

Paper 385-2008

New Architecture and Speed with a Netezza Data Warehouse Appliance

Dana Rafiee, Destiny Corp, Rocky Hill, CT

ABSTRACT

One of the biggest challenges organizations have is how they store and retrieve for analysis large amounts of data. This problem has plagued industries for years. As data sizes grow into Terabytes and Petabytes, organizations continue to face these challenges as users want to analyze and report on 'the whole data'. SAS has partnered with Netezza, a Data Warehouse Appliance vendor. This vendor is a pioneer of the self-contained warehouse appliance, which contains a database, storage and SQL processing. Each appliance contains 108 computers and storage disks to offer intelligent, massive parallel processing with load rates of 500GB per hour, all at an economical price in comparison to Teradata. The largest worldwide data customers such as Google, Yahoo, AOL and the NYSE had no other choices but to look to new, innovative technology to allow them to not only store enormous amounts of data, but to also reduce their query times from 24 hours down to 4 minutes. They outgrew the traditional Oracle, DB2 types of environments. SAS R&D has also jumped on board to work with this vendor to perform traditional SAS processing such as scoring 'IN' the database, hence gaining incredible speed for SAS customers.

INTRODUCTION

This presentation will discuss and demonstrate this technology and how SAS is offering new methodologies for processing in the database against large amounts of data.

BUSINESS PROBLEM

Many organizations have stated various challenges when accessing data:

- Access to Data is slow
- Long running queries

THE RESULT

Data analysis is constrained by the time it takes to load and retrieve the data. Organizations adapt and only analyze the data they can get access to within a reasonable amount of time and effort; they forego critical analysis because it is too time consuming or not possible to achieve given their current infrastructure.

IMAGINE AN INNOVATIVE SOLUTION

Times have changed. Now an environment that offers access and analysis of data that is 25 – 500 times faster than traditional methods exists. This allows Business Intelligence and Analytics software the ability to get access to the data and analyze it as quickly as the business can use it. Imagine the ability to summarize and get statistics from the organization's data in a few minutes or seconds instead of many hours. Imagine the ability to set up this environment in a day, not require a team of Database Administrators to run it and allow the business community to concentrate on the analysis instead of the hassles of getting access to and organizing the data. Imagine the ability to have real-time profile analysis of the customer against the warehouse of data while the customer is still on the phone.

The excuse 'we cannot get access to the data fast enough to support the business' is now history.

DESTINY PERSPECTIVE

As a Business and Information Technology Consulting firm, we constantly assess the vendors in the industry and market niche they support. We work for our customers to help them answer their business questions and support their business needs. We have found a proven, simple, cost effective way to solve the problem of not being able to get fast access to data to make decisions – the Netezza Data Warehouse Appliance.

The Netezza Data Warehouse Appliance looks like a refrigerator that rolls into the Data Center. Each rack holds 12.5 Terabytes of storage. For more storage, just chain a bunch of these 'refrigerator looking' racks together. We believe it is simpler because all of the pieces are contained in a single rack. Many vendors on the market require a complement of software, hardware, database technology and networking to make their solutions work.

How can it load data at 500 Gigabytes an hour and retrieve data in seconds? It is due to its design.

DESIGN

Each Data Warehouse Appliance (refrigerator) contains (108) computers called Snippet Processing Units (SPU). Each SPU is an integrated circuit board with a CPU, 400-Gigabyte hard disk, memory and 1 Gigabyte Network Interface Card. Each Appliance (cabinet) contains 108 of these SPUs. This offers parallel processing across 108 computers inside of each cabinet.

HOW IT WORKS

As the data is loaded into the Appliance, it intelligently separates each table across the 108 SPUs. Typically, the hard disk is the slowest part of a computer. Imagine 108 of these spinning up at once, loading a small piece of the table. This is how Netezza achieves a 500 Gigabyte an hour load time.

After a piece of the table is loaded and stored on each SPU (computer on an integrated circuit card), each column is analyzed to gain descriptive statistics such as minimum and maximum values. These values are stored on each of the 108 SPUs, instead of indexes, which take time to create, updated and take up unnecessary space. Imagine your environment without the need to create indexes.

When it is time to query the data, a master computer inside of the Appliance queries the SPUs to see which ones contain the data required. Only the SPUs that contain appropriate data return information, therefore less movement of information across the network to the Business Intelligence/Analytics Server.

For joining data, it gets even better. The Appliance distributes data in multiple tables across multiple SPUs by a key. Each SPU contains partial data for multiple tables. It joins parts of each table locally on each SPU returning only the local result. All of the 'local results' are assembled internally in the cabinet and then returned to the Business Intelligence/Analytics Server as a query result. This methodology also contributes to the speed story.

The key to all of this is 'less movement of data across the network'. The Appliance only returns data required back to the Business Intelligence/Analytics server across the organization's 1000/100 MB network. This is very different from traditional processing where the Business Intelligence/Analytics software typically extracts most of the data from the database to do its processing on its own server. The database does the work to determine the data needed, returning a smaller subset result to the Business Intelligence/Analytics server.

BACKUP AND REDUNDANCY

To understand how the data and system are set up for almost 100% uptime, it is important to understand the internal design. It uses the outer, fastest, one-third part of each 400-Gigabyte disk for data storage and retrieval. One-third of the disk stores descriptive statistics and the other third stores hot data back up of other SPUs. Each Appliance cabinet also contains 4 additional SPUs for automatic fail over of any of the 108 SPUs.

