Database Are Not Toasters: A Framework for Comparing Data Warehouse Appliances

Omer Trajman[1], Alain Crolotte[2], David Steinhoff[3], Raghunath Othayoth Nambiar[4], and Meikel Poess[5]

[1] Vertica Systems, 8 Federal St., Billerica, MA-01821, USA
Omer@Vertica.com
[2] Teradata, 17095 Via del Campo, San Deigo, CA-92127, USA
Alain.Crolotte@teradata.com
[3] ParAccel, 9920 Pacific Heights Blvd. Suite 450, San Diego, CA-92121, USA
david.steinhoff@paraccel.com
[4] Hewlett-Packard Company, 11445 Compaq Center Dr. W, Houston, TX-77070, USA
Raghu.Nambiar@hp.com
[5] Oracle Corporation, 500 Oracle Parkway, Redwood Shores, CA-94065, USA
Meikel.Poess@oracle.com

Abstract. The success of Business Intelligence (BI) applications depends on two factors, the ability to analyze data ever more quickly and the ability to handle ever increasing volumes of data. Data Warehouse (DW) and Data Mart (DM) installations that support BI applications have historically been built using traditional architectures either designed from the ground up or based on customized reference system designs. The advent of Data Warehouse Appliances (DA) brings packaged software and hardware solutions that address performance and scalability requirements for certain market segments. The differences between DAs and custom installations make direct comparisons between them impractical and suggest the need for a targeted DA benchmark. In this paper we review data warehouse appliances by surveying thirteen products offered today. We assess the common characteristics among them and propose a classification for DA offerings. We hope our results will help define a useful benchmark for DAs.

Keywords: Appliances, Benchmark Development, Databases, Data Warehousing, Database Systems Standard.

1 Introduction

Business Intelligence continues to be a top priority of Chief Information Officers today because organizations gain a competitive edge by successfully leveraging integrated enterprise business intelligence. This capability helps them better understand customers, streamline their supply chain, and improve financial performance. Enterprise BI systems require larger data warehouses to support them, meaning more data, tables, users and query complexity. In today's market vendors

R. Nambiar and M. Poess (Eds.): TPCTC 2009, LNCS 5895, pp. 31–51, 2009.

offer three types of data warehouse configurations: traditional custom configurations, reference architectures configurations and data warehouse appliances.

Traditional data warehouse systems are custom-built to meet individual customer requirements. Such configurations are highly flexible and provide high performance to meet demanding data warehouses while providing for potentially unlimited scalability for future growth. Custom-built configurations have a high pre-sale, implementation and support price tags.

Reference architecture configurations are best practice building blocks that are pre-tested and documented. Such components are typically designed to support varying scenarios at multiple levels of data and query scaling. Some immediate benefits for deploying data warehouse solutions based on reference configurations include:

- Simplified product selection and sizing process
- A predefined set of core components to simplify purchasing
- Simplified, rapid deployment with predictable performance
- Off-the-shelf components to enable repurposing and lower risk

Alternatively, data warehouse appliances are designed around scalable design principles, delivered with preconfigured hardware and software pre-installed, and are ready to run at a specified performance level out of the box.

In this paper we attempt to analyze the product offerings, not as a marketing exercise, but to attempt and discern the properties that data warehouse appliance vendors themselves offer and believe are lacking in the traditional offerings. The remainder of section one covers a brief history of how database appliances have been portrayed by analysts and vendors. Section two is an analysis of the common properties of database appliances based on a market survey conducted by the authors of appliances available today. The survey is included in full as Appendix. Section three attempts to classify and construct a test for what customers may consider an appliance. Section four describes future work including the need for an industry standard benchmark by which the market can assess key properties that are unique to appliances.

The primary motivation of this study is to determine what features data appliances have in common in order to facilitate the process of defining a data appliance. Wikipedia proposes the following definition for a data warehouse appliance: “a data appliance consists of an integrated set of servers, storage, operating system(s), DBMS and software specifically pre-installed and pre-optimized for data warehousing.” [1] They go on crediting Teradata and Britton-Lee as the first appliances on the market – both companies were founded on 1979. Britton-Lee was acquired by Teradata which itself was acquired later on by AT&T/NCR and then spun off. The company who can be credited with “popularizing” the concept of data appliance is Netezza in the early 2000s. There has been a flurry of start-ups in the current decade most benefiting from open source PostgreSQL or Ingres. They implement traditional relational architectures or vertical technologies. All traditional database vendors have recently joined the bandwagon with special-purpose bundles either software running on some specific vendor hardware or proprietary hardware/software combinations.

Numerous analysts as well as vendors have attempted to find a single definition for a database appliance (or an IT appliance in general). Foster Hinshaw, arguably the father of the modern data warehouse appliance, used the benefits of an appliance to

attempt a definition [2]. Gartner in 2006 found that vendors began misusing the word appliance starting in the late 1990s and created a set of questions to evaluate an appliance [3]. Microsoft published a review of many database appliances in the market in the context of comparing them to SQL Server 2008 using openness and cost-effectiveness as their criteria [4]. The whitepaper is a market positioning paper that argues that systems from older vendors such as Teradata, IBM and Oracle are costly to operate. The paper further maintains that newer systems are too specialized and immature, and vendors have provided little or no public support network. The whitepaper fails to acknowledge or analyze market demands driving appliance vendors to introduce new products that address limitations in traditional data warehousing solutions.

These definitions are relatively recent, but already appear to be outdated as the pace of new product offerings continues to develop. The increasing number of vendors offering dedicated appliances or appliance variations of their solutions makes it clear there is a wide market demand for an alternative to classic “do it yourself” data warehouse solutions. Despite some common characteristics across the vendors described in this paper, the rapid progress of appliance market presence continues without clear industry wide technology convergence.

