

Netezza Corporation

**200 Crossing Boulevard
Framingham, MA 01702
(508) 665-6800
www.netezza.com**

Netezza — An Extreme Analytic Framework in a Box

Preface

All but a few enterprise-scale analytics systems are hitting technology-imposed ceilings. And the business and its information technology (IT) department are paying for these limitations. IT organizations are adding (expensive) database administrators (DBAs) whose sole task is to process a day's worth of incoming data within 24 hours. But the very techniques that enterprises rely on to give analytic solutions the breathing room to handle skyrocketing data volumes constrain the system's — and ultimately the enterprise's — ability to address changing business demands.

Schema optimization and continual tune-ups are two prominent examples. The costs — measured in both time and money — of using these techniques to reorient an analytical solution so that it can handle new dimensions and views are exploding. The costs of being caught in the headlights may be even greater.

To escape this catch-22, enterprises with the greatest speed, depth, and flexibility requirements have invested in a class of technologies — what Aberdeen calls extreme analytic frameworks (EAFs) — to revolutionize the underpinnings of business intelligence architectures. These approaches bring a new set of challenges, particularly in the deployment and user training stage, which must be conquered before EAF software solutions achieve mainstream acceptance.

Netezza, a start-up provider of high-speed, high-volume analytic appliances, has taken on the challenge of mainstreaming the EAF model. Combining software innovations — such as parallel processing — with customized hardware, the company has wrapped the analytical revolution into a single easily managed box.

Executive Summary

At the high-end of the analytics market, the insertion of a mechanism into the database primitives is the clearest sign that an analytical technology has “arrived.” OLAP (online analytical processing) and, most recently, data mining are two

examples. Netezza's NPS 8000 series appliances take a different approach — and take a step further — by imprinting optimization into the low-level building blocks of hardware as well as software. With an innovative use of processing, memory, and storage, Netezza has found ways to transform a number of common instructions into reflex actions and thus avoid the bottlenecks that cripple many relational database management systems (RDBMSes) in high-volume analytics.

The company has not ignored software-driven performance either. Its architecture, combining elements of two different parallel processing methodologies and proprietary techniques within an embedded DBMS, melds together several ways to supercharge analytics. Although performance varies in Netezza's best-of-breed approach depending on usage patterns, the company has demonstrated more than 100-fold improvements in speed over general-purpose servers. For example, in an early pilot deployment of the system, Netezza has been able to shrink the time needed for a common data aggregation task from 12 hours to four minutes.

These innovations would find adherents even if Netezza were only offering a methodology and relying on the end-user or a systems integrator for implementation. But Netezza has lowered the bar to adoption by bundling the methodology in an integrated, standards-compliant appliance. To make the solution more attractive to a broad and diverse audience, the company has made the migration to the Netezza appliances a relatively painless task for any application using Java Database Connectivity (JDBC) or Open Database Connectivity (ODBC).

With price tags ranging from slightly more than \$600,000 to \$2.5 million, the NPS 8000 servers are not for every pocket. But the investment quickly becomes worthwhile, especially for businesses that expend vast sums on DBAs and are searching for a way to squeeze an extra 10% or 20% of performance or an extra feature from an overworked server.

The Optimizer's Dilemma

With data volumes increasing at an unprecedented rate and the pace of business demanding rapid adaptation to change, large enterprises are at an analytic technology crossroads. They face a seemingly irresolvable conflict between speed and flexibility.

The speed problem is the natural result of an increasingly overwhelming volume of information. Raw Web traffic data alone can reach a terabyte per day. Analyzing this and other information floods requires a highly streamlined and structured data flow, with careful tuning for optimal performance to mitigate bottlenecks inherent in the underlying hardware and software configurations.

At the same time, nearly every sizable corporation has seen both new opportunities and new challenges emerge as a result of the e-business explosion of the last several years. Enterprises need flexibility to evaluate and react to opportunities with a

nanced representation of the business (i.e., data model). Compounding the difficulty is the need to deliver discrete views into the model that must reach the screens of multiple audiences within and outside the corporation.

When faced with choosing between speed and flexibility, technologists almost always favor speed. Failing to process a day's worth of data in a day is a highly visible shortcoming, whereas the impact of inflexibility is harder to pinpoint. As a result, the IT department throws raw hardware power, DBA man-hours, and every optimization trick in the book at the task of squeezing an extra few ticks out of its batch cycles. The choice typically leaves the IT staff with few resources available for flexibility — and insufficient capacity to satisfy business users' demands for new types of analytic functionality. Often, the very improvements used to optimize a task “lock in” specific data models and analytic processes, making it costly and perilous to radically shift them — and through them, the enterprise — in any way.

