The future looks bright considering all the innovation

State of the Data Warehouse Appliance

Krish Krishnan

Krish Krishnan is a recognized expert worldwide in the strategy, architecture and implementation of high performance data warehousing solutions. He is a visionary data warehouse thought leader and an independent analyst, writing and speaking at industry leading conferences, user groups and trade publications. His conclusion: The future looks bright for the data warehouse appliance.

We have been seeing in the past 18 months a steady influx of data warehouse appliances, and the offerings include the established players like Oracle and Teradata. There are a few facts that are clear from this rush of options and offerings:
- The data warehouse appliance is now an accepted solution and is being sponsored by the CxO within organizations;
- The need to be nimble and competitive has driven the quick maturation of data warehouse appliances, and the market is embracing the offerings where applicable;
- Some of the world's largest data warehouses have augmented the data warehouse appliance, including: New York Stock Exchange has multiple enterprise data warehouses (EDWs) on Netezza and Greenplum Appliances; Capital equity firm Arsenal Partners has an EDW on HP's Neoview, and HP has other customers that have gone public, including Walmart; Trade Doubler, a European web marketing firm, is using InfoBright's Brighthouse appliance to analyze web clickstreams; Corporate Express, the office supply giant, is running a data mart on Netezza; MySpace.com runs AsterData; Subex runs a revenue assurance data warehouse on Oracle-Dataupia; ATT runs a revenue assurance data warehouse on Exadata (LGR).

A lot of confusion and lack of clarity remains in the market about the subject of data warehouse appliances. Before we start looking at the data warehouse appliances, the different players and where the market is heading, let us pause and look at why we needed data warehouse appliances.

The Background

What is the biggest issue that keeps you awake at night? I pose this question in discussions with IT and C-level executives alike. Their answers range from:
- "My query runs for 10 minutes in a 100,000 records database runs for 10 hours in a 1,000,000 records database";
- "My users keep reporting that the ETL processes failed due to large data volumes in the DW";
- "My CFO wants to know why we need another $500,000 for infrastructure when we invested a similar amount just six months ago";
- "My DBA team is exhausted trying to keep the tuning process online for the databases to achieve speed."

The consistent theme throughout the marketplace focuses on the need to keep performance and response time up without incurring extra spending due to higher costs.
The key factors that are driving performance in a data warehouse are:
Data loading. Data loading into the data warehouse is one of the longest processes. The reason for this is manifold and has been discussed in depth in various ETL articles and case studies. To summarize the process of extracting various data feeds, processing them through data quality and data profiling systems and loading them with or without transformations to a final destination is time-consuming. This is especially true when the input volumes are low and you have smaller bursts of data since the speed is impacted by the volume of data in the data warehouse.
Data availability. Data availability service level agreements (SLAs) have a profound impact on the need for a high-performance environment. For data to be pristine, integrated and available for downstream applications such as reporting and analytics, end-user needs must be clearly documented. Another area where organizations often fall short concern data growth projections, data demand projections, data retention cycles and associated SLAs that have not been documented.
Data volumes. Data volumes in the data warehouse have been exploding, adding gigabytes every day. Growth rates for trans-

actional data at its granular level have increased by 50 percent in the past three years. Reasons for this data volume explosion include compliance, expanding pool of business users and increasing importance of analytics.

Compliance. SOX, HIPAA, GLBA, PCI regulations have mandated that structured data must be retained online long enough to meet eDiscovery and compliance requirements. Unfortunately, the overall performance and response time of current major RDBMS platforms suffers as the database size increases. This impacts databases larger than 1TB, which is now rapidly becoming the norm for a typical data warehouse with two to three years of data.

Legal mandate. Recent class action lawsuits in the pharmaceutical, manufacturing and tobacco industries have left companies finding a legal need to retain data for a longer period. The data structures and associated metadata need to be maintained whether online or offline to satisfy these needs.

