: What is the potential size of the NSEERS database?

The U.S. Immigration and Naturalization Services (INS) is creating a number of very large databases to support a variety of operations and processes. In addition to the NSEERS transaction database discussed in DSS News, Vol. 3, No. 19, INS is developing the Student and Exchange Visitor Information System (SEVIS). SEVIS is an Internet-based system that will be accessed at U.S. Ports of Entry and by more than 1900 schools and colleges. Biometric border crossing cards will also be required of Mexican border crossers as of October 1, 2002. These projects are massive in scope and it is not clear how data will be shared between systems. This Ask Dan! continues the discussion.

How do we "size" these databases? Is a data warehouse needed? What platform and software is needed? How will this data collection effort support decision making at INS? Some of these questions were addressed in my prior column, but Marc Demarest, former Chairman and CEO DecisionPoint Applications and current President of Noumenal (http://www.noumenal.com/marc/), offered his analysis and insights and I accepted. Marc began with my assumptions for the NSEERS transaction processing system -- 35 million visitors per year, 45 KB for a fingerprint, 10 KB for a photo and 5 KB for alphanumeric string data per visitor.

Marc writes "I think we need to double or triple your alphanumeric string count. For US citizens and green card holders, the 5K is probably right, but have you seen the forms foreign nationals have to fill out? They're huge, and, after all, this is John Ashcroft we're talking about, so, by the time we add in the foreign national's 'home information' and all of the data on where the foreign national will be traveling in the US and what they will be doing and who they will be seeing, I'd bet the foreign national alpha information tops 20K easily. So let's say 12K as an average for alpha data."

Second "I didn't see a discussion of what would be captured on exit, but something will surely be captured, yes? That would be the easiest way to catch visa overstays. And if Ashcroft is worried about people masquerading as other people, he'll capture about as much on the way out as he did on the way in. Let's say a photo, a fingerprint, and 5K on the way out."

"Now, how would such an application work? The 'decision support' is going to be largely automated, I'd imagine. I'd assume this application is going to (a) compare photographs, (b) compare fingerprints, (c) analyze the alpha data provided and then (d) weight the outcomes and (e) make a recommendation to the official at point-of-capture based on that data. That is the system model (at least I hope it is). Now, they may also do other things with the data, including uploading it into other (INS, NSA, DCA) systems for other kinds of analysis, but this is a closed loop capture-and-analyze system, I'd bet, not two systems: (1) a transaction processing app and (2) a DSS app that is loaded from the TP app according to a schedule."

"Since the data in the database is analyzed programmatically and not by a person, it doesn't have to be inherently legible at the schema level, so we're probably not talking a 'star' style schema -- we're talking some kind of normal form schema. Raw data loaded into any schema creates a loaded set size larger than the raw data -- because of DBMS storage mechanisms, indexing overhead, etc. They'll have to be using a conventional RDBMS because this is a INSERT-AND-QUERY system: Teradata- and nCube-style DBMSs and OLAP engines won't take the INSERTs fast enough or elegantly enough. For a star implemented in a conventional RDBMS, one usually sees a 2.5X growth in the raw load set size once it's loaded, and I think it'd be close to the same growth factor in this case: maybe 2.8X. We're also not talking about needing to maintain a lot of history in this system -- the system-of-record for pictures and fingerprints will be elsewhere, because that system

will be used for multiple purposes, and data will be migrated out of this system into the real 'data warehouse' for the Department of Homeland Security as soon as an individual alien's entry-exit loop is closed, so I'd bet there will never be more than the equivalent of 1 year's worth of data in the system. In other words, this system won't be extracted INTO; it will be extracted FROM."

As an aside Marc noted "The most difficult technical bit for the system will be indexing strategy: the more indexes they add to cut query time, the longer insert time will take. The fewer the indexes, the longer the complex set of queries they are going to have to run will take."

Based on his assumptions and analysis, Marc calculated NSEERS is "about a 12 TB system ... easily within the range of Oracle running on a nice cluster of Sun or IBM SMP/NUMA boxes." Marc concludes "You're right, however, that in the final analysis it will be hard to implement." Thanks Marc for letting me quote so extensively from your analysis. A 12 Terabyte database is huge and the more I reflect on the INS projects the more I can see the databases expanding in size. Perhaps these two Ask Dan! columns will stimulate more thinking and discussion about the important decision support issues associated with monitoring visitors to the United States. These new INS systems are mission critical and the projects provide us the opportunity to think innovatively about providing decision support from very large databases. As always your comments and questions are welcomed. If you want a challenge, reflect on how you would support decision making at INS. If you teach DSS or database, try asking your students the questions raised in DSS News, Vol. 3, Nos. 19 and 20.

References

Demarest, Marc, Email message, Monday, September 16, 2002.

Power, D. J., "Is it feasible to track all visitors to the United States and then build a Data-driven DSS?" DSS News, Vol. 3, No. 19, September 15, 2002.

The above is based upon Power, D., What is the potential size of the NSEERS database? DSS News, Vol. 3, No. 20, September 29, 2002.

Author: Daniel Power
Last update: 2005-08-07 11:27