Data Management Appliances

Claudia Imhoff, Intelligent Solutions, Inc
Colin White, BI Research

August 2007
Sponsored by Dataupia

TABLE OF CONTENTS

EXECUTIVE SUMMARY

BI and data warehousing technologies continue to evolve and innovate producing more efficient and cost effective ways to deliver this critical functionality. The latest innovations in this area are IT appliances specific to the performance and data management required for the BI analytics of sophisticated enterprises. These appliances can be quickly defined by these four characteristics:

- They have *one purpose* only.
- They are contained within *one package*
- They require *one installation* of their technology
- There is only *one vendor* to call for support and maintenance.

This paper compares and contrasts data management appliances with other data-centric appliances

To understand IT appliances, it is worthwhile to review their history and evolution. The examples described in this paper include database machines and storage management appliances, which are followed with descriptions of the data warehouse and data management appliances. For data management appliances, the paper compares and contrasts this new approach with other appliance solutions, focusing specifically on differences and similarities (at least in terms of their requirements) with fellow data-centric appliances.

The newest data management appliance vendor is Dataupia, founded in 2005 by Foster Hinshaw. The company's flagship product is the Dataupia Satori Server 12000. This appliance uses a massively parallel processing architecture to run data-intensive operations for its host system(s). These include IBM DB2, Microsoft SQL Server, and Oracle databases. Each Dataupia Satori Server contains one or more hardware blade servers with each blade holding 2 terabytes of disk space at a starting cost of $20,000.

The paper continues with a description of how to get started using a data management appliance by assessing your company's need for the appliance's performance, scalability, and cost features while analyzing its ease of use, flexibility, and agility in your environment. A list of best practices to consider when selecting this new form of IT technology is also included to help ensure a successful implementation.

INTRODUCTION

IT appliances have been around for many years helping companies implement and manage their IT environments. Examples include network appliances, storage appliances, database machines, etc. In recent years, we have seen the advent of new types of appliances that serve the data-intensive environments like data warehousing. This paper discusses appliances for data-intensive operations, and takes a detailed look at the role of a data management appliance in BI and data warehousing.

WHAT IS AN APPLIANCE?

To explain what a data management appliance is, let's start with a general definition of appliances. 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 a toaster, the user simply wants the bread toasted to the preferred specific color, and really does not care how the toaster performs this feat. The bottom line is that this appliance is designed to do one thing and to do it very well.

One purpose
One package
One install
One support vendor

Another example of an appliance is a stereo amplifier. Most of us have no idea how it boosts the performance of the other stereo components, 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 our other stereo components (DVD, CD, video players, and so on). We don't have to configure it or tweak it once everything is plugged in to get good performance. It is certainly much cheaper than if we had to buy the individual amplifier pieces and put them together ourselves. 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. Finally if something goes wrong with it, we have only one phone number to call. Therefore, the characteristics of a good appliance are:[1]

- One purpose – designed to perform a widely recognized function or set of closely related functions. The simplest case of an appliance – the toaster – demonstrates this point well. While it is tempting to try to make it do other things (Remember our college days?), it is best to stick to toasting different forms of bread.

- One package – tested, ordered, and delivered as a single unit. An appliance comes in a single box, no assembly needed, no involved instruction manual, and certainly no advanced degree needed to understand its function.

- One install – installed and maintained as a single unit. Most appliances need little more effort than to remove them from the box and plug them in. Maintenance

[1] From the B–EYE-Research.com paper titled "Data Warehouse Appliances: Evolution or Revolution?" by Colin White, Richard Hackathorn

consists of occasionally dusting it and removing any debris that might have accumulated from its usage.

- One vendor to call – single point of service provided by a single vendor. Should the appliance malfunction, it is either repaired or replaced by a single call to a vendor.

One purpose means the appliance has been optimized to perform a single function

Such a device comes about as a result of the maturation of the prevailing technologies. The toaster or amplifier of today resulted from years of research and development and was built with its specific purpose in mind. This "one purpose" 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 toast your bread or amplify your other stereo components. Basically the evolution of these technologies allowed appliance companies to "pave an existing cow path" of function with clever, componentized systems.

These appliances were 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 can easily integrate into anyone's environment.

Finally an appliance like the toaster or amplifier is much cheaper to purchase than if we had to buy the individual pieces and construct the complete function from them. The synergy of controlling all the "stuff" in the black box and configuring it for the user has reduced not only the overall cost of the appliance but also the maintenance of it. Unless you are an overly inquisitive person, you would rarely take apart your toaster to see what makes it "tick".

With this as an introduction to appliances in general, let's now turn our attention to IT appliances, the different types available today and a closer look at the data management appliance specifically.

TYPES OF APPLIANCES

The use of appliances has grown rapidly over recent years, and there are now products for most types of IT processing. This paper focuses on data-centric appliances that support high performance data-intensive operations. Two important product types in this area are data management appliances and data warehouse appliances. Before taking a detailed look at these two types, however, it is worthwhile to briefly review the history and evolution of appliances.

Database Machines

Appliances for commercial use first appeared in 1980s with the release of *database machines* from Britton Lee (the BL8000) and Teradata (the DBC/1012). Competing products from companies such as Kendall Square Research, nCube, and Thinking Machines also joined the fray during this period. Most of these machines were based on relational database technology, which was beginning to gain traction at that time, especially for decision processing workloads.

Wikipedia defines a database machine as “A computer or special hardware that stores and retrieves data from a database. It is specially designed for database access and is coupled to the main (front-end) computer(s) by a high-speed channel.” One of the main objectives of database machines was to offload processing from mainframe computers, which were expensive to operate. They also established massively parallel processing (MPP) as a sound approach for doing analytical processing against large amounts of data.

Database machine vendors enjoyed moderate success for a few years, but with the exception of Teradata, they ultimately failed. The advent of low-cost client/server computing reduced the price/performance benefits of database machines, and their proprietary hardware architecture was no match for off-the-shelf commodity hardware used by client/server systems. David DeWitt and Jim Gray, well known database experts, summed up the situation in a 1992 paper [2] “In retrospect, specialized database machines have indeed failed; but parallel database systems are a big success.”

Storage Management Appliances

The next generation of appliances in the 1990s focused on storage management. As with a database machine, the objective of a *storage management appliance* is to offload processing from more expensive systems running large-scale enterprise workloads.

Another goal of a storage management appliance is to share data between multiple host systems. One of the first companies in this area was Network Appliance with its series of network attached storage (NAS) devices. Since then additional storage management solutions from companies such as EMC have been introduced. Examples include storage area networks (SAN) and content addressable storage (CAS). Today, storage management appliances represent a multi-billion dollar business.

Data Warehouse Appliances

Several vendors have introduced data warehouse appliances

Over the past decade, the IT industry has also seen growth in the use of *data warehouse appliances*. A number of startup vendors and several large system vendors have introduced a variety of data warehouse appliances into the marketplace. Some of these products have evolved from database machine hardware technology, while others have been developed by packaging together and optimizing existing commercial hardware and relational database software products.

[2] D. J. DeWitt, J. Gray, “Parallel Database Systems: the Future of High Performance Database Systems,” ACM Communications, June 1992.

Data Management Appliances

Like a database machine, the objective of a *data management appliance* is to offload certain data-intensive operations from a host computer. In a database management appliance environment, the host systems could do operational processing, while the appliance could handle analytical processing on behalf of the host systems. Dataupia (the sponsor of this paper) is an example of a vendor that provides this type of appliance.

