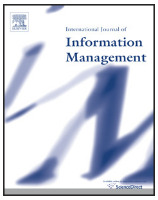
International Journal of Information Management 36 (2016) 1231–1247

Contents lists available at ScienceDirect

International Journal of Information Management

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



Big data: From beginning to future

Ibrar Yaqoob[a,*], Ibrahim Abaker Targio Hashem[a], Abdullah Gani[a,*], Salimah Mokhtar[a], Ejaz Ahmed[a], Nor Badrul Anuar[a], Athanasios V. Vasilakos[b]

[a] *Centre for Mobile Cloud Computing Research (C4MCCR), Faculty of Computer Science and Information Technology, University of Malaya, 50603 Lembah Pantai, Kuala Lumpur, Malaysia*

[b] *Lulea University of Technology, Sweden*

ARTICLE INFO

Article history:
Received 19 November 2014
Received in revised form 29 June 2016
Accepted 31 July 2016

Keywords:
Big data
Parallel and distributed computing
Cloud computing
Internet of things
Social media
Analytics

ABSTRACT

Big data is a potential research area receiving considerable attention from academia and IT communities. In the digital world, the amounts of data generated and stored have expanded within a short period of time. Consequently, this fast growing rate of data has created many challenges. In this paper, we use structuralism and functionalism paradigms to analyze the origins of big data applications and its current trends. This paper presents a comprehensive discussion on state-of-the-art big data technologies based on batch and stream data processing. Moreover, strengths and weaknesses of these technologies are analyzed. This study also discusses big data analytics techniques, processing methods, some reported case studies from different vendors, several open research challenges, and the opportunities brought about by big data. The similarities and differences of these techniques and technologies based on important parameters are also investigated. Emerging technologies are recommended as a solution for big data problems.

Contents

[*] Corresponding authors.

E-mail addresses: ibraryaqoob@siswa.um.edu.my, ibraryaqoob@yahoo.com (I. Yaqoob), targio@siswa.um.edu.my (I.A.T. Hashem), abdullah@um.edu.my (A. Gani), salimah@um.edu.my (S. Mokhtar), imejaz@siswa.um.edu.my (E. Ahmed), badrul@um.edu.my (N.B. Anuar), athanasios.vasilakos@ltu.se (A.V. Vasilakos).

http://dx.doi.org/10.1016/j.ijinfomgt.2016.07.009
0268-4012/© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Since the invention of computers, large amounts of data have been generated at a rapid rate. This condition is the key motivation for current and future research frontiers. Advances in mobile devices, digital sensors, communications, computing, and storage have provided means to collect data (Bryant, Katz, & Lazowska, 2008). According to the renowned IT company Industrial Development Corporation (IDC; 2011), the total amounts of data in the world has increased nine times within five years (Gantz and Reinsel, 2011). This figure is expected to double at least every two years (Chen, Mao, & Liu, 2014). Big data is a novel term that originated from the need of large companies, such as Yahoo, Google, and Facebook, to analyze large amounts of data (Garlasu et al., 2013). Various explanations from 3V Volume, Variety, and Velocity to 4V Volume, Velocity, Variety and Veracity have been provided to define big data (Gandomi & Haider, 2015; Philip Chen & Zhang, 2014; Rodríguez-Mazahua et al., 2015; Hashem et al., 2015).

Doug Laney (presently with Gartner) described big data through three Vs, namely, volume, velocity, and variety. The term volume refers to the size of the data, velocity refers to the speed of incoming and outgoing data, and variety describes the sources and types of data (Philip Chen & Zhang, 2014). IBM and Microsoft added veracity or variability as the fourth V to define big data. The term veracity refers to the messiness and trustworthiness of data. McKinsey & Co. added value as the fourth V to define big data (Chen et al., 2014). Value refers to the worth of hidden insights inside big data. Commonly, big data is a collection of large amounts of complex data that cannot be managed efficiently by the state-of-the-art data processing technologies (Philip Chen & Zhang, 2014).

Off-the-shelf technologies utilized to store and analyze large-scale data cannot operate satisfactorily (Siddiqa et al., 2016). Only advanced data mining and storage techniques can make the storage, management, and analysis of enormous data possible. The major challenges for researchers and practitioners arise from the exponential growth rate of data, which surpasses the current ability of humans to design appropriate data storage and analytic systems to manage large amounts of data effectively (Begoli & Horey, 2012). All the acronyms along with their definitions have been provided in Table 1.

The contributions of this survey are as follows: (a) A broad overview of the genesis of big data applications and its current trends, (b) A discussion of big data processing technologies and methods, (c) A discussion of analysis techniques, (e) We look at different reported case studies (f) We explore opportunities brought about by big data and also discuss some of the research challenges remain to be addressed, (g) A discussion of emerging technologies for big data problems. These contributions are given in separate Sections from 2 to 7 respectively; the conclusion is provided in Section 8.

2. Genesis of big data applications

To get to know the origins of big data applications, we considered the application architecture, chronological development, and gradual evolution of major application models, namely, standalone, desktop, the web, rich Internet, and big data applications (Abolfazli et al., 2014a). We then extrapolated our findings through two paradigms: structuralism and functionalism. These paradigms help analyze, characterize, comprehend, and interpret a phenomenon. "*Structuralism examines the evolution of a phenomenon, compares its structural characteristics, and unveils its limitations while generally maintaining its ontology and epistemology* (Burrell & Morgan, 1997). *Structuralism aims to identify the underlying building blocks of a phenomenon and the relationships among these blocks to better comprehend the phenomenon. Functionalism analyzes the current and future roles and functionalities of a phenomenon in a certain environment to identify its characteristics and behavior* (Burrell & Morgan, 1997)." Five metrics, namely, storage architecture, computing distribution, storage technology, analytics technology, and user experience, are utilized to evaluate these applications. These metrics are discussed below.

- Storage architecture refers to stored data in a computing environment. It offers criteria for data processing operations that can be employed to control the flow of data in the system. It also provides standards for data systems and the interactions between these systems.
- Computing distribution refers to numerous software components located in networked computers that perform as a single system. These computers can be remote from one another and connected by a wide area network or physically close together and connected by a local network.
- Storage technology refers to the location where data is held in an electromagnetic or optical form. Storage technology has changed the landscape of digital media in a profound manner. Most current storage technologies rely on tape backup equipment (e.g., Large Hadron Collider) and software to manage storage systems.