

恩
門

BILLINMON.COM

NETEZZA: NEW ARCHITECTURE RISING!

By

W. H. Inmon and R.H. Terdeman

DECEMBER 2002

WWW.BILLINMON.COM

Introduction..... 3

A Short Course in Database History..... 3

Teradata and Its Architecture..... 4

Technical Challenges that Set the Stage for a New Player: Netezza..... 5

A New Paradigm: Two-Tiered Architecture within the Hardware..... 5

The Appliance Model and Netezza..... 7

Bandwidth and the Netezza Performance Server System..... 7

The Netezza Performance Server System and Storage..... 8

New Horizons in Information Magnitude..... 8

Netezza Performance Server System as a Foundation Architectural..... 9

Netezza as a Technology Replacement Strategy 9

Acquisition Pricing 10

Total Cost of Ownership..... 11

Conclusions..... 11

Introduction

Occasionally a new product comes along which represents a breakthrough in architectural thinking. The Netezza Performance Server™ system represents this type of product. Targeted to the large-scale data warehousing DSS environment, Netezza addresses design limitations that have challenged existing products. These design challenges evolved from historical limitations in technology or ways of thinking about long-standing problems. Real advancement in technology only comes when a new paradigm arises. These new paradigms of technology, contrary to popular thought, are evolutionary not revolutionary. They occur when a company or individual is able to synthesize a new idea with ongoing known problems in order to come up with a new approach that is significantly better than the previous approaches, and provides better value. This is exactly what Netezza has done.

A Short Course in Database History

Once upon a time the first databases were designed to serve On Line Transactional Processing (OLTP) needs. These databases, sensitive to online performance needs, were developed in the time of the mainframe. In the days when the mainframe was king (some still think it is!) the financial equation of technology was that machines were very expensive but people (programmers) were cheap. With that mindset, the proper approach to design was to optimize machine time. Optimizing machine time was accomplished by using the fewest number of cycles in order to process the largest number of transactions. The thought of buying a second mainframe represented a real financial obstacle, except for the biggest of institutions.

Over the next few decades there were a few attempts to use the mainframe outside of the traditional purview. In these cases, the mainframe was used for intelligent decision-making but it was almost exclusively reserved for addressing problems like high-energy nuclear physics or academically funded research projects. For the average enterprise, using the mainframe for intelligent decision-making was simply out of the question.

Then two key breakthroughs occurred which changed the paradigm of computing. The first key event was the appearance of the mid-range system and the subsequent evolution of what was first known as the microcomputer. The second key event was the advent of decision support theory in the form of data warehousing and then ultimately in the framework known as the 'Corporate Information Factory'. Simultaneously, the arrival of cheap CPU cycles enabled firms (for the first time) to apply computer technology affordably to business problems beyond basic financial business functions run on transactional processing systems. The advent of data warehousing and the extension to the Corporate Information Factory gave companies a way to harness the information to run the business instead of running the business with the information available. Thus, at least in theory, new doors were opened with the arrival of data warehousing.

But there was a problem. While business people knew what they wanted, they still did not have an 'engine' that could deliver on the promise of data warehousing. The Codasyl databases of the time were still based upon procedural language approaches from the mainframe world and did not have the ability to handle the data sizes or the workloads required

to meet the growing volume of business data. When data warehousing first appeared, database management systems still fell into only three categories: hierarchical, inverted list or the newly successful network structure. While there was some new database theory available in the form of Date and Codd's relational database, that theory was in its infancy and there were no practical business implementations.

In the early years of this market, IBM's attempts at positioning DB2 were not as a decision support engine but as a replacement for its aging IMS database and its VSAM file access method. Initially DB2 was solely an overlay product to the VSAM file system that offered a relational structure. DB2 was targeted to the OLTP environment, not the decision support or data warehousing environment. There was enough energy and innovation in the air for a new synthesis to occur and thus a new paradigm and a new 'King Of The Hill'.

Teradata and Its Architecture

About 20 years ago, a band of ex-Citibank employees created an experiment in technology. The company and its experimental product were known as Teradata. This experiment involved linking the new 8080 based personal computers together via a common bus architecture call the 'ynet'. The ynet allowed these computers to asymmetrically approach computation. However, as innovative as the hardware was, there was no known database at the time capable of operating in such an environment. In addition, while there were standard disks that were available for these machines, the disks were small and operated on each of the parallel machines in isolated fashion. Further, these new SCSI disks were in their early infancy and were clearly subject to rapid evolution.

In order to link these disparate and newly emerging technologies together, all the components had to be integrated. First, a database management system was written to operate asynchronously in this new world. Second, recognizing that many small computers were needed, a hash-based approach was used to spread the data evenly over many engines. Finally, an operating system was written that used the high-speed proprietary merge capability of the ynet and efficiently managed the space available. This proprietary operating system became known as the Teradata Operating System (TOS).

Over time TOS was in some respects Teradata's biggest asset and greatest liability. It was Teradata's biggest asset in that it made the whole system work and thus become saleable. But at the same time it was Teradata's greatest liability in that it was proprietary and had endless possible conditions that lead to 'restarts'. These restarts (which were acceptable in the beginning of the life of the product) became intolerable as the machines began to take on their role as the platform of choice for high-end, large data warehouses. The more mission critical Teradata became, the less acceptable were the restarts.

In the mid 1990s Teradata was acquired by NCR. Cash poor Teradata had found a deep pocket partner to further develop the technology. But further technological developments made sense only if robustness issues with the software could be resolved. The solution to the TOS problem of being a bottomless money pit was to adopt the NCR flavor of UNIX, namely UNIX V.4. It should be clearly noted by this time that UNIX V.4 was far from the most popular version of UNIX. Sun had already captured major market share with Solaris, HP had

propagated UX, IBM was at its infancy with AIX and finally there was still a happy SCO community. NCR caused the untimely disappearance of the true MPP version of Teradata and replaced it with a hybrid base of standard NCR SMP nodes connected by a new faster Bynet. This new configuration caused the prior symmetrical configuration of MPP Teradata systems to migrate into a non-symmetrical configuration requiring a separate parsing engine. Thus, the simplicity of the original Teradata design was lost in a complex series of requirements, each brought on by the application of increasing levels of software complexity in order to assure reliability.

It is much easier to understand a new architecture if an existent architecture is used as a baseline. In the Teradata software an SQL request is taken apart by a program called a parser. The decomposed SQL request is broken into steps, which are recognized by the software running on the database control portion of Teradata historically called AMPs (Access Module Processors). These software processors run on the SMP hardware units. The processing steps are then run in parallel, where appropriate, by a piece of software called the Optimizer. The Optimizer applies a rule base to determine the best way to execute the steps. There are various optimization approaches but rule-based or performance-based are the primary approaches. It is important to understand that regardless of approach, all database components in the Teradata design are executed at the SMP level on all processors simultaneously. Thus if a step is to obtain raw space for a database insert, this is done at the SMP level and on every SMP processor. Whether the step is to create a temp file, delete a row, restructure a row or obtain a sector, every function is done at the high level of the SMP processor.

Technical Challenges that Set the Stage for a New Player: Netezza

At the heart of technical innovation is always a problem in search of a solution. There are four classic problems related to large-scale architecture that have been at the heart of much of the technical innovation in the last decade. The first is the need for a high performance, low overhead operating system. The second problem is the need for a technical solution that can scale upward with the growth of the business without causing disruptive change. The third is for very high reliability. Reliability requirements start at the electrical plug in the wall and extend to every component in the architecture. Finally, there is a need to provide adequate data movement capacity so as not to constrict the business usage of the information.

A New Paradigm: Two-Tiered Architecture within the Hardware

When one examines all existing data warehousing architectures they can best be described as one-tier. Even if there is special hardware to perform a function like parsing, they are of the same design as the database nodes in terms of hardware, and operate on a peer-to-peer basis. The Netezza architecture is radically different. It is two-tiered and uses standard SMP Linux nodes to provide the higher-level functions such as SQL compilation, Query Planning, Optimization and Administration. **All lower level functions such as: record operations, field operations, locking, logging and storage management are preformed by intelligent storage controllers known as Snippet Processing Units (SPUs).**

This two-tiered architecture is unique in the industry and accounts for the Netezza Performance Server system's outstanding performance characteristics. By joining the SPUs to their SMP hosts in a gigabit network, the matrix scales to a power, speed and size potential unmatched in the industry. Never before has a database been designed to specifically spread in a vertically integrated hardware matrix as well as horizontally across components. Why then is this architecture so important?

The first and most obvious advantage of the Netezza Performance Server system is that it is a server specifically designed to meet the challenges of large scale data warehousing with low overhead. Netezza starts with the choice of an operating system with the least overhead and yielding the most relative throughput, and that is LINUX. It is an established fact that each generation of operating system represents progress over its antecedent. The errors made on the mainframe operating system were corrected in the mid range by UNIX, NT and LINUX operating systems. Every major financial institution in New York has a LINUX initiative. While LINUX may have its limitations, it has successfully driven down the core cost of an operating system. This is particularly important in times when enterprises are attempting to reduce or at least limit the cost of their operating system overhead.