Data management appliances compared with other types of appliance

The best way of explaining the role and benefits of a data management appliance is to review how it differs from other types of appliances:

Differences from database machines. Data management appliances represent a new generation of database machine. Like their predecessors they provide better price/performance for certain large-scale workloads than that provided by the host systems to which they are connected. This cost advantage is becoming increasingly more important as companies face scalability issues in data volumes, workload complexity, and the number of concurrent users to be supported. Three key differences from database machines are 1) they are built using commodity (non-proprietary) hardware, 2) they can be connected to multiple hosts, which enable data sharing across systems, and 3) they provide all the advantages of the appliance approach, including ease of use and administration, reliability, and expandability.

Differences from storage management appliances. With a storage management appliance, the database system software resides on the host system, and access to data on the appliance is done using file-based and block-level protocols. With a data management appliance, the database software resides on the appliance itself, and interaction between the host and the appliance occurs at the database statement level. This moves the processing of data much closer to where the data is stored and managed, which can improve performance significantly.

Differences from data warehouse appliances. A data warehouse appliance is a standalone computer system that supports cost-effective processing for large-scale data integration and analytical workloads. To do this, the complete workload and all its associated tools must be migrated to, and run on, the data warehouse appliance. This means that the organization must carefully decide ahead of time what workloads will run on the appliance. With a data management appliance, the workload and its associated tools run on the host computer, and database requests are routed to the host and appliance database systems as required. This potentially reduces the amount of work involved in deploying the appliance because applications and tools do not have to be installed on the appliance itself.

With both data warehouse and data management appliances, the required databases on the appliance must be defined and loaded. With both types of appliances, full SQL compatibility with existing applications and tools is essential to avoid recoding.

Data management appliances have the advantage that they allow multiple heterogeneous hosts to be connected to the appliance and share appliance data. Although data management appliances support operational workloads, for most projects they will be used primarily for analytical, archival, and disaster recovery workloads. This is because most organizations do not have major scalability issues with their operational workloads.

Users need to select an appliance that best matches the requirements of the project

Each of the appliance solutions reviewed above has its strengths and weaknesses, and all of them have a role to play in IT systems. Prospective users therefore need to select an appliance that best matches the requirements of the project under development. The next part of the paper summarizes some of the main selection criteria for data management appliances.

REQUIREMENTS FOR A DATA MANAGEMENT APPLIANCE

Many of the requirements for a data management appliance are the same as those for other types of data-centric appliances. Key requirements here include:

- One integrated package, one install, one vendor support point
- Low total cost of ownership
- High performance and scalability
- High availability, reliability, and fault tolerance
- Easy administration
- Flexible expansion
- Open architecture for enterprise integration
- Support for data security, compression, and encryption
- Support by third-party software vendors for appliance interfaces

Certain requirements are unique to data management appliances. Important ones to consider here are:

- Support for operational and analytical host workloads
- Full compatibility with host DBMSs including SQL support (SQL manipulation and definitional syntax, data types, stored procedures, triggers, referential integrity), database utilities (load interfaces, backup, recovery) and administration tools (security, auditing, workload manager, SQL optimization and statistical tools, common administration workbench)
- Appliance specific tools for database loading, backup, and recovery
- Multi-host read-only and read-write processing with full integrity and recovery
- Support for third-party systems management tools

Now that the objectives, positioning, and requirements for a data management appliance have been discussed, we are now in a position to review the new data management appliance from Dataupia.

VENDOR EXAMPLE: DATAUPIA

COMPANY OVERVIEW

Dataupia was founded in 2005. Its President and CEO is Foster Hinshaw, who was one of the original co-founders of the data warehouse appliance company, Netezza. The company is privately held, and has venture funding from Polaris Venture Partners and Valhalla Partners. It has some 50 employees and is based in Cambridge, MA. More information on Dataupia can be found at *www.dataupia.com*.

PRODUCT OVERVIEW

Dataupia's flagship product is the Dataupia Satori Server 12000. (Satori is a key concept in Zen Buddhism that refers to deep or lasting enlightenment.) The Dataupia Satori Server bundles processors, storage, and system software into a single data management appliance. The appliance is connected to one or more host systems by a gigabit Ethernet network. When connected to multiple hosts, the server guarantees full transactional integrity for all read and write operations.

