Homeland Security Challenges in Data Warehousing and Analytics

Keynote presentation, Data Warehousing 2004, June 7, 2004
(http://dmforum.org/portal/content.asp?contentid=202)

Seth Grimes

Introduction

Good morning.

It's a thrill to be speaking before you. This is a new talk for me although as you can see from the handouts, which include a number of articles I've written for Intelligent Enterprise magazine, I've written on a number of subjects related to homeland security that I will cover today in more breadth and in most cases more depth. I've also spoken, less formally, on today's topic on a number of occasions including for a local data warehousing and analytics group, the Capital Area Business Intelligence Network, that I'd love to tell you about later if you're interested.

I'm going to start by taking issue with the conference organizers. They billed me as an expert on Homeland Security and Intelligence systems. I'm not, not even considering only the topic immediately at hand, homeland security challenges in data warehousing and analytics. Actually, I don't know any experts. Admittedly, I set the bar VERY high, but then a lot is at stake and a false pretense of expertise can be quite damaging. We launched and have conducted the Iraq war based on mistaken analysis and interpretation.

Here I've made what's essentially a political statement just a minute into my talk. That was fast, wasn't it? Politics would have come up sooner or later because homeland security efforts are designed to protect the building blocks of our society and therefore must be woven into the fabric of our everyday lives. So regarding expertise, I'll just say that I don't know anyone who understands both the operational context – that is, the politics – and simultaneously how to apply analytical technology to do more than create a line of defense. And that's our first challenge, responsibly contributing our technical knowledge and skills in a constructive and a socially appropriate fashion, becoming homeland security experts to the extent we can within our own workplaces.

We have to work collaboratively. The field of operations is far too broad for any one person to have mastered. It's broad in geographic extent: it's global, covering nations with a multitude of languages and cultures and ethnicities and religions. It's simultaneously intensely local, requiring (if done right) monitoring and understanding, and with the participation of, small groups and even individuals. And it involves the whole spectrum of organizations between the two extremes, from international to individual.

Next, the field of operations is similarly fragmented. (Actually – and I'll dip once more into analysis of geopolitical dynamics – the word “fragmented” describes the New World Order proclaimed by the first Bush administration in the wake of the 1991 Gulf War as well as any word I can think of.) But back to information technology: There's quite a variety of computing architectures and techniques one might apply in creating Homeland Security and Intelligence systems, many of them quite sophisticated. A path forward has developed only in the last couple of years. Homeland Security analytics is too new for anyone to be considered a master, an expert.

Lastly, many of the supposed experts have, to be blunt, failed, often spectacularly, in major aspects of their tasks. Granted those failures have been due almost exclusively to errors of omissions – due to things that people didn't do rather than things they did – but that means they've been failures of conception: of design, planning, implementation, and execution. And on the technical side they haven't been the fault of any particular individuals. Rather they've been organizational failures, political failures, failures to properly and adequately allocate resources, failure to apply sufficient foresight.

Modern American society tends to throw technology at problems. We'd rather put a machine on the job than deal with root societal issues. And it's information technology first and foremost that we're applying to homeland security problems. The vast majority of homeland security work generates, manipulates, or analyzes data – massive volumes of data with lots of detail. And where there's big data, there are data warehouses and analytical systems front and center along with lots of plumbing to connect those systems, move data around, and integrate and report results.

So given the recognition that organizational issues are key and that application of information technology is and will continue to be a core tactic, our second challenge is to align our technology applications with homeland security goals. In the context of today's forum, that means thinking about how we can use existing, available data and systems to meet homeland security needs, how we can protect our systems that have been created for routine but important everyday tasks, and how we can innovate to meet today's and tomorrow's security challenges.

As an entrée to all that, let's talk about financial instruments.

Evidence

An option is a form of purchase or sale contract. It's the right to buy or sell company shares, agricultural commodities, real estate, and other goods. If you purchase a *put* option, you've bought the right to sell something on or after some future date for a price fixed in the contract. If you buy a *call* option, you can buy the underlying good on or after some future date, again for a price fixed in the contract. Since we're talking about options that can be executed at a future date, this form of option is referred to with the term *futures*.

An option is a *derivative* financial instrument, in essence a wager. Prices in free markets fluctuate and they react to events that are likely to affect the value of the item priced. For people who think they know which way the price will move, an option is a way of monetizing their hunch or knowledge. Through futures, you can make a lot of money from a small but highly leveraged investment.

