The Selection of Netezza for a Data Warehouse Platform

McKnight Associates, Inc.
BUILDING BUSINESS INTELLIGENCE

Tel: 214.514.1444
www.mcknight-associates.com

July 2005

The Decision Landscape

When making product decisions for a data warehousing (DW) environment, the database platform is of utmost importance. It should be chosen with care and with active discernment over the issues and marketing messages. The landscape for this selection has changed. It is now much more of a value-based proposition. The architectural offerings have moved beyond the traditional MPP or clustered SMP, which have been the standards for many years.

The platform decision could come about based on new initiatives. However, it is equally viable to reassess the platform when your current one is going to require major new investment or simply is not reaching the scale that you require. It is also prudent for every shop to periodically reevaluate the marketplace to make sure that the current direction is the right one in light of the new possibilities. Now is a time when the possibilities, with data warehouse appliances, merit such a reevaluation.

You will create a culture around your selected platform. You will hire and train your people to support it. It will become the primary driver for hardware and other software selections. Your people will attend user group meetings and interact with others using the DBMS for similar purposes. You will hire consultancy on the DBMS and you will research how to most effectively exploit the technology. You will need vendor support and you will want the vendor to be adding relevant features and capabilities to the DBMS that are needed for data warehousing in the future.

Data warehouses grow over time. Data volumes will soar over time as history accumulates, third-party data is added, clickstream data is added and new uses require new data. According to Giga research director Philip Russom "the number and diversity of BI users will continue increasing steadily throughout the decade, achieving a penetration of between 25 percent to 40 percent of all enterprise users." With all the investment and value, you'll want and need to leverage your data warehouse for customers, supply chain partners and possibly selectively to the broader Internet. You want to make sure you choose a proven platform not just for the initial, known requirements but also for the future, to-be-determined requirements.

As the greatest contributor to overall cost, the platform should be chosen that will deliver the requirements at the lowest cost. Budget dollars are flowing back to business intelligence, but there is a difference between the budgets I work with today versus those of late 1990s levels. They are more targeted and risk adverse. While ROI as a mechanism for establishing the value of business intelligence has not become mainstream, the linkage between business intelligence activities such as building data warehouses and making clean data available for analysis to increases in sales and decreases in expenses have never been more required. Companies everywhere, of all sizes, are engaging in business intelligence. When done properly, nothing adds value to business in today's economy than business intelligence. It's the modern competitive landscape.

The biggest challenge to achieving the business intelligence vision is query performance. Many shops manage this issue by actually limiting user queries to a predefined list or predetermined times!

The vendor mantra for business intelligence query performance seems to be "soldier on" or that the end user is not fully exploiting their existing technology. However, the only real answer to the problem is more hardware. This answer is seldom acknowledged initially, but frequently arrived at after learning another reality of today's business intelligence – most systems are already fully optimized for the hardware in place.

The quality of database administration and related disciplines to fully exploit database technologies has never been higher. There is a finite set of non-hardware-based activities ("physical modeling") that improve query performance, either directly or indirectly. Yet many still cannot deliver on the business intelligence promise of "the right data to the right people at the right time" due to query performance issues. Hardware upgrades and additions are left as seemingly the only means left to improve performance without changing the underlying architecture radically. Many end up scaling hardware in excess proportion to its delivered value. Performance gains tend to be sub-linear.

Of the major inputs to a data warehouse architecture specification and DBMS selection, eventual data size is one of the most important. At certain levels of data size (in the terabytes), data warehouse programs tend to gravitate towards similar levels of usage and complexity. There are programs in the multiple terabytes today with thousands of users and this represents the eventuality of many data warehouses starting out today.

Make sure your platform selection can accommodate a true eventual production data warehouse environment.

Criteria for a Platform Selection

The data warehouse platform selection is critical and acts as a catalyst for all other technology decisions. The technology needs to support both the immediate as well as future, unspecified and unknown requirements. Ideally the platform selection should be the first technology decision made for a data warehouse project.

Given the state of the marketplace, described above, the technical architecture should be:

- **Scalable –** In both performance capacity and incremental data volume growth. Make sure the proposed solution scales in a near-linear fashion and behaves consistently with growth in database size, number of concurrent users and complexity of queries. Understand additional hardware and software required for each of the incremental uses.
- **Powerful –** Designed for complex decision support activity in a multi-user mixed workload environment.
- **Manageable –** Through minimal support tasks requiring DBA/System Administrator intervention. It should provide a single point of control to simplify system administration. You should be able to create and implement new tables and indexes at will.
- **Extensible –** Provides flexible database design and system architecture that keeps pace with evolving business requirements and leverages existing investment in hardware and applications. What is required to add and delete columns? What is the impact of repartitioning tables?
- **Available –** Supports mission critical business applications with minimal down time. Check on "hot pluggable" components, understand system down time requirements and any issues that might deny or degrade service to end users. These can include batch load times, software/hardware upgrades, severe system performance issues and system maintenance outages.
- **Interoperable –** Integrated access to the web, internal networks, and corporate mainframes.
- **Affordable –** Proposed solution (hardware, software, services, required customer support) providing a low total cost of ownership (TCO) over multi-year period.
- **Proven –** You don't want to risk a critical decision regarding a fundamental underpinning of the data warehouse environment on an unproven solution.
- **Flexible –** Provides optimal performance across the full range of normalized, star and hybrid data schemas with large numbers of tables. Look for proven ability to support multiple applications from different business units, leveraging data that is integrated across business functions and subject areas.

The Vendor

There are few vendors who understand what it means to build production, mission critical data warehouse systems.

This decision process should go well beyond the usual feature/function comparisons that are done. The vendor itself should be a major category in the selection of a data warehouse DBMS. The vendor's financial stability, the importance of data warehousing to their overall business strategy and their continued research and development in the area of data warehousing towards a well developed and relevant vision are all key components of a vendor's viability in this critical decision.

While these criteria may seemingly shut out a newer entrant, this paper's vendor-of-interest, Netezza fits very well.

Data Warehouse Platform Trends

Numerous moves by major vendors are signaling the race to standards and full suite availability. This is partially in response to client receptivity to the "buy" (versus "build") model for business intelligence.

The "buy" model means that there is frustration with seemingly fruitless endeavors with tuning data warehouses for query performance, which is the biggest challenge to achieving the business intelligence vision of "right data to the right people at the right time." Limiting user queries to a predefined list or predetermined times perpetuates non- and under-consumption.

Mostly, the methods are already properly used in most business intelligence programs. Hardware upgrades and additions are left as seemingly the only means left to improve performance without changing the underlying architecture radically. Many end up scaling hardware in excess proportion to its delivered value. Carrying costs tend to be linear but performance gains tend to be sub-linear.

The appliance model, discussed below, alleviates many of these concerns.

A data warehouse platform consists of many components such as CPU, RAM and disk working together. Together, the components will determine processing power, I/O bandwidth, the upper limit on concurrent usage and the complexity of that usage and how much data can be stored. This makes the platform of paramount importance to business intelligence. The platform must be adaptable to growing requirements over time that may be unknown initially. It will also dictate, in large part, the TCO of business intelligence.

The mainstream business intelligence platform has evolved over the years through the following standards:

Symmetric Multiprocessing and Clustering

One of the early forms of parallel processing was Symmetric Multi-Processing or SMP. The programming paradigm was the same as that for uniprocessors. However, multiple CPUs could share the load using one or more of the forms of parallelism. A least-recently-used (LRU) cache, kept in each CPU, makes this option more viable. SMP tended to hit a saturation point around 32-64 CPUs, when the fixed bandwidth of the bus became a bottleneck.

Clustering became a way to scale beyond the single node by using an interconnect to link several intact nodes with their own CPUs, bus and RAM. The set of nodes is considered a "cluster". The disks could either be available to all nodes or dedicated to a node. These models are called "shared disk" and "shared nothing" respectively. This was great for fault tolerance and scalability, but eventually the interconnects, with their fixed bandwidth, became the bottleneck around 16-32 nodes.

Massively Parallel Processing

Massively Parallel Processing, MPP, is essentially a large cluster with more I/O bandwidth. There can be up to thousands of processors in MPP. The nodes are still either shared disk or shared nothing. The interconnect is usually in a "mesh" pattern, with nodes directly connected to many other nodes through the interconnect. The MPP interconnect is also faster than the Clustered SMP interconnect. Remote memory cache, called NUMA, was a variant introduced to MPP. DBMSs quickly adopted their software for MPP and while the interconnect bottleneck was eroding, MPP became a management challenge. And it is expensive.

We can see that each step was an evolutionary advancement on the previous step. With this paper, I propose that we may now be entering another advancement in business intelligence architecture – that of the data warehouse appliance. I'll define a data warehouse appliance as preconfigured hardware, operating system, DBMS, storage and the proprietary software that makes them all work together.

Introducing Netezza

Netezza is a leading data warehouse appliance vendor. Their preconfigurations range from 1.0 TB to 100 TB. Netezza's philosophy is that parallelism is a good thing and they take parallelism to a new level. They utilize an SMP node and up to 896 single-CPU SPUs ('snippet processing units') configured in an MPP arrangement in the overall architecture, referred to as "AMPP" for asymmetric massively parallel processing. The SPUs are connected by a Gigabyte Ethernet, which serves the function of the interconnect.

There are 112 SPUs in a rack. Each rack fully populated contains 4.5 TB. The racks stand 6'3" tall. The DBMS is a derivative of Postgres, the open source DBMS, but has been significantly altered to take advantage of the performance of the architecture.

What's inside are Hitachi drives, 2- or 4-way HP/Intel host CPUs and Red Hat Linux Operating System. These are commodity components and this is where the cost savings to a customer come from. Cost savings also come from the lowered staff requirements for the DBA/System Administration roles. The use of commodity components is one important introduction from Netezza.

The architecture is a shared nothing but there is a major twist. The I/O module is placed adjacent to the CPU. The disk is directly attached to the SPU processing module. More importantly, logic is added to the CPU with a Field Programmable Gate Array (FPGA) that performs record selection and projection, processes usually reserved for relatively much later in a query cycle for other systems. The FPGA and CPU are physically connected to the disk drive. This is the real key to Netezza query performance success – filtering at the disk level. This logic, combined with the physical proximity, creates an environment that will move data the least distance to satisfy a query. The SMP host will perform final aggregation and any merge sort required.

Enough logic is currently in the FPGA to make a real difference in the performance of most queries. However, there is still upside with Netezza as more functionality can be added over time to the FPGA.

All tables are striped across all SPUs and no indexes are necessary. Indexes are one of the traditional options that are not provided with Netezza. All queries are highly parallel table scans. Netezza will clearly optimize the larger scans more. Netezza does provide use of highly automated "zone map" and materialized view functionality for fast processing of short and/or tactical queries.

Appliance Maturity

Most customers of Netezza are value purchasers and purchase for the price/performance. Specifically, the low price and the high query performance. This paper, so far, has reviewed some of the relevant points about the Netezza architecture that drive the low cost and high query performance – a peek under the hood if you will. To some shops, it is important to know why Netezza performs. To others, the end results speak all the volumes necessary.

Netezza has put together components that offer a solid price/performance and are solving the data flow problem at a fundamental level. It's the structure, integration and interoperability of the parts that form its platform and value proposition. If you understand the platform is a TCO and performance buy, owning the fastest and most expensive parts are not a requirement.

Netezza has full ODBC tool integration, which does not historically have a positive connotation in regards to performance. Again however, it's the overall performance that counts.

As of the time of this writing, Netezza has a self-reported 30 customers and has shipped over 60 systems.

Customer Success Story

One of those customers, a marketing services provider, is chartered with the management of its clients' data. These clients are in various industries such as financial services, travel, retail, telecommunications and insurance, and they rely on the marketing service provider for CRM analysis and marketing campaigns. The company realized that its current data management infrastructure was not sufficient to support its clients' growing need for information analysis, but was also constrained by the need to keep data center investment and operating costs as low as possible. Many client queries were returning in hours and this was limiting in-depth understanding of customer segmentation and optimization of marketing offers.

The company chose a Netezza 8150 configuration, which accommodates up to 4.5 TB of data, after an evaluation process based on performance, price and openness. The company was able to get their primary BI application, Business Objects, running on Netezza in 2 hours, and the bulk of their existing data was loaded and accessible in 2 days.

In addition to Business Objects, the company uses popular industry tools Ascential for their ETL and Unica, Doubleclick, SAS and MicroStrategy for data access. Netezza allows them to continue to use these tools, which they have developed expertise and code in.

Netezza helped them solve their performance and cost issues. By providing faster reaction to market forces, their clients were able to get their campaigns out the door more quickly, utilizing the data capture and query performance capabilities of Netezza. With better and more iterative analyses, marketing dollars are allocated to the right efforts and therefore allow more money to be allocated to potential high value customers.

Retail banking clients, for example, over the last few years have gone through multiple mergers and acquisitions and that has created larger customer data warehouses and prospecting databases. Netezza allows them to do the customer profiling to quickly learn about the new customers they gain through those acquisitions and what type of programs they then may be able to offer them. They now can take advantage of the processing power of Netezza at a price point that had previously been unobtainable.

The Netezza platform allows them to react to the market forces that they face, i.e., mergers or new competition within a region. It ultimately, again, allows them to allocate their marketing dollars to the right efforts, whether that's acquisition, retention, up-sell or cross-sell.

They have marketing processes that used to take 6 or more hours on a typical relational database that now take about 14 minutes. This allows marketers to do multiple iterations of campaigns. If they get to the end point and they want to go back and tweak some of the inclusion or exclusion to get a different size audience, the Netezza platform allows them to do this multiple times where in the past they could get only one run per day.

As a creative company, they could appreciate the creativeness inherent in the appliance approach. The service provider has characterized Netezza as a "hardware approach to a problem previously tackled only by software…that allows us to meet or exceed all of the business requirements and assist our clients in gaining maximum value for their dollar."

Conclusion

The Netezza data warehouse appliance is a significant architectural advancement over the traditional MPP data warehouse architectures. It is likely to set new standards for data warehousing cost and query performance. Business intelligence environments must get beyond the "personal touch" that is possible when the user community is small. Yet, there are challenges achieving this vision. The biggest challenge is query performance. Managing this issue by limiting user queries to a predefined list or predetermined times is not acceptable. Netezza meets the criteria for data warehouse platform selection and is achieving acceptance due to its value proposition.