

The Data Warehouse Appliance: The Evolution of the Data Warehouse Market

By John Ladley



Simply Complex

Although shelves-worth of books have been written about data warehousing, a data warehouse is quite simple in concept: extract and integrate data from throughout (and maybe outside) the organization, and load it into a database. This concept came about when some bright people realized that the price/performance metrics of hardware and storage had fallen to a point where loading (and essentially duplicating) what was then perceived as absurd amounts of data was economically feasible.

This is a key concept – nearly everyone who had come in contact with computers or written a program prior to 1988 felt it would be really cool to load as much data as possible into a database to do analysis and reports. The economics of hardware and software prevented this until the late 1980's.

To the Hammer, (even the retooled one), Everything Looks Like a Nail

As long as the data warehouse concept has been in play, IT organizations and consulting firms have been designing and developing data warehouses using technologies with roots in operational, transaction-oriented processing, not analytic processing. This sub-optimal technology stack has not only resulted in awkwardly extended infrastructure components and architectures, but has also burdened users with artificially limiting the business value potential of their “ideal” analytic system. Granted, all of the large ‘merchant’ DBMS vendors have released their data warehouse “versions” to some extent or another. And there have been targeted, dedicated DBMSs only for data warehouse.

Any generation technology can be pushed and extended for a period of time to accommodate business problems and architectural concepts. But at some point, new infrastructure components, (or new configurations of existing granular infrastructure components) emerge as an alternative. These solutions typically are driven not only by traditional technical limitations, but more so by achievable quantum leaps in business value.

No Fool with a New Tool

Recent developments in hardware technology strangely echo the changes in the late 1980's that made data warehousing possible. The dawn of the so-called “data warehouse appliance” leverages a new wave of hardware price/performance changes. Data warehouse appliances are specialized database/hardware hybrids (pre-integrated hardware and software) that can act as the primary DBMS for a data warehouse, or act as a back-end supplement to an existing data warehouse. They tend to be built from commoditized, off-the-shelf components that have proven reliability, stable suppliers, and throughput and capacities that would have made a 1990 data warehouse developer green with envy. Current vendors of this technology are Datallegro and Netezza. One

might liken this to the price/performance leaps made by Apple in the 1970s to integrate and consolidate functional chips on their flagship PC motherboard.

Any time an IT sub sector experiences a sharper growth in services than in technology components themselves, it's high-time for a set of specialized technologies to emerge that tempers the consulting/staffing machinations typically needed to integrate components or work-around technology limitations. For example, in the late 1980's as organizations came to realize the insurmountable programmatic challenges of integrating their accounting, manufacturing, HR, customer support and other business functions, along came pre-integrated ERP solutions. Similarly CRM solutions emerged over the past ten years to integrate and coordinate contact management, call center, sales, marketing and other customer-oriented functions. This is now the case in the data warehouse arena – data warehouse consultants are relied upon more than the DBMS engines to increase performance and throughput.

Data Warehouse Dollars and Sense

While the concept of data warehouse is simple, the reality has become much more complicated. This complication is expressed in terms of the billions of dollars per year spent on services, hardware and software. Data warehousing is its own industry. The development and maintenance costs and impact on the organizations doing the data warehouse projects can be substantial.

The data warehouse at its core is more about performance than functionality. Software-only solutions such as extending the DBMS with analytic functions and multidimensional schema support, or integrating ETL (extract, transform, load) functions into the DBMS engine all fail to recognize the critical importance of hardware in the data warehouse technology stack.

Organizations supporting or developing a data warehouse need to consider the lessons learned of the last 20 or so years. The main emphasis of this white paper is that, like the changes that made data warehousing feasible decades ago, new developments in technology permit consideration of data warehouse appliances as feasible applications of technology. By examining the price/performance arena, it will become obvious that there is more here than just another way to gather data. In addition, there are contemporary challenges in the data warehouse arena that are addressed poorly by so-called "merchant" DBMS'. Finally, the market that will be the acquirer of the data warehouse appliance technology is turning out to be open minded to the solution, which may encourage others to consider data warehouse appliances if they are appropriate.

The context of this paper is for large sets of data – i.e. over 2 TB and beyond, and whether or not the data warehouse appliance is an appropriate alternative for consideration. Most commercial database technology can now be tweaked to deal with up to a terabyte. However, there seems to be no limit to the amount of storage data warehouse projects now require.

