

WHITE PAPER

W I N T E R C O R P O R A T I O N

LARGE SCALE USE OF THE NETEZZA PERFORMANCE SERVER[®] SYSTEM

The Large Scale Data Management Experts



WinterCorp

SPONSORED
RESEARCH
PROGRAM



1 Introduction

For a few years now we have been hearing about data warehouse appliances: a new class of products designed to make data warehousing simpler and less expensive.

The Netezza Performance Server® System (NPS®) was the first of these products to be developed and has been available commercially since 2003.

WinterCorp, an expert in large scale data management and a consultant to leading enterprises on data warehouse architecture and scalability, has closely tracked the development of data warehouse appliances with interest—and with a lot of questions. One of the most interesting questions to us was this: on just how large a scale are these appliances going to be any good?

Paraphrasing the line from the movie Jerry Maguire, our slogan has been “Show me some evidence”.

Recently, we were approached by Netezza with the proposition that we interview some of their customers engaged in large scale use and satisfy ourselves as to just what they are doing. We agreed to do so on three conditions: (a) only WinterCorp and the customers would participate in the interviews; (b) Netezza would pay WinterCorp a fee for conducting the research; and, (c) if a publication resulted, we (WinterCorp) would have editorial control. Any findings or conclusions would be entirely our own.

Part of the arrangement, also agreed up front in the interest of making this information available as rapidly as possible, was that the customers would not be named in the publication. However, WinterCorp would of course know their identities and be able to validate them.

The companies we interviewed are all successful, sizable companies with substantial experience in data warehousing. Two of them provide marketing services. One is a leading web retailer.

Sections 2-4 of this report review the experience of each company, case by case.

Section 5 presents our overall findings and conclusions with regard to Netezza.

2 Retail Data Services Case—Company A

BACKGROUND

Company A provides data warehouse services to retailers based on information collected at tens of thousands of stores. It operates over 75 data warehouses, each containing point of sale information for a specific retailer for the past 27 months. The data warehouses are updated nightly on the basis of data that has been collected from the stores over the prior 24 hours.

The total size of this operation is very large: hundreds of billions of retail transaction scans are online at all times. Every day, about a half billion new transaction scans are incorporated into the data warehouses and a similar volume is purged.

EXPANSION OF SERVICES

In connection with a new, more powerful SMP hardware platform installed in 2002, *Company A* decided to upgrade its services to users. Previously most reporting was done offline, in batch. OLAP was offered via an in house developed user interface.

The improved version of its service would have two key features. First, online reporting and query would be offered to users. Second, the in house interface would be replaced with MicroStrategy, a commercially available end user query and reporting tool. This would enable users to create a wider range of queries and reports themselves. These two changes placed a substantially increased workload on the system.

In particular, online query and reporting required that data be analyzed and summarized at night, with the results stored in the database (referred to as “aggregates” of the data). For example, if a given customer had many queries and reports concerning the daily sales of products by store, then each night the system would calculate such figures for every product and store. This would then make it faster and easier for the system to handle such queries during the day.

CRITICAL ISSUES

Not long after upgrading the services it offered its clients, *Company A* found that its difficulties were increasing.

The nightly job of loading the data and updating the aggregates was growing very rapidly. And, each set of aggregates that was built only helped with some types of queries. For example, if a query that was counting distinct values violated the grain of the aggregate then the aggregate would be bypassed altogether. In order to satisfy its customers’ needs to be able to run a changing mix of reports, *Company A*’s technical staff was devoting more and more effort to analyzing performance problems and revising the aggregates or indexes built for a particular customer.

Even with all this attention to database design and tuning, some customers were unhappy with the response time to their queries. It was common for users to have to alter their business questions to avoid unreasonably long delays. These alterations often compromised the business value received, in addition to creating more work—and less satisfaction—for the customers.

