

Numeric Components and Data Mining: Part 2

By NAG for DStar Numeric Components and Data Mining Part 2: Case Study - Informix NAG Financial Datablade

During 2001, Informix, the database company which has pioneered Object Relational technology, formed a partnership with a world leader in numerical software components, the Numerical Algorithms Group, to release a unique analytical tool geared for the investment banking industry. The Informix NAG Datablade provides a new and exceptional way to handle time-series data. It includes 50 of the Numerical Algorithm Group's mathematical functions that assist the accurate analysis of high volumes of data. This is also a good example of how numeric components are transforming the capabilities of classic data mining and automated knowledge discovery tools.

In this interview, Terry Ralph, Executive Director, Database Business Development for Informix, explains why numeric components are critical to the success of this new species of data mining tools launched by Informix in 2000.

What are the special challenges of providing data mining tools for the finance industry?

Nearly all commercial and investment banks now rely on the expertise of Quantitative Analysts to guide them to better trades and sales of higher-margin products to their corporate customers. These "Quants" build mathematical models of how a particular security, a complex trade, or an entire market will behave in the future. A key input in this analysis is the historic price of an asset, and it is not uncommon to utilize 20 or more years worth of pricing data with a model. This time series data is used to backtest models to see just how effective they would have been to predict an asset's historical price.

Not so long ago, Quants looked at daily data - opening and closing prices plus daily volumes - to get a snapshot of a financial asset. These daily data sets were typically several gigabytes in size. These were some of the larger data sets being processed for business intelligence. The problem of dataset size has increased by several orders of magnitude because Quants want to plumb tick data. For example, an analyst studying equities now wants a data set containing the price and volume for each and every trade for a particular stock over a number of years. Ten years worth of tick history for global equities is about one terabyte in size.

How does Informix manage this problem of data set size in time-series data?

While others look to fat client solutions, we realized that a server-centric model could provide a unique solution to handling data sets of this magnitude.

Firstly, because we store data in a time series, the size of the data is reduced significantly to around 1/5 of the original size that would ordinarily be stored in a relational database. The Informix solution puts all the data into store contiguously, which allows you to extract the data ten times quicker.

Secondly, we have a unique Real Time Data Loader, which can swallow tens of thousands of ticks per second and make up-to-the-second time series data immediately available for analysis as part of a historical dataset going back many years. Using the Informix NAG Financial Datablade technology, the statistical routines are executed directly on the server. This set-up gives you the advantage of the relatively high-speed link between a disk drive and a server.

This contrasts with the classic model where the data would be held in some sort of database, then extracted and delivered to a client and moved across the network to a fat client. Unlike the Informix system, the link between the server and client is relatively slow. That, coupled with the very high quantity of data that needs to be moved across the network, slows the process down immensely. Furthermore, the Quant is probably using an offline statistical routine, again slowed down by moving off and on line. Ordinary relational databases are unable to capture tick data in real time, so single queries or analyses cannot encompass historic and real-time data.

So the server-centric model is much more efficient, anywhere from 20 times faster to 2,500 times faster when you are dealing with very large datasets. Typically, a Quant is processing data a thousand times faster than they would otherwise and this means that analysis is being done within the industry standard definition of real time. In finance houses, this means that the types of analyses that used to be on desks the next morning are now appearing within a few seconds.

A further benefit of the server-centric approach is that one Quant's analysis can be delivered across an Intranet to many traders, and stored longterm to support compliance auditing.

Why and how are mathematical and statistical routine components important to the success of this model?

Time-series data and statistical analysis go together like bacon and eggs. The Informix relational database handles time series data in an elegant and unique way and it is natural that we pair with the world's expert in numerical calculations, the Numerical Algorithms Group, and include their routines in the software. This delivers an incomparably powerful toolset to anyone who wants to do a statistical rather than a relational analysis.

The real business intelligence comes from the statistical analysis that is done on the time series data. Informix's object relational databases and time series Datablade can handle many different sorts of datatypes, unlike the classic business intelligence databases that only allow a relationship analysis. These sorts of databases, familiar to many as image databases or for video, can give users ways to handle vectors, matrices, lattices and other datatypes. Using the Informix IDS 2000 relational database coupled with the Informix NAG Financial Datablade, Quants get everything they need to do high-powered analysis in one package.

What other benefits are available to users, beyond speed?

Another aspect that comes to bear on developing real business intelligence from tick data is the inclusion of NAG's powerful 3D visualization software tools. Thus, there are four components to Informix' solution: 1) Informix Time Series Database; 2) Informix NAG Datablade; 3) IRIS Explorer for data visualization; and 4) Informix Real Time Data Loader. This latter eliminates the choke that is typical in classic systems when you have very high data rates. Instead, this tool allows huge amounts of time series data to be initially loaded into a memory resident datastore and then very efficiently moved into a relational database store. All this data is available for single analyses or queries. It smashes the usual limit on data loading wide open. We can load over 40,000 stock trades per second where typically other relational databases would only be able to handle hundreds of trades per second.

What other applications can use the Informix NAG datablade technology?

This is more than a tool for finance. It actually can deliver the same sort of analytical power to any industry where there are huge volumes of data that have regular and repeated readings. For example, there are many manufacturing environments that need to take multiple temperature measurements or comparable sensor data and then make sense of it on a real-time basis. So whenever you have the combination of time-stamped data in large volume and the ability to make sense of it with statistical analysis tools, you will be able to use this technology with great success.

In the oil and gas industry we have worked to implement this technology with two major companies, Sensa and Telegnomic. Sensa has a very advanced method of taking temperature readings out of oil/gas wells. Telegnomic has a very good technology for collecting that sort of data, conditioning it, and transmitting it to a central point where NAG technology is used to store, analyze and visualize it before it is sent out to petroleum engineers across the Web. This all happens within seconds of the information being collected from the well, compared to the classic environment where people are lucky to get that sort of analysis on a weekly basis.

ALM Communications Inc,
1454 West Glenlake Chicago,
IL 60660-1802,
773-973-2077,
alm@almcommunications.com