"Data! Data! Data!" — an interview with Tim Williams of Index Engines: massive search power, unified process and audit trails, and more

Jan 19th, 2010 | By Gregory P. Bufithis, Esq.

This interview is part of our new series *"Data! Data! Data!" — Cures for a General Counsel's ESI Nightmares".* For our introduction to the series *click here*.

Tim Williams is Chief Executive Officer and co-founder of Index Engines. He founded CrosStor Software, and served as its Chairman and Chief Executive Officer. CrosStor was a pioneer in enterprise-class purpose-built NAS operating systems and SAN/NAS convergence, with over 24 OEM customer including EMC, IBM, and HP. CrosStor was sold to EMC in November 2000. After the acquisition, Tim served for a time as EMC's VP of storage operating systems for their mid-range division.

In 2001, Tim led a group of angel investors to restart Tacit Networks (sold to Packeteer in 2006). He refocused their technology, reinvented their business model, attracted new institutional investors, rebuilt the management team and served as Interim CEO until a permanent CEO was found.

Prior to founding CrosStor, Tim served as a consultant for a variety of computer and telecommunications companies. He participated in the development of the UNIX operating system at Bell Laboratories, and has held a number of key engineering positions at high-growth and startup computer companies. Tim has a Masters of Computer Science from New York University's Courant Institute. Index Engine's flagship product, Unified Discovery Platform, won the Gold LTN Award for Best New Product in 2009.

We caught up with Tim at the Masters Conference and in the offices of Index Engines.

TPL: First of all, congratulations on the Gold LTN Award.

TW: Thank you. We are very proud of that. Our Unified Discovery Platform was also recognized for its litigation support and records management capabilities. We are particularly proud, because the selection process for these awards is based on the publication's more than 40,000 subscribers voting for technology that represented outstanding achievement in legal technology.

TPL: Ok, we'll let you plug the product now.

TW: [laughing] Thanks. In short (you can get more from our **website**) by enabling a unified, quick, easy and streamlined access to enterprise data, regardless of where it is stored or in what format it is stored, our product makes both e-discovery and records management processes more efficient and cost-effective. As IT and legal teams work together more, solutions that solve problems for both departments, like the Unified Discovery Platform, will continue to be in great demand.

TPL: You have a direct link we can give folks for more information?

TW: Sure. For an overview go ***here***.

TPL: Actually, you folks take what we'll call a "fresh" approach.

TW: We agree. Our product was designed from the ground up for the needs of the modern, data-inundated enterprise. It's the only solution on the market to offer a complete view of all electronic data assets – not just online data, but offline backup data as well. That's unique to us – a single system with a unified process and audit trail. Keep in mind that online data is indexed, de-duped, and de-nisted in-stream in it's native storage format at wire speeds of up to one terabyte per hour, and similarly, offline data can be processed in it's native backup formats directly from the backup tape as fast as the tape can be read eliminating any need for a time-consuming restoration process.

TPL: So it's comprehensive?

TW: Absolutely. We provide the only comprehensive discovery platform across both online and offline data, saving time and money when managing enterprise information for discovery of ESI in the enterprise.

TPL: And all this only since 2003, when you started?

TW: Correct. Our mission was — and still is — to organize enterprise data assets, making them immediately accessible, searchable and easy to manage. Enterprise data is growing at 100% per year. Historically, most of that has been offline data. Keeping all of that data organized is a significant challenge. Businesses require timely and cost efficient access to all that data, and at the same time, must maintain compliance to regulations governing it.

Talk to any global enterprise and they will tell you that the total data they have data under management is far greater than the total data available on the Internet. That implies that if they

used Internet-class technology to manage it, they would need to build internet-class data centers in the process. That's not practical. If the goal is to get a handle on all that data, to organize it and make it accessible, a fundamentally different process, a "fresh approach" as you say, was clearly necessary.

TPL: Ok, bingo. You have hit on the purpose of this series of interviews. The "tsunami of data" as Ralph Losey says. A volume of data (and cost of discovery) which seems to be exponentially greater by the minute. In a nutshell, how do you help clients cope, get organized?

TW: At a conceptual level, we haven't done anything unique – we built a rich, searchable index of all the data – we are just doing it far faster (at 1 Terabyte per Hour per engine), far more scalable (up to 1 Billion files and emails per engine) and in place, directly in the native storage format, without copying or preprocessing.

Given that class of power, enterprise-wide search, classification and data management suddenly becomes practical. Our customers can finally see everything they have, and then make the decisions on what to keep and how to store it. They can survive that data tsunami, and scale cost effectively, process it quickly, and manage it efficiently, and most importantly, know what they have, what are the real assets and liabilities contained in their enterprise data. As we like to say, we sell "Power Over Information".

TPL: So, we now have a new lexicon, funky technology — and not necessarily technologically astute lawyers. Are most lawyers technophobic or perhaps they don't see technology like those of us in the industry?

TW: Actually, these days, most of the general counsel and law firms we deal with are more technical than they get credit for. I don't believe they are afraid of technology as much as they are afraid of the unknown. It's understandable. How can you manage risk on behalf of your company or client if you can't size it quickly and accurately? So in the past their jobs have largely been focused on avoiding data discovery, as it was too difficult and costly, and generally too hard to control. Index Engines changes that.

TPL: So it's really a lack of knowledge, a lack of familiarity? How do you help?

TW: Getting the word out that things have changed, that the traditional "burden argument" around discovery of ESI is no longer going to fly, and in fact isn't even defensible any more – that's a full time job for us. But as Index Engines gains more and more market awareness, through things like the LTN awards, the legal community will begin realize that the burden argument no longer holds water – it's now affordable and practical. We've transitioned a process that once required specialized skills, significant infrastructure, and many may man hours into an automated process taking only a fraction of the time and requiring only our appliance.

It still amazes us how often online sources of data are used for eDiscovery instead of offline sources. This makes no sense. Online sources rarely have forensic integrity. How easy is it to modify or delete an email stored on a company server? Litigation hold is often happening way too late – like closing the bard door after the horse is long gone.

Compare that to the point in time forensic copy of the same email stored on a backup tape. This turns the concept of litigation hold on its head. It's no longer explicitly necessary, because it's being done every night as a routine byproduct of the company's existing disaster recovery processes.

TPL: And your technology works no matter what — potential litigation, government investigation, and internal investigation, whatever?

TW: That's right. What's more, our technology also makes tape remediation feasible. You can typically pay for it for less than a year of the cost of storing that data for a year with a third party provider.

TPL: And what do you think is at the forefront of the discovery process, the most important thing, the biggest challenge?

TW: Its obsolescence. The word discovery presupposes an initial state of ignorance, and that is what nurtures all the fear of the unknown. The challenge must be to remove the need of going through a long internal discovery process in the first place. If you have a solid knowledge management system in place, then you don't have to discover, you already know. Know if you are in compliance or not, know which employees are increasing your corporate liability, know if your cases are worth settling or should be fought. Know all these things immediately as a matter of standard management practice.

If you don't know, if you don't believe knowing is possible or practical, you stay forever in a reactive mode, live your corporate life in a state of denial, forever putting off the inevitable, at an enormous cost in time, money, wasted resources and lost opportunities. The costs of ignorance are enormous.

Let me put it this way: ***knowing requires really getting complete power over your data – if you don't have power over it, it has power over you.*** I cannot emphasize that enough. To know, you need to integrate data knowledge management deeply into your organization proactively. And key to all that is a comprehensive indexing strategy, one tightly integrated into the company's backup and storage management policies.

TPL: There is a feeling among in-house counsel (gleaned from the ACC meetings we attended) that a direct relationship with e-discovery vendors is best, rather than through outside counsel. Do your law firm clients perceive this as a threat to their business?

TW: I guess it depends on the relationship. Many law firms we know operate as an extension to the enterprise in-house counsel, and their expert opinion is valued. We have a significant internal investment in programs to educate outside counsels about our product and find that they are very receptive to learning about ways to create more value for their clients, especially in tough economic times like these.

TPL: E-discovery vendors have also had much success the last 2 years moving into the e-discovery space across the whole EDRM model, especially in the area of document review (the

"right side") and that success is due to the continuing move by corporations to move EDD directly in-house. Document review is a nice piece of change. Is this a move you contemplate?

TW: We are focused on solving the problems on the left side of the EDRM model – identification collection, culling – which all have to do with speed, scale and complexity of data. Our value comes into play when large amounts of data need to be identified, searched, collected, and culled down either for litigation review or for knowledge management projects.

TPL: Ok, news flash. There is a myriad of software out there — review software, early case assessment software, ESI management software, etc. How do you distinguish Index Engines from the pack?

TW: We feed them. We agree, there are a lot of good review tools out there, and for the most part, our customers have already chosen their favorites. However, none of these tools can access data at the speeds we can and over the various data formats and container that we do. Our speed and scalability are unique and set us apart from the pack. We take that mountain of data and turn it back into a manageable molehill.

TPL: E-discovery costs are skyrocketing. Yet much of EDD is now a commodity – and that has changed the structure of the market. Prices are — shall we say — more predictable and probably more realistic. E-discovery vendors have capped fees, set flat fees or worked with various forms of pricing estimators. Have you changed your pricing?

TW: The challenge in 2010 for everybody — vendors, law firms, etc. — is going to be to stay flexible, stay focused on the needs of the customer, know what value they require, how much capacity they need and charge them only for that value and no more.

This hasn't required that we change our pricing, but rather that we broaden our pricing and licensing options, and in fact, we've invested considerably in that throughout this past year. Our product can be purchased outright by end users based upon the amount of indexing, search and ingestion capacity needed by the customer. We find that when measured against other products at similar capacities, we are always the price leader.

In addition, it can be purchased with term licenses, and with capacity metering licenses. There are a dozen different combinations of capacity metering in the current release.

Finally, we have a large network of partners that offer our product as part of a larger eDiscovery service offering. It's not unusual for companies to start working with one of those partners and latter decide on a more proactive approach and bring the technology in house.

