TANEJA GROUP
TECHNOLOGY ANALYSTS

MARKET PROFILE

Barbarians at the Gate – Data Warehouse Appliances Challenge Existing Storage Paradigm

May 2007

THE NET NET

Despite all the marketing talk about “intelligence” in the storage network, we still have a ways to go as an industry. The truth is that most storage today is simply not as aware of applications, data access patterns, and workflows, as they should be. While established vendors have built general purpose block-based storage capable of running a wide spectrum of workloads, these systems are not optimized for any particular workload and possess no intelligence about the application, its data formats, and access patterns. On the other end of the spectrum, especially over the past 5 years, we have witnessed an increasing drive toward more specialized storage appliances and devices. These specialized storage appliances and systems combine application intelligence or workload optimizations with core data storage capabilities in order to deliver a tailored turnkey solution for a particular application or business need.

While Network Attached Storage (NAS) is probably the oldest example of specialized storage appliances replacing general purpose computers, more recently, Content Addressable Storage (CAS) has evolved into a new specialized class of storage focused on the requirements of archival and compliance data. Further, with the growth in High Performance Computing (HPC) applications, Data Direct Networks (FC and IB-based block storage) and Isilon (clustered NAS) among others have delivered data storage solutions optimized to a specific I/O profile, such as large block, sequential I/O, and carved out lucrative market niches in the industry. At present, we are also witnessing this trend manifest itself as companies, like Teneros, pioneer storage appliances tailored to delivering email storage continuous available.

Five Observations Changing Database Storage Architectures

This trend toward specialized storage architectures and devices is operating in the database world too. In fact, several key drivers are transforming how large scale databases (greater than 1 TB) are stored, managed, and scaled. Ultimately, five drivers are leading to the emergence of a new class of database storage optimized around data warehousing and business intelligence workload. The implications for storage managers and administrators will be significant:

1. **Users are Facing a Tsunami of Structured Data.** From our research, end users tell us that their databases, particularly data warehouses, are doubling in size every year. The primary driver for this growth in the database size is coming from the line of business. Business decision makers recognize the value for maintaining more historical data online longer for better analytics and

decision making purposes. A secondary driver fueling the size of databases is due to a tightening regulatory and compliance environment from such legislation as Sarbanes-Oxley (SOX), Healthcare Insurance Portability and Accountability Act (HIPAA), Gramm-Leach Bliley Act (GLBA), and Payment Card Industry (PCI). The drive to keep more data online longer in databases exacerbates issues of database performance, scalability, and management and makes the current general purpose storage approaches less attractive.

2. **The Need for Speed.** The need for more database performance is insatiable. DBA and storage administrators are being asked to greatly expand and manage much larger database and storage environments, while improve data loading times, query responses, and deliver deeper data analytic. Unfortunately, the overall performance and response time of current RDBMS systems is impacted as the database size increases. This fact is particularly true as the database grow in excess of 1TB. Techniques like database archiving allow IT to prune the size of a database to improve performance, but don't necessarily allow that data to be kept online and fully queryable. IT faces huge challenges in coaxing significant I/O throughput and response times of the underlying storage system in order to meet the insatiable requirements of a large data warehouse implementation. Clearly, the overall throughput and response time of the underlying storage infrastructure directly affects what end user see in terms of response time.

3. **Current Database Scalability Approaches Have Significant Drawbacks.** Three architectural approaches to scaling database performance have emerged: buy a larger Symmetric Multi-Processor (SMP) server to run the database, implement a clustered, shared disk database architecture like Oracle Real Application Clusters (RAC), or deploy a Massively Parallel Processing (MPP) architecture like Teradata. SMP systems (multiple CPUs or cores tightly interconnected in a single server) are by far the most common deployment model for OLTP databases and small data warehouses or data marts, but a big iron SMP server can cost in excess of $1M and cannot be modularly scaled on demand. Clustered databases offer the promise of near linear scalability, but require laborious partitioning to reduce synchronization overhead and achieve optimum performance for data intensive workloads. MPP systems that partition data and parallelize queries have emerged as the de facto approach for large scale data warehouses. However, traditional MPP systems require constant tuning and repartitioning and as a result, on going OPEX cost can run into the tens of millions of dollar for a large scale data warehouse. There is no silver bullet approach that offers low acquisition cost, scalability, and ease of management.

