ENTERPRISE INTELLIGENCE: DATA FINDS THE DATA AND RELEVANCE FINDS THE USER

NEXTGENS TECHNOLOGIES • SEATTLE, WASHINGTON • DECEMBER 5-6, 2006

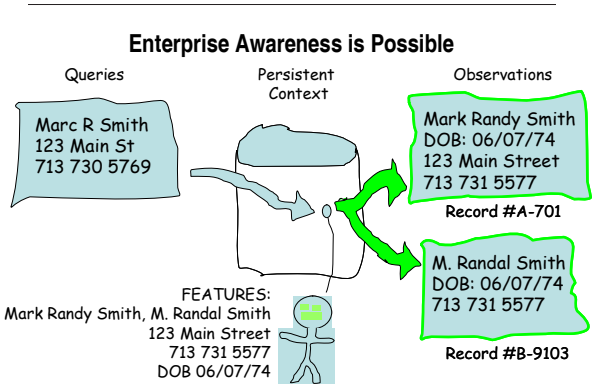
TTI VANGUARD
The advanced technology forum for senior executives

"There has been an error along the way that data is one thing and queries are another. I think that if you treat them more similarly you get a different kind of result."

"Enterprise intelligence requires consistent context. It is kind of like you need a brain to think," analogizes Jeff Jonas, founder of Systems Research and Development (SRD) and now of IBM Entity Analytics. Yet context across most enterprises is anything but consistent. Consider the case of keeping track of people—important when rooting out fraud, hunting down terrorists, or even maintaining an up-to-date customer database. Even minor variations in the spelling of a name or address, transposition of month and day in date of birth, or empty data field could result in mistaking a costly connection or cross-identification. In Las Vegas, where Jonas makes his home, mistakes can quickly run into the millions of dollars. For instance, a dealer and a high roller might be in cahoots, with the player switching decks midgame, resulting in a high-stakes payout. From the outside, it might look like good luck, but if the casino had recognized that the dealer and player shared an address and phone number, they would have been forbidden to sit down at the same table with chips in front of them. Although, in this example, both sets of personal data were held by the casino, they were kept in separate databases, causing "perception isolation," and were thus not co-identified, much to the organization's detriment. "The state of the union is corporate amnesia," Jonas asserts. To solve problems of this sort, he developed the nonobvious relationship awareness (NORA) system to correlate seemingly distinct records. A recent extension demonstrates that it is also possible to perform correlations solely from anonymized (encrypted) data. His basic tenet is to treat data and queries on an equal footing, placing them within the same data structure for continuous, real-time comparisons with incoming streams of data. In the case of anonymized records, he works his magic by attaching a finite number of variants (e.g., Robert for each of Rob, Robby, Bob, etc.; or 05122006 for each of 5 Dec 2006, 12/5/06, 05/12/06, December 5, 2006, etc.) prior to submitting the cleartext into the meat grinder of a one-way hash function.

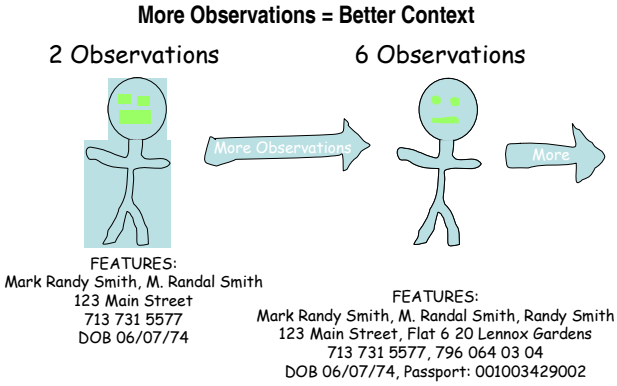
"If you do not first treat a new piece of data like a question, you will not know if it matters until someone asks," says Jonas, but the asking might not happen until sometime in the future, when that piece of data would have been long forgotten. However, in Jonas's system, each datum—and each query—remains active throughout the lifetime of the data store. In the examples Jonas presents, where do data originate? From transactions, each occurring in its own timeframe: a prospective employee submits a job application, a miscreant commits fraud against the company, and a member of the board gets married and moves to a new address. Clearly the enterprise would benefit from knowing that job applicant Mark Randy Smith—of 123 Main Street, phone number 713.731.5577, and birthdate 06/07/74—has something in common with M. Randal Smith, born 06/07/74, who can be reached at 713.731.5577 while on parole following his fraud trial, and that this self-same person just shared nuptials with the board member whose new phone number is 713.731.5577. "People identify themselves in different ways in different places, and the identity itself is not observable," notes Jonas. The sooner the company recognizes the connections, the better, but the siloization of data in most organizations precludes the growth of enterprise intelligence—until NORA and its offshoots.