TPL: The big "new new" thing all of last year — at every event we covered — was early case assessment and winnowing relevant data down to reduce the number of documents to review. As the stats bear out, it is the most expensive part of the process. But now we have predictive coding, plus the work being done in computer assisted review as evidenced by Patrick Oot and Anne Kershaw's study "Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review ", plus the work being done by Google and Microsoft on auto-

categorization and auto-coding. Is the technology getting to the point where we can also winnow out the eyeballs — contract attorney reviewers?

TW: What's important is that companies cull down data as early as possible. Ideally, they should cull before any collection is done. That's another unique capability of our product. Since we can index data very rapidly and in place without first copying it, it allows you to do your first cull prior to collection. That can be an enormous savings.

The technologies that you mention are also promising, particularly during the later review stages of the process, but ideally, the goal should be moving as little data into to those late stages as possible.

TPL: You recently announced that Index Engines had joined the EMC Velocity2 Technology and ISV Program. In addition, you were jointly conducting test results of indexing performance and capabilities on EMC® Data Domain® deduplication storage systems. Your press release said "that the Index Engines 3.0 platform is able to perform full content and metadata indexing of backup data stored on EMC Data Domain systems, enabling users to avoid time consuming restores prior to the indexing process. During recent tests conducted by Index Engines with support from EMC, a single indexing engine of the Index Engines 3.0 platform achieved sustained rates of over 1 Terabyte per hour on a Data Domain DD690 storage system". For the non-techies (like me) in our audience, work us through that.

TW: Well, that is part of an on-going program to qualify popular storage vendors at our 1 Terabyte per Hour per Engine indexing speed. You may have noticed that we recently announced the EMC Celerra NAS platform as well. Others are in the works.

What's interesting about those two announcements is what's different about them. Data Domain systems store data in backup format – same format as that used on backup tapes. Celerra systems store data in file system formats. What's unique about us is that one of our Engines can process both in excess of t the 1 Terabyte per Hour speed. No one else can do that.

In the first case in particular, the format of backup data, by design, is optimized for restores versus the indexing process that accompanies eDiscovery efforts. For all the reasons I stated above, backup data, as the most forensically sound copy of data, will inevitably become more important in the eDiscovery process. Other processes a full restoration of backup data to prior to indexing — an arduous, multi-step process that requires a large cache of disk, and usually a recreation of the backup and email environment. With the Index Engines platform, indexing systems like EMC's Data Domain deduplication storage systems, backup data can be indexed directly at high speed without having to restore multiple versions of backups to disk. And since the backup data is online, we can index as fast as the network allows – we are not bottlenecked on tape technology. Instead of restoration, a comparably small index (roughly 5% of the data) is created that users can use to query the data, cull it and then choose what to extract.

The combined solution is unique in its capability to enable customers to index de-duplicated backup data and enables a markedly more efficient approach to the eDiscovery process. What's more we can even allow you to search for individual files and emails on the Data Domain system

stored in backup format, and move those individual files and emails to the Celerra NAS file system. Think of it – an email stored inside an Exchange or Notes database, on a compressed, multiplexed backup can be found and individually moved without restoring the rest of the tape, or even the rest of the database. That's what we mean by "Unified". We've eliminated the brick wall that previously existed between backup formats typically used for offline data and file system formats, typically used of online data. That's really game changing.

Helpful?

TPL: Yes, very. And I actually understood it. But how do you do it? What is the secret sauce?

TW: As you can see by my background and the background of our co-founder, Gordon Harris, and other early employees, we are really storage guys, not eDiscovery guys. The key to the "fresh" approach we took was to see that the bottlenecks in all other indexing products were storage related, or more specifically, were I/O related.

TPL: Hold on. I/O related?

TW: Sorry. That's an abbreviation of Input / Output and refers to the transfer of data to or from an application. The key to fixing this was to take control of I/O and eliminate any inefficiencies.

So unlike traditional project-based indexing products, Index Engines started by creating purpose-built indexing operating system that was designed from the ground up to meet the demands of indexing for the enterprise. By approaching the problem at that level, we were able to achieve enormous improvements in speed and scalability. Most importantly, we eliminated the need to make a copy of the data prior to indexing, or the need to access it randomly, and we created a system that could index data serially in one pass, regardless of the complexity of that data's format. That's the secret to our ability to index backup formats as well as file system formats.

TPL: Tim, we greatly appreciate your time.

TW: Thanks. I enjoyed the conversation. We'll see you at LegalTech.

Postscript: *LegalTech is one of the premier events in the industry. It will be February 1, 2 and 3 in NYC (for details click here) and you can find Index Engines in the Exhibit Hall at Booth #2119.*

For all interviews in this series click here.

Gregory P. Bufithis is the founder and chairman of The Posse List and its sister sites The Electronic Discovery Reading Room (http://www.ediscoveryreadingroom.com) and The Posse Ranch (www.theposseranch.com). He is also founder and chairman of Project Counsel (www.projectcounsel.com).