



Fermat's Penultimate Theorem and the Conundrum of Big Data

Statisticians and modelers now have a solution.

(May 14, 2002) Fermat's second to last (penultimate) theorem relates to another unsolved mystery - modelers and their use of data. It states that:

Amount of useful data needed = (Available space) * 10

While not as famous as his subsequent theorem, solving this problem has proved far more elusive for the systems industry. The inability to support modelers and analysts and realize the benefits tangled up in today's massive amounts of data has continued to strain the IT/business relationship.

Why does the 'Universal Constant of Modeling' dictate that modelers spend 80% of their time setting up the environment for the other 20% of their work? All they want is access to as much data as possible and to develop and execute their models quickly. Loading data on and off overused machinery takes time and invites a greater and greater chance of error. If a relational database is also involved, the ridiculous time/cost overhead simply multiplies. All this added time and expense is spent preparing an environment that contains bells and whistles not even used by anyone. As budgets shrink or remain flat, there is no extra money to throw at more hardware. So how do you make better use of what you've got?

Why doesn't someone come up with a way to compress all this data (tightly enough that it makes a huge difference) and allow you to execute against it without the need to decompress it to disk first? You could maintain (or even decrease) the hardware budget, while simultaneously enabling the modelers to work with 3 years of data versus 4 months worth. With chip speed increases far outpacing any growth in I/O speed, the compression and decompression work is offset with the improved I/O throughput. For once, you might actually start making business use of the increased chip speed your hardware vendor has been pushing on you for the past few years.

Well, as you may have guessed, someone has developed a product that is already there. Just as Andrew Wiles toiled in obscurity for 7 years while developing and perfecting his proof of Fermat's Last Theorem, so has Corworks labored for the past 6 years perfecting their product for market. Gradually building upon very large-scale successes at Equifax, Sears Roebuck, Acxiom and Mastercard, Corworks is now positioned to capture the world of big data modeling and analysis.

With links to and from 'C' and the ability to work with SAS, the Corworks Knowledge Server (CKS) language delivers a simple, clean programming language with unique constructs designed for analysis and modeling. One of these, the Time Series Structure, takes a near universal modeling requirement, namely processing data across numerous time periods, and with elegance and efficiency removes the programming and testing costs associated with organizing these time periods for quick, clear analysis. A number of other tools built into the product are also available, but it's the simplicity of use, combined with a track record of amazing cost savings, that makes Corworks well worth looking at, especially if you are being squeezed to deliver more accurate and useful analysis from an ever growing universe of data.