexible investigation of the data along various dimensions, applying operations such as time series analysis, and enabling interactive drill-down capabilities. Data mining tools use a discovery-driven approach where sophisticated data mining algorithms are automatically applied to detect trends, patterns, and correlation hidden in the data. Enterprise Reporting is information distribution for the masses, usually via a Web browser that provides collection and distribution of data to large numbers of people throughout the enterprise, sometimes referred to as publish and subscribe.

**Data delivery**

This element refers to how the resultant data is presented to the end user, and has two dimensions, the physical and logical conduit by which it reaches the end user, and the mechanisms by which the data may be visualized. Geographic dispersion, security and data volumes will dictate the use of various conduits such as Local Area Networks, Internet, Wide Area Networks (public and private), etc. Query complexity and output volume will dictate the rendering of the data. Options range from simple tables such as spreadsheets to simple two-dimensional graphics such as bar graphs and pie charts to very sophisticated visualization technologies that utilize three-dimensional landscapes to portray the results.

**Management**

Management of the Warehouse represents an indispensable architectural element. Once the Warehouse is set up, refreshes of the data must be scheduled and efficiently and correctly executed. Backups of the Warehouse must be made at reasonable and predetermined frequencies. Users must be added and deleted. Security must be adhered to. The list goes on. The larger the warehouse, the more difficult the task becomes, and therefore the more important that this be done well. A recent article in Computerworld stressed the fact that some Warehouses have reached a size such that data refreshes cannot be done without taking the system down during operational hours, because of 18 and 20 hour upload processes.

**Methodology**

Methodologies are key ingredients for any complex project. A good methodology will define:

- A logical stepwise approach to solving a problem where each step builds on the previous one.
- A complete set of deliverables and measurements against goals so that nothing is forgotten in the project.
- A project structure including roles and responsibilities of each participant.

Use of a good methodology will greatly increase the chances of success of any project, but in Data Warehousing it is essential due to the integrative (and therefore complex) nature of the problem.

**Project management and administration**

Although this technically falls under the purview of a good methodology, it is important to emphasize the need for solid project management with both IT and Business units involved in developing and setting up the Warehouse. Assembling the team, setting and managing expectations and working an effective communications plan are all a part of a good project management team.

---

**Appendix B—Case studies**

IBM’s principal solution for generating and managing Data Warehouse and Data Mart systems is Visual Warehouse. Other IBM offerings such as the Data Replication family and DB2 DataJoiner (for multi-vendor database access) can complement Visual Warehouse as data is moved from source to target systems. IBM also partners with companies such as Evolutionary Technologies International for more complex extract capabilities and Valley Technology Inc. for data cleansing technology. In addition, IBM has key partnering arrangements with Brio Technology, Business Objects, and Cognos for query and reporting as well as with Hyperion for OLAP technology.

For the warehouse database, IBM offers industry-leading DB2. The DB2 family spans Nettinsky systems, AS/400 systems, RISC System/6000 hardware, IBM mainframes, non-IBM machines from Hewlett-Packard and Sun Microsystems, and operating systems such as OS/2, Windows (9x & NT), UNIX, OS/400, and OS/390. When DB2 DataJoiner is used in conjunction with Visual Warehouse, non-IBM databases, such as those from Oracle, Sybase, and Informix serve as the warehouse database.

More case studies and information is available on IBM’s Web site at www.ibm.com/software/bi.

**Visual Warehouse helps McDonald’s Canada get to the meat of its marketing data**

Behind every mouthful of a beefy Big Mac is a huge organization, working long hours to make sure customers get the quality and service that has become McDonald’s hallmark. In Canada, more than 1,000 McDonald’s restaurants do brisk business. Nevertheless, growing competitive pressures and customers’ demand for new values have prompted McDonald’s Canada to aggressively expand its market presence with a larger number of strategically located restaurants. At the same time, the company constantly strives to curtail its cost of operations. Oswald Edwards, manager for strategic & architectural planning at McDonald’s Canada, explains,
“Extensive discussions with our business executives revealed that they needed detailed, accurate, and timely information for strategic planning and decision making.”

Today, a new DB2-based data warehouse, created and run by IBM Visual Warehouse, is providing key transaction information to market analysts at McDonald’s Canada. It includes information that is helping them answer questions such as what combination of products sells the most at a given time of day, which day of the week a new campaign should be launched, the success of promotional campaigns linked with other leading brands, and much more. Says Edwards, “We were able to convince our business users that IBM’s data warehouse solution was the one for us,” adding, “We’ve achieved enormous returns from our investment in Visual Warehouse. It will give us access to information that will help us to substantially increase restaurants sales and reduce operating expenses.”

Simplifying development, driving down maintenance costs

The data warehouse, in DB2 for AIX, resides on a four-node RS/6000 SP server. Driving this is Visual Warehouse. Running on a Windows NT server, Visual Warehouse captures transactional data that is scattered on a number of different enterprise business systems, and loads the data warehouse. Says Serge Tremblay, database administrator at McDonald’s: “The beauty of Visual Warehouse is that it automates data collection from all our operational systems, refines it and feeds it to the data warehouse. There is no comparable product in the market that combines all these functions in one box.”

The company knew that maintenance—requiring extensive database administration resources—is one of the most challenging and expensive aspects of data warehousing. And that was another reason it chose Visual Warehouse. “Visual Warehouse takes over the headache of administration and maintenance,” says Edwards, referring to the software’s automation of monitoring, performance logging, and other administrative tasks. “It wins in terms of cost/performance, because you need to dedicate fewer personnel resources for warehouse management.”

Summary data from the warehouse is extracted onto four department-specific data marts, which reside in DB2 for AIX on different nodes of the RS/6000 SP. Users access the data marts using Cognos PowerPlay and Impromptu clients. “Visual Warehouse supports seamless integration with a wide range of sophisticated decision support tools. All the data we have would be meaningless if not for Visual Warehouse’s openness in allowing us to use other tools,” says Tremblay.

Mining for hidden value in new found data

As the data warehouse continues to evolve, by 1998 it will cater to the purchase department’s needs, with the restaurant, and operations and franchise departments joining in over the next two years. Information on inventory turnaround, geographical distribution of sales, and restaurant profitability, that will be stored in the data warehouse, will help users from these departments in managing supply chains and planning new locations. Seeing how existing business users are benefiting, Edwards adds, “We understand the tremendous advantage in mining the newly found data to discover hidden correlation and will be reviewing IBM Intelligent Miner.”

The scalable trio

During the initial planning and testing, data from a combination of 10 restaurants was stored in the data warehouse. By year-end, data from 100 restaurants will be available, and eventually McDonald’s aims to get all 1,000 restaurants in the loop. In the meantime, McDonald’s will move its data warehouse to DB2 Universal Database: Extended Enterprise Edition, as well as migrate to Version 3.1 of Visual Warehouse. Says Edwards, “We believe that the scalable, parallel processing capabilities of DB2 Universal Database will help us make the best use of our RS/6000. By bringing the two together with Visual Warehouse, we’ll have the most scalable and effective system.”

Testing Visual Warehouse Version 3.1 at McDonald’s, Tremblay is excited about the new functionality it provides. “Visual Warehouse Version 3.1 now includes new and improved features that make it an even more powerful data warehousing tool,” he says. “For example, in the new version it is easier to make changes to source data and have these changes flow through the business views.” Tremblay explains that since data warehousing is by nature an iterative process, changes are always being made to database objects. This version of Visual Warehouse makes it much easier to change objects. “Now, changes can be made on the fly. This means the database administrator doesn’t constantly have to worry about the status of database data,” Tremblay notes. Listing some of the other features, Tremblay adds, “Multiple operations can be performed simultaneously, rather than sequentially. Indexes are generated automatically, and Visual Warehouse can perform star joins and even display them graphically.”