(Allow me to quote one of my articles that you've been provided as a hand-out:)

> Something anomalous happened at the Chicago Board Options Exchange (CBOE) in the weeks leading up to Sept. 11, 2001. The *Chicago Sun-Times* reported on Sept. 20 [2001] that five days before the attacks, U.S. markets recorded four times daily average trading of options to sell shares in UAL, the parent of United Airlines, and sixteen times average trading in American Airlines' parent.
> [You'll recall the United's and American's airliners were hijacked.]
> "Most experts said the wide swings and the timing indicate some trades were made with advance knowledge of the attacks. "The trading 'is so striking that it's hard to attribute it to chance,' said University of Chicago finance professor George Constantinides. 'So something is definitely going on.'"

Putting it together

The Chicago options trading was one of quite a few mid-2001 clues that, in Constantanides words, "something is definitely going on": visa violations, suspicious international money transfers, foreign individuals attending U.S. flight school without much interest in learning to take off or land an airplane, groups of foreign men repeatedly flying certain domestic air routes on first-class tickets bought with cash. We've learned in recent months from the public release of the federal government's 9/11 commission report that the president was briefed in August 2001 about real threats of attacks using airliners. Then on September 9, – noting that it's important for an attacking force to secure its base by neutralizing potential threats and that the Afghan Taliban were playing host to Osama bin Laden and al Qaeda – the leading anti-Taliban Afghan warlord, Ahmad Shah Massoud, was assassinated.

Looking narrowly at Chicago options-board trading –

- Airline-options trading was measured and recorded. Of course: you can't run a modern financial system without tracking transactions.

- It was understood immediately post-9/11 that the airline-options trading was anomalous. But was that a conclusion reachable *only* through forensic analysis, that is after-the-fact when someone started looking for clues that would have suggested that something extraordinary was underway? No – keying in on "*only* through forensic analysis" – that seems highly unlikely, that we couldn't have detected the pattern in real-time if we had been looking for such patterns. Arbitrage has been around for years with automated systems that monitor financial markets and compare behavior in real-time to nominal, model-predicted behaviors. These systems aim to tease out a divergence between price and value that can lead to profit when price and value are realigned due to exogenous, external factors. So-called "program trading" systems may even respond without human intervention to such divergences by automatically triggering sales or purchases to ensure profits or stem losses or simply hedge the bet that is inherent in ownership of a financial instrument.

- Third point: Should we have expected anyone not linked to the attacks – ANYONE – to have inferred that the trading was a form of profiteering by associates of would-be terrorists? I'd say no. At the time, such a conclusion based solely on that evidence would have lacked convincing precedent.

So we missed clues. Let's not make the same mistakes, and let's innovate, let's figure out how to exploit untapped but promising approaches that could prove useful.

Policy Analysis Marketplace

The Chicago options trading pattern may have contributed to the establishment of something called the Policy Analysis Marketplace, PAM, which would have let traders buy and sell contracts on specific events and derivative indicators. The site administrators would "make a market" by establishing what could be traded. If you own a contract on an event that occurs, you make a profit that depends on what you paid for the contract.

PAM's government sponsor was the Defense Advanced Research Projects Agency (DARPA) Futures Markets Applied to Prediction (FutureMAP) program. DARPA as we all know created the Internet.

Let's take a look at a graphic from the PAM Web site courtesy of the Google archives. **[First slide.]** This futures market dedicated to terrorism research would aggregate diverse data and indicators in a single interface. PAM would serve the role of an analytic portal whose look-and-feel would be similar to the dashboard displays that have now become common in the business intelligence world.

Politics and political insensitivity killed PAM, which died a quick death after word of one of the contractor's illustration cases got out. **[Second slide.]** It's politically unacceptable for the government to sponsor public speculation about the overthrow of an ally as you see they did in this slide. PAM's death was further hastened by the involvement of Admiral John Poindexter, a key figure who had been dogged by outcry against the invasiveness of DARPA's Terrorism [formerly Total] Information Awareness program. (I'm going to talk more about TIA, which was – again past tense – a research project aimed at developing technologies to mine vast, distributed data collections for evidence of terrorist activity.) Given that Poindexter was a Reagan-administration national security advisor whose criminal convictions for his role in the Iran-Contra scandal and for lying to Congress were overturned on a technicality, his PAM involvement helped make PAM a target.

[Third slide.] PAM would have provided a form of collaborative analytics, collecting and weighting input from diverse expert players. But rather than relying on a deterministic, fixed model, PAM would facilitate dynamic "Electronic Market-Based Decision Support." PAM concepts owe much to Delphi Methods, which were similarly developed to synthesize diverse expert opinions in a quantitative manner that would avoid individual bias. To cite a less esoteric comparison, PAM isn't so dissimilar from a market researcher's focus group.

