NETEZZA

The Power to Question Everything™

Unleashing Enterprise Business Intelligence in an Era of Affordable Computing Resources

Whitepaper

1 866 NETEZZA

WWW.NETEZZA.COM

NOVEMBER 2005

About the Author

Neil Raden is the founder of Hired Brains, Inc., www.hiredbrains.com. Hired Brains provides consulting, systems integration and implementation services for clients across North America and Europe. Hired Brains also provides consulting, market research and advisory services to the Business Intelligence, Data Warehousing and Decision Support industry. Based in Santa Barbara, CA, Neil is an active consultant and widely published author and speaker on data warehousing, business intelligence and information technology strategy. He welcomes your comments at nraden@hiredbrains.com.

Executive Summary

Well into its second decade as a mainstream concept, data warehousing has grown into a sizeable industry but, in many ways, it has not lived up to its promise. Only a small percentage of knowledge workers[1] in organizations use Business Intelligence (BI) tools as a regular part of their work and a large portion of analytical work to support decision-making is still done elsewhere.

Existing tools and methodologies are not leveraging the latent value of data warehousing investments because they are not:

- Current enough
- Flexible enough
- Useful enough

To a large extent, these problems derive from an approach to data warehousing that predates many innovations we take for granted today. Complex designs, rigid models and restricted access are enduring artifacts, not unlike the missing century digits in the Y2K problem, from a time when computing resources were drastically more limited and more expensive.

Economic realities are ushering in a new era in Enterprise Business Intelligence, disseminating analytical processing throughout and beyond the organization.

Economic realities are ushering in a new era in Enterprise Business Intelligence, disseminating analytical processing throughout and beyond the organization, not just to a handful of analysts. With the accumulated effect of Moore's Law[2], the relentless doubling of computing and storage capacity[3] every two years, with a corresponding drop in prices, plus engineering innovation and Linux-based and open-source software, a redrawing of the data warehouse model is needed.

Current "best practices" in data warehousing call for a multi-level structure, with raw data extracted from sources, an "enterprise" data warehouse of very fine-grained data, transformed, integrated and cleansed, many smaller, summarized schema, data marts and others[4]. Updating all of these separate structures takes time, leading to high levels of latency (not current enough). Designing and maintaining all of these structures creates an instant maintenance lag as even small modifications require extensive research, modification and testing of dependent structures (not flexible enough). Forcing users into aggregated data solves the performance problem, at the expense of data depth (not useful enough). Aggregating or "cubing" data is like reducing the resolution of a picture – the detail is lost forever. This approach is a relic from the days of precious computing resources. Detailed, transactional, even sub-transactional data is endowed with rich qualities that have to be exploited. Roping off the "enterprise" data warehouse is no longer acceptable in an era of Enterprise BI.

Detailed, transactional, even sub-transactional data is endowed with rich qualities that have to be exploited.

A great deal of important analysis requires access to very fine-grained data. Even though the questions that people have can be phrased rather simply, and the results they view may be very condensed, the detail needed to answer those questions cannot be simplified and summarized in advance. Solutions that cannot leverage all of the data drain value from an implementation. When business people see that data warehouse and BI tools are not capable of solving their end-to-end problems, they revert to spreadsheets to do their analytic work. Despite the accepted view that spreadsheets are not adequate for enterprise BI, data warehousing has failed to slow spreadsheet propagation. BI is currently too expensive, too limited in scope and too rigid.

The economics of computing are not reflected in current data warehousing best practices. The solution to problems of latency and depth of data is the application of the vast computing and storage resources now available at reasonable prices. Due to Moore's Law and innovations in query processing techniques, workarounds and limitations that have become "best practices" can be abandoned. Forward-looking companies like Netezza are positioned to provide the kinds of capacity, easy scalability, affordability (both purchase cost and TCO[5]) and simplicity that allow for broadly inclusive environments with far greater flexibility. Tedious design work and endless optimization that slow a project can be eliminated with new data warehouse "appliances", at a much lower cost.

In addition to servicing the reporting and analysis needs of business people, data warehouses will soon be positioned to support a new class of applications - unattended decision support. So-called hybrid or composite systems, that mix operational and analytical processing in real-time are already in limited production and will become mainstream within the next few years. Most BI architectures today are not equipped to meet these needs.