2 Survey Analysis

Surveying the vendor offerings available in the market today we distilled a catalog of common properties attribute by the various vendors to their appliance offerings. We describe these properties and introduce a matrix (see Table 1) that summarizes which solutions offer each property. This analysis is not a validation of the properties that each vendor promotes, rather an extrapolation of how vendors perceive appliance offerings differently from custom configurations and reference configurations. This classification is also not intended as an exhaustive enumeration of the properties that the market demands of data warehouse appliances. This list does highlight some of the key properties which are later analyzed in the context of existing benchmarks.

2.1 Efficiency and Energy Efficiency

Data warehouse appliances are preconfigured not only for performance and cost, but also for energy efficiency and physical size. Many appliances are designed to offer "out of the box" efficiency relieving the user of balancing different hardware and software configurations. Whereas an assembled system may have additional CPU that the software cannot take advantage of or more disks that are necessary for the workload, appliances are balanced to maximize resources. For example, a system that requires ten disks for warehouse type workloads on 1TB of data would ship with ten appropriately configured disks of the correct capacity and speed. A custom-built system could easily be over or under provisioned leading to degradation in performance or unnecessary space and power consumption.

2.2 Large Volumes/Capacity of Data

Data warehouse appliances are primarily geared towards large volumes of data. While some appliances focus at smaller volumes (less than 1TB) or medium volumes (1-10TB), many are designed for environments with 100s of terabytes. Constructing data warehouses at the larger scales is a complex project that can consume system design resources over an extended period of time. To date only the most sophisticated organizations have been able to construct 100TB warehouses, always with substantial assistance from hardware and software vendors. As smaller organizations look to store and analyze more data, an appliance model - pre configured and tested at scale can be very attractive. A pre-built system with a shipping capacity of 100TB simplifies much of the provisioning, design and installation required in a self-built system.

2.3 Expandable in Modular Units

Data warehouse appliances are designed with integrated modular components. An expandable appliance is well suited to environments where data volume, numbers of users, workload or performance requirements grow over time. Some appliances are complete (closed) configurations that have fixed capacity with limited or no expansion capability. Other appliances are designed for limited or continual modular expansion at the storage system, server or rack level. A modular system allows expansion starting at a smaller size such as 1TB system all the way up to some maximum capacity such as 100TB or even 1PB+. Similar to the advantages of a fixed appliance configuration, an expandable appliance alleviates the ongoing system design burden from the user that needs to scale the system.

2.4 Single SKU

While it may appear trivial, the value of a single SKU when assembling a complex data warehousing system can greatly simplify the purchasing and maintenance process for a large system. Ordering separate hardware and software can often lead to weeks of analysis, comparing the different components that are available at any given point in time. Often times the disks, additional CPUs and memory upgrades are priced separately. Some RDBMS software requires packages on top of the base software when deploying specifically for data warehousing. Similarly when upgrading a system, different components may be available resulting in dozens of changes to a purchase order.

2.5 Single Vendor/Support

Traditional data warehouse systems are composed of various combinations of hardware and software components. For instance, a general purpose RDBMS may run on any suitable hardware/OS platform chosen by the user. In a single-vendor system the hardware and the software are sold by the same vendor. The single vendor is responsible for the entire system delivery and support. A single-vendor system resolves many of the contentious issues that can arise when troubleshooting a large complex system. For example, data warehouse workloads often involve large disk

scans. Poor performance may result from a software issue that has incorrectly laid out data on disk, a disk firmware issue that is incorrectly buffering data. A single vendor takes responsibility for resolution regardless of the source of the issue.

2.6 Specialized Hardware

While some data warehouse appliances are constructed using industry standard hardware that is also used for other applications, some appliances are constructed with specialized hardware. This may be a chip that is designed specifically for warehousing queries or a disk that is programmed to primarily read and write data sequentially. Since the specialized hardware is design and built specifically for the appliance it is typically not sold as a separate component. For certainly types of workloads, specialized hardware may provide a particular advantage or may complement the software provided in the appliance.

2.7 Specialized Software

General purpose DBMS software is designed to run on a variety of different types of hardware. This flexibility in software that can run on a variety of hardware platforms can lead to missed opportunities to optimize for a specific tested configuration. Some data warehouse appliances offer specialized software, designed only for their appliance offerings. The specialized software is tuned and embedded into a custom hardware design to maximize performance and scalability of the appliance solution. The specialized software is not available other then as part of the data warehouse appliance.

2.8 Special Purpose or Purpose Built

Often Data warehouse appliances are specifically designed to handle certain types of workloads. For example data warehouses usually have large volumes of data, are primarily ready only, compared to transaction processing systems, and handle workloads ranging from reporting to ad-hoc analytic queries. To accommodate these use cases, some appliances are designed from scratch or tuned in such a manner that the software and hardware configuration is optimized for warehousing workloads. This may require sacrificing performance for non-warehousing workloads, such as transaction processing.

2.9 Pre-installed

Different major versions and minor revisions of database software, operating system, drivers and interoperability can cause confusion and interoperability challenges. Specific bug fixes addressed by software may be unnecessary following hardware revisions (and vice versa). Commonly with data warehouse appliances, all software and hardware components are pre-installed and tested for compatibility with specific version of each component. While not necessarily an official certification, the packaging of each version combined with the support (and presumably quality assurance tests) result in a higher guarantee of function and performance for the appliance user.

2.10 Pre-configured or Fast Deployment

Custom Data Warehouses and warehouses built from reference architectures require significant planning, design and implementation effort. The larger the system the more complex these requirements become and the longer they will take to design and implement. Most data warehouse appliances are delivered as configurable software or complete software/hardware systems that require only basic connection and setup installation. Unlike traditional custom-built data warehouse configurations, data warehouse appliance hardware is pre-installed (servers, storage, disk arrays) to specification from the factory and shipped directly to the customer. Pre-configured application software may be bundled with the system, as well. The complete appliance speeds everything from provisioning to installation and deployment.