The Dataupia Satori Server employs a massively parallel processing architecture to run relational database intensive operations that have been offloaded from the host systems to which it is connected. The appliance is built using one or more Linux-based hardware *blade* servers. Today, each blade server consists of dual 64-bit AMD Opteron processors and 2 terabytes of disk space managed on eight RAID-5 drives. Each blade server has hot spares, mirroring, and record level multi-versioning.

Additional 2 terabyte blade servers can be added to a Dataupia Satori server configuration as required without disrupting existing operations. A configuration can potentially scale up to hundreds of terabytes. Systems start just under $20,000.

Supports the offloading of selected DB2, SQL Server, and Oracle host workloads

The appliance supports host applications that run against IBM DB2, Microsoft SQL Server, and Oracle. Dataupia claims full compliance with the SQL dialects of all three of the database system products. Dataupia software on the host computer captures required SQL operations, parses them, and executes them on the Dataupia Satori Server. Dataupia utilities provide the ability to extract or replicate data from existing databases, and load it into a Dataupia Satori database. While the data is being loaded, a fast indexing capability runs in the background. Appliance database backup with real-time and restore utilities are also provided. A management console features dashboards statistics and configurable alerts to monitor performance and manage Dataupia Satori Servers in the network.

DATAUPIA CUSTOMERS

At present, Dataupia has several OEM partners, who package the Dataupia Satori Server with their own products to make new solutions. These include:

Focus Data Services provides a secure portal that is used by telephone companies in the United Kingdom to provide information to police and emergency services. When building the portal, Focus had the option of consolidating multiple data stores, or using federation approaches to access the required information *in situ*. Neither method offered an acceptable solution. The consolidation approach was cost prohibitive, and the federation approach had limited performance. The solution was to use the Dataupia Satori Server to reduce the cost of consolidating the information from multiple back-end systems. Focus states that they are now able to supply its offering to customers at slightly more than a tenth of the original price.

Tektronix develops applications for telecommunications companies to manage their networks, services, and customers. The amount of data involved in these applications is significant. Tektronix's telecommunications customers need to quickly and affordably analyze many months worth of failed call logs to effectively manage, troubleshoot, and optimize their networks. Keeping all of this data online and readily available was becoming cost prohibitive. The solution to this problem was to install a Dataupia Satori Server to manage this data at a much lower cost.

Sendio develops and markets enterprise e-mail spam blocking appliances that employ sender address verification to block spam. The storage and retrieval requirements in this environment are substantial. Sendio is using the Dataupia Satori Server to reduce the cost of managing these large amounts of data, and to ensure it can support scalability requirements in the future.

ANALYSIS

There are two ways to analyze the Dataupia Satori Server. The first is to compare it against its predecessor, the database machine. The second is to contrast it with data warehouse appliances.

Database machines failed for two reasons. The first is that they began to lose their price/performance competitive edge as companies began to use lower cost client/server systems instead of mainframe computers. The second reason is that IT departments felt database machines were proprietary and didn't integrate well into the data center environment. These departments were also reluctant to rely on start-up companies and products for enterprise computing.

Data management appliances provide better price/ performance than database machines

As companies face ever-increasing demands for data, they are facing the same cost issues that occurred in the days of mainframe computers. Data management appliances provide better price/performance than database machines ever did. They also make more use of commodity hardware, but some aspects of their architecture, however, are still proprietary.

There is still a certain amount of opposition to the use of appliances by IT departments, but this is easier to overcome since appliances often enable projects to be developed that are simply not possible from an economic standpoint using more traditional approaches. The projects involved are also more analytical than operational in nature, and are sometimes considered (incorrectly) by IT to be less mission-critical to the business.

Data management appliances compared with data warehouse appliances

The second, and more important, aspect of data management appliances is how they compare with data warehouse appliances. Outlined below are factors to consider when comparing the two approaches.

1. Both types of appliance are designed to reduce the cost of ownership of IT solutions that support data-intensive operations. This is achieved not only through the use of lower cost hardware and software, but also by an environment that is easier to administer.

2. The more generalized design of a data management appliance allows it to support a broader range of applications and workloads than a data warehouse appliance. The MPP architecture of the data management appliance, however, is better suited to analytical workloads than operational ones. Other possible workloads include data archiving and disaster recovery.

3. Data management appliances support data sharing between multiple host systems. Data warehouse appliances do not support data sharing.

4. Data warehouse appliances are highly tuned for processing simple queries against very large data warehouses. Their ability to support more complex workloads varies considerably by product. Data management appliances are able to support more complex workloads, but they may not provide the same level of performance (as data warehouse appliances) for more simple queries against very large databases.

5. One of the biggest advantages of a data management appliance is that existing applications and tools do not have to be migrated to the appliance and can continue to run unchanged on host systems. It is still necessary to move the required data to the appliance, and it is crucial that the appliance support the full SQL syntax of the applications it handles. It is very difficult for any appliance vendor to provide full compatibility with another commercial database product, and this ultimately could become an important differentiator between products.

6. At present, the cost of the Dataupia Satori Server is lower than its data warehouse counterparts. This price coupled with its more generalized architecture make it particularly attractive to medium sized companies.

Benefits of Dataupia are its low cost, ability to handle existing applications, and data sharing between hosts

The data-centric appliance marketplace is becoming increasingly crowded. The advantages of the Dataupia Satori Server are its low cost, its capability to handle existing applications, and its support for data sharing between host systems. The ability of Dataupia to compete with data warehouse appliance vendors will rely heavily on these three important differentiators, and it will be crucial for Dataupia to demonstrate that its product can integrate seamlessly with existing database applications without any changes being required.

GETTING STARTED WITH DATA MANAGEMENT APPLIANCES

As with any new initiative, the first step a project manager must do is create a business case for that initiative. It is no different for a project that involves a data management appliance. This paper lists a number of reasons why a company should use such technology; the task of the project manager is to make these pertinent to his or her initiative. Start by examining the need for the following features or functions:

- Performance – Many data warehouse projects fail due to the technology's inability to deliver results to complicated queries in a timely manner. A project that uses massive amounts of data (or will eventually be using massive amounts in the near future) must use technology that has a proven track record for delivering results quickly.

- Scalability – Performance is just a start though. Projects using large volumes of data typically continue to add huge amounts of data requiring the technology to scale to sizes unimaginable a decade ago. Therefore, you can easily justify the need for technology that must be easily expanded to accommodate the growing data volumes.

The starting cost of Dataupia is less than $20,000

- Cost – Of course if the cost of the increased storage and performance breaks the bank, then the project will not get off the ground at all. Fortunately, as you have read, there is technology today to not only perform well, store large volumes of data, but do so with a remarkably small price tag. At a starting cost of less than $20,000, Dataupia is certainly a viable option to consider.

These are certainly good places to start but the project manager must go further and consider after-the-project aspects in the business case. These consist of the various maintenance features in the technology:

- Ease of use – While certainly of concern during the start of any project, ease of use becomes even more important to data-intensive ones as staff members turn over, new ones come on board, and new requirements are implemented into the existing environment. The one install, one package, one vendor to call characteristics of data management appliances ensure that ease of use remains paramount in the minds of the Dataupia technologists. As an example of simplicity as data requirements increase, the ability to add another 2 terabytes by plugging in another blade is simplicity at its best.

- Flexibility and agility – Flexibility comes from the ability to insert new technology without having to revamp an entire environment. As mentioned earlier, existing applications and tools continue to run unchanged on host systems. No migration is needed other than to move the data to the appliance environment. And because a data management appliance is not just for data warehouse applications, it can be used for multiple purposes including archival and disaster recovery.

- Reliability – It is paramount upon the project manager to ensure that the chosen technology is made from standard, proven pieces. This is no different for a data management appliance. Under its covers, this appliance must be composed of established technology that is reputable and highly regarded. In addition, it is recommended that the project manager focus on the *one purpose, one package, one install, one support* aspects described earlier. These have a great deal to do with not only reliability but also ease of maintenance and use.

BEST PRACTICES

In selecting a new technology for an IT environment, there are a few best practices to consider:

Vendors should perform a POC for a subset of the project's requirements

- In view of the six aspects above, the top vendors should perform a proof of concept (POC) for a given subset of the project's requirements. Each vendor should demonstrate the technology's ability to handle the difficult aspects of the project's requirements. The project manager should choose those activities that are considered to be beyond the existing technological environment for testing by the new appliance vendor. All six features above should be examined in the POC in close detail.

- Until the IT support personnel are familiar with the new technology, it is advisable that the project that will use the technology be relatively small in scope and not mission-critical – at least at first. Both of these can change once the team is adept with and knowledgeable about the data management appliance. For the scope, the project can start by "productionalizing" the POC. These usually have the small scope and non-critical characteristics needed. Once in production, the scope can expand and mission-criticality brought in.

Make the vendor your partner

- Make the vendor your partner. The project should bring the vendor's technical staff into the project as full members at first. The vendor should have a significant stake in the success of the project. Once the environment is stable and in production, the vendor should ensure a full transfer of knowledge to the client team. The project manager should take advantage of any training or education made available to the team.

- Understand where the data management appliance can and cannot speed up your project. Many business intelligence project managers do not realize that the most time-consuming aspects of their initial project come not from the usage of the data warehouse data but from the data integration and data quality activities needed to get the data into the warehouse. Unfortunately the data management appliance can't help with these activities. It helps after the fact – once the data is loaded and ready for analysis. The project manager must set parameters on these latter activities to judge the true value that the appliance brings.

SUMMARY

IT departments have many technological choices to make today and environments that require data-intensive support are particularly challenging. A complicated analytic processing involving a massive amount of data is most challenging for modern BI environments. These types of difficult environments demand new thinking and new technologies from our vendors.

The innovative form of today's data management appliance technology is well suited to support these and other applications requiring high-performing, data-intensive, complex manipulations in an easy to use and reliable environment. Cost-effective solutions like Dataupia's are changing the traditional paradigms by removing conventional barriers to large data warehouses and sophisticated analytics and by creating cost effective solutions to today's complex data management environments.

Understand where a data management appliance can be used most effectively

It is necessary to understand where the data management appliance can be used most effectively and what types of data manipulations are most suitable to this technology. This requires an in-depth understanding of the appliance's features and functions as well as solid knowledge of your business users' needs and requirements. This white paper is a good start toward your understanding of the concepts.

A proof of concept is highly recommended to further determine the business needs that should be followed with an initial project of limited scope. Both give the technology the best chance of succeeding in your environment. Once completed, the environment can be rapidly expanded to meet the growing needs of the business community and the increasing demands of information technologies.

In summary, the data management appliance offers a fast, affordable way to ensure your data-intensive applications run in an optimal environment that is easy to deploy, use, and maintain. Data management and data-intensive applications have never had it so good.

About BI Research

BI Research is a research and consulting company whose goal is to help companies understand and exploit new developments in business intelligence and business integration. When combined, business intelligence and business integration enable an organization to become a smart and agile business.

About Intelligent Solutions, Inc.

Intelligent Solutions provides data management consulting, education, and literature to business and government organizations worldwide. It specializes in business intelligence and customer analytics, and in the systems required to enable and support them.

BI Research
Post Office Box 398
Ashland, OR 97520
Telephone: (541)-552-9126
Internet URL: www.bi-research.com
E-mail: info@bi-research.com

Intelligent Solutions, Inc.
PO Box 4587
Boulder, CO 80306
Telephone: (303) 444-2411
www.intelsols.com
moreinfo@intelsols.com

Data Management Appliances
Version 1, August 2007