NEW TRENDS

Several Business Intelligence and Analytics vendors are now exploring the ability to push most of their processing to the Netezza database, further reducing the movement of data. One of the initial offerings is in the area of scoring in the database. This is typically a very simple process where an algorithm calculates across a set of columns in the database. The problem is that it typically does this across an entire table that has first been relocated to the Business Intelligence/Analytics server, hence movement of the large table across the network. Quite often, after the scoring process completes, the table is loaded back into the database requiring a second move of the data across the network.

Destiny Corporation, a member of the Netezza Developer Network, is developing one of the first scoring processes that automatically takes a Business Intelligence/Analytics vendor's algorithms and transforms them to a compiled process that runs directly inside the Netezza database.

CASE STUDIES

NEW YORK STOCK EXCHANGE

Recently, the New York Stock Exchange purchased a 100 Terabyte Netezza Data Warehouse Appliance. One of their challenges is not only the fact that their data needs are growing exponentially, but they need to be able to analyze different types of data daily. Every day the Securities Exchange Commission goes onsite to the NYSE to audit trade data on different publicly traded firms across several years. The NYSE must prepare the data for the SEC Auditors to analyze. Using Netezza's ability to load 500 Gigabytes an hour allows the NYSE to load different, ad hoc data upon request, making it available to the SEC Auditors. The NYSE offloads the data once the SEC has completed their work and new data is loaded for the next business day's audit. The NYSE has stated that no other warehouse infrastructure was able to support this type of business need.

TRADITIONAL SAS BASED ORGANIZATION

One of our recent clients wanted to incorporate a Netezza Data Warehouse Appliance to reduce their multi hour run times. They are a SAS shop and heard that SAS is looking at doing more work inside of the Netezza Appliance. They also learned about the SAS/Access to Netezza software recently released. We analyzed their overall processes that ran on both z/OS and Unix using traditional SAS data sets and a relational database. Queries against the data were slow, loading data was slow and the organization adapted to this 'slow' response time by 'modifying' their business needs. The organization grew into accepting that working with large data was cumbersome and therefore evolved into a culture of running very little 'what-if' analysis. Some jobs would never even finish. Old reporting processes that ran for hours remained their standard information set.

We went through several steps with the client to understand what they do and how they could best convert to a simpler, Netezza based environment to gain the fast performance they expect. The steps were as follows:

1. Destiny met with the client to understand their current architecture and business processes. We looked at the size of the data, number of tables, keys, cardinality, indexes, politics and more.
2. The key design team members from both Destiny and the client came together to whiteboard how these processes should be designed with a Netezza Data Warehouse Appliance and industry best practices.
3. We defined benchmark targets we expected to achieve such as load and retrieval times desired, e.g. 200 times faster than current processing.
4. We analyzed the client's return on investment. For example, if they invest in this platform and get drastic reductions in overall processing time, what does that mean to the business? Increased revenue, decreased expenses, speed to market, etc.
5. We designed a best of breed approach using SAS and Netezza as a complement to each other.
6. We determined ideal load and retrieval strategies, defining what Netezza does best and what SAS does best.
7. We finalized the IT process design to support the business needs.

8. We implemented the new design.
9. The client saw results that were 75 – 250 times faster than their current environment.
10. The client purchased their first Netezza Data Warehouse Appliance. The box arrived on a Tuesday and was integrated into their system by Thursday.

COMPETITIVE ALTERNATIVES

In our studies of other vendor solutions, we have seen a trend. After Netezza coined the phrase 'Data Warehouse Appliance' at the turn of the century, other vendors are trying to follow. The offerings are software based, hardware based or a combination of the two, leading to complicated solutions. Software companies can devise the logic, but need an array of hardware devices chained together in parallel from hardware vendors. The large hardware vendors have billions of dollars already invested in their legacy storage technology. This is not easy to transform overnight. We have seen several vendors offering their version of the 'Data Warehouse Appliance' to the marketplace, but to date, the solutions are still very complex, require several vendors, a combination of skill sets not readily available in the industry and several database and architect resources. Some are simply repositioning legacy hardware designs.

TEST DRIVE WITH A PROOF OF CONCEPT

Netezza is a very aggressive organization and constantly puts its reputation on the line, standing behind its promises. Netezza will offer a proof of concept to any qualified organization with a large data business need. The way it works is simple. The business benefit and ROI for fast data loading and retrieval are defined and the customer identifies the data they desire to test on the appliance. The data is loaded on a Netezza box at the customer site or at Netezza headquarters in Framingham, Massachusetts. The customer runs their normal Business Intelligence/Analytics processes. The benchmarks are documented. Customers always see incredible results.

CONCLUSION

In conclusion, Destiny Corporation has tracked trends in Business and Information Technology since 1987. We have heard our customer's issues from both a Business and an IT perspective. Living and working with large data has been a growing challenge for our customers as data acquisition continues yielding masses of stored data available to the organization for analysis. The storage and analysis mechanisms must adapt. Our analysis has shown that Netezza is currently the leader as it elegantly answers today's large data business needs.

CONTACT INFORMATION

Dana Rafiee
Destiny Corporation
2075 Silas Deane Highway
Rocky Hill, CT 06067 USA
Work Phone: 860-721-1684
Fax: 860-721-9784
E-mail: drafiee@destinycorp.com
Web: www.destinycorp.com



SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies.