STEPHEN E ARNOLD APRIL 22, 2008

Privacy: Balancing on a Knife Edge

Cloud-computing remains a promising solution to the challenges of on-premises installation of enterprise applications. The increasing interest in SaaS or software as a service as a way to control some information technology costs is evident in the Google-Salesforce.com tie up.

For those of you unfamiliar with the jargon and the players in the growing trend to shift from traditional on-premises software to cloud-based software, let's work through some terminology.

First, cloud-computing means that an application sits on a server in a data center. The twist is that the software vendor assumes responsibility for maintaining the application and ensuring quality of service. The company using a cloud-based service pays a monthly fee. The burden on the company's in-house information technology staff is greatly reduced. The idea is a variation of outsourcing in order to reduce certain costs. Cloud-computing is what you do when you use Internet services for word processing (Google Docs or Zoho, for example). Microsoft's cloud-based services are a reaction to push back about the complexity of installation, recovering from crashes, and updating applications. Companies are beginning to test the cloud computing model, but organizations are conservative, and some risks are hard to quantify.

Second, is SaaS or Software as a Service. This acronym is 2008 synonym for application service provider and managed services. It's a two-fisted acronym that relies on cloud computing and the pick up truck of acronyms for the technology that allows an employee behind a firewall to use a service located somewhere “out there” in a data center and have access to information whenever it may reside, including on a server in the company's own behind-the-firewall system. SaaS, in short, eliminates the notional boundaries that once separated in-house information from external information. You don't need me to remind you that in this type of set up, security is job number one.

The next term is “on premises”. In its simplest form, a company has a server, and it is down the hall from the boss's office. In addition to mental peace of mind, an on-premises server and software running on it give the company the illusion of control. Over-the-network upgrades and remote diagnostics undermine this naïve assumption about control of on-premises software installations. To muddy the conceptual waters, “on premises” may not mean in a facility owned and maintained by the software licensee. Enterprise application vendors allow a licensee to install the application in a data center that is under contract to the software licensee. The idea boils down to control. And control is at the core of cloud computing for the enterprise.

Finally, the company Salesforce.com is the poster child for the cloud computing and SaaS business sector. The company was the brain child of a former Oracle executive, and the Oracle relational database management system is the data management heart of the company. Salesforce.com was among the first cloud computing companies to market to companies unable or unwilling to pay for traditional enterprise customer relationship management systems from Oracle Corp., Microsoft, and RightNow, among others. In the last year, Salesforce.com has become a development platform so its customers can build more robust applications or customize the basic Salesforce.com services.

Now into this mix cometh the Google.

A year or two ago, a Google employee told me when I asked about Salesforce.com, “We really like those guys.” Google does like Salesforce.com, but it doesn't love them; otherwise, Google would have purchased the company. Maybe Google will buy Salesforce.com at some point in the future. What's

happened in the spring of 2008 is that Google has asked Salesforce.com to go steady.

The two companies have cooperated to make it easy for a Salesforce.com customer to participate in Google advertising. Salesforce.com explains the expanded "going steady" relationship this way: With Salesforce for Google Apps, you can now run your favorite desktop applications and your Salesforce applications side by side by accessing Gmail, Google Calendar, Google Talk, and Google Docs all seamlessly from within Salesforce." (http://blogs.salesforce.com/blogs/2008/04/announcing-sale.html)

The implications of this deal are significant for Microsoft. Google is exerting more pressure on the Microsoft hegemony in the enterprise. The deal also signals the IBMs, Oracles, and SAPs of the world that Google—despite its protestations of beta testing and exploring—is circling the Enterprise market in preparation for a larger-scale invasion. My hunch is that if Google finds that going steady with Salesforce.com is comfortable, Google may acquire Salesforce.com and use the company as a platform for a full-scale assault on the enterprise market.

Google's Achilles' heel may be privacy. Cloud computing asks an enterprise to cede more control to the vendor. This applies not just to Google but to Amazon, AT&T, or Pageflakes—any company offering cloud-based services. A licensee must believe the vendor who says, "We don't keep track of any personal or proscribed information."

I'm a trusting soul, but I have worked in and around online systems for more than 30 years. I have sitting not 10 feet from me a person who can write a script and suck content from any system to which she can get or has access. Privacy just like security is only as good as its weakest link. Cloud computing increases the risk that a breach could occur. Note, please, I am not saying will occur. We're dealing with risk here and one's definition of acceptable risk.

The muddled information in the April 21, Financial Times's story "Google Resolve Crumbles on 'Cookies' Pledge", April 21, 2008. The journalist, Richard Waters in San Francisco, points out that Google is not able to resolve some of the tracking and monitoring issues associated with the small text files placed on a user's computer. These "cookies" make it trivial for a cloud-based vendor to know who does what, when, how, and even where the user navigates when leaving one site to visit an unrelated url (uniform resource locater). My colleague—the one who can script dance with the best of them—said, "You can do lots of interesting tricks with cookies like sucking data off the user's computer. Snort. Snort. Giggle."

One newspaper story doesn't mean Google is going to take any chances with its enterprise customers. I think Google is much better about privacy than some other firms with which I am familiar. But Google faces some push back on privacy from the European Commission. Google's public policy Web log does a good job of keeping me current on Google's policies. The Web log is here: http://googlepublicpolicy.blogspot.com/.

In an April 7, 2008, post titled "The European Commission's Data Protection Findings", Google said:

> ...The European Commission's Article 29 Data Protection Working Party -- named after the rules they are monitoring –- has been conducting a lengthy inquiry into the question of online privacy. While the working party has welcomed our decision to anonymise [sic] data logs after 18 months as a positive privacy protective step, it suggested in findings released today that this period might still be too long.[sic]We believe that data retention requirements have to take into account the need to provide quality products and services for users, like accurate search results, as well as system security and integrity concerns. We have recently discussed some of the many ways that using this data helps improve users' experience, from making our products safe, to preventing fraud, to building language models to improve search results. This perspective -- the ways in which data is used to improve consumers' experience on the web -- is unfortunately sometimes lacking in discussions about online privacy.

STEPHEN E ARNOLD APRIL 22, 2008

The one piece of information that is not widely known that I find helpful when thinking about cloud-based computing and privacy is the mundane data schema for a user's behavior. In Google's patent and patent application corpus, I identified 15 open source documents that explain the systems and methods disclosed by the company for usage tracking. The first patent application is dated 2002, US 2002/0123988. Work on patent applications typically consumes several months, so it's reasonable to conclude that usage tracking was of interest prior to this patent application's filing in March 2001 to Google in its earlier days. The most helpful information to me on the issue of privacy is this diagram which appears in Google US 2006/0224583, titled “Analyzing a User's Web History”. Take a look at the

diagram that I found so interesting.

This is Figure 5: User Record in User Information Database in US 2006/0224583. It's clear that this record structure keeps track of the user by a “user identifier”. This record structure is also designed to monitor advertising. Of particular interest to me were the two column heads “Derived Data” and “Additional Data”. It occurred to me that with a little editing, this record structure would be useful in tracking the activities of a user of other types of Google services; for example, enterprise cloud-based services. I don't think that will happen, but it is interesting to look at this fine-grained data model in the context of the other Google inventions that track, manipulate, aggregate, and analyze stateful and stateless user actions.

My thought is that as Google moves into enterprise cloud-based services in a more significant way than it has to date, Google will want to reassure its enterprise customers that the concerns about “crumbling cookies,” the European Union's nagging, and the powerful usage tracking inventions disclosed Google's patent documents are lined up with soldiers, polished, and outfitted for battle in the enterprise wars ahead. IBM, Microsoft, Oracle, and SAP may find that Google's own policies, procedures, statements, and inventions are sometimes at sixes and sevens.

Stephen Arnold, April 22, 2008

Caption for Figure: “The data model shows fine-grained tracking of a wide range of user actions. Of particular interest are the “to be defined” functions that allow Google to expand the type of actions captured about each user. The document is available from http://www.uspto.gov.