Big Data Issues and Challenges

B

Stephen Kaisler
SHK and Associates, USA & George Washington University, USA

Frank Armour
Kogod School of Business, American University, USA

William Money
School of Business, George Washington University, USA

J. Alberto Espinosa
Kogod School of Business, American University, USA

INTRODUCTION

"Big Data" originally meant the volume of data that could not be processed (efficiently) by traditional database methods and tools (Wikipedia, 2013). Each time a new storage medium was invented, the amount of data accessible exploded because it could be easily accessed. The original definition focused on structured data, but researchers and practitioners have realized that most of the world's information resides in unstructured data, largely as data are text and imagery. Leslie Johnston (2012) stated that "big data can most definitely mean small data files but a lot of them." She notes that the Library of Congress (as of 2012) has over 6 billion files in its Web archive which are all small. Further, she argues, they do not possess exotic data formats, but are simple text files or Excel spreadsheets or what have you. Today, big data refers to data volumes in the range of petabytes (10^{15}) and beyond. Such volumes exceed the capacity of most current on-line storage and processing systems. An emerging critical problem is the tension between the ability to collect and process big data and use the results versus the right to personal privacy.

BACKGROUND

Big Data was originally described by the 3Vs (Laney, 2001), but Kaisler, Armour, Espinosa, and Money (2013) have suggest two more.

Big Data has been often used to represent a large volume of data of one type, such as text or numbers or pixels. Recently, many organizations are creating blended data from data sources with varied types through analysis. These data come from instruments, sensors, Internet transactions, email, social media such as Twitter, YouTube, Reddit, Pinterest, Tumblr, and clickstreams. New data types may be derived through analysis or joining different types of data.

One of the biggest challenges is the day that machine-generated data exceeds the data created by humans. More sensors are being introduced into products, such as bicycles, coffee pots, washing machines, and thermostats that we use in everyday living. The data collected by these sensors is used to create other data which begets even more data and so on. Another example is Twitter tweets, where many original tweets are retweeted automatically based on profiles set up by receiving users. All retweets are stored someplace. In fact, it is likely that retweets are retweeted in a cascading effect that distributes copies around the world.

BIG DATA: THE CHALLENGES

Kaisler, Armour, Money and Espinosa (2013) identified some major challenges confronting users of Big Data. These challenges represent open research problems – some technical, some managerial, and some social and legal. This section discusses these challenges.

DOI: 10.4018/978-1-4666-5888-2.ch035

Table 1. Five Vs of Big Data

V	Description
Data Volume	The amount of data collected and available. It is estimated that over 2.5 Exabytes (10^{18}) of data are created every day as of 2012 (Wikipedia, 2013).
Data Velocity	The rate at which data is accumulated or the speed at which the data arrives, and how quickly it gets purged, how frequently it changes, and how fast it becomes outdated.
Data Variety	The types of data required for analysis, either *structured*, such as RDF files, databases, and Excel tables or *unstructured*, such as text, audio files, and video.
Data Value	The value derived from processing the data that contributes to decision making and problem solving. A large amount of data may be valueless if it is perishable, late, imprecise, or has other weakness or flaw.
Data Veracity	The accuracy, precision and reliability of the data. A data set may have very accurate data with low precision and low reliability based on the collection methods and tools.

Table 2. Selected fields of Big Data impact

Physical Sciences	Astronomy, Particle Physics, Weather Prediction, Signal Processing. CERN processed exabytes of data to determine the existence of the Higgs Boson.
HealthCare	Imaging, Medical Records. HealthCare accounts for 17% of GDP and employs ~11% of the USA's workers. McKinsey and Co estimates about $300 billion in value creation per year (Manyika, Chui et al., 2011).
Artificial Intelligence	Natural Language Processing of Unstructured Text, Computer Vision
Political Science	Agent-based analysis and prediction of regime change
Forensics	Fraud detection, human/drug/CBRN trafficking, Cybersecurity.
Biology	Genomics, Proteonomics, Ecology.
Cultural Studies	Human terrain assessment, cultural geography, demography
Business	Marketing, Social media influence, Retail operations. McKinsey and Co estimates various levers may yield 10 -35% operating margins across different sectors (Manyika, Chui et al., 2011).

Big Data: Impact vs. Value

Big Data has an impact in every field of human endeavour, if the data are available and can be processed. Impact is different from value. Impact helps to advance a field with new knowledge whereas value affects how useful the resulting actionable information is – whether predicting events, making a profit, discovering a new particle, or improving the lot of our fellow humans. Big Data can add value in several ways: (1) It can make information transparent and usable at a higher frequency; (2) As more accurate data is collected, it allows organizations to conduct more controlled experiments to assess efficiency and refine business operations; (3) It can focus attention on narrower segments of the customer community for precisely specifying products and/or services; and (4) given usage data, it can be used for the specification of new products and services. The National Academy of Science (2013) has noted multiple fields where Big Data is already having an impact and adding value.

Big Data: Acquisition and Storage

Jacobs (2009) summarized this challenge very succinctly by noting that it was easier to get the data into a system, than to get it out. He showed that data entry and storage can be handled with processes currently used for relational databases. But, the tools designed for transaction processing that add, update, search for, and retrieve small to large amounts of data are not capable of extracting the huge volumes and cannot be executed in seconds or a few minutes.

The capacity of disk drives seems to be doubling about every 18 months due to new techniques, such as helical recording, that lead to higher density platters. However, disk rotational speed has changed little over the past 20 years, so the bottleneck has shifted – even with large disk caches – from capacity to getting data on and off the disk. In particular, as NAP (2013) noted, if a 100-TByte disk requires mostly random-access, it was not possible t do so in nay reasonable time.

Network communications speed and bandwidth has not kept up with either disk capacity or processor performance. Of the 3Es – exabytes, exaflops, and exabits – only the first two seem attainable within the next ten years. The National Academy of Science (2103) has noted that data volumes on the order of petabytes mean that the data cannot be moved to where the computing is; instead, the analytical processes must be brought to the data.

Going forward, big data and information provenance will become a critical issue. Buneman and Davidson