NPS 8000 — Obliterating the Speed Bumps on the Road to Success

Plenty of solutions promise to reduce the burden of the speed-flexibility dilemma, usually via database optimization tools. Netezza takes a dramatically different approach. The company wants to remove the speed issue from the picture altogether. Netezza's analytics-specific appliance delivers performance improvements measured in orders of magnitude for some of the most demanding tasks in analytically centered data processing.

A History of Hardware Bottlenecks

Although an average desktop now has more computing power than a Fortune 500 company of two decades past, multimillion-dollar analytics systems still have performance issues. The reason for this seeming incongruity is simple: Performance improvements cluster around any given technology's dominant dimensions. For example, processors may get twice as fast without doubling the width of the data pipes leading to them or the size of the internal cache. Similarly, the capacity growth of storage devices outpaces incremental improvements in data access speed.

Performance of any system is only as good as that of its weakest component; for most data-intensive applications, the culprit is input/output (I/O) speed. During the process of loading data into memory, pushing it through the processor, and returning it to the disk, the vast majority of time is expended on the first and the last steps. In other words, storage and system interconnects dissipate speed gains of the processing core.

To overcome or at least combat the data-intensive nature of analytical processes, technology suppliers have tackled the I/O bottleneck by using techniques such as:

- Loading entire data sets into memory before performing any operations
- Compressing data into numerical shorthand

- Pre-fetching data during processing lulls

Netezza: Short-Circuiting the I/O Problem

Netezza's primary response to the I/O bottleneck is to short-circuit the storage access process. The company uses commodity components to build its products, but pays special attention to their configuration. This approach is best demonstrated in I/O, where Netezza combines ordinary hard drives and processors into purpose-built devices called snippet processing units (SPUs).

An SPU is an enhanced storage device with an embedded processor and random access memory (RAM) that is primed with algorithms to handle many data operations commonly used in analytics. Netezza performs these operations right on the storage device, so an analytical task that needs data from only one SPU barely pushes any information into the system's central processing and memory resources.

Enhancing with Software Best Practices

Taking advantage of the custom-configured hardware's capabilities, Netezza's proprietary *Asymmetric Massively Parallel Processing* methodology enlists variants of massively parallel processing (MPP) and symmetric multiprocessing (SMP) methodologies to handle different parts of the I/O problem.

Tasks with high locality of reference respond well to MPP, a technique to control multiple processors without shared memory. This methodology takes precedence in data management and security, including locking, logging, and mirroring, as well as record- and field-level procedures, including most join, sort, and aggregate operations. Netezza directs these tasks to the pre-processors located on storage devices in an SPU (Figure 1).

Netezza uses the more familiar SMP techniques for other procedures, particularly for functions in which the final result depends on interaction with the entire data set. With SMP, multiple processors share the same memory resources. Netezza's appliances perform query optimization, plan execution, and some table-level operations in the central SMP host.

The routing and control of these tasks are delegated to the embedded DBMS, aided by Netezza's proprietary *Intelligent Query Streaming* methodology. In addition to utilizing the hardware for maximum analytic throughput, the DBMS incorporates some software optimization techniques. Two examples are caching of recently or frequently accessed data and processing increments of data as they come off the disk, rather than waiting for the whole data set to load. The result is a further speed enhancement targeting complex and frequently run queries.