[Fourth slide.] PAM would have used mostly commercially available data including information from the Economist Intelligence Unit or EIU, which, quoting from the defunct PAM Web site "continuously assesses and forecasts political, economic and business conditions in 195 countries" via "a global network of over 500 analysts." EIU "is working ...

to collect and process the data on which the securities in PAM are based, and then to assess the value of the securities when they mature." You can obtain similar data and analyses from sources ranging from the World Bank to vendors such as EcoWin, an economic and financial-market data supplier, or for that matter, Acxiom, Bloomberg, Choicepoint, Dun & Bradstreet, Lexis-Nexis, and the like.

What of PAM's goal of quantifying the importance of indicators that could forecast terrorism and the likelihood of terrorist events? Every organization knows that it should attempt to quantify and mitigate risk by trying to foresee and simultaneously hedging against disruptive events and also preparing disaster-recovery plans. That's standard business practice.

And every significant organization concerns itself with external trends that can indirectly or directly affect its suppliers, logistics, markets, and customers. That's why, for instance, organizations shell out big bucks for Gartner reports that prognosticate about emerging technologies and markets and even assign probabilities to possible outcomes... even if those reports often seem to go unread.

Although PAM was using common techniques, it lacked explanatory power. A high contract value for, say, weakening of Saudi religious fundamentalism suggests a likely outcome without illuminating the events that preceded and led to the end result, that is, without regard for the bidders' reasoning. PAM-based predictions would be opaque, contributing little to the goal of preventing terrorism and lending nothing to the construction of predictive models. Last, it appears that PAM would, like most of today's business systems, have been unable to accommodate unexpected events, which would neither have had been listed on the market nor had the precedents necessary for pricing.

But PAM was on the right track: innovative and non-invasive.

It's important to identify and assess risk, exploiting the wealth of available data and modern, Web-based technologies that can facilitate aggregation, analysis, and visualization. Doing that is another challenge for you, one I'll explore some more through other program examples. **[Fifth, placeholder slide.]**

Interpretation

National security organizations – those that engage in what we now term Homeland Security, the military, the intelligence agencies, an extended community of government, academic, and think-tank analysts – look at broad bodies of clues. Often these are in the form of indicators like those used in PAM, that is, derived, composite values. Indicators are not directly measured quantities. The higher you get up the decision pyramid – executives at the top of the pyramid operate strategically, with wide scope of responsibility and therefore the need to comprehend a broad spectrum of information and understand the significant repercussions of their choices – the higher you climb the decision pyramid, the farther you are from direct measurement and the more you depend on abstraction and interpretation.

Interpretation takes many forms. It can involve reconciling conflicting or contradictory information or choosing to neglect particular types or fields of information –

even whole sectors or domains – and particular sources. It involves understanding correlations and linkages, that is, the ability to infer, forecast, and project. Interpretation can involve assessing the reliability, timeliness, and importance of different facts. It certainly involves deciding how to aggregate and integrate knowledge of disparate character.

An interpretation can involve decisions how to structure data, and present indicators, conclusions, and recommendations for appropriate action. For example, say I'm federal-sector vice president for a systems integrator and I have a dashboard interface with profitability displayed in a speedometer form with red, yellow, and green slices. The color-coding indicates an unacceptable level that is "alarming" in a technical sense, a borderline level, and a desired level. The depicted profitability value is of course not "atomic"; rather it is a derived value calculated from quantities, cost and revenue, that are themselves aggregates of lower-level measurable items. High-level decision interfaces tend to be highly reductive. The aggregate and interpret a wide variety of inputs and a large volume of data into a small set of action points. The more input, the better, right? No, not necessarily.

Total Information Awareness

The Policy Analysis Marketplace was a relative flash in the pan: not a lot of money involved, born and died in the wink of any eye. Technology on the bleeding edge, creativity in general, they work like that. You try a lot of things, some of them stick, some don't. DARPA does that a lot. PAM wasn't their only brainstorm that failed. And it was far from their most notorious failure. That was Total Information Awareness.

You likely know about about TIA: another Poindexter brainchild. The idea was to look at transactional information from every aspect of everyday life that creates an electronic record: banking, commerce, education, medical care, travel, you name it. Poindexter talked about assimilating and analyzing petabytes of data. A petabyte is a thousand terabytes. There are commercial organizations nowadays that have data warehouses with tens of terabytes. Those figures will likely continue to grow explosively with the adoption of RFID – radio-frequency identification – technology for supply-chain, shipping, and inventory control. Add to those figures governmental data related to customs, immigration, licensing, police arrests, commercial reporting, and so on. And add, especially, data generated by satellite surveillance and video surveillance and by monitoring communications traffic of all types – audio, video, e-mail – much of it in so-called unstructured form.