Business users. Another major driver for the growth in the database size is coming from the business users. Business leaders are increasingly seeing the value of information gathered from exercises in data mining and exploration, historical analysis and predictive intelligence. Because of the desire to save the history of previously conducted activities such as promotions, campaigns and other revenue-generation opportunities to find business value, there is a demand for storing data longer.

Apart from these requirements, there are two more challenges that are being faced by the IT organizations:

Storage performance. Disk and storage systems have been consistently improving over the years both in terms of speed and performance, while costs have been relatively stable or less expensive. Architecturally, storage is shared across the data warehouse, making it a highly constrained area in terms of availability and performance. ETL and BI queries consume a lot of space and network bandwidth. If multiple queries are fired off on this shared storage architecture, even the best-in-class hardware and large disk capacity are not going to enable faster query processing and lightning response times for the results set. In addition, adding mixed query workloads to the storage architecture will produce slower performance cycles, resulting in a poor query performance and highly constrained network in terms of bandwidth. While strides are being made, even improving the overall storage performance will not create optimal conditions for data warehousing. *Faster is better* does not apply in this situation.

Operational Costs

The operational cost of running and maintaining a data warehouse is monumental for many organizations. Especially with the granularity of the data growing deeper and the increasing amounts of historical data to store, this two-way explosion has resulted in an unmanageable amount of information that needs to be handled by the data warehouse. In addition, related activities such as data mining, predictive analytics and heuristics analysis have placed a heavy demand on the resources in both hardware and IT administration. The overall cost of running and maintaining the data warehouse has left IT feeling numb.

A path-breaking alternative to reduce the cost of data warehousing while providing sustainable response time in mixed workload conditions is what we call the data warehouse appliance

The Data Warehouse Appliance

This is my definition of a data warehouse appliance is: *A data warehouse appliance is an integrated set of servers, storage, operating system, database and interconnections specifically preconfigured and tuned for the rigors of data warehousing.*

Data warehouse appliances offer an attractive price/performance value proposition and are frequently a fraction of the cost of traditional data warehouse solutions.

Netezza was the first commercially available data warehouse appliance in 2001, and since then we have seen a steady flow of new generations of data warehouse appliances in the market. Some of today's solutions include (this is not a complete listing): Netezza; Microsoft (DATAllegro); HP (Neoview and HP-Oracle DB Machine); Kognitio; ParAccel; KickFire; Aster Data; IBM; SUN; Greenplum; Oracle (HP-DB Machine, Exadata); Dataupia; Infobright; Exasol.

All the appliances have been built on the common basic principles:

- Massively parallel processing (MPP) architecture;
- Shared-nothing architecture;
- Commodity hardware and storage;
- Open source DBMS/commercial RDBMS platforms;
- Linux/UNIX – Open source operating system.

Type	Definition
Modular Appliance	A single integrated hardware and software packaged offering. The software and hardware are not individually licensed and cannot be separated.
Custom Appliance	Commercial software and hardware is optimized for data warehousing. The package offering is supplied by a single vendor and is installed and maintained as a single system.
Data Management Appliance	Can be augmented under the RDBMS to offload data intensive operations from a host computer. Serves as a slave appliance to the master database.
Software Appliance	Commercial or open source relational DBMS software is enhanced and optimized for data warehouse processing. The software supports hardware solutions purchased from one or more thirdparty vendors.

Figure 1: Classifications of Appliances.

Benefits

There are several benefits which are listed below.
Architecture Benefits of the Appliance: Commodity hardware and open source operating system; Intelligent storage with MPP built in for query optimization; Minimal DBA requirements for performance tuning and optimization; Built-in failover and fault tolerance; System administration requires minimal resources.
Performance Benefits: Queries will perform faster (50% to 1,000%) on an average compared to the traditional RDBMS platforms; Command-line options are available to adjust any data skew in the different partitions; A minimal indexing requirement allows data loading to occur faster; Resource allocation and marshalling on shared nothing architecture provides quicker query response;
Cost Savings: License costs many times lower than leading RDBMS; No additional cost for MPP engine; Minimal DBA requirements for performance tuning and optimization; Built-in failover and fault tolerance, reduces outages; Modular scalability, build as you go keeps costs in control.

