February 2013 Volume 41 Issue 1

by Mike Nelson

The big picture on big data

Big data was the technology buzzword of 2012 and, despite a lack of agreement on its meaning, it is driving investments of billions of dollars in data storage systems, analytics software, and telecommunications services. The term “big data” is quite misleading. It actually covers:

- Big data - Millions and billions of measurements (e.g., cash register data)
- Fat data - Thousands of measurements per person (e.g., cellphone geolocation data)
- Fast data - Rapidly, constantly changing data (e.g., highway traffic flow)
- Distributed data - Information collected and stored in thousands of places that must be combined
- Messy data - “Noisy” or improperly entered numerical data and unstructured data (e.g., text, audio, video)

But while “big, fat, fast, distributed, messy data” would be a far more accurate term, it does not fit in a headline. What is different about big data? First, it is too large, changing too fast, or too dispersed to be managed using traditional database software and computer systems. Second, big data and the tools used to analyze it are giving companies and governments new insights to new markets, new threats, and new ways to better serve their clients. And the opportunities are increasing; the flood of digital data more than doubles every two years, according to most estimates.

Big data does not just mean numbers in spreadsheets, it is also text, video and images. It includes structured data in table form and unstructured data, such as e-mail messages, tweets and Facebook posts. According to a 2010 Gartner report, 80 percent of the business data produced each year is unstructured.

Why the sudden surge in interest in big data? The dramatic reduction in the cost of computing and storage made possible by cloud computing services and the spread of easy-to-use, open-source analytic tools have enabled thousands of startups and established companies to create new big data tools for research, business, government, and the non-profit sector. Examples include:

- Marketing and online advertising (e.g., targeted coupons, Twitter trends and Google AdWords)
- Health care (e.g., epidemic tracking using Google search terms, treatment outcome analysis and fraud reduction)
- Utilities (e.g., smart grid to manage electricity production and consumption and water quality data)
- Logistics and transportation (e.g., Walmart’s supply chain management and highway traffic alerts based on smartphone data)
- Government services (e.g., more detailed weather forecasts and faster disaster response after Hurricane Sandy)
- Security, intelligence and law enforcement (e.g., spotting crime trends or tracking terrorists).

Big data is forecast to drive $34 billion in worldwide information technology spending in 2013, a 21 percent increase from $28 billion in 2012, according to an October 2012 report from Gartner (2012). Most of this new spending will go to buy the IT systems and software needed to store, manage and process the flood of data, which worldwide will increase by a factor of 50 between the start of 2010 and the end of 2020, according to a study by IDC (2012). The big winners will be established IT vendors such as IBM, EMC, Google, and NetApp, as well as big data start-ups such as Splunk, Kaggle, and Cloudera.

Startup companies that develop big data technologies and services will see rapid revenue growth and more market opportunities - if they can keep pace with the flood of data and make it more useful. Information technology companies that can integrate big data solutions into their offerings will give clients a new reason to invest more in computer hardware and software. Firms in sectors ranging from finance and telecommunications to petroleum and agriculture could reduce the cost of operations, improve marketing, reduce risk and launch new data-driven services. For many businesses, data could soon be as important to the bottom line as labour and capital.

A comprehensive approach is needed

Alan Blatecky, director of the Office of Cyberinfrastructure at the National Science Foundation (NSF), describes what he calls the big data cycle (see Table 1), a comprehensive approach that addresses each step of the cycle needed to harness the full power of big data.

Most of the media hype, much of the venture capital invested in big data start-ups and much of the government funding for research and development has focused on step four, the analysis of big data. A recent Bloomberg analysis found that non-defence agencies and the Pentagon spent $3.15 billion on analytics in fiscal year 2011. However, without a comprehensive approach that includes funding for all nine steps, big data will fail to reach its potential. Analysis is only useful if the data being analyzed is properly collected, tagged, managed and explained

The role of governments in big data

Governments could and should play a critical role in spurring the growth of big data tools and services, just as they played a key role in spurring the growth and spread of the Internet and the World Wide Web.

The US government has funded core technologies for analyzing, visualizing and assessing Big Data. Federal research grants have generated new technologies that the private sector could exploit. In March 2012, the White House announced a Big Data Research and Development Initiative (White House, 2012) to encourage and coordinate development of tools and technologies needed to collect, analyze, and curate the massive quantities of data generated by government agencies and researchers. Government research programmes not only fund development and testing of new technologies and they also fund the education and training of data scientists needed to build and use big data applications. According to a report by the McKinsey Global Institute (2011), the US will face a shortage of 140,000 to 190,000 big data analysts by 2018, which could hinder the ability of businesses and governments to leverage the power of big data.