Making data warehouses more current, flexible and useful will lead to much greater propagation of BI, improving comprehension and retention of business information, facilitating collaboration and increasing the depth and breadth of business intelligence.

Introduction

Bringing data warehouse practices up to speed with the state of technology requires a reversal in polarity: the focus of data warehousing is still solidly on data and databases and needs to shift to business users and practices that need its functionality. The preponderance of effort in delivering BI solutions today is still highly technical and dominated by technologists. For the effort to be relevant, the focus and control must be turned over to the stakeholders. This is impossible with current best practices.

Unlike systems designed to support business processes such as supply chain management, fulfillment, accounting or even human resources, where the business processes can be clearly defined (even though it may be arduous), Business Intelligence (BI) isn't really a business process at all. Traditional systems architecture and development methodologies presuppose that a distinct set of requirements can be discovered at the beginning of an implementation. Asking knowledge workers to describe their needs and "requirements" for BI is often a fruitless effort because the most important needs are usually unmet and largely unknown. Thinking out the box is extremely difficult in this context. The entire process of "requirements gathering" in technology initiatives, and BI in particular, is questionable because the potential beneficiaries of the new and improved capabilities are operating under a mental model formed by the existing technology.

Current best practices are no longer necessary given the entrance of players like Netezza.

When it comes to BI, the industry is largely constrained by this drag of technology. What passes as acceptable BI in organizations today is rarely much more than re-platforming reports and queries that are ten- to twenty-years old. In the typical requirements gathering effort, people who have been laboring in this information-poor environment fail to mention the most basic and important informational needs they have. Requirements in BI are accretive, they emerge over time and it is the people in organizations who have not been exposed to BI who have the most expansive view of their informational needs and are capable of visualizing a closed loop of query, analysis and action that is not only possible with today's technology, it is absolutely critical.

For data warehousing and BI to truly pay off in organizations, IT needs to shift its focus from deciding the informational needs of the organization through technical architecture and discipline, to one of responding to those needs as quickly as they arise and change. Current practices lead to rigid structures and exclusion, measures taken to control the development at a time when needs far outstripped resources, practices that are no longer necessary given the entrance of players like Netezza that exercise the real economics of commodity hardware and Linux-based and open-source software.

The Challenge in Data Warehousing

As a maturing technology, data warehousing is subject to the drag of "best practices." Though derived of good intentions, the industry's focus on data instead of solutions has actually become a barrier to good solutions on the front end. It's a classic case of the tail wagging the dog – the data architecture that was meant to serve the needs of BI is now the limiting function in what can be done. Massive data warehouses are built at great expense, using costly hardware and software, but to protect their performance, summarized "data marts" are extracted for BI purposes, limiting the depth, breadth and usefulness of the data warehouse. In other words, many data warehouses are built that are too large to be queried. What this really means is that, in a given configuration, the database servers cannot efficiently process analytical queries against the large databases they can build. Adding more computing power is an expensive proposition because the hardware and database software are proprietary offerings that only reluctantly respond to continually declining prices according to Moore's Law. In many cases, the hardware is "maxed-out" and cannot be extended at any price, but instead requires a completely new platform. For all of these reasons, the seemingly inexplicable restriction of access to data warehouses is widespread.

"Best practices" today call for careful modeling of the relational database schema first, a process that is as fluid as poured concrete – fluid until set. It adds instant inflexibility to the BI landscape. All subsequent development work of populating the data warehouse, (including multiple steps of downstream schemas, "cubes," extracts, etc. and development of BI applications), references these physical schemas directly. This creates an immediate maintenance burden and resistance to modifications. Historically, developers could point to the prohibitive costs of managing such large volumes of data as the driving need for so much handwork and structure, but that is no longer the case in today's market.

One thing organizations learned in the business process re-engineering of the nineties is that adding steps to a process adds cost, complexity, latency and error. By separating knowledge workers from the analytical models they work with by two or three layers of complexity (requirements gatherers, data modelers, DBA's), current data warehousing best practices drain value from the effort at the beginning, middle and, ultimately, end results of the initiative.

Expensive to Build and Maintain

The reality today is that moderately priced, high-performance hardware is widely available. The cost of CPUs, memory and storage devices have plummeted in the past five to ten years, but data warehouse design, completely out of step with market realities, is still based on getting extreme optimization from expensive and proprietary hardware. A great deal of the effort involved in data modeling a data warehouse is performance-related. In some popular methodologies, six or seven different schema types are employed, with as many as twenty or thirty or more separate schemas.

This has the effect of providing adequate query performance for mixed workloads, but the cost is limited access to the data and extreme inflexibility borne of the complexity of the design. Under today's cost structure, it makes more sense to ramp up the hardware and to simplify the models for the sake of access and flexibility.

Annual data warehouse maintenance costs are considerably higher than operational systems. This is partly due to the iterative nature of analytics, with changing focus based on business conditions, but that is only part of the equation. The major cause is that the architecture and best practices are no longer aligned with the technology. It takes a great deal of time to maintain complex structures and it vastly complicates the understanding and usability of the utility on the front end. For those BI tools that do provide a useful semantic layer, allowing business models to be expressed in business terms, not database terms, user queries quickly overwhelm resources with complex queries (complex from the database perspective. Many BI tools still operate without a semantic layer between the user and the physical database objects, forcing non-technical users to deal with concepts like tables, columns, keys and joins which makes navigation nearly impossible. All of these problems can be solved today by taking advantage of the tools made possible by new market economics.

Business Participation Is Low

BI has only penetrated 5%[6] of people in organizations who could use it at the same time that spreadsheet use continues to grow, even in organizations that spend millions on data warehousing. Though spreadsheets are a useful personal productivity tool, they are consistently misapplied as workgroup or collaborative tools, a role to which they are poorly suited. The continued and even growing use of them is evidence that BI solutions, to date, have failed to solve the end-to-end problems that business people need.

The reason people continue to use spreadsheets, despite their inherent lack of consistency and control across an enterprise, is that they work. They are "expressive," meaning that business people can simply and declaratively express their business models, rules and logic. By contrast, most BI tools lack this expressiveness and to the degree they possess it, require data to be extracted from a data warehouse and stored in downstream repositories, usually proprietary ones. And even in those cases, the BI tools rarely allow business people to develop business models as easily as they can in a spreadsheet.

Aggregating or "cubing" data is like reducing the resolution of a picture - the detail is lost forever.

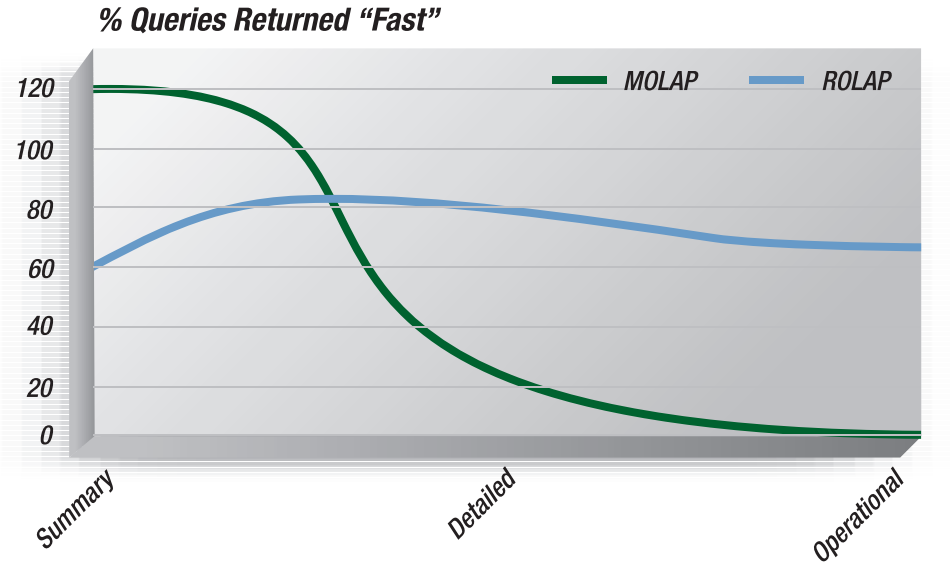
In today's BI environments, with "best practices" data warehouses, the business models are burned into the circuitry of the relational database schemas. The range of analysis possible is limited by the ability of the database server to process the queries. Some BI software has the capability to do much more, but it is constrained by the computing resources in the database server. These tools can present a visual model to business users, allow them to manipulate, modify, enhance and share their models. When a request for analysis is invoked, the tools dynamically generate thousands of lines of SQL per second, against the largest databases in the world, asking the most difficult questions. The use of these tools, though they are precisely what knowledge workers need, is very limited today because most data warehouses are built on proprietary platforms that are an order of magnitude more expensive and an order of magnitude slower than those available from newer entrants into the market like Netezza.

