Brought to you by

What Does Big Data Mean for the Smaller Business?

By:
Dave Turbide
CFPIM, CIRM, CSCP, CMfgE
Production Solutions
www.daveturbide.com

Introduction

One of the most hyped buzzwords around, “Big Data” is causing excitement, confusion, and more than a little apprehension among companies large and small. Already buried in more data than most people and organizations can effectively use, the prospect of an accelerating onrush of even more data can indeed be daunting.

Nevertheless, more data, and more kinds of data, are coming your way and bringing with them new challenges, new tools, and new threats and opportunities for the small and mid-sized company with limited resources and limited ability to deal with a fast-changing technological landscape.

So let’s look at Big Data, see how it is changing the competitive landscape, and think about its impact on smaller companies in the coming years.

What is Big Data?

The growth of data is not something that developed just recently – the availability and proliferation of data is a long-established phenomenon that has driven (or perhaps been driven by) the rapid acceleration of computing and technology in general. Every new electronic device, application, sensor technology, engineering achievement, and communications advancement creates new sources of data and new ways to collect, disburse and exploit it. And we all know that technologies in general, and computing in particular, are growing at an accelerating pace. Moore's Law states that electronic circuit technology capabilities double every 12 to 18 months, setting the pace for an exponential increase of information and data. This contention is borne out by some startling facts[1]:

- 90 percent of the data in the world today was created within the past two years.
- At the end of 2011, it was estimated that there were more networked devices than people on earth.
- At that same time, 20 typical households generated more Internet traffic than the entire Internet in 2008.
- In 2011, 1.8 zettabytes of data were created – expected to increase to more than 7.8 zettabytes by 2015. A zettabyte is a 1 followed by 21 zeroes.
- For comparison, the U.S. Library of Congress Web Archiving Team reports that "As of April 2011, the Library has collected about 235 terabytes of data" (a zettabyte is a billion terabytes) and that it adds about 5 terabytes per month.
- Over the next 10 years, the number of servers will grow by a factor of 10, information by a factor of 50, the number of files by 75X, while the number of IT professionals in the world will grow by less than 1.5X.

The simplest definition of Big Data is data sets (blocks of information) too large for traditional database management tools. When the amount of data is so large, it becomes difficult to capture and store, search and analyze, and parse useful information from the overwhelming flood of detail.

There are three dimensions of Big Data that describe the challenge: volume, velocity and variety.

Volume – The factoids listed above give some indication of the size of the data universe today and its phenomenal growth. Handling and storing this mass of data is getting more problematic every day. Many companies are opting for offsite (Cloud) storage to reduce the burden on in-house

[1]*Factoids from IBM, Cisco, siliconangle.com, columnfivemedia.com, EMC, edCetra training.*

resource, but that's just a resourcing decision and doesn't address the fundamental issue: What are we going to do with all this data to change it from a cost to a valuable asset?

Velocity – Data is arriving so fast now that it defies normal handling processes. It is important to analyze the data and respond or react to get full benefit, but current systems may not have the capacity to do so at the speed required by the flood of input. Additionally, the volume taxes storage capability and much of the data is ultimately not of much or any use so should be filtered out before being stored. But, again, the velocity of input strains our ability to sort the wheat from the chaff before relegating the data stream to storage so we might end up storing a lot more than we can profitably exploit.

Variety – Most of the data that companies have maintained and used up to this point has been structured data, and there is still a lot of that in the mix, but new forms of unstructured data are an expanding portion of evolving data streams and they pose an additional challenge to IT resources and infrastructure. Structured data is defined and organized – data field and record formats that can be stored efficiently in relational databases and acted upon by application programs. Unstructured data in the form of text, audio, video, click streams, log files and more is not so easily handled and used. New database paradigms and new processing approaches are needed to interpret the information contained in these unstructured streams.

Working with Big Data

Big Data is, by definition, too much for traditional data management tools to handle effectively. There's just too much of it, coming at you too fast, and it's disorderly – much of it not easily tucked away like the structured data we are all used to.

The three dimensions identified above – volume, velocity and variety – all factor into the challenges of Big Data. Since the volume can quickly overwhelm typical available storage capacity, there might be a temptation to summarize or sift out the important data from the flood. This can be problematic because the data is coming so quickly that application programs can't keep up with it, and even if they were capable of doing this kind of real-time analysis, it would be difficult to know what's really important and what isn't, because what is considered trash today could be the foundation of tomorrow's big analytical breakthrough. Variety complicates it all because unstructured data doesn't fit well in traditional database management systems and is difficult for application programs to sort through and make sense from; new tools are needed.

The volume issue is being addressed by a technology known as massively parallel processing databases. Like massively parallel computing, this approach distributes the workload among an array of commodity resources (processors and disk drives) networked together to multiply the processing speed (computing) and storage capacity (database) of individual processors/drives. Combined with Cloud-based storage services, massively parallel technology offers even smaller companies the storage they need to take advantage of Big Data.

