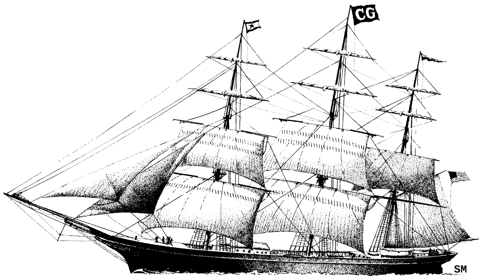
THE CLIPPER GROUP
Navigator™



Navigating Information Technology Horizons℠

Published Since 1993 Report #TCG2005047 July 28, 2005

IBM DB2 Anonymous Resolution Restores Privacy to the Need to Know

Analyst: Anne MacFarland

Management Summary

Shakespeare once wrote, *What's in a name? A rose by any other name would smell as sweet*. Similarly, in ascertaining that someone is who they say they are, the moniker is only the most convenient and easily-alphabetized hook. The salient points of identity, like the sweet scent of a rose, come from its background, its *provenance*. When more of us lived in small towns, and companies partnered less and evolved more slowly, more was known, either directly or through trusted sources. Lies, though easily promulgated, were usually exposed easily. Now, most of us are exposed daily to strangers as a part of city life. Travel is cheap, and the world is a more transparent and a smaller space conceptually.

Globalism affects all businesses, regardless of their scope. Small niche businesses have added a Web incarnation to provide a low-cost channel to deliver information about goods or services, as well as to provide a buying opportunity. Globalism offers great promise, but also comes with some clear threats. In an e-business, like any environment where a large component of strangers are a routine part of life, quick verification of a person's identity is important – but this capability often runs up against concerns of both privacy and confidentiality.

IBM bought a company named SRD in January of 2005. Based in Las Vegas, SRD had its roots in the world of casinos, providing technology that allowed properties to recognize black-listed or criminal patrons with murky or changing identities. Their identity and relationship resolution software could determine, by aggregating information from various sources, whether a "new" customer was, in fact, an old and well-known troublemaker.

IBM felt that the software that is useful to casinos could be useful more broadly in organizational space. This kind of identity resolution provides the quicker time-to-market and increased organizational resilience through multi-sourcing supplies and partnerships that globalism offers enterprises of all kinds. You only get benefit from this larger world if you know, for sure, with whom you are dealing. But, first, IBM had to come to find a way to protect the privacy of individuals (and, pragmatically, organizations as well), before the software would be socially acceptable to the world at large. Electronic records offer endless occasions to snoop – but they also offer a way to compare the entity characteristics that, aggregated, provide the characteristics of identity without exposing names. Any ID card that gives enough information to determine identity can also be stolen or misused. Central registries are high-upkeep, and beg to be politically misused in a variety of ways. But, **in open societies and outward-facing institutions, there needs to be a way to tell that a person is who they purport to be without violating that person's privacy**. This is not possible in a world of paper lists – but it is in a world of electronic records. If you irrevocably hash the primary identifier (usually a name) to obscure it, and compare only entity characteristics, you can still determine, as a rose lover can determine the variety of a rose by its color and scent, who is what, and by comparison, if someone is who they say they are. By this comparison and match of characteristics, organizations can share anonymized information to gain insight without having to relinquish control of source data.

Sounds a little arcane? Well, yes, but it is well worth exploring, because it could be a way around the rip tide of justifiable concerns about privacy that is eroding the benefits of the inexorable flood of globalism.

IN THIS ISSUE

- IBM DB2 Entity Analytics 2
- Some Technical Details 3
- Some Uses of Anonymous Resolution 3
- Conclusion 3

IBM DB2 Entity Analytics

DB2 Anonymous Resolution is built on the two modules of the original SRD product, now called *Identity Resolution* and *Relationship Resolution.* Together, the three modules comprise *IBM DB2 Entity Analytics*. Together, they provide a way to facilitate information sharing in environments where privacy and security concerns have previously precluded such activity. They protect privacy and confidentiality. They let an enterprise keep control of its data, and, in that process, define clearly who owns what data with all the attendant responsibilities to protect it, something that has been less than clear ever since the introduction of easy photocopying. This is much needed if we are to benefit from informatics and other massive information projects like the genome, without compromising the concepts of privacy that underlie most human societies and commercial associations.

