



February 22, 2008

## The Streaming Analytic Appliance

*Ready When You Are*

By Justin Lindsey

### **The Early Days of Data Warehouse Appliances**

When data warehouse appliances were first unleashed on the scene back in 2002, industry observers viewed this new technology with a bit of skepticism, and rightly so. Whenever an entirely new technology enters a market - or creates its own market, in this case - it can take a while before businesses truly begin to adopt it. Now, five years later, data warehouse appliances operate at the center of some of the top global enterprises. Companies like Corporate Express US, Neiman Marcus, NYSE Euronext and Virgin Media have deployed data warehouse appliances as the cornerstone of their business analytics operations.

The concept was quite simple when data warehouse appliances first appeared in the marketplace. Traditional data warehouse architectures were sufficient for storing large amounts of information, but as data volumes exploded in the new millennium, these systems were unable to offer the detailed analysis that businesses needed in order to compete in a timely fashion. SQL queries performed on these traditional data warehouse systems usually took hours, days or even weeks to process. In a world where business agility is key, having information that was weeks out-of-date simply didn't cut it.

Data warehouse appliances offered a new alternative - integrated architectures comprised of database, server and storage. These systems were built to be powerhouses in the realm of data analysis, speeding queries by orders of magnitude when compared to traditional data warehouse systems. Queries that once took days or weeks to process were cut down to seconds or minutes. Data analytics no longer meant getting a view of where your company had been a day or a week ago, but where it was in the present. Data warehouse appliances enabled business to see what was possible down the road, meaning that they could compete on the analytics locked inside their data.

Over the past few years, data warehouse appliances have begun to shift from their role as supplemental to traditional deployments and move into the starter gate in many business environments. Companies like Michaels Stores, a leading arts and crafts retailer, have implemented data warehouse appliances as their primary enterprise data warehouses. Expanding amounts of data and the ability of these appliances to scale to accommodate these large data volumes have emerged as key factors in their adoption. Companies no longer need to deploy lumbering architectures to store tens to hundreds of terabytes of

data and can instead leverage both the storage and speedy data analysis capabilities of appliances to realize their business needs.

### **Streaming Analytics: The Next Milestone in Business Analytics**

As data warehouse appliances continue to gain prominence in business analytics, a new wave of these systems has begun to appear that makes real-time data analysis a true possibility. Just recently, the term “analytic appliances” has come into play with systems that leverage streaming architectures to process data right off the disk as it feeds into these systems. Instead of shuttling data from component to component in a data warehouse environment, these streaming analytic systems have moved the processing power right to the data, so there is the smallest physical distance possible between the information and the analytic components, and processing is done “on stream.”

Imagine if Wall Street firms could run complex risk analyses iteratively, hundreds of times throughout the day rather than overnight, for better trading decisions. Imagine if telecommunications companies could analyze call data records (CDRs) in real time to offer customers the best packages and rates. Imagine if retailers could optimize SKU-level pricing by capturing dynamic in-store data. This new wave of analytic appliances will play a key role in many organizations where fast, in-depth analysis of data impacts business performance.

### **Innovation at the Core: FPGA Inside**

So what has made this evolution in appliances possible? One key component to analytic appliances is the implementation of hardware accelerators to make streaming processes possible. Many applications require high-performance processing of data streams, where massive amounts of image or signal data flowing into a system must be processed immediately. Streaming processing allows you to watch videos on your computer and is also used in industrial controls, medical imaging, telecommunications devices, missile guidance systems and many other applications. Bringing this technology into an analytic appliance architecture can aid in the processing of business data as well, just as it has sped data flow in other environments.

One common hardware accelerator, first used in consumer electronics like video game consoles and high-definition televisions, is called Field Programmable Gate Arrays (FPGAs). An FPGA is a semiconductor chip with a series of internal gates that can be programmed for different tasks – hence the term “field programmable.” Introducing these hardware accelerators into the appliance environment allows the system to analyze only the relevant data for each query, greatly speeding analytics. With gate counts of over one million, modern FPGAs can implement much of the functionality in today’s systems. These large gate counts and the reconfigurable nature of FPGAs make them an attractive component for a growing range of streaming applications.

The analytic appliance architecture is based on a fundamental computer science principle: when operating on large data sets, do not move data unless you absolutely have to.

