Big data technologies and management: What conceptual modeling can do

Veda C. Storeya, Il-Yeol Songb

J. Mack Robinson College of Business, Dept. of Computer Information Systems, Georgia State University, Atlanta, Georgia
College of Computing & Informatics, Drexel University, Philadelphia, PA 19104, USA

Abstract

The era of big data has resulted in the development and applications of technologies and methods aimed at effectively using massive amounts of data to support decision-making and knowledge discovery activities. In this paper, the five Vs of big data, volume, velocity, variety, veracity, and value, are reviewed, as well as new technologies, including NoSQL databases that have emerged to accommodate the needs of big data initiatives. The role of conceptual modeling for big data is then analyzed and suggestions made for effective conceptual modeling efforts with respect to big data.

1. Introduction

Big data is widely recognized as referring to the very large amounts of data, both structured and unstructured, that organizations are now capable of capturing and attempting to analyze in a meaningful way so that data-driven decision analysis and actionable insights can be obtained. Doing so has required the development of techniques and methods for analysis, new ways to structure data, and interesting applications in science and in management (e.g., [19,8,1]). Although the value of big data has sometimes been challenged, the big data landscape continues to grow [28].

The objective of this paper is to examine the progression of big data in an effort to: identify the challenges that exist; and specify the role that conceptual modeling can play in advancing work in this important area. The next section defines and describes big data and its recognized, inherent characteristics. Then, new and emerging big data technologies are presented before analyzing the specific role that conceptual modeling can play in understanding and advancing research and applications of big data.

2. Big Data

The volume of data has grown exponentially over the past decade, to the point where the management of the data asset by traditional means is no longer possible (Ribeiro et al. 2015). As shown in Fig. 1, big data trends have been enabled by advances in computing technologies, which facilitated the sudden explosion of data from various sources such as the Web, social media, and sensors. The flood of data brought about the emergence of a data-driven paradigm to take advantage of the newly available computing technologies. Big data technologies materialized the data-driven paradigm, making it increasingly sophisticated and useful.

* Corresponding author.
E-mail addresses: VStorey@gsu.edu (V.C. Storey), song@drexel.edu (I.-Y. Song).

http://dx.doi.org/10.1016/j.datak.2017.01.001
0169-023x/© 2017 Elsevier B.V. All rights reserved.

Please cite this article as: V.C. Storey, I.-Y. Song, (2017), http://dx.doi.org/10.1016/j.datak.2017.01.001
Big data refers to the high volume, velocity, and variety of information assets that demand new, innovative forms of processing for enhanced decision making, business insights, and process optimization\cite{23}. As a relatively new concept, the basic notion of big data includes the techniques and technologies required to manage very large quantities of data. In addition to the technologies, skilled professionals are needed with analysis and design skills to appropriately manage this resource\cite{2,21}.

Mayer-Schonberger and Cukier (2013) argue that big data will change the way people live, work and think, although it requires that many obstacles be overcome. The data must be obtained, processed, and effectively used, raising related issues on how big data will be represented and modelled. Understanding the challenges associated with big data representation and modeling, though, first requires an understanding of the characteristics of big data.

2.1. The Vs of big data

Big data as traditionally characterized by the “3Vs” of volume, variety, and velocity, have emerged from advances in sensing, measuring, and social computing technologies (Gartner.com). In addition to these Vs, veracity (accuracy) and, especially, value, are important. Each of the Vs has its own unique challenges. The volume is too big, the variety requires both structured and unstructured analysis, and the velocity is so fast that we might not even have time to identify reasonable questions to ask\cite{14}. The veracity leads to uncertainty, and the volume competes with velocity\cite{34}. It is the value, however, that is the most time-consuming to extract, and difficult to ascertain. Fig. 2 summarizes the “5 V” challenges dominant in big data practice and research efforts.

Volume: The large volume of data has resulted in data availability coming from diverse, often location-dependent, data streams containing various kinds of data that are being generated at a very high velocity from huge banks of physical, digital, and human sensors\cite{16}. The data sources include wearable technologies, cloud-based service (e.g., Amazon web services), enterprise data warehouses (EDW), and NoSQL databases. The scale is now terabytes, petabytes, and exabytes. The volume challenge is being addressed, technologically, by using commodity hardware and the Hadoop Distributed File System (HDFS).

Velocity: The velocity is the speed to create, capture, extract, process, and store data. A semi-technology solution is needed to deal with the velocity challenge, with the software solution portion having real-time processing, streaming and in-memory computing.

Variety: Different data types and sources provide relations (from relational databases), documents, Web data, XML files, sensor data, multimedia files, and so forth. The variety challenge is primarily addressed by software solutions because the integration of heterogeneous data requires an extensive software effort to handle the variety.

Veracity: The veracity refers to the imprecision of the data. It raises issues of quality, reliability, uncertainty, incompleteness, and the meaning in the data itself (e.g., word variation, weather data, translation of hand-written data). Eventually, though, the veracity must be consistent in order to be processed in an automated manner. The veracity challenge should be addressed by software solutions.

Value: The value of big data can be difficult to ascertain. Reasonable measures focus on: identifying “actionable” knowledge; computing, if possible, a return on investment (ROI); identifying relevancy to customers (as evidenced by much research on text mining and sentiment analysis); and other measures. The needed analysis of big data to identify such value may occur in various ways including: traditional SQL-type queries, machine learning techniques, data mining, statistics, optimization, and decision support analysis. The results may be represented in different forms, including traditional, standard and ad hoc report generation, and visualization. The value challenge is most difficult to achieve as its software
دریافت فوری
متن کامل مقاله
امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات