Data Appliances
–
A Smart Alternative to Traditional BI Technologies

February 2003

Prepared by

Dr. Claudia Imhoff
Intelligent Solutions, Inc.

Table of Contents

Introduction

For many organizations, the experience of building a first data warehouse is similar to a child receiving a puppy. The anticipation and expectation are high, the rewards are many, and costs are worth it "out of the box". What's more, given the proper care, feeding, and training, the initial cost / benefit ratio is sustainable over time.

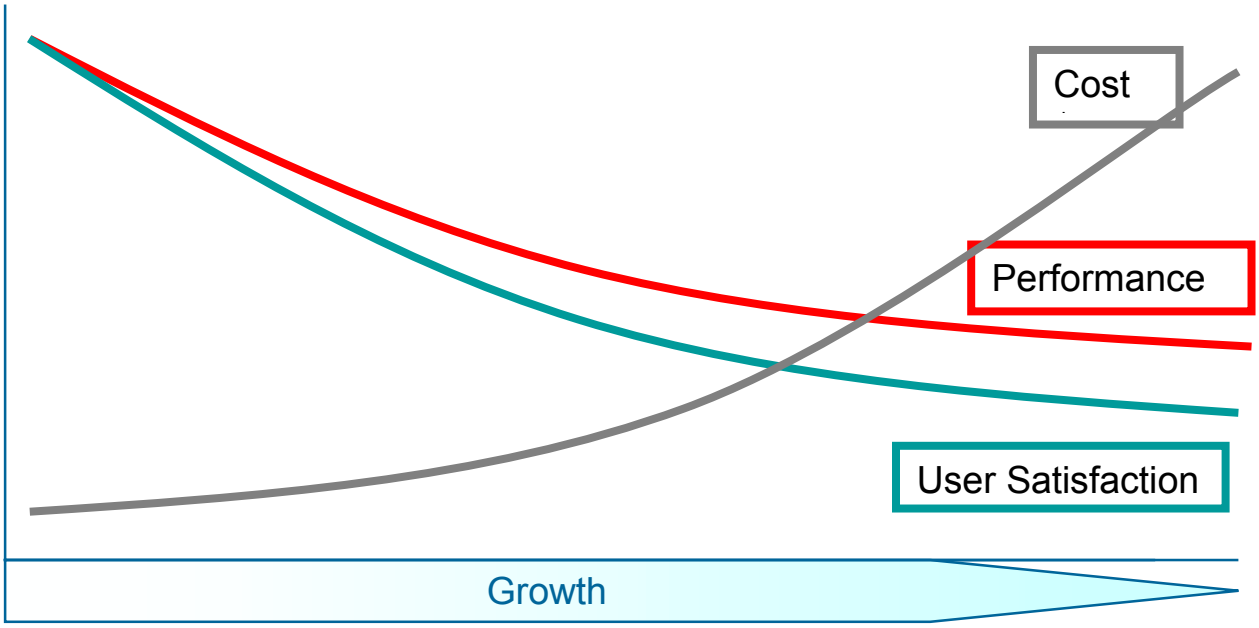
Most pet owners will tell you, though, that the care, feeding, and training process requires constant monitoring and effort, especially during the growth phase. So will most data warehouse owners.

As your implementation of the corporate information factory (CIF) grows, you will find a number of common "behaviors" hold true. First, like puppies, your first data marts will attract the attention of the others around you. If your data mart puppy is effective and cute, others will compliment you on it and want one, too. Before you know it, your implementation will include additional subject areas and data marts, and your data warehouse will grow into its enterprise data model "feet".

Furthermore, enterprises need to analyze more and more data to even greater depths. That is, as more and more of the business community come to depend on their analyses, the data warehouse will truly become the support mechanism for these growing marts. Its role becomes critical to the health and well-being of the entire Business Intelligence (BI) environment. Multi-terabyte sized data warehouses are becoming commonplace.

The growth is inevitable and welcomed in this environment. The downside to this progress is that, as your data warehouse and data mart puppy increases in size and usage, its reach will increase to the kitchen counter – to elsewhere in the organization. At this point, performance starts to decrease and the cost of the entire environment becomes an eye-opener. If you don't swat it on the nose with a rolled-up newspaper, ultimately, user satisfaction and overall support for the environment will begin to suffer (See **Figure 1**).

Figure 1: Trend in a Growing BI Environment



So – how do you improve or maintain the performance this critical component of your enterprise's strategic decision-making capability? There are four possibilities:

- Limit or remove certain functionalities
- Limit or remove the depth or breadth of the information available for analysis
- Increase the hardware/software horsepower
- Adopt a new paradigm for the technological solution – use a data appliance

The first two alternatives in today's enterprises are nonsensical in today's competitive world – restricting the business community's ability to perform certain analyses (e.g., data mining or exploration) that require massive amounts of data is simply not acceptable. Business problems today have big data requirements that cannot simply be ignored. Examples of problems requiring massive amounts of data are in **Table 1**.

Table 1: Type of Analysis versus Data Needed

Type of Analysis	Data Needed
Predicting Markets	Point of Sales, Market, and Customer
Business Process Optimization	Event and Process Monitoring, Geographic and Spatial
Risk Management	Sub-transactional and/or POS
Customer Behavior	Demographic and POS
Text Processing and Mining	Sales, Service, POS
Advanced Analytics and KPIs	Customer, POS, Financial

Competitive companies have no choice today; they must be able to analyze detailed data efficiently and effectively with reasonable response times. The response time is mandatory since the next query will be based on the results of the first query. Stale results caused by non-responsive systems are useless, or worse, misleading.

So that leaves the last two alternatives – loading up on traditional technologies or moving to a completely different model for BI technology.

This paper begins with a discussion of the traditional hardware and software approach to these problems, that is, adding more and more hardware and software to improve performance. Then we define what a data appliance is, the new paradigm it represents, and its characteristics and specifications. We follow with a discussion of the benefits and challenges to this new approach. Finally, we include a section on getting started in which we discuss the parts of the Corporate Information Factory architecture that the data appliance supports, and a good starting point for its installation in your BI environment.

The Traditional Approach to a Growing BI Environment

In the current economic environment, companies are examining very closely where they spend every dollar. They are wringing every bit of performance they can from their existing analytic platforms, which tend to be generic hardware and software components supporting standard relational databases and storage devices. These components come from multiple vendors who must support multiple platforms themselves. While making these technologies more ubiquitous and universal in their usage, the manufacturers must also optimize them to the least common denominator across the disparate technologies. In many cases, this means that the technology is sub-optimized across the board for all other vendors' technologies.

Traditional Architectures

Many companies have implemented their BI environments using general purpose hardware (e.g., IBM, Sun, HP, etc.) and DBMS software (Oracle, SQL Server, DB2, etc.). These combinations can be further clustered to improve performance, using various forms of configurations for high-end computing (e.g., Symmetric multiprocessing, massively parallel processing, and non-uniform memory access).

Solutions to Performance Problems

The solution to increasingly poor performance and unhappy users in the traditional approach is to throw more machine power at the problem. This increased machine power usually takes the form of:

- Purchasing more CPUs and memory
- Purchasing bigger, faster disks and controllers
- Purchasing upgrades to servers and new DBMS versions
- Replicating or redistributing databases across multiple platforms

Each and all of these steps result in dollars being spent on vendors' products. But it does not stop there. With each new purchase, there are the maintenance costs and associated administrative expenditures that ensure the ongoing performance of the new components. Fortunately, most enterprises standardize on one or possibly two of the platform combinations mentioned in the last section. This standardization can help to mitigate overall upgrade and resource costs but, even with standardization, these costs can become quite staggering.

There is another significant source of cost. The administration expenditures may not be obvious at first. These include not only the set up costs of the new technologies but also the continuing effort to fine-tune each component in the mix to obtain its optimal performance given specific circumstances. Multiple DBAs may be needed in environments that have tens of terabytes of data just adding to the overall cost. For query performance in these environments, data must be indexed intelligently, I/O is still at the record level and, even with sophisticated and knowledgeable DBA "tweaking", performance is only incrementally improved.

> Doug Laney, Vice President at Meta Group, states: "Throwing incremental hardware and DBMS software at analytic performance problems may be not more than an expensive stopgap solution." He continues that "incremental or add-on solutions to address business performance frequently cannot provide the ROI that alternate/specialized approaches do." [1]

What is a Data Appliance?

To explain what a data appliance is, let's work at general definition of appliances and then briefly look at the history of the data appliance. The American Heritage Dictionary defines an appliance as "A device or instrument designed to perform a specific function, especially an electrical device, such as a toaster." The inner workings of the appliance are irrelevant to the ultimate user. In the case of the toaster, the user simply wants the bread toasted to the preferred a specific color, and really does not care how the toaster performs that feat. The bottom line is that an appliance is designed to do one thing and to do it very well.

Another example of an appliance is a stereo amplifier. Most of us have no idea how it boosts performance, no clue as to the inner workings inside the metal box. More importantly, we do not care. An amplifier is a black box into which we plug other stereo components (DVD, CD and video players, and so on). We do not configure it. We do not have to tweak it once everything is plugged in to get good performance. And if we want to upgrade the amplifier, it is a simple process of buying another one and dropping it into the slot where the old one resided. The ultimate in Plug and Play appliances!

Therefore, the characteristics of a good appliance are that it:

- Is transparent to the user
- Yields an obvious performance boost
- Is easy to install and administer
- Has very low maintenance costs

Such a device comes about as a result of the maturation of the technology. The amplifier of today is the result of years of research and development and was built with a specific purpose in mind. This "purpose built" specification means that it has been thoroughly optimized to perform a well-defined and documented purpose. You would not think about using it to do anything else but amplify your other stereo components.

This appliance was also made possible by the creation of many electronic industry standards. Things plug into it and it can be easily replaced because of these standards. By taking advantage of mature industry standards, a purpose-built appliance such as the amplifier can easily integrate into anyone's home entertainment center.

Finally an appliance such as the amplifier is much cheaper to purchase than if we had to buy the individual pieces and build them ourselves. The synergy of controlling all the "stuff" in the black box and configuring it for user has greatly reduced the overall cost of the appliance.

---

[1] "Accelerating Analytics: Alternatives Abound", November 6, 2002, Meta Group Inc.

A New Paradigm for BI Environments – the Data Appliance

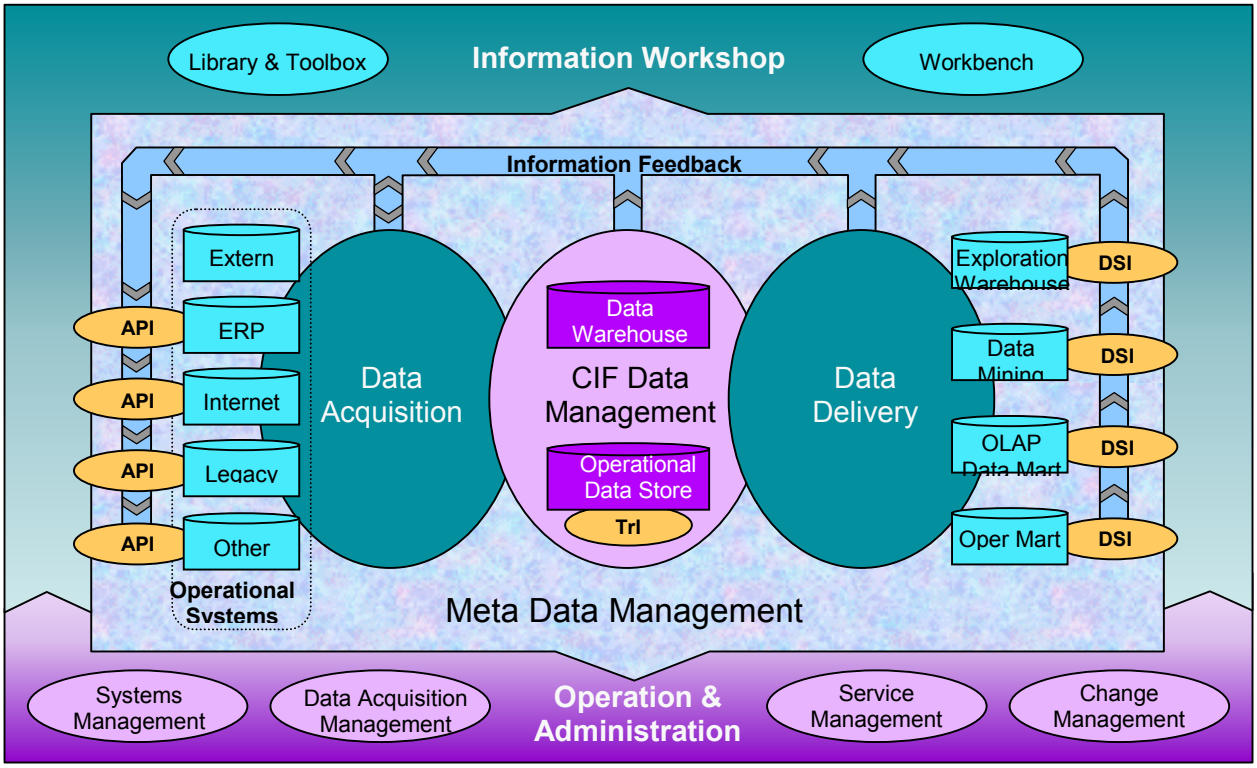
Appliances in the computer world have been around for many years. In fact, they are so ubiquitous that we often don't realize we are using them. An appliance in our world is defined as an integrated box that can retrieve information at the request of external applications. Similar to the amplifier, its inner workings are hidden to maintain simplicity and ease of use. Examples of purpose built technology appliances are the network router, hubs and switches. These devices require minimal configuration, are basically plug and play, greatly increase the performance of network transport but are relatively transparent to the user.

For years, we have attempted to apply these same notions presented by the technology appliances to databases. The result was the advent of the database machine. Researchers found that they could develop a technology that combined all the components – the hardware, DBMS software, and other components – to perform a specific set of functions or applications. Several attempts have been made over the years at creating such a general-purpose database machine but none have really caught the attention of most IT departments.

Perhaps the problem is that these database machines were too general, too unfocused, and therefore did not offer any real advantages over the traditional hardware and software configurations they competed with. Or it could be that the database machine was not created by any single vendor; it was a collaboration of several vendors rather than a single vendor creating the appliance. This is similar to buying your amplifier and finding out that much of the interior was outsourced to various other companies. Many of the benefits are missed because these third party components may not be completely integrated with the hosting company's components. In any case, the database machine has yet to find its traction in the IT world.

Now let's turn our attention to the BI industry. We are fortunate today to have more than a decade of building BI environments behind us. Just as in the electronics industry, we, too, have developed standards for use in creating BI systems. First, we have a proven and stable architectural roadmap – the Corporate Information Factory (CIF). See **Figure 2**.

Figure 2: The Corporate Information Factory



The CIF has been well documented and has been implemented in a great number of Global 2000 companies. There are several books and a plethora of articles describing it in detail; therefore, this paper will not spend more time describing it than is necessary to determine where the data appliance fits into it.

In addition to this standard architecture, we also have our own mature industry standards such as ODBC, JDBC, XML, and SQL. These standards allow use to leverage their usefulness to create a new paradigm specifically created for the technological infrastructure of the BI environments. The confluence of these two factors, a solid architecture and established industry standards, has made it possible for a viable data appliance to be built for the BI environment.

Definition of a Data Appliance for BI

The data appliance for BI is a purpose-built database machine that is specifically used to manage analytical data and to retrieve the results from massive data analyses with impressive performance – a matter of seconds or minutes instead of hours or even days. The data appliance for BI is a combination of hardware, software, DBMS and storage, all under one umbrella – a black box that yields high performance in both speed and storage – making the BI environment simpler and more useful to the users.

To be even more effective, this black box should use off-the-shelf components to speed up critical BI functions such as data loads and query performance. The off-the-shelf nature of the components also has the added advantage of reducing the overall costs of the appliance.

Well thought-out data appliances use the best aspects from both an SMP and MPP configurations, incorporating them into a new technical architecture that processes queries in the most optimal fashion possible. Most of the bottlenecks found in the traditional architecture of separate hardware and DBMS software components can be either eliminated or significantly mitigated in this new model.

The result is a streamlined environment that is efficient, inexpensive, and simple to use and maintain. With such an environment, we can now revisit the needs of the business community. As stated earlier, there is great demand for in-depth, complicated queries requiring massive amounts of data. These types of queries yield far more insight and true business intelligence than the simpler "slice and dice" queries supported by multidimensional databases and designs. With a data appliance created specifically to support very large data warehouses, very complex queries, and sophisticated analysts, we can now satisfy these critical business needs easily.

Characteristics of a company that needs a data appliance

There are several obvious characteristics of companies that should consider using a data appliance in their BI architectures. These include:

- Companies that have very large and fast-growing amounts of data, for example, hundreds of gigabytes to tens of terabytes in total.
- Business users who want to do in-depth analysis on their full data sets, not just sampling.
- Business users that require real-time analyses, not environments that have large latency periods between the execution of a query and the returning results set.

What to Look For in a Data Appliance

Let's reexamine the four features for any appliance. These were that the appliance:

- Is transparent to the user
- Yields an obvious performance boost
- Is easy to install and administer
- Has very low maintenance costs

Does the data appliance concept support these as well? And what are the benefits of each of these features? Let's see.