IBM Inspires Confidence

Even more important than the products themselves was the selection of the vendor. Tremblay recalls, “There was a slew of database companies on the market. But we decided on IBM because their technical and support staff were superior to anyone else out there.” Adds Edwards, “Without the commitment shown by McDonald’s senior management—people like George Mencke, our chief financial officer, and our IT staff—as well as IBM’s willingness to partner with us in building our warehouse, our project would not have been as successful as it is.”

DAMAN converts information into knowledge with IBM’s suite of business intelligence products and ETI•EXTRACT

As business becomes increasingly information-driven, more and more corporations are moving towards new information system architectures such as data warehouses and enterprise-wide applications. One of the toughest and most frequently under-estimated challenges in implementing these new applications is the timely and accurate conversion and migration of data between legacy systems and the new ones that will replace or augment them. DAMAN Consulting, an IBM Business Partner, is making that challenge much less painful by integrating IBM’s business intelligence suite of products into its data migration and conversion solutions.

DAMAN Consulting, a systems integrator based in Austin, Texas, specializes in providing data migration, data warehousing, and decision support solutions developed around IBM business intelligence products. “Our approach to data migration and warehousing combines IBM business intelligence products with our in-house methodologies,” explains Gita Lal, executive vice president for DAMAN.
IBM and ETI add value to DAMAN's solutions

In the data management consulting industry, DAMAN Consulting has earned a reputation for quick delivery of solutions based on IBM's business intelligence products. Alfredo R. Ramirez, Jr., president of DAMAN Consulting attributes much of this, not only to DAMAN's knowledge and ability to quickly and accurately perform data migration implementation, but also to IBM's comprehensive package of business intelligence products and integration with complementary products such as ETI•EXTRACT. "IBM's support for products such as ETI•EXTRACT gives us a considerable advantage, because we don't have to go out and test the integration components or develop integration processes among the different tools. They're already integrated into the solution," explains Ramirez. "IBM's package includes reporting tools, meta data tools, tools for propagating and migrating the data from the source application to the warehouse, and also replication technologies." He adds, "Having an integrated suite of tools that supports all of our reporting needs definitely increases our ability to sell such solutions to our customers."

Tightly integrated products mean rapid, low-cost implementation

For DAMAN's customers, quick implementation and conversion from legacy systems to their new data warehouses is a major concern. "Using IBM's business intelligence products in conjunction with propagation tools such as ETI•EXTRACT enhances our ability to deliver interface and programmatic conversion solutions under tight deadlines," says Ramirez. "ETI•EXTRACT's tight integration with DB2 and Visual Warehouse hastens the migration effort and provides the customer an environment that can support their reporting needs very quickly."

One of DAMAN's most recent implementations was for a large property and casualty insurance provider, where the business objectives were to migrate a legacy policy management system to a newer system running on an IBM mainframe using DB2 for MVS. DAMAN also developed several programmatic data interfaces from the new policy management system on DB2 to hundreds of disparate sub-systems that support numerous decision support functions. At another site, DAMAN designed a datamart on DB2 for OS/390 to maintain an insurance fraud decision support system. ETI•EXTRACT was used for the data extraction and load processes, and Visual Warehouse was used to design and build the datamart.

"The cost of implementation and deployment is significantly reduced by using the Visual Warehouse prepackaged solution, largely because implementation time is reduced, as is the level of skill required for implementation," says Lal. "The combination of Visual Warehouse, ETI•EXTRACT, and DB2 delivers a synergy that provides the basis for an advanced decision support environment that is easy to learn, use, and maintain."

Data migration’s triple threat: IBM, ETI, and DAMAN

In addition to providing comprehensive IBM data migration and warehousing solutions, DAMAN has developed tools that assist in meta data management for these solutions, including DAMAN's InfoManager. Meta data is information about the enterprise data and is a critical element in effective data management. DAMAN uses Visual Warehouse's Information Catalog to index the meta data in the application. InfoManager then gathers information from the Information Catalog and other sources of meta data (including disparate applications and data sources external to the organization), and creates an integrated meta data model that can be used for performing impact analyses on the entire computing or decision support environment. The accuracy and currency of the meta data model is further enhanced by InfoManager's Intelligent Agent technology (supporting management of operational activities and change notification).

"A big advantage of using IBM's business intelligence solutions is the integration of the meta data arena. Most products are sold independent of each other and then it's left to the clients to create their own integrated model. With IBM's business intelligence suite of products, there's DataPropagator and ETI•EXTRACT sharing meta data with the Information Catalog, which allows users to automate meta data integration. This provides huge benefits in terms of the ongoing maintenance of the data warehouse," Ramirez explains. "That's a key component that makes IBM's business intelligence suite of products particularly appealing."

Lal concludes, "In addition to being an end-to-end provider of business intelligence solutions comprised of world class technology, IBM adds value to its partnerships, and that's something that can't be said about the competition. If you make partnering difficult and you don't add value to the partnering process, you'll lose your partners very quickly. With IBM, partnering is a desirable thing."

Blue Cross & Blue Shield prescribe formulas for real-time cost analysis with DB2 OLAP Server and Visual Warehouse

Founded in 1939, Blue Cross & Blue Shield of Rhode Island (BCBS) provides health coverage for one out of every two Rhode Islanders. This translates into well over 450,000 insurance policies. Imagine having to track and analyze every administrative cost associated with selling, implementing, and processing each one of these claims. Quite a tedious task. Yet, the BCBS cost accounting department does it every day. Ensuring that all costs are properly allocated across the organization means monitoring business results and interacting with data on a day-to-day, hour-to-hour, and even minute-to-minute basis.

Keeping pace with these demands and the hundreds of thousands of policies was nearly impossible in the past, when analysts had to manually sift through mountains of printed reports and key in the required data into spreadsheets. Today, BCBS has empowered its business analysts with online analytical processing (OLAP) tools to enable sophisticated, multi-dimensional analysis of large volumes of data. Having such capabilities is imperative to BCBS as it allows them to transform data into useful
information for analysis and decision making, and to address the corporate need to understand what is driving business performance.

George Trudel, an internal business and technology consultant at BCBS, agrees: “When you take a look at the demands coming out of organizations for real time information to meet the challenges of competition in today’s electronic environment, decision support without OLAP is a very unwieldy task.”

**Letting analysts get down to business**

Since 1993, BCBS, Rhode Island’s largest health insurance provider, has been using Hyperion Software’s Essbase OLAP server on Windows NT in its cost accounting department. This department plays a critical role in coding, tracking, reviewing, and analyzing all administrative expenses for its 75 diverse lines of health insurance products. Prior to implementing Hyperion Essbase, BCBS cost accountants had to thumb through hundreds of pages of reports to dig up the specific information needed to prepare an expense report and re-enter the data into Excel spreadsheets. Once Essbase was deployed, cost accountants were able to acquire the information they needed almost instantaneously, leaving them more time for their main task—cost analysis.

Recently, BCBS took its OLAP capabilities to a new level by giving its OLAP users direct access to the company’s relational data stores with DB2 OLAP Server. DB2 OLAP Server integrates the Hyperion Essbase OLAP engine and application program interfaces (APIs) with DB2 Universal Database. “With IBM DB2 OLAP Server leveraging the Essbase OLAP engine, we will not only take a great burden off of our IT resources,” says Trudel, “but also give the business user access to vast amounts of summary information that would have normally taken a very long time to produce.”

**Essbase and DB2 working side-by-side**

BCBS uses IBM Visual Warehouse to build and maintain portions of the DB2 data warehouse on its S/390 server. “We have a large volume of data that needs to interact with other SQL-based decision support tools, so we want to maintain those in DB2,” Trudel explains. DB2 OLAP Server enables the Essbase OLAP engine to operate on top of this relational store.

Trudel emphasizes that accessing the relational warehouse has not come at the expense of its Essbase multidimensional stores. DB2 OLAP Server and the native Essbase system work side-by-side at BCBS, fulfilling important and complementary roles. “With DB2 OLAP Server, we’ll be able to do everything that we’ve done with Hyperion Essbase, and in addition, have the ability to easily access data from our DB2 data sources,” says Trudel.

**Managed OLAP**

Blue Cross & Blue Shield uses both DB2 OLAP Server and Essbase as data marts. “Combining Visual Warehouse with DB2 OLAP Server and Essbase provides us with a managed OLAP environment, in which we can automate the process of extracting data from different sources, loading it into the data warehouse, and performing ongoing maintenance of the warehouse,” says Trudel. “This reduces the need for our IT staff to constantly manage the whole system.”

**The bottom line**

Microsoft Excel is the primary desktop tool for BCBS’s business analysts. Predefined Excel templates created internally are used to retrieve the data from Hyperion Essbase or DB2 OLAP Server and automatically load it into the appropriate cost accounting reports. According to Trudel, what used to take several days to create now takes three and a half minutes. “With Hyperion Essbase and DB2 OLAP Server, we’re able to get a much better feel for how costs are being allocated across the organization, the effect of the allocations on our final cost objectives, and the impact of administrative expenses on the business,” says Trudel. “Since our cost accountants now have more time to produce analytical reports and build scenarios, they can give us a better handle on controlling costs and identifying issues in that process.”

**Essbase engine enables excellence**

It’s no secret how Trudel feels about the Hyperion Essbase engine. “I think it’s an excellent tool. It gives business users a technology that enables them to get to the information they need when they need it,” he explains. “And from the IBM side, we have massive amounts of information in IMS and DB2 databases that we need to get at for multi-dimensional analysis. Being able to get at that information with the OLAP tools enables us to hypothesize about other potential revenue streams that we couldn’t address before.”

**Glossary of terms**

**Ad hoc query**—Any query that cannot be determined prior to the moment the query is issued.

**Business directory**—The portion of the Meta data Repository which deals with information about data pertinent to the business users. For example, meta data in the Business Directory could tell users what factors are included in the calculations of gross vs. net sales.

**Centralized data warehouse**—A Data Warehouse implementation in which a single warehouse serves the needs of the multiple business units simultaneously with a single data model which spans the needs of the multiple business divisions.

**Change data capture**—The process of capturing changes made to a production data source. Change data capture is typically performed by reading the source DBMS log. It consolidates units of work, ensures data is synchronized with the original source, and reduces data volume in a data warehousing environment.

**Copy management**—The process of taking a snapshot of data from a source and copying it to a target environment.

**Crosstab**—A process or function that combines and/or summarizes data from one or more sources into a two dimensional format for analysis or reporting.

**Data administration (DA)**—The processes and procedures by which the integrity and currency of the data in the warehouse are maintained.

**Data aggregation**—A type of data derivation where a data value is derived from the aggregation of different data occurrences of the same subject data. For example, yearly sales data can be aggregated from monthly sales data.

**Data cleansing**—The process of removing errors and resolving inconsistencies in source data before loading the data into a target environment.
Data dictionary—It is a collection of definitions and specifications for data categories and their relationships. It is a database of data about data (meta data).

Data extract—The process of copying a subset of data from a source to a target environment.

Data mart—A type of data warehouse designed to meet the needs of a specific group of users such as a single department or part of an organization. Typically a data mart focuses on a single subject area such as sales data. Data marts may or may not be designed to fit into a broader enterprise data warehouse design.

Data mining—A process of analyzing large amounts of data to identify hidden relationships, patterns, and relationships. This is often called “discovery-driven” data analysis.

Data model—A logical map that represents the inherent properties of the data independent of software, hardware or machine performance considerations. The model shows data elements grouped into records, as well as the associations between those records.

Data propagation/replication—A process for distributing data from a source database to target databases while usually keeping the databases synchronized.

Data scrubbing/transformation—The process of filtering, merging, decoding, and translating source data to create validated data for the data warehouse. For example, a numeric regional code might be replaced with the name of the region.

Data warehouse—a subject oriented, integrated, time-variant, non-volatile collection of data in support of management’s decision making process. A repository of consistent historical data that can be easily accessed and manipulated for decision support.