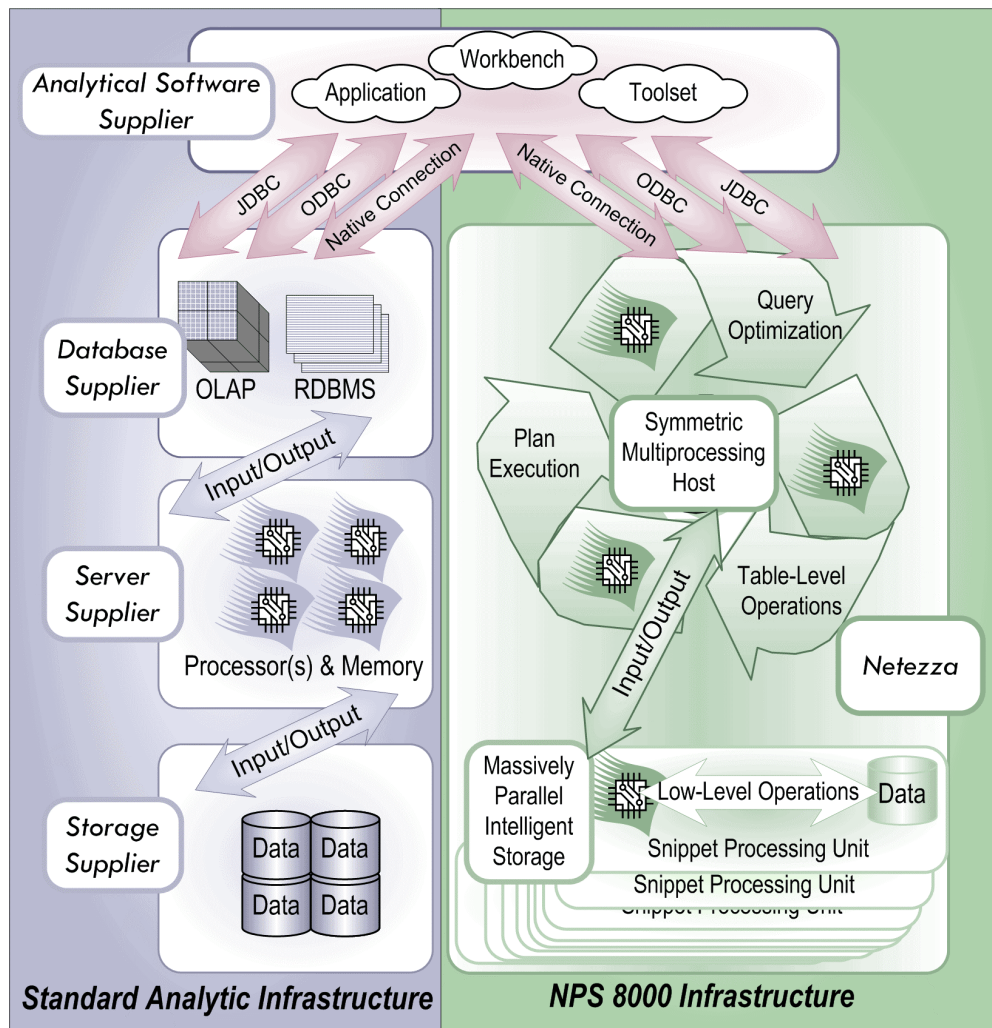
The analytical function is becoming more integral to the day-to-day activities of many businesses, making stability and reliability paramount. Netezza uses Red Hat Linux as its internal operating system and RAID (redundant array of independent

disks) storage techniques to minimize system downtime and data loss. With rack-based hardware, the company can also supply spare components to be kept at the user's site for quick recovery from a physical failure.

Wrapping in a Standards-Compliant Interface

Netezza has been particularly sensitive to — and has taken many steps to obviate — the reluctance among enterprises to abandon their existing DBMS strategies. The wariness is not without just cause. With changes to both the hardware and the software underpinnings of any technology, the IT department has been conditioned to expect a steep learning curve. In the analytical domain, where a well-compensated cadre of experts still do many database administration and tuning

Figure 1. Traditional and NPS 8000-based Analytic Architectures



Source: Aberdeen Group, December 2002

tasks, a radical realignment of technology almost certainly means that the unique talents of the existing staff will go to waste. Furthermore, this devaluation of internal expertise would come at the most vulnerable time — during application migration projects that offer more potential for disaster than any other IT activity. It is no wonder that many enterprises would rather stay with the familiar framework than venture down the path of radical DBMS change.

To overcome this hesitancy, Netezza has obscured the internal processing logic of the NPS 8000 behind standards-compliant external interfaces. As a result, if the enterprise wants to treat the appliance as a generic JDBC-, ODBC-, or SQL-compliant DBMS, it can do so without foregoing many of the speed benefits of the underlying architecture.

As DBAs continue working with the NPS 8000, they will inevitably want to tweak the architecture-specific internal processes to further improve performance. However, during the initial transition period, their existing knowledge will be sufficient for smoothing out the inevitable bumps. The bottom line: the enterprise does not have to take a step back before it leaps forward, making the transition faster, easier, and nearly risk free.

Packaging for a Broad Audience

The importance that Netezza places on fast deployment is underscored by its choice of the appliance as the company's sole deliverable product. By controlling the entirety of the low-level analytical architecture, Netezza ensures that all the kinks in the interaction between the DBMS and hardware are worked out prior to deployment, and the user's sole burden is deploying the application on a sturdy platform. By minimizing its professional service involvement, the company imposes the discipline of getting it right the first time around on its own engineers.

Not all enterprises are created equal, and not all have equivalent speed, depth, and dynamism requirements. Recognizing this fact, Netezza has avoided the "one-size-fits-all" pitfall that occasionally troubles appliance providers. The NPS 8000 series covers three sizes. The 8100 is optimized for roughly 1.5 terabytes of user data; the 8400 quadruples the capacity to 6 terabytes without loss of speed; and the 8200 server occupies the middle ground, handling 3 terabytes of user data. Furthermore, the architecture has shown near-linear scalability at these multi-terabyte data sizes, ensuring that an enterprise that outgrows even the largest server will have a viable upgrade path.