TIA and DARPA's Information Awareness office would have developed technologies to harvest and analyze all that data, but they're dead, killed by Congress in light of privacy concerns and public gaffes similar to those that killed the Policy Analysis Marketplace and half-measure responses to criticism such as renaming the program Terrorism Information Awareness.

TIA's promises didn't wash – the claim wasn't convincing that the government wouldn't actually hold all that data but would find ways distribute the processing and analysis out to the data custodians, that TIA wouldn't be used inappropriately. Personally, I'm all for responsible research. What if DARPA's Information Awareness office hadn't been headed by a lightening rod figure like John Poindexter and if they had convincingly addressed privacy concerns? Well, TIA was highly suspect nonetheless.

I love to relate current events to Star Trek, in particular to Captain Kirk's Star Trek that first promised to "boldly go where no man has gone before" rather than to the wishy-washy Next Generation trekking of that moral-relativist Frenchman Captain Jean-Luc Picard. In one episode, Kirk defeated a genetic superman, Khan Noonian Singh, played by Ricardo Montalban, who lamented that "It would have be glorious." So would TIA, no? Think of the grand challenges: transforming video and audio into a form that can be married for analysis to fielded, numeric data; training machines to discern important concepts hidden in document stores and create linkages between documents; parsing dozens of human languages; identifying terrorists, their supporters, their resources, their recruits; creating global predictive models that reliably warn of probable attacks. "It would have been glorious."

Kirk defeated Khan, but then Khan comes back in one of those forgettable Star Trek movies so the similarity of TIA to Star Trek breaks down. Instead, we can compare TIA to the Hydra of Greek mythology, a creature that generates two heads when one is cut off. TIA work is continuing, albeit in a less grandiose, less unified form. Steve Aftergood of the Federation of American Scientists said that "the whole congressional action looks like a shell game. There may be enough of a difference for them to claim TIA was terminated while for all practical purposes the identical work is continuing." It's continuing through the National Foreign Intelligence Program, through the Advanced Research and Development Activity, or ARDA, through the Department of Homeland Security Advanced Research Projects Agency, DHSarpa, through other organizations. And some of this type of work is not just research: it's being operationalized, for instance in CAPPS-II, more on which in a few minutes.

Last month, the government's General Accounting Office issued a report that found – and here I'm going to quote Robert Pear's article in the May 27 New York Times – the report found "more than 120 programs that collect and analyze large amounts of personal data on individuals to predict their behavior." Fifty-two agencies "reported 199 data-mining projects of which 68 were planned and 131 were in operation.... Of the 199 data-mining projects, 54 use information from the private sector like credit reports and records of credit-card transactions. Seventy-seven projects use data obtained from other federal agencies like student-loan records, bank account numbers, and taxpayer identification numbers." Here are four particular extant programs as identified by the ACLU:

- Verity K2 Enterprise - Defense Intelligence Agency (DIA). Mines data "to identify foreign terrorists or U.S. citizens connected to foreign terrorism activities." (Page 30 of GAO report)

Verity is a search vendor; I covered them among others in my Enterprise Search article in the June 1 issue of Intelligent Enterprise.

- Analyst Notebook I2 - Department of Homeland Security. "Correlates events and people to specific information." (p. 44)

This is another commercial product. That's pretty vague. You can do a lot under that cover.

- PATHFINDER - DIA. "Can compare and search multiple large databases quickly" and "analyze government and private sector databases." (p. 30)

The DIA again. Again vague, covering a lot.

- Case Management Data Mart - DHS. “Assists in managing law enforcement cases” Using private-sector data. (p. 44)

So information-awareness type of activities are continuing. I won't offer any views on their appropriateness but I will observe that the technical challenges inherent in many of them are quite significant:

- scaling up, handling ever increasing data volumes,
- scaling out, patching into distributed data sources,
- scaling across, integrating heterogeneous systems that manage data that originated in a disparate variety of unstructured forms.
- scaling through, deriving structure and linkages – interrelationships – that could not be discerned in isolated systems or through deterministic means.

I find the technologies being developed and applied to be quite fascinating. First, on the scale-up/scale-out end, we have grid computing and heterogeneous clustering via Web Services and the like. We also have a growing trend to fat databases where growing analytical power is moved into the database management system and made accessible through standard SQL interfaces, that is, aggregations, stuff like moving averages. Most recently, IBM and Oracle have implemented the ability to define and/or import data-mining models via SQL and/or PMML, the Predictive Modeling Markup Language, and Microsoft will build on the basics they currently have in place when they deliver SQL Server 2005, the much-delayed *Yukon* release, next year.