Analytic appliances exploit this principle by processing data extremely efficiently, as early in the data stream as possible. By optimizing the use of FPGAs and other commodity components, this new architecture has revolutionized the data warehousing and analytic industry, enabling it to deliver tremendous performance in a compact, low-power system that is fast to install and incredibly simple to operate.

These multiple levels of optimization in hardware and software filter out extraneous information as early in the data flow as possible, greatly reducing the processing burden downstream. This approach eliminates the I/O bottlenecks that occur when general-purpose processing architectures are used to examine massive amounts of data, making the analytic appliances dramatically faster than conventional systems. By implementing FPGA technology, companies are able to process only the relevant data with each query, drastically cutting down on query time.

### **Solving the Non-SQL Dilemma**

Until now, many types of analytic processing were simply impossible within a database. Complex analytic problems that could not be expressed in SQL had to be solved outside the data warehouse on a separate SMP cluster or grid computing array. Such non-SQL processing frequently requires large, time-consuming data extractions from the data warehouse – potentially up to several billion rows that have to be exported from the database to the analytics server. This approach works directly against the principal of minimizing data movement – a return to the inefficient world before analytic appliances. Users and their companies pay the price: in the cost and complexity of a separate analytics server as well as analyses based on stale or incomplete data.

Bringing analytics into the appliance is a natural evolution of the performance and innovation that have made data warehouse appliances so popular. The same innovative architecture used for executing SQL commands is now enabled for the non-SQL algorithms used by complex analytic applications. Analytic applications are run on stream – at streaming speeds within the database, with no data movement. By moving analytic functions to the data itself, such massively parallel streaming processing operates on data where it resides, with unprecedented efficiency and a corresponding leap in performance.

### **The Power to Question Everything with Streaming Analytics**

Streaming analytic appliances provide a distinct advantage in a broad range of applications: for spatial analysis, text mining, risk-profiling, real-time pricing, network monitoring, fraud detection and many others. But in addition, the ability to run a complex analysis against a huge live database, without the delays and costs of moving data to separate hardware, opens up new types of analyses previously out of reach. Here are just a few examples of how analytic appliances can be utilized by companies:

Geospatial analytics: Geospatial analysis performs operations such as combining multiple maps or map layers according to predefined rules, or identifying regions within a

specified distance of one or more features, such as roads or rivers. Geospatial analysis is used for solving problems like: “*Find all properties within 10 miles of Hurricane Katrina’s eye path*” or “*Find all properties in Massachusetts that physically straddle county boundaries.*”

Predictive model scoring: Many companies use predictive modeling to finely segment their customers and make real-time decisions about promotions, pricing, fraud and other applications. Whereas this typically involves a time-consuming process that can often take hours, the entire process can now be done within an analytic appliance in a fraction of the previous time, providing real-time offers and promotions to the right customers.

Fingerprinting with hashing algorithms: The Message-Digest algorithm 5 (MD5) is a standard cryptographic hash function with a 128-bit hash value. It is commonly used to store passwords and ensure that files transferred are intact. It is also used in chain of custody document fingerprinting. By performing the hash directly on stream, analytic appliances can run hash algorithms on millions or even billions of records in seconds, which is typically hundreds of times faster than today’s method.

Fuzzy text search analytics: Fuzzy text search analysis uses algorithms that provide a “best guess” of most likely results. One example is the Levenshtein edit distance algorithm, which calculates how many text edits would be required to manipulate, for example, “Denver” into “Dover.” This type of algorithm is used by many text searching scenarios which require analysis of billions of text records, such as data cleansing of names and addresses for marketing campaigns, or national security applications for complex analysis of names in port of entry data.

In an industry where the boat is very seldom rocked, data warehouse appliances shook up the scene in 2002 when they offered greatly enhanced business analytics for a fraction of the cost of traditional architectures. A similar seismic shift is on the horizon in the form of the next generation of these systems: analytic appliances that leverage streaming architectures for advanced analytic applications. While companies are enjoying the benefits of having business analytics faster than previously possible, they will now be able to perform analyses even faster, and in some cases run queries never before dreamed possible on even data warehouse appliances. Companies truly are looking for the power to question anything, and it appears that the capability may soon be within their grasp.

*Justin Lindsey is the Chief Technology Officer of Netezza. He is responsible for overall product direction and spearheads the Netezza Developer Network. Prior to Netezza, Lindsey served as the Deputy CIO and CTO at the United States Department of Justice, and prior to that was the CTO of the Federal Bureau of Investigation.*