Meanwhile, the data volumes continued to grow. By 2004, some customers had chains of thousands of stores, each delivering data to the central operation; data volume and workload were increasing at the same time; and, the data warehouse operation was showing signs of strain. In fact, for the largest chains, the system could no longer complete a day’s transactions in the nightly batch window. When users arrived at work in the morning, they wanted a complete view of the previous day’s activity. For the largest chains, with tables as large as twenty billion rows, that was no longer consistently possible.

The difficulty in staying up to date was partly due to the sheer data volume—*Company A* needed to load over 450 million new records each night—but the need to perform daily updates on large numbers of aggregates and indexes compounded the problem greatly.

CAPABILITY NEEDED

Company A badly needed a better solution.

But, executive management was not going to be receptive to proposals to greatly increase either the cost of the system or the size of the staff. Such investments would not be consistent with the business model of the organization.

Rather, a solution was wanted that would simplify the operation, improve the performance *and* lower system cost—a tall order.

EVALUATION PROCESS

Because the Netezza Performance Server (NPS) appeared to provide just what they were looking for, the evaluation focused primarily on Netezza.

Company A decided to subject Netezza to the most realistic and stringent test it could readily devise. Of its eight major decision support applications, it chose the largest and most demanding—the one with which they were having the most difficulty. They challenged Netezza to install a system on *Company A*'s premises and run through every step of their normal production cycle with this application—and with 17 billion records of real data. The steps in the production cycle included: load, referential integrity check, primary key enforcement, query, analyze, dump and purge—all on hundreds of billions of rows of data.

A common type of query in this application is a count of the customers in a certain set. An example is: “How many customers purchased frozen food products today in our downtown stores?” One such dashboard query that ranged across 17 billion records took ten hours and six minutes on the customer's existing data warehouse system. The same query took one minute and seven seconds on Netezza—over 540 times faster. Even considering that the existing platform was two years old and the Netezza system was brand new; that Netezza was involved in the test and the incumbent vendor was not—that is still a very large difference.

Going into the proof-of-concept, *Company A* was concerned about the ability of NPS to handle multiple concurrent queries. They had heard that the query concurrency “just wasn't there”. But after seeing such large performance gains—on the average, queries ran 30 times faster than on their existing system—*Company A* became convinced that their concurrency requirements would be much reduced. In fact, they tested this hypothesis in the POC. They set up a large workload which would have resulted in many queries and reports running at the same time on their existing system. NPS was required to accept all the queries; it was acceptable to queue them, but not to reject any. All the queries were completed well within response time requirements.

Through this onsite proof-of-concept with real data, real queries and a complete workload, *Company A* became convinced that the Netezza Performance Server had the functional capability and the performance to satisfy their needs. It demonstrated large performance improvements on their toughest application.

In addition, *Company A* monitored system stability during the proof-of-concept and tested failure and recovery.

In all respects, *Company A* found the system to be satisfactory.

Company A's evaluation favored Netezza for simplicity of installation, database implementation, and management; and, for its low cost of acquisition and ownership.

ACQUISITION & MIGRATION

In mid-2005, *Company A* acquired three Netezza 8250s. *Company A* migrated its data to the Netezza servers in three months. Most of the time was spent dealing with peripheral infrastructure issues, not with the process of implementing the databases in Netezza—that was remarkably simple. In the three

month period, *Company A* defined and implemented backup, recovery, security, network connectivity and all the other processes needed to support production. Three months after acquisition, they were in production with about 300 billion rows of data.

To accomplish migration, they moved the fact tables and dimension tables. They left the aggregates behind. On Netezza *Company A* runs without aggregate tables, “the most liberating thing you can imagine”, according to *Company A*’s Executive Director of Database Architecture.

Production Experience. After a year of production experience, *Company A* reports that the most remarkable aspect of NPS is that they really don’t have to tune the database. They get the data to be logically and structurally correct; define it to NPS; and, load it. NPS does its work—loading and querying—rapidly and efficiently. They don’t have to spend energy on file management, tuning, selecting indexes or selecting aggregates.

As a result, the staff spends its time solving business problems and helping customers—not struggling with vexing design and performance problems.