Carrying out analyses close to the data is a good thing. There's less data crossing the network, which is good for performance and security. You're also avoiding duplication – the need to maintain a given data item in multiple data stores and the need to multiply define security and usage policies on a given item. And you're off-loading a computational burden to distributed systems that may be dedicated to and tuned for particular tasks.

On the scale-out/scale-across front, we have federated databases and developing metadata management standards and techniques. I'll mention one other advantage to federation: that you can allow a system to use data, for instance by computing derived indicator values or a score from a statistical model identified and fitted via data mining, without providing that system access to the underlying data.

And I said that it's a challenge to create systems that “scale through,” that derive structure and linkages – interrelationships – that could not be discerned in isolated systems or through deterministic means. I'd contrast probabilistic or statistical systems with deterministic ones. Probabilistic systems allow for uncertainty and error in measurement, data reliability, modeling, and so on. They apply statistical techniques to identify the best models for a given problem, to compute coefficients that provide the best model fit, and to create predictions and explore scenarios.

I'd also contrast systems that learn, applying techniques such as neural networks, as an improvement on deterministic system.

I'm particularly fascinated by text mining: software that generates and/or applies taxonomies that embody hierarchies of concepts that can be used to describe bases of knowledge. This is stuff from vendors that include Autonomy, ClearForest, IBM, Inxight, SAS, SPSS, and others. I've researched and written about this technology and you should too, and look in particular into ways of integrating textual analyses with analysis of fielded, numeric data.

I'll flash up a screen shot from one vendors in that space, ClearForest. Their products are being used by a number of agencies working in homeland security and so are their rivals' products. The ClearResearch product can discern names of individuals and organizations, accounting for different languages and spellings, and assess and report relationships and their strength. ClearForest and its rivals didn't originate this technology. Like many others it's an automated realization of something you can do manually if you have enough time and attentiveness. **[Sixth slide.]** I love these diagrams, created by an artist, now deceased, named Mark Lombardi, who became obsessed with graphically depicting complex, obscure relationships. You'll recall that Bill Clinton got into some heat over his relationship ... with the Lippo Group, which made some apparently illegitimate campaign contributions to the Democratic Party. Clinton had a variety of relationship problems, didn't he? Here's Mark Lombardi's mapping of the network of relationships that links Clinton and Lippo and others.

But this is a non-partisan talk, **[Seventh slide.]** so here's a Lombardi diagram – with a detail – that includes a 1990s friend-of-a-friend-of-a-friend path between George W. Bush and Osama bin Laden based on oil dealings. And finally on to the ClearForest screen shot. **[Eigth slide.]** I'm not a shill for ClearForest, but I'll say that they make up for the comparative lack of artistic merit of their relationship depiction with a high degree of interactivity. In case it's not obvious: I really like these diagrams. **[Ninth slide.]** Here's one from another vendor, Autonomy. **[Clippings.]** Here are a couple of diagrams that appeared in print in the Washington Post. One concerns former Defense Department official Richard Perle and his government and private sector involvements, some of which have led to allegations of conflict of interest. The other graphically maps President Bush's fund-raising network. It'll be hard for you to see detail in these graphics, but I'll describe that they do a nice job of classifying relationships and creating a relationship hierarchy and simultaneously depicting the type of relationship: the graphs are multi-dimensional.

So here's your next challenge: don't rely solely on business analysts and subject-matter experts who think they fully understand a problem. Apply statistical and machine-learning technology that complement the experts and look for advanced visualization that can help you recognize the identified patterns. Extend your analyses to other than fielded, numeric data. **[Tenth, placeholder slide.]**

CAPPS-II

Not all innovation is good. Here's an item from a Florida newspaper, the St. Augustine Record, dated March 28, 2004:

> A self-proclaimed psychic's warning that a bomb might be on a Dallas-bound passenger jet at Southwest Florida International Airport prompted federal and local officials to search it with bomb-sniffing dogs.
>
> Nothing suspicious was found on American Airlines Flight 1304, but the delay Friday caused the flight to be canceled because some crew members had exceeded their work hours when the search was finished, officials said.
>
> Doug Perkins, the local Transportation Security Administration director, said the psychic's call was "unusual."
>
> "But in these times, we can't ignore anything. We want to take the appropriate measures," he said.

