Google's MapReduce, Chubby, and the Hadoop the Loop

Energy drink Red Bull Sponsors air races. In New York, colorful, propeller air craft raced around and through inflated gates. The gates looked like the blow-up animals in Macy's Thanksgiving Day Parade. When a steel-nerved air race miscalculated and nicked an inflated course marker, the blow up sagged like a deflating hot air balloon.

The traditional database market is in an air race with what seem to be faster, more agile, and more streamlined air craft. Structured query language or SQL databases owe their popularity to Edwin Cobb's insight that data could be held in digital form in a structure reminiscent of a piece of ledger paper. Some of you reading this column may be unfamiliar with the green ruled tablets that accountants once used to record debits and credits.

The row-and-column method still works. Even super-wizards at Google use Dr. Codd's traditional database to fiddle with small chunks of structured data. IBM, Microsoft, and Oracle have dominated the database market. Larry Ellison executed a strategic arabesque and purchased the Codd-clone MySQL, an open source competitor to the Oracle database management system. When Oracle's hands circled the throat of MySQL, one could hear gasps and squeals from MySQL users who feared for the future of MySQL. Oracle's approach to open source has been to own it and then share. So far, Oracle's approach seems to be working.

Not far from Oracle's black towers, Google's sprawling junior college-type campus cheerleads for a different approach. Google, like Oracle, is a for-profit enterprise. But Google's management team has been active in the open source world. In sharp contrast to Oracle's approach, Google releases code to the open source community. To make sure that the open "sourciness" of Google is recognized, Google has an open source ambassador, open source Web pages at http://code.google.com/opensource, an open source blog, open source programs like the GSoc or Google Summer of Code, and a stealth challenge to the Codd-loving incumbents.

Some of those following Google's business focus on advertising which accounts for 99 percent of Google's revenue. However, Google launched what amounts to a tactical probe of the commercial RDBMS market. Since 2004, Google has described its MapReduce technology in a handful of publicly-accessible technical papers, presentations, and YouTube.com videos. Navigate to www.youtube.com and run this query, "Google MapReduce." You can kick back at enjoy hours of lectures about Google's engineering marvel. MapReduce and its pal, Chubby and GFS (the Google File System) are among Google's core innovations. These systems make Google's massively-parallel, distributed, high-performance computing infrastructure a reality I use several times a day. My hunch is that you rely on Google and these components as well.

Is Google just generous? Was the firm making good on its promise to not be evil? Did Google's management team have a specific goal in mind? Was Google just being Googley?

Google's motivations are tough to figure out. From the point of view of a marketer, "open source" is deep voodoo. The interest in open source software seems to be increasing. In search, for instance, Lucene/Solr has captured the attention of a number of high-profile organizations, including Cisco Systems, eHarmony, Mitre, and Twitter. In content management, I stumbled upon a conference several months ago with throngs of people entranced by Drupal. I listened to a podcast from IT Conversations that focused on Drizzle (http://drizzle.org), an open source database server, and learned that a mini-revolution is taking place against some traditional database methods and limitations.

But the buzz word of the summer is Hadoop. Bloggers and poobahs have done loop-the-loops around Hadoop. According to Doug Cutting's "Hadoop: A Brief History" http://research.yahoo.com/files/cutting.pdf , today's open source wunderkind had its roots in Nutch, a Web-scale open source search system. Between 2004 and 2006, Google disclosed details of its Google File System and its MapReduce method. Yahoo hired Mr. Cutting, and the Hadoop project was "split out of Nutch." By 2008, Hadoop "hit Web scale", and the interest accelerated.

Apache Hadoop, therefore, inherited some of the Google MapReduce bloodline, but the muscle behind the software framework comes from the open source community with support from Yahoo's wizards, including Mr. Cutting, who named the framework after his child's stuffed animal.

The basic idea behind Hadoop is that the engineering methods avoid some of the well-known problems inherent in Edwin Codd's RDBMS invention. Hadoop supports common file systems, including Amazon's cloud. Hadoop relies on data nodes with a "name node" providing some of the Google innovations for knowing where data are across a distributed system and keeping performance free of RDBMS-style bottlenecks. A quick trip to Amazon will provide you with one-click access to books that explain Hadoop in considerable detail. I recommend Chuck Lam's *Hadoop in Action* and Tom White's *Hadoop: The Definitive Guide*, but there are a number on offer as well as a wealth of information at http://hadoop.apache.org/.

Hadoop is used by Amazon, Facebook, IBM, and Rackspace, among others. Commercial vendors have embraced Hadoop. IBM, for example, has several applications, including an analytics service, running on Hadoop. IBM and Google teamed in 2007 to offer university courses about Hadoop to computer science majors.

What interests me is the emergence of an open source software for "reliable, scalable, distributed computing." Translating this, Hadoop is "a direct alternative to existing commercial operating systems, RDBMS that masquerade as high-performance data management systems, and extremely-expensive proprietary solutions that have been designed to lock-in licensees."

In short, Hadoop along with Lucene/Solr and the NoSQL data management tools represent an outright attack on traditional enterprise-grade, commercial-carrier solutions. Companies like IBM are savvy and have embraced open source in order to reduce certain costs without

compromising IBM's ability to sell engineering support services and for-fee software that connects, extends, or amplifies the basic functions of the open source solution. IBM and a handful of other companies have figured out how to comply with the open source community approach and build a business. RedHat, Cloudera, and Lucid Imagination are three firms that have built solid businesses via open source.

My view is that Google is committed to open source software. The reason is that as open source gains momentum, Google weakens some of its staunchest competitors in the enterprise market. Instead of attacking one outfit head on, Google seems to take a five-year or long-march approach. As commercial solutions face increased competition, Google can selectively target specific business opportunities.

I no longer doubt Google's commitment to open source. The reason is that Google has made its Android operating system available as open source. Although there are risks associated with forks in Android, the benefit of the approach is that Google can gain "shelf space" without spending much on marketing. Who can beat free? Android is available on a range of devices. These include mobile phones which Motorola seems committed to burn into my brain every time I watch a television program to set top boxes that promise a personalized television guide.

Apple and Microsoft will have to find a way to blunt Android. Nokia already is struggling to maintain its revenues and further disruption in mobile and embedded device markets is a certainty.

The downside to Google's approach is that Google may find itself fighting a company using Google-inspired technology. Facebook is a good example. Mark Zuckerberg has hired Xooglers (former Google employees) and uses some open source software. Google, as I write this, has no current product that puts Facebook at a competitive disadvantage. Microsoft invested more than $200 million in Facebook in 2007. Google, on the other hand, has suggested that it will have its own Facebook-type product sometime in 2010. If Google's next Facebook killer follows the trajectory of Google Buzz, Google will be in a unique position. Its own open source initiatives have put steel in the backbone of a competitor like Facebook.

The irony is delicious. Google's open source efforts and former Google employees become the nemesis of Google. That's one of the risks of what I call the ultimate Red Bull-charged air race--"The open source card." Google may not be playing solitaire. In Google's high stakes race, open source-centric companies may win a trophy and a series of races. Although hard to accept, Google could be an also ran. Open source is a disruptive and Google's open source tactics may lack the horsepower to power to victory after victory.

Stephen E Arnold, August 6, 2010

Stephen E Arnold is a consultant. His Web site is www.arnoldit.com. His blog "Beyond Search" is at http://www.arnoldit.com/wordpress.