4. **OPEX Costs Mount for Tuning & Managing Big Databases.** As the

database size grows, the administrative overhead of managing a database grows exponentially along two dimensions – database management and tuning and storage management and tuning. The type of tuning and management required to maintain a large scale database requires some of the most highly skilled professionals. As can be imagined, as the database grows in scale, the amount a business must spend in terms of OPEX to maintain and grow it, increases dramatically. The OPEX cost of administering a large scale database does not grow linearly or in proportion to the database size, instead the OPEX costs scale exponentially as the size of the database grows. OPEX costs can be the number one inhibitor to further growing a very large database.

5. **Database & Storage Are Becoming Increasingly Intertwined.** Increasingly the storage administrator must have a working knowledge of the database architecture, table layout, how data is placed on disk by the database in order to deliver the desired performance SLA. As a result, we have seen database vendors like Oracle increasingly incorporate core storage features like automatic volume management into their database kernels as a way to more tightly couple storage with the databases engine. A data warehouse appliance takes this convergence to the ultimate endpoint – collapsing database intelligence and moving it closer to the physical storage in order to minimize network roundtrips and gain performance. This convergence of storage design and host level software is not unprecedented. File systems have evolved to the point where they are now considered extensions of the underlying storage infrastructure. Furthermore, NAS appliances ultimately subsumed file systems as a key component of a NAS system. It is natural for databases and storage to become more tightly coupled as the need for optimum performance grows.

Enter the Data Warehouse Appliance

Given this environment, Taneja Group has begun to track how this historical trend toward specialized storage appliances is being applied to structured data. In fact, we have witnessed a new emerging category of data warehouse appliances come to market over the past three years and gain traction at the high end of the data warehousing market. Although the term "data warehouse appliance" is recognized in DBA circles, the term has almost no meaning or mindshare within the storage community. The reality is that data warehouse appliances have far reaching implications to how structured data will be managed and how access to that data will be scaled in the future. Ultimately, we see data warehouse appliances morphing into a new class of storage in much the same fashion NAS and CAS have become synonymous with new types of storage today.

The origins of the term "data warehouse appliance" can be traced back to 2002 or 2003 when Foster Hinshaw, the founder of Netezza and now founder and CEO of Dataupia, coined the term. Essentially, data warehouse appliances are defined to be a turnkey, fully integrated stack of CPU, memory, storage, Operating System (OS), and Relational Database Management System (RDBMS) software that is purpose built and optimized for data warehousing and business intelligence workloads. It utilizes massive parallelism like MPP architectures to allow queries to be processed in the most optimized way possible. Through its knowledge of SQL and relational data structures, a data warehouse appliance is architected to remove all the bottlenecks to data flow so that the only remaining limit is the disk speed. Through standard interfaces like SQL and ODBC, it is fully compatible with existing Business Intelligence (BI) and packaged 3rd party applications, tools and data.

At its core, data warehouse appliances simplify the deployment, scaling, and management of the database and storage infrastructure. As a result, ease of use and operation are considered paramount qualities of a data warehouse appliance. Ultimately, the vision of a data warehouse appliance is to provide a self-managing, self-tuning, plug-and-play database system that can be scaled out in a modular and cost effective manner. To that end, data warehouse appliances are defined by four criteria:

- **Workload Optimized** - A data warehouse appliance is optimized to deliver excellent performance for large block reads, long table scans, complex queries, and other common activities in data warehousing.

- **Extreme Scalability** - A data warehouse appliance is designed to scale and perform well on large data sets. In fact, the sweet spot for all data warehouse appliances on the market today is databases over 1TB in size.

- **Highly Reliable** – A data warehouse appliance must be completely fault tolerant and not be susceptible to a single point of failure.

- **Simplicity of Operation** – A data warehouse appliance must be extremely simple to install, setup, configure, tune, and maintain. In fact, a data warehouse appliance promises to eliminate or greatly minimize mundane tuning, data partitioning, and storage provisioning tasks.

Data Warehouse Appliance Landscape

Although the notion of a data warehouse appliance may sound like nirvana, the vendor community has been particularly active in defining and innovating around this vision of a data warehouse appliance. The grandfather and original data warehouse appliance is Netezza. However, since Netezza's market entry, several other firms, such as DataAllegro, Dataupia, and Kognitio have entered the market with variations and improvements on the original concept of a data warehouse appliance.