A Start-up with Swagger

Netezza initially launched the NPS 8000 series in September 2002, having spent most of the preceding months working stealthily. Yet the company is far from a typical start-up. With the former founder and CEO of Applix at the helm and the ex-president of Sun Microsystems on its board, Netezza has put together an impressive

management team. Venture firms such as Matrix Partners, Charles River Ventures, and Battery Ventures have pulled together some \$28 million to fund the company, with the bulk of the cash coming in early 2002 — an anomaly in a time noted for shut-tight venture wallets.

None of these factors is an absolute guarantee of future success. However, the early returns have largely been favorable. With a growing number of analytic solution partners, including such prominent names as MicroStrategy, SPSS, and Unica, Netezza is pushing even further toward plug-and-play analytics. MicroStrategy has taken the boldest step and formally certified Netezza as a hardware option for its customers. Partnerships with these analytics specialists, as well as systems integrators, have seeded interest in many high data volume sectors, such as financial services and telecommunications. These early deployments have produced results that are sure to be envied by industry peers.

Vibrant, one of Netezza's first customers, cited an example of processing billions of call detail records — a task that typically took more than 36 hours using a Sun- and Sybase-based platform — in just 26 minutes using the NPS 8000. Another pilot deployment in financial services also showed extraordinary improvement:

- A data merge task on 250 million records ran in four minutes on Netezza, compared with 11 hours, 53 minutes on a Sun e5500.
- A reporting task on a subset of that data took 40 seconds on Netezza's platform, compared with five hours and 13 minutes on the Sun platform.

Other industries are catching on as well. For example, Netezza's sales force is targeting analytical and marketing services providers, which can use the company's offerings to avoid being overwhelmed by their own success and ensure that speed is not an obstacle to developing new services. One of the early adopters in this field, Epsilon, has already improved the time for a 250-gigabyte data load task from 11 hours using Oracle 8i on a Sun e6500 to three hours on the NPS 8000.

Aberdeen Conclusions

The best reason to buy into a low-latency analytic architecture is not system failure, but the expenditures associated with preventing that failure. Many organizations have displayed unquestionable skill and dedication in holding back the overwhelming tide of data, and they will continue pouring that energy into traditional analytic architectures for quite some time. But emerging analytic applications are starting to tilt the labor-to-benefit balance of traditional RDBMS architectures.

From another perspective, if the same energy is dedicated to satisfying emerging demands from users, the enterprise will become significantly more agile and responsive to changing business conditions. A product like Netezza's NPS 8000

server, which radically speeds up the data processing and analysis functions of a large enterprise, frees up scarce IT resources for tasks that really matter.

The speed increments that the NPS 8000 offers, while certainly impressive, are not necessarily unique. A variety of software-only approaches to radically accelerating data processing have sprung up in the past few years. These extreme analytical frameworks typically get large speed gains by extending or replacing the relational database as the foundation of analytical architectures. Many even match comparable relational databases on base price. However, the training, deployment, and maintenance costs associated with them are often uncertain, giving an appliance provider like Netezza room to claim an advantage in total cost of ownership (TCO).

The biggest challenge to Netezza's success is establishing credibility. Put bluntly, extraordinary claims from any start-up require extraordinary proof. On the technological side, this typically means formal benchmarks or at least a wide range of customer testimonials — anecdotal evidence can rarely buttress a “speeds-and-feeds” message adequately. On the business side, the company must track the real TCO of its own and competing products on an ongoing basis to make sure that it can justify its price tag even as the competitive field continues to get more crowded.

In the end, Netezza will succeed if, by holding back the deluge of data, it helps IT organizations offer better service to their business customers. There is no guarantee that data volumes will not eventually catch up with the performance gains the company offers. However, for the foreseeable future, an enterprise that has spent its scarce IT resources to optimize analytical processes such as data merging and aggregation can use the breathing room afforded by the NPS 8000 to refocus on answering new challenges from the business user.

To provide us with your feedback on this research, please go to www.aberdeen.com/feedback.

*Aberdeen Group, Inc.
260 Franklin Street, Suite 1700
Boston, Massachusetts
02110-3112
USA*

*Telephone: 617 723 7890
Fax: 617 723 7897
www.aberdeen.com*

*© 2002 Aberdeen Group, Inc.
All rights reserved
December 2002*

Aberdeen Group is a computer and communications research and consulting organization closely monitoring enterprise-user needs, technological changes and market developments.

Based on a comprehensive analytical framework, Aberdeen provides fresh insights into the future of computing and networking and the implications for users and the industry.

Aberdeen Group performs projects for a select group of domestic and international clients requiring strategic and tactical advice and hard answers on how to manage computer and communications technology. This document is the result of independent research initiated and performed by Aberdeen Group. Aberdeen Group believes its findings are objective and represent the best analysis available at the time of publication.