Business Horizons (2017) 60, 293—303

Available online at www.sciencedirect.com

ScienceDirect

KELLEY SCHOOL OF BUSINESS
INDIANA UNIVERSITY

www.elsevier.com/locate/bushor

Big data: Dimensions, evolution, impacts, and challenges

In Lee

School of Computer Sciences, Western Illinois University, 1 University Circle, Macomb, IL 61455-1390, U.S.A.

KEYWORDS
Big data;
Internet of things;
Data analytics;
Sentiment analysis;
Social network analysis;
Web analytics

Abstract Big data represents a new technology paradigm for data that are generated at high velocity and high volume, and with high variety. Big data is envisioned as a game changer capable of revolutionizing the way businesses operate in many industries. This article introduces an integrated view of big data, traces the evolution of big data over the past 20 years, and discusses data analytics essential for processing various structured and unstructured data. This article illustrates the application of data analytics using merchant review data. The impacts of big data on key business performances are then evaluated. Finally, six technical and managerial challenges are discussed.

1. The day of big data

The emerging technological development of big data is recognized as one of the most important areas of future information technology and is evolving at a rapid speed, driven in part by social media and the Internet of Things (IoT) phenomenon. The technological developments in big data infrastructure, analytics, and services allow firms to transform themselves into data-driven organizations. Due to the potential of big data becoming a game changer, every firm needs to build capabilities to leverage big data in order to stay competitive. IDC (2015) forecasted that the big data technology and services market will grow at a compound annual growth rate of 23.1% over the 2014—2019 period, with annual spending reaching $48.6 billion in 2019.

While structured data is an essential part of big data, more and more data are created in unstructured video and image forms, which traditional data management technologies cannot process adequately. A large portion of data worldwide have been generated by billions of IoT devices such as smart home appliances, wearable devices, and environmental sensors. Gartner (2015) forecasted that 4.9 billion connected objects would be in use in 2015—up 30% from 2014—and will reach 25 billion by 2020. To meet the ever-increasing storage and processing needs of big data, new big data platforms

E-mail address: i-lee@wiu.edu

0007-6813/$ — see front matter
http://dx.doi.org/10.1016/j.bushor.2017.01.004

are emerging, including NoSQL[1] databases as an alternative to traditional relational databases and Hadoop as an open source framework for inexpensive, distributed clusters of commodity hardware.

In this article, I start with a discussion of big data dimensions and trace the evolution of big data since 1995. Then, I illustrate the application of data analytics using a scenario involving merchant review data. In the following section, I discuss impacts of big data on various business performances. Finally, I discuss six technical and managerial challenges: data quality, data security, privacy, data management, investment justification, and shortage of qualified data scientists.

2. Dimensions of big data

Laney (2001) suggested that volume, variety, and velocity are the three dimensions of big data. The 3 Vs have been used as a common framework to describe big data (Chen, Chiang, & Storey, 2012; Kwon, Lee, & Shin, 2014). Here, I describe the 3 Vs and additional dimensions of big data proposed in the computing industry.

Volume refers to the amount of data an organization or an individual collects and/or generates. While currently a minimum of 1 terabyte is the threshold of big data, the minimum size to qualify as big data is a function of technology development. Currently, 1 terabyte stores as much data as would fit on 1,500 CDs or 220 DVDs, enough to store around 16 million Facebook photographs (Gandomi & Haider, 2015). E-commerce, social media, and sensors generate high volumes of unstructured data such as audio, images, and video. New data has been added at an increasing rate as more computing devices are connected to the internet.

Velocity refers to the speed at which data are generated and processed. The velocity of data increases over time. Initially, companies analyzed data using batch processing systems because of the slow and expensive nature of data processing. As the speed of data generation and processing increased, real time processing became a norm for computing applications. Gartner (2015) forecasted that 6.4 billion connected devices would be in use worldwide in 2016 and that the number will reach 20.8 billion by 2020. In 2016, 5.5 million new devices were estimated to be connected every day to collect, analyze, and share data. The enhanced data streaming capability of connected devices will continue to accelerate the velocity.

Variety refers to the number of data types. Technological advances allow organizations to generate various types of structured, semi-structured, and unstructured data. Text, photo, audio, video, clickstream data, and sensor data are examples of unstructured data, which lack the standardized structure required for efficient computing. Semi-structured data do not conform to specifications of the relational database, but can be specified to meet certain structural needs of applications. An example of semi-structured data is Extensible Business Reporting Language (XBRL), developed to exchange financial data between organizations and government agencies. Structured data is predefined and can be found in many types of traditional databases. As new analytics techniques are developed, unstructured data are generated at a much faster rate than structured data and the data type becomes less of an impediment for the analysis.

IBM added *veracity* as a fourth dimension, which represents the unreliability and uncertainty latent in data sources. Uncertainty and unreliability arise due to incompleteness, inaccuracy, latency, inconsistency, subjectivity, and deception in data. Managers do not trust data when veracity issues are prevalent. Customer sentiments are unreliable and uncertain due to subjectivity of human opinions. Statistical tools and techniques have been developed to deal with uncertainty and unreliability of big data with specified confidence levels or intervals.

SAS added two additional dimensions to big data: variability and complexity. *Variability* refers to the variation in data flow rates. In addition to the increasing velocity and variety of data, data flows can fluctuate with unpredictable peaks and troughs. Unpredictable event-triggered peak data are challenging to manage with limited computing resources. On the other hand, investment in resources to meet the peak-level computing demand will be costly due to overall underutilization of the resources. *Complexity* refers to the number of data sources. Big data are collected from numerous data sources. Complexity makes it difficult to collect, cleanse, store, and process heterogeneous data. It is necessary to reduce the complexity with open sources, standard platforms, and real-time processing of streaming data.

Oracle introduced *value* as an additional dimension of big data. Firms need to understand the importance of using big data to increase revenue, decrease operational costs, and serve customers better; at the same time, they must consider the investment cost of a big data project. Data would be low value in their original form, but data analytics will transform the data into a high-value

[1] Interpreted as Not Only SQL