New data management tools have evolved to supervise the distribution of processing and storage across multiple resources, the best known of which is the Yahoo-developed open source software library Apache Hadoop and its companion applications. The Hadoop family includes a distributed file system (HDFS), MapReduce, Hive, Pig and others. MapReduce, an approach pioneered by Google, maps the data and compiles indexes (map) and recombines partial results (Reduce) that are used by Hive to provide some of the SQL-like access functionality that users need to make use of the data. Pig is defined by the Apache Software Foundation as "a high-level data-flow language and execution framework for parallel computation." Be aware that this is a batch operation environment for analysis and does not support interactive tasks.

And that brings us to the velocity issue. As Big Data is streaming in, we have to do something with it besides just channeling it into storage. A large part of Big Data's value comes from immediate use and interaction with the source. A great illustration of this is the way online stores can watch a site visitor's click-stream and provide guidance, suggestions or assistance while the user is still on the site. New tools for online analysis and interaction include IBM InfoSphere Streams, Twitter Storm, and Yahoo S4.

The storage issue is also a concern on the velocity dimension. Big Data is just so, well, big that we may not want to or even be able to store it all – and a lot of the data may actually be useless "noise" that should be filtered out before sending the good data to storage. The challenge is in knowing what is truly valueless – today's random noise has an annoying habit of becoming useful at a later time when new tools and new ideas emerge that can make it valuable. Nevertheless, real-time analysis and filtering/consolidation are often necessary to reduce the volume of data stored for later data mining.

Relational databases are rigid in that the data structure is pre-defined and embedded in the tables. That makes it easy to access and manage, but presents some problems with unstructured data like video, audio, raw sensor feed, and text, full of inconsistencies and errors. According to IBM, and they should know, 80 percent of the world's data is unstructured, and most businesses don't even attempt to use this data to their advantage. A new class of database known as NoSQL has evolved to

accommodate unstructured Big Data. These tools use what's called "key value stores" to manage the data without a pre-defined structure (schema).

Threat, Burden or Opportunity?

In addition to the natural growth of data from traditional sources – the proliferation of sensors and connected devices – Internet-based enterprises like Google, Yahoo, Amazon and Facebook have been the stimulus for an explosion of these new data types. They have also pioneered the effective uses of Big Data and the development of tools and approaches that make it useful and valuable. The second wave in the exploitation of Big Data is occurring in the retail and consumer product space where point-of-sale, click-stream and buying pattern data are being woven into sales strategies and marketing activities. Manufacturers are just beginning to understand the potential of Big Data and starting to explore ways to use the new tools to improve performance and competitiveness.

While the use of Big Data is not yet a significant threat in most industries, it is making itself felt in retail and e-business, where consumer data is helping companies target marketing and sales activities to an extent that couldn't even be imagined a few years ago. While there is some concern by privacy advocates, this train has already left the station and there's no stopping the spread of Big Data as the ultimate key to understanding the customer and tailoring activities to ever more finely defined niche groupings.

Lessons learned and tools developed in the retail space are proliferating upstream to business-to-business e-commerce and will soon pervade other areas of buying and selling. After that, we will find new ways to use these tools to better understand and manage the dynamics of the entire supply chain, including plant operations, transportation and warehousing.

As with most new technologies, the biggest companies with the deepest pockets are the first to adopt Big Data techniques for competitive advantage. Smaller companies will generally have to wait until the tools and techniques are packaged and proliferated in affordable solutions that don't require an army of technologists and statisticians to drive the effort.

Smaller companies should be vigilant, however, as technology cycles are shrinking and new technologies emerge, evolve, and reach critical mass much faster than most of us are ready to accept them and put them to use. Keep a close eye on competitors, sales patterns, markets, and any unusual or unexpected changes. They could be the visible manifestation of Big Data.

One final piece of advice comes from Christer Johnson, IBM's leader for advanced analytics in North America: First decide what problem you want to solve. Data by itself is not especially useful. Only when the data is put into context and illustrates a business situation or issue does it become of value. When considering the impact of Big Data in your business, first ask what value you hope to glean from the data. Only then can you put together a viable plan with costs and benefits to steer your team and keep it focused on value to the business.

About the author:

David A. Turbide, CFPIM, CIRM, CSCP, CMfgE

Author of six books, hundreds of magazine articles, and numerous white papers and reports, Dave is an independent consultant, analyst and freelance writer serving both the developers and the users of software and systems for manufacturers. For thirty years, Dave has helped manufacturers to select, implement, and get better results from their systems. He writes the Enterprise Insights column for APICS magazine and is a regular contributor to various industry blogs and discussion threads. For more information contact dave@daveturbide.com