Database—A collection of data which are logically related.

Database management systems (DBMS)—A software system for creating, maintaining and protecting databases.

Decision support system (DSS)—Systems which allow decision makers in organizations to access data relevant to the decisions they are required to make.

Derived data—Warehouse data that results from calculations or processing applied to the source data before it is stored in the Warehouse environment. For example, source data might be used to calculate ROI (Return On Investment) which is stored as derived data in the Warehouse.

Drill down—A method of exploring detailed data that was used in creating a summary level of data. Drill down levels depend on the granularity of the data in the data warehouse.

Enhanced data—Warehouse data that has been cleansed, scrubbed, transformed, derived, summarized, or aggregated.

Enterprise data model—A blueprint for all of the data used by all departments in the enterprise. An Enterprise Data Model has resolved all of the potential inconsistencies and parochial interpretations of the data used and presents a consistent and commonly understood and accepted view and definition of the enterprise data.

Enterprise data warehouse—An Enterprise Data Warehouse is a Centralized Warehouse which services the entire enterprise. Enterprise Data Warehouse are sometimes used to populate data marts.
Executive information system (EIS)—Tools programmed to provide canned reports or briefing books to top-level executives. They offer strong reporting and drill-down capabilities. Today these tools allow ad-hoc querying against a multi-dimensional view of data, and most offer analytical applications along functional lines such as sales or financial analysis.

Extract—The process of copying a subset of data from a source to a target environment.

Information directory—A system for browsing Data Warehouse meta data. Sub-components usually include a business directory and a technical directory. See the entries for Business Directory and Technical Directory.

Information mining—The process of extracting previously unknown, comprehensible, and actionable information from any source—including transactions, documents, e-mail, web pages, etc., and using the information to make crucial business decisions.

Information Warehouse™—IBM’s approach to developing an architecture for a data warehouse that supports the implementation of either functional, centralized, or decentralized warehouses.

Informational applications—Applications which are written to analyze data from a Data Warehouse for Decision Support purposes.

Informational data—Data which has been extracted, summarized, and stored in a Data Warehouse for purposes of supporting Informational Applications.

Meta data—Data about data. For example, information about where the data is stored, who is responsible for maintaining the data, how often the data is refreshed, etc.

Middleware—A communications layer that allows applications to interact across hardware and network environments.

Multi-dimensional analysis (MDA)—Informational Analysis on data which takes into account many different relationships, each of which represents a dimension. For example, a person doing an analysis of retail may want to understand the relationships among sales by region, by quarter, by demographic distribution (income, education level, gender, or by product). Multi-Dimensional Analysis will yield results for these complex relationships. Multi-Dimensional Analysis is sometimes referred to as On-Line Analytical Processing or OLAP.

On-line analytical processing (OLAP)—Processing that supports the analysis of business trends and projections. It is also known as Multi-Dimensional Analysis.

Operational applications—Applications which support the daily operations of the enterprise. Usually included in this class of applications are Order Entry, Accounts Payable, Accounts Receivable, etc.

Query—A request for information from the Data Warehouse posed by the user or tool operated by the user.

Relational database management system (RDBMS)—A database system built around the relational model based on tables, columns and views.

Replication—The process of keeping a copy of data.

Source database—The database from which data will be extracted or copied into the Data warehouse.

Star schema—A modeling scheme that has a single object in the middle connected to a number of objects around it radially - hence the name star. A fact such as sales, compensation, payment, or invoices is qualified by one or more dimensions such as by month, by product, by geographical region. The fact is represented by a fact table and the dimensions are represented by dimension tables.

Target database—The database in which data will be loaded or inserted.

Technical directory—The portion of the Meta data Repository which deals with the technical information about the data. Such information may include the field designation (alphanumeric, etc.), the number of characters, range checks, etc.
For more information about IBM Visual Warehouse please contact your IBM marketing representative or IBM authorized software reseller, or visit our Web site at www.ibm.com/software/vw or www.ibm.com/software/data