2.11 Massively Parallel

Massively parallel processing refers to a system in which independent units execute operations on separate data in parallel. In the context of a database system MPP generally equates to a "shared-nothing" architecture where units have their own memory and their own data on disk. For redundancy purposes however, certain processing units and I/O subsystems can be arranged in clusters sharing data. A combination of modularity and scalability, MPP is a common architecture for data warehousing and data warehouse appliances.

2.12 Packaged Database and Application

Data warehouse solutions require not only hardware, operating system and database software, but also application software to load and query the database. A few data warehouse appliances have entered the market with combined database and application software. These are sometimes loosely coupled with joint marketing, packed together and per-configured. Others have specific optimizations and are sold as a single turnkey system.

2.13 Packaged Database and Hardware

Most Data warehouse appliances are pre-packaged with database software, operating system and hardware. From design through sales, marketing and support these appliances function as a single solution rather than a combination of products.

2.14 Bundled Solution vs. Appliance Only

Data warehouse appliances may be configured and delivered the pre-installed OS and database product components. Bundled solutions extend beyond the basic appliance to include general purpose or pre-configured vertical product software. These may include general or specialty applications, connection/interface components for variable data environments or external storage components. Appliance only solutions are comprised of components that are not sold separately. Often appliance only solutions include specialized hardware and software.

2.15 Self Managing or Self Tuning

While the data warehouse appliance that arrives from the factory is tuned for a specific size and workload. Both the composition of the appliance and the target workload may change over time. A self-managing system is able to re-balance based on changing workloads and data. Similar to a device that can self-align, the self managing or self tuning appliance stand in contrast to a user assembled data warehouse that often requires manual maintenance and tuning both out of the box and as an ongoing task.

2.16 Bundled Pricing

Bundled pricing refers to a commercial practice whereby a single line item is provided to the client for the entire system including hardware, software and maintenance. The appliance bundle can be accompanied by a detailed list of items that constitute the bundle (a requirement for TPC pricing for example) though not necessarily for single-vendor bundles. A typical bundle could be for instance: System X with software version Y and 3-year maintenance for Z dollars.

Table 1. Analysis of different appliances

A. Greenplum/Sun
B. HP/Neoview
C. HP/Oracle
D. HP/Vertica
E. IBM
F. Ingres
G. Kickfire
H. Microsoft/DATAllegro
I. Netezza
J. Oracle
K. ParAccel
L. Sybase
M. Teradata
N. Vertica

System Characteristics	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Efficiency and Energy Efficiency	X	X		X		X	X				X	X		X
Large Volumes/Capacity of Data	X	X	X	X	X		X	X	X	X	X		X	X
Expandable in modular units	X	X	X	X	X			X	X		X		X	X
Single SKU	X	X	X				X	X	X		X		X	X
Single Vendor/Single Support	X	X		X	X	X	X	X	X		X	X	X	X
Specialized hardware			X				X	X	X				X	
Specialized software			X			X		X	X		X		X	
Special purpose or purpose built	X	X	X	X		X	X	X	X	X	X	X	X	X
Pre-installed	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Pre-configured / Fast deployment	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Massively Parallel	X	X	X	X				X	X	X	X		X	X
Packaged database plus application						X					X	X		
Packaged database plus hardware	X	X	X	X	X		X	X	X	X	X	X	X	
Bundled Solution vs. Appliance		X	X			X					X	X		
Self managing or self tuning				X					X	X	X			X
Bundled pricing	X	X		X							X	X		

3 A Data Warehouse Appliance Test

In the first two sections of this paper we reviewed the existing market demand for data warehouse appliances, appliance offerings from over a dozen vendors and the properties that each of these vendors ascribed to their appliance offerings. In this section we extrapolate from these properties a test that consumers can use to identify whether an offering meets a commonly accepted criterion for an appliance. We do not claim an absolute definition of a data warehouse appliance rather we have relied on the collective research of these vendors to make our assessment.

A brief glance at table 1 shows some clear consensus among vendors as to the qualities of a data warehouse appliance. DAs are targeted at large volumes starting at 100GB or greater and expandable in modular units. Appliance vendors provide a single source of support or at least a single point of contact for purchase and return. Most appliances are specially designed or customized for data warehousing and all are pre installed and pre configured for warehousing type workloads. Finally, nearly all appliances are offered as packaged database and hardware.

While these seven most popular criteria are not an absolute definition of an appliance, it seems that most consumers expect the scalability and simplicity of an appliance to include most if not all of these features. Where vendors differ is in the nine other properties – either due to design decisions or to offer competitive differentiation. For example, a common but not universal property of data warehouse appliances is massively parallel processing. Ten out of fourteen solutions employ this architecture in their appliance offerings. Similarly, seven vendors have some specialized software or hardware that is designed specifically for their appliance solution while only three vendors offer additional bundled application software that is not included in their non-appliance offering. Though bundled pricing is not espoused as an appliance feature, an informal inquiry revealed that discounts among appliances are common.

The trend among these key properties is that at first glance most are not easily compared quantitatively. Yet the quantitative measures tested in existing benchmarks including performance and price/performance are absent from the list of key appliance properties. While many data warehouse vendors often claim price and price/performance measures of their custom-built configurations, only those that exclusively sell appliances do so of their DA offerings. This dichotomy leads us to speculate that there is a need for some evaluation of the defining properties for data warehouse appliances.

Existing benchmarks such as the TPC-H decision support benchmark and its predecessors emphasize query performance and overall price performance. These benchmarks require that submitters record load times but do not factor them into the results. Nor do the benchmarks measure setup time or scalability of a single solution. In fact, comparison across scales is expressly forbidden [15]. While we cannot claim

that existing benchmarks provide a disincentive to run appliances, it is telling that only three appliances have been run since the concept first became popularized ten years ago. In particular vendors who sell both appliance and non-appliance solutions have generally opted to run their non-appliance solutions.

4 Need for a Data Warehouse Appliance Benchmark Specification

The distinguishing factors in data warehouse appliances fall under two broad categories: scalability and simplicity. The consensus among vendors appears to be that the market demands a solution that can grow in modular units from 100s of gigabytes to 100s of terabytes or greater by simply plugging in new components purchased from a single vendor. The proliferation of data warehouse appliances from both new and established vendors indicates that this demand is not met by traditional custom configurations or even reference configurations.

We believe there is a strong market demand for a suite of tests that benchmark data warehouse appliances. The warehousing decision support benchmarks that exist today, while satisfactory for comparing custom and reference configurations do not test for the properties that embody data warehouse appliances. Rather than focus exclusively on traditional workloads and timing aspects, this assessment should also focus on appliance specific factors and account for the properties outlined in this paper.

A simple proposal could begin with extending or supplementing an existing benchmark to create quantitative measures that capture the essence of scalability and simplicity. For example, the TPC-H benchmark could be augmented to encourage comparisons across scale as a measure of modularity and change in price/performance across different volumes of data. Scalability testing might be performed without reloading instead appending new data as a customer would. Simplicity could be measured by including the load timings, including a random query generation phase or even comparing the amount of configuration information provided to the system for setup.

We encourage bodies such as the TPC to take on this future work to define a data warehouse appliance benchmark specification that assess scalability and simplicity of a packaged solution. The benchmark should also definitively articulate the un-measurable qualities of an appliance solution to facilitate evaluation by consumers.

Acknowledgements

The authors would like to thank Karl Huppler, Michael Corwin, Kannan Govindarajan for participating in the survey.

References

1. Wikipedia contributors, Data warehouse appliance. Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Data_warehouse_appliance (accessed June 20, 2009)
2. Hinshaw, F.D.: Tera-Scale Data Appliances for Business Intelligence. The Data Administration Newsletter (April 1, 2003), http://www.tdan.com/view-articles/5075
3. MacDonald, et al.: Findings: Not All Appliances' Are Appliances. Gartner Research (September 12, 2006)
4. SQL Server Technical Article: Data Warehouse in the Enterprise, A Competitive Review of Enterprise Data Warehouse Appliances and Technology Solutions, Microsoft web site (January 2009), http://tinyurl.com/l8b4x6
5. Greenplum website: The Data Warehouse Appliance, http://www.greenplum.com/partners/data_warehouse_appliance/ Sun, Greenplum Unveil Data Warehouse Appliance, http://www.hpcwire.com/offthewire/17886374.html?viewAll=y
6. HP Neoview Enterprise Data Warehouse, http://www.hp.com/go/neoview
7. HP Oracle Database Machine, http://www.hp.com/go/exadata, http://www.hp.com/go/exadata
8. A Next-Generation Teramart™ Solution from Vertica, HP & Red Hat, http://www.vertica.com/appliance
9. Shanley, K.: History and Overview of the TPC, http://www.tpc.org/information/about/history.asp
10. IBM Business Intelligence, http://ibm.com/bi
11. InfoSphere Balanced Data Warehouse, http://www-01.ibm.com/software/data/infosphere/balanced-warehouse/
12. Microsoft DATAllegro: DATAllegro Data Warehouse Appliance, http://www.datallegro.com/data_warehouse_appliances/data_warehouse_overview.asp
13. Netezza Marketing Collateral: Data Warehouse Appliance, http://www.netezza.com/data-warehouse-appliance-products/dw-appliance.aspx
14. The ParAccel Analytic Database, http://www.paraccel.com/data_warehouse_resources/library.php
15. TPC Policies v5.17, http://tinyurl.com/l88ykx

Appendix Data Warehouse Appliance Survey

This section surveys an array of solutions that may be viewed as appliances in today's market place. In order to perform an unbiased survey we contacted twelve vendors and their partners to contribute a one-page description of their appliance solution, which have been included in unedited form, except for formatting changes to comply with publisher guidelines.

Table A.1. Contacted Vendors/Partners and Products offered

Vendor/Partner	Product Offering
Greenplum/Sun	The Greenplum Data Warehouse Powered by Sun
Hewlett Packard	HP Neoview Enterprise Data Warehouse
Hewlett Packard/Oracle	HP Oracle Database Machine
Hewlett Packard/Vertica	The Vertica Analytic Database Appliance
IBM	IBM InfoSphere Balanced Warehouse
Ingres	Ingres Icebraker Appliance
Kickfire	Kickfire Analytic Appliance for the Data Warehouse Mass Market
Microsoft	DATAllegro Appliance
Netezza	Data Warehouse Appliance
Oracle	Oracle Optimized Warehouse
ParAccel	ParAccel Analytic Database
Sybase	Sybase Analytic Appliance
Teradata	Teradata Data Warehouse Appliance
Vertica	The Vertica Virtualized Analytic Database

We received product descriptions from Hewlett Packard, IBM, Kickfire, Oracle, ParAccel, Sybase, Teradata and Vertica. For those vendors who did not contribute a description we have researched their website and press releases and compiled a short description. These descriptions may not fully reflect the features and capabilities of these products. The intention is to come up with a list of appliance properties commonly claimed by vendors.

A.1 The Greenplum Data Warehouse Powered by Sun

Sun-Greenplum Data Warehouse Appliance [5] is open source software and general purpose systems powered by the Solaris 10 Operating System (OS) and PostgreSQL. The solution combines Sun Fire X4500 data server powered by dual-core AMD Opteron processors with Greenplum's massively parallel distribution of PostgreSQL, Bizgres MPP, in a single turnkey appliance. With its Query-In-Storage design, the solution is capable of scanning 1 terabyte of data in 60 seconds and can scale to hundreds of terabytes of usable database capacity. Sun claims the data warehouse system is one of the most energy efficient solutions in the industry, at only 90W per terabyte. The solution offers the following key attributes:

- Performance - Massively parallel processing, made possible by the performance scalability of the AMD Opteron processor with Direct Connect Architecture, leverages a high-performance interconnect and moves processing to where the data resides.
- Openness - Powered by open source software including the Solaris Operating System, PostgreSQL, and Solaris ZFS. Supports industry standards and interfaces (SQL, ODBC, JDBC).

- Value - An acquisition and ongoing administration cost of less than $20,000 per usable terabyte. Small footprint (up to 50 TB per rack) and low power requirements (4.5 kW per rack) reduce operating costs.
- Ease-of-use - Integrated and turnkey appliance reduces system complexity, for easy implementation and maintenance.

A.2 HP Neoview Enterprise Data Warehouse

HP Neoview enterprise-class data warehousing platform [6] is designed to meet the needs of a 24x7 operational BI environment, where massive amounts of information are analyzed in seconds. This pre-configured, pre-integrated and pre-tested platform is built to help simplify the deployment process. It includes startup and deployment services that enable users to begin loading data as soon as the system is powered up.

Table A.2. Neoview features and benefits

Neoview feature	Key customer benefit
Parallel execution across hundreds of servers	Handles complex queries, mixed workloads, and high concurrency
Massive scaling to hundreds of terabytes	Allows complete analysis of large volumes of data, maximizing business insight from all information assets
Innovative optimization engine	Delivers unparalleled performance
Built-in system-wide fault tolerance	Provides mission-critical 24x7 capabilities and mitigates risk
Remote management and monitoring from HP	Simplifies administration and reduces maintenance costs
Single vendor solution	Provides accountability and simplicity so that you achieve your goals and mitigate risk
Priority driven workload execution and user defined service levels and rules	Gives consistent performance to meet service levels of varied mixed workloads
Industry-standard components	Protects investment and facilitates data center integration. Allows component reuse.
Completely integrated and preconfigured hardware, software, and services	Assures faster time to operation. Provides simplicity that lowers cost of operations.
Fully compatible with leading data integration, ETL, query, analysis, and reporting tools	Allows easy integration into an existing environment and preserves investment in tools, training, and process

There are fourteen HP Neoview platform models. The Cxxx model family uses 146 GB user data disks and the Exxx model family uses 300 GB user data disks.

A.3 HP Oracle Database Machine

The HP Oracle Database Machine [7] is a complete system, including software, servers, networking and storage designed to run multi-terabyte data warehouses. At the heart of this system, is the HP Oracle Exadata Storage Server, which has smart

storage software that offloads data-intensive query processing from database servers closer to the data. This results in much less data getting sent over fast InfiniBand interconnects, dramatically improving both query performance and concurrency. It's simple and fast to implement, very cost-effective, and can linearly scale storage capacity, processing power and network bandwidth as your data warehouse grows.

HP Oracle Database Machine is pre-configured for performance, pre- tuned, and certified for Oracle Business Intelligence Enterprise Edition tools and Oracle Real Application Clusters. Complete configurations can be ordered from Oracle, with hardware support by HP. The HP Oracle Database Machine is a high-performance system configured for data warehousing that includes a grid of eight database servers featuring: 64 Intel processor cores, and Oracle Enterprise Linux; and a grid of 14 HP Oracle Exadata Storage Servers that include up to 168 terabytes of raw storage and 14 GB/sec data bandwidth to the database servers, when accessing compressed data, the effective throughput for a single rack can be 50 GB per second or more. Up to eight Database Machines can be networked together without requiring additional InfiniBand switches, and larger configurations can be built with the addition of external switches.

A.4 HP Vertica Analytic Database Appliance

The HP Vertica Analytic Database appliance [8] is a pre-configured HP c-class BladeSystem delivered as a single unit with integrated compute and storage blades. The HP BladeSystem c3000 or c7000 enclosure is designed as a balanced system with either 4 or 8 database nodes, each comprised of a compute blade paired with a storage blade. The Vertica software and RedHat Linux operating system are pre-installed and configured, delivering an instant out of the box database experience. In addition to the benefits of a single fully configured system, the HP Vertica appliance also delivers a dense and efficient real time data warehousing solution. A single c3000 can accommodate up to 20TB of user data with a c7000 scaling to 40TB or approximately 2TB per 1U of rack space[1].

The HP c-class BladeSystems are expandable by simply plugging in new pairs of blades or new enclosures and registering them with the running system. For example, a c7000 enclosure can be purchased with half capacity (c3000 equivalent) and then expanded by adding compute and storage blade pairs. A full c3000 or c7000 based Vertica Appliance can be connected over Ethernet or InfiniBand to any number of additional appliances for additional capacity. Since each enclosure is self contained, each compute blade and storage blade forms a single unit and each Vertica nodes functions in a shared-nothing configuration, appliances can be sized as small or as large as the customer requires, up to four c7000 enclosures or 32 Vertica nodes per full rack capable of warehousing a full 160TB of user data.

As with all configurations of the Vertica Analytic Database, the HP Vertica appliance is designed to be always on. Since each of the nodes functions as a peer in the database operation so users can connect to any node and see a single system image, regardless of the number of nodes. The nodes all support high availability so

[1] 20TB c3000 or 40TB c7000 is fully configured with pairs of BL460c and an SB40c with 6x 300GB SFF SAS.

that one or more blade failures do not bring down ten rest of the cluster. When adding and removing nodes or while the database is self-tuning, users can continuous load and query data against the appliance.