Best Practices Are Not Optimal

Many of the "common sense" approaches to designing data warehouses are tainted by an over emphasis in traditional systems architecture, relational data modeling and data management to the exclusion of business value and practice.

Aggregating or "cubing[7]" data is like reducing the resolution of a picture – the detail is lost forever. Detailed, transactional, even sub-transactional data is endowed with rich qualities that have to be exploited. For example, changing the "grain[8]" of the data from Product to Customer can increase the size of a data warehouse one or two orders of magnitude. Having access to a data mart summarized at the product level deprives the analyst of all of the rich attributes of a customer. Summarizing the data to the customer level likewise obscures the attributes of product. There is no logical reason for choosing one or the other, though this is a typical trade-off to shrink the size of a data mart. Multi-layered data warehouse architectures that lock away crucial atomic data rope off an essential part of BI. The only reason for restricting access to this data is insufficient computing resources.

Proprietary hardware and database software are a "sticky" economic proposition – market forces that



drive down the cost of data warehousing are blocked by the locked-in nature of these tools. Database professionals continue to build structures to maintain adequate performance, such as data marts, cubes, specialized schemas tuned for a single application and multi-stage ETL that slow the update of information and hinder innovation. As shown in the chart[9], users' satisfaction with aggregated solutions (MOLAP in this chart) drops precipitously when the requirements move from Summary to Detailed. Solutions that are based on leveraging SQL against large databases (ROLAP) exhibit fairly flat responses across the whole spectrum of needs.

It is convenient for database professionals to assume that business people only need access to aggregated or "cubed" data. Otherwise, the total population of users would have free ranging access to the largest databases, and performance tuning would be a critical issue. Instead, the very audiences that are served by this vast resource are cordoned off in a summarized set of data. Access to detailed data, if allowed at all, is through a process known as "drill-through," seeing the lower-level data that rolls-up an aggregated datum. None of the analytical capabilities of the BI tool, such as navigation, metric formation, filtering by characteristics or a host of other functions is available to use with this "drilled-through" data.

What's needed is a new paradigm that breaks open the data warehouse to business users.

Prescription for Realignment

No amount of interviewing or "requirements gathering" can break through and find out what sorts of questions people would have if they allowed themselves to think freely. The alternative is to ramp up the capacity, allow exploration and discovery, guiding it over a period of time. What's needed is a new paradigm that breaks open the data warehouse to business users, enabling ad hoc and iterative analyses on full sets of detailed and dynamic data.

Walk in Green Fields

Many ordinary business questions require sifting through massive volumes of detail data, not summarized cubes. The amount of data sifted per query may range from small to enormous, but the common characteristic is that the result is dependent on examining attributes of the most atomic pieces of data. These bits of information are lost as the data is summarized. Some examples of queries that are simple in appearance but cannot be resolved with summarized data are:

- Complementarity (a.k.a. market basket analysis): What options are most likely to be purchased, sorted by profitability, in connection with others? Extra credit: which aren't?
- Metric Filters (a metric derived within a query is used as a filter): Who are the top ten salespersons based on the number of times they have been in the top ten monthly (inclusion based % of quota)?
- Lift Analysis (what do we get from giving something away): How many cardholders will we send a gift to for making a purchase greater than $250 on their VISA card in the next 90 days and which of them did so because they were given an incentive? What if the amount is $300? Extra credit: greater than triple their average charge in the trailing 90 days before they were sent the promotion?
- Portfolio Analysis (to evaluate concentration risk, for example): Aggregating data obscures information, especially in a portfolio where the positive or negative impact of one or a small proportion of entities can adversely affect the whole collection. Evaluating and understanding all of the attributes of each security, for example, is essential, not an average.
- Valuations (such as asset/liability matching in life insurance): Each coverage and its associated attributes (face amount, age at issue, etc.) is required to do a complete valuation of a book of business.