The Transportation Security Administration or TSA runs a program known as CAPPS-II, the Computer Assisted Passenger Prescreening System. The idea is that suspect passengers will be flagged and stopped from traveling by air – and possibly even be arrested when they try to travel – if they're scored by a data-mining model that uses credit, criminal, and other information to assign a threat level. **[Eleventh slide.]** Here's a slide that while taken from a protest site, Alaska Freedom, seems like a fair depiction although I doubt that implemented CAPPS-II stop-lighting will ever take the form of bears.

CAPPS-II combined with watch lists, which have proven problematical because of abuse that targets non-terrorist political activists and because they apply a very broad brush that paints many travelers as suspects because they share a name with an individual validly on the watch list, is supposed to help keep airliners from being used as weapons.

If you rely on a psychic you're liable to come up with a false positive that flags an innocent individual. One of the biggest criticisms of CAPPS-II and similar program is that they will similarly generate an unacceptably large number of false positives. The ACLU has pointed out, "[e]ven if we assume an unrealistic accuracy rate of 99.9%, mistakes will be made on approximately one million transactions, and 100,000 separate individuals." Congress articulated serious concerns about CAPPS-II's accuracy and effectiveness, about lack of testing and evaluation, about privacy. A General Accounting Office study earlier this year found that the TSA had not adequately addressed most of those concerns. So far as I can tell, however, it's full speed ahead. **[Twelfth, placeholder slide.]**

There are a variety of ways to address accuracy and correctness concerns although they can be costly and degrade responsiveness. Good, clean data helps. But it's well known that the commercial data used for many of these security decisions is far from clean. The marketers for whom commercial data is collected can tolerate a relatively high proportion of errors. Government data and analyses aren't always perfect either. Think of Portland Oregon attorney Brandon Mayfield, who was accused by the FBI of complicity in the Madrid train bombings based on a fingerprint found on a plastic bag that contained bomb parts. The FBI locked him up, a decision possibly encourage by Mayfield's being a Muslim convict, and held him for two weeks despite Spanish government insistence that his print didn't match. Mayfield was freed only after the Spanish arrested an Algerian suspect.

Redundancy is chief among the approaches one might apply to improving accuracy. Find and apply several ways to evaluate a situation, that is, set up a high standard of proof.

But redundancy is expensive as is clean data, which might not be available in any case, as is careful attention to design and testing. I'll close this accuracy discussion by saying that I'm glad that government and public sector organizations – Congress, the General Accounting Office, the American Statistical Association, the Association for Computing Machinery, and others – are taking on these problems.

Department of Homeland Security

I've talked so far about two DARPA initiatives and one, CAPPS-II, run by the Transportation Security Administration. I haven't mentioned except in passing TSA's parent, the 800-pound gorilla of a discussion like this one, the Department of Homeland Security. You all know the basics: established by the Homeland Security Act of 2002, consolidates 22 agencies from a variety of cabinet-level departments – the Customs Service from the Treasury Department, the Nuclear Incident Response Team from the Energy Department, and so on – with a couple of them, the Secret Service and the Coast Guard, running fairly autonomously

DHS has a handful of top-level directorates, one of which is Information Analysis & Infrastructure Protection but all of which necessarily rely heavily on information technology to run their operations and achieve their goals.

Let's talk about one particular DHS program, one that's gotten a lot of press in recent days because it's a homeland-security keystone and because it's big: big in impact, big to the tune of a projected $10 billion over ten years, big in its potential to change American culture. I'm referring to US-VISIT, United States Visitor and Immigrant Status Indicator Technology . US-VISIT boils down to creation of a so called “virtual border” with the idea that it's easiest and most effective to interdict terrorists before they enter the United States. The border is virtual because it will exist largely in cyberspace.

What's US-VISIT? According to the DHS, “the system will be designed to ... collect, maintain, and share information, including biometric identifiers, through a dynamic system, on foreign nationals.” DHS wants the US-VISIT system to “be capable of capturing and reading a biometric identifier” and of “scanning travel documents and taking fingerprints and pictures of foreign nationals, which then could be checked against databases.” DHS documentation continues, “At a minimum, the US-VISIT system will utilize the existing fingerprint and photographic technology. Other biometric identifiers, such as facial recognition and iris scan, are still being studied.”

US-VISIT is important. The General Accounting Office – this is the third time I'm citing an investigation of theirs – reported last September that "the program is a very risky endeavor.... The missed entry of one person who poses a threat to the United States could have severe consequences.” Here's a warning about false negatives – missed real threats – that is complementary to the the false positives issue I discussed earlier under the CAPPS-II heading. Both should be concerns.