A.5 IBM InfoSphere Balanced Warehouse

The IBM® InfoSphere Balanced Warehouse™ is the complete data warehousing solution comprised of pre-tested, scalable and fully-integrated system components of InfoSphere Warehouse, Server and Storage. It takes the best attributes of appliance-like solutions, while still maintaining the flexibility of typical relational database servers. Like an appliance it offers the ease of deployment and configuration. The building-block approach in Balanced Warehouse™ allows for easy growth to handle additional workload and data in a data warehouse. This is unlike many "appliance" data warehouse solutions. And, while the configuration comes with a set of tools designed to make the implementation and maintenance of a data warehouse easy to accomplish, the customer also has the complete power of DB2 available to use as they need. Included in the IBM® InfoSphere Balanced Warehouse™ are:

- InfoSphere Warehouse software: Including tooling and infrastructure to help data warehouse architects and administrators efficiently design, deploy and maintain an enterprise data warehouse.
- A preconfigured, fully installed server
- A fully installed storage subsystem
  - Currently, the IBM® InfoSphere™ Balanced Warehouse™ has three basic building blocks:
- IBM® InfoSphere™ Balanced Warehouse™ C4000, comprised of:
  - InfoSphere Warehouse Intermediate Edition software
  - IBM System x3850 M2
  - IBM System Storage DS3200
  - Optional EXP3000 Storage
- The IBM® InfoSphere Balanced Warehouse™ D5100 includes:
  - InfoSphere Warehouse software
  - IBM System x 3650
  - IBM System Storage DS3400
  - Optional EXP3000 Storage
- The IBM® InfoSphere Balanced Warehouse™ E7100 includes:
  - InfoSphere Warehouse software
  - IBM System p 570 (on POWER6)
  - IBM System Storage DS4800
  - Optional EXP810 Storage

The D5100 and E7100 are available in the form of five flexible modules, designed to deliver affordable scaling that meets a customer's needs:

- Foundation (includes all the function needed to manage the data warehouse)
- Data (for flexible scaling of data and data access capabilities)
- User (for extending the user access capabilities)

- Failover (for high availability solutions)
- Application (portals to a variety of data warehousing applications)

For more information on the IBM® InfoSphere Balanced Warehouse™ see [10][11].

A.6 Ingres Icebraker ECM Appliance

Ingres web site states that today's general-purpose databases and operating systems include a huge array of features, most of which are not used by all applications. Features that are not marketed as optional are included in products by default. Customers have to pay, install and maintain them. This may contribute not only to an unnecessarily high license fee, but also to an increased cost for deployment, configuration maintenance and security.

The Icebreaker Software Appliances try to simplify installation, reduce security risks and, simplify deployment to improvements in efficiency and cost savings. Simplicity is achieved by reducing the layers in the appliance to the bare essentials. Additionally, Ingres takes responsibility for the integration of the various components within the appliance. The provided setup utility installs and configures the complete set of technologies as one unit. Ingres provides maintenance for the entire Icebraker Appliance.

Reducing the components used in a database reduces the risk of exposure to security vulnerabilities. Additionally, securing the environment is greatly simplified because of less number of components that may interact. Lastly, the Ingres Icebraker Appliance only opens ports that it requires, only creates the minimum number of accounts and only starts the services and demons that are required.

Being designed from the ground up, Ingres' Software Appliances take advantage of the latest advances in hardware and software solutions, which are preinstalled and configured on a particular hardware configuration.

Ingres' appliances take advantage of Virtualization Technologies resulting in high efficient systems while also reducing cost. Especially since they are supplied as a unit, they greatly reduce the cost of installing, configuring, deploying and maintaining these solutions.

A.7 Kickfire Analytic Appliance for the Data Warehouse Mass Market

Kickfire™, Inc., delivers the first analytic appliance that affordably delivers the high-performance capabilities of large commercial database systems to the mass market. The data warehousing mass market constitutes those deployments in the gigabytes to low terabytes which, according to IDC, represent over three quarters of the total market.

This market presents significant challenges to traditional vendors. Customers need performance but are very price-sensitive. They have limited data warehousing expertise and few IT resources with which to deploy solutions. Finally early-stage data warehousing deployments, typical in the mass market, often include mixed workload characteristics.

Kickfire has addressed these challenges the first analytic solution based on a parallel-processing SQL chip. Kickfire chose to package the solution as an appliance

because this has proven to be an attractive deployment model for customers in the mass market. The key characteristics of the Kickfire appliance are summarized below.

Kickfire's appliance combines the industry's first SQL chip, which packs the power of 10's of CPUs, and an advanced column-store engine with full ACID compliance, it achieves low price/performance based on rigorous benchmark tests and ensures complete data integrity. The small form factor and low power consumption of the device are designed for today's cost- and green-conscious customers. Some properties of Kickfire's appliance are:

- An integrated and optimized solution down to the operating system level with features like Active System Monitor which notifies users of any potential system anomaly.
- Runs standard MySQL Enterprise™.
- Built to scale to the needs of the mass market in terms of data volumes, user concurrency and workload type, handping data sizes from the gigabytes to the terabytes.
- Offers high-user concurrency on a single node, supporting 100 concurrent users and 1,000 active users.
- Fast bulk loading as well as fast incremental loading to handle traditional as well as operational workloads.

As highlighted above, Kickfire's solution delivers all the benefits of an appliance, namely price/performance, ease of use, and manageability. Additionally, the Kickfire Appliance enables the mass market to scale in terms of data capacity, user concurrency, and workload type.

A.8 Microsoft DATAllegro Appliance

Prior to acquisition by Microsoft in July of 2008, DATAllegro sold a data-warehouse appliance named DATAllegro v3 [12].The DATAllegro appliance was created to enable rapid deployment of hundreds of terabytes of data at relatively low cost. The architecture was designed to offer both flexibility and scalability using a modular and open, standards-based technology. DATAllegro v3 utilized EMC storage, Dell servers and Cisco InfiniBand switches. Each server contained Intel multi-core CPUs and was powered by the Ingres open source database. Combined, these technologies allowed DATAllegro to offer a low cost, high performance appliance as well as a reliable and scalable solution.

Traditionally, SAN based data warehouse storage solutions offers slow query performance and reduced disk space available for user data. This is because traditional warehouses were optimized to put data only on the fastest parts of the disk and to deploy as many small capacity spindles as possible to maximize IOPs. DATAllegro employed at RAIDW technology to maximize overall I/O performance as well as provide built in fault tolerance across all system components, not just the storage.

