International Journal of Information Management 35 (2015) 137–144

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Beyond the hype: Big data concepts, methods, and analytics

Amir Gandomi*, Murtaza Haider

Ted Rogers School of Management, Ryerson University, Toronto, Ontario M5B 2K3, Canada

ARTICLE INFO

Article history:
Available online 3 December 2014

Keywords:
Big data analytics
Big data definition
Unstructured data analytics
Predictive analytics

ABSTRACT

Size is the first, and at times, the only dimension that leaps out at the mention of big data. This paper attempts to offer a broader definition of big data that captures its other unique and defining characteristics. The rapid evolution and adoption of big data by industry has leapfrogged the discourse to popular outlets, forcing the academic press to catch up. Academic journals in numerous disciplines, which will benefit from a relevant discussion of big data, have yet to cover the topic. This paper presents a consolidated description of big data by integrating definitions from practitioners and academics. The paper's primary focus is on the analytic methods used for big data. A particular distinguishing feature of this paper is its focus on analytics related to unstructured data, which constitute 95% of big data. This paper highlights the need to develop appropriate and efficient analytical methods to leverage massive volumes of heterogeneous data in unstructured text, audio, and video formats. This paper also reinforces the need to devise new tools for predictive analytics for structured big data. The statistical methods in practice were devised to infer from sample data. The heterogeneity, noise, and the massive size of structured big data calls for developing computationally efficient algorithms that may avoid big data pitfalls, such as spurious correlation.

© 2014 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

1. Introduction

This paper documents the basic concepts relating to big data. It attempts to consolidate the hitherto fragmented discourse on what constitutes big data, what metrics define the size and other characteristics of big data, and what tools and technologies exist to harness the potential of big data.

From corporate leaders to municipal planners and academics, big data are the subject of attention, and to some extent, fear. The sudden rise of big data has left many unprepared. In the past, new technological developments first appeared in technical and academic publications. The knowledge and synthesis later seeped into other avenues of knowledge mobilization, including books. The fast evolution of big data technologies and the ready acceptance of the concept by public and private sectors left little time for the discourse to develop and mature in the academic domain. Authors and practitioners leapfrogged to books and other electronic media for immediate and wide circulation of their work on big data. Thus, one finds several books on big data, including *Big Data for Dummies*, but not enough fundamental discourse in academic publications.

The leapfrogging of the discourse on big data to more popular outlets implies that a coherent understanding of the concept and its nomenclature is yet to develop. For instance, there is little consensus around the fundamental question of how big the data has to be to qualify as 'big data'. Thus, there exists the need to document in the academic press the evolution of big data concepts and technologies.

A key contribution of this paper is to bring forth the oft-neglected dimensions of big data. The popular discourse on big data, which is dominated and influenced by the marketing efforts of large software and hardware developers, focuses on predictive analytics and structured data. It ignores the largest component of big data, which is unstructured and is available as audio, images, video, and unstructured text. It is estimated that the analytics-ready structured data forms only a small subset of big data. The unstructured data, especially data in video format, is the largest component of big data that is only partially archived.

This paper is organized as follows. We begin the paper by defining big data. We highlight the fact that size is only one of several dimensions of big data. Other characteristics, such as the frequency with which data are generated, are equally important in defining big data. We then expand the discussion on various types of big data, namely text, audio, video, and social media. We apply the

* Corresponding author. Tel.: +1 416 979 5000x6363.
E-mail addresses: agandomi@ryerson.ca (A. Gandomi), murtaza.haider@ryerson.ca (M. Haider).

http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007
0268-4012/© 2014 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

analytics lens to the discussion on big data. Hence, when we discuss data in video format, we focus on methods and tools to analyze data in video format.

Given that the discourse on big data is contextualized in predictive analytics frameworks, we discuss how analytics have captured the imaginations of business and government leaders and describe the state-of-practice of a rapidly evolving industry. We also highlight the perils of big data, such as spurious correlation, which have hitherto escaped serious inquiry. The discussion has remained focused on correlation, ignoring the more nuanced and involved discussion on causation. We conclude by highlighting the expected developments to realize in the near future in big data analytics.

2. Defining big data

While it is ubiquitous today, however, 'big data' as a concept is nascent and has uncertain origins. Diebold (2012) argues that the term "big data . . . probably originated in lunch-table conversations at Silicon Graphics Inc. (SGI) in the mid-1990s, in which John Mashey figured prominently". Despite the references to the mid-nineties, Fig. 1 shows that the term became widespread as recently as in 2011. The current hype can be attributed to the promotional initiatives by IBM and other leading technology companies who invested in building the niche analytics market.