Since the initial acquisition, *Company A* has expanded its Netezza production operation from three 8250’s to four 8250s and subsequently, to yet larger systems.

SCALE OF OPERATION

A total of 31 terabytes of user data is managed online in the Netezza production servers. There are over 100 customer data warehouses running on this combined system. These data warehouses service about 350 Microstrategy users and about 50 SAS users. The individual data warehouses contain as many as 55 billion rows of transaction data. On a typical night there are about 40,000 SQL statements executed in batch. On a typical day, there are about 300 analytic Microstrategy queries and reports. Each of these can execute as many as 400 SQL statements—thus the Microstrategy queries and reports—as well as the SAS jobs—can be quite extensive in their demands on the database.

SCHEMA COMPLEXITY

Company A’s data warehouses—each for a separate customer—employ a classic star schema. The central fact table is the transaction table, typically containing billions of rows of data. There are dimension tables for product, store, customer and other components of reference information. However, it is worth noting that some dimensions can also be very large.

The product dimension can be as large as 20 million rows. The customer dimension can be 100 million rows. These large dimensions must be joined to the fact table in certain queries. *Company A* reduces the complexity involved by distributing the fact table and the customer table on the same key.

QUERY COMPLEXITY

According to *Company A*, joins involving as many as ten tables are not uncommon—and such joins execute efficiently. Analytically and statistically, the queries are reportedly quite complex. Since Microstrategy queries routinely result in the execution of hundreds of SQL statements, these are clearly not trivial queries.

One executive at *Company A* says, “We have customers who will upload ten million customer IDs and 500,000 UPCs (product identifiers) and say, ‘analyze this combination for the last six months’. They aren’t asking us for the price of milk—they are asking for deep, complex analysis on a large scale.”

VENDOR RELATIONSHIP

One executive at *Company A* sums up the relationship as follows, “Netezza is not perfect, but they are solid. They are responsive and a good partner focused on data warehousing. It has been very beneficial for our organization to find a technology and a partner to work with us as we move forward. Any way you look at it, the pros greatly outweigh the cons.”

Company A has discovered bugs in the product: Netezza has consistently handled them professionally

and corrected them promptly. *Company A* has at times requested product improvements—and Netezza has delivered them.

BUSINESS VALUE

Paraphrasing slightly, the comments of *Company A* on business value can be summed up: before we brought in Netezza, our difficulties were impacting our ability to deliver. Our data was not always up to date. Our customers could not get answers to many of the questions that were important to them—or they had to engage in complex work arounds to get the results they needed.

Now our customers run virtually any query they want to. They are absolutely doing things they could not do before. System response is fast. Our DBAs are five to ten times as productive. We have no stored aggregates or indexes—and we get much better query performance than before. Our DBAs and our customers are able to put their energy into solving business problems.

3 Web Retailing Case—Company B

BACKGROUND

Company B is a large, rapidly growing—and rapidly diversifying—web retailer. Business growth is so rapid that the data warehouse doubles in size every 12-24 months.

The data warehouse is the core business reporting platform across the entire company. It is expected to contain all relevant data for all areas of business, and contains over 70 terabytes of data.

Company B's existing Oracle/RAC architecture delivers good performance for a broad range of straightforward needs. However, very large volume and very high complexity queries are often a problem.

About three years ago, *Company B* started seeking more insight into the long term trends in the behavior of its customers. The Company believed such insight could be developed by doing more extensive analysis of its clickstream data.

One year of clickstream comprised roughly 40% of the data in the data warehouse. There were tens of billions of rows in a single table, where each row represented one view of one page by one visitor to the website.

The clickstream data was already used extensively for short term analysis on Oracle, and *Company B* was happy with that capability.

But it was impractical to do long term analysis of this data on *Company B*'s existing data warehouse. It took more than 24 hours to scan the whole clickstream table—and attempts to do so often failed because they required too much temporary database space or main memory.