All of these examples are just extrapolations of classic analytics. In addition to these examples, the following are examples of hybrid processing, involving the coordination of analytics and operational processing:

- Manufacturing Quality: Real-time feeds of shop floor are monitored and combined with historical data to spot anomalies, resulting in drastically reduced turnaround times for quality problems.
- Structured/Unstructured Analysis: The Web is monitored for press releases that indicate price

changes by competitors, causing the release to be combined with relevant market share reports from the data warehouse and broadcast to appropriate people, all on an unattended basis.

- Personalization: B2B and B2C eCommerce dialogues are dynamically customized based on analysis of rich integrated data from the data warehouse in real-time.
- Dynamic Pricing: Operational systems, like airline reservations, embed real-time analysis to price each transaction (also known as Yield Management).

This list is endless. Providing people (and processes) access to this level of detail is considered too difficult and too expensive, which is true in typical environments of proprietary databases and servers. Porting these applications to newer, commodity-based platforms that reflect the real economics of the market today can empower a multitude of powerful new BI applications.

Make It Current

The latency in data warehouses, with updates occurring overnight at best, and even more infrequently for the downstream data structures that users actually see, is unacceptable for the emerging hybrid applications discussed in the previous section. In addition, unattended analytical processes, where queries to the data warehouse are initiated by other systems, without human involvement, are increasingly relying on real-time or near-real-time data. The elaborate designs of today's data warehouse simply cannot accommodate these requirements. Sourcing transaction-level and even sub-transaction-level data, at enormous volume and extremely low latency demands extreme computing power, driving the cost up when using proprietary platforms and software to unacceptably high levels. Alternatives are now available.

With a high-performance, highly scalable architecture, at reasonable prices, appliances like Netezza's can solve the problem by simply applying the additional horesepower.

Using high-performance, scalable data warehouse appliances like Netezza's solves a major challenge in data warehousing. In order to provide adequate query performance on constrained hardware platforms, design techniques like star schemas, aggregates, heavy indexing and data-marting are employed. Unfortunately, these designs are notoriously poor at being updated, except in batch. Normalized schemas, optimized for transaction processing, are more appropriate for trickle-feeding real-time updates and catching messages from queues, EAI transactions, web services, logs of running programs and a host of other sources -- but these schemas typically offer very poor query performance. With a high-performance, highly scalable architecture, at reasonable prices, appliances like Netezza's can solve the problem by simply applying the additional horsepower.

Make It Simple

Whether man or machine initiated, any question that can be posed against a data model should be resolvable within a reasonable time at a reasonable cost. The business user should be able to think broadly about business questions to be answered and then use a simple interface to turn these into queries. Clearly, controls are needed to deal to with naïve queries that can cause degradation or complex queries that are so resource-intensive they must be isolated in order not to affect the other users of the system, but this should be the exception, not the rule. Moreover, simplicity should not be limited to the business users of the system. Managing and enhancing the physical elements should also be simple. Adding new CPU, memory and storage capabilities, performing load balancing, in short, all of the administrative and systems functions should be just as simple as posing a question to the system.

Make Change Painless

Relationships, mappings, and models are all conceptual. In any organization, they change constantly. The best set of models in a relational database, with views, indexes, demoralizations and partitioning, is only a good solution for a period of time. Unfortunately, organizations are often unable to apply the same level of discipline to the maintenance of these structures as they did to the original development effort, particularly given the multi-departmental and cross-functional nature of BI. Simplifying the structure by replacing physical optimization with more computing power makes a compelling case. Another compelling case is to employ software that manages the physical optimization of the data warehouse, by monitoring the usage patterns and constantly re-configuring the physical layer without effort or incident above it.

A BI system must be easy to update in near-real-time to reflect real-time changes in the business environment. Unless business people are provided with a BI solution that is at least as easy to use as a spreadsheet, they will continue to shun it. These are the minimal requirements:

- Declarative modeling (no programming or scripts)
- Intelligence about the data resources (implies active metadata)
- Collaboration capabilities (the ability to share work, distribute it)
- Complete abstraction from physical data such as tables, columns, files, directories
- Zero-impact from changes in the physical model
- Ability to build models from scratch, assemble models from components and be assured of referential integrity over time

Make It More Expressive

Relational database models are not business models. They are business data models. They lack the essential depth and richness that are expected by business people, such as the sequential logic found in financial statements. The actual logic for calculating Profit and Loss, Balance Sheet and Cash Flow statements involves certain items in sequence, making decisions (if profit, use interest rate #1, if loss use #2) and sometimes even solving simultaneous equations, all of which can be handled easily in a spreadsheet. Another type of analysis involves separating things into buckets, where the discriminant for bucketing is an attribute of the lowest level of detail available, such as action on a claim. Describing the logic for a model using this approach is very simple, but nearly impossible in a relational data model.