I'll quote business columnist Steven Perlstein, who wrote in last Friday's Washington Post (“How Accenture Seized Tomorrow,” June 4, 2004, page E1) that the proposal managers at the winning US-VISIT contractor, Accenture, “figured that both Lockheed and

CSC" – that's Lockheed-Martin and Computer Sciences Corporation – would take a tech-centric 'systems engineering' approach in preparing their bids. So the Accenture team decided to differentiate itself by focusing on the management and operational challenges involved in checking 125 million border crossings a year, working from those to its technology solutions." Alignment. Perlstein continued that the Accenture managers "offered to put a sizable portion of their fee at risk, to be paid only if performance criteria were met."

To digress for just a moment: Here's a chance to bring in what's probably the hottest topic in the analytics world these days, Business Performance Management, or BPM. BPM is about pulling data from your data warehouse and your operational systems, creating key performance indicators (KPIs) based on methodologies such as Balanced Scorecard, and displaying the those KPIs in dashboard displays. The technology is really about performance measurement and reporting rather than performance management, but regardless, those Accenture managers are recognizing that the federal government is nowadays placing huge importance on performance-based contracting. So we have another challenge: Apply organizational and performance-management best practices and emerging BPM techniques to encourage optimal results.

Performance methodologies like Balanced Scorecard look at a variety of indicators of organizational health but they also look at the unity, at cross-functional indicators. So here's where we might bring in a few terms that are commonly used in the technology world. For instance, "stovepipe" is used to refer to a process with no connectedness to other processes. A stovepipe process is an example of a "vertical" process, one that incorporates a number of different functions to move something from one end of the pipe to another. But all the time the contents of the pipe are insulated from the outside world – even invisible and inaccessible – at any point except the extremes, the entry and exit points.

Let's take it as a given that a stovepipe system, as efficient and effective as it may be for some purposes, may not deliver all the benefits it might were its design different. First and foremost, there may be users other than those for whom the system was designed who could profitably use it in total or in part but who don't have access.

The creation of the Department of Homeland Security was all about opening up stovepipe systems, about bridging islands of information, about coordinating similar functions among different agencies and also with the private sector, which controls and operates the majority of America's critical infrastructure. DHS does have to get its own act together however. They have lots of back-office systems to integrate, and then they, a security-focused organization, go and do something that produces chuckles: they chose to standardize on the Microsoft Windows platform, which has had its share of security problems, hasn't it?

Homeland Security involves all levels of government: local, state, county; private-sector organizations; and different functional entities within government and the private-sector. DHS's coordination function does not preclude these other organizations from working together outside the DHS umbrella, however, nor should it, and DHS isn't the only player in the game, not by any means.

I talked about some DARPA programs, and outside government, as observed, the bulk of the nation's critical infrastructure is owned and operated by the private sector. Then there are the local agencies that must prepare for and deal with crises – the "first responders" – and the states. The US after all is a federal system. Our government is much less centralized than those of most countries, whether industrial powers like France or Germany or Japan or, needless to say, dictatorships like China or Saudi Arabia.

So we have to connect all the players. You do that with networks.

What's the hot network topology of the last half dozen years and, if you think about it, the three dozen years counting from 1969? I'm talking computing networks.

- Large, monolithic systems: we shouldn't even consider them, right? A mainframe supporting a bunch of dumb terminals: that's not even a network, and so '70s, gone the way of disco, gone the way of gas lines. Yeah, well just as we still have John Travolta and just as we're having a new, improved fuel crisis, we still have mainframes. But nowadays they're not so monolithic. A modern mainframe – and actually I'd extend my description here to Windows Citrix hosts – may support hundreds or thousands of jobs each processing stream running in its own virtual machine, a technology IBM pioneered back in the '70s. And those machines are networked, they're no longer isolated.
- Small monolithic systems? I'm talking about PCs. They're also all networked nowadays, and of course the downside is that when Bill Gates turned my 79-year-old mother and my 9-year-old son into systems administrators. As a result we have a fertile breeding ground for malware of all sorts that has serious implications for all computer users.
- Client-server is no longer a hot architecture. We're into the '80s and early '90s now: a large host carrying out functions that are computationally intensive or require a lot of memory or storage with other machines, usually PCs, running the user interfaces. Client-server is good for some functions but evolved into...
- N-tier architectures with application servers and middleware connecting clients and servers. Now I'm finally going to answer my question.

The hot network topology in recent years is peer-to-peer where nodes are multiply wired, that is, with direct connections to several other machines and where there is a potentially huge number of paths between any two nodes. N-tier and peer-to-peer architectures are embodied in the software layer that brings us the World Wide Web

Matrix

What's the least useful network? It's one where people won't join or they drop out and leave the network. And that's Matrix, the Multistate Anti-Terrorism Information Exchange, a sort-of state-based TIA backed by the Department of Homeland Security and the Department of Justice. In an application of the *network effect*, the idea that the value of a network grows in proportion to the square of the number of participants, participating states would share and integrate and mine data from disparate public and private sources.