The disk patters underlying most data warehouse solutions resemble random I/O, requiring a high number of IOPs, achieved with a very large number of spindles. This

is due to the number of indexes, and the complex disk infrastructure required to support different hard ware platforms and operating systems. By developing Direct Data Streaming technology, DATAllegro was able to optimize query execution to use sequential disk I/O and minimize the need for tuning.

DATAllegro also reduced network traffic using their USN parallel database design. USN configured the database to maximize co-located joins across all nodes in the system. This minimizes network traffic and improves performance for most query types. Additionally, the use of multi-level partitioning reduces the amount of data read for each query. DATAllegro also supported indexes as required for queries that accessed a small number of rows each.

A.9 Netezza

Netezza Performance Server is a data warehouse appliance combining database, server, and storage, based on Linux and PostgreSQL. It uses a patented massively parallel architecture. The following is a copy the first page of Netezza's description of a data warehouse appliance [13]: "To eliminate the need for constant tuning and the technology bottlenecks caused by slow disk transfer I/O rates, slow network transfer rates and inefficient caching, Netezza developed a system designed specifically for analytical processing on extremely large amounts of data. We call our system an "appliance" because, like a network or storage device, it was designed to do one thing extremely well. It is very straightforward – when you plug it in and load your data, it analyzes it very fast. And because the Netezza appliance is an integrated database-server-storage system, you are no longer faced with building your own system out of disparate components, or maintaining these different piece parts.

Netezza's patented massively parallel architecture takes a different approach than traditional database architectures to processing queries. Where standard computing architectures are targeted at operations on individual data elements, the Netezza architecture is all about "streaming" processing. Rather than shuttling data between disk and memory for processing once a query comes in, which creates the bottleneck, data streams off the disk and through query logic loaded into an FPGA (field programmable gate array). The FPGA and processor (a PowerPC chip), together with 400 GB of disk storage, reside on each of the massively parallel nodes that Netezza calls snippet processing units (SPUs). Each of our Netezza racks contains 112 of these SPUs. Queries are optimized across the SPUs for maximum performance and power efficiency A Linux host server aggregates SPU results and manages query workload and the results are returned to the user. It is this different approach to the growing analytic processing challenge that provides our customers with a high-performance database engine that brings simplicity to what has become a very complicated process.

The Netezza 10000 product line starts with the 10050 with 56 SPUs (half a cabinet) rated at 6.25 TB of user data (12TB with compression). A basic 10100 (one cabinet) is rated at 12.5TB of user data with 112 SPUs. The other models are multipliers of the 10100 – a 10x00 with x=2, 4, 6 or 8 is composed of x 10100 cabinets and the number of SPUs and the amount of user data are x times that of the 10100. In summary Netezza emphasizes modularity, ease of use and lack of tuning chores

A.10 Oracle Optimized Warehouse Initiative

The Oracle Optimized Warehouse Initiative is a joint development between Oracle and its partners to provide customers with a choice of easy-to-implement, high performance and scalable data warehousing solutions. It provides optimized data warehouses that combine the world's most popular database for data warehouses, Oracle Database, with server, storage and networking components from HP, IBM, SUN, Dell, EMC and SGI. The Oracle Optimized Warehouse Initiative provides customers flexibility of choice to meet their different scale of requirements:

- Reference Configurations are a best practice guide to choosing the right server, storage and networking components to build an Oracle data warehouse. These best practice guides help customers take the risk out of implementing a data warehouse as Oracle and its partners have encapsulated years of configuration expertise.
- Oracle Optimized Warehouses provide customers with a pre-built, optimized data warehouse, complete with Oracle software, servers, storage and networking components ready to load data and run queries and reports. Available from Oracle's partners, Optimized Warehouses have been fully tested and validated.

There are a wide range of Reference Configurations available from HP, IBM, SUN, Dell, EMC and SGI to suit the different needs of all customers. These range in scale from 500GB to over 100TB data warehouses available on single and clustered servers to support tens to thousands of users with a wide choice of operating systems including Linux and Unix depending on specific partner reference configurations and Oracle Optimized Warehouses available.

Configurations and Oracle Optimized Warehouses are designed with modular scalability in mind to incrementally add more storage and processing power as demand grows.

Any Business Intelligence tool that is supported with Oracle Database can be used with the Optimized Data Warehouse. These include, OLAP, Data Mining, Oracle Business Intelligence Enterprise and Standard Editions, Hyperion, Business Objects, Cognos and many more.

Oracle Warehouse Builder, a feature of Oracle Database, is a core component of Oracle Optimized Warehouse Initiative enabling transformation of raw data from disparate sources into high-quality information for business reporting and analytics.

Oracle Enterprise Manager Database Control, a feature of Oracle Database, is used in the Oracle Optimized Warehouse Initiative solutions to automatically monitor the data warehouse environment and to help proactively resolve issues before they turn into emergencies.

The HP Oracle Exadata Storage Server, a combination of the HP DL180 G5 storage server and smart software from Oracle optimized for use with Oracle Databases, is part of the Oracle Optimized Warehouse Initiative. Its massively parallel architecture delivers outstanding query performance for data warehousing applications.

A.11 ParAccel Analytic Database

The ParAccel Analytic Database (PADB) is a purpose-built analytic DBMS that is available as both traditional enterprise software or as a software appliance that can be implemented on a variety of configurations of any brand of industry standard hardware. The Ultra-Fast Data Warehouse powered by Sun and ParAccel, was the world's first analytic appliance to publish TPC-H benchmarks (October 2007).

As a software appliance, PADB offers both the flexibility of server choice that people enjoy with traditional DBMSs and the convenience of a purpose-built appliance with its simplicity and lower cost of ownership (e.g., load-and-go installation, single point of contact for troubleshooting, etc.).