There are numerous articles written on these benefits. But in the real world, how does one understand all the different offerings and choose the appliance that is best suited to their unique requirements?
Appliances can be broadly classified into four categories as shown in figure 1. Netezza, Teradata are the two truly large-scale modular vendors, Exasol is a small, medium-sized business (SMB) modular vendor. Kickfire, Microsoft, Sun and HP offer solutions based on custom appliance models. Exadata and Dataupia are data management and storage optimization appliances. Aster Data, ParAccel, Infobright and Kognitio largely are software-only appliances.
How do you determine which class of appliance to select? To answer this question, look at figure 2. If you need to reduce the latency at any of the layers as mentioned in this figure, examine your current data warehouse environment and determine where your biggest issues lie in terms of latency and process complexity:

- If the issues are with I/O and storage, the answer lies in a data management appliance, which can be augmented into the architecture. Dataupia and Oracle Exadata fall into this category;
- If you are a current Teradata customer, they have appliance augmentation options that you can examine. In this situation, you will also need to examine the cost of replatforming versus the business needs from an ROI perspective;
- If you are in need of high-volume analytical processing that cannot be handled in your current data warehouse environment, Aster Data, Netezza and Kognitio offer solutions that can be implemented with your current architecture and hardware platform and you might benefit from a software appliance;
- If you are looking to implement high speed and low volume processing and are willing to consider columnar databases, Sybase, ParAccel and Vertica have solutions that can be considered for a proof of concept (POC);

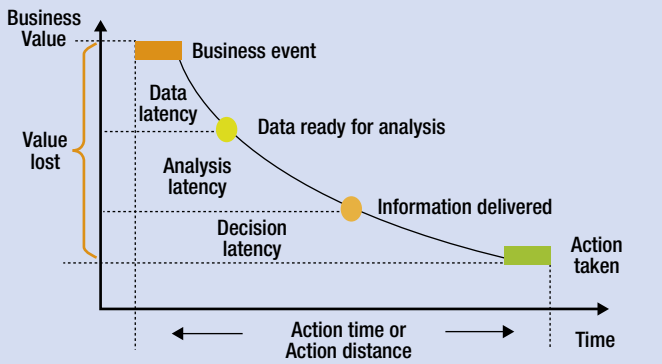

Figure 2: Latency Determination. (Courtesy Dr. Richard Hackathorn).

- Netezza, Teradata, Greenplum and Oracle can support deploying a data warehouse or a data mart on their platform. Determining your choice for the POC and final selection depends on your budget and needs;
- There are a number of companies that offer appliances specific to a niche market. For MySQL, you have a choice of Kickfire or Infobright as vendors to consider;
- Exasol is a modular platform and is popular in Europe;
- HP Neoview has solutions and the customers that will adopt to this platform will be looking for very large database (VLDB) types of solutions;
- IBM Balanced Configuration is another emerging contender.

Conclusion

While reading this article you may wonder why I have not recommended any one platform over others. The answer lies in your question, your situation and the requirements that will drive the choice of the platform that will be best suited for you. In the world of data warehouse appliances, "there is no one size fits all."
What is the current state of the data warehouse appliance? Well, the absence of a strong case against appliances, proves that only maturity and product diversity stand between the data warehouse market as we know it today and one dominated by appliances in the future. Not only will the future be dominated by the data warehouse appliance, you will also see BI appliances, Google appliances and others in the ranks.
The future looks bright for the data warehouse appliance, considering all the innovation in hardware, storage, processors and networks. The newer and better technologies will provide a better platform for the appliance market and will certainly increase price wars and market shares.

Krish Krishnan is the president of Sixth Sense Advisors Inc, a Chicago based independent analyst company covering the Data Warehouse and Business Intelligence areas, and teaches regularly at TDWI, DAMA, IRM UK and other conferences and publishes with www.beyenetwork.com\kkrishnan. Krish also serves as Associate Vice President of Programs for DAMA Chicago, and Ethics and Governance Advisor to DAMA International.