Prized Price-Performance

When it comes to data warehouse and business intelligence (BI), performance of the environment is a key concern. Research indicates, in many cases, it is one of the highest concerns. As database sizes grow larger, the labor expended as a proportion of total effort for data warehouse success tends to migrate from functionality to performance. For the database and servers, implementation speed, cost of ownership and functionality are the top considerations. Of course, the data warehouse appliance is a server and DBMS in one package.

Handling raw data is only one way to look at the efficacy of the data warehouse appliance. Even if the data warehouse appliance architecture meets the top technical concerns of data warehouse managers, there has to be consideration of cost of acquisition, development and on-going costs of ownership.

Development Cost

Typically, when an environment for a new large data warehouse is developed, the merchant or specialized DBMS has to be “tweaked.” Rarely is there a project where the vendor does not have to supply gurus to contribute experience and insight to coax the DBMS to higher performance levels.

At times this is akin to building a custom car versus one off the show room floor. Without the addition of the special talent, you get average or mundane performance. And the talent is not cheap.

A typical scenario is where a cluster of servers is configured into an SMP or MPP cluster, and the parallel features of a merchant DBMS are “turned on.” The historical result has been adequate (not sterling) throughput, after a period of time. In many cases, however, the period of time taxes the patience of the data warehouse team and the entire user community. If there is an excessive amount of effort required to get the database ‘right’ to extract proper performance, the amount of effort expended is, at times, far in excess of original expectations for tuning and development. This expenditure often counterweights any benefits expected from the data warehouse. None of the mainstream data warehouse DBMS vendors have been able to provide “world-class” teams to all of their data warehouse clients satisfactorily.

It is even being suggested by some technical types that a data warehouse appliance can be developed in-house, given the tendency to use commoditized parts and DBMS kernels. Obviously the users of mainstream DBMS are not going to develop a data warehouse DBMS on their own. Certain shops may be tempted. They should be made aware that there are many backplane and disk communications issues that are being addressed by the data warehouse appliance class of vendor and in no way does the typical DBA department have the horsepower to tackle that programming task.

However, this idea brings to light an interesting thought process. When using a

merchant DBMS many data warehouse managers are in reality making a build vs. buy decision. Does the manager stick with the merchant DBMS, which offers perhaps lower political risk, but then face extraordinary effort in building the right indices, schemas, tuning queries and ETL processes, in addition to the usual items, e.g. business needs, culture change?

This 'build' scenario has resulted in cost overruns at numerous data warehouse projects. The 'buy' scenario (acquiring a data warehouse appliance) becomes relevant when the manager compares the costs of tuning the merchant DBMS and assigning numerous DBAs.

Maintenance Cost

After the environment is up and running, there is the on going need to maintain indices, adjust, refine, and monitor data warehouse content and usage. At one time, the industry rule of thumb was one DBA for every 500GB of logical data. The merchant DBMS vendors have improved this, and most likely today the ratio is 1 per TB. However, this does not count maintaining downstream data marts, Universes, materialized views, etc.

As volume grows, there is also the matter of adding capacity. Granted "disk is cheap," but it seems it can only be acquired in \$250,000 increments. This is not cheap.

The on going management of ETL and data quality is also a cost issue. Nearly all ETL processes require "babysitting" to some extent. Historically, data warehouse teams are lax at data controls and spending required time to correctly engineer ETL jobs. In addition, DBMS idiosyncrasies require a lot of tuning of ETL processes.

On going management of the query tools occupies considerable effort after data warehouse development – mostly sustaining performance and managing users

Contemporary Challenge: Splitting the Atom

Enterprises that were once content with capturing so-called atomic data about major business entities (e.g. customers, suppliers, partners, employees), are now requiring the capture, storage and analysis of minute business activities between discernable, traditional business events. They also need to capture and analyze data about individual transactions. Being able to capture and analyze this level of granular, "sub transactional" data (e.g., network monitoring, click stream data) can help organizations *affect* imminent transactions, not just record and analyze them *ex post facto*. Traditional infrastructure components make achieving sub transactional data collection and analytics nearly impossible.

The competitive drive that spawns requirements for capturing and leveraging sub transactional data, introduces a quantum leap in feeds (data) and requisite speeds (processing) that taxes traditional data management infrastructure com-

ponents beyond their limits. In effect, a data warehouse manager should assume that for every transaction generated, there are 10 sub transactional events that happen 10 times as rapidly. The ability to sense and respond to sub transactional customer or /prospect behaviors, or manufacturing/distribution activities, or even external economic factors, offer enterprises significant competitive advantage. A traditional data warehouse infrastructure and platform can probably be made to deal with this opportunity. However, the data warehouse appliance offers a legitimate alternative.

Key Buyer Trends

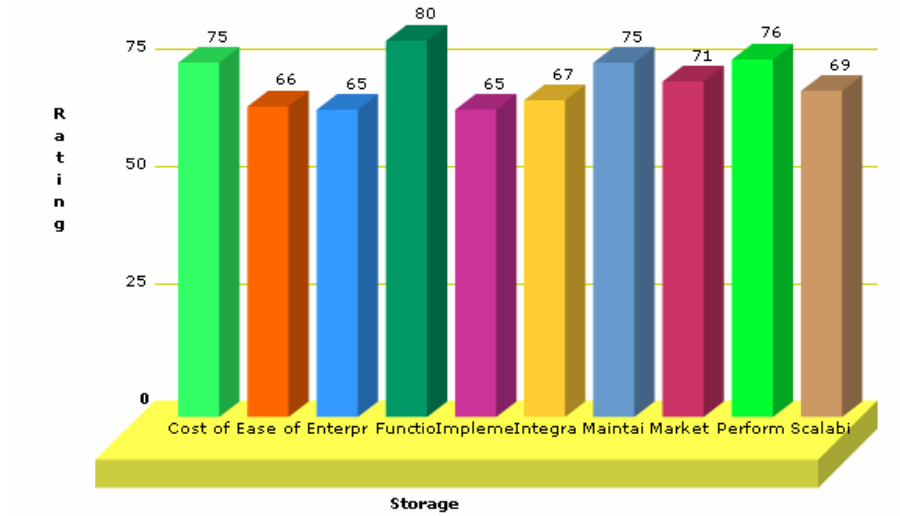
According to June 2005 results from Evalubase Research's ongoing study of the enterprise technology market, buyers currently rank performance as the most important selection criteria when considering DBMSs (see chart below). Any solution, therefore, that provides augmented DBMS performance, either through intrinsic DBMS advances or more so via specialized DBMS-hardware integration, would be particularly attractive to most buyers. With cost-of-ownership, functionality and implementation speed also being of high importance according to the study, an appliance solution that reduces overall operational and setup costs (not just technology acquisition costs), while maintaining expected DBMS levels of functionality should also greatly appeal to most organizations.

Notice also how much buying decisions discount whether the DBMS is an enterprise standard or what the vendors' market share is. This means that those selecting DBMSs care little about popularity contests or whether their ultimate vendor is a household name.

Similarly with storage components (see chart below), organizations participating in the Evalubase study rate functionality, performance and cost of ownership as being the most important buying considerations. Although buyers of storage are more concerned with market share than the buyers of DBMSs, they remain least concerned about them being an existing standard within their enterprise.

When selecting servers however, organizations are more concerned with ease-of-use, than with performance and scalability. The introduction of scalability concerns into this data warehouse-stack equation suggests that current servers are running out of headroom to crunch data as fast as organizations require. If this is true in general, it is critical for data warehouse solutions. Organizations will apparently overlook existing enterprise standards and vendor market share to accommodate these other more important characteristics.

Evalubase's findings should pave the way for any IT organization to select combined DBMS-storage-server solutions such as analytic appliances that buck enterprise standards and marketplace ubiquity in favor of enabling required performance, functionality and reduced cost-of-ownership.



Source: Evalubase Research

Summary

The data warehouse appliance is a relevant and legitimate technology option for a data warehouse platform add-on or as part of the initial infrastructure. Granted, there may be a struggle with “standards”. However, the advantages in cost of ownership, the ability to tackle problems and growth curves that the main stream merchant products have problems with far outweigh the disadvantages. Finally, history points to the natural evolution of this type of technology. While the technology is nascent, it is by no means experimental.

Many thanks to Doug Laney and Evalubase Research for the data and assistance with this article.

John Ladley is an internationally known practitioner and a popular speaker on information and knowledge management. He is a Director with Navigant Consulting, which acquired KI Solutions in 2005. John is widely published and has several regular columns. Prior to founding KI Solutions, John was Senior Program Director of Data Warehouse Strategies and a Research Fellow at Meta Group. Mr. Ladley is an authority on information architectures, business performance measurement architectures, knowledge management, collaborative applications, and information resource management. He can be reached at jladley@navigantconsulting.com.