There is another important reason why the choice of LINUX as the operating system makes sense. The specific Linux flavor of software chosen by Netezza is Red Hat, which holds the dominant position in the market place. The choice of Red Hat by Netezza as the operating system of choice is a very savvy move. Netezza has offloaded the associated operating system maintenance and improvement costs on a provider who is the dominant vendor in the Linux space and has specialized in this and only this operating system. Unlike Teradata, which first wrote its own operating system and then did a forced retrofit, Netezza has gone with 'the brightest and the best' from the outset. But there is another reason why Red Hat Linux was a clever choice – Red Hat as an operating system fits well with the appliance hardware model. The appliance model is based on a conception that is robust and easy to use. Red Hat meets those criteria.

The second problem addressed in the Netezza Performance Server system is that of scalability. In order to scale, Netezza's Asymmetric Massively Parallel™ (AMPP™) architecture uses modular components at two architectural levels. At the higher level, the components are standard rack-mounted SMP units. At the lower level are a series of intelligent controllers and redundant components that include disk, controllers and power making for a package that is small, easily managed and extended.

The third problem addressed by Netezza - the need for reliability - is accomplished by full redundancy at every level. Finally, adequate data transfer capacity is assured between all components in the architecture by Gigabit switching technology.

The Netezza two-tiered AMPP architecture is very different from the traditional Teradata one-tiered architecture. In the Teradata architecture there is a Parsing Engine. The Teradata Parsing Engine decomposes the query that is passed to it into multiple execution steps, which are then executed in parallel on all SMP nodes. In essence, the parsing engine exists in a peer relationship to the execution engines all on the same SMP architecture. In Netezza's AMPP architecture, an SMP host connects to hundreds of MPP drives called Snippet Processing Units (SPUs). It is at this level that activities, which should be done at the physical

database level, are performed. This movement of processing into the hardware allows for processing decomposition into many more processors than any other technology, including Teradata, can provide. Further, Netezza's Intelligent Query Streaming™ technology takes full advantage of this two dimensional two-tiered architecture, thereby enabling an order of magnitude improvement in performance over even Teradata, without needing additional human capital. All of these advancements create an attractive value proposition.

The Appliance Model and Netezza

The term 'appliance' is one of the most misused (and abused!) terms in all of information technology. There have been hardware and software appliances. 'Appliance' has been used to describe anything from a data storage solution on a network to a device that turns on lights at home. The basic concept behind an appliance is to produce a modular unit that provides immediate functionality without a great deal of work or overhead. This appears to be at the very core of the Netezza Performance Server system design. The concept behind the design is a modularity of hardware that is rack-based rather than node-based.

Rack-based modularity allows for small 'bricks' of scaling rather than the ten-ton 'cinder blocks' that were required for upgrade strategy in the mainframe world as well as in the early Teradata architecture. What we have with Netezza is a very maintainable computing framework, which does not require rocket scientists for maintenance.

Bandwidth and the Netezza Performance Server System

Most aficionados of data warehousing know that there are only three problems at the heart of data warehousing infrastructure. Those three problems are: horsepower, bandwidth, and storage. The previous section has discussed how Netezza has solved the adequacy of horsepower issue through its appliance design. The issue of bandwidth communication between processing units is handled by the use of gigabit Ethernet switch technology. When multiprocessing technology became available for relational database use, the only way to communicate data was by channel technology on the mainframe or the still evolving early Ethernet technology.

The Teradata ynet and bynet technology has a glorified TCP/IP underpinning. While the bandwidth has improved going from wire-based structures to fiber-based technology, the fact remains that in the realm of massive data movement, linear and ring-based architectures for communication are subject to bus saturation. The switch technology that has recently arrived has found two uses. The first use of switching technology is in the world of Ethernet LAN applications as expected. The second use was brought on by the creation of a new entity called 'enterprise storage'. The ability to share a vast amount of storage among many processors led to the need for 'non-blocking' switch technology to assure optimization of the resource. While this switch technology for storage has become wide spread among the largest firms, it is only now taking root among the slightly smaller enterprises.