Transparency to the User

Because the data appliance is, in essence, a black box and the plug and play nature that results from this configuration, the user need not worry about constantly optimizing, tweaking, fiddling with the separate pieces. The appliance comes specifically tuned for the BI environment. This aspect alone serves to significantly reduce the overall complexity of BI data management, especially in large enterprise environments.

Performance Boost

The increase in sophisticated consumers of analytical data and the demand for far more complex queries has put unprecedented stress on the technical infrastructure for BI. A data appliance for BI is designed to deliver business intelligence quickly and efficiently, especially for very large data warehouse environments. Benchmarks comparing traditional infrastructures against data appliances such as Teradata's and Netezza's, have demonstrated the considerable improvement in performance from the data appliance results. Complex queries are run on the order of magnitudes faster using the data appliance environment (e.g., from 10 to 20 times increase in performance!).

A second performance boost for the data appliance comes from its ability to scale linearly. Said another way, the addition of extra storage should not adversely affect the performance of the complex queries. A well-built data appliance can scale to support multi-terabyte data warehouses with minimal performance degradation.

Easy to Install and Administer

Because a data appliance for BI integrates the critical components of a BI infrastructure – hardware, DBMS software and storage – into a single unit, it greatly simplifies the challenges most companies face in managing disparate technologies. Similar to a generic appliance, the data appliance for BI was designed to do one thing and to do it well. It has been optimized for that purpose and is highly efficient. There is nothing to "manage".

This has a number of benefits to it. First it provides a level of reliability inherent in the appliance construction. By removing the inner workings from being adjusted by administrators, the potential for de-optimizing the system or causing failures is decreased. The administrator does not have to integrate disk arrays, operating systems, database software and the hardware from multiple vendors with the hope that they will work perfectly.

Second, it increases the "Plug and Play" nature of the overall environment. The data appliance is compatible with all common open-source database and operating system standards making it easy to integrate into any existing BI environment. This also means that there is support for the transparent access for heterogeneous BI tools and applications, which serves to reduce ongoing support issues. The features and functions of the data appliance are accessible to all other tools and applications within the BI system. No one part is "captive" to any particular BI tool or application. This effectively increases the value of the data appliance because it can be used where it is more needed at all times.

Low Maintenance Costs

One of the metrics used by IT departments everywhere is the gigabytes managed per person – of the efficiency ration of data administrators. With the amount of data in data warehouses easily doubling every year, an IT department either has to double their staff or double their productivity every year – neither approach is welcomed. Therefore, a significant reduction in the total cost of ownership comes from the simplicity of a data appliance for BI, not just the cost of the products. This simplicity translates into a reduced requirement for systems and database administrators.

In addition, because the data appliance is purpose-built, it uses off-the-shelf or commodity components, thus reducing the overall costs of ownership. Just like the amplifier example, the data appliance takes the advantage of the technological advances of recent years to create a cost efficient solution to what has traditionally been a very expensive and complex environment.

Challenges

Now that we have covered the features and benefits of a data appliance, a look at the challenges to this new technology is in order. These can be classified as:

- Resistance to the new paradigm
- Limited offerings from vendors
- Point of failure
- The double-edged sword of standards

Resistance to the New Paradigm

Whenever a new paradigm is introduced into the marketplace, there will be resistance from the entrenched paradigm. Certainly this will be no different for the data appliance for BI model. DBAs and administrators who are used to fiddling and fine-tuning their multi-vendor environments may be uncomfortable with a black box-type technology. The analogy here is the Apple operating system. When the Apple computer first came out with its proprietary OS, many hard core programmers were put off because Apple did not give them the same control over the technical parameters of the OS. These programmers were uncomfortable with the “leap of faith” Apple required from them and resisted using the Apple for that reason.

The same may be said of the data appliance. Administrators of BI environments who are used to constantly tuning and tweaking the technology will need to make a similar leap of faith. The belief that this black box is fully optimized without human intervention requires a new way of thinking from the administrators.

Limited Offerings from Vendors

There are few vendors offering such data appliances for BI. The best known today are Teradata and Netezza. Certainly the strides made by these two companies in furthering the industry's understanding of data appliances has been considerable in the last few years. Indeed, it should be understandable that competition is welcomed and that it leads to lower prices, accelerated technology advances and a greater diversity of approach. The fact that Teradata now has a viable competitor in Netezza has already caused a significant shift in price points and increased technological innovation.

Point of Failure

As in any technological architecture, the points of failure must be examined and mitigated as much as possible. The data appliance's unique integration of hardware, DBMS software and storage shares the same problems that other architectures do in that it has points of failure in a BI environment. This can be overcome by mirroring or clustering the technology and the usage of an SMP and MPP architecture. In addition, the vendors supplying data appliances have employed best practices and have leveraged the decades of industry experience in building reliability into the hardware and software.

The Double-edged Sword of Standards

The use of ANSI SQL standards has the benefits of making the data appliance fully compatible with existing BI tools and applications, as mentioned in the last section. However, it also means that access may be slower than the native formats for each interface. In choosing a data appliance, you should look for ones that user a driver such as an ODBC or JDBC driver that is fast enough to make the limiting factor only the network connection or the speed of the client machine.

Getting Started

Companies that require a BI environment that yields high performance, both in speed and storage of strategic data, are likely candidates for a data appliance for BI. Today, companies and their analysts are moving beyond the relatively simple multidimensional and OLAP forms of analyses. These were good for beginning BI implementations. They gave corporations a good handle on the possibilities of BI but have been found to be limiting in their ability to deliver real in-depth, sophisticated analytics.

The new and far more insightful analyses require massive volumes of detailed data and take on the characteristics of an ad hoc environment. Waiting for hours or even days for the results of these complicated and complex queries is unacceptable; business analysts must have the results from one query quickly so that they can then change the direction of the next query based on the results in a real-time mode.

Data Appliances and the Corporate Information Factory Architecture

So in a mature BI environment, where does the data appliance for BI fit in? In the Corporate Information Factory (CIF), the data appliance has two obvious "homes". One is as the infrastructure for the data warehouse itself. The proven ability of the data appliance to handle large volumes of data with relative simplicity and low cost makes it a very appealing alternative to the traditional hardware and relational DBMS combination we see today.

A second home within the CIF is for a new component called the exploration warehouse. The exploration warehouse was introduced in 1998 as a new component to the Corporate Information Factory. Its purpose is to provide a unique capability that is generally very difficult to create – a safe haven for the exploration of data and ad hoc analytical processing. Although the exploration warehouse is a valid and necessary part of the CIF, in practice a difficulty arose due to the limited technology available to support this capability.

As a final section in this paper, we discuss the utilization of a data appliance for BI as the infrastructure supporting an exploration warehouse.

Exploration Warehouse

To facilitate the unstructured and ad hoc nature of these queries, the exploration warehouse must be based on technology that can deal with the unknown in a BI environment. It must be able to handle the unstructured processing of the exploring analyst with suitable performance. In general, the nature of the exploration warehouse is temporary and transitory. It is rare that the exploration warehouse remains a permanent structure and it is constantly being constructed and reconstructed. The question then is what technology can be used for this unusual but extremely valuable component of the Corporate Information Factory?

To accommodate the "ad hoc" nature of the exploration warehouse, a technology must support an easy and speedy development. Second, the technology must be able to handle large amounts of data with reasonable response times. This dichotomy has been a problem in the past with traditional relational databases. Tearing down and rebuilding these data structures is no easy feat. Therefore, new and different structures in support of exploration are needed. The data appliance suits this environment nicely because:

- It can handle large volumes of data quickly,
- It can accept design changes handily,
- It provides superb response times to complex and massive queries, and
- It is reasonably priced

The exploration warehouse has now become such an important part of the overall Corporate Information Factory architecture for unstructured, ad hoc queries that many corporations have teams of implementers specializing in its construction and maintenance. The data appliance can be used to more easily and more cost efficiently created this invaluable BI environment.

Once established, the data appliance is then in a perfect position to become the infrastructure for the data warehouse itself. It is not a far cry to move from exploration to being the stable repository for all BI data. The simplicity of the appliance and the low total costs associated with it make it an ideal infrastructure for the warehouse. Then various data marts can be built from the data warehouse in rapid order and with minimal administrative overhead.

Summary

Data appliances are becoming a viable option as a part of a coherent enterprise BI program. Data appliances are popular due to their effective means of delivering critical intelligence along with the key advantages of simple installation and troubleshooting ease. The benefits of this simplicity, lowered costs, and performance boosts plainly cannot be ignored.

About the author:

Claudia Imhoff, Ph.D., *is President and Founder of Intelligent Solutions, Inc., a leading provider of business intelligence and customer-focused technology and strategy. Dr. Imhoff is a popular speaker and internationally recognized expert on customer relationship management and the infrastructure required to support this initiative – the Corporate Information Factory. Dr. Imhoff has co-authored four books on these subjects and writes monthly columns for technical and business magazines. She can be reached at CImhoff@IntelSols.com.*