Table 1 The big data cycle and key challenges

Step one - Collecting the data	
Get data from sensors, clickstream data and Web surfing behavior, sales data, satellite and aerial imagery, government databases, and computer models	Challenges: Volume of data; combining and reconciling disparate data types
Step two - Verifying the data	
Determine the source (provenance) and the collection technique(s) used; assess the accuracy	Challenges: Develop standards for provenance and quality
Step three - Tagging the data	
Use standards to indicate the category, quality and provenance of data	Challenges: Develop standards and tools for metadata ("data about the data") and automated curation tools
Step four - Analyzing the data	
Use statistical methods and visualization to understand the data	Challenges: Create easy-to-use, affordable software that scales to terabytes of data (e.g., Hadoop and MapReduce)
Step five - Applying the data	
Ensure managers and executives can understand the power and limits of big data	Challenges: Develop curriculum for MBA and executive education programmes to avoid being "blinded by the data"
Step six - Sharing the data	
Use online collaboration tools, social media and other tools to get results to people who can use them	Challenges: Manage access control, digital identity
Step seven - Protecting the data	
Ensure the data is not stolen or corrupted	Challenges: Develop affordable, interoperable tools to protect large and distributed online databases
Step eight - Archiving the data	
Determine what data to keep and what to toss - and when	Challenges: Create standard procedures to assess value of data, legal requirements, redundancy of systems
Step nine - Repeat, reuse, remix (return to step one)	

Governments can also spur the development of big data services by accelerating efforts to make more government data readily accessible in useful formats. Since early 2009, the Obama administration has made more than 440,000 databases have been put on the data.gov website along with more than 1200 applications for using that data. The UK government has also been a leader in make its data more useful and, in December of 2012, the European Union announced a new EU-wide Data Portal (EC, 2012).

While new applications of big data are being announced every week, there are several barriers to wider use of big data tools and services. The biggest obstacles to widespread application of big data are issues related to privacy, security and transparency and public perception that big data equates to big government, big business, and Big Brother. It is clear that the growth of big data will require new approach to privacy protection. Unfortunately, in the European Union, 20- to 30-year-old data privacy laws have held back applications of big data in marketing and fraud reduction. Regulations that require personal data be collected for a specific purpose and not be reused for other purposes will block some of the most exciting, new, big data applications that involve remixing different data sets to spot surprising trends and correlations. Concerns extend beyond protecting and restricting the use of personally identifiable information. Some privacy advocates worry that companies could essentially decode information intended to be private or anonymous by cross-matching data sets. This would allow a company to develop a detailed digital dossier that could be used to infer what someone buys, where they travel, and with whom they associate. The solution will probably require new government imposed disclosure requirements so consumers can know what information businesses and governments are collecting about them and how it is being used.

Governments and the courts will need to address liability issues if data are incomplete or inaccurate and leads to conclusions that may be costly or even dangerous.

Laws need to be clarified on how databases are licensed, what intellectual property rights database creators have, and how data can be reused. In 2005, the European Union evaluated its data protection Directive 95/46/EC and found that it had discouraged investment in database services while the US database industry has grown rapidly.

Reconciling legal and regulatory treatment of databases will be a major challenge for companies trying to create integrated, global big data solutions. Fortunately, European officials are discussing these issues with their US counterparts. Companies such as IBM, SAP and Google have been at the forefront of efforts to encourage governments to adopt policies to foster growth of big data applications, often working through trade associations such as TechAmerica. The policy issues are difficult and new challenges are arising as new technologies and new applications develop.

Michael R Nelson is a policy analyst with Bloomberg Government (http://bgov.com), a subscription service providing analysis of how government actions affect business. This article is adapted from several of his recent BGOV reports on big data. The views expressed are his own.

You can follow him on @MikeNelson

Bibliography

EC. (2012). *The open data hub of the European Union.* Retrieved January 24, 2013, from European Commission: http://open-data.europa.eu/open-data/

Gartner. (2012, October 17). *Big data drives rapid changes in infrastructure and $232 billion in IT spending through 2016.* Retrieved January 24, 2013, from Gartner Inc.: http://www.gartner.com/it/page.jsp?id=2200815

IDC. (2012, December). *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East.* Retrieved January 24, 2012, from IDC: http://www.emc.com/leadership/digital-universe/iview/executive-summary-a-universe-of.htm

McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition, and productivity.* Retrieved January 24, 2013, from McKinsey Global Institute: http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

White House. (2012, March 29). *Big data is a big deal.* Retrieved January 24, 2013, from Office of Science and Technology Policy: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

Events Diary

• February 2013

UNESCO: Towards knowledge societies: First WSIS+10 review meeting
25-27 FEBRUARY
Paris, France
http://tinyurl.com/a5kcpkf

• March 2013

IIC: Telecommunications and Media Forum
19-20 MARCH
Brussels, Belgium
www.iicom.org

• May 2013

WSIS Forum 2013
13-17 MAY
Geneva, Switzerland
www.wsis.org/forum

ITU World Telecommunication/ICT Policy Forum
1-15 MAY
Geneva, Switzerland
www.itu.int/WTPF

• June 2013

IIC: Telecommunications and Media Forum
12-13 JUNE
Istanbul, Turkey
www.iicom.org

CommunicAsia 2013
18-21 JUNE
Singapore
http://www.communicasia.com/

2013 IAMCR Conference
25-29 JUNE
Dublin, Ireland
http://iamcr2013dublin.com/

• July 2013

ITU 13th Global Symposium for Regulators (GSR)
3-5 JULY
Warsaw, Poland
www.itu.int/gsr13

• September 2013

TPRC41
27-29 SEPTEMBER
Arlington VA, USA
www.tprcweb.com

• October 2013

IIC International Regulators Forum
7-8 OCTOBER
London, UK
www.iicom.org

IIC Annual Conference
9-10 OCTOBER
London, UK
www.iicom.org

• November 2013

ITU Telecom World 2013
18-21 NOVEMBER
Bangkok, Thailand
www.itu.int/TELECOM

• March 2014

World Telecommunication Development Conference 2014
31 MARCH-11 APRIL
Sharm-el-Sheikh, Egypt
www.itu.int/en/ITU-D/Conferences/WTDC