As a result, *Company B* knew how many times a customer visited in a week, but didn't know how many times each customer visited in a year. And many other important business questions simply remained unanswered. Either such questions would take too much in the way of resources or they could not be answered soon enough.

EVALUATION

In 2003, *Company B* did a periodic reassessment of data warehouse direction, looking at Oracle/RAC, DB2 and the Netezza Performance Server (NPS).

Netezza seemed to present an opportunity for a step change in performance. At that time, Netezza looked interesting but was very new. It outperformed the others, but was functionally incomplete.

In early 2005, Netezza approached *Company B* again, saying they had added the features *Company B* was looking for. Netezza had made considerable progress and *Company B* decided to bring the platform in house for a proof of concept (POC).

In the POC, *Company B* became convinced that Netezza would provide a good platform for analysis of large volumes of data. An NPS 8650 system was acquired: 660 processors capable of managing a total of 27 terabytes of user data.

KEY USES

The current role of the NPS is as a specialty analysis platform. In that role, it delivers millions of dollars of business value each year. It is typically used for interactive analysis: situations where the timeliness of the response is important. Three examples are credit pricing, fraud detection and promotion planning.

CREDIT PRICING

In one instance, *Company B* wanted to establish a pricing regime for a new co-branded credit card. The credit partner proposed to segment pricing by customer, depending on credit risk and other factors. *Company B* had 24 hours to negotiate a deal to put the credit card program in place for the holidays.

In order to understand the impact of the proposed pricing scheme on its business, *Company B* needed to analyze the credit risk, credit behavior and purchase behavior of its 60 million customers. It needed to project the impact of the proposed scheme. *Company B* then needed to develop alternative pricing schemes that were better suited to its business objectives—as well as developing the data driven analysis to support the negotiations.

This is the type of problem that is not solved with a single query. Rather, the team had to form hypotheses; test them on the data; analyze the results; and, then try again. The queries involved joins on clickstream, customer data and order history—all large tables. On NPS, the team was able to perform the key query ten times in 24 hours. The first version of the query took 20 minutes, even though it had to analyze 15-20 terabytes of data, and perform large joins.

Company B was able to renegotiate the credit agreement successfully and realize millions of dollars of incremental revenue as a result. On its main data warehouse platform, *Company B* believes that each iteration of the key query would have taken many hours. It would not have been possible to obtain the same favorable results in the negotiation.

FRAUD DETECTION

In another instance, *Company B* came to suspect that an attack was underway on its website one evening about 6pm, when it identified a user session that appeared to fit a malicious pattern of behavior.

In a series of queries of the clickstream data, *Company B* was able to confirm this pattern and several variations of it that were in use. By midnight they had a complete picture of history of fraudulent behavior of this individual. At that point, they had run many large queries, each of which would have taken 45 minutes or more on the existing data warehouse. On NPS, each of these queries took a few minutes. Once the full pattern and history were known, they were then able to take action and protect *Company B*'s valuable brand.

PROMOTION TESTING

NPS is used regularly to test new business ideas, promotions and discount programs. By analyzing past customer behavior—using clickstream data, customer data and purchase history—*Company B* is able to project the likely cost and benefit of a proposed campaign.

All three of these examples—credit pricing, fraud detection and promotion testing -- are like others that arise regularly at *Company B*. They require a complex analysis of a very large volume of data in a short time frame. A business decision is about to be made—a decision that could produce a high yield—or a large loss. “There is great business value in not having to guess when there are millions of dollars on the line”.

PRODUCTION EXPERIENCE

Company B has about a year and a half of production experience with the NPS 8650. The system has worked well and delivered exceptional query performance.

To support the system, *Company B* reports it employs one part time system administrator (SA) and no database administrators (DBAs).

SCALE OF OPERATION

Company B has about 27 terabytes of data online. New data is loaded throughout much of the day at a rate of about 60 gigabytes per day. Concurrently, large queries performing wide ranging, deep analysis of the data are running much of the time. Typically, a single query or a few queries are running at any one moment. They have rarely experienced query contention. The most remarkable scale factor in the operation is the amount of data scanned in many of the queries. It is not unusual for a query to scan 10 terabytes, 20 terabytes, even all 27 terabytes of the data.

SCHEMA COMPLEXITY

There are four large tables in the NPS data warehouse: page views (billions), purchase history (hundreds of millions), customer (sixty million) and product (millions). No two of these are distributed on the same key. There are also smaller dimension tables.

QUERY COMPLEXITY

It is common for queries to join all four large tables and several of the smaller dimension tables. Queries range over large areas of the database and are complex analytically.

Queries involving only one of the large tables run “at blinding speed”—typically 100 to 1000 times faster than on the main data warehouse system.

According to one manager at *Company B*, queries that join multiple large tables that are not distributed on the same key—meaning that one of the sets of data to be joined must be redistributed across the SPUs—typically run more slowly than single table scans. Most are still 5-10 times as fast as the main DW.

Overall, *Company B* feels that Netezza handles complex queries well, including analytic functions. Early in the Netezza deployment, *Company B* identified a performance problem for queries that contained multiple left-outer-join constructs. Netezza has subsequently fixed this issue in a later release.

SYSTEM AVAILABILITY

Within the first few months of production use, *Company B* experienced multiple SPU failures. One *Company B* manager says, “We were concerned. But Netezza had been tracking these failures before we became aware of the problem. They briefed us on what they knew. With our consent, they installed some diagnostic software on our system. Shortly, they called us back with a diagnosis of the problem—it was a bad lot of chips. They had already contacted the manufacturer. They replaced about 5% of our SPU’s, needing only 15 minutes of downtime. The problem was fixed.” Since that time, system hardware has been highly reliable and highly available.

Space management is sometimes an issue and can affect system availability and performance. Storage of sparse data (data with many null values) is less efficient than on Oracle (about 30% worse in one important case). As the system approaches its storage capacity, query performance can suffer. WinterCorp believes that Netezza will soon develop product enhancements that will help with this issue.

There is some table maintenance required in the case of a “rolling window” and this can be an issue in some situations. Deleted data is not removed at the time of deletion. Rather, a separate process reclaims space. *Company B* reclaims about 9 GB per day from one main table, which takes about 7 seconds. This effectively allows the Netezza to function with no maintenance window.

While acknowledging that NPS is not perfect, one manager at *Company B* says that it is still remarkably well suited to their current uses. The system handles extremely demanding queries with remarkable performance and with few problems. Data analysts in the company love it.

VENDOR RELATIONSHIP

The relationship with Netezza has been excellent. *Company B* has been very impressed with Netezza’s customer service. As a vendor, Netezza has been professional and responsive. There are not many problems; those that arise get addressed quickly.

Summing up, one manager says, “We haven’t found anything like Netezza in the marketplace. We are accomplishing many things—things that are important to our business—that we could not have done any other way. The clickstream analysis application has been a huge success.”

4 Market Research Case—Company C

BACKGROUND

Company C provides market research services to manufacturers of consumer packaged goods (CPG) and to retailers—based on information collected from stores and newspaper advertisements. At the time of our interview, it was operating 20 data marts on a Netezza Performance Server, each one for a specific client.

The client data warehouses typically contain between 500 gigabytes (GB) and one terabyte (TB) of data. *Company C* collects the data for each client; organizes and aggregates it according to client specifications; hosts it; and, provides desktop tools to help the client analyze the data.

In addition the client-specific data marts, *Company C* has about 7 TB of base data on the Netezza system.

CRITICAL ISSUES

About a year before our interview, *Company C* found itself struggling with daily operations, facing steadily increasing complexity and cost maintaining its clients' data warehouses. The worst part of the problem related to aggregation. The clients wanted the data aggregated many different ways and their interests changed frequently.

