# Big Data: The New Information Challenge

Big data. I hear the phrase at conferences, in meetings, and on podcasts. Those using the term often have degrees in non technical field, so “big” means physical size like the Baltimore Ravens’ Ray-Ray or a never diminishing credit card debt of $20,000.

Both Ray-ray and the financial problem of $20,000 at 18 percent per year are “big”, but “big data” is different.

In today’s digital world, an organization can expect the amount of digital data to double every six or seven months. An insurance agency with a boss and three underwriters can start 2011 with plenty of hard drive space. By June 2011, the company will need external storage devices or a cloud-based storage service. Data just grows.

The problem is even more severe for large corporations. In the same six month period, the Fortune 500 company will have to store twice the data and figure out how to locate a specific document or comply with an eDiscovery request. Consider the problem: 100,000 employees, several gigabytes of email per employee and an email archive. When the court documents arrive, those data have to be secured against spoliation, searched for specific information germane to the legal matter, and then be made available so specific documents and attachments are easy to email to a feisty 30 year old Harvard law grad whose job it is to win the case for her client. Humans can’t do the email work because there is not enough time to read millions of documents.

But this is not “big data”. In fact, the Fortune 500 eDiscovery data are like a tugboat next to the Carnival Splendor. The problems of “big data” are as big a headache as the one the captain of the ill-fated cruise ship faced after feeding upscale passengers Spam and canned crab for three days.

Amazon, eBay, Facebook, Google, and telecommunications companies handle hundreds of millions of data events every hour. Google, to cite one example, receives in one minute 35 hours of digital video. Amazon processes retail orders and supports thousands  of users and their data via the Amazon cloud service. Facebook and its 650 million members generate petabytes of content in the time it takes to fly from New York to Chicago. Twitter, the messaging service that makes no sense to people over the age of 30, processes upwards of 8,000 tweets a second. The European Hadron collider dwarfs the Google with exabytes of data generated when mini-Big Bangs are created beneath placid Swiss cows.

These are examples of big data, and the problems require a different way of engineering, organizing, accessing, and analyzing them.

The buzz of interest around an open source technology or, more accurately, software system is amplified by the growing need to handle “big data.”

Some background. Google learned from the mistakes that burned pre 1996 Web search systems to smoldering ashes. The company leaned on learnings from engineers with deep knowledge of high-performance systems. The early Googlers ferreted out ideas for dealing with big data from research papers, Ivory Tower research labs at leading universities, and smart men and women sitting in a room with their mobile phones and laptops.

The Hadoop Wiki at http://wiki.apache.org/hadoop/ explains the technology this way:

> *Apache Hadoop is a framework for running applications on large clusters built of commodity hardware. The Hadoop framework transparently provides applications both reliability and data motion. Hadoop implements a computational paradigm named Map/Reduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system (HDFS) that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both Map/Reduce and the distributed file system are designed so that node failures are automatically handled by the framework.*

Simplifying and translating, Hadoop makes it possible to use a large number of commodity servers to handle some of the back office work required to deal with petabytes or exabytes of data. Commodity translates to lower cost. Map/Reduce refers to a Google method released for anyone with some serious computer savvy to use for free. And distributed means the computers don’t have to be in one data center or even on the same continent. The subtext is that Hadoop and some supporting open source methods make IBM’s DB2, the Oracle database technology, Microsoft SQL Server, and other traditional relational databases look pretty clunky.

Hadoop has been incorporated into a number of commercial systems that provide tools to manipulate and extract hitherto unknown nuggets of information from huge data collections. Examples range from figuring out how to connect mobile phone calls and provide location-based searches to flashy products like the Google and Microsoft geospatial services. In real time, both companies can provides hundreds of thousands of simultaneous users with augmented maps that show the pizza joint closest to the user at 5 pm as a soccer mom rushes to pick up her kid.

The challenge Hadoop presents to established software vendors is significant. One example is a company that recently received $30 million in venture funding. Aster Data is one of a small number of specialist firms that are using open source Hadoop and certain proprietary systems and methods to deal with the analysis of big data. Aster Data’s CEO Quentin Gallivan (www.asterdata.com) told me:

> *Aster Data is breaking new ground in big data management and processing as recognized by the World Economic Forum and others. We've brought to market a new platform for big data management and advanced processing of data that provides a fundamental shift in how data will be stored and processed the next decade. We tap into the existing $20 billion market opportunity and open a new $7 billion market opportunity – take these numbers together and solutions like ours draw a lot of attention.*

Setting aside the rocket science, I asked Mr. Gallivan, "What makes Aster Data different from a traditional data management company?" He said:

> *Aster Data's solution is unique in that it allows complete processing of analytic applications 'inside' the Aster Data MPP database. This means you can now store all your data inside of Aster Data's MPP database that runs on commodity hardware and deliver richer analytic applications that are core to improving business insights and providing more intelligence on your business. To enable richer analytic applications we offer both SQL and MapReduce.*
>
> *I think you know that MapReduce was first created by Google and provides a rich parallel processing framework. We run MapReduce in-database but expose it to analysts via a SQL-MapReduce interface. The combination of our MPP DBMS and in-database MapReduce makes it possible to analyze and process massive volumes of data very fast.*

Aster Data is not alone in its use of open source technology. Palantir, a specialist content processing company, received $90 million in venture funding in the summer of 2010. Fast growing Digital Reasoning, based in Franklin, Tennessee, is undergoing explosive growth because the company can chop big data down to size, delivering near real time analytic outputs without the cost burdens of the 50 year old traditional relational database technology.

How will information companies benefit from open source "big data" ? In my opinion, information companies will be able to use Hadoop and related tools to create new information products. The Hadoop revolution is one that will yield new content sources. Second, information companies are often bastions of the on premises, old-style approaches of the IBMs, Oracles, and Microsofts. The Hadoop and open source options now allow publishers to think about alternatives to traditional information management tools. Third, the new technology will increase the stakes for content-centric products and services. New tools mean new competitors. As difficult as it may be to believe, the competitive environment will undergo additional rapid change.

Big data is now the new normal. As Yogi Berra allegedly said, "If the world was perfect, it wouldn't be."

Stephen E Arnold, December 1, 2010

Mr. Arnold is a consultant. More information about his practice is available at www.arnoldit.com and in his Web log at www.arnoldit.com/wordpress.