Although architectural approaches to data warehouse appliances vary widely, there are four main pivot points for assessing different vendor's approaches. First, does the data warehouse appliance replace existing third party database software with it own purpose built kernel? Most of the data warehouse appliances replace traditional database kernels like Oracle, IBM DB2, and Microsoft SQL Server with their own optimized database kernel. One exception to this rule is Dataupia. Dataupia, unlike the other data warehouse appliances, interoperates and does not replace existing database systems and instead complements those investments.

Second, does the data warehouse appliance use low cost industry building blocks or customized ASICs and FPGAs to achieve high levels of scalability and performance? Netezza makes heavy use of custom ASICs and FPGAs to increase performance and scalability, while other designs (Datallegro, Dataupia, and Kognitio) utilize industry standard building blocks in order to offer the best price/performance possible. The total cost and overall price performance of the solution can be directly affected by the underlying components a vendor chooses to use.

Third, does the data warehouse appliance make use of a highly parallelized design to gain greater scalability and performance? All vendors leverage some degree of parallelism to deliver the requisite performance and scalability desired. However, with any highly complex product, the devil is in the details. End users would be well served to scrutinize and understand the various architectural tradeoffs and benefits taken by each vendor and assess whether those tradeoffs are well suited to their database workload.

Fourth, what is the overall entry price point of the solution and can a user scale storage capacity in increments that match how their data warehouse is growing? Data warehousing appliance vendors have widely divergent price points. Several solutions start at $100,000s of dollars and can easily top out at several million dollars for an average implementation. Moreover, several solutions require end users to purchase their storage capacities in relative large chunks (greater than 10 TBs). As a result, some appliances may be cost prohibitive for smaller scale or slower growing data warehousing deployments.

Table 1. Data Warehouse Appliance Comparison

	Datallegro	Dataupia	Netezza	Kognitio
Optimized Workload	Data Warehouse	Data Warehouse	Data Warehouse	Data Warehouse
Works with Existing Databases (e.g. Oracle, DB2, SQL Server)	No	Yes	No	No
Industry standard component design	Yes	Yes	No	Yes
Massively Parallel Processing architecture	Yes	Yes	Yes	Yes
Entry Price Point	$110,000	$19,500	$200,000	N/A
Capacity Scaling Increments	15-20 TBs	2 TBs	12-25TBs	N/A

Taneja Group Opinion

We are strong believers that over next five years workload optimized storage appliances, such as data warehouse appliances, will emerge to become key elements of the storage infrastructure in most data centers. In fact, we have seen storage appliances such as NAS and CAS define new classes and types of storage and become a part of the overall industry lexicon. Data warehouse appliances represent another data point in this historical trend toward more specialized, workload optimized storage infrastructure. However, that is not to say, general purpose storage will be replaced or rendered obsolete by these optimized storage appliances. We see workload optimized storage devices carving out specific market niches where application specific scaling, performance, and management requirements are unique and not easily met by general purpose storage designs.

Large scale data warehousing represents a significant headache for IT today. The continuing data tsunami, need to keep more structured data online longer, and the insatiable need for faster, more responsive databases are driving users to consider new storage alternatives. Add to the mix that the current database scaling technologies are too cost prohibitive, inflexible, or too operationally expensive to meet the ever increasing demands of the business. Specialized storage approaches, such as the data warehouse appliance, offer a novel approach that provides cost-effective scalability and simplified management of structured content. End users must realize that the new requirements of structured content and data warehouses are creating a brave new world. They must be willing to embrace new approaches to solve the vexing problems of scaling and managing large scale data warehouse implementations today and in the future.

NOTICE: *The information and product recommendations made by the TANEJA GROUP are based upon public information and sources and may also include personal opinions both of the TANEJA GROUP and others, all of which we believe to be accurate and reliable. However, as market conditions change and not within our control, the information and recommendations are made without warranty of any kind. All product names used and mentioned herein are the trademarks of their respective owners. The TANEJA GROUP, Inc. assumes no responsibility or liability for any damages whatsoever (including incidental, consequential or otherwise), caused by your use of, or reliance upon, the information and recommendations presented herein, nor for any inadvertent errors which may appear in this document.*