IBM DB2 Identity Resolution

Identity resolution resolves ambiguities of names and nicknames and of changeable addresses, zip codes and phone numbers that leave most of us festooned with historical remnants inconsistent and contradictory information. Within the context of a repository, it can merge identities that are determined to be the same person, yet keep separate identities of persons with the same name and address but different characteristics, (such as John Doe, Sr. and his son, John Doe, Jr.). The pearl of wisdom within this product is the concept of *full attribution*, which preserves all information about where identity elements came from (sort of like *provenance* for museum objects). Full attribution helps sort out deterministic elements from the clutter of conflicting information. Full attribution also keeps the ownership of information clear.

IBM DB2 Relationship Resolution

Relationships provide still more deterministic elements of who a person is – or is not. Over the years, most of us accumulate roommates, business partners, spouses and/or children, and billing histories with various businesses. The more relationship elements we aggregate, the easier our uniqueness is to prove, given access to the necessary information[1]. All this relationship information forms the basis for the second module, *DB2 Relationship Resolution*[2]. Relationship resolution software can discriminate between apartment houses and single-family dwellings (co-residence in the latter being more deterministic of a relationship). It can recognize back-door neighbors with addresses on different streets. Since no information is thrown away, each entity identity becomes a rich broth of many diverse elements that the software can use to find commonalities and linkages to other unique identities and organizations.

Why There is a Need for the Third Element

The development of a rich soup of entity information within an organization may be an excellent thing, but when it comes to sharing or exposing this knowledge beyond its bounds, all sorts of alarms go off, mostly because the information is keyed to a name. This heady brew of aggregated, fully-attributed links begs an anonymizing element to make it broadly usable:

- Without compromising the privacy of individuals or organizations
- Without exposing irrelevant information, and
- Without exposing the organizations that contribute information to an enormous liability risk.

There are a number of cases, as in medical research, where it is the characteristics that are important and the identifier is important only as the point of aggregation that associates different characteristics in a meaningful way. There, obscuring patient identity renders a medical record more generally useful. In some circumstances, like the due diligence process that is part of a merger or acquisition, when customer lists are being compared for overlap, the degree of congruence is very important, but the congruent identities must remain confidential, in case the expected event does not go through. The ability to determine congruence without identity exposure is a subtle, but enormously important, capability.

Then there are the cases where it is the identity of the individual that stands before you that is in question. There, comparison of identity information with an anonymized database can determine the identity to be valid, without unduly exposing information. This does not address privacy protection for miscreants on a watch list, whose identities are meant to be publicized, but it can help the rest of us from being confused with these people. In addition, for any individual, it allows identity to be based on associations and history rather than on racial profiling, bigotry or chance.

IBM DB2 Anonymous Resolution

The DB2 Anonymous Resolution Module can use any hashing algorithm to randomize the digital signature for a particular identity. This is an irreversible, one-way hash. The relationship elements and their full attribution are attached to the now-random identifier. You cannot get from that hash back to a name by any undo. Privacy is protected.

The data can be purposed specifically to a particular use, or it can be shared more generally, because the only thing that is compared is the attributes and the hash algorithm (which will be different if the information is from different sources). Thus, this is not a *search*, but a *compare*, with all the clean, quick efficiency and scalability that that implies. If there is evidence of a real match, the two list-owners might want to share information specific to that entity to enhance their information bases. The actual sharing of

[1] Obviously, this module is not so effective when dealing with loners.

[2] Its original name was NORA (short for Non-Obvious Relationship awareness).

that information, by creating a link between the two hashed identity values, is a separate process. **Throughout, the original owners of the data maintain control of their data. With anonymization, the identity management process does have to involve snatching the data from one source and exposing it to another.** Instead, it is an arbitrated comparison. It is the match between data sources, and the quality of that match, that is useful. Nevertheless, a match, *per se*, does not lead to witch-hunts and other misuses of the information. It is the combination of *identity exposure* with the *comparison information* revealed in the process that causes the trouble. If you can anonymize the identities, you add much-needed identity and privacy protection into the process.