Here's a Matrix splash graphic from their Web site, matrix-ag.org. **[Thirteenth slide.]** That site says the Matrix "pilot project leverages proven technology to assist criminal investigations by implementing factual data analysis from existing data sources and integrating disparate data from many types of Web-enabled storage systems. This technology helps to identify, develop, and analyze terrorist activity and other crimes for investigative leads. Information accessible includes criminal history records, driver's license data, vehicle registration records, and incarceration/corrections records, including digitized photographs, with significant amounts of public data records." Sound familiar?

I'll underscore that wording: "factual data analysis." The core Matrix application is called Facts, the Factual Analysis Criminal Threat Solution. I feel like someone's playing with us in calling Matrix data analysis "factual."

Note the five participating states. The Matrix program originated in Florida in the wake of the September 11 attacks. At its peak, the program had 16 member states. States including most pointedly New York have withdrawn initially because of privacy concerns and then concerns that the program had lost its forward momentum; that it would be unable to attain a critical mass of participation. The technical architecture, which utilizes a central data repository residing in Florida, created a legal barrier for some states that were not allowed to ship data outside their control. Again, true peer-to-peer with no master system is a superior architecture. That the founder of the company that is the lead Matrix developer, Seisnet, was involved in '80s drug smuggling didn't help matters.

Matrix isn't the only program of its type however. It was preceded by a the Regional Information Sharing Systems (RISS) Program and, in particular, RISSNET, an intranet currently used by thousands of law-enforcement agencies at all levels of government for a variety of purposes. Matrix uses RISSNET as a communications backbone. **[Final, placeholder slide.]**

Challenges

I've perhaps sounded fairly negative about the homeland security situation given that I've used as examples several failed or failing programs and reported harsh evaluations of others. We're all interested in seeing successful homeland security efforts, which will be strengthened by continued open discussion and frank criticism of approaches.

I've come up with a number of homeland security challenges for data warehousing and analytics. We're working in an environment that's highly politicized with asymmetric, non-traditional threats and operations and systems that are extremely heterogeneous and distributed. There are very significant cultural issues involving coordination and cooperation, business process alignment, and privacy and confidentiality. And chief among the technical needs are new, innovative ways of collecting, integrating, analyzing, and delivering information.

With this backdrop, I'm going to close with a few more challenges, some of which restate earlier ones and some of which are essentially conclusions to points I raised in discussing programs and technologies:

Challenge #1: Do more with what you have. Filter it better. Aggregate better. Weight it better. Make connections. And pay attention.

Challenge #2: Realign systems and organizational structures with objectives.

Challenge #3: Accelerate. Work in real time. Move the analyses closer to the data. Distribute processing; federate databases.

Challenge #4: Expand your scope. Assimilate information from new sources and in new forms, in particular "unstructured information." Detect patterns. Create models. Predict.

Challenge #5: Expect the unexpected. Adopt non-deterministic, statistically based approaches that accommodate, that anticipate, uncertainty (measurement errors, processing errors, analytical and interpretive errors).

Challenge #6: Prepare for disaster, for abrupt change. Model, conduct scenario analyses, design-in redundancy with fail-over capabilities.

Challenge #7: Share appropriately and judiciously. Protect. Build in policy-based data and systems security.

Challenge #8: Design in performance metrics. Monitor and optimize. Be accountable.

I'd say that the key challenge is one I did state earlier, that each of use should responsibly contribute our technical knowledge and skills in a constructive and a socially appropriate fashion, assuming personal and organizational responsibility for security concerns to the extent we can within our own workplaces.

Thanks very much for attending, and I'd be pleased to take a questions or let you present your views on the topic in the time remaining.

Publication references – selected Intelligent Enterprise magazine articles by Seth Grimes

Futures Shock, October 10, 2003: DARPA's ill-fated "terrorism futures" market (http://www.intelligententerprise.com/031010/616decision1_1.shtml)

Shared Risk, Shared Rewards: homeland security and IT innovation, September 1, 2003. (http://intelligententerprise.com/030901/614feat2_1.shtml)

Government Equivocates on TIA (Terrorism Information Awareness), July 18, 2003. (http://www.intelligententerprise.com/030718/612news3.shtml)

Look Before You Leap, June 17, 2003: will government application of analytics to national-security problems be effective? (http://www.intelligententerprise.com/030617/610decision1_1.shtml)