In the *before* version, a query for Marc R Smith, 123 Main St, 713.730.5769 would at best register a possible correspondence with the innocuous record of Mark Randy Smith, but not with the others that indicate obvious risk to the enterprise. The *after* version—that of Jonas's NORA—preconstructs the context by assembling the various observations across the entire enterprise and persisting the context, thus readying it for an incoming query or new data, and thus enabling enterprise awareness. "Because it has been precomputed and persisted, the query now finds what you persisted, and you end up finding all the observations in the enterprise," explains Jonas.



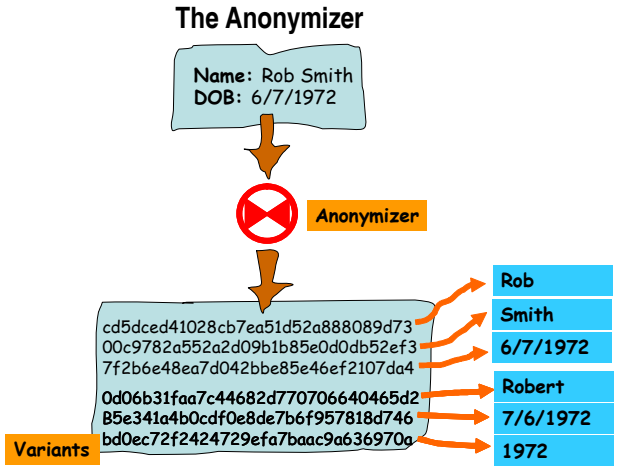
Similarly, by treating each incoming piece of data (i.e., each perception) as a query in its own right, and each query as a datum, all persist and have the opportunity for an intelligence-enhancing future collision. With queries and observations persisting together, not only do queries find data, but data find data, data find queries, and queries even find one another—all in real time as streams of data and queries enter the system. "I'm not talking about stored queries held in

separate lists where on weekends it gets banged against the data," says Jonas, "I'm talking about putting the query and the data in the same space." When matches occur, data types in the persistent context point back to the original observation in the same manner that the Dewey Decimal catalog entry points to a book, and query types point back to the source of the question.



"Just date of birth would mean nothing, just name would mean nothing, or just address would mean nothing. It is the plurality of things, and you have to look for things that are true and things that are not true."

The question arises of whether such a data-accumulating system can scale. "I have seen one of my systems with 3B rows of data describing 600M unique people, and ingesting 2000 new observations per second, streaming," reassures Jonas. While the larger the system, the greater its value, greater too is the risk associated with data theft or misuse. Looked at differently, the larger the system, the greater its value when repurposed. To address issues of this sort that touch on privacy and civil liberties, Jonas sought to perform his analytics in an anonymized data space instead of on cleartext data. While the use of variants helps the process of associating unencrypted records, it becomes essential if the products of one-way hashing are subject to cross-correlation. "The names and addresses and phone numbers are not human-readable and not mathematically reversible," explains Jonas. "They do have the Dewey Decimal system—you know who produced it—but you have to go back and ask them for it, and in a national security setting that means you have to present a Foreign Intelligence Surveillance Act (FISA) subpoena."