PADB installs with minimal setup overhead in environments with data volumes ranging from 100s of gigabytes to multi-terabytes, on configurations with variable numbers of servers based on standard multi-core x86 compatible CPUs. Despite the variable degree of configuration options, PADB uses a failover strategy that minimizes downtime and guarantees data integrity through intelligent mirroring, and the appliance's integrated SAN storage option. The SAN configuration also offers guaranteed disaster recovery capability in the unlikely event of multiple simultaneous systems failures.

PADB differs from most traditional DBMSs in that it assumes an analytic processing environment. PADB is built on a linearly scalable MPP shared-nothing architecture that uses columnar storage to optimize query performance, and data compression to minimize storage and access overhead. PADB includes support for advanced analytics defined by the SQL 2003 standard, complex multi-level aggregations, correlated sub-queries, and a patent-pending join strategy that allows the system to handle hundreds of tables in a single query. Another difference between PADB and many traditional DBMSs is that they were often designed for operational work and thus require many redundant structures or specialized design techniques (star schema, materialized view, indexes, summaries) to offer reasonable analytic query ability. PADB does not require these performance enhancement and tuning mechanisms to deliver record-setting performance for applications based on a wide range of schema, data and query requirements.

ParAccel is also available as vertical application appliance (e.g., Autometrics Pulse). Finally, PADB differs from all other databases (appliance and non-appliance) in how it integrates into SAN environments. Patent-pending blended scan features leverage both server- and SAN-based storage for higher overall performance than either storage type alone. The scalable analytical appliance (SAA), bundled with SAN, is available as a single SKU.

A.12 Sybase

In the analytics space, Sybase offers both an appliance and a software-only solution. Sybase opted for this route to cater to what they believe are two almost independent markets. Increasingly, some customers are exploring "black-box" solutions to analytic and data mart types of use cases. Sybase believes there are several value propositions, which are driving some of their customers to choose an appliance solution over separate components. Briefly, these are:

- Faster time to “Value”
  - No system integration work required, especially with respect to storage;
- Full solution stack:
  - the server and storage appliance already appropriately sized & configured (multiple options available based on data volumes and concurrent user estimates;
  - the DBMS software appropriately sized & configured;
  - Data modeling and development software to automate modeling and acquisition of source schemas as well as the creation of target schemas within the appliance;
  - ETL software to automate the extraction, transfer and load of production data into the appliance;
  - Top business intelligence tool to gain immediate business insight from the data;
  - a unified systems management console for the entire appliance.
- Single vendor support model:
  - No additional database tuning required
  - One vendor services the entire solution

Sybase also believe this market is particularly well suited to column-based RDBMS products because the vast majority of customers are really trying to get solutions to the business faster than IT can deliver. In these environments, Sybase find that the degree of ad-hoc, what-if scenario queries is significantly higher than in data warehouse / data mart use cases with traditional IT-based (non-COTS) applications. Column-stores inherent “index-everything” model is generally better suited to these types of environments.

A.13 Teradata

Teradata entered the appliance market in 1984 with the original Teradata DBC/1012, a Database Computer with integrated storage, servers, and RDBMS all supplied by a single vendor and purpose built for Decision Support processing. Today, although best known for the customized Active Data Warehouse 5550 platform, Teradata offers three appliance models.

- Data Mart Appliance 551P, suitable for entry-level data warehousing, data mart, or test and development systems up to 2.6 TB user data.
- Extreme Data Warehouse 1550, designed for very high-volume (100+ TB), non-enterprise data/analysis requirements for a small number of power users.
- Data Warehouse Appliance 2550, designed for entry-level data warehouses or analytical data marts.

Although all run the same Teradata RDBMS and SuSE Linux OS, each appliance has different disk subsystems and packaged software to suit their different purposes. All Teradata appliances are pre-configured and staged for quick installation. System options are minimized and simplified for easy configuration and ordering.

A.14 Vertica Analytic Database

The Vertica Analytic Database is a high performance SQL database, written from scratch to provide the best performance possible for data warehousing, business intelligence and analytic applications. Vertica's database management system is a massively parallel, compressed columnar system with a sophisticated design that provides built in high availability, continuous load and query capabilities and self-tuning technologies in an always-on system.

Vertica offers two version of the Vertica Analytic Database configured as a database appliance. Vertica's Analytic Database Appliance is a pre-configured system that includes industry standard hardware from vendors such as HP, Dell or Sun with an operating system and the Vertica Analytic Database pre-installed. Customers can purchase the Vertica Analytic Database Appliance and experiences a plug and play high performance database. Additional capacity can be added on the fly but plugging in new appliance nodes and registering them with the system.

The Vertica Virtualized Analytic Database is a VMware virtual machine image that includes a pre-installed copy of the Vertica Analytic Database. With a pre-installed operating system, a fully configured Vertica Analytic Database and the Vertica Webmin Administration tools, customers can deploy a full database solution near instantly. Using VMware vSphere, users can easily add capacity by starting new virtual machines.

Vertica's Web-based administration tools allow point and click addition of new nodes to the cluster – either physical or virtual. Adding and removing nodes is a fully on-line operation and data can be loaded, offloaded and segmented across the cluster all while users are continuously loading and querying data. As a shared nothing database, each Vertica nodes is connected to independent storage. This is directly attached to the nodes of a Vertica Analytic Database Appliance and provisioned at deployment time for the Vertica Virtualized Analytic Database.

All versions of the Vertica Analytic database include the automatic Database Designer, which creates an optimized physical designs using the data and query workloads as users begin to load and query data. The designer can run in the background as users continue to manage data and will optimize for new data or new workload patterns.

The Vertica Analytic Database appliances also include built in management and monitoring tools. With SNMP for critical events and remote monitoring capabilities, the Vertica Analytic Database appliance runs in a completely headless configuration. Vertica requires little to no administration and can be backed up using standard enterprise backup solutions.