Big data definitions have evolved rapidly, which has raised some confusion. This is evident from an online survey of 154 C-suite global executives conducted by Harris Interactive on behalf of SAP in April 2012 ("Small and midsize companies look to make big gains with big data," 2012). Fig. 2 shows how executives differed in their understanding of big data, where some definitions focused on what it is, while others tried to answer what it does.

Clearly, size is the first characteristic that comes to mind considering the question "what is big data?" However, other characteristics of big data have emerged recently. For instance, Laney (2001) suggested that *Volume*, *Variety*, and *Velocity* (or the *Three V's*) are the three dimensions of challenges in data management. The Three V's have emerged as a common framework to describe big data (Chen, Chiang, & Storey, 2012; Kwon, Lee, & Shin, 2014). For example, Gartner, Inc. defines big data in similar terms:

> "*Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.*" ("Gartner IT Glossary, n.d.")

Similarly, TechAmerica Foundation defines big data as follows:

> "*Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.*" (TechAmerica Foundation's Federal Big Data Commission, 2012)

We describe the Three V's below.

Volume refers to the magnitude of data. Big data sizes are reported in multiple terabytes and petabytes. A survey conducted by IBM in mid-2012 revealed that just over half of the 1144 respondents considered datasets over one terabyte to be big data (Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012). One terabyte stores as much data as would fit on 1500 CDs or 220 DVDs, enough to store around 16 million Facebook photographs. Beaver, Kumar, Li, Sobel, and Vajgel (2010) report that Facebook processes up to one million photographs per second. One petabyte equals 1024 terabytes. Earlier estimates suggest that Facebook stored 260 billion photos using storage space of over 20 petabytes.

Definitions of big data volumes are relative and vary by factors, such as time and the type of data. What may be deemed big data today may not meet the threshold in the future because storage capacities will increase, allowing even bigger data sets to be captured. In addition, the type of data, discussed under variety , defines what is meant by 'big'. Two datasets of the same size may require different data management technologies based on their type, e.g., tabular versus video data. Thus, definitions of big data also depend upon the industry. These considerations therefore make it impractical to define a specific threshold for big data volumes.

Variety refers to the structural heterogeneity in a dataset. Technological advances allow firms to use various types of structured, semi-structured, and unstructured data. Structured data, which constitutes only 5% of all existing data (Cukier, 2010), refers to the tabular data found in spreadsheets or relational databases. Text, images, audio, and video are examples of unstructured data, which sometimes lack the structural organization required by machines for analysis. Spanning a continuum between fully structured and unstructured data, the format of semi-structured data does not conform to strict standards. Extensible Markup Language (XML), a textual language for exchanging data on the Web, is a typical example of semi-structured data. XML documents contain user-defined data tags which make them machine-readable.

A high level of variety, a defining characteristic of big data, is not necessarily new. Organizations have been hoarding unstructured data from internal sources (e.g., sensor data) and external sources (e.g., social media). However, the emergence of new data management technologies and analytics, which enable organizations to leverage data in their business processes, is the innovative aspect. For instance, facial recognition technologies empower the brick-and-mortar retailers to acquire intelligence about store traffic, the age or gender composition of their customers, and their in-store movement patterns. This invaluable information is leveraged in decisions related to product promotions, placement, and staffing. Clickstream data provides a wealth of information about customer behavior and browsing patterns to online retailers. Clickstream advises on the timing and sequence of pages viewed by a customer. Using big data analytics, even small and medium-sized enterprises (SMEs) can mine massive volumes of semi-structured data to improve website designs and implement effective cross-selling and personalized product recommendation systems.

Velocity refers to the rate at which data are generated and the speed at which it should be analyzed and acted upon. The proliferation of digital devices such as smartphones and sensors has led to an unprecedented rate of data creation and is driving a growing need for real-time analytics and evidence-based planning. Even conventional retailers are generating high-frequency data. Wal-Mart, for instance, processes more than one million transactions per hour (Cukier, 2010). The data emanating from mobile devices and flowing through mobile apps produces torrents of information that can be used to generate real-time, personalized offers for everyday customers. This data provides sound information about customers, such as geospatial location, demographics, and past buying patterns, which can be analyzed in real time to create real customer value.

Given the soaring popularity of smartphones, retailers will soon have to deal with hundreds of thousands of streaming data sources that demand real-time analytics. Traditional data management systems are not capable of handling huge data feeds instantaneously. This is where big data technologies come into play. They enable firms to create real-time intelligence from high volumes of 'perishable' data.

Download English Version:

https://daneshyari.com/en/article/7429105

Download Persian Version:

https://daneshyari.com/article/7429105

Daneshyari.com