DATA MINING:
NAGGING THAT IT REALLY ADDS UP

by Nancy Cohen

Software marketers promoting the wonders of data mining often use the *Tale of the Diapers* to show what data mining can mean to any merchant. Without data mining, a merchant isn’t even close to leveraging what customers want and will buy. That story is proof.

The *Tale of the Diapers* is about information seekers at a Midwest grocery chain who, in using data mining software to sift through scanner data to analyze buyer behavior, were left with a head-scratcher upon seeing correlations between sales of diapers and beer. When doing their fatherly runs to pick up baby diapers, men seized the opportunity to stock up on beer at the same time. The savvy store marketers were able to profit from that discovery by moving the diapers and the beer closer together.

My, how data mining has grown. In size, complexity, and in software development problems. Data mining is to be found in applications like bio-informatics, web analytics, as well as retail. Businesses large and small are using data mining on some level, whether to simply monitor website traffic or conduct elaborate investigations to discover customer patterns that would otherwise go unrecognized.

We have all witnessed the retail giant Wal-Mart pioneering massive data mining with a 7.5 terabyte data warehouse. Every day, Wal-Mart processes enormous numbers of complex data queries coming from 2,900 stores. Tracking buying trends shelf by shelf and item by item, they perfect their market grasp and supplier relationships.

And we have seen how banking uses sophisticated data analysis techniques for complex global trading, risk assessments, and customer acquisition. Herb Edelstein, the president of Two Crows, a data mining consultant company, has pointed out that the time seems long gone when data mining software is falling short of competitive advantage.

Specialized data mining vendors still compete for market share, while vendor giants like Oracle and IBM jockey for data mining market recognition as well. The Aberdeen Group pegged considerable growth in data mining, saying as of the year 2000, the market was growing at about 200% a year and predicting a $4B market in 2004.

Meanwhile, software vendors providing analytic software are discovering something on their own: Good numbers matter. And they know something that to the non-mathematician sounds like a silly joke: Computers can’t count. But when it’s the Numerical Algorithms Group saying that, you know it’s no joke.

Founded in 1970 as a University of Nottingham project in the UK, NAG team members moved to Oxford and then spread out to its current UK headquarters plus offices in Germany, Japan, and North America, also lining up distributors worldwide. These are the “algorithm people.” A more learned rendering might be that NAG is a group of experts who tame mathematical constructs into terms that can be understood by computing machines. They sell data-mining software components and tools for developers. Their customers range from finance, engineering, and scientific-research firms to commercial software vendors like IBM/Informix, Intel, and PeopleSoft.

NAG’s success is tied to developers’ problems: Computers are inherently flawed in their arithmetic capabilities. While these limitations are always a potential problem, they are especially important in calculations with enormous data sets such as those in data mining applications. With the increasing sophistication of data mining and automated knowledge discovery tools, one nasty ISV hurdle is being able to ensure that software is going to produce consistently correct results.

Take round-off errors as only one of numerous examples (see “When Good Computers Make Bad Calculations: A Cautionary Tale” at http://www.nag.com). A computer can only retain a finite number of significant digits to represent an operation’s results and if result can’t be expressed
exactly, round-off errors can occur. NAG sells the solution: software designed for the rigors of mining massive data sets. Their numeric components are designed to eliminate concerns about the accuracy of computer arithmetic.

Q: NAG has Data Mining Components, and describes these as statistical and machine-learning software intended for developers to integrate into a broad range of applications. You say it can eliminate the weeks and months it often takes to develop, debug, test, and maintain analytical methods. What specific signs were you getting from the real world that got this product going?

A: NAG received requests for data mining functionality from corporate practitioners (AC Nielsen) and software developers, including those at consultancy companies that are offering CRM and various business analytics applications. We polled customers early on in the design process, to see which data mining techniques and interfaces would be most useful.

Q: You’ve acknowledged that there are other companies selling libraries of functions for data mining. What makes you unique?

A: Products overlap in functionality and others have similar platform coverage, but we aren’t aware of any competitors selling a wide range of data mining functions as a static or shareable library with the data cleaning, data preparation, and model building functionality of NAG Data Mining Components.

Q: Any future plans to work with an Open Source group?

A: No. In data mining applications, the code tends to be complex, special-purpose, and users desire assurance of technical support for critical applications. These factors mitigate against Open Source versions of the code.

Convinced of strong demand for algorithms specific to the data mining industry, NAG in November launched its Data Mining Components, a collection of numerical algorithms for data-mining modeling processes. Those processes range from data cleaning through data transformation up to model building. Their customer target: software product managers and developers. The data mining components, including robust algorithms, can be integrated into existing business analytic and knowledge-discovery products. The software is positioned by NAG to take the angst out of having to research methods, code, debug, test, and document routines.

The algorithms can be called in “novice” mode for fast results or in “expert” mode for precise control. Aware that some organizations have numerical experts and a good deal do not, the NAG product also has hyper-linked documentation that guides novice users to data mining solutions applicable to their data. The components are available for Windows, Solaris, or Linux platforms. They can extract data from flat files or OBDC-compliant databases.