BI developed along two parallel paths that are still evident. Some tools are aligned with relational databases and are SQL-based. Their conceptual models range from simple row-column manipulations to very expressive analytical tools, but in the latter case, a very powerful SQL-generating engine is used to mask the complexity and difficulty of conducting analysis against relational tables. Other tools developed to allow more comprehensive conceptual models, but use proprietary languages and data management approaches, such as multidimensional databases. It is an open question whether all of these solutions can coexist under the umbrella of a unified conceptual model with a conceptual-to-physical interpreter. One thing is clear – taking advantage of high-performance resources in today's market will power these decisions.

Conclusion

Too long the domain of technologists and IT, BI needs to propagate throughout organizations and deliver the latent potential of the information stored in data warehouses. For this to happen, BI must deliver solutions that are more useful to the stakeholders. These must be more expressive, allowing business people to develop their own models as easily as they can in a spreadsheet. And they must help business users solve end-to-end problems, closing the loop by tying analytics directly to action and vice-versa.

In order to reach this new level performance, it is necessary to have enough horsepower in the database engine that the endless optimizing of database resources becomes unnecessary, giving way to simplicity and flexibility. This performance has to be delivered with a level of affordability commensurate with the advances of Moore's Law and innovation in database software. Most important, the administration of the database system has to mask the underlying complexity of the technology and offer a simple, seamless method for adding capacity as needed.

Solutions like the Netezza Performance Server® appliance are poised to provide this infrastructure. By integrating database, server and storage and placing processing power at the disk level, Netezza is able to harness the power of Moore's Law and combine it with the advantages of massively parallel processing, Linux and open-source systems. Together, this new approach delivers the performance, currency, flexibility and business value demanded by today's BI users.

---

[1] Knowledge worker is a term coined by Peter Drucker and defined roughly as anyone who works for a living at the tasks of developing or using knowledge.

[2] Moore's Law: In 1965, Gordon Moore, co-founder of Intel, observed that the number of transistors per square inch on integrated circuits had doubled every year since the integrated circuit was invented. Moore predicted that this trend would continue for the foreseeable future. Current thinking now is every eighteen months.

[3] Though technically not a result of Moore's Law, the capacity and cost for storage follows a similar and even more dramatic curve

[4] W.H. Inmon, Claudia Imhoff, et. al, The Corporate Information Factory, Wiley, 1997

[5] Total Cost of Ownership

[6] Proprietary research from Hired Brains, Inc.

[7] Multidimensional "cubes" are data structures that are aligned along dimensions, like Time, Customer or Product, and get their name from the analogy to a 3-dimensional cube. Most of these structures are more than three dimensions, so the cube metaphor is misleading, but the name prevails. Cubes can be real, virtual or in-between, so-called hybrids, but they generally lack the scalability of relational databases, thus the data is usually a summary.

[8] Ralph Kimball, The Data Warehouse Toolkit, Wiley, 1996

[9] Proprietary research Hired Brains, Inc.

About Netezza

Netezza is the market-leading provider of enterprise-class data warehouse appliances that deliver breakthrough performance and ease-of-use at a fraction of the cost of traditional data warehouses. The Netezza Performance Server system enables Fortune 1000 customers with terabytes of dynamic, detailed data to dramatically simplify even the most complex Business Intelligence (BI) initiatives. By architecturally integrating database, server and storage within a single appliance, the NPS system delivers 10 to 50 times the performance at half the cost of existing systems. Founded in 2000 and based in Framingham, Mass., Netezza has offices in Washington, DC, the United Kingdom and Asia Pacific. The Company has raised more than $68M from leading venture capital firms, including Matrix Partners, Charles River Ventures, Battery Ventures, Orange Ventures, Sequoia Capital and Meritech Capital Partners. For more information about Netezza, please visit www.netezza.com

Netezza Corporation : 200 Crossing Boulevard : Framingham, MA : 01702-4480
+1 508 665 6800 tel : +1 508 665 6811 fax : www.netezza.com

NETEZZA
The Power to Question Everything™

WP.ENTERBI.11.05