Since hashes are wildly different for character strings that are nearly—but not quite—identical (e.g., "Rob " (note the trailing space) and "Rob"), Jonas achieves useful consistency by adding variants, as previously detailed and as depicted in the accompanying image. "You don't throw in all the possible variants before you anonymize," says Jonas. "Instead, you do rooted names." If every instance of Rob, Robby, Robbie, Bob, and Bobby is not only encrypted in its own right, but also appended with the hash for Robert, then their anonymized records will have a mutual point of contact. (Note that a reload might be necessary when novel variants first appear.) With a sufficient number of such contact points—with the criteria for sufficiency defined by a set of rules appropriate to the context—discovery ensues, and it does so without disclosing the cleartext nature of the data, the queries, or the matching attributes, even to the librarian function. "It is the librarian's function to report that the vendor and the employee live together, but the librarian cannot tell you the name or the address or the phone number. All the librarian can see is the pedigree of the Dewey Decimal system." That is, when the librarian encounters a match that the rules suggest bears significance, pointers lead back to the source data and/or questioner(s), and those with need-to-know status are duly informed or connected, as per the governing rules. "The purpose of anonymization is to hide from the librarian the collection of data," says Jonas. "You are trying to reduce the transfer of data." Implicit in this scheme is the separation of roles and responsibilities between those who perform the analytics and those who perform intelligence with the sensitive cleartext data. Further protecting privacy is the oft-used practice of adding "salt" to the hash, whereby a community of interest will append each datum with a short character string to alter the output of the hash function from identical input by another community of interest, thus inhibiting dictionary attacks against the database.

Jonas's systems, both cleartext and anonymized NORA, have been implemented in large-scale real-world applications, including in governmental intelligence in the USA and abroad. One such application is cross-compartment exploitation, where two groups under a single umbrella organization each possess sensitive data; for instance, one is an anti-money-

"If information can be shared in an anonymized form whereby a materially similar result can be achieved, why would an organization share information any other way?"

laundering division and the other works on counterterrorism. Although the divisions cannot freely share sets of data, with anonymized analytics they can mutually discover common records and proceed to cooperate in a targeted fashion. "They are not worried about attacks, because it is done internally," says Jonas, "and they know what FISA records to ask each other about." Jonas emphasizes that the goal is not to sidestep any legal constraints, but rather, "when there are data that are already being transferred between organizations, this is an example of something you can use that is just better."

Essential to the efficacy of a system that purports to resolve identities is the ability to scale. Jonas has constructed his flavor of analytics to enhance its ability to scale gracefully. He relies on rules-based determinism in lieu of probabilistic thresholding to avoid the need to initially train the system on the data set of interest—and to retrain it as the data set evolves, grows, and changes in character over time. To avoid error creep as new information updates and corrects prior analysis, real-time assessment of incoming data (and queries) yields modifications to the persistent context, both for the primary and associated entries. Moreover, this is done with sequence neutrality. "Whatever order the data arrive, your end state is the same," says Jonas. "Most systems lack this. If you get records *A*, *B*, and *C*, in that order, you get one end state; if you get *C*, *B*, and *A*, in that order, you get a different end state." By building sequence neutrality into his analytics, Jonas has effectively circumvented the greatest challenge of unchecked data set growth. As an increasing amount of data is incorporated into the persistent context, identities that initially appeared dissimilar begin to coalesce into distinct entities, thus truncating the expansion of the database. "As it loads new data, maybe some observations are of new people, but more often a new observation would be the same as somebody it knew," he explains. "You end up having this collapse phenomenon, where, the more observations you get, you end up overstating the universe and it starts to collapse." Taken to the extreme, a large enough collection of records brought together with effective analytics should, in the long run, enumerate the extent of independent individuals described by the system.

"If somebody asks about Billy the Kid the bank robber and somebody else asks about Willy the Old Guy who used to rob banks, and there's no data, the queries find each other, and then you can put users into communication."