For example, one week a client might be interested in analyzing the performance of a class of products in grocery stores which have a deli counter. For that purpose, the client would want data aggregated from those stores. Having resolved the business problem behind that interest, the client might next want to analyze the performance of another set of products in stores which have a pharmacy.

And, the next client may be interested in the performance of certain products in free-standing grocery stores that are near a Wal*Mart one week; the next week this client may want to analyze the performance of products in stores in malls.

In all of these cases, the client will do extensive analysis asking many questions about the same sets of stores. Thus, it makes sense to organize and aggregate the data to make it easy to ask the questions and efficient to produce the answers.

In general, the problem was that each project would often require its own special hierarchy of stores and products. Very large volumes of data were thus constantly being reorganized in rather complex ways. And, *Company C* was finding it increasingly difficult to keep up with the requirements to aggregate data in different ways for each project that each client undertook.

Performance problems were common; each performance problem could be identified, diagnosed and overcome; but, as time went on, the resulting infrastructure became progressively more complicated, expensive and difficult to change.

CAPABILITY NEEDED

Company C wanted a better way of operating its data warehouses—a way that would be simpler and more nimble, while reducing performance problems.

EVALUATION PROCESS

In September of 2005, *Company A* decided to evaluate four relational database systems, including Netezza's Performance Server.

Following a benchmark, Netezza was selected, primarily for the simplicity with which databases could be designed, implemented and supported. As it turned out, this came with large improvements in performance over *Company C*'s prior platform.

For example, *Company C* builds clusters of stores on request (e.g., reorganizing and aggregating the data

of a client according to a custom store hierarchy) to support specific analysis needs. Time to build one set of clusters was in one case reduced from ten days on the prior platform to three hours on Netezza.

Some of the other products evaluated outperformed Netezza on specific tests. But Netezza had the best performance overall.

SCALE OF OPERATION

The scale of operation has increased rapidly since Netezza was adopted as the production platform. In six months, the number of customer data warehouses has increased from 5 to 21. There is now 7 TB of raw data in production. *Company C* has handled this rapid increase in scale while decreasing staff. With the old platform, *Company C* had about 3 full time staff equivalents supporting the system—which was servicing 5 clients. It now has less than one full time staff equivalent supporting a Netezza architecture that services 20 clients—each with its own data warehouse.

Company C believes that, if it had stayed on the old platform, it would have needed to add 2-3 full time staff equivalents—and probably would not have been able to deliver the same level of customer service.

PRODUCTION EXPERIENCE

Initially, *Company C* acquired an NPS 8250 for its production system. After an initial period of positive experience, this was expanded to an 8650.

It used to take two hours each day to update each customer's data. It now takes 7 minutes. End user response time has dropped dramatically.

On the prior platform, *Company C* was always concerned about the adding to the database workload because increased workload tended to result in performance problems. Now, *Company C* is confident that additional workload will be readily handled by the platform—and they tend to push work onto the database instead of pushing it off the database onto other platforms. This makes the operation more integrated and simpler to manage.

At the time of the interview, *Company C* had been in production for about 6 months. Experience with NPS has been very good. There have been a few performance problems, which were readily resolved with the help of Netezza.

SCHEMA COMPLEXITY

Company C's data warehouses are classic star schema designs. The key dimensions are product, store and time.

QUERY COMPLEXITY

While join complexity is not high, the queries are large and demanding, ranging across large volumes of data and doing extensive analysis.

WORKLOAD

The overall database workload is large and complex, mostly due to the demanding requirements for updating, organizing and aggregating the data for each client.

VENDOR RELATIONSHIP

The vendor relationship has been “really good.” Netezza is very responsive and is generally able to resolve problems quickly. Performance problems are usually diagnosed remotely. Netezza is efficient in how it works with customers.

BUSINESS VALUE

As a result of switching to Netezza, *Company C* feels that its client data warehouses are simpler to operate and manage. *Company C* is able to be more responsive to the needs of its customers, giving better and faster service. Customers are able to do things more quickly and easily themselves. Customers get better response time to their queries.

Company C is able to build databases faster and more readily accommodate custom requests.

Company C believes it is saving money on total cost of ownership with Netezza—at least the equivalent of 2-3 skilled full time database administrators.

5 Conclusions

WinterCorp's principal conclusions from these three customer interviews, and from other information we have gleaned in tracking Netezza since its inception, are:

1. DATA SIZE: There is no question that Netezza can handle sizeable volumes of data.

Company A has 31 terabytes of data online in production use and is moving toward yet more. *Company B* has 27 terabytes of data in production.

Note that *Company A's* 31 terabytes is comprised of some 100 separate databases. The largest is probably 5-10 terabytes—and most are smaller. However, *Company B's* 27 terabytes is stored in a single integrated database with a 20 terabyte fact table.

2. DATABASE QUERY PERFORMANCE: Customers report “jaw dropping” query performance.

The customers we interviewed could not stop talking about the query performance. It was hard work to get them to move on to other subjects we wanted to cover! Over and over, they told us about queries—mostly analytical queries that range over large volumes of data—that used to take hours and now take minutes.

In certain areas of their business, this is game changing: fraud detection; promotion planning; credit negotiations; and, other areas. *Company B* emphatically describes solving business problems that were out of reach before. And, the end users simply love it.

The customers perceive these big gains in query performance to apply in many cases. Not every query runs much faster on Netezza, but many do. Even if some run more slowly than on other platforms, which at least one customer acknowledged, that is ok because users are so thrilled with the performance they get on most of their queries.

3. QUERY FREQUENCY: Online query frequencies are moderate at these sites.

At *Company A*, there are about 1300 SQL statements processed as online, interactive queries on a typical day. Each of these statements can itself comprise a substantial amount of work.

4. QUERY SIZE: Many of the queries are large, in terms of the amount of data they range over.

As we were told at *Company A*, their customers are seeking in depth analysis of large volumes of retail data. So, the daily query workload translates into perhaps a thousand large scans a day, many scanning terabytes of data. Each of these large scans would take hours to complete on a typical data warehouse, where there is much lower degree of parallelism.

The pattern is similar at *Company B*, where the application was moved onto the Netezza platform precisely because the individual queries require analysis enormous amounts—terabytes—of data.

5. ANALYTICAL QUERY COMPLEXITY: Many queries are analytically quite demanding, meaning that large volumes of data and complex calculations are involved.

The customers we spoke with have queries running in parallel across 220 to 660 processors. An equivalent degree of parallelism is not available or not practical on the data warehouse platforms these customers used in the past. Clearly, it is the effective exploitation of this high degree of parallelism that makes it possible for these customers to run hundreds of such queries a day.

6. STRUCTURAL QUERY COMPLEXITY: Structural query complexity is low to moderate, meaning that the simpler types of joins predominate in the workloads described to us.

At the same time, workload at both *Company B* and *Company C* presents evidence that Netezza does handle more demanding joins on a large scale.

One of the most demanding operations in a large scale data warehouse is a join between sets of data

distributed in storage on different keys. In this case, the database engine must redistribute at least one of the data sets so that matching rows can be brought together. Accomplishing the join and/or data redistribution efficiently, as the join sets grow very large, can involve challenging problems in performance and scalability. In our opinion, some of the most widely used data warehouse products are quite limited in this respect.

Not all data warehouse applications exhibit a requirement for this demanding type of join. In particular, applications in the widely popular star schema model often do not, as they typically have only one large table.

The Netezza products are functionally capable of performing these more complex joins. Netezza has told us that the data redistribution is accomplished in parallel on the SPUs. *Company B* and *Company C* are performing some joins between large tables distributed on different keys.

7. DATA INGEST: These customers are loading large volumes of new data into their databases at least daily.

At *Company A*'s site, the four 8250's ingest a combined 450 million rows a day of new data. *Company B* loads yet higher volumes of data throughout the day, concurrently with large scale demanding queries.