Netezza has integrated this new technology into its design to increase the bandwidth between processors by at least an order of magnitude compared to all other competitors. This clearly

gives Netezza a design advantage by using an emerging technology, which is only at the beginning of its developmental cycle. Why is this important? In the early days of SCSI (small computer standard interfaces) the prevailing data rate of transfer was 10 MB/second. As the technology matured the rate went from 10 to 20 to 40 to 80 and finally 160 MB/second. This technology has 'maxed out'. The importance of a 'new' technology is that its baseline is established in its early release. Thus gigabit technology was introduced at 100 MB/second. The 200 MB/second upgrades are already available. It is anticipated that this technology will continue on the same growth curve as the SCSI technology, providing clear benefits as data sizes continue to grow.

Now let's turn our attention to data warehousing and the final piece of the puzzle - storage.

The Netezza Performance Server System and Storage

In the early days of data warehousing, storage was an afterthought. While a lot of storage was necessary (in orders of magnitude never seen before in a single system), what was used for the data warehouse was what was available. In the beginning, there were a mix of controller technologies CMD, IPI, SMD and eventually SCSI. The controllers were initially 'dumb' controllers. Dumb controllers buffered tracks and did data management basics; but they had no advanced functionality. The modern RAID controller came after the fact and only when it became clear that the disk was the weakest link in the computing chain. About the same time, another company developed functionality affecting disk integrity and reliability but not in the disk controller. Instead the functionality was added as a combination of micro code loaded to a memory buffer on a custom controller with a memory buffer. This micro code running in the controller cache plus a large global memory in the disk storage bank created 'fast disk'.

This fast disk did sequential pre-fetch to take down access time into the sub-second response and added many enhanced functions like replication in storage but did little to change the role of the disk controller as an off-the-shelf device. The controller was, after all, just a controller with memory added and the real work was done in a micro code routine, which made storage even better. From an application standpoint, no radically new functionality was at hand. Netezza has implemented a new architecture that elevates the role of the controller from a relatively dumb device into a device which partners with the SMP platform to disperse the functions normally performed by the SMP host alone. This is the key Netezza breakthrough.

New Horizons in Information Magnitude

In 1993, less than 10 years ago, the largest known Oracle database was 80 Gigabytes and ran precariously. At the same time the largest known Teradata machine (in spite of its name) had barely cleared the 300 Gigabyte threshold. How things have changed! Today the mega databases do not even begin until the 10 Terabyte and up range is reached. In the 1 to 30 Terabyte range, IBM, Oracle and Teradata are all players. However, recent world events have taught us that for many reasons, it is necessary to carry much larger amounts of data, perhaps into the Petabyte range, in a single store. It is questionable whether any of the older

architectures are capable of running in that size range. Issues of latency in the bus architecture make it questionable. Also, the older architectures take long periods of time to integrate the fastest processors.

There is no such question in the case of Netezza's architecture. Faster switch and chips can be easily tested and integrated as they arrive upon the scene. Most of the components are generic, leveraging 'best of breed' technology. Like any upstart player, there are not a lot of legacy problems to be addressed. Therefore the full thrust of the company is to provide a forward-looking architecture. How then do you position the Netezza Performance Server system in a bewildering array of solutions?

Netezza Performance Server System as a Foundation Architectural

If you are looking for a panacea plug-in solution including the application, Netezza is not it. Unlike the Teradata approach, Netezza does not claim to be expert in retail market segmentations or in Telco call detail record analysis. Netezza supplies what best can be described as emerging foundation architecture. This architecture will flow well with your Information Services organization. It is an architecture that will scale with the business. It is architecture that will decrease in cost, relative to power, over time. It is also an architecture that fits into the framework of most existing IS organizations. In short, as a solution it will allow the user community and the Information Services organization to walk hand-in-hand in terms of requesting funding. This has not been the case in the past where certain solutions vendors backed the Information Services organizations in to a corner with a black box they could not easily manage or understand and flowed against their basic skills and knowledge base. Further, the cost of the 'high end' solution has always been an issue with the IS organization. The Netezza Performance Server system minimizes the need for data movement by keeping the low-level technology activities where they belong - on the SPU unit. At the same time the Netezza Performance Server solution frees more bandwidth at the SPU level for use in the query process. In short, it appears that with this approach, you get the same horsepower you are currently using for a lot less cost, or as an alternative you enable a lot more computing at the current cost. Any way you approach the problem, it is a win-win solution for both the IS department and the end user community.

Netezza as a Technology Replacement Strategy

Every technology architect's dream is to build a data warehousing architecture from the ground up. However, in this day and age this rarely happens. More frequently existing architectures reach limits of reliability, scalability and usability. This is a known issue with the various pure software database solutions, most notably Oracle, Informix and Sybase. While all have demonstrated very large single installations in the multi-Terabyte range, little is said about the pain in keeping these very large instances up and working. Almost all software database vendors require multiple software products in their software suites to build very large stores. While they would lead the customer to believe that their initial purchase of the software would then result in an easy upward scalability, this is far from the case. In the case of Informix, it meant a move to Informix XPS. In the case of Oracle, it was OPS that became a dreaded obstacle to overcome. And finally, Sybase has its IQ product. In each case there is a

requirement to add expertise and training in order to implement and support the higher level of product.

Further, in virtually every case, the cost of human capital to support the product far exceeds the cost of purchasing the product. In short, the choice becomes one of higher entry point cost but dramatically lower TCO or lower entry point and higher TCO. As knowledge about the real cost of information technology has spread, virtually all institutions that require large data have opted for the lower TCO models. Today TCO is widely available both from independent analysts as well as the vendors themselves. Clearly, Netezza has TCO as part of its strategy. So if the organization is faced with too many systems administrators supporting too many platforms with too many DBAs, the Netezza Performance Server system is worth serious consideration as a replacement strategy.

Acquisition Pricing

Based on Netezza's list prices, the pricing of the product appears to be about 1/3 of the prevailing cost of the high-end provider, Teradata. The published price of a fully configured Netezza Performance Server 8100 is \$622,000 for 4.5 TB of storage and complete reliability features. This number is a pure acquisition number based on data size and not on throughput. The numbers based on a pure throughput basis appear to be more along the lines of an order of magnitude improvement (**10x**) based upon the available data. This opens up the realm of VLDB data warehousing applications that up to now have been unable to even begin the process of developing this type of business capability. Thus, many mid-size companies with large decision support needs, or departments of large enterprises can afford to start with a relatively small implementation without the risk of a multimillion-dollar hardware commitment so frequently associated with data warehousing in the past. Netezza's return to a focus on 'the database' means that the Netezza Performance Server 8000 Series becomes an appliance as intended rather than a career path as so frequently occurs with other vendors. The developer can stay focused on the business needs and he can design to specification as opposed to force fitting the business requirements to the limits of the solution as provided by various vendors.

The pricing is very important from another business aspect, namely risk mitigation. ***For the cost of the high-end solutions, a firm could purchase two full Netezza Performance Server systems and still save almost 1/3 of their total expenditure!*** In this time of strained resources this issue is critical. Many of the old arguments for the status quo architecture do not hold. In a post 911 world, it means that high-end firms can afford a primary system and a business continuance system in the same budget and still save almost 1/3! Mid-tier firms can afford to design their emerging data warehousing application in business continuance mode from the beginning of the project and not as an afterthought.

Total Cost of Ownership

One of the earliest claims of the industry leader, Teradata, was that from a personnel cost standpoint, their product had the lowest Total Cost of Ownership (TCO) based upon the need for fewer individuals. This savings was based upon the database performing functions that were previously labor-intensive manual tasks. As the leader moved from a technology vendor to a total solutions provider, this claim was dropped. In fact, the multi-tiered software of Unix V.4, RAS and then Teradata led to a rather sophisticated support matrix. So sophisticated that virtually all of the large sites now require dedicated staff from both the customer and the vendor full time.

While business partners are a good thing, most organizations prefer to have some limits to the relationship and not live in a 'shot gun' marriage arrangement. Netezza's design assures that the NPS system provides technical capability without massive personnel overhead. Once again, this is an appliance approach, not an end in itself. Ironically, Netezza has a better claim on lowest personnel overhead than Teradata. The Netezza architecture shows no pretension to any other deliverable than a high performance, reliable appliance platform to be used at the discretion of the organization.

Conclusions

There are a series of 'take away' conclusions and recommendations.

- First, Netezza delivers just what it promises, a scalable very large database solution.
- Second, the solution is a net new synthesis of technology and not just another 'me too' solution.
- Third, Netezza is an excellent choice for the mid-tier business unit or enterprise that has enough technical skill to manage the application deliverable but needs a robust, scalable engine.
- Fourth, Netezza is an alternative for the very large-scale database (VLDB) customer who needs costs to be brought under control where data growth has subsided, but business issues require enhanced functionality.
- Finally, Netezza represents an opportunity to architect into an organization an 'appliance' approach to databases as opposed to building a massive knowledge base, which has become a significant overhead for information-intensive organizations.

It is worth the time and effort to evaluate Netezza's Performance Server system as a core component in infrastructure to support the Corporate Information Factory model.