Some Technical Details

This software can analyze data from any relational database. Like the rest of DB2, it runs on many platforms, and is priced per processor[3]. Because this is a software-only, turnkey solution that develops a repository of links to identity elements, it scales easily and new information can be quickly compared to all the identities in the repository. These modules support thousands of data sources[4]. Like many data services, after the first large ingestion and analysis of information, it can crawl sources and extract identity-specific information as a background process. Depending on the size of the identity information trove, Entity Analytics can be up and running in a couple of months. Because of full attribution[5], you never have to reload these elements. They are in there with a track to the source. While the unique identifier, or name, is what is hashed and anonymized by this software, other elements can also be hashed, either reversibly or irreversibly, as is appropriate.

The system scales to many millions of identities and to terabytes of data. It does not amass this data into a repository; it just builds an aggregation of links to information about people (and organizations). It resolves ambiguities but it neither searches nor mines. No, it is not a form of cleansing a la ETL (extract, transform, load) software. It is gleaning, documenting, and hashing. If you want to be trendy, it is a logical-level construct that could be considered a virtual repository of physically-distributed, variously-owned assets. Moreover, for many forms of public-facing data, over time, it may become a core data utility.

Some Uses of Anonymous Resolution

In health care environments, personal health information specifics can be shared without revealing identity, or other elements. Perhaps more important, patient attributes (but not specific identities) can be aggregated across health-care domains. Because additional elements not germane to the research can be hashed, even the privacy of limited populations, such as those suffering from a rare disease, can be protected.

In law enforcement, *Anonymous Resolution* has been very helpful in allowing the sharing of information about undercover activities between organizations not known for collaborative zeal. With the knowledge gleaned from the anonymized records, one enforcement branch does not arrest – or expose – another branch's operatives. Of course, in the classic case of verifying the identity of an individual, this historic approach would also showcase assumed identities whose identity elements all were generated in the recent past.

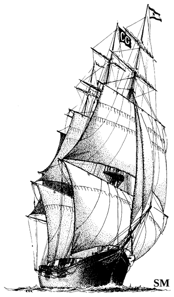
More prosaically, the use of *Anonymous Resolution* allows better sharing of customer data between strategic partners[6] for joint marketing campaigns. It allows an organization, by characterizing projects as entities, to discover similar projects without having access to the details of the project. In general, *Anonymous Resolution* lets you analyze sensitive data of all types to a fare-thee-well with less disclosure risk[7].

In general, the ability to hide the specifics of data while allowing similarities to be discerned makes trust easier to ascertain – not only at the start of a relationship but on an ongoing basis. The process of anonymizing information allows organizations to be more safely transparent, but it also may allow them to evolve more gracefully. And, it corrals the *can't do*'s that prevent organizations from helping one another.

Conclusion

These modules have obvious uses in companies qualifying new customers, employees, or partners. Entity Analytics can uncover applicants for jobs who previously worked at the company under a different name, or who have unarticulated relationships with competitors. On a larger scale, entity analytics promotes analysis of aggregate populations in a very granular way, while protecting privacy and confidentiality.

More broadly, IBM's Entity Analytics establishes a differentiation between ownership of information and use of information. It makes sharing information about identity or other things an incremental proposition, not all-or-nothing. Think about how entity analytics could extend organizational self-knowledge and market intelligence, while protecting privacy as an intrinsic part of the process.



[3] This pricing is based on the processors supporting the application, not the processor count of the server hosting the application.

[4] Identity elements can even be gleaned from point of sale environments.

[5] XML is used to articulate the elements of each identity element.

[6] These days, these partners are often competitors in another part of their business.

[7] Think how financial service enterprises could use selective hashing within their organization to minimize their data exposure risks.

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Anne MacFarland **is Director of Infrastructure Architectures and Solutions for The Clipper Group.** Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- ***Reach Anne MacFarland via e-mail at Anne.MacFarland@clipper.com or at 781-235-0085 Ext. 128. (Please dial "128" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group** ***Captain's Log,*** and "*clipper.com*" are trademarks of The Clipper Group, Inc., and the clipper ship drawings, "*Navigating Information Technology Horizons*", and "*teraproductivity*" are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.