8. DATABASE LOAD PERFORMANCE: Customers report very good database load performance. This is partly due to the highly parallel architecture and partly due to the fact that product does not rely on indexing.

Company A did away with stored data aggregates (totals, averages, medians, etc.) when switching to Netezza, something described as, "The most liberating thing you can imagine." Some users of other data warehouse products have databases in which the aggregates occupy far more storage than the base data.

If you don't have to maintain aggregates and indexes, then database loading and updating is going to be much simpler and faster. Also, there is less time and effort in accommodating many new requirements.

9. EASE OF IMPLEMENTATION AND TUNING: Customers say that it is remarkably simple to create, administer and maintain databases on Netezza.

We heard, "there is no tuning"; "there is practically nothing to tune"; "we don't have a full time DBA"; "it is nothing to create a new database"; and, similar remarks. *Company C* describes a large savings in staff due to the simplicity of working with Netezza and, after reducing staff, says it is able to be more responsive to customers.

We find ourselves wondering: is it really that good?

But, whatever the details may be, you can't miss that customers perceive a major advantage here. They are very pleased with this aspect of the product. They have redeployed staff to more productive activities. They report they are more agile in serving their customers or reacting to business developments.

And, with *Company A*, it is more dramatic than that. *Company A* couldn't tune fast enough to keep up with the needs of its customers. And now, it has practically no tuning—and its customers can run a wider range of queries to boot.

So, comment from these three customers is very strong and very consistent on this point.

10. SYSTEM STABILITY: System stability is good.

These customers report no major problems with up time. There are bugs from time to time, but Netezza fixes them promptly and professionally.

11. PRODUCT MATURITY: The Netezza product line is not fully mature.

Rather, it is a new, highly innovative product with some rough edges. At the same time, though, it

has users operating on a large scale, loading a lot of data and handling a substantial amount of query, analysis and reporting—day after day. Customer A says that Netezza has been responsive to requests for enhancement.

12. VENDOR RELATIONSHIP: All three customers reported that Netezza provides very good customer service. All three described Netezza as a good partner.

BOTTOM LINE

So, how can we boil it down? On the basis of what we have heard so far, we believe that Netezza has happy and successful customers operating on a scale that reaches into the 35+ terabyte range. That is 35+ terabytes of data, not disk (these customers have three times that much disk).

We did hear about a few technical limitations and problems. Customers view neither the company nor the product as perfect. The product is certainly not fully mature, and this has some consequences. But these customers have found that all of the problems and limitations are manageable and—for what they are trying to do—the pros greatly outweigh the cons.

Of course, none of the companies interviewed is operating an enterprise data warehouse on Netezza. But, if you apply the Netezza product to an application that fits its capabilities, you may well get results like the customers we talked to.

And what is their story? These customers are operating data warehouses that have been in production for six to eighteen months. They load huge volumes of new data at least daily. Their Netezza systems work well, managing as much as 35 terabytes of data. They deliver spectacular query performance. And, they deliver results day in and day out. All three of the customers we spoke with are contemplating -- if not actually implementing—further expansions of their Netezza data warehouse. All have resounding evidence of excellent business value and happy end users. And, finally, all report that Netezza has been an excellent partner and delivered very good customer service.

WinterCorp is an independent consulting firm that specializes in the performance and scalability of terabyte- and petabyte-scale data management systems throughout their lifecycle.

Since our inception in 1992, we have architected many of the world's largest and most challenging databases in production today. Our consulting services help organizations define business-critical database solutions, select their platforms, engineer their implementations, and manage their growth to optimize business value.

With decades of experience in large-scale database implementations and in-depth knowledge of database products, we deliver unmatched insight into the issues that impede performance and the technologies that enable success.



WinterCorp

411 WAVERLEY OAKS ROAD, SUITE 327
WALTHAM, MA 02452
781-642-0300

visit us at www.wintercorp.com

©2006 Winter Corporation, Waltham, MA. All rights reserved.
Duplication only as authorized in writing by Winter Corporation.