Association for Information Systems
AIS Electronic Library (AISeL)

MWAIS 2013 Proceedings | Midwest (MWAIS)

5-24-2013

Using Big Data for Analytics and Decision Support

Daniel Power
University of Northern Iowa, daniel.power@uni.edu

Using Big Data for Analytics and Decision Support

Daniel J. Power
University of Northern Iowa
Daniel.Power@uni.edu

ABSTRACT

People and the computers they use are generating large amounts of varied data. The phenomenon of capturing and trying to use all of the semi-structured and unstructured data has been called by vendors and bloggers "Big Data". Organizations can capture and store data of many types from almost any source, but capturing and storing data only adds value when it has a useful purpose. Big Data must be used to provide input to analytics and decision support capabilities if it is to create real value for organizations. Some bloggers, industry leaders and academics have become disillusioned by the term Big Data. It is a marketing term and not a technical term. More descriptive terms like unstructured data, process data and machine data are more useful for IT professionals. Researchers need to study and document use cases that explain how specific, novel data, so-called Big Data, can be used to support decision-making.

Keywords

Analytics, Big Data, decision support, machine data.

INTRODUCTION

Big Data is a colorful phrase for a significant change in data capture, storage and retrieval. Each day, every one of us generates very large amounts of digital data. We send and receive email, visit Web sites and make online purchases, use tools like Google Docs, make phone calls, upload photos to Facebook, use Google Search, chat with friends, take our cars for service, work out at a gym on a machine with an Internet connection, pay bills online, and our utilities, Internet and cable usage is digitally monitored and captured. This data and much more from our activity is recorded and often backed-up in the computing cloud. Now we can capture, store and perhaps analyze the data incidental to personal and organization activities and actions. Also, extensive machine generated data can be stored and analyzed. Organizations now have very large data sets stored in many files and databases.

A Google search for the phrase -- Big Data -- in quotations in April 2013 returns about 17,700,000 results with ads from SAS, Intel and EMC about Big Data. On the first page of results, the McKinsey & Company link is to a 2011 report titled "Big data: The next frontier for innovation, competition, and productivity". When the search is narrowed to the phrase -- define "Big Data" -- there are about 2,680,000 results with the same three ads. The phrase -- What is big data -- returns about 24,100,000 results. Big Data has attracted extensive interest and created high expectations for positive outcomes.

This paper examines the usefulness of the term "Big Data" and more broadly examines the potential for using the expanding data that can now be captured and analyzed. The following section reviews the definition of "Big Data" used in the popular media, and the increasing skepticism about the term in the practitioner community. The next section briefly discusses analysis of these Big Data sources, possible use cases for new and novel data, and the need to operationalize Big Data for research. The concluding section explores the role of Information Systems researchers in studying this phenomenon.

BIG DATA DEFINED

Provost and Fawcett (2013) define Big Data as “datasets that are too large for traditional data-processing systems and that therefore require new technologies” with names like Hadoop, Hbase, MapReduce, MongoDB or CouchDB. Ehrenberg (2012) notes that when he first used the term "big data" in lower case in 2009 to label a new ventures fund, the term "implied tools for managing large amounts of data and applications for extracting value from that data". Cloudera CEO Mike Olson describes Big Data as complex data at volume, but he admits to not really liking the term Big Data (see Scoble interview, 2010).

Machine data is a major contributor to the Big Data revolution. Machine data is all of the data generated by a computing machine while it operates. Examples of machine data include application logs, clickstream data, sensor data and Web access logs (cf., Power, 2013b). The O'Reilly Radar definition of Big Data is a situation where the size of the data itself becomes part of the problem. According to Cooper and Mell (June 2012), "Big data is where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of significant horizontal scaling for efficient processing."

Digital data is massive. For example, an *Economist* magazine special report (2010) notes that Wal-Mart "handles more than 1 million customer transactions every hour, feeding databases estimated at more than 2.5 petabytes — the equivalent of 167 times the books in America's Library of Congress ..." Data comes from both new and old sources and the increased volume of data has led some vendors and industry observers to proclaim a new era of Big Data. IBM researchers (Zikopoulos et al, 2013) describe Big Data in terms of four dimensions: 1) Volume, 2) Velocity, 3) Variety, and 4) Veracity. Gartner defines Big Data as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." Carter (2011) a researcher at International Data Corporation (IDC) asserts "Big Data refers to data sets whose volume, variety, velocity and complexity make it impossible for current databases and architectures to store and manage." IDC defines Big Data technologies as "a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis."

Based on Gartner, IDC, IBM and SAS web pages and papers, the following are the 5 dimensions for data that are creating new challenges for data management and analysis. The Big Data metaphor is supposedly at the extreme end of one or more of the following dimensions. 1. Data volume - measures the units of data storage on various media. 2. Data variety - refers to the many formats of digital data including photos, email and text documents. 3. Data velocity - according to Gartner "means both how fast data is being produced and how fast the data must be processed to meet demand." 4. Data variability - according to SAS means "data flows can be highly inconsistent with periodic peaks". 5. Data complexity - according to SAS means data is from multiple sources and it is difficult and challenging to link, match, cleanse and transform data across systems.

Andrew Brust noted in a March 2012 post to inaugurate his Big Data blog that "The excitement around Big Data is huge; the mere fact that the term is capitalized implies a lot of respect." Rust acknowledges the term is not well-defined. He asserts "Big Data is about the technologies and practice of handling data sets so large that conventional database management systems cannot handle them efficiently, and sometimes cannot handle them at all." Pundits and vendors have combined many topics and technologies both new and old under the Big Data label including Business Intelligence and analytics.

Recently some bloggers have become disillusioned by the term Big Data, but realize the enormous potential of analyzing non-traditional data sources. For example, Barry Devlin (2013) argues, "Big data as a technological category is becoming an increasingly meaningless name." De Goes (2013) further asserts, "The phrase 'big data' is now beyond completely meaningless." Sorofman (2013) considers Big Data, "a cute way of describing the idea of data processed at massive scale and speed, where the trail thrown off by all of our varied digital interactions and experiences becomes the fuel for decisions, insights and actions."

ANALYZING BIG DATA

Some information technology vendors regularly over-promote technology opportunities, and that has happened with Big Data and analytics. Some managers quickly get disillusioned, and that is happening with the ambiguous concept of Big Data. Venture Capitalist Bryce Roberts (2012) reminds us "Data, big, medium or small, has no value in and of itself. The value of data is unlocked through context and presentation."

Managers need to understand what to do with new data sources and few managers want to blindly hire high salary data scientists to work magic and find new strategic insights. Managers want to understand what a data scientist will do and why someone is needed in that role. Managers also seem reluctant to purchase more expensive hardware and software to store data that may not be useful. Big Data is not necessarily needed or better data. Aziza (2013) in his critique notes, "we need a different and more mainstream way to think about Big Data".

Information technology educators need to help prepare data analysts and scientists who have the skills of a database designer, software programmer, statistician and storyteller. Davenport and Patil (2012) describe the job of a data scientist in more detail. In general, we can prepare three major types of analyses with these new data sources and data manipulation technologies (see Power, 2013a):

1) Retrospective data analyses — using historical data and quantitative tools to understand patterns and results to make inferences about the future. This is the area of business intelligence.

2) Predictive data analyses — using simulation models to generate scenarios based on historical data to understand the future. Predictive means "looking forward" and making known in advance.

3) Prescriptive data analyses - using planned, quantitative analyses of real-time data that may trigger events. Prescriptive analyses recommend actions.

Academic information technology researchers need to define concepts to study them meaningfully and to communicate results. Rigorous research starts with operationalizing concepts. Operationalizing a concept, construct or variable means identifying a valid, quantifiable measure. That is not possible with an amorphous term like Big Data.

A major ongoing challenge for decision support and information technology researchers is identifying use cases and user examples related to analyzing large volumes of semi- and unstructured data. It is important to document what data was used and how it was collected and analyzed for decision support. The SAS website briefly identifies nine possible uses of Big Data with appropriate analytics: 1) analyze millions of shop keeping units (SKUs) to determine optimal prices that maximize profit and clear inventory, 2) recalculate entire risk portfolios in minutes and understand future possibilities to mitigate risk, 3) mine customer data for insights that drive new strategies for customer acquisition, retention, campaign optimization and next-best offers, 4) quickly identify customers who matter the most, 5) generate retail coupons at the point of sale based on the customer's current and past purchases, ensuring a higher redemption rate, 6) send tailored recommendations to mobile devices when customers are in the right location to take advantage of offers, 7) analyze data from social media to detect new market trends and changes in demand, 8) use clickstream analysis and data mining to detect fraudulent behavior, and 9) determine root causes of failures, issues and defects by investigating user sessions, network logs and machine sensors. Many more use cases certainly can be identified and documenting them is especially important.

CONCLUSIONS

A significant technology change has occurred. On March 7, 2012, IDC released a "worldwide Big Data technology and services forecast showing the market is expected to grow from $3.2 billion in 2010 to $16.9 billion in 2015." Aziza, Ehrenberg, Franks (2013), Morris (2012) and others argue that the potential of Big Data for improving our personal lives, helping businesses compete, and governments provide services is unbounded. According to Ehrenberg (2012), "Greater access to data and the technologies for managing and analyzing data are changing the world." Somehow Big Data will lead to better health, better teachers and improved education, and better decision-making. Researchers need to study these claims.

Also, information systems and decision support researchers need to study the implementation of evolving technologies like Hadoop, investigate more use cases and investigate the claims made for using new data sources and new technologies. The advent of new data sources and new processing technologies might indeed lead to beneficial outcomes, but we need to demonstrate that the desired outcome is occurring and that unintended negative consequences are not occurring. For example, there are growing concerns about the misuse of Big Data (cf., Hall, 2013).

An *Economist* (2010) special report cautions us about misanalysis of Big Data. The report explains that "During the recent financial crisis it became clear that banks and rating agencies had been relying on models which, although they required a vast amount of information to be fed in, failed to reflect financial risk in the real world. This was the first crisis to be sparked by big data — and there will be more." Researchers need to examine both misuse and misanalysis of novel data sources.

Big Data is useful only if we use the data in analyses. The term has limited usefulness as a descriptive label for managers or researchers. Big Data is a marketing term and not a technical term. Descriptive terms like unstructured data, process data and machine data are more useful in practice and in research. Information technology researchers need to explore and clearly document business use cases and help prepare professionals to manage and analyze a wide array of data. Database skills with both traditional and newer technologies are especially important to many organizations. The term Big Data seems increasingly meaningless and the expectations for decision support with Big Data may be too high, but extensive digital data can be captured and analyzed, and in many companies there likely exist uses that justify the expense.

Analyzing new data sources is possible and might be helpful, but beware of the hyperbole. Colorful terms like Big Data create excitement and interest in a practical research area, but they can distract from meaningful scientific inquiry by confusing a marketing term with a technical or research term.

ACKNOWLEDGMENTS

The author thanks the anonymous reviewers and Professors Dale Cyphert and Rex Karsten for their comments. Some of the content in this paper appeared in blog posts of the author at DSSResources.com.

REFERENCES

1. Aziza, B. (2013) Big Data "A-Ha" Moment? *Forbes*, February 25 at URL http://www.forbes.com /sites/ciocentral/2013/02/25/big-data-a-ha moment/?goback=%2Egde_4732551_member_218623809 .
2. Brust, A. (2012) Big Data: Defining its definition, *ZDNet*, March 1 at URL http://www.zdnet.com/blog/big-data/big-data-defining-its-definition/109 .
3. Carter, P. (2011) Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO, *IDC Whitepaper*, September at URL http://free-doc-lib.com/book/big-data-analytics-future-architectures-skills-and-roadmaps.pdf
4. Cooper, M. and P. Mell (2012) Tackling Big Data, Federal Computer Security Managers' Forum, June at URL http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/fcsm_june2012_cooper_mell.pdf .
5. Davenport, T. H., P. Barth & R. Bean (2012) How 'Big Data' is Different, *MIT Sloan Management Review*, 54, 1, 22-24.
6. Davenport, T.H. and D. J. Patil (2012) Data Scientist: The Sexiest Job of the 21st Century, *Harvard Business Review*, October, 70-76.
7. De Goes, J. (2013) "'Big data' is dead. What's next?" Venturebeat.com guest blog post, February 22 at URL http://venturebeat.com/2013/02/22/big-data-is-dead-whats-next/?goback=%2Egde_62438_member_217099766
8. Devlin, B. (2013) Big Analytics rather than Big Data, *B-eye-Network* blog, February 5 at URL http://www.b-eye-network.com/blogs/devlin/archives/2013/02/big_analytics_r.php
9. Ehrenberg, R. (2012) What's the big deal about Big Data? *InformationArbitrage.com* blog post, January 19 at URL http://informationarbitrage.com/post/16121669634/whats-the-big-deal-about-big-data.
10. Franks, B. (2013) Taming the Big Data Tidal Wave, Hoboken, NJ: Wiley.
11. Gartner (2012) Big Data at URL http://www.gartner.com/it-glossary/big-data/, accessed April 18, 2013.
12. Gourley, B. (2012) Big Data Defined for 2013: A Definition That Can Help in Your Interaction with the IT Community, December 31 at URL http://smartdatacollective.com/bobgourley/94151/big-data-defined-2013-definition-can-help-your-interaction-it-community.
13. Hall, Robert E. (2013) In Big Data We Hope and Distrust, *DSSResources.COM*, April 14.
14. International Data Corporation (March 7, 2012) IDC releases first worldwide Big Data technology and services market forecast [Press Release]. Retrieved from http://dssresources.com/news/3749.php .
15. Morris, J. (2012) Top 10 categories for Big Data sources and mining technologies, *ZDNet*, July 16 at URL http://www.zdnet.com/top-10-categories-for-big-data-sources-and-mining-technologies-7000000926/ .
16. Power, D. (2011) What is Hadoop? *Decision Support News*, 12, 23, November 13 at URL http://dssresources.com/faq/index.php?action=artikel&id=235.
17. Power, D. (2012) What is analytics? *Decision Support News*, 13, 18, September 16.
18. Power, D. (2013a) Decision Support, Analytics, and Business Intelligence, New York, NY: Business Expert Press.
19. Power, D. (2013b) What is machine data? *Decision Support News*, 14, 02, January 20.
20. Power, D. (2013c) Does the term big data have utility for managers? *Decision Support News*, 14, 06, March 17.
21. Provost, F. and T. Fawcett (2013) Data Science for Business: Fundamental principles of data mining and data-analytic thinking, *O'Reilly*.
22. Roberts, B. (2012) Data Data Everywhere and Not a Drop of Value, http://bryce.vc blog, February at URL http://bryce.vc/post/15300645787/data-data-everywhere-and-not-a-drop-of-value .
23. Scoble, R. (2010) Interview with Cloudera CEO Mike Olson, What is Hadoop?, YouTube video, March 4 at URL http://youtu.be/S9xnYBVqLws.
24. Sorofman, J. (2013) Data, Data Everywhere, *Gartner Blog Network*, February 28 at URL http://blogs.gartner.com/jake-sorofman/data-data-everywhere/ .
25. SAS (2013) Big Data – What Is It? last accessed March 5, 2013 at URL http://www.sas.com/big-data/.
26. Special Report (2010) Data, data everywhere, *Economist*, February 25.
27. Zikopoulos, P., D. deRoos, K. Parasuraman, T. Deutsch, J. Giles, and D. Corrigan (2013) Harness the Power of Big Data: The IBM Big Data